IBM SPSS Modeler Text Analytics 18.1.1 - Guide d'utilisation

Remarque
Avant d'utiliser le présent document et le produit associé, prenez connaissance des informations figurant à la section «Remarques», à la page 235.

Informations produit

LE PRESENT DOCUMENT EST LIVRE EN L'ETAT SANS AUCUNE GARANTIE EXPLICITE OU IMPLICITE. IBM DECLINE NOTAMMENT TOUTE RESPONSABILITE RELATIVE A CES INFORMATIONS EN CAS DE CONTREFACON AINSI QU'EN CAS DE DEFAUT D'APTITUDE A L'EXECUTION D'UN TRAVAIL DONNE.

Ce document est mis à jour périodiquement. Chaque nouvelle édition inclut les mises à jour. Les informations qui y sont fournies sont susceptibles d'être modifiées avant que les produits décrits ne deviennent eux-mêmes disponibles. En outre, il peut contenir des informations ou des références concernant certains produits, logiciels ou services non annoncés dans ce pays. Cela ne signifie cependant pas qu'ils y seront annoncés.

Pour plus de détails, pour toute demande d'ordre technique, ou pour obtenir des exemplaires de documents IBM, référez-vous aux documents d'annonce disponibles dans votre pays, ou adressez-vous à votre partenaire commercial.

Vous pouvez également consulter les serveurs Internet suivants :

- http://www.fr.ibm.com (serveur IBM en France)
- http://www.ibm.com/ca/fr (serveur IBM au Canada)
- http://www.ibm.com (serveur IBM aux Etats-Unis)

Compagnie IBM France
Direction Qualité
17, avenue de l'Europe
92275 Bois-Colombes Cedex

Cette édition s'applique à la version 18.1.1, édition 0, modification 0 d'IBM SPSS Modeler Text Analytics et à toutes les éditions et modifications ultérieures sauf mention contraire dans les nouvelles éditions.

Table des matières

Avis aux lecteurs canadiens

Le présent document a été traduit en France. Voici les principales différences et particularités dont vous devez tenir compte.

Illustrations

Les illustrations sont fournies à titre d'exemple. Certaines peuvent contenir des données propres à la France.

Terminologie

La terminologie des titres IBM peut différer d'un pays à l'autre. Reportez-vous au tableau ci-dessous, au besoin.

IBM France	**IBM Canada**
ingénieur commercial	représentant
agence commerciale	succursale
ingénieur technico-commercial	informaticien
inspecteur	technicien du matériel

Claviers

Les lettres sont disposées différemment : le clavier français est de type AZERTY, et le clavier français-canadien de type QWERTY.

OS/2 et Windows - Paramètres canadiens

Au Canada, on utilise :

- les pages de codes 850 (multilingue) et 863 (français-canadien),
- le code pays 002,
- le code clavier CF.

Nomenclature

Les touches présentées dans le tableau d'équivalence suivant sont libellées différemment selon qu'il s'agit du clavier de la France, du clavier du Canada ou du clavier des États-Unis. Reportez-vous à ce tableau pour faire correspondre les touches françaises figurant dans le présent document aux touches de votre clavier.

France	Canada	Etats-Unis
(Pos1)		Home
Fin	Fin	End
(PgAr)		PgUp
(PgAv)		PgDn
Inser	Inser	Ins
Suppr	Suppr	Del
Echap	Echap	Esc
Attn	Intrp	Break
Impr écran	ImpEc	PrtSc
Verr num	Num	Num Lock
Arrêt défil	Défil	Scroll Lock
(Verr maj)	FixMaj	Caps Lock
AltGr	AltCar	Alt (à droite)

Brevets

Il est possible qu'IBM détienne des brevets ou qu'elle ait déposé des demandes de brevets portant sur certains sujets abordés dans ce document. Le fait qu'IBM vous fournisse le présent document ne signifie pas qu'elle vous accorde un permis d'utilisation de ces brevets. Vous pouvez envoyer, par écrit, vos demandes de renseignements relatives aux permis d'utilisation au directeur général des relations commerciales d'IBM, 3600 Steeles Avenue East, Markham, Ontario, L3R 9Z7.

Assistance téléphonique

Si vous avez besoin d'assistance ou si vous voulez commander du matériel, des logiciels et des publications IBM, contactez IBM direct au 1 800 465-1234.

Préface

IBM® SPSS Modeler Text Analytics propose de puissantes fonctionnalités d'analyse de texte qui utilisent des technologies linguistiques avancées et le traitement du langage naturel (NLP - Natural Language Processing) pour traiter rapidement une grande variété de données texte non structurées et, à partir de ce texte, extraire et organiser les concepts clés. De plus, IBM SPSS Modeler Text Analytics peut regrouper ces concepts par catégories.

Environ 80 % des données détenues dans une organisation se présentent sous la forme de documents de texte, par exemple, des rapports, des pages Web, des messages électroniques et des notes de centre d'appels. Le texte est un facteur clé qui permet à une organisation de mieux comprendre le comportement de ses clients. Les systèmes dotés du traitement du langage naturel extraient les concepts de manière intelligente, y compris les expressions composées. En outre, grâce à la maîtrise du langage sous-jacent, ils classent les termes en groupes d'informations similaires (produits, entreprises ou personnes, par exemple), s'aidant du sens et du contexte. Par conséquent, vous pouvez rapidement savoir si les informations du document présentent un intérêt pour vous. Ces concepts extraits et les catégories peuvent ensuite être combinés aux données structurées existantes, telles que les données démographiques, et appliqués à la modélisation grâce à la gamme complète d'outils de Data mining de IBM SPSS Modeler, afin de favoriser une prise de décision précise et efficace.

Les systèmes linguistiques sont sensibles à la connaissance. Plus leurs dictionnaires contiennent d'informations, plus la qualité des résultats obtenus est élevée. IBM SPSS Modeler Text Analytics est fourni avec un ensemble de ressources linguistiques, telles que des dictionnaires de termes et de synonymes, des bibliothèques et des modèles. Ce produit vous permet également d'approfondir ces ressources linguistiques et de les adapter à votre contexte. La mise au point des ressources linguistiques est souvent un processus itératif nécessaire pour assurer avec précision l'extraction et la catégorisation des concepts. Enfin, des modèles personnalisés, des bibliothèques et des dictionnaires spécialisés dans des domaines précis, tels que la gestion de la relation client et la génomique, sont également fournis.

A propos d'IBM Business Analytics

Les logiciels IBM Business Analytics aident les entreprises à mesurer, comprendre et anticiper leur performance financière et opérationnelle en fournissant des informations exactes, cohérentes et complètes. Un porte-feuilles étendu de veille économique, d'analyses prédictives, de gestion des performances et de stratégie financières et d'applications analytiques vous offre des informations claires, immédiates et décisionnelles sur les performances actuelles et vous permet de prévoir les résultats futurs. Ce logiciel intègre des solutions dédiées à l'industrie, des pratiques éprouvées et des services professionnels qui permettent aux organisations de toute taille de maximiser leur productivité, d'automatiser leurs décisions sans risque et de proposer de meilleurs résultats.

Intégrée dans ce portefeuille, la solution logicielle IBM SPSS Predictive Analytics permet aux entreprises de prévoir les événements et d'agir proactivement en fonction de ces informations, afin d'obtenir de meilleurs résultats. Les clients des secteurs privé, public et universitaire du monde entier font appel à la technologie IBM SPSS, qui les dote d'un atout concurrentiel pour attirer, fidéliser et développer leur clientèle, tout en réduisant les fraudes et en atténuant les risques. L'intégration du logiciel IBM SPSS aux opérations quotidiennes transforme les organisations en entreprises prédictives, capables de guider et d'automatiser leurs décisions de manière à répondre aux objectifs métier et à obtenir un avantage concurrentiel mesurable. Pour plus d'informations ou pour contacter un représentant, visitez le site http://www.ibm.com/spss.

Assistance technique

L'assistance technique est réservée aux clients ayant signé un contrat de maintenance. Les clients peuvent contacter l'assistance technique pour obtenir de l'aide concernant l'utilisation des produits IBM Corp. ou l'installation dans l'un des environnements matériels pris en charge. Pour contacter l'assistance technique, rendez-vous sur le site Web IBM Corp. à l'adresse http://www.ibm.com/support. Votre nom, celui de votre société, ainsi que votre contrat d'assistance vous seront demandés.

Chapitre 1. Présentation d'IBM SPSS Modeler Text Analytics

IBM SPSS Modeler Text Analytics propose de puissantes fonctionnalités d'analyse de texte qui utilisent des technologies linguistiques avancées et le traitement du langage naturel (NLP - Natural Language Processing) pour traiter rapidement une grande variété de données texte non structurées et, à partir de ce texte, extraire et organiser les concepts clés. De plus, IBM SPSS Modeler Text Analytics peut regrouper ces concepts par catégories.

Environ 80 % des données détenues dans une organisation se présentent sous la forme de documents de texte, par exemple, des rapports, des pages Web, des messages électroniques et des notes de centre d'appels. Le texte est un facteur clé qui permet à une organisation de mieux comprendre le comportement de ses clients. Les systèmes dotés du traitement du langage naturel extraient les concepts de manière intelligente, y compris les expressions composées. En outre, grâce à la maîtrise du langage sous-jacent, ils classent les termes en groupes d'informations similaires (produits, entreprises ou personnes, par exemple), s'aidant du sens et du contexte. Par conséquent, vous pouvez rapidement savoir si les informations du document présentent un intérêt pour vous. Ces concepts extraits et les catégories peuvent ensuite être combinés aux données structurées existantes, telles que les données démographiques, et appliqués à la modélisation grâce à la gamme complète d'outils de Data mining de IBM SPSS Modeler, afin de favoriser une prise de décision précise et efficace.

Les systèmes linguistiques sont sensibles à la connaissance. Plus leurs dictionnaires contiennent d'informations, plus la qualité des résultats obtenus est élevée. IBM SPSS Modeler Text Analytics est fourni avec un ensemble de ressources linguistiques, telles que des dictionnaires de termes et de synonymes, des bibliothèques et des modèles. Ce produit vous permet également d'approfondir ces ressources linguistiques et de les adapter à votre contexte. La mise au point des ressources linguistiques est souvent un processus itératif nécessaire pour assurer avec précision l'extraction et la catégorisation des concepts. Enfin, des modèles personnalisés, des bibliothèques et des dictionnaires spécialisés dans des domaines précis, tels que la gestion de la relation client et la génomique, sont également fournis.

Déploiement. Vous pouvez déployer les flux d'exploration de texte à l'aide de IBM SPSS Modeler Solution Publisher pour le scoring en temps réel des données non structurées. Cette possibilité permet de garantir une mise en oeuvre réussie d'opérations d'exploration de texte en boucles fermées. Par exemple, votre organisation peut désormais analyser de manière plus pertinente les notes consignées dans le Bloc-notes, issues des appelants entrants et sortants, en appliquant vos modèles prédictifs. Votre communication marketing en temps réel est ainsi mieux adaptée.

Pour exécuter IBM SPSS Modeler Text Analytics avec IBM SPSS Modeler Solution Publisher, ajoutez le répertoire `<install_directory>/ext/bin/spss.TMWBServer` à la variable d'environnement `$LD_LIBRARY_PATH`.

Remarque : L'adaptateur japonais de IBM SPSS Modeler Text Analytics est obsolète depuis la version 18.1.

Mise à niveau vers IBM SPSS Modeler Text Analytics Version 18.1.1

Mise à niveau à partir des versions précédentes de PASW Text Analytics ou Text Mining for Clementine.

Avant d'installer IBM SPSS Modeler Text Analytics version 18.1.1, enregistrez et exportez tous les TAP, modèles et bibliothèques de votre version actuelle que vous souhaitez utiliser dans la nouvelle version. Nous vous recommandons d'enregistrer ces fichiers dans un répertoire qui ne sera pas effacé ou remplacé lors de l'installation de la nouvelle version.

Après avoir installé la dernière version de IBM SPSS Modeler Text Analytics, vous pouvez charger le fichier TAP enregistré, ajouter toutes les bibliothèques enregistrées ou importer et charger les modèles enregistrés pour les utiliser dans la dernière version.

Important : Si vous désinstallez votre version actuelle sans commencer par enregistrer et exporter les fichiers dont vous avez besoin, le travail effectué dans les TAP, modèles et bibliothèques publiques de la version précédente sera perdu et ne pourra pas être utilisé dans IBM SPSS Modeler Text Analytics version 18.1.1.

A propos de Text Mining

De nos jours, de plus en plus d'informations sont stockées dans des formats non structurés et partiellement structurés (messages électroniques de clients, notes de centre d'appel, réponses ouvertes à des enquêtes, actualités, formulaires Web, etc.). Ce flot d'informations pose problème à de nombreuses organisations qui souhaitent trouver la méthode leur permettant de collecter, d'étudier et d'exploiter ces informations.

Le processus de *Text Mining* consiste à analyser des ensembles de documents textuels afin de capturer les concepts et thèmes-clés, et de découvrir les relations et les tendances cachées. Il ne nécessite pas que vous connaissiez les mots ou les termes précis utilisés par les auteurs pour exprimer ces concepts. Bien qu'il s'agisse de processus très différents, l'exploration de texte est parfois confondue avec la récupération d'informations. Si l'extraction et le stockage précis des informations représentent un défi considérable, l'extraction et la gestion efficaces du contenu, de la terminologie et des relations compris dans ces informations jouent un rôle vital.

Text Mining et Data Mining

Pour chaque élément du texte, le système de Text Mining linguistique renvoie un index de concepts, ainsi que des informations à propos de ces concepts. Ces informations simplifiées et structurées peuvent être combinées à d'autres sources de données afin de répondre aux questions du type :

- Quels concepts sont associés ?
- A quel autre élément sont-ils liés ?
- Quelles sont les catégories de niveau supérieur pouvant découler des informations extraites ?
- Quels résultats les catégories ou les concepts permettent-ils de prédire ?
- De quelle façon les catégories ou les concepts prédisent-ils les comportements ?

Par une utilisation conjointe de Text Mining et de Data mining, vous obtenez des résultats plus probants que sur la base des données structurées ou non structurées seules. Ce processus comprend généralement les étapes suivantes :

1. **Identification du texte à explorer.** Préparation du texte avant exploration. Si le texte apparaît dans plusieurs fichiers, enregistrez-les tous au même endroit. Dans le cas de bases de données, déterminez le champ contenant le texte.
2. **Analyse et extraction des données structurées.** Appliquez les algorithmes de Text Mining au texte source.
3. **Génération de modèles de concept et de catégorie.** Identifiez les principaux concepts et/ou créez des catégories. Généralement, le système renvoie de nombreux concepts à partir de données non structurées. Identifiez les meilleurs concepts et catégories en vue du scoring des catégories.
4. **Analyse des données structurées.** Utilisez les techniques standard du Data mining (comme le clustering, la classification et la modélisation prédictive) pour connaître les relations unissant les concepts. Fusionnez les concepts extraits avec d'autres données structurées afin de prévoir le comportement sur la base des concepts.

Analyse de texte et catégorisation

L'analyse de texte, sorte d'analyse qualitative, est l'extraction d'informations utiles d'un texte, de manière à regrouper les principaux concepts ou idées qui figurent dans ce texte dans un nombre approprié de catégories. Vous pouvez effectuer une analyse de texte sur tout type et toute longueur de texte, bien que l'approche analytique varie quelque peu.

Etant donné que les enregistrements ou les documents courts sont moins complexes et contiennent généralement moins de mots et de réponses ambigus, leur catégorisation est plus simple. Par exemple, si nous posons des questions ouvertes et courtes au cours d'une enquête sur les trois activités préférées des personnes interrogées lorsqu'elles sont en vacances, leurs réponses seront pour la plupart courtes : *aller à la plage*, *visiter des parcs nationaux* ou *ne rien faire*. Des réponses ouvertes plus longues risquent, par contre, d'être plutôt complexes et démesurées, en particulier si les personnes interrogées sont instruites, motivées et qu'elles disposent de suffisamment de temps pour remplir un questionnaire. Si nous interrogeons des personnes sur leurs opinions politiques dans le cadre d'une enquête ou si nous mettons au point un flux de blogue concernant la politique, nous nous attendons à recevoir de très longs commentaires sur une grande variété de problèmes et de prises de position.

La possibilité d'extraire les principaux concepts et de créer des catégories avec pertinence à partir de ces longues sources textuelles en très peu de temps est un avantage-clé de l'utilisation d'IBM SPSS Modeler Text Analytics. Pour obtenir les résultats les plus fiables à chacune des étapes du processus d'analyse de texte, des techniques statistiques et linguistiques automatiques sont associées.

Traitement linguistique et traitement du langage naturel

Le principal problème lié à la gestion de ces données textuelles non structurées est l'absence de règles standard de rédaction permettant aux ordinateurs de comprendre les textes. La langue, et par conséquent le sens des mots, varie d'un document à l'autre et même au sein d'un même document. Pour pouvoir récupérer et organiser efficacement ces données non structurées, vous devez analyser la langue et découvrir la signification du texte. Il existe plusieurs méthodes automatisées permettant l'extraction des concepts d'informations non structurées. Ces méthodes peuvent être réparties en deux types : linguistiques et non linguistiques.

Certaines entreprises ont tenté d'employer des solutions non linguistiques automatisées basées sur des statistiques et des réseaux de neurones. Grâce aux technologies informatiques, ces solutions permettent d'analyser et de catégoriser les principaux concepts plus rapidement qu'un être humain. Le degré d'exactitude de ces solutions est malheureusement relativement faible. La plupart des systèmes basés sur les statistiques comptent simplement le nombre d'occurrences des mots et calculent leur proximité statistiques vis-à-vis des concepts associés. Ils produisent un grand nombre de résultats non pertinents (« bruit ») et passent à côté de ceux qu'ils doivent trouver. On parle alors de « silence ».

Pour compenser leur exactitude limitée, certaines solutions intègrent des règles non linguistiques complexes permettant de distinguer les résultats pertinents des résultats non pertinents. Cette technique est appelée *Text Mining basé sur des règles*.

La technique du *Text Mining basé sur la linguistique* associe les principes de traitement du langage naturel (analyse assistée par ordinateur des langues humaines) et l'analyse des mots, des phrases, de la syntaxe et de la structure du texte. Les systèmes dotés du traitement du langage naturel extraient les concepts de manière intelligente, y compris les expressions composées. En outre, grâce à la maîtrise du langage sous-jacent, ils classent les concepts en groupes d'informations similaires (produits, organisations ou personnes, par exemple), s'aidant du sens et du contexte.

Le Text Mining basé sur la linguistique détermine la signification d'un texte à la manière d'une personne humaine, en reconnaissant un certain nombre de formes de mots comme ayant une signification semblable et en analysant la structure de la phrase de manière à fournir un canevas permettant de

comprendre le texte. Tout en garantissant la rapidité et la rentabilité des systèmes statistiques, cette méthode offre un degré d'exactitude nettement supérieur et exige une intervention considérablement moindre de l'utilisateur.

Pour illustrer la différence entre la méthode statistiques et la méthode linguistique pendant le processus d'extraction, examinons le mode d'action de chacune de ces méthodes dans le cadre d'une requête concernant l'expression `reproduction de documents`. La solution statistiques et la solution linguistique doivent toutes les deux étendre le mot `reproduction` à ses synonymes (`copie` et `duplication`, par exemple). Sinon, des informations pertinentes risquent d'être ignorées. Toutefois, si une solution statistiques tente d'effectuer une recherche sur les synonymes et donc, sur des termes ayant la même signification, elle peut également inclure le terme `naissance`, générant ainsi un certain nombre de résultats non pertinents. Comme la compréhension de la langue permet de lever toute ambiguïté dans le texte, l'exploration de texte linguistique reste par définition la méthode la plus fiable.

Si vous comprenez le fonctionnement du processus d'extraction, vous êtes plus à même de prendre les décisions-clés lorsque vous affinez vos ressources linguistiques (bibliothèques, types, synonymes, etc.). Les principales étapes du processus d'extraction sont les suivantes :

- Conversion des données source en un format standard
- Identification des termes susceptibles d'être extraits
- Identification des classes d'équivalence et intégration des synonymes
- Affectation d'un type
- Indexation et, si nécessaire, mise en correspondance de motifs avec un deuxième analyseur

Etape 1. Conversion des données source en un format standard

Au cours de cette première étape, les données que vous importez sont converties dans un format uniforme pouvant être utilisé pour effectuer d'autres analyses. Cette conversion, qui s'effectue en interne, ne modifie pas les données d'origine.

Etape 2. Identification des termes susceptibles d'être extraits

Il est important de comprendre le rôle des ressources linguistiques dans l'identification des termes susceptibles d'être extraits lors de l'extraction linguistique. Les ressources linguistiques sont utilisées lors de chaque exécution d'une extraction. Elles se présentent sous la forme de ressources compilées, de bibliothèques et de modèles. Les bibliothèques comportent des listes de mots, des relations et des informations complémentaires qui permettent de spécifier ou d'affiner l'extraction. Vous ne pouvez pas afficher ni éditer les ressources compilées. Toutefois, les autres ressources peuvent être modifiées dans l'Editeur de modèle ou, si vous êtes dans une session de plan de travail interactif, dans l'Editeur de ressources.

Les ressources compilées sont des composants internes essentiels du moteur du programme d'extraction de IBM SPSS Modeler Text Analytics. Ces ressources comportent un dictionnaire général qui répertorie les formes de base avec un code concernant la catégorie grammaticale (nom, verbe, adjectif, etc.).

Outre ces ressources compilées, plusieurs bibliothèques sont fournies avec le produit et peuvent être utilisées pour compléter les types et les définitions de concept figurant dans les ressources compilées, ainsi que pour proposer des synonymes. Ces bibliothèques et celles que vous pouvez créer sont constituées de plusieurs dictionnaires. dictionnaires de types, dictionnaires de synonymes et dictionnaires d'exclusions.

Une fois les données importées et converties, le moteur du programme d'extraction commence à identifier les termes susceptibles d'être extraits. Ces termes sont des mots ou des groupes de mots qui permettent d'identifier des concepts du texte. Pendant le traitement du texte, les mots uniques (*unitermes*) et les mots

composés (*multitermes*) sont identifiés à l'aide d'extracteurs de motifs de catégorie grammaticale. Par conséquent, les mots-clés de sentiment susceptibles d'être extraits sont identifiés à l'aide de l'analyse des liens du texte de sentiment.

Remarque : Les termes du dictionnaire général compilé susmentionné représentent une liste de tous les mots susceptibles d'être insignifiants ou linguistiquement ambigus en tant qu'unitermes. Ces mots sont exclus de l'extraction lorsque vous identifiez les unitermes. Ils font toutefois l'objet d'une réévaluation lorsque vous déterminez les catégories grammaticales ou que vous recherchez des mots composés (expressions multitermes) plus longs, susceptibles d'être extraits.

Etape 3. Identification des classes d'équivalence et intégration des synonymes

Une fois les expressions unitermes et multitermes susceptibles d'être extraites identifiées, le logiciel utilise un dictionnaire de normalisation afin d'identifier des classes d'équivalence. Une classe d'équivalence désigne la forme de base d'une expression ou la forme unique de deux variantes d'une même expression. Pour déterminer quel concept utiliser pour la classe d'équivalence le moteur d'extraction applique, dans l'ordre, les règles suivantes :

- Forme définie par l'utilisateur dans une bibliothèque.
- La forme la plus fréquente, comme définie par les ressources précompilées.

Etape 4. Affectation d'un type

Des types sont ensuite affectés aux concepts extraits. Un type correspond à un regroupement sémantique de concepts. Les ressources compilées et les bibliothèques sont utilisées au cours de cette étape. Les types comprennent des éléments tels que des concepts de niveau supérieur, des mots positifs et négatifs, des prénoms, des lieux, des organisations, etc. Pour plus d'informations, voir «Dictionnaires de types», à la page 187.

Les systèmes linguistiques sont sensibles à la connaissance. Plus leurs dictionnaires contiennent d'informations, plus la qualité des résultats obtenus est élevée. Modifier le contenu du dictionnaire, les définitions de synonyme par exemple, permet de simplifier les informations obtenues. Souvent itératif, ce processus est nécessaire pour obtenir une extraction précise des concepts. Le traitement du langage naturel est un élément fondamental d'IBM SPSS Modeler Text Analytics.

Fonctionnement de l'extraction

Au cours de l'extraction des concepts clés et des idées à partir de vos réponses, IBM SPSS Modeler Text Analytics se base sur une analyse de texte linguistique. Cette approche a la même efficacité en temps et en argent que les systèmes statistiques. Mais elle offre un plus grand degré d'exactitude tout en ne nécessitant que peu d'intervention humaine. L'analyse de texte linguistique se base sur un domaine d'étude appelé processus de langage naturel, également connu sous le nom de linguistique computationnelle.

Si vous comprenez le fonctionnement du processus d'extraction, vous êtes plus à même de prendre les décisions-clés lorsque vous affinez vos ressources linguistiques (bibliothèques, types, synonymes, etc.). Les principales étapes du processus d'extraction sont les suivantes :

- Conversion des données source en un format standard
- Identification des termes susceptibles d'être extraits
- Identification des classes d'équivalence et intégration des synonymes
- Affectation d'un type
- Indexation
- Mise en correspondance de l'extraction des motifs et des événements

Etape 1. Conversion des données source en un format standard

Au cours de cette première étape, les données que vous importez sont converties dans un format uniforme pouvant être utilisé pour effectuer d'autres analyses. Cette conversion, qui s'effectue en interne, ne modifie pas les données d'origine.

Etape 2. Identification des termes susceptibles d'être extraits

Il est important de comprendre le rôle des ressources linguistiques dans l'identification des termes susceptibles d'être extraits lors de l'extraction linguistique. Les ressources linguistiques sont utilisées lors de chaque exécution d'une extraction. Elles se présentent sous la forme de ressources compilées, de bibliothèques et de modèles. Les bibliothèques comportent des listes de mots, des relations et des informations complémentaires qui permettent de spécifier ou d'affiner l'extraction. Vous ne pouvez pas afficher ni éditer les ressources compilées. Toutefois, les autres ressources (modèles) peuvent être modifiées dans l'Editeur de modèle ou, si vous êtes dans une session de plan de travail interactif, dans l'Editeur de ressources.

Les ressources compilées sont des composants internes essentiels du moteur du programme d'extraction de IBM SPSS Modeler Text Analytics. Ces ressources comportent un dictionnaire général qui liste les formes de base avec un code concernant la catégorie grammaticale (nom, verbe, adjectif, adverbe, participe, coordinateur, déterminant ou préposition). Les ressources comprennent également des types intégrés et réservés qui permettent d'affecter de nombreux termes extraits aux types suivants : `<Location>`, `<Organization>`, ou `<Person>`. Pour plus d'informations, voir «Types intégrés», à la page 188.

Outre ces ressources compilées, plusieurs bibliothèques sont fournies avec le produit et peuvent être utilisées pour compléter les types et les définitions de concept figurant dans les ressources compilées, ainsi que pour proposer d'autres types et synonymes. Ces bibliothèques et celles que vous pouvez créer sont constituées de plusieurs dictionnaires. Cela inclut des dictionnaires de types, des dictionnaires de substitutions (synonymes et éléments optionnels) et des dictionnaires d'exclusions. Pour plus d'informations, voir Chapitre 15, «Utilisation des bibliothèques», à la page 177.

Une fois les données importées et converties, le moteur du programme d'extraction commence à identifier les termes susceptibles d'être extraits. Ces termes sont des mots ou des groupes de mots qui permettent d'identifier des concepts du texte. Lors du traitement du texte, les mots uniques (*unitermes*) qui ne figurent pas dans les ressources compilées sont considérés comme étant des termes susceptibles d'être extraits. Les mots composés (*multitermes*) susceptibles d'être extraits sont identifiés à l'aide d'extracteurs de motifs de partie du discours. Par exemple, l'expression multiterme `belle voiture`, qui suit le motif de catégorie grammaticale adjectif nom, possède deux composants. L'expression multiterme `belle petite voiture`, qui suit le motif de catégorie grammaticale "adjectif adjectif nom", possède trois composants.

Remarque : Les termes du dictionnaire général compilé susmentionné représentent une liste de tous les mots susceptibles d'être insignifiants ou linguistiquement ambigus en tant qu'unitermes. Ces mots sont exclus de l'extraction lorsque vous identifiez les unitermes. Ils font toutefois l'objet d'une réévaluation lorsque vous déterminez les catégories grammaticales ou que vous recherchez des mots composés (expressions multitermes) plus longs, susceptibles d'être extraits.

Enfin, un algorithme spécial est appliqué pour traiter les chaînes en majuscules (intitulés de postes, par exemple), de telle sorte que ces motifs puissent être extraits.

Etape 3. Identification des classes d'équivalence et intégration des synonymes

Une fois les expressions multitermes et les unitermes susceptibles d'être extraits identifiés, le logiciel utilise un ensemble d'algorithmes pour les comparer et identifier des classes d'équivalence. Une classe d'équivalence désigne la forme de base d'une phrase ou la forme unique de deux variantes d'une même phrase. L'affectation de phrases à des classes d'équivalence a pour objectif de veiller à ce que, par exemple, `président de l'entreprise` et `président d'entreprise` ne soient pas traités comme des

concepts distincts. Pour déterminer quel concept utiliser pour la classe d'équivalence, c'est-à-dire si le terme principal est `président de l'entreprise` ou `président d'entreprise`, le moteur d'extraction applique, dans l'ordre, les règles suivantes :

- Forme définie par l'utilisateur dans une bibliothèque.
- Forme la plus fréquente dans l'ensemble du corps du texte.
- Forme la plus courte dans l'ensemble du corps du texte (ce qui correspond généralement à la forme de base).

Etape 4. Affectation d'un type

Des types sont ensuite affectés aux concepts extraits. Un type correspond à un regroupement sémantique de concepts. Les ressources compilées et les bibliothèques sont utilisées au cours de cette étape. Les types comprennent des éléments tels que des concepts de niveau supérieur, des mots positifs et négatifs, des prénoms, des lieux, des organisations, etc. Vous pouvez définir d'autres types. Pour plus d'informations, voir «Dictionnaires de types», à la page 187.

Etape 5. Indexation

L'indexation de l'ensemble des documents ou des enregistrements s'effectue en définissant un pointeur entre une position de texte et le terme représentatif de chaque classe d'équivalence. Cela suppose que toutes les instances de forme infléchies d'un concept susceptible d'être extrait sont indexées en tant que forme de base susceptible d'être extraite. La fréquence globale est calculée pour chaque forme de base.

Etape 6. Mise en correspondance de l'extraction des motifs et des événements.

IBM SPSS Modeler Text Analytics peut non seulement détecter les types et les concepts, mais également les relations qui existent entre eux. Plusieurs algorithmes et bibliothèques sont fournis avec ce produit ; ils permettent d'extraire les motifs de relations entre les types et les concepts. Ils s'avèrent particulièrement utiles lorsque vous tentez de détecter des opinions spécifiques (relations entre des produits, par exemple) ou les liens relationnels entre des personnes ou des objets (liens entre des groupes politiques ou des génomes, par exemple).

Fonctionnement de la catégorisation

Lorsque vous créez des modèles de catégories dans IBM SPSS Modeler Text Analytics, vous disposez de plusieurs techniques. Etant donné que chaque ensemble de données est unique, le nombre de techniques et leur ordre d'application peuvent varier. Dans la mesure où votre interprétation des résultats peut être différente de celle d'une autre personne, vous pouvez tester les différentes techniques afin de déterminer celle qui génère les meilleurs résultats pour vos données texte. Dans IBM SPSS Modeler Text Analytics, vous pouvez créer des modèles de catégories dans une session de plan de travail, puis, dans un second temps, explorer et affiner ces catégories.

Dans ce manuel, la **génération de catégories** fait référence à la génération de définitions et de classification de catégories à l'aide d'une ou de plusieurs techniques intégrées et la **catégorisation** fait référence au scoring, ou à l'étiquetage, processus par lequel des identificateurs uniques (nom/ID/valeur) sont affectés aux définitions de catégorie de chaque enregistrement ou de chaque document.

Pendant la génération de catégories, les concepts et les types qui ont été extraits sont utilisés en tant que blocs de construction de vos catégories. Lorsque vous créez des catégories, les documents ou les enregistrements sont automatiquement affectés aux catégories s'ils contiennent du texte qui correspond à un élément d'une définition de catégorie.

IBM SPSS Modeler Text Analytics offre plusieurs techniques de génération automatique de catégories afin de vous aider à catégoriser plus rapidement vos documents ou vos enregistrements.

Techniques de regroupement

Chaque technique disponible convient à certains types de données et de situation. Cependant, il est souvent judicieux de combiner des techniques dans la même analyse afin de capturer le maximum de documents ou d'enregistrements. Aussi est-il possible de voir un concept figurer dans plusieurs catégories ou de rencontrer des catégories redondantes.

Dérivation des racines de concept. Cette technique crée des catégories en sélectionnant un concept et en recherchant les autres concepts y étant associés. Pour cela, elle analyse si certains composants du concept sont liés d'un point de vue morphologique ou s'ils partagent des racines. Cette technique est très utile pour l'identification de concepts de mots composés synonymes, car les concepts de chaque catégorie générée sont synonymes ou leur signification est proche. Elle utilise des données de différentes longueurs et génère un nombre inférieur de catégories compactes. Par exemple, le concept `opportunités d'avancer` serait regroupé avec les concepts `opportunité d'avancement` et `opportunité d'un avancement`. Pour plus d'informations, voir «Dérivation des racines de concept», à la page 111.

Réseau sémantique. Cette technique commence par identifier les sens possibles de chaque concept à partir de son index complet de relations existant entre les mots, puis crée des catégories en regroupant les concepts associés. Cette technique est plus performante lorsque les concepts sont connus dans le réseau sémantique et qu'ils ne sont pas trop ambigus. Son efficacité est cependant amoindrie lorsque le texte contient des termes spécialisés dont le réseau n'a pas connaissance. Par exemple, le concept `pomme granny smith` pourrait être regroupé avec `pomme gala` et `pomme golden` car il s'agit de soeurs de la granny smith. Pour donner un autre exemple, le concept `animal` pourrait être regroupé avec `chat` et `kangourou` car il s'agit d'hyponymes d'`animal`. Cette technique est uniquement disponible pour les textes en anglais dans cette édition. Pour plus d'informations, voir «Réseaux sémantiques», à la page 113.

Inclusion de concept. Cette technique crée des catégories en regroupant les concepts multitermes (mots composés) selon qu'ils contiennent ou non des mots qui sont des sous-ensembles ou des super-ensembles d'un mot dans l'autre. Par exemple, le concept `siège` serait regroupé avec `siège de sécurité`, `siège couchette` et `commande de siège éjectable`. Pour plus d'informations, voir «Inclusion de concepts», à la page 112.

Co-occurrence. Cette technique crée des catégories à partir des co-occurrences trouvées dans le texte. Ainsi, lorsque des concepts ou des motifs de concept apparaissent souvent ensemble dans des documents et des enregistrements, la co-occurrence reflète une relation sous-jacente qui a vraisemblablement de l'intérêt dans vos définitions de catégorie. Lorsque des mots font l'objet d'une co-occurrence de manière significative, une règle de co-occurrence est créée et peut être utilisée comme descripteur de catégorie pour une nouvelle sous-catégorie. Par exemple, si de nombreux enregistrements contiennent les mots `prix` et `disponibilité`, (mais que peu d'enregistrements contiennent l'un sans l'autre) alors ces concepts peuvent être regroupés dans une règle de co-occurrence, (`prix & disponible`) et ils peuvent être affectés à une sous-catégorie de la catégorie `prix` par exemple. Pour plus d'informations, voir «Règles de co-occurrence», à la page 114.

Nombre minimum de documents. Pour aider à déterminer l'intérêt des co-ocurrences, déterminez le nombre minimum de documents ou d'enregistrements devant contenir une co-occurrence donnée pour être utilisés en tant que descripteurs dans une catégorie.

Noeuds IBM SPSS Modeler Text Analytics

Outre les nombreux noeuds standard fournis avec IBM SPSS Modeler, vous pouvez utiliser les noeuds Text Mining afin de pouvoir effectuer des analyses de texte poussées dans vos flux. IBM SPSS Modeler Text Analytics met à ce titre plusieurs noeuds à votre disposition. Ces noeuds sont stockés dans l'onglet IBM SPSS Modeler Text Analytics de la palette des noeuds.

Les noeuds suivants sont inclus :

- Le **noeud source Liste fichiers** génère une liste de noms de documents utilisée comme entrée pour le processus de Text Mining. Cela s'avère utile lorsque le texte se trouve dans des documents externes, et non dans une base de données ou un fichier structuré. Le noeud génère un champ unique comportant

un enregistrement pour chaque document ou dossier répertorié, pouvant servir d'entrée dans un noeud Text Mining suivant. Pour plus d'informations, voir «Noeud liste fichiers», à la page 11.

- Grâce au **noeud source Flux de nouvelles**, vous pouvez lire du texte issu de flux de nouvelles, tels que des blogues ou de l'actualité au format RSS ou HTML, et utiliser ces données dans le processus de Text Mining. Le noeud génère un ou plusieurs champs pour chaque enregistrement figurant dans les flux. Ces champs peuvent servir d'entrée dans un noeud Text Mining suivant. Pour plus d'informations, voir «Noeud Flux de nouvelles», à la page 13.
- Le **noeud identificateur de langue** est un noeud de processus qui analyse le texte source pour déterminer dans quelle langue ce dernier est écrit, puis indique cette langue dans un nouveau champ. Initialement conçu pour être utilisé avec de grandes quantités de données, ce noeud est particulièrement utile si vous disposez de sources de données multilingues et que vous voulez traiter une langue seulement. Pour plus d'informations, voir «Noeud Langue», à la page 17.
- Le **noeud Text Mining** applique des méthodes linguistiques pour extraire les principaux concepts du texte, permet de créer des catégories avec ces concepts et d'autres données, et offre la possibilité d'identifier les relations et les associations existant entre les concepts en fonction de motifs connus (analyse des liens du texte). Grâce à ce noeud, vous pouvez explorer le contenu de données textuelles ou générer soit un modèle de concepts, soit un modèle de catégories. Les concepts et les catégories peuvent être combinés avec les données structurées existantes, telles que des données démographiques, et appliqués à la modélisation. Pour plus d'informations, voir «Noeud de modélisation Text Mining», à la page 20.
- Le **noeud analyse des liens du texte** extrait des concepts et identifie également les relations existant entre les concepts en fonction de motifs connus dans le texte. L'extraction de motifs permet de révéler les relations existant entre vos concepts, ainsi que tout qualificatif ou opinion associé à ces concepts. Le noeud analyse des liens du texte propose une méthode plus directe pour identifier les motifs, les extraire du texte et ajouter leurs résultats à l'ensemble de données figurant dans le flux. Cependant, vous pouvez également effectuer une analyse des liens du texte en lançant une session de plan de travail interactif dans le noeud modélisation Text Mining. Pour plus d'informations, voir «Noeud Analyse des liens du texte», à la page 47.
- Lors de l'exploration du texte contenu dans des documents externes, vous pouvez utiliser le **noeud de sortie Text Mining** pour générer une page HTML qui comporte des liens pointant vers les documents desquels les concepts ont été extraits. Pour plus d'informations, voir «Noeud Afficheur de fichiers», à la page 55.

Applications

En règle générale, IBM SPSS Modeler Text Analytics profite à toutes les personnes amenées à rechercher régulièrement des éléments-clés dans de gros volumes de documents.

Voici quelques exemples d'applications :

- **Recherche scientifique et médicale.** Explorer des documents de recherche divers, tels que des rapports sur les brevets, des articles et des publications relatives aux protocoles. Identifier des associations jusque-là inconnues (par exemple, l'association d'un docteur à un produit particulier), ouvrant la voie à de nouvelles explorations. Réduire le délai nécessaire au processus de découverte de médicaments. Utiliser le programme dans le cadre de recherches en génomique.
- **Recherche dans le domaine des investissements.** Passer en revue les rapports d'analyse quotidiens et les articles des journaux afin d'identifier les changements de stratégie et les évolutions du marché. A partir de ces informations, il est possible d'analyser les tendances, et de détecter les problèmes et opportunités que rencontre une société ou un secteur sur une période donnée.
- **Détection des fraudes.** Dans le secteur bancaire et dans le domaine de la santé, ce logiciel peut servir à détecter les anomalies et les alertes dans de gros volumes de texte.
- **Etude de marché.** Dans le domaine de l'étude de marché, cette application permet d'identifier les rubriques essentielles contenues dans les réponses ouvertes formulées à l'occasion d'enquêtes.

- **Analyse de flux de nouvelles et de blogues.** Cette application permet de générer et d'explorer des modèles en s'appuyant sur les principales idées figurant dans l'actualité, les blogues, etc.
- **CRM.** Créer des modèles sur la base des données issues de l'ensemble des points de communication avec la clientèle (messages électroniques, transactions et enquêtes).

Chapitre 2. Lecture du texte source

Dans l'exploration de texte, les données utilisées peuvent se trouver dans tout format standard utilisé par IBM SPSS Modeler, notamment les bases de données et autres formats dits "rectangulaires", qui représentent les données sous forme de lignes et de colonnes, ainsi que dans les formats de document (par exemple, les formats Microsoft Word, Adobe PDF ou HTML), qui n'obéissent pas à cette structure.

- Pour lire du texte à partir de documents qui n'obéissent pas à la structure de données standard (Microsoft Word , Microsoft Excel et Microsoft PowerPoint, en plus des formats Adobe PDF, XML et HTML, entre autres), vous pouvez utiliser le noeud liste fichiers afin de générer une liste de documents ou de dossiers qui servira d'entrée au processus Text Mining. Pour plus d'informations, voir «Noeud liste fichiers».
- Pour lire du texte à partir de flux de nouvelles, tels que des blogues ou des actualités au format RSS ou HTML, il est possible d'utiliser le noeud Flux de nouvelles afin de formater des données de flux de nouvelles pour les convertir en entrée dans le processus Text Mining. Pour plus d'informations, voir «Noeud Flux de nouvelles», à la page 13.
- Pour lire un texte à partir d'un format de données standard utilisé par SPSS Modeler, tel qu'une base de données comprenant un ou plusieurs champs de texte dédiés aux commentaires des clients, vous pouvez utiliser n'importe lequel des noeuds source SPSS Modeler. Pour plus d'informations, reportez-vous à la documentation sur les noeuds SPSS Modeler.
- Si vous traitez de grandes quantités de données, qui peuvent inclure du texte dans différentes langues, utilisez le noeud Langue pour identifier la langue utilisée dans un champ spécifique. Pour plus d'informations, voir «Noeud Langue», à la page 17.

Noeud liste fichiers

Pour lire du texte à partir de documents non structurés, enregistrés dans des formats tels que Microsoft Word, Microsoft Excel et Microsoft PowerPoint, mais aussi Adobe PDF, XML et HTML (entre autres), vous pouvez utiliser le noeud liste fichiers afin de générer une liste de documents ou de dossiers qui servira d'entrée au processus de Text Mining. Ceci est nécessaire car les documents texte non structurés ne peuvent pas être représentés par des champs et des enregistrements (lignes et colonnes), comme c'est le cas des autres données utilisées par IBM SPSS Modeler.

Le noeud liste fichiers fonctionne comme un noeud source.

Vous pouvez trouver ce noeud dans l'onglet IBM SPSS Modeler Text Analytics de la palette de noeuds en bas de la fenêtre IBM SPSS Modeler. Pour plus d'informations, voir «Noeuds IBM SPSS Modeler Text Analytics», à la page 8.

Important : Les noms de répertoire et de fichier contenant des caractères qui ne sont pas inclus dans le codage local de la machine ne sont pas pris en charge. Lorsque vous essayez d'exécuter un flux contenant un noeud liste fichiers, les noms de fichier ou de répertoire contenant ces caractères font échouer l'exécution du flux. Cela peut arriver avec des noms de répertoire ou de fichier en langues étrangères, comme un nom de fichier allemand dans un paramètre local français.

Prise en charge des données locales. Si vous êtes connecté à un serveur IBM SPSS Modeler Text Analytics Server distant et que vous avez un flux avec un noeud liste fichiers, les données doivent résider sur le même ordinateur qu'IBM SPSS Modeler Text Analytics Server . Vous pouvez également vous assurer que le serveur a accès au dossier dans lequel les données source du noeud liste fichiers sont stockées.

Remarque : Vous pouvez utiliser le noeud Liste de fichiers pour le scoring dans une configuration IBM SPSS Collaboration and Deployment Services - Scoring.

Noeud Liste fichiers : onglet Paramètres

Dans cet onglet, vous définissez les répertoires, les extensions de fichier et l'entrée de ce noeud.

Remarque : L'extraction ne traite pas les fichiers Microsoft Office et Adobe PDF sous des plateformes non-Microsoft Windows. Néanmoins, les fichiers XML, HTML et les fichiers texte peuvent toujours être traités.

Les noms de répertoire et de fichier contenant des caractères qui ne sont pas inclus dans le codage local de la machine ne sont pas pris en charge. Lorsque vous essayez d'exécuter un flux contenant un noeud liste fichiers, les noms de fichier ou de répertoire contenant ces caractères font échouer l'exécution du flux. Cela peut arriver avec des noms de répertoire ou de fichier en langues étrangères, comme un nom de fichier allemand dans un paramètre local français.

Répertoire : spécifie le dossier racine qui contient les documents à répertorier.

- **Inclure les sous-répertoires** : indique que les sous-répertoires doivent également être analysés.

Types de fichier à inclure dans la liste : Vous pouvez sélectionner ou désélectionner les types de fichiers et les extensions que vous voulez utiliser. Si une extension de fichier est désélectionnée, les fichiers présentant cette extension sont ignorés. Vous pouvez appliquer un filtre sur les extensions suivantes :

Tableau 1. Filtres de type de fichier par extension.

• *.rtf, .doc, .docx, .docm*	• *.xls, .xlsx, .xlsm*	• *.ppt, .pptx, .pptm*	• *.txt, .text*
• *.htm, .html, .shtml*	• *.xml*	• *.pdf*	• *.$*

Remarque : Pour plus d'informations, voir «Noeud liste fichiers», à la page 11.

si vous avez des noms de fichiers sans extension ou se terminant par un point (par exemple, `Fichier01` ou `Fichier01.`)utilisez l'option **Aucune extension** pour les sélectionner.

Codage des entrées Si la zone de sortie doit contenir du texte exact, choisissez la valeur appropriée dans la liste suivante :

- Automatique (européen)
- UTF-8
- UTF-16
- ISO-8859-1
- ISO-8859-2
- Windows-1250
- US ASCII

La sortie est affichée au format texte UTF-8.

Important : Depuis la version 14, l'option **Liste des répertoires** n'est plus disponible ; la seule sortie possible est une liste de fichiers.

Noeud Liste fichiers : autres onglets

L'onglet Types est un onglet standard dans les noeuds IBM SPSS Modeler, tout comme l'onglet Annotations.

Utilisation du noeud Liste fichiers dans Text Mining

Le noeud Liste fichiers est utilisé lorsque les données textuelles résident dans des documents non structurés externes dans des formats tels que Microsoft Word, Microsoft Excel et Microsoft PowerPoint, ainsi que Adobe PDF, XML et HTML (entre autres).

Supposons que l'on connecte un noeud Liste fichiers à un noeud Text Mining afin de fournir des données texte depuis des documents externes :

1. **Noeud Liste fichiers (onglet Paramètres).** Nous avons tout d'abord ajouté ce noeud au flux pour indiquer l'emplacement de stockage des documents texte. Nous avons sélectionné le répertoire contenant tous les documents sur lesquels nous souhaitons effectuer l'exploration de texte.
2. **Noeud Text Mining (onglet Champs).** Nous avons ensuite ajouté et connecté un noeud Text Mining au noeud liste fichiers. Dans ce noeud, nous avons défini le format d'entrée, le modèle de ressources et le format de sortie. Nous avons choisi le nom de champ créé à partir du noeud liste fichiers, le champ de texte et d'autres paramètres. Pour plus d'informations, voir «Utilisation du noeud Text Mining dans un flux», à la page 30.

Pour plus d'informations sur l'utilisation du noeud Text Mining, voir «Noeud de modélisation Text Mining», à la page 20.

Noeud Flux de nouvelles

Vous pouvez utiliser le noeud Flux de nouvelles afin de préparer des données textuelles à partir de flux de nouvelles pour le processus de Text Mining. Ce noeud accepte les flux de nouvelles dans les deux formats suivants :

- Format RSS. RSS est un format normalisé simple, basé sur le langage XML et destiné au contenu Web. Pour ce format, l'URL pointe vers une page qui présente un ensemble de liens vers d'autres articles tels que les sources d'actualité et les blogues publiés. Le format RSS étant normalisé, chaque lien d'article est automatiquement identifié et traité comme un enregistrement séparé dans le flux de données obtenu. Aucune autre donnée n'est nécessaire pour que vous puissiez identifier les données textuelles et les enregistrements importants du flux de nouvelles sauf si vous souhaitez appliquer une technique de filtrage au texte.
- Format HTML. Vous pouvez définir une ou plusieurs URL vers des pages HTML dans l'onglet Entrée. Ensuite, dans l'onglet Enregistrements, définissez la balise de début d'enregistrement et identifiez également les balises qui délimitent le contenu cible et assignez ces balises aux champs de sortie de votre choix (description, titre, date modifiée, etc.) Pour plus d'informations, voir «Noeud Flux de nouvelles : onglet Enregistrements», à la page 14.

Important ! Si vous essayez de récupérer des informations sur le Web avec un serveur proxy, vous devez activer ce serveur dans le fichier `net.properties` pour le serveur et le client IBM SPSS Modeler Text Analytics. Suivez les instructions détaillées dans ce fichier. Cette procédure s'applique lorsque vous accédez au Web via le noeud Fil de nouvelles ou lorsque vous récupérez une licence SDL Software as a Service (SaaS), car ces connexions s'effectuent via Java™. Ce fichier est situé par défaut dans *C:\Program Files\IBM\SPSS\Modeler\18.1.1\jre\lib\net.properties*.

La sortie de ce noeud est un ensemble de champs utilisés pour décrire les enregistrements. Le champ **Description** est le plus fréquemment utilisé car il contient la plus grosse partie du contenu textuel. Toutefois, d'autres champs sont intéressants, tels que la description courte de l'enregistrement (champ **Description courte**) ou le titre de l'enregistrement (champ **Titre**). Il est possible de sélectionner n'importe quel champ de sortie en tant qu'entrée pour un noeud Text Mining suivant.

Remarque : Vous ne pouvez pas utiliser le noeud Flux de nouvelles pour le scoring dans une configuration IBM SPSS Collaboration and Deployment Services - Scoring.

Vous pouvez trouver ce noeud dans l'onglet IBM SPSS Modeler Text Analytics de la palette de noeuds en bas de la fenêtre IBM SPSS Modeler. Pour plus d'informations, voir «Noeuds IBM SPSS Modeler Text Analytics», à la page 8.

Noeud Flux de nouvelles : onglet Entrée

L'onglet Entrée permet de spécifier une ou plusieurs adresses Web, ou URL, afin de capturer les donnés textuelles. Dans le contexte du processus de Text Mining, vous pouvez indiquer des URL pour des flux qui contiennent des données textuelles.

Important : Si vous travaillez avec des données non-RSS, vous préférerez peut-être utiliser un outil d'extraction de contenu Web tel que WebQL®, afin d'automatiser la collecte de contenu et référencer le résultat en utilisant un noeud source différent.

Vous pouvez définir les paramètres suivants :

Saisir ou coller les URL. Dans ce champ, vous pouvez entrer ou coller une ou plusieurs URL. Dans le cas de plusieurs URL, entrez une seule URL par ligne et utilisez la touche **Entrée/Retour** pour séparer les lignes. Saisissez le chemin de l'URL vers le fichier. Ces URL peuvent correspondre à des flux apparaissant dans l'un des deux formats suivants :

- `Format RSS.` RSS est un format normalisé simple, basé sur le langage XML et destiné au contenu Web. Pour ce format, l'URL pointe vers une page qui présente un ensemble de liens vers d'autres articles tels que les sources d'actualité et les blogues publiés. Le format RSS étant normalisé, chaque lien d'article est automatiquement identifié et traité comme un enregistrement séparé dans le flux de données obtenu. Aucune autre donnée n'est nécessaire pour que vous puissiez identifier les données textuelles et les enregistrements importants du flux de nouvelles sauf si vous souhaitez appliquer une technique de filtrage au texte.
- `Format HTML.` Vous pouvez définir une ou plusieurs URL vers des pages HTML dans l'onglet Entrée. Ensuite, dans l'onglet Enregistrements, définissez la balise de début d'enregistrement et identifiez également les balises qui délimitent le contenu cible et assignez ces balises aux champs de sortie de votre choix (description, titre, date modifiée, etc.) Si vous travaillez avec des données non-RSS, vous préférerez peut-être utiliser un outil d'extraction de contenu Web tel que WebQL®, afin d'automatiser la collecte de contenu et référencer le résultat en utilisant un noeud source différent. Pour plus d'informations, voir «Noeud Flux de nouvelles : onglet Enregistrements».

Nombre des entrées les plus récentes à lire (par URL). Ce champ spécifie le nombre maximal d'enregistrements à lire pour chaque URL répertoriée, en commençant par le premier enregistrement détecté dans le flux. La quantité de texte influe sur la vitesse de traitement pendant l'extraction dans un noeud Text Mining ou Analyse des liens du texte, en aval.

Si possible, enregistrer et réutiliser les flux de nouvelles précédents. Cette option permet d'analyser les flux de nouvelles et de mettre en cache les résultats traités. Puis, après plusieurs exécutions de flux, si le contenu d'un flux donné reste inchangé ou si le fil est inaccessible (interruption d'Internet, par exemple), la version mise en cache est utilisée pour accélérer le temps de traitement. Tout nouveau contenu détecté dans ces flux est également mis en cache pour la prochaine exécution du noeud.

- **Libellé**. Si vous sélectionnez l'option **Si possible, enregistrer et réutiliser les flux de nouvelles précédents**, vous devez spécifier un nom de libellé pour les résultats. Ce libellé permet de décrire les flux mis en cache sur le serveur. Si aucun libellé n'est spécifié ou si le libellé n'est pas reconnu, la réutilisation n'est pas possible. Vous pouvez gérer ces caches de flux de nouvelles dans la table de session d'IBM SPSS Text Analytics Administration Console incluse dans IBM SPSS Deployment Manager. Pour plus d'informations, reportez-vous au guide utilisateur de Deployment Manager.

Noeud Flux de nouvelles : onglet Enregistrements

L'onglet Enregistrements permet de définir le contenu textuel de flux non-RSS en identifiant le début de chaque nouvel enregistrement, ainsi que d'autres informations intéressantes concernant chaque

enregistrement. Si vous savez qu'un flux non-RSS (HTML) contient du texte se trouvant dans plusieurs enregistrements, vous devez identifier ici la balise de début d'enregistrement ou bien le texte sera traité comme enregistrement indépendant. Pendant la normalisation des flux RSS qui ne nécessitent aucune spécification de balise dans cet onglet, vous pouvez prévisualiser le contenu dans l'onglet Aperçu.

Important ! Si vous travaillez avec des données non-RSS, vous préférerez peut-être utiliser un outil d'extraction de contenu Web tel que WebQL®, afin d'automatiser la collecte de contenu et référencer le résultat en utilisant un noeud source différent.

URL. Cette liste déroulante contient la liste des URL entrées dans l'onglet Entrée. Sont présents à la fois des flux au format HTML et RSS. Si l'adresse URL est trop longue pour la liste déroulante, elle est automatiquement coupée au milieu et le texte coupé est remplacé par des points de suspension, par exemple *http://www.ibm.com/exemple/début-de-l'adresse...reste-de-l'adresse/chemin.htm*.

- Avec les **flux au format HTML**, si le fil contient plusieurs enregistrements (ou entrées), vous pouvez déterminer les balises HTML qui contiennent les données correspondant au champ indiqué dans le tableau. Par exemple, vous pouvez définir la balise de début (qui indique le commencement d'un nouvel enregistrement), une balise de date modifiée ou un nom d'auteur.
- Avec les **flux au format RSS**, il ne vous est pas demandé d'entrer des balises car le format RSS est normalisé. Cependant, vous pouvez prévisualiser les exemples de résultats dans l'onglet Aperçu si nécessaire. Tous les flux RSS reconnus sont précédés de l'image du logo RSS.

Onglet Source. Dans cet onglet, vous pouvez visualiser le code source des flux HTML. Ce code n'est pas modifiable. Vous pouvez utiliser le champ Rechercher pour localiser des balises ou des informations spécifiques dans cette page ; vous pouvez ensuite copier et coller ces dernières dans le tableau situé en dessous. Le champ Rechercher ne distingue pas les majuscules des minuscules et fournit des correspondances partielles.

Onglet Aperçu. Dans cet onglet, vous pouvez prévisualiser la façon dont un enregistrement est lu par le noeud Flux de nouvelles. Ceci est particulièrement utile pour les flux HTML car vous pouvez modifier ces catégories de lecture en définissant des balises HTML dans le tableau situé sous l'onglet Aperçu.

Balise de début d'enregistrement non RSS. Cette option ne s'applique qu'aux flux non RSS. Si votre flux HTML contient beaucoup de texte que vous souhaitez séparer en plusieurs enregistrements, indiquez ici la balise HTML qui signalera le début d'un enregistrement (un article ou un billet de blogue par exemple). Si vous ne définissez pas de balise de début pour un flux non RSS, la totalité de la page est traitée en tant qu'enregistrement unique, la totalité du contenu apparaît dans le champ **Description** et la date d'exécution du noeud est utilisée à la fois comme **date de modification** et **date de publication**.

Table des zones. Cette option ne s'applique qu'aux flux non RSS. Dans cette table, vous pouvez découper le contenu textuel en champs de sortie spécifiques en saisissant une balise de début pour chacun des champs de sortie prédéfinis. Entrez uniquement la balise de début. Toutes les correspondances sont obtenues via l'analyse du code HTML et la mise en correspondance du contenu de la table et des noms et attributs de balise détectés dans le code HTML. Vous pouvez utiliser les boutons situés sous la table pour copier les balises définies et les réutiliser pour d'autres flux.

Tableau 2. Champs de sortie possibles pour flux non RSS (formats HTML)

Nom du champ de sortie	Contenu de balise prévu
Titre	La balise délimitant le titre de l'enregistrement (facultatif).
Description courte	La balise délimitant la description courte ou le libellé (facultatif).
Description	La balise délimitant le texte principal. Si ce champ n'est pas renseigné, il inclura le contenu de la balise `<body>` (s'il existe un enregistrement unique) ou le contenu détecté au sein de l'enregistrement actuel (lorsqu'un séparateur d'enregistrement a été spécifié).
Auteur	La balise délimitant l'auteur du texte (facultatif).

Tableau 2. Champs de sortie possibles pour flux non RSS (formats HTML) (suite)

Nom du champ de sortie	Contenu de balise prévu
Collaborateurs	La balise délimitant les noms des collaborateurs (facultatif).
Date de publication	La balise délimitant la date de publication du texte. Si ce champ n'est pas renseigné, il contiendra la date de lecture des données par le noeud.
Date modifiée	La balise délimitant la date de modification du texte. Si ce champ n'est pas renseigné, il contiendra la date de lecture des données par le noeud.

Lorsque vous entrez une balise dans le tableau, le flux est analysé à l'aide de cette balise, en vue d'obtenir une correspondance minimale plutôt qu'une correspondance exacte. Ainsi, si vous entrez `<div>` dans le champ Titre, la correspondance est établie avec toutes les balises `<div>` du flux, y compris celles comprenant des attributs spécifiques (par exemple, `<div class="post three">`), de sorte que `<div>` est égal à la balise racine (`<div>`) et à toutes les dérivées incluant un attribut et utilisant ce contenu pour le champ de sortie Titre. Si vous entrez une balise racine, tous les autres attributs sont également inclus.

Tableau 3. Les exemples de balises HTML utilisés identifient le texte des champs de sortie.

Si vous entrez :	Cela renvoie :	Et également :	Mais cela ne renvoie pas :
`<div>`	`<div>`	`<div class="post">`	toutes les autres balises
`<p class="auth">`	`<p class="auth">`	`<p color="black" class="auth" id="85643">`	`<p color="black">`

Noeud Flux de nouvelles : onglet Filtrer le contenu

L'onglet Filtrer le contenu permet d'appliquer une technique de filtre au contenu des flux RSS. Cet onglet ne s'applique pas aux flux HTML. Le filtrage peut être utile lorsque le flux contient beaucoup de texte sous forme d'en-têtes, de bas de page, de menus, de publicités, etc. Vous pouvez utiliser cet onglet pour supprimer les balises HTML, JavaScript et des lignes ou des mots courts du contenu.

Filtrage de contenu. Si vous ne souhaitez pas appliquer de technique de nettoyage, sélectionnez **Aucun**. Sinon, sélectionnez **Nettoyeur de contenu RSS**.

Options du nettoyeur de contenu RSS. Si vous sélectionnez **Nettoyeur de contenu RSS**, vous pouvez choisir d'ignorer des lignes basées sur certains critères. Une ligne est délimitée par une balise HTML telle que `<p>` et `<li>` mais pas par des balises incorporées telles que `<span>`, `<b>` et `<font>`. Veuillez noter que les balises `<br>` sont exécutées comme des sauts de ligne.

- **Ignorer les lignes courtes**. Cette option ignore les lignes qui ne contiennent pas le **nombre de mots minimum** défini ici.
- **Ignorer les lignes avec des mots courts**. Cette option ignore les lignes qui contiennent plus que **la longueur de mot minimum moyenne** définie ici.
- **Ignorer les lignes avec beaucoup de mots à caractère unique**. Cette option ignore les lignes qui contiennent plus d'une certaine **proportion de mots à caractère unique**.
- **Ignorer les lignes contenant des balises spécifiques**. Cette option ignore le texte des lignes contenant une des balises spécifiées dans ce champ.
- **Ignorer les lignes contenant du texte spécifique**. Cette option ignore le texte des lignes contenant une partie du texte spécifié dans ce champ.

Utilisation du noeud Flux de nouvelles dans le processus de Text Mining

Vous pouvez utiliser le noeud Flux de nouvelles afin de préparer des données textuelles à partir de flux de nouvelles Internet pour le processus de Text Mining. Ce noeud accepte les flux de nouvelles au format HTML ou RSS. Ces flux servent d'entrée au processus de Text Mining (un noeud Text Mining ou Analyse des liens du texte suivant).

Si vous utilisez le noeud Flux de nouvelles, veillez à spécifier que le champ de texte correspond au **texte réel** dans le noeud Text Mining ou Analyse des liens du texte pour indiquer que ces flux conduisent directement à chaque article ou billet de blogue.

Important ! Si vous essayez de récupérer des informations sur le Web avec un serveur proxy, vous devez activer ce serveur dans le fichier `net.properties` pour le serveur et le client IBM SPSS Modeler Text Analytics. Suivez les instructions détaillées dans ce fichier. Cette procédure s'applique lorsque vous accédez au Web via le noeud Fil de nouvelles ou lorsque vous récupérez une licence SDL Software as a Service (SaaS), car ces connexions s'effectuent via Java. Ce fichier est situé par défaut dans *C:\Program Files\IBM\SPSS\Modeler\18.1.1\jre\lib\net.properties*.

Exemple : noeud Flux de nouvelles (flux RSS) avec noeud modélisation Text Mining

Supposons que l'on connecte un noeud Flux de nouvelles à un noeud Text Mining afin de fournir des données texte depuis un flux RSS vers un processus Text Mining.

1. **Noeud de flux de nouvelles (onglet Entrée).** Nous avons tout d'abord ajouté ce noeud au flux afin d'indiquer l'emplacement du contenu du flux et de vérifier la structure du contenu. Dans le premier onglet, nous avons fourni l'URL d'un flux RSS. Etant donné que notre exemple concerne un flux RSS, le formatage est déjà défini et il n'est pas nécessaire d'apporter des modifications dans l'onglet Enregistrements. Un algorithme de filtrage de contenu facultatif est disponible pour les flux RSS mais n'a pas été utilisé dans cet exemple.
2. **Noeud Text Mining (onglet Champs).** Ensuite, nous avons ajouté et connecté un noeud Text Mining au noeud Flux de nouvelles. Dans cet onglet, nous avons défini la sortie du champ de texte par un noeud de flux de nouvelles. Dans ce cas, nous voulions utiliser le champ **Description**. Nous avons également sélectionné l'option champ de texte correspond au **texte réel**, ainsi que d'autres paramètres.
3. **Noeud Text Mining (onglet Modèle).** Ensuite, dans l'onglet Modèle, nous avons choisi le mode et les ressources de génération. Dans cet exemple, nous avons choisi de créer un modèle de concepts directement à partir de ce noeud à l'aide du modèle de ressources par défaut.

Pour plus d'informations sur l'utilisation du noeud Text Mining, voir «Noeud de modélisation Text Mining», à la page 20.

Noeud Langue

Vous pouvez utiliser le noeud Langue pour identifier la langue naturelle d'un champ de texte dans vos données source.

La sortie de ce noeud est un champ dérivé qui contient le code de langue détecté.

Remarque : Vous ne pouvez pas utiliser le noeud Langue pour le scoring dans une configuration IBM SPSS Collaboration and Deployment Services - Scoring.

Vous pouvez trouver ce noeud dans l'onglet IBM SPSS Modeler Text Analytics de la palette de noeuds en bas de la fenêtre IBM SPSS Modeler. Pour plus d'informations, voir «Noeuds IBM SPSS Modeler Text Analytics», à la page 8.

Noeud Langue : onglet Paramètres

Dans cet onglet, vous indiquez comment générer les détails linguistiques d'un champ texte sélectionné.

Champ Texte Sélectionnez le champ texte dont vous souhaitez identifier la langue.

Dériver champ Entrez un nom pour le champ dérivé qui contiendra le code de langue détecté. La valeur par défaut est *Langue*.

Valeur par défaut si la langue ne peut pas être identifiée Spécifiez le nom du champ à créer si la langue ne peut pas être identifiée. Les options disponibles sont les suivantes :

- **Non défini** Si cette option est sélectionnée, le champ dérivé contient des valeurs null.
- **Pris en charge** Si cette option est sélectionnée, vous pouvez choisir l'une des langues ISO suivantes prises en charge :
  - Anglais (EN)
  - Allemand (DE)
  - Espagnol (ES)
  - Français (FR)
  - Italien (IT)
  - Néerlandais (NL)
  - Portugais (PT)
- **Personnalisé** Si aucune langue prise en charge ne convient, utilisez cette option pour indiquer qu'une valeur personnalisée doit être utilisée. Il s'agit généralement d'un code de langue ISO sur deux lettres, mais toute chaîne de texte dont vous avez besoin est requise.

Chapitre 3. Text Mining pour les concepts et les catégories

Le noeud modélisation Text Mining est utilisé pour générer l'un des deux nuggets de modèle Text Mining :

- *Les nuggets de modèles de concepts* explorent et extraient les concepts fondamentaux de données textuelles, structurées ou non.
- *Les nuggets de modèles de catégories* scorent et attribuent des documents et des enregistrements à des catégories, qui sont constitués des concepts (et des motifs) extraits.

Les concepts, les motifs et les catégories extraits de vos nuggets de modèle peuvent tous être combinés aux données structurées existantes, telles que les données démographiques, et appliqués grâce à la gamme complète d'outils de IBM SPSS Modeler, afin de favoriser une prise de décision plus précise et plus efficace. Par exemple, si des clients signalent fréquemment des problèmes de connexion comme étant l'entrave principale aux tâches de gestion de compte en ligne, vous serez peut-être amené à inclure “problèmes de connexion” dans vos modèles.

En outre, le noeud modélisation Text Mining est totalement intégré dans IBM SPSS Modeler. Vous pouvez ainsi déployer des flux de Text Mining via IBM SPSS Modeler Solution Publisher pour le scoring des données non structurées en temps réel dans des applications comme PredictiveCallCenter. Cette possibilité permet de garantir une implémentation réussie des opérations d'exploration de texte en boucles fermées. Par exemple, votre organisation peut désormais analyser de manière plus pertinente les notes consignées dans le Bloc-notes, issues des appelants entrants et sortants, en appliquant vos modèles prédictifs. Votre communication marketing en temps réel est ainsi mieux adaptée. L'utilisation des résultats des modèles Text Mining dans les flux améliore l'exactitude des modèles de données prédictifs.

Pour exécuter IBM SPSS Modeler Text Analytics avec IBM SPSS Modeler Solution Publisher, ajoutez le répertoire `<install_directory>/ext/bin/spss.TMWBServer` à la variable d'environnement `$LD_LIBRARY_PATH`.

Dans IBM SPSS Modeler Text Analytics, il est souvent fait référence aux concepts et aux catégories extraits. Il est important de comprendre la signification des concepts et des catégories car ils peuvent faciliter la prise de décisions plus avisées lors du travail exploratoire et de la création de modèles.

Concepts et nuggets de modèles de concepts

Lors du processus d'extraction, les données texte sont analysées pour identifier des mots isolés intéressants ou pertinents, par exemple `élection` ou `paix` et des groupes de mots, par exemple `élection présidentielle`, `élection du président` ou `traités de paix`. Ces mots et groupes de mots sont appelés des *termes*. En utilisant les ressources linguistiques, les termes pertinents sont extraits et les termes similaires sont regroupés sous un terme principal appelé **concept**.

Ainsi, un concept peut représenter plusieurs termes sous-jacents en fonction de votre texte et des ressources linguistiques utilisées. Par exemple, prenons une enquête de satisfaction destinée à des employés et supposons que le concept `salaire` a été extrait. Supposons également que lorsque vous avez examiné les enregistrements associés à `salaire`, vous avez noté que `salaire` n'est pas toujours présent dans le texte mais qu'au lieu de cela certains enregistrements contiennent des termes similaires, tels que `revenu`, `revenus`, et `salaires`. Ces termes sont regroupés sous `salaire` car le moteur du programme d'extraction a déterminé qu'ils étaient similaires ou qu'il s'agissait de synonymes en fonction des règles de traitement ou des ressources linguistiques. Dans ce cas, les documents ou les enregistrements contenant ces termes sont traités de la même manière que s'ils contenaient le mot `salaire`.

Si vous souhaitez prendre connaissance des termes regroupés sous un concept, vous pouvez explorer le concept dans un plan de travail interactif ou consulter les synonymes indiqués dans le modèle de concepts. Pour plus d'informations, voir «Termes sous-jacents dans les modèles de concepts», à la page 34.

Un **nugget de modèle de concepts** contient un ensemble de concepts pouvant être utilisés afin d'identifier des enregistrements ou des documents qui contiennent également le concept (notamment ses synonymes ou des groupes de termes). Un modèle de concepts peut être utilisé de deux manières. La première consiste à explorer et à analyser les concepts rencontrés dans le texte source d'origine ou à identifier rapidement les documents intéressants. La seconde consiste à appliquer ce modèle aux nouveaux enregistrements ou documents texte afin d'identifier rapidement les concepts-clés similaires contenus dans ceux-ci (par exemple, la recherche en temps réel de concepts-clés dans les notes d'un centre d'appel).

Pour plus d'informations, voir «Nugget de Text Mining : Modèle de concept», à la page 31.

Catégories et nuggets de modèles de catégories

Vous pouvez créer des **catégories** représentant, essentiellement, des concepts de niveau supérieur ou des rubriques pour capturer les principales idées, les connaissances et les attitudes exprimées dans le texte. Les catégories sont constituées d'un ensemble de descripteurs, tels que des *concepts*, des *types* et des *règles*. Ensemble, ces descripteurs permettent d'identifier si un enregistrement ou un document appartient ou non à une catégorie. Un document ou un enregistrement peut être analysé afin de déterminer si un texte qu'il contient correspond à un descripteur. Si une correspondance est détectée, le document/ l'enregistrement est attribué à cette catégorie. Ce processus est appelé **catégorisation**.

Les catégories peuvent être créées automatiquement à l'aide des techniques fiables et automatisées du produit, ou manuellement, en utilisant les informations supplémentaires dont vous disposez concernant les données, ou à l'aide d'une combinaison des deux. Vous pouvez aussi charger un ensemble de catégories prédéfinies à partir d'un pack d'analyse de textes grâce à l'onglet Modèle de ce noeud. La création manuelle de catégories ou l'affinement de catégories ne peuvent être effectuées que par l'intermédiaire du plan de travail interactif. Pour plus d'informations, voir «Noeud Text Mining : onglet Modèle», à la page 24.

Un **nugget de modèle de catégories** contient un ensemble de catégories et les descripteurs associés. Le modèle peut être utilisé afin de regrouper en catégories un ensemble de documents ou d'enregistrements en fonction du texte contenu dans chaque document/enregistrement. Chaque document ou enregistrement est lu et affecté à chaque catégorie pour laquelle une correspondance de descripteur a été identifiée. De cette manière, il est possible d'attribuer un document ou un enregistrement à plus d'une catégorie. Vous pouvez utiliser des nuggets de modèles de catégories pour visualiser les idées essentielles contenues dans des réponses ouvertes à des enquêtes ou dans un ensemble d'entrées de blogue, par exemple.

Pour plus d'informations, voir «Nugget de Text Mining : Modèle de catégorie», à la page 40.

Noeud de modélisation Text Mining

Le noeud Text Mining applique des techniques linguistiques et de fréquence pour extraire les principaux concepts du texte et créer des catégories avec ces concepts et d'autres données. Grâce à ce noeud, vous pouvez explorer le contenu de données textuelles ou générer soit un nugget de modèle de concepts, soit un nugget de modèle de catégories. Lorsque vous exécutez ce noeud modélisation, un moteur d'extraction linguistique interne extrait et organise les concepts, les motifs et/ou les catégories à l'aide de méthodes de traitement du langage naturel.

Vous pouvez exécuter le noeud Text Mining et générer automatiquement un nugget de modèle de concepts ou de catégories grâce à l'option **Générer directement**. Sinon, vous pouvez aussi utiliser une

approche plus pragmatique et exploratoire grâce au mode **Créer de manière interactive** dans lequel vous pouvez non seulement extraire des concepts, créer des catégories et affiner vos ressources linguistiques mais aussi procéder à une analyse des liens du texte et explorer des clusters. Pour plus d'informations, voir «Noeud Text Mining : onglet Modèle», à la page 24.

Vous pouvez trouver ce noeud dans l'onglet IBM SPSS Modeler Text Analytics de la palette de noeuds en bas de la fenêtre IBM SPSS Modeler. Pour plus d'informations, voir «Noeuds IBM SPSS Modeler Text Analytics», à la page 8.

Conditions requises. Les noeuds modélisation Text Mining acceptent des données textuelles d'un noeud Flux de nouvelles, d'un noeud Liste fichiers ou de noeuds source standard. Le noeud est installé avec IBM SPSS Modeler Text Analytics et est accessible sur la palette IBM SPSS Modeler Text Analytics.

Remarque : Ce noeud remplace le noeud Extraction de texte fourni avec les versions précédentes du produit. Si vous disposez de flux plus anciens qui utilisent les anciens noeuds ou nuggets de modèle, vous devez recréer vos flux à l'aide du noeud Text Mining.

Noeud Text Mining : onglet Champs

Utilisez l'onglet Champs pour indiquer les paramètres de champ des données dont vous allez extraire les concepts. Pour accélérer la durée du traitement, pensez à utiliser un noeud échantillon en amont de ce noeud lorsque vous travaillez avec de grands ensembles de données. Pour plus d'informations, voir «Echantillonnage en amont pour gagner du temps», à la page 30.

Vous pouvez définir les paramètres suivants :

Champ ID Sélectionnez le champ contenant l'identificateur des enregistrements textuels. L'identificateur doit être un entier. Le champ d'ID sert d'index aux enregistrements textuels. Utilisez un champ ID si le champ de texte correspond au texte à explorer.

Champ Texte. Sélectionnez le champ contenant le texte à explorer. Ce champ dépend de la source de données.

Champ Langue Sélectionnez le champ contenant l'identificateur de langue ISO sur deux lettres. Si vous ne sélectionnez pas de champ, la langue du modèle fourni est utilisée pour chaque document.

Type de document. Le type de document indique la structure du texte. Sélectionnez l'un des types suivants :

- **Texte réel.** S'utilise pour la plupart des documents ou des sources textuelles. L'ensemble de texte entier est analysé pour l'extraction. Contrairement aux autres options, il n'existe pas de paramètres supplémentaires pour cette option.
- **Texte structuré.** Utilisez ce format pour les bibliographies, les brevets et les fichiers contenant des structures standard qui peuvent être identifiées et analysées. Ce type de document est utilisé pour éviter tout ou une partie du processus d'extraction. Il permet de définir des séparateurs de termes, d'affecter des types et d'imposer une valeur de fréquence minimale. Si vous sélectionnez cette option, vous devez cliquer sur le bouton **Paramètres** et entrer les séparateurs de texte dans la zone **Formatage de texte structuré** de la boîte de dialogue Paramètres du document. Pour plus d'informations, voir «Paramètres de document de l'onglet Champs», à la page 22.

Unité de texte. Sélectionnez le mode d'extraction parmi les choix suivants :

- **Mode de document.** Utilisez ce mode pour les documents courts et homogènes d'un point de vue sémantique (par exemple, dans le cas d'articles issus d'agences de presse).
- **Mode de paragraphe.** Utilisez ce format pour les pages Web et les documents sans balise. Le processus d'extraction divise les documents en unités sémantiques, sur la base de certaines caractéristiques, comme des balises internes et des éléments syntaxiques. Si ce mode est sélectionné, le scoring est

appliqué paragraphe par paragraphe. Par conséquent, la règle `pomme & orange` est vraie uniquement si `pomme` et `orange` se trouvent dans le même paragraphe, par exemple.

Remarque : En raison du mode d'extraction du texte depuis les documents PDF, le **mode Paragraphe** ne fonctionne pas sur ces documents. La raison en est que l'extraction supprime le marqueur de retour chariot.

Paramètres de mode de paragraphe. Cette option n'est disponible que si vous affectez à l'option Unité de texte la valeur **Mode de paragraphe**. Indiquez les nombres maximal et minimal de caractères à utiliser dans les extractions. La taille employée est arrondie à la période la plus proche. Pour vous assurer que les associations de mots obtenues à partir du texte du groupe de documents sont représentatives, n'indiquez pas de taille d'extraction trop petite.

- **Minimum.** Indiquez le nombre minimum de caractères à utiliser dans les extractions.
- **Maximum.** Indiquez le nombre maximal de caractères à utiliser dans les extractions.

Partitionnement Le partitionnement permet d'opter pour un partitionnement basé sur les paramètres du noeud type ou pour un autre type de partitionnement. Le partitionnement répartit les données en échantillons d'apprentissage et de test.

Paramètres de document de l'onglet Champs

Formatage de texte structuré

Si vous souhaitez éviter une partie ou tout le processus d'extraction parce que vous avez des données structurées ou que vous souhaitez imposer des règles sur la façon de traiter le texte, utilisez l'option de type de document **Texte structuré** et déclarez les champs ou les balises contenant le texte dans la section **Formatage de texte structuré** de la boîte de dialogue Paramètres de document. Les termes extraits sont calculés à partir du texte contenu entre les champs ou les balises déclarés (et entre leurs balises enfant). Tout champ ou balise non déclaré sera ignoré.

Dans certains contextes, le traitement linguistique n'est pas obligatoire et le moteur d'extraction linguistique peut être remplacé par des déclarations explicites. Dans un fichier bibliographique où les champs de mot-clé sont séparés par des séparateurs, comme un point-virgule (;) ou une virgule (,), il suffit d'extraire la chaîne comprise entre deux séparateurs. C'est pourquoi il est possible de suspendre le processus d'extraction complet et de définir des règles de traitement spécifiques pour déclarer les séparateurs de termes, affecter les types au texte extrait, ou imposer un effectif de fréquences minimal pour l'extraction.

Utilisez les règles suivantes pour déclarer des éléments de texte structuré :

- Un seul champ, balise ou élément peut être déclaré par ligne. Il n'est pas nécessaire de les faire figurer dans les données.
- Les déclarations font la distinction entre majuscules et minuscules.
- Si vous déclarez une balise avec des attributs, par exemple `<title id="1234">` et que vous voulez inclure toutes les variations (dans ce cas, tous les ID) ajoutez la balise sans attribut ou sans le crochet fermant (>) de la manière suivante : `<title`
- Ajoutez le signe deux-points après le nom du champ ou de la balise pour indiquer qu'il s'agit de texte structuré. Ajoutez le signe deux-points directement après le nom du champ ou la balise mais avant le séparateur, le type ou la fréquence, par exemple `author:` ou `<place>:`.
- Pour indiquer que plusieurs termes sont contenus dans le champ ou la balise et qu'un séparateur est utilisé pour désigner les termes individuels, déclarez le séparateur après le signe deux-points, comme `auteur:,` ou `<section>:;`.
- Pour assigner un type au contenu trouvé dans la balise, déclarez le nom du type après le signe deux-points et un séparateur comme `author:,Person` ou `<place>:;Location`. Déclarez le type à l'aide des noms tels qu'ils apparaissent dans l'éditeur de ressources.

- Pour définir un effectif de fréquences minimal pour un champ ou une balise, déclarez un chiffre à la fin de la ligne, comme `author:,Person1` ou `<place>:;Location5`. Où `n` est l'effectif de fréquences : les termes trouvés dans le champ ou la balise doivent se produire au moins `n` fois dans l'ensemble entier de documents ou d'enregistrements à extraire. Il vous faut également définir un séparateur.
- Si vous avez une balise qui contient un signe deux-points, une barre oblique inverse doit précéder ces deux points afin que la déclaration ne soit pas ignorée. Par exemple, si vous avez un champ nommé `<topic:source>`, saisissez-le sous la forme `<topic\:source>`.

Pour illustrer la syntaxe, imaginons les champs bibliographiques récurrents suivants :

```
author:Morel, Kawashima
abstract:This article describes how fields are declared.
publication:Text Mining Documentation
datepub:March 2010
```

Dans cet exemple, si nous souhaitions que le processus d'extraction considère l'auteur et l'extrait mais ignore le reste du contenu, nous déclarerions uniquement les champs suivants :

```
author:,Person1
abstract:
```

Dans cet exemple, la déclaration de champ `author:,Person1` indique que le traitement linguistique du contenu des champs a été suspendu. Au lieu de cela, il spécifie que le champ auteur contient plus d'un nom, lequel est séparé du suivant par le séparateur virgule, et que les noms doivent être attribués au type Personne et que si le nom apparaît au moins une fois dans l'ensemble complet des documents ou des enregistrements, il doit être extrait. Le champ`abstract:` étant listé sans aucune autre déclaration, il sera analysé au cours de l'extraction et soumis à un traitement linguistique et de détermination de type standard.

Formatage de texte XML

Si vous souhaitez limiter le processus d'extraction au texte contenu entre des balises XML spécifiques uniquement, utilisez l'option de type de document **texte XML** pour déclarer les balises contenant le texte dans la section **formatage de texte XML** de la boîte de dialogue Paramètres du document. Les termes extraits sont calculés à partir du texte contenu entre ces balises ou entre leurs balises enfant.

Important ! Si vous souhaitez éviter le processus d'extraction et imposer des règles sur les séparateurs de termes, affecter des types au texte extrait ou imposer un effectif de fréquences aux termes extraits, utilisez l'option **texte structuré** suivante.

Lorsque vous déclarez des balises pour le formatage de texte XML, utilisez les règles suivantes :

- Une seule balise XML peut être déclarée par ligne.
- Les éléments de balise respectent la casse.
- Si une balise a des attributs, par exemple `<title id="1234">` et que vous voulez inclure toutes les variations (dans ce cas, tous les ID) ajoutez la balise sans attribut ou sans le crochet fermant (>) de la manière suivante : `<title`

Pour illustrer la syntaxe, imaginons le document XML suivant :

```
<section>Règles de la route
     <title id="01234">Feux de circulation</titre>
     <p>Les panneaux de signalisation sont utiles.</p>
</section>
<p>L'apprentissage des règles est important.</p>
```

Pour cet exemple, déclarons les balises suivantes :

```
<section>
<title
```

Dans cet exemple, parce que vous avez déclaré la balise `<section>`, le texte à l'intérieur de cette balise et de ses balises imbriquées, `Feux de circulation` et `Les panneaux de signalisation sont utiles`, est analysé pendant le processus d'extraction. Toutefois, `L'apprentissage des règles est important` est ignoré, car cette balise`<p>` n'a pas été explicitement déclarée et qu'elle n'a pas été imbriquée dans une balise déclarée.

Noeud Text Mining : onglet Modèle

L'onglet Modèle est utilisé pour indiquer la méthode de création et les paramètres de modèle généraux pour la sortie du noeud.

Vous pouvez définir les paramètres suivants :

Nom du modèle Vous pouvez générer le nom du modèle automatiquement sur la base du champ cible ou ID (ou du type de modèle si aucun de ces champs n'est spécifié) ou spécifier un nom personnalisé.

Utiliser les données partitionnées. Si une zone de partition est définie, seules les données d'apprentissage sont utilisées pour la création du modèle.

Mode Créer Indique la façon dont les nuggets de modèle sont générés lorsqu'un flux contenant ce noeud Text Mining sera exécuté. Sinon, vous pouvez aussi utiliser une approche plus pragmatique et exploratoire grâce au mode **Créer de manière interactive** dans lequel vous pouvez non seulement extraire des concepts, créer des catégories et affiner vos ressources linguistiques mais aussi procéder à une analyse des liens du texte et explorer des clusters.

- **Créer de manière interactive** Lorsqu'un flux est exécuté, cette option lance une interface interactive dans laquelle vous pouvez extraire des concepts et des motifs, explorer les résultats extraits afin de construire et d'affiner les catégories et les ressources linguistiques (modèles, synonymes, types, bibliothèques, etc.)et construire des nuggets de modèle de catégories. Pour plus d'informations, voir «Créer de manière interactive», à la page 25.
- **Générer directement** Cette option indique, que lorsque le flux est exécuté, un modèle doit être automatiquement créé et ajouté à la palette Modèles. A la différence du plan de travail interactif, aucune manipulation supplémentaire n'est nécessaire de votre part au moment de l'exécution outre les paramètres définis dans le noeud. Si vous sélectionnez cette option, des options propres au modèle apparaissent et vous permettent de définir le type de modèle que vous souhaitez produire. Pour plus d'informations, voir «Générer directement», à la page 26.

Store large models in AS Si vous disposez d'une connexion à IBM SPSS Analytic Server, sélectionnez cette option pour stocker vos modèles à distance sur le serveur.

Remarque : Tout modèle créé et stocké sur un serveur ne peut être évalué que sur ce dernier. Pour reprendre une session de plan de travail interactif qui contient un tel modèle, vous devez disposer d'une connexion au serveur d'origine utilisé pour créer la session.

Copier les ressources depuis Lors de l'exploration de texte, l'extraction est basée sur les paramètres de l'onglet Expert mais également sur les ressources linguistiques. Ces ressources servent de base au traitement et à l'exécution du texte pendant l'extraction afin d'obtenir des concepts, des types et parfois des motifs. Vous pouvez copier les ressources dans ce noeud à partir d'un modèle de ressources ou d'un pack d'analyse de texte. Sélectionnez-en un puis cliquez sur **Charger** pour définir le pack ou le modèle depuis lequel les ressources seront copiées. Au moment du chargement, une copie des ressources est stockée dans le noeud. Par conséquent, si vous souhaitez utiliser un modèle mis à jour ou un TAP (pack d'analyse de texte), vous devez le recharger ici ou dans une session de plan de travail interactif. Pour faciliter votre travail, la date et l'heure de copie et de chargement des ressources sont indiquées dans le noeud. Pour plus d'informations, voir «Copie des ressources à partir de modèles et de TAP», à la page 26.

Langue du texte. Identifie la langue du texte exploré. Les ressources copiées dans le noeud contrôlent les options de langue présentées. Sélectionnez la langue pour laquelle les ressources ont été optimisées.

Créer de manière interactive

Dans l'onglet Modèle du noeud modélisation Text Mining, vous pouvez choisir un mode de génération pour vos nuggets de modèle. Si vous sélectionnez **Créer de manière interactive**, une interface interactive s'ouvre lorsque vous exécutez le flux. Dans ce plan de travail interactif, vous pouvez effectuer les opérations suivantes :

- Extraire et explorer les résultats de l'extraction, y compris des concepts et typages pour découvrir les idées fondamentales dans vos données de texte.
- Utilisez différentes méthodes pour créer et étendre des catégories provenant de concepts, de types, de motifs TLA et de règles et pouvoir scorer vos documents et enregistrements dans ces catégories.
- Affinez vos ressources linguistiques (modèles de ressources, bibliothèques, dictionnaires, synonymes, etc.) pour pouvoir améliorer vos résultats grâce à un processus itératif dans lequel les concepts sont extraits, examinés et affinés.
- Procéder à une analyse des liens du texte (TLA) et utiliser les motifs TLA découverts afin de créer de meilleurs nuggets de modèle de catégories. Le noeud analyse des liens du texte ne propose pas les mêmes options exploratoires ni les mêmes capacités de modélisation.
- Générez des clusters pour découvrir de nouvelles relations ou explorer des relations entre les concepts, les types, les motifs et les catégories dans la sous-fenêtre Visualisation.
- Générez des nuggets de modèles de catégories affinés dans la palette Modèles de IBM SPSS Modeler et utilisez-les dans d'autres flux.

Remarque : Vous ne pouvez pas créer un modèle interactif si vous créez un travail IBM SPSS Collaboration and Deployment Services.

Ré-utiliser les données sauvegardées. Lorsque vous travaillez dans une session de plan de travail interactif, vous pouvez mettre à jour le noeud avec les données de session (paramètres d'extraction, ressources, définition de catégories, etc.). L'option **Utiliser le plan de travail interactif** permet de redémarrer la session interactive en utilisant les données de la session enregistrées. Cette option est désactivée la première fois que vous utilisez ce noeud, car aucune donnée de session n'a pu être enregistrée. Pour déterminer comment mettre à jour le noeud avec des données de session afin de pouvoir utiliser cette option, voir «Mise à jour des noeuds modélisation et enregistrement», à la page 78.

Si vous lancez une session *avec* cette option, alors les paramètres d'extraction, les catégories, les ressources et tout autre travail effectué lors de la dernière mise à jour du noeud depuis une session de plan de travail interactif seront disponibles au prochain lancement de session. Les données de session enregistrées étant utilisées avec cette option, certains contenus, comme les ressources copiées dans le modèle ci-dessous et d'autres onglets, sont désactivés et ignorés. Mais si vous lancez une session *sans* cette option, seuls les contenus du noeud comme définis actuellement sont utilisés, ce qui signifie que tout travail précédent effectué dans l'utilitaire ne sera pas disponible.

Remarque : si vous modifiez le noeud source de votre flux après la mise en cache des résultats de l'extraction à l'aide de l'option **Ré-utiliser les données sauvegardées**, vous devez lancer une nouvelle extraction lorsque vous démarrez la session de plan de travail interactif afin d'obtenir des résultats d'extraction à jour.

Ne pas ré-extraire et ré-utiliser les données cachées et les résultats. Vous pouvez ré-utiliser les résultats et les données cachés de l'extraction dans la session de plan de travail interactif. Cette option est particulièrement utile lorsque vous souhaitez gagner du temps et ré-utiliser les résultats de l'extraction plutôt que d'attendre l'exécution d'une toute nouvelle extraction lors du lancement de la session. Afin d'utiliser cette option, vous devez d'abord avoir mis à jour ce noeud depuis une session de plan de travail interactif et avoir choisi l'option **Conserver la session interactive et les données texte cachées avec les résultats de l'extraction pour une utilisation ultérieure**. Pour déterminer comment mettre à jour le noeud avec des données de session afin de pouvoir utiliser cette option, voir «Mise à jour des noeuds modélisation et enregistrement», à la page 78.

Démarrer la session par une. Sélectionner l'option indiquant la vue à afficher et l'action à effectuer en premier lors du lancement d'une session de plan de travail interactif. Quelle que soit la vue dans laquelle vous commencez, une fois la session ouverte, vous pouvez choisir n'importe quelle vue.

- **Utilisation des résultats d'extraction pour créer des catégories.** Cette option lance le plan de travail interactif dans la vue Catégories et concepts et, le cas échéant, effectue une extraction. Dans cette vue, il est possible de créer des catégories et de générer un modèle de catégories. Vous pouvez également afficher une autre vue. Pour plus d'informations, voir Chapitre 7, «Mode Plan de travail interactif», à la page 67.
- **Exploration des résultats de l'analyse des liens du texte (TLA).** Au démarrage, cette option commence par extraire et identifier les relations entre les concepts contenus dans le texte, comme les opinions ou les autres liens de la vue Analyse des liens du texte. Vous devez sélectionner un modèle ou un pack d'analyse de texte qui contient des règles de motifs TLA pour utiliser cette option et obtenir des résultats. Si vous travaillez avec des ensembles de données plus importants, l'extraction TLA peut prendre du temps. Dans ce cas, vous pouvez envisager d'utiliser un noeud échantillon en amont. Pour plus d'informations, voir Chapitre 11, «Exploration de l'analyse des liens du texte», à la page 149.
- **Analyse de classification en cluster des co-termes.** Cette option est lancée dans la vue Clusters et met à jour les résultats d'extraction obsolètes. Dans cette vue, vous pouvez effectuer une analyse en cluster des co-termes, ce qui produit un ensemble de classes. La classification en cluster des co-termes est un processus qui commence par évaluer la force de la valeur du lien entre deux concepts d'après leur co-occurrence dans un enregistrement ou un document spécifique et se termine par le regroupement des concepts fortement liés dans des clusters. Pour plus d'informations, voir Chapitre 7, «Mode Plan de travail interactif», à la page 67.

Générer directement

Dans l'onglet Modèle du noeud modélisation Text Mining, vous pouvez choisir un mode de génération pour vos nuggets de modèle. Si vous sélectionnez **Générer directement**, vous pouvez définir les options dans le noeud puis simplement exécuter votre flux. La sortie est un nugget de modèle de concepts qui a été directement placé dans la palette Modèles. A la différence du plan de travail interactif, aucune manipulation supplémentaire n'est nécessaire de votre part au moment de l'exécution outre les paramètres de fréquence définis pour cette option dans le noeud.

Nombre maximum de concepts à inclure dans le modèle. Cette option, valable uniquement lorsque vous créez un modèle automatiquement (non interactif), indique que vous souhaitez créer un modèle de concepts. Elle indique également que ce modèle ne doit pas contenir plus que le nombre indiqué de concepts.

- **Sélectionner les concepts basés sur la fréquence la plus élevée. Plus grand nombre de concepts.** Il s'agit du nombre de concepts qui seront cochés, en partant de celui dont la fréquence est la plus élevée. Le terme Fréquence fait ici référence au nombre d'occurrences d'un concept (et de tous ses termes sous-jacents) dans l'ensemble des documents/enregistrements. Il est parfois supérieur au nombre d'enregistrements car un concept peut figurer plusieurs fois dans un même enregistrement.
- **Désactiver les concepts apparaissant dans trop d'enregistrements. Pourcentage d'enregistrements.** Désactiver les concepts apparaissant dans un pourcentage supérieur à celui indiqué pour le nombre d'enregistrements. Cette option permet d'exclure les concepts qui figurent fréquemment dans le texte ou les enregistrements, mais qui ne présentent pas d'intérêt pour l'analyse.

Optimiser la vitesse de scoring. Sélectionnée par défaut, cette option permet de créer un modèle compact qui procède à un scoring à grande vitesse. Si vous désélectionnez cette option, le modèle sera beaucoup plus important et le scoring plus lent. Toutefois, un modèle plus important permet de s'assurer que les scorings affichés initialement dans le modèle de concept généré sont identiques à celles obtenues via un scoring du même texte avec le nugget de modèle.

Copie des ressources à partir de modèles et de TAP

Lors de l'exploration de texte, l'extraction est basée sur les paramètres de l'onglet Expert mais également sur les ressources linguistiques. Ces ressources servent de base au traitement et à l'exécution du texte

pendant l'extraction afin d'obtenir des concepts, des types et parfois des motifs. Vous pouvez copier des ressources dans ce noeud depuis un *modèle de ressources*, et si vous êtes dans le noeud Text Mining, vous pouvez également sélectionner un *pack d'analyse de texte* (TAP).

Par défaut, les ressources sont copiées dans le noeud du modèle de base de la langue dont vous possédez la licence pour votre produit lorsque vous ajoutez le noeud à l'espace de travail. Si vous possédez des licences pour plusieurs langues, la première langue sélectionnée est utilisée pour déterminer le modèle à charger automatiquement.

Au moment du chargement, une copie des ressources sélectionnées est stockée dans le noeud. Seul le contenu du modèle ou du TAP est copié, mais le modèle ou TAP n'est pas lui-même lié au noeud. Par conséquent, si ce modèle ou TAP est par la suite mis à jour, ces mises à jour ne sont pas automatiquement disponibles dans le noeud. Pour résumer, les ressources chargées dans le noeud sont toujours utilisées sauf si vous rechargez une copie d'un modèle ou d'un TAP ou si vous mettez à jour un noeud Text Mining et sélectionnez l'option **Utiliser le travail d'une session**. Pour des informations sur l'option **Utiliser le travail d'une session**, consultez la rubrique suivante.

Lorsque vous sélectionnez un modèle ou TAP, choisissez-en un ayant le même langage que vos données texte. Vous ne pouvez utiliser que les modèles ou TAP définis dans les langues pour lesquelles vous détenez une licence. Pour effectuer une analyse des liens du texte, vous devez sélectionner un modèle contenant des patrons TLA. Si un modèle contient des patrons TLA, une icône s'affiche dans la colonne TLA de la boîte de dialogue Charger le modèle de ressources.

Remarque : Vous ne pouvez pas charger les TAP dans le noeud d'analyse des liens du texte.

Modèles de ressources

Un modèle de ressources est un ensemble prédéfini de bibliothèques et de ressources linguistiques et non linguistiques avancées qui ont été affinées pour un domaine ou une utilisation spécifique. Dans le noeud modélisation Text Mining, une copie des ressources d'un modèle de base est déjà chargée dans le noeud lorsque vous ajoutez le noeud au flux, mais vous pouvez modifier les modèles ou charger un pack d'analyse de texte en sélectionnant **Modèle de ressources** ou **pack d'analyse de texte** puis en cliquant sur **Charger**. Pour les modèles, vous pouvez ensuite sélectionner le modèle dans la boîte de dialogue Charger un modèle de ressources.

Remarque : Si le modèle que vous voulez utiliser n'apparaît pas dans la liste, mais que vous en avez une copie exportée sur votre machine, vous pouvez l'importer à ce stade. Vous pouvez également exporter depuis cette boîte de dialogue pour un partage avec d'autres utilisateurs. Pour plus d'informations, voir «Import et export des modèles», à la page 175.

Packs d'analyse de texte (TAP)

Un pack d'analyse de texte (TAP) est un ensemble prédéfini de bibliothèques et de ressources linguistiques et non linguistiques avancées qui ont été regroupées dans un ou plusieurs ensembles de catégories prédéfinies. IBM SPSS Modeler Text Analytics contient plusieurs TAP préconfigurés pour le texte anglais, lesquels sont affinés pour un domaine particulier. Vous ne pouvez pas éditer ces TAP mais vous pouvez les utiliser pour commencer votre création de modèle de catégorie. Vous pouvez également créer vos propres TAP dans la session interactive. Pour plus d'informations, voir «Chargement des packs d'analyse de texte», à la page 137.

Remarque : Vous ne pouvez pas charger les TAP dans le noeud d'analyse des liens du texte.

Utilisation de l'option « Utiliser le travail d'une session » (onglet Modèle)

Alors que les ressources sont copiées dans le noeud de l'onglet Modèle, vous pouvez également effectuer des modifications ultérieures sur les ressources dans une session interactive et mettre à jour le noeud

modélisation Text Mining avec ces dernières modifications. Dans ce cas, vous pouvez sélectionner l'option **Utiliser le travail d'une session** dans l'onglet Modèle du noeud modélisation Text Mining.

Si vous sélectionnez **Utiliser le travail d'une session**, le bouton **Charger** est désactivé dans le noeud pour indiquer que ces ressources qui provenaient du plan de travail interactif seront utilisées à la place des ressources chargées précédemment.

Après avoir sélectionné l'option **Utiliser le travail d'une session**, vous pouvez modifier vos ressources directement dans la session de plan de travail interactif avec la vue Editeur de ressources. Pour plus d'informations, voir «Mise à jour des ressources d'un noeud après le chargement», à la page 173.

Noeud Text Mining : onglet Expert

L'onglet Expert contient des paramètres avancés ayant une incidence sur le mode d'extraction et de traitement du texte. Les paramètres de cette boîte de dialogue déterminent le fonctionnement de base du processus d'extraction, ainsi que quelques procédures avancées. Cependant, ils ne représentent qu'une partie des options disponibles. Il existe également un certain nombre de ressources linguistiques et d'options ayant une incidence sur les résultats de l'extraction, qui sont contrôlées par le modèle de ressources sélectionné dans l'onglet Modèle. Pour plus d'informations, voir «Noeud Text Mining : onglet Modèle», à la page 24.

Remarque : Cet onglet est désactivé si vous avez sélectionné le mode **Créer de manière interactive** en utilisant les informations du plan de travail interactif sauvegardé de l'onglet Modèle ; dans ce cas, les paramètres d'extraction sont ceux de la dernière session de plan de travail sauvegardée.

Vous pouvez définir les paramètres suivants lors d'une extraction :

Limiter l'extraction aux concepts ayant une fréquence globale supérieure à [n]. Spécifie à partir de combien d'occurrences un mot ou un groupe de mots présent dans un texte doit être extrait. Ainsi, une valeur de 5 limite l'extraction aux mots ou groupes de mots figurant au moins cinq fois dans l'ensemble des enregistrements ou des documents.

Dans certains cas, modifier cette limite peut faire une grande différence dans les résultats d'extraction et par conséquent, dans les catégories. Imaginons que vous travaillez avec des données concernant un restaurant et que vous n'avez pas augmenté la limite au-dessus de 1 pour cette option. Dans ce cas, vos résultats d'extraction pourront contenir *pizza (1)*, *pizza fine (2)*, *pizza épinards (2)*, et *pizza préférée (2)*. Mais si l'extraction était limitée à une fréquence globale de 5 ou plus et que vous recommenciez l'extraction, trois de ces concepts ne seraient pas renvoyés. Vous obtiendriez *pizza (7)*, car *pizza* est la forme la plus simple et que ce mot existait déjà comme candidat possible. Et en fonction du reste du texte, vous pourriez obtenir une fréquence supérieure à 7, si le texte contient d'autres phrases avec le mot pizza. De plus, si *pizza épinards* était déjà un descripteur de catégorie, vous pourriez le remplacer par*pizza* pour pouvoir capturer tous les enregistrements. C'est pour cette raison que lorsque des catégories ont déjà été créées, la modification de cette limite doit être effectuée avec prudence.

Veuillez noter qu'il s'agit d'une fonction d'extraction uniquement ; si votre modèle contient des termes (ce qui est généralement le cas) et qu'un terme pour le modèle est trouvé dans le texte, alors le terme sera indexé quelle que soit sa fréquence.

Par exemple, supposons que vous utilisez un modèle Ressources de base qui inclut "los angeles" sous le type `<Location>` dans la bibliothèque principale ; si votre document contient une seule occurrence du terme "los angeles", celui-ci fera partie de la liste des concepts. Pour éviter cela, vous devrez définir un filtre pour afficher les concepts se produisant au moins le même nombre de fois que la valeur saisie dans le champ **Limiter l'extraction aux concepts ayant une fréquence globale supérieure à [n]**.

Adapter les erreurs de ponctuation. Cette option normalise temporairement le texte contenant des erreurs de ponctuation (par exemple, usage incorrect) au cours de l'extraction pour améliorer l'extraction

des concepts. Cette option s'avère extrêmement utile lorsque le texte est court et de qualité médiocre (réponses ouvertes, messages électroniques, données CRM, etc.) ou qu'il contient de nombreuses abréviations.

Traitement des fautes de frappe. Nombre de caractères minimum requis : [n] Cette option applique une technique de regroupement flou qui permet de regrouper les mots mal orthographiés ou d'orthographe similaire sous un seul concept. L'algorithme de regroupement flou retire temporairement toutes les voyelles (à l'exception de la première) et les consonnes doubles/triples des mots extraits pour comparer ces derniers et voir s'ils sont identiques, afin par exemple que `modélisation` et `modéllisation` soient regroupés. Néanmoins, si chaque type est affecté à un type différent et que vous excluez le type `<Unknown>`, le regroupement flou n'est pas appliqué.

Vous pouvez également définir le nombre minimum de caractères *racines* requis avant d'utiliser le regroupement flou. Le nombre de caractères racine d'un terme est calculé en ajoutant l'ensemble des caractères et en soustrayant les caractères formant des suffixes inflexionnels et, dans le cas des termes apparaissant sous la forme de mots composés, les déterminants et les prépositions. Par exemple, le terme `exercices` sera considéré comme contenant 8 caractères racines dans la forme “exercice,” la lettre *s* de fin étant une inflexion (marque du pluriel). De même, `sauce soja` contient 9 caractères racines (“sauce soja”) et `usine de voitures` en contient 12 (“usine voiture”). Cette méthode de calcul est uniquement utilisée pour vérifier si le regroupement flou doit être appliqué, mais n'influe pas sur la mise en correspondance des mots.

Remarque : Si vous constatez plus tard que certains mots sont regroupés par erreur, vous pouvez exclure des paires de mots en les déclarant explicitement dans la section **Regroupement flou : Exceptions** de l'onglet Ressources avancées. Pour plus d'informations, voir «Regroupement flou», à la page 203.

Enlever les expressions unitermes Cette option extrait les mots uniques (unitermes) dans la mesure où le mot ne fait pas déjà partie d'un mot composé et si c'est un nom ou une catégorie grammaticale non reconnue.

Extraire les entités non linguistiques Cette option extrait les entités non linguistiques, telles que les numéros de téléphone, numéros de sécurité sociale, heures, dates, devises, chiffres, pourcentages, adresses électroniques, adresses HTTP, etc. Vous pouvez inclure ou exclure certains types d'entités non linguistiques dans la section **Entités non linguistiques : Configuration** de l'onglet Ressources avancées. Désactivez les entités dont vous n'avez pas besoin pour éviter au moteur d'extraction un temps de traitement inutile. Pour plus d'informations, voir «Configuration», à la page 208.

Algorithme des majuscules Cette option extrait les termes simples et composés ne figurant pas dans les dictionnaires intégrés, si la première lettre est une majuscule. Cette option offre une bonne manière d'extraire la plupart des noms propres.

Regrouper si possible les noms de personnes partiels et complets Cette option regroupe les noms qui apparaissent différemment dans le texte. Cette fonction est utile car les noms sont souvent cités dans leur forme complète au début du texte puis uniquement en version abrégée. Cette option essaye de faire correspondre tout uniterme ayant le type `<Unknown>` avec le dernier mot de tout terme composé entré comme `<Person>`. Par exemple, si *martin* est trouvé et initialement saisi comme `<Unknown>`, le moteur d'extraction vérifie si un terme composé de type `<Person>` contient *martin* comme dernier mot, tel que *pierre martin*. Cette option ne s'applique pas aux prénoms, car la plupart ne sont jamais extraits comme unitermes.

Taille maximale pour la permutation des mots utiles Cette option indique le nombre maximal de mots utiles pouvant être présents lorsque s'applique la technique de permutation. Cette technique de permutation regroupe les expressions similaires qui ne diffèrent les unes des autres que par la présence de mots utiles (par exemple, de et la), quelle que soit leur inflexion. Par exemple, si vous avez paramétré cette valeur sur au moins deux mots et si les deux expressions `représentants d'entreprise` et

représentants de l'entreprise ont été extraites. Dans ce cas, les deux termes extraits sont regroupés dans la liste de concepts finale car les deux termes sont considérés comme étant les mêmes lorsque de l' est ignoré.

Utiliser la dérivation des termes (grouper par composants multitermes) Pour le traitement de Big Data, sélectionnez cette option pour regrouper les multitermes à l'aide de règles de dérivation.

Remarque : Pour activer l'extraction des résultats de l'analyse des liens du texte, vous devez démarrer la session avec l'option **Exploration des résultats de l'analyse des liens du texte** et également choisir les ressources qui contiennent des définitions TLA. Vous pouvez toujours extraire les résultats TLA ultérieurement pendant une session de plan de travail interactif à partir de la boîte de dialogue Paramètres d'extraction. Pour plus d'informations, voir «Extraction de données», à la page 82.

Echantillonnage en amont pour gagner du temps

La durée de traitement d'une grande quantité de données peut aller de plusieurs minutes à plusieurs heures, particulièrement lors de l'utilisation d'une session de plan de travail interactif. Plus la taille des données est importante, plus la durée des processus d'extraction et de catégorisation est longue. Pour travailler plus efficacement, il est possible d'ajouter un noeud Echantillon d'IBM SPSS Modeler en amont de votre noeud Text Mining. Utilisez ce noeud échantillon pour prendre un échantillon aléatoire à l'aide d'un sous-ensemble de documents ou d'enregistrements moins important et d'effectuer les premiers transferts.

Un échantillon de taille moins importante est souvent parfaitement indiqué pour décider de la façon dont vous modifiez vos ressources et même créez la plupart, voire toutes, vos catégories. Une fois l'opération exécutée sur ce petit sous-ensemble de données et les résultats escomptés obtenus, vous pouvez appliquer la même technique pour créer des catégories dans l'ensemble de données entier. Vous pouvez ensuite rechercher des documents et des enregistrements qui ne font pas partie des catégories créées et faire les ajustements nécessaires.

Remarque : Le noeud Echantillon est un noeud IBM SPSS Modeler standard.

Utilisation du noeud Text Mining dans un flux

Le noeud modélisation Text Mining permet d'accéder aux données et d'extraire des concepts dans un flux. Vous pouvez utiliser n'importe quel noeud source pour accéder aux données, comme un noeud SGDB, Délimité, Flux de nouvelles ou Fixe. Un noeud liste fichiers peut être utilisé pour les textes résidant dans des documents externes.

Exemple 1 : Noeud liste fichiers et noeud Text Mining pour générer directement un nugget de modèle de concept

L'exemple suivant indique comment utiliser le noeud Liste fichiers, ainsi que le noeud modélisation Text Mining, pour générer la sortie du modèle de concepts. Pour plus d'informations sur l'utilisation du noeud liste fichiers, reportez-vous à «Noeud liste fichiers», à la page 11.

1. **Noeud Liste fichiers (onglet Paramètres).** Nous avons tout d'abord ajouté ce noeud au flux pour indiquer l'emplacement de stockage des documents texte. Nous avons sélectionné le répertoire contenant tous les documents sur lesquels nous souhaitons effectuer l'exploration de texte.
2. **Noeud Text Mining (onglet Champs).** Nous avons ensuite ajouté et connecté un noeud Text Mining au noeud liste fichiers. Dans ce noeud, nous avons défini le format d'entrée, le modèle de ressources et le format de sortie. Nous avons choisi le nom de champ créé à partir du noeud liste fichiers et sélectionné le champ de texte, ainsi que d'autres paramètres. Pour plus d'informations, voir «Utilisation du noeud Text Mining dans un flux».
3. **Noeud Text Mining (onglet Modèle)**. Dans l'onglet Modèle, nous avons ensuite sélectionné le mode de génération pour générer un nugget de modèle de concepts directement à partir de ce noeud. Vous pouvez sélectionner un modèle de ressources différent ou conserver les ressources de base.

Exemple 2 : Noeud fichier Excel et noeud Text Mining pour générer un modèle de catégorie de manière interactive

Cet exemple indique comment le noeud Text Mining peut également lancer une session de plan de travail interactif. Pour plus d'informations sur le plan de travail interactif, voir Chapitre 7, «Mode Plan de travail interactif», à la page 67.

1. **Noeud source Excel (onglet Données).** Nous avons tout d'abord ajouté ce noeud au flux pour indiquer l'emplacement de stockage du texte.
2. **Noeud Text Mining (onglet Champs).** Nous avons ensuite ajouté et connecté un noeud Text Mining. Dans le premier onglet, nous avons défini le format d'entrée de notre choix. Nous avons sélectionné un nom de champ du noeud source.
3. **Noeud Text Mining (onglet Modèle).** Dans l'onglet Modèle, nous avons ensuite décidé de créer un nugget de modèle de catégories de manière interactive et d'utiliser les résultats d'extraction pour créer automatiquement des catégories. Dans cet exemple, nous avons chargé une copie des ressources et un ensemble de catégories à partir d'un pack d'analyse de texte.
4. **Session de plan de travail interactif**. Nous avons ensuite exécuté le flux et l'interface du plan de travail interactif s'est ouverte. A l'issue d'une extraction, nous avons commencé à explorer nos données et à améliorer nos catégories.

Nugget de Text Mining : Modèle de concept

Un nugget de modèles de concepts Text Mining est créé lorsque vous parvenez à exécuter un noeud de modèle Text Mining pour lequel vous avez sélectionné l'option **Générer directement le modèle** dans l'onglet Modèle. Un nugget de modèles de concepts Text Mining est utilisé pour la recherche en temps réel de concepts-clés dans d'autres données textuelles, telles que les notes d'un centre d'appel.

Le nugget de modèle de concepts lui-même comprend une liste de concepts qui ont été affectés à des types. Vous pouvez sélectionner n'importe lequel des concepts de ce modèle pour le scoring en fonction d'autres données. Si vous exécutez un flux contenant un nugget de modèle Text Mining, de nouveaux champs sont ajoutés aux données en fonction du mode de génération sélectionné dans l'onglet Modèle du noeud modélisation Text Mining avant la création du modèle. Pour plus d'informations, voir «Modèle de concepts : onglet Modèle», à la page 32.

Si le nugget de modèle a été généré à l'aide de documents traduits, le scoring sera effectué dans la langue de traduction. De la même manière, si le nugget de modèle a été généré avec la langue Anglais, vous pouvez indiquer une langue de traduction dans le nugget de modèle, puisque les documents seront ensuite traduits en anglais.

Les nuggets de modèles Text Mining se trouvent dans la palette de nuggets de modèles (dans l'onglet Modèles situé dans la partie supérieure droite de la fenêtre IBM SPSS Modeler) lorsque ceux-ci sont générés.

Visualisation des résultats

Pour obtenir des informations sur le nugget de modèle, cliquez avec le bouton droit de la souris sur le noeud de la palette de nuggets de modèles, puis sélectionnez **Parcourir** dans le menu contextuel (ou **Editer** pour les noeuds du flux).

Ajout de modèles aux flux

Pour ajouter le nugget de modèle au flux, cliquez sur l'icône correspondante dans la palette de nuggets de modèles, puis dans l'espace de travail de flux à l'endroit où vous souhaitez placer le noeud. Vous pouvez également cliquer sur l'icône avec le bouton droit de la souris et sélectionner **Ajouter au flux** dans le menu contextuel. Il vous suffit alors de connecter votre flux au noeud pour pouvoir transmettre des données et générer des prévisions.

Avertissement : Si vous désirez utiliser un nugget de scoring pour régénérer un noeud de modélisation contenant à la fois le modèle de catégorie et le canevas utilisés, il est recommandé de créer un pack d'analyse de texte (TAP) et de l'utiliser dans une session interactive à la place du noeud de modélisation avant de générer le nugget de scoring.

Modèle de concepts : onglet Modèle

Dans les modèles de concepts, l'ensemble des concepts ayant fait l'objet d'une extraction figurent dans l'onglet Modèle. Les concepts sont présentés sous forme de tableau, avec une ligne par concept. Cet onglet permet de sélectionner les concepts qui seront utilisés pour le scoring.

Remarque : si vous avez généré un nugget de modèle de catégories, cet onglet contient des résultats différents. Pour plus d'informations, voir «Nugget de modèle de catégories : onglet Modèle», à la page 40.

Tous les concepts sont sélectionnés pour le scoring par défaut, comme l'indiquent les cases à cocher de la colonne située à l'extrême gauche. Si la case est cochée, le concept est utilisé pour le scoring. Si elle n'est pas cochée, il est exclu du scoring. Vous pouvez cocher plusieurs lignes à la fois en les sélectionnant toutes et en cliquant sur l'une des cases de la sélection.

Pour en savoir plus sur chaque concept, vous pouvez consulter les informations supplémentaires fournies dans chacune des colonnes suivantes :

Concept. Il s'agit de l'expression ou du mot principal extrait. Dans certains cas, ce concept représente le nom du concept, ainsi que d'autres termes sous-jacents associés à ce concept. Pour connaître les termes sous-jacents qui font partie d'un concept, affichez la sous-fenêtre des termes sous-jacents dans cet onglet et sélectionnez le concept pour consulter les termes correspondants situés au bas de la boîte de dialogue. Pour plus d'informations, voir «Termes sous-jacents dans les modèles de concepts», à la page 34.

Global. Global (fréquence) fait ici référence au nombre d'occurrences d'un concept (et de tous ses termes sous-jacents) dans l'ensemble des documents/enregistrements (fréquence).

- **Graphique à barres.** Fréquence globale de ce concept dans les données textuelles, présentée sous la forme d'un graphique à barres. La barre prend la couleur du type auquel le concept est affecté pour distinguer les types de manière visuelle.
- **%.** Fréquence globale de ce concept dans les données textuelles, présentée sous la forme d'un pourcentage.
- **N.** Nombre d'occurrences de ce concept dans les données texte.

Docs. Fait ici référence aux effectifs des documents ou des enregistrements dans lesquels le concept (et tous ses termes sous-jacents) apparaît.

- **Graphique à barres.** Effectifs des documents de ce concept présentés sous la forme d'un graphique à barres. La barre prend la couleur du type auquel le concept est affecté pour distinguer les types de manière visuelle.
- **%.** Effectifs des documents de ce concept présentés sous la forme d'un pourcentage.
- **N.** Nombre de documents ou d'enregistrements contenant ce concept.

Type. Type auquel le concept est affecté. Pour chaque concept, les colonnes Global et Docs arborent une couleur pour représenter le type auquel le concept est affecté. Un **type** correspond à un regroupement sémantique de concepts. Pour plus d'informations, voir «Dictionnaires de types», à la page 187.

Utilisation des concepts

Cliquez avec le bouton droit de la souris sur une cellule du tableau pour afficher un menu contextuel proposant les options suivantes :

- **Sélectionner tout.** Toutes les lignes du tableau sont sélectionnées.

- **Copier.** Le ou les concepts sélectionnés sont copiés dans le Presse-papiers.
- **Copier avec les champs** Les concepts sélectionnés sont copiés dans le Presse-papiers, de même que l'en-tête de colonne.
- **Cocher les éléments sélectionnés.** Coche les cases de toutes les lignes de tableau sélectionnées incluant ainsi ces concepts pour le scoring.
- **Désélectionner les éléments sélectionnés.** Désélectionne les cases de toutes les lignes de tableau sélectionnées.
- **Tout sélectionner.** Coche toutes les cases du tableau. Par conséquent, tous les concepts sont utilisés dans les résultats finaux.
- **Tout désélectionner.** Désélectionne toutes les cases du tableau. Les concepts désélectionnés ne sont pas utilisés dans les résultats finaux.
- **Inclure les concepts.** Affiche la boîte de dialogue Inclure les concepts. Pour plus d'informations, voir «Options pour inclure des concepts pour le scoring».

Options pour inclure des concepts pour le scoring

Pour activer ou désactiver les concepts qui seront utilisés pour le scoring, cliquez sur le bouton **Inclure les concepts** de la barre d'outils.

Figure 1. Bouton de la barre d'outils Inclure les concepts

En cliquant sur ce bouton de la barre d'outils, la boîte de dialogue Inclure les concepts s'ouvre et permet de sélectionner des concepts en fonction de règles. Tous les concepts cochés dans l'onglet Modèle seront inclus dans le scoring. Appliquer une règle dans cette boîte de dialogue secondaire afin de modifier les concepts qui seront utilisés pour le scoring.

Les options suivantes sont disponibles :

Sélectionner les concepts basés sur la fréquence la plus élevée. Plus grand nombre de concepts. Il s'agit du nombre de concepts qui seront cochés, en partant de celui dont la fréquence globale est la plus élevée. Le terme Fréquence fait ici référence au nombre d'occurrences d'un concept (et de tous ses termes sous-jacents) dans l'ensemble des documents/enregistrements. Il est parfois supérieur au nombre d'enregistrements car un concept peut figurer plusieurs fois dans un même enregistrement.

Sélectionner les concepts basés sur le nombre de documents. Nombre minimal. Il s'agit du nombre minimal de documents nécessaires pour l'activation des concepts. Le terme effectifs des documents fait ici référence au nombre de documents/d'enregistrements dans lesquels le concept (et tous ses termes sous-jacents) apparaît.

Vérifier les concepts affectés à ce type. Choisissez un type dans la liste déroulante pour sélectionner tous les concepts affectés à ce type. Les concepts sont automatiquement affectés aux types durant le processus d'extraction. Un **type** correspond à un regroupement sémantique de concepts. Les types comprennent des éléments tels que des concepts de niveau supérieur, des qualificatifs et des mots positifs et négatifs, des qualificatifs contextuels, des prénoms, des lieux, des organisations, etc. Pour plus d'informations, voir «Dictionnaires de types», à la page 187.

Désactiver les concepts apparaissant dans trop d'enregistrements. Pourcentage d'enregistrements. Désactiver les concepts apparaissant dans un pourcentage supérieur à celui indiqué pour le nombre d'enregistrements. Cette option permet d'exclure les concepts qui figurent fréquemment dans le texte ou les enregistrements, mais qui ne présentent pas d'intérêt pour l'analyse.

Désélectionner les concepts assignés au type. Désélectionne les concepts correspondant au type sélectionné dans la liste déroulante.

Termes sous-jacents dans les modèles de concepts

Vous pouvez consulter les termes sous-jacents définis pour les concepts que vous avez sélectionnés dans le tableau. Vous pouvez cliquer sur le bouton-bascule des termes sous-jacents dans la barre d'outils pour afficher le tableau des termes sous-jacents dans une sous-fenêtre distincte, située en bas de la boîte de dialogue.

Ces termes sous-jacents incluent les synonymes définis dans les ressources linguistiques (que ceux-ci aient été rencontrés ou non dans le texte) ainsi que toutes les formes extraites au pluriel/singulier trouvées dans le texte qui sont utilisées pour générer le nugget de modèles, les termes permutés, les termes provenant du regroupement flou, etc.

Figure 2. Bouton de la barre d'outils Afficher les termes sous-jacents

Remarque : vous ne pouvez pas éditer les termes sous-jacents. Cette liste est générée par des substitutions, des définitions de synonymes (dans le dictionnaire de substitution), des regroupements par ressemblance, etc., définis dans les ressources linguistiques. Pour modifier la manière dont les termes sont regroupés sous un concept et comment ceux-ci sont manipulés, vous devez effectuer les modifications directement dans les ressources (dans l'Editeur de ressources dans le plan de travail interactif ou dans l'Editeur de modèle puis recharger dans le noeud) et réexécuter le flux pour obtenir un nouveau nugget de modèle comprenant les résultats mis à jour.

Cliquez avec le bouton droit de la souris sur une cellule comprenant un terme sous-jacent ou un concept pour afficher un menu contextuel proposant les options suivantes :

- **Copier.** La cellule sélectionnée est copiée dans le presse-papiers.
- **Copier avec les champs.** La cellule sélectionnée est copiée dans le presse-papiers, de même que les en-têtes de colonne.
- **Sélectionner tout.** Toutes les cellules du tableau sont sélectionnées.

Modèle de concepts : onglet Paramètres

L'onglet Paramètres sert à définir la valeur du champ de texte des nouvelles données d'entrée, si nécessaire. Il permet également d'indiquer le modèle de données des résultats (mode de scoring).

Remarque : Cet onglet n'apparaît que si le nugget de modèle est placé sur le canevas. Il n'existe pas si vous accédez à cette boîte de dialogue directement à partir de la palette Modèles.

Mode de Scoring : concepts en tant qu'enregistrements

Grâce à ce mode de scoring, un nouvel enregistrement est créé pour chaque paire de `concepts/documents`. Généralement, la sortie comporte plus d'enregistrements que n'en comportait l'entrée.

Outre les champs d'entrée, les nouveaux champs suivants sont ajoutés aux données :

Tableau 4. Champs de sortie de l'option Concepts en tant qu'enregistrements.

Champ	Description
`Concept`	Contient le nom de concept extrait qui figure dans le champ des données textuelles.
`Type`	Indique le type du concept sous la forme d'un nom de type complet, comme *Location* ou *Person*. Un **type** correspond à un regroupement sémantique de concepts. Pour plus d'informations, voir «Dictionnaires de types», à la page 187.
`Count`	Affiche le nombre d'occurrences de ce concept (et de ses termes sous-jacents) dans le corps du texte (enregistrement/document).

Lorsque vous sélectionnez cette option, toutes les autres options à l'exception de **Traitement des erreurs de ponctuation** sont désactivées.

Mode de Scoring : concepts en tant que champs

Dans les modèles de concepts, pour chaque enregistrement d'entrée, un enregistrement est créé pour chaque concept trouvé dans un document donné. Ainsi, les enregistrements de sortie sont aussi nombreux que les enregistrements d'entrée. Désormais, chaque enregistrement (ligne) comporte toutefois un nouveau champ (colonne) pour chaque concept sélectionné (coché) dans l'onglet Modèle. La valeur du champ de chaque concept dépend de la valeur de champ que vous sélectionnez dans cet onglet (**Indicateurs** ou **Effectifs**).

Remarque : Si vous utilisez des jeux de données particulièrement volumineux, par exemple avec une base de données Db2, l'utilisation de **Concepts en tant que champs** peut rencontrer des problèmes de traitement en raison du volume des données. Dans ce cas, il est recommandé d'utiliser à la place **Concepts en tant qu'enregistrements**.

Valeurs de champs. Sélectionnez si vous souhaitez que le nouveau champ de chaque concept contienne un effectif ou une valeur d'indicateur.

- **Indicateurs.** Cette option permet d'obtenir des indicateurs avec deux valeurs distinctes en sortie, par exemple *Oui/Non*, *Vrai/Faux*, *V/F*, ou *1* et *2*. Les types de stockage sont automatiquement définis pour refléter les valeurs choisies. Par exemple, si vous entrez des valeurs numériques pour les indicateurs, elles seront automatiquement traitées comme valeurs entières. Les indicateurs disposent des types de stockage suivants : chaîne, entier, nombre réel ou date/heure. Entrez une valeur d'indicateur pour **Vrai** et pour **Faux**.
- **Comptages.** Valeur utilisée pour obtenir le nombre d'occurrences du concept dans un enregistrement donné.

Extension nom de champ. Spécifiez l'extension du nom de champ. Les noms de champ générés reprennent le nom de concept et l'extension.

- **Ajouter en tant que.** Spécifiez à quel emplacement du nom de champ l'extension doit être ajoutée. Choisissez **Préfixe** pour ajouter l'extension en début de chaîne. Choisissez **Suffixe** pour ajouter l'extension en fin de chaîne.

Adapter les erreurs de ponctuation. Cette option normalise temporairement le texte contenant des erreurs de ponctuation (par exemple, usage incorrect) au cours de l'extraction pour améliorer l'extraction des concepts. Cette option s'avère extrêmement utile lorsque le texte est court et de qualité médiocre (réponses ouvertes, messages électroniques, données CRM, etc.) ou qu'il contient de nombreuses abréviations.

Modèle de concepts : onglet Champs

L'onglet Champs définit la valeur du champ de texte des nouvelles données d'entrée, si nécessaire.

Remarque : Cet onglet n'apparaît que si le nugget de modèle est placé dans le flux. Il n'apparaît pas lorsque vous accédez à cette sortie directement à partir de la palette Modèles.

Champ Texte. Sélectionnez le champ contenant le texte à explorer. Ce champ dépend de la source de données.

Type de document. Le type de document indique la structure du texte. Sélectionnez l'un des types suivants :

- **Texte réel.** S'utilise pour la plupart des documents ou des sources textuelles. L'ensemble de texte entier est analysé pour l'extraction. Contrairement aux autres options, il n'existe pas de paramètres supplémentaires pour cette option.
- **Texte structuré.** Utilisez ce format pour les bibliographies, les brevets et les fichiers contenant des structures standard qui peuvent être identifiées et analysées. Ce type de document est utilisé pour éviter tout ou une partie du processus d'extraction. Il permet de définir des séparateurs de termes, d'affecter des types et d'imposer une valeur de fréquence minimale. Si vous sélectionnez cette option, vous devez cliquer sur le bouton **Paramètres** et entrer les séparateurs de texte dans la zone **Formatage de texte structuré** de la boîte de dialogue Paramètres du document. Pour plus d'informations, voir «Paramètres de document de l'onglet Champs», à la page 22.

Codage en entrée. Cette option n'est disponible que si vous indiquez que le champ de texte correspond au **chemin d'accès des documents**. Elle indique le codage de texte par défaut. Le codage spécifié ou reconnu est converti en un codage `ISO-8859-1`. Ainsi, même si vous indiquez un autre codage, le moteur d'extraction le remplace par le codage `ISO-8859-1` avant de traiter le texte. Les caractères qui ne figurent pas dans la définition du codage `ISO-8859-1` sont convertis en espaces.

Langue du texte. Identifie la langue du texte en cours d'exploration. Il s'agit de la langue principale détectée pendant l'extraction. Si vous souhaitez acquérir la licence d'une langue prise en charge à laquelle vous n'avez pas accès actuellement, contactez votre représentant commercial.

Modèle de concepts : onglet Récapitulatif

L'onglet Récapitulatif contient des informations sur le modèle lui-même (dossier *Analyse*), sur les champs utilisés dans le modèle (dossier *Champs*), sur les paramètres utilisés pour la construction du modèle (dossier *Créer des paramètres*), ainsi que sur l'apprentissage du modèle (dossier *Récapitulatif de l'apprentissage*).

Lorsque vous accédez pour la première fois à un noeud modélisation, l'arborescence des dossiers de l'onglet Récapitulatif est réduite. Pour afficher les résultats qui vous intéressent, utilisez la commande de développement à gauche du dossier ou cliquez sur le bouton **Développer tout** pour afficher tous les résultats. Pour masquer les résultats après les avoir consultés, utilisez la commande de développement pour réduire le dossier voulu ou cliquez sur le bouton **Réduire tout** pour réduire tous les dossiers.

Utilisation des nuggets de modèle de concepts dans un flux

Lors de l'utilisation d'un noeud modélisation Text Mining, vous pouvez générer soit un nugget de modèle de concepts soit un nugget de modèle de catégories (dans la session de plan de travail interactif). L'exemple suivant indique comment utiliser un modèle de concepts dans un flux simple.

Exemple : noeud Fichier statistiques avec le nugget de modèle de concept

L'exemple suivant indique comment utiliser le nugget de modèle de concepts Text Mining.

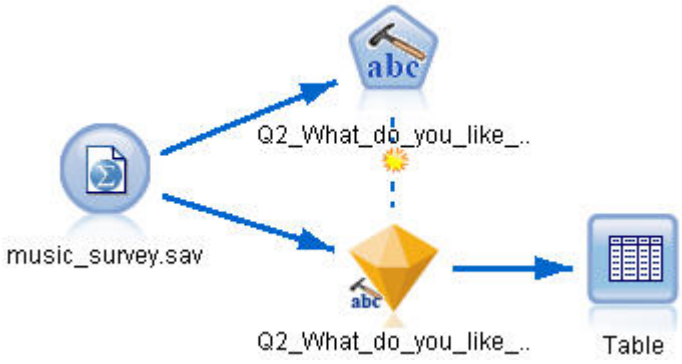


Figure 3. Exemple de flux : noeud Fichier statistiques avec un nugget de modèle de concept Text Mining

1. **Noeud Fichier statistiques (onglet Données).** Nous avons tout d'abord ajouté ce noeud au flux pour indiquer l'emplacement de stockage des documents texte.

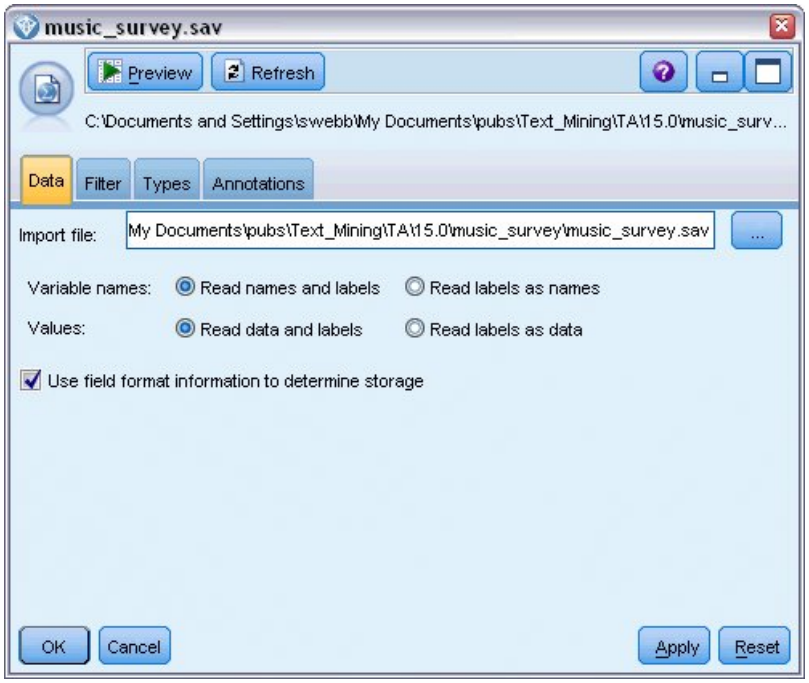


Figure 4. Boîte de dialogue Noeud Fichier statistiques (onglet Données).

2. **Nugget de modèle de concept Text Mining (onglet Modèle).** Nous avons ensuite ajouté et connecté un nugget de modèle de concepts au noeud Fichier Statistiques. Nous avons sélectionné les concepts que nous souhaitions utiliser pour scorer nos données.

Figure 5. Boîte de dialogue Nugget de modèle Text Mining (onglet Modèle)

3. **Nugget de modèle de concept Text Mining (onglet Paramètres).** Nous avons ensuite défini le format de sortie et sélectionné *Concepts en tant que champs*. Un nouveau champ sera créé dans la sortie pour chaque concept sélectionné dans l'onglet Modèle. Chaque nom de champ sera composé du nom du concept et du préfixe « *Concept_* »

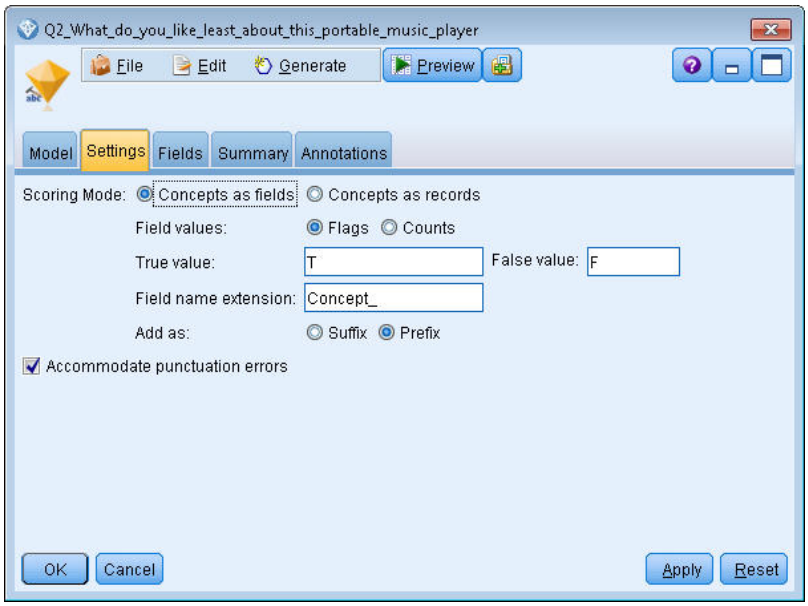


Figure 6. Boîte de dialogue de nugget de modèle de concepts Text Mining (onglet Paramètres)

4. **Nugget de modèle de concepts Text Mining (onglet Champs).** Nous avons ensuite sélectionné le champ de texte, **Q2_What_do_you_like_least_about_this_portable_music_player**, qui est le nom de champ provenant du noeud Fichier Statistiques. Nous avons également sélectionné l'option **Le champ Texte représente : Texte réel**.

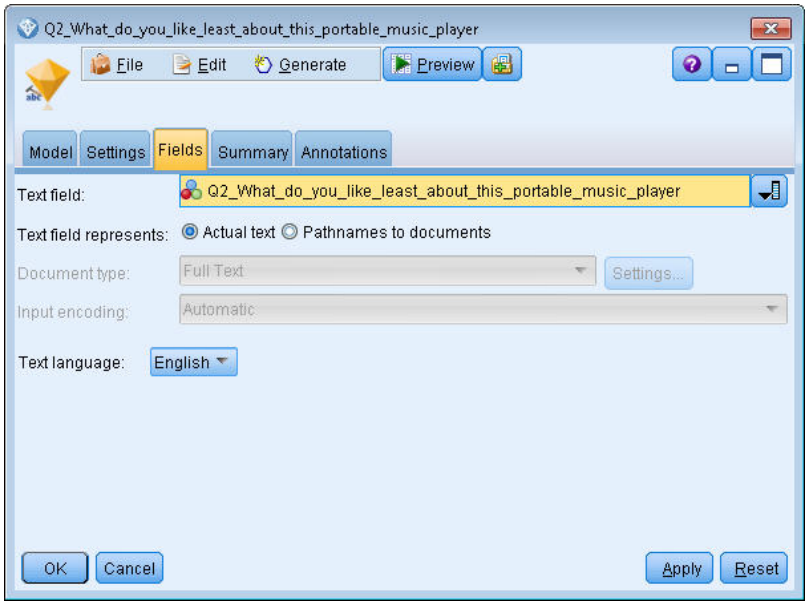
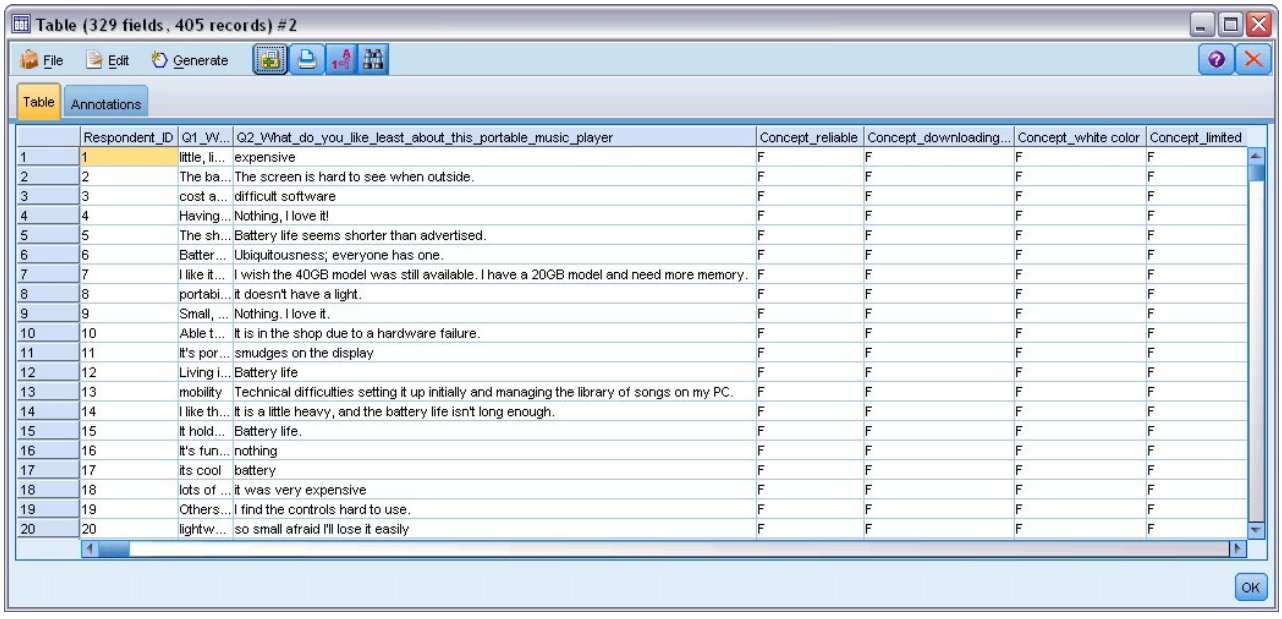


Figure 7. Boîte de dialogue de nugget de modèle de concepts Text Mining (onglet Champs)

5. **Noeud Table.** Nous avons ensuite associé un noeud de table pour afficher les résultats et exécuté le flux. La sortie apparaît à l'écran.



	Respondent_ID	Q1_W...	Q2_What_do_you_like_least_about_this_portable_music_player	Concept_reliable	Concept_downloading...	Concept_white color	Concept_limited
1	1	little, li...	expensive	F	F	F	F
2	2	The ba...	The screen is hard to see when outside.	F	F	F	F
3	3	cost a...	difficult software	F	F	F	F
4	4	Having...	Nothing, I love it!	F	F	F	F
5	5	The sh...	Battery life seems shorter than advertised.	F	F	F	F
6	6	Batter...	Ubiquitousness; everyone has one.	F	F	F	F
7	7	I like it...	I wish the 40GB model was still available. I have a 20GB model and need more memory.	F	F	F	F
8	8	portabi...	it doesn't have a light.	F	F	F	F
9	9	Small, ...	Nothing. I love it.	F	F	F	F
10	10	Able t...	It is in the shop due to a hardware failure.	F	F	F	F
11	11	It's por...	smudges on the display	F	F	F	F
12	12	Living i...	Battery life	F	F	F	F
13	13	mobility	Technical difficulties setting it up initially and managing the library of songs on my PC.	F	F	F	F
14	14	I like th...	It is a little heavy, and the battery life isn't long enough.	F	F	F	F
15	15	It hold...	Battery life.	F	F	F	F
16	16	It's fun...	nothing	F	F	F	F
17	17	its cool	battery	F	F	F	F
18	18	lots of ...	it was very expensive	F	F	F	F
19	19	Others...	I find the controls hard to use.	F	F	F	F
20	20	lightw...	so small afraid I'll lose it easily	F	F	F	F

Figure 8. Nous avons fait défiler les tables de résultats pour afficher les indicateurs de concepts

Nugget de Text Mining : Modèle de catégorie

Un nugget de modèle de catégories Text Mining est créé lorsque vous générez un modèle de catégories dans le plan de travail interactif. Ce nugget de modélisation contient un ensemble de catégories dont la définition se compose de concepts, de types, de motifs TLA et/ou de règles de catégorie. Le nugget permet de regrouper en catégories des réponses à des enquêtes, des entrées de blogue, d'autres flux de nouvelles ou d'autres données textuelles.

Si vous lancez une session de plan de travail interactif dans le noeud modélisation, vous pouvez explorer les résultats de l'extraction, adapter les ressources, mettre au point vos catégories avant de générer des modèles de catégories. Si vous exécutez un flux contenant un nugget de modèle Text Mining, de nouveaux champs sont ajoutés aux données en fonction du mode de génération sélectionné dans l'onglet Modèle du noeud modélisation Text Mining avant la création du modèle. Pour plus d'informations, voir «Nugget de modèle de catégories : onglet Modèle».

Si le nugget de modèle a été généré à l'aide de documents traduits, le scoring sera effectué dans la langue de traduction. De la même manière, si le nugget de modèle a été généré avec la langue Anglais, vous pouvez indiquer une langue de traduction dans le nugget de modèle, puisque les documents seront ensuite traduits en anglais.

Les nuggets de modèles Text Mining se trouvent dans la palette de nuggets de modèles (dans l'onglet Modèles situé dans la partie supérieure droite de la fenêtre IBM SPSS Modeler) lorsque ceux-ci sont générés.

Visualisation des résultats

Pour obtenir des informations sur le nugget de modèle, cliquez avec le bouton droit de la souris sur le noeud de la palette de nuggets de modèles, puis sélectionnez **Parcourir** dans le menu contextuel (ou **Editer** pour les noeuds du flux).

Ajout de modèles aux flux

Pour ajouter le nugget de modèle au flux, cliquez sur l'icône correspondante dans la palette de nuggets de modèles, puis dans l'espace de travail de flux à l'endroit où vous souhaitez placer le noeud. Vous pouvez également cliquer sur l'icône avec le bouton droit de la souris et sélectionner **Ajouter au flux** dans le menu contextuel. Il vous suffit alors de connecter votre flux au noeud pour pouvoir transmettre des données et générer des prévisions.

Avertissement : Si vous désirez utiliser un nugget de scoring pour régénérer un noeud de modélisation contenant à la fois le modèle de catégorie et le canevas utilisés, il est recommandé de créer un pack d'analyse de texte (TAP) et de l'utiliser dans une session interactive à la place du noeud de modélisation avant de générer le nugget de scoring.

Nugget de modèle de catégories : onglet Modèle

Dans le cas des modèles de catégories, l'onglet Modèle affiche, à gauche, la liste des catégories du modèle de catégories et, à droite, les descripteurs d'une catégorie sélectionnée. Chaque catégorie comprend un certain nombre de descripteurs. Pour chaque catégorie sélectionnée, les descripteurs associés apparaissent dans la table. Ces descripteurs peuvent comporter des concepts, des règles de catégorie, des types et des motifs TLA. Le type de chaque descripteur, ainsi que des exemples de ce que chaque descripteur représente, y figurent également.

Dans cet onglet, l'objectif est de sélectionner les catégories que vous souhaitez utiliser pour le scoring. Dans un modèle de catégories, le scoring des documents et des enregistrements s'effectue par catégorie. Si un document ou un enregistrement contient au moins un descripteur dans son texte ou des termes sous-jacents, ce document ou cet enregistrement est alors affecté à la catégorie auquel le descripteur

appartient. Ces termes sous-jacents incluent les synonymes définis dans les ressources linguistiques (que ceux-ci aient été rencontrés ou non dans le texte) ainsi que tous les termes extraits au pluriel/singulier trouvés dans le texte qui sont utilisées pour générer le nugget de modèles, les termes permutés, les termes provenant du regroupement flou, etc.

Remarque : si vous avez généré un nugget de modèle de concept, cet onglet contient des résultats différents. Pour plus d'informations, voir «Modèle de concepts : onglet Modèle», à la page 32.

Arborescence des catégories

Pour en savoir plus sur chaque catégorie, sélectionnez la catégorie de votre choix et passez en revue les informations correspondant aux descripteurs de cette catégorie. Pour chaque descripteur, vous pouvez consulter les informations suivantes :

- Nom du **Descripteur**. Ce champ contient une icône représentant le type de descripteur et indiquant le nom de ce dernier.

Tableau 5. Icônes du Descripteur

	Concepts		Motifs TLA
	Types		Règles de catégorie

- **Type.** Ce champ contient le nom du type du descripteur. Les types sont des regroupements de concepts similaires (regroupements sémantiques), tels que les noms d'organisation, les produits ou les opinions positives. Les règles ne sont affectées à aucun type.
- **Détails.** Ce champ contient une liste du contenu du descripteur. En fonction du nombre de correspondances, la liste complète de chaque descripteur risque de ne pas s'afficher entièrement en raison de la taille limitée de la boîte de dialogue.

Sélection et copie de catégories

Toutes les catégories de niveau supérieur sont sélectionnées par défaut pour le scoring, comme l'indiquent les cases à cocher de la sous-fenêtre de gauche. Si la case est cochée, la catégorie est utilisée pour le scoring. Si elle n'est pas cochée, la catégorie sera exclue du scoring. Vous pouvez cocher plusieurs lignes à la fois en les sélectionnant toutes et en cliquant sur l'une des cases de la sélection. Ainsi si une catégorie ou une sous-catégorie est sélectionnée mais que l'une de ses sous-catégories n'est pas sélectionnée, la case affiche un fond bleu indiquant que la sélection des enfants est seulement partielle dans la catégorie sélectionnée.

Cliquez avec le bouton droit de la souris sur une catégorie de l'arborescence pour afficher un menu contextuel proposant les options suivantes :

- **Cocher les éléments sélectionnés.** Coche les cases de toutes les lignes de tableau sélectionnées.
- **Désélectionner les éléments sélectionnés.** Désélectionne les cases de toutes les lignes de tableau sélectionnées.
- **Tout sélectionner.** Coche toutes les cases du tableau. Par conséquent, toutes les catégories sont utilisées dans les résultats finaux. Vous pouvez également utiliser l'icône de la case correspondante sur la barre d'outils.
- **Tout désélectionner.** Désélectionne toutes les cases du tableau. Si vous désélectionnez une catégorie, celle-ci ne sera pas utilisée dans les résultats finaux. Vous pouvez également utiliser l'icône de la case vide correspondante sur la barre d'outils.

Cliquez avec le bouton droit de la souris sur une cellule du tableau Descripteurs pour afficher un menu contextuel proposant les options suivantes :

- **Copier.** Le ou les concepts sélectionnés sont copiés dans le Presse-papiers.
- **Copier avec les champs.** Le descripteur sélectionné est copié dans le Presse-papiers, de même que les en-têtes de colonne.
- **Sélectionner tout.** Toutes les lignes du tableau sont sélectionnées.

Nugget de modèle de catégories : onglet Paramètres

L'onglet Paramètres sert à définir la valeur du champ de texte des nouvelles données d'entrée, si nécessaire. Il permet également d'indiquer le modèle de données des résultats (mode de scoring).

Remarque : Cet onglet n'apparaît dans la boîte de dialogue du noeud que si le nugget de modèle est placé sur le canevas ou dans un flux. Il n'apparaît pas lorsque vous accédez à ce nugget directement à partir de la palette Modèles.

Mode de scoring : catégories en tant que champs

Grâce à cette option, les enregistrements de sortie sont aussi nombreux que les enregistrements d'entrée. Désormais, chaque enregistrement comporte toutefois un nouveau champ pour chaque catégorie sélectionnée (cochée) dans l'onglet Modèle. Pour chaque champ, entrez une valeur d'indicateur pour **Vrai** et pour **Faux**, par exemple*Oui/Non*, *Vrai/Faux*, *V/F* ou *1* et 2. Les types de stockage sont automatiquement définis pour refléter les valeurs choisies. Par exemple, si vous entrez des valeurs numériques pour les indicateurs, elles seront automatiquement traitées comme valeurs entières. Les indicateurs disposent des types de stockage suivants : chaîne, entier, nombre réel ou date/heure.

Remarque : Si vous utilisez des jeux de données particulièrement volumineux, par exemple avec une base de données Db2, l'utilisation de **Catégories en tant que champs** peut rencontrer des problèmes de traitement en raison du volume des données. Dans ce cas, il est recommandé d'utiliser à la place **Catégories en tant qu'enregistrements**.

Extension nom de champ. Vous pouvez choisir de spécifier une extension préfixe/suffixe pour le nom de champ ou vous pouvez choisir d'utiliser les codes de catégories. Les noms de champ générés reprennent le nom de catégorie et l'extension.

- **Ajouter en tant que.** Spécifiez à quel emplacement du nom de champ l'extension doit être ajoutée. Choisissez **Préfixe** pour ajouter l'extension en début de chaîne. Choisissez **Suffixe** pour ajouter l'extension en fin de chaîne.

Si une sous-catégorie n'est pas sélectionnée. Cette option permet de déterminer comment seront manipulés les descripteurs appartenant aux sous-catégories qui n'étaient pas sélectionnées pour le scoring. Il y a deux options.

- Avec l'option **Exclure complètement ses descripteurs de l'évaluation**, les descripteurs des sous-catégories qui ne sont pas cochées (qui ne sont pas sélectionnées) seront ignorés et ne seront pas utilisés pendant le scoring.
- Avec l'option **Agréger les descripteurs avec ceux de la catégorie parent**, les descripteurs des sous-catégories qui ne sont pas cochées (qui ne sont pas sélectionnées) seront utilisés comme descripteurs pour la catégorie parent (la catégorie au-dessus de cette sous-catégorie). S'il existe plusieurs niveaux de sous-catégories et que celles-ci ne sont pas sélectionnées, les descripteurs seront reportés sous la première catégorie parent disponible.

Adapter les erreurs de ponctuation. Cette option normalise temporairement le texte contenant des erreurs de ponctuation (par exemple, usage incorrect) au cours de l'extraction pour améliorer l'extraction des concepts. Cette option s'avère extrêmement utile lorsque le texte est court et de qualité médiocre (réponses ouvertes, messages électroniques, données CRM, etc.) ou qu'il contient de nombreuses abréviations.

Mode de Scoring : catégories en tant qu'enregistrements

Grâce à cette option, un nouvel enregistrement est créé pour chaque paire de `catégories/documents`. Généralement, la sortie comporte plus d'enregistrements que n'en comportait l'entrée. Outre les champs d'entrée, de nouveaux champs sont également ajoutés aux données en fonction du type de modèle dont il s'agit.

Tableau 6. Champs de sortie de l'option Catégories en tant qu'enregistrements.

Nouveau champ de sortie	Description
`catégorie`	Indique le nom de la catégorie à laquelle le document texte a été affecté. Si la catégorie est une sous-catégorie d'une autre catégorie, alors le chemin complet au nom de catégorie est contrôlé par la valeur choisie dans cette boîte de dialogue.

Valeurs des catégories hiérarchiques. Cette option contrôle le mode d'affichage des noms de sous-catégories dans les résultats.

- **Chemin d'accès complet à la catégorie.** Cette option va générer le nom de la catégorie et le chemin complet aux catégories parents le cas échéant en utilisant des barres obliques pour séparer les noms de catégories des noms de sous-catégories.
- **Chemin d'accès court à la catégorie.** Cette option va générer seulement le nom de la catégorie mais utilise des point de suspension pour afficher le nombre de catégories parents pour la catégorie en question.
- **Catégorie de niveau le plus bas.** Cette option va générer seulement le nom de la catégorie sans afficher le chemin complet ou les catégories parents.

Si une sous-catégorie n'est pas sélectionnée. Cette option permet de déterminer comment seront manipulés les descripteurs appartenant aux sous-catégories qui n'étaient pas sélectionnées pour le scoring. Il y a deux options.

- Avec l'option **Exclure complètement ses descripteurs de l'évaluation**, les descripteurs des sous-catégories qui ne sont pas cochées (qui ne sont pas sélectionnées) seront ignorés et ne seront pas utilisés pendant le scoring.
- Avec l'option **Agréger les descripteurs avec ceux de la catégorie parent**, les descripteurs des sous-catégories qui ne sont pas cochées (qui ne sont pas sélectionnées) seront utilisés comme descripteurs pour la catégorie parent (la catégorie au-dessus de cette sous-catégorie). S'il existe plusieurs niveaux de sous-catégories et que celles-ci ne sont pas sélectionnées, les descripteurs seront reportés sous la première catégorie parent disponible.

Adapter les erreurs de ponctuation. Cette option normalise temporairement le texte contenant des erreurs de ponctuation (par exemple, usage incorrect) au cours de l'extraction pour améliorer l'extraction des concepts. Cette option s'avère extrêmement utile lorsque le texte est court et de qualité médiocre (réponses ouvertes, messages électroniques, données CRM, etc.) ou qu'il contient de nombreuses abréviations.

Nuggets de modèle de catégories : autres onglets

L'onglet Champs et l'onglet Paramètres du nugget de modèle de catégories sont identiques à ceux du nugget de modèle de concepts.

- Onglet Champs. Pour plus d'informations, voir «Modèle de concepts : onglet Champs», à la page 35.
- Onglet Récapitulatif. Pour plus d'informations, voir «Modèle de concepts : onglet Récapitulatif», à la page 36.

Utilisation des nuggets de modèle de catégories dans un flux

Le nugget de modèle de catégories Text Mining est généré à partir d'une session de plan de travail interactif. Vous pouvez utiliser ce nugget de modèle dans un flux.

Exemple : le noeud Fichier statistiques avec le nugget de modèle de catégories

L'exemple suivant indique comment utiliser le nugget de modèle Text Mining.

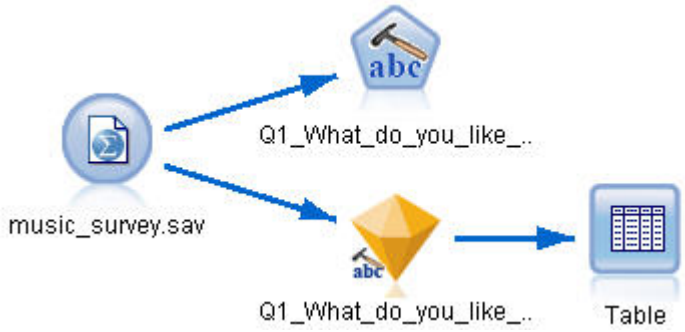


Figure 9. Exemple de flux : noeud Fichier statistiques avec un nugget de modèle de catégories Text Mining

1. **Noeud Fichier statistiques (onglet Données).** Nous avons tout d'abord ajouté ce noeud au flux pour indiquer l'emplacement de stockage des documents texte.

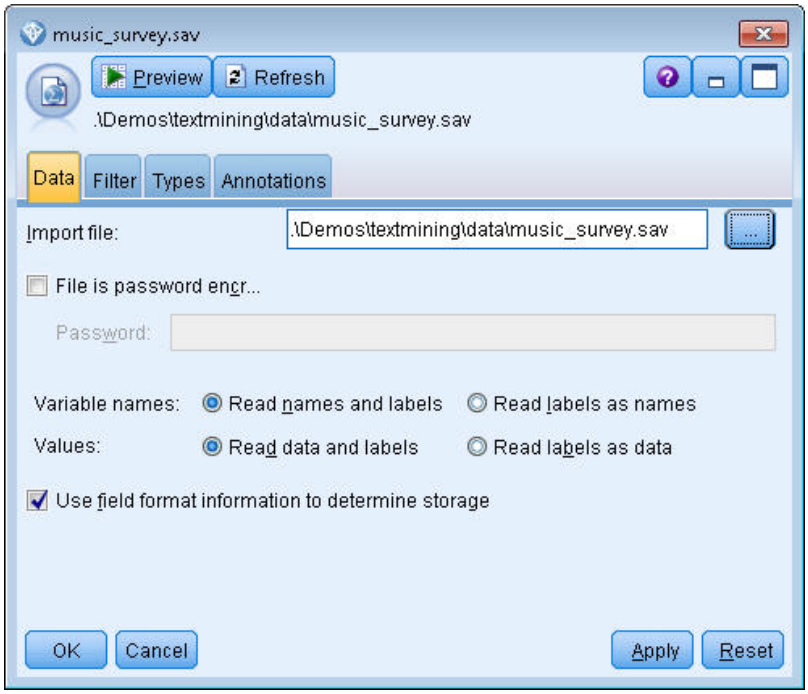


Figure 10. Boîte de dialogue Noeud Fichier statistiques (onglet Données).

2. **Nugget de modèle de catégories Text Mining (onglet Modèle).** Nous avons ensuite ajouté et connecté un nugget de modèle de catégories au noeud Fichier Statistiques. Nous avons sélectionné les catégories que nous souhaitions utiliser pour scorer nos données.

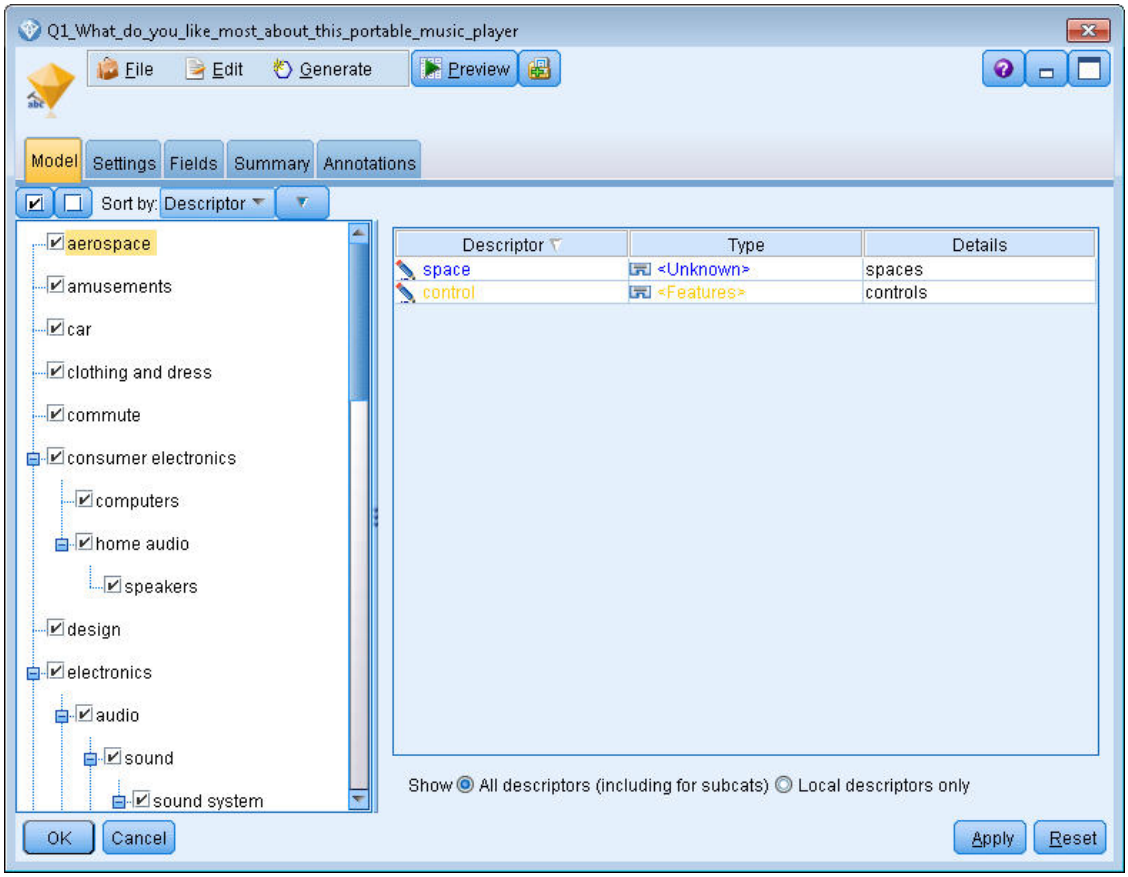


Figure 11. Boîte de dialogue Nugget de modèle Text Mining (onglet Modèle)

3. **Nugget de modèle Text Mining (onglet Paramètres).** Nous avons ensuite défini le format de sortie **Catégories en tant que champs**.

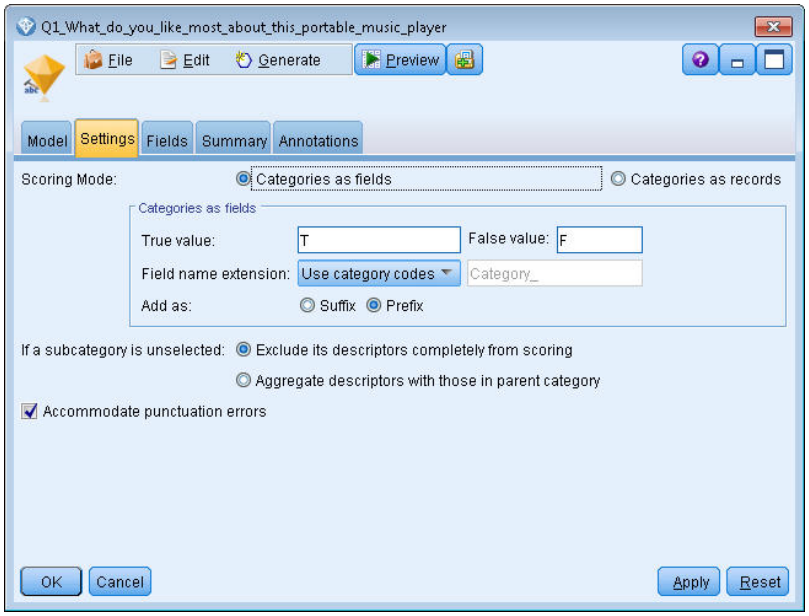


Figure 12. Boîte de dialogue Nugget de modèle de catégories (onglet Paramètres)

4. **Nugget de modèle de catégorie Text Mining (onglet Champs).** Nous avons ensuite sélectionné la variable de champ de texte, correspondant au nom du champ issu du noeud Fichier statistiques, et sélectionné l'option Champ de texte qui correspond au **Texte réel**, ainsi que d'autres paramètres.

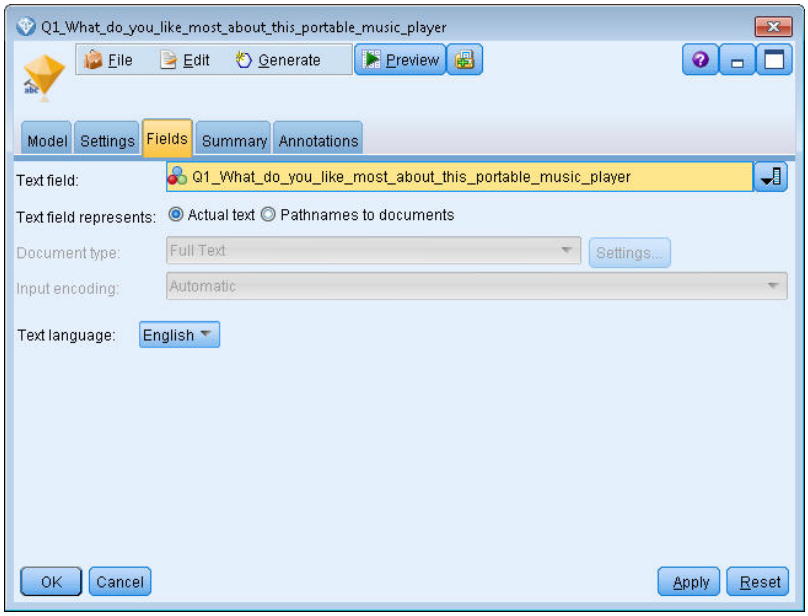
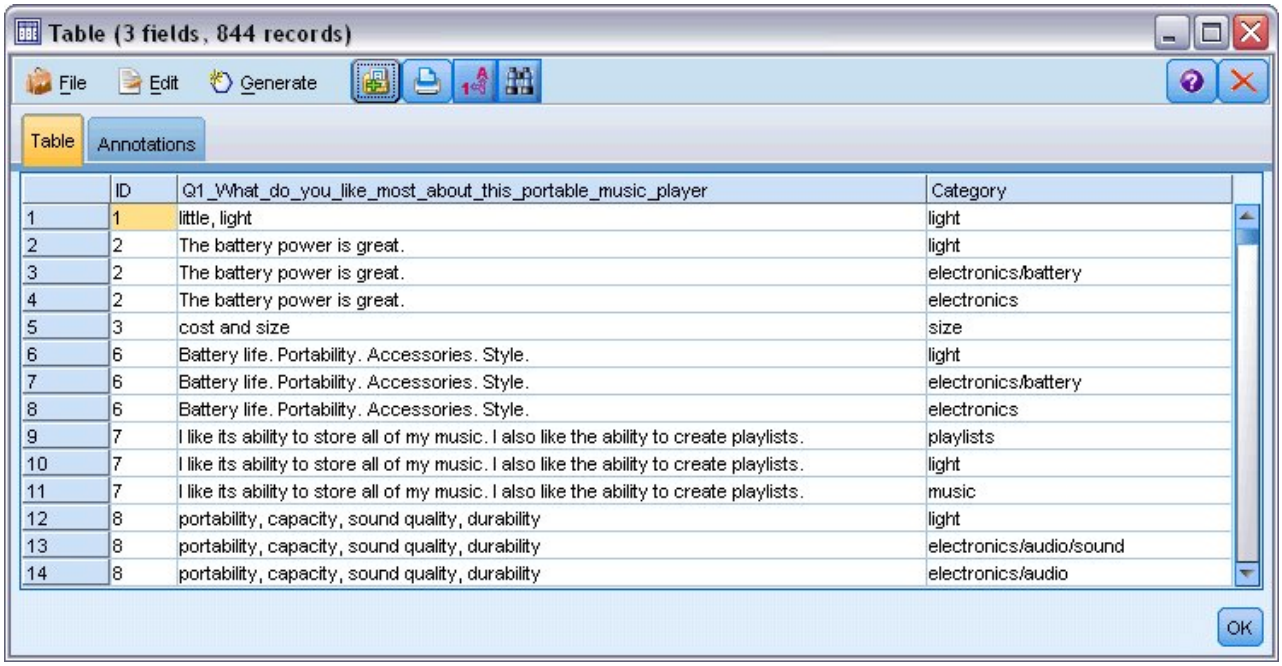


Figure 13. Boîte de dialogue de nugget de modèle Text Mining (onglet Champs)

5. **Noeud Table.** Nous avons ensuite associé un noeud de table pour afficher les résultats et exécuté le flux.



	ID	Q1_What_do_you_like_most_about_this_portable_music_player	Category
1	1	little, light	light
2	2	The battery power is great.	light
3	2	The battery power is great.	electronics/battery
4	2	The battery power is great.	electronics
5	3	cost and size	size
6	6	Battery life. Portability. Accessories. Style.	light
7	6	Battery life. Portability. Accessories. Style.	electronics/battery
8	6	Battery life. Portability. Accessories. Style.	electronics
9	7	I like its ability to store all of my music. I also like the ability to create playlists.	playlists
10	7	I like its ability to store all of my music. I also like the ability to create playlists.	light
11	7	I like its ability to store all of my music. I also like the ability to create playlists.	music
12	8	portability, capacity, sound quality, durability	light
13	8	portability, capacity, sound quality, durability	electronics/audio/sound
14	8	portability, capacity, sound quality, durability	electronics/audio

Figure 14. Table de résultats

Chapitre 4. Exploration des liens du texte

Noeud Analyse des liens du texte

Le noeud analyse des liens du texte (TLA) ajoute une technologie de mise en correspondance de motifs à l'extraction de concepts d'exploration de texte, de façon à identifier les relations entre les concepts des données textuelles sur la base de motifs connus. Ces relations peuvent décrire l'impression d'un client vis-à-vis d'un produit, les organisations qui travaillent ensemble et même les relations entre des gènes ou des agents pharmaceutiques.

Par exemple, l'intérêt de l'extraction du nom du produit de votre concurrent peut être limité. Grâce à ce noeud, vous pouvez également savoir comment les gens perçoivent le produit, du moins si ces opinions sont exprimées dans les données. Pour identifier et extraire les relations et les associations, les données textuelles sont comparées à des motifs connus.

Vous pouvez utiliser les règles de motifs TLA de certains modèles de ressources livrés avec IBM SPSS Modeler Text Analytics ou créer/modifier vos propres motifs. Les règles de motifs sont constituées de macros, de listes de mots et d'intervalles de mots pour former une requête booléenne, ou règle, qui est comparée à votre texte d'entrée. Lorsqu'une règle de patron TLA correspond au texte, il est possible d'extraire ce texte sous la forme d'un résultat TLA et de le restructurer sous la forme de données de sortie. Pour plus d'informations, voir Chapitre 18, «A propos des règles des liens du texte», à la page 213.

Le noeud analyse des liens du texte propose une méthode plus directe pour identifier les résultats de motifs TLA, les extraire du texte et ajouter leurs résultats à l'ensemble de données figurant dans le flux. Mais le noeud analyse des liens du texte ne constitue pas le seul moyen d'exécuter une analyse des liens du texte. Vous pouvez également lancer une session de plan de travail interactif dans le noeud modélisation Text Mining.

Dans le plan de travail interactif, vous pouvez explorer les résultats de motifs TLA et les utiliser sous la forme de descripteurs de catégorie et/ou en apprendre davantage sur les résultats à l'aide de défilements et de graphiques. Pour plus d'informations, voir Chapitre 11, «Exploration de l'analyse des liens du texte» , à la page 149. L'utilisation du noeud Text Mining pour extraire les résultats TLA est une très bonne méthode pour explorer et affiner vos modèles en prévision d'une utilisation ultérieure directe dans le noeud TLA.

La sortie peut être représentée sous la forme de 6 propriétés ou parties maximum. Pour plus d'informations, voir «Sortie du noeud TLA», à la page 51.

Vous pouvez trouver ce noeud dans l'onglet IBM SPSS Modeler Text Analytics de la palette de noeuds en bas de la fenêtre IBM SPSS Modeler. Pour plus d'informations, voir «Noeuds IBM SPSS Modeler Text Analytics», à la page 8.

Conditions requises. Le noeud Analyse des liens du texte accepte les données texte lues dans un champ par l'intermédiaire de n'importe quel noeud source standard (noeud Base de données, noeud Fichier plat, etc.) ou lues dans un champ listant des chemins vers des documents externes générés par un noeud Liste de fichiers ou un noeud Liste de nouvelles.

Puissance. Le noeud analyse des liens du texte ne se limite pas à une simple extraction de concepts permettant de fournir des informations sur les relations *entre* les concepts, ainsi que des opinions ou des qualificatifs associés susceptibles d'apparaître dans les données.

Noeud Analyse des liens du texte : onglet Champs

Utilisez l'onglet Champs pour indiquer les paramètres de champ des données dont vous allez extraire les concepts. Vous pouvez définir les paramètres suivants :

Champ ID. Sélectionnez le champ contenant l'identificateur des enregistrements textuels. L'identificateur doit être un entier. Le champ d'ID sert d'index aux enregistrements textuels. Utilisez un champ ID si le champ de texte correspond au texte à explorer.

Champ Texte. Sélectionnez le champ contenant le texte à explorer. Ce champ dépend de la source de données.

Champ Langue. Sélectionnez le champ contenant l'identificateur de langue ISO sur deux lettres. Si vous ne sélectionnez pas de champ, la langue du modèle fourni est utilisée pour chaque document.

Type de document. Le type de document indique la structure du texte. Sélectionnez l'un des types suivants :

- **Texte réel.** S'utilise pour la plupart des documents ou des sources textuelles. L'ensemble de texte entier est analysé pour l'extraction. Contrairement aux autres options, il n'existe pas de paramètres supplémentaires pour cette option.
- **Texte structuré.** Utilisez ce format pour les bibliographies, les brevets et les fichiers contenant des structures standard qui peuvent être identifiées et analysées. Ce type de document est utilisé pour éviter tout ou une partie du processus d'extraction. Il permet de définir des séparateurs de termes, d'affecter des types et d'imposer une valeur de fréquence minimale. Si vous sélectionnez cette option, vous devez cliquer sur le bouton **Paramètres** et entrer les séparateurs de texte dans la zone **Formatage de texte structuré** de la boîte de dialogue Paramètres du document. Pour plus d'informations, voir «Paramètres de document de l'onglet Champs», à la page 22.

Unité de texte. Sélectionnez le mode d'extraction parmi les choix suivants :

- **Mode de document.** Utilisez ce mode pour les documents courts et homogènes d'un point de vue sémantique (par exemple, dans le cas d'articles issus d'agences de presse).
- **Mode de paragraphe.** Utilisez ce format pour les pages Web et les documents sans balise. Le processus d'extraction divise les documents en unités sémantiques, sur la base de certaines caractéristiques, comme des balises internes et des éléments syntaxiques. Si ce mode est sélectionné, le scoring est appliqué paragraphe par paragraphe. Par conséquent, la règle `pomme & orange` est vraie uniquement si `pomme` et `orange` se trouvent dans le même paragraphe, par exemple.

  Remarque : En raison du mode d'extraction du texte depuis les documents PDF, le **mode Paragraphe** ne fonctionne pas sur ces documents. La raison en est que l'extraction supprime le marqueur de retour chariot.

Paramètres de mode de paragraphe. Cette option n'est disponible que si vous affectez à l'option Unité de texte la valeur **Mode de paragraphe**. Indiquez les nombres maximal et minimal de caractères à utiliser dans les extractions. La taille employée est arrondie à la période la plus proche. Pour vous assurer que les associations de mots obtenues à partir du texte du groupe de documents sont représentatives, n'indiquez pas de taille d'extraction trop petite.

- **Minimum.** Indiquez le nombre minimum de caractères à utiliser dans les extractions.
- **Maximum.** Indiquez le nombre maximal de caractères à utiliser dans les extractions.

Copier des ressources depuis. Lors de l'exploration de texte, l'extraction est basée sur les paramètres de l'onglet Expert mais également sur les ressources linguistiques. Ces ressources servent de base au traitement et à l'exécution du texte pendant l'extraction afin d'obtenir des concepts, des types et des motifs TLA. Vous pouvez copier les ressources dans ce noeud à partir d'un modèle de ressources.

Un modèle de ressources est un ensemble prédéfini de bibliothèques et de ressources linguistiques et non linguistiques avancées qui ont été affinées pour un domaine ou une utilisation spécifique. Ces ressources servent de base à la gestion et le traitement des données lors de l'extraction. Cliquez sur **Charger** et sélectionnez le modèle dans lequel copier vos ressources.

Les modèles sont chargés lorsque vous les sélectionnez et non lorsque le flux est exécuté. Au moment du chargement, une copie des ressources est stockée dans le noeud. Par conséquent, si vous souhaitez utiliser un modèle mis à jour, vous devez le recharger ici. Pour plus d'informations, voir «Copie des ressources à partir de modèles et de TAP», à la page 26.

Langue du texte. Identifie la langue du texte exploré. Les ressources copiées dans le noeud contrôlent les options de langue présentées. Sélectionnez la langue pour laquelle les ressources ont été optimisées.

Noeud Analyse des liens du texte : onglet Expert

Dans ce noeud, l'extraction des résultats des motifs TLA (analyse des liens du texte) est automatiquement activée. L'onglet Expert contient des paramètres supplémentaires ayant une incidence sur le mode d'extraction et de traitement du texte. Les paramètres de cette boîte de dialogue déterminent le fonctionnement de base du processus d'extraction, ainsi que quelques procédures avancées. Il existe également un certain nombre de ressources linguistiques et d'options ayant une incidence sur les résultats de l'extraction, qui sont contrôlées par le modèle de ressources sélectionné.

Limiter l'extraction aux concepts ayant une fréquence globale supérieure à [n]. Spécifie à partir de combien d'occurrences un mot ou un groupe de mots présent dans un texte doit être extrait. Ainsi, une valeur de 5 limite l'extraction aux mots ou groupes de mots figurant au moins cinq fois dans l'ensemble des enregistrements ou des documents.

Dans certains cas, modifier cette limite peut faire une grande différence dans les résultats d'extraction et par conséquent, dans les catégories. Imaginons que vous travaillez avec des données concernant un restaurant et que vous n'avez pas augmenté la limite au-dessus de 1 pour cette option. Dans ce cas, vos résultats d'extraction pourront contenir *pizza (1)*, *pizza fine (2)*, *pizza épinards (2)*, et *pizza préférée (2)*. Mais si l'extraction était limitée à une fréquence globale de 5 ou plus et que vous recommenciez l'extraction, trois de ces concepts ne seraient pas renvoyés. Vous obtiendriez *pizza (7)*, car *pizza* est la forme la plus simple et que ce mot existait déjà comme candidat possible. Et en fonction du reste du texte, vous pourriez obtenir une fréquence supérieure à 7, si le texte contient d'autres phrases avec le mot pizza. De plus, si *pizza épinards* était déjà un descripteur de catégorie, vous pourriez le remplacer par*pizza* pour pouvoir capturer tous les enregistrements. C'est pour cette raison que lorsque des catégories ont déjà été créées, la modification de cette limite doit être effectuée avec prudence.

Veuillez noter qu'il s'agit d'une fonction d'extraction uniquement ; si votre modèle contient des termes (ce qui est généralement le cas) et qu'un terme pour le modèle est trouvé dans le texte, alors le terme sera indexé quelle que soit sa fréquence.

Par exemple, supposons que vous utilisez un modèle Ressources de base qui inclut "los angeles" sous le type `<Location>` dans la bibliothèque principale ; si votre document contient une seule occurrence du terme "los angeles", celui-ci fera partie de la liste des concepts. Pour éviter cela, vous devrez définir un filtre pour afficher les concepts se produisant au moins le même nombre de fois que la valeur saisie dans le champ **Limiter l'extraction aux concepts ayant une fréquence globale supérieure à [n]**.

Adapter les erreurs de ponctuation. Cette option normalise temporairement le texte contenant des erreurs de ponctuation (par exemple, usage incorrect) au cours de l'extraction pour améliorer l'extraction des concepts. Cette option s'avère extrêmement utile lorsque le texte est court et de qualité médiocre (réponses ouvertes, messages électroniques, données CRM, etc.) ou qu'il contient de nombreuses abréviations.

Traitement des fautes de frappe. Nombre de caractères minimum requis : [n] Cette option applique une technique de regroupement flou qui permet de regrouper les mots mal orthographiés ou d'orthographe similaire sous un seul concept. L'algorithme de regroupement flou retire temporairement toutes les voyelles (à l'exception de la première) et les consonnes doubles/triples des mots extraits pour comparer ces derniers et voir s'ils sont identiques, afin par exemple que `modélisation` et `modéllisation` soient regroupés. Néanmoins, si chaque type est affecté à un type différent et que vous excluez le type `<Unknown>`, le regroupement flou n'est pas appliqué.

Vous pouvez également définir le nombre minimum de caractères *racines* requis avant d'utiliser le regroupement flou. Le nombre de caractères racine d'un terme est calculé en ajoutant l'ensemble des caractères et en soustrayant les caractères formant des suffixes inflexionnels et, dans le cas des termes apparaissant sous la forme de mots composés, les déterminants et les prépositions. Par exemple, le terme `exercices` sera considéré comme contenant 8 caractères racines dans la forme “exercice,” la lettre *s* de fin étant une inflexion (marque du pluriel). De même, `sauce soja` contient 9 caractères racines (“sauce soja”) et `usine de voitures` en contient 12 (“usine voiture”). Cette méthode de calcul est uniquement utilisée pour vérifier si le regroupement flou doit être appliqué, mais n'influe pas sur la mise en correspondance des mots.

Remarque : Si vous constatez plus tard que certains mots sont regroupés par erreur, vous pouvez exclure des paires de mots en les déclarant explicitement dans la section **Regroupement flou : Exceptions** de l'onglet Ressources avancées. Pour plus d'informations, voir «Regroupement flou», à la page 203.

Enlever les expressions unitermes Cette option extrait les mots uniques (unitermes) dans la mesure où le mot ne fait pas déjà partie d'un mot composé et si c'est un nom ou une catégorie grammaticale non reconnue.

Extraire les entités non linguistiques Cette option extrait les entités non linguistiques, telles que les numéros de téléphone, numéros de sécurité sociale, heures, dates, devises, chiffres, pourcentages, adresses électroniques, adresses HTTP, etc. Vous pouvez inclure ou exclure certains types d'entités non linguistiques dans la section **Entités non linguistiques : Configuration** de l'onglet Ressources avancées. Désactivez les entités dont vous n'avez pas besoin pour éviter au moteur d'extraction un temps de traitement inutile. Pour plus d'informations, voir «Configuration», à la page 208.

Algorithme des majuscules Cette option extrait les termes simples et composés ne figurant pas dans les dictionnaires intégrés, si la première lettre est une majuscule. Cette option offre une bonne manière d'extraire la plupart des noms propres.

Regrouper si possible les noms de personnes partiels et complets Cette option regroupe les noms qui apparaissent différemment dans le texte. Cette fonction est utile car les noms sont souvent cités dans leur forme complète au début du texte puis uniquement en version abrégée. Cette option essaye de faire correspondre tout uniterme ayant le type `<Unknown>` avec le dernier mot de tout terme composé entré comme `<Person>`. Par exemple, si *martin* est trouvé et initialement saisi comme `<Unknown>`, le moteur d'extraction vérifie si un terme composé de type `<Person>` contient *martin* comme dernier mot, tel que *pierre martin*. Cette option ne s'applique pas aux prénoms, car la plupart ne sont jamais extraits comme unitermes.

Taille maximale pour la permutation des mots utiles Cette option indique le nombre maximal de mots utiles pouvant être présents lorsque s'applique la technique de permutation. Cette technique de permutation regroupe les expressions similaires qui ne diffèrent les unes des autres que par la présence de mots utiles (par exemple, de et la), quelle que soit leur inflexion. Par exemple, si vous avez paramétré cette valeur sur au moins deux mots et si les deux expressions `représentants d'entreprise` et `représentants de l'entreprise` ont été extraites. Dans ce cas, les deux termes extraits sont regroupés dans la liste de concepts finale car les deux termes sont considérés comme étant les mêmes lorsque `de l'` est ignoré.

Utiliser la dérivation des termes (grouper par composants multitermes) Pour le traitement de Big Data, sélectionnez cette option pour regrouper les multitermes à l'aide de règles de dérivation.

Sortie du noeud TLA

Après l'exécution du noeud analyse des liens du texte, les données sont restructurées. Il est important de comprendre comment l'exploration de texte restructure les données. Pour obtenir une structure de Data mining différente, vous pouvez avoir recours aux noeuds de la palette Opérations sur les champs. Par exemple, si vous travaillez avec des données dans lesquelles chaque ligne représente un enregistrement textuel, une ligne est créée pour chaque motif découvert dans les données textuelles source. Pour chaque ligne de la sortie, il existe 15 champs :

- Six champs (**Concept#**, tels que **Concept1**, **Concept2**, etc. jusqu'à **Concept6**) représentent les concepts découverts dans la correspondance de motifs.
- Six champs (**Type#**, tels que **Type1**, **Type2**, etc. jusqu'à **Type6**) représentent le type de chaque concept.
- **Nom de la règle** représente le nom de la règle des liens du texte utilisée pour renvoyer le texte et générer la sortie.
- Un champ qui utilise le nom du champ ID spécifié dans le noeud et qui représente l'ID de l'enregistrement ou du document tel qu'il apparaissait dans les données d'entrée.
- **Texte mis en correspondance** représente la partie des données textuelles de l'enregistrement ou du document d'origine qui a été mise en correspondance avec le patron TLA.

Remarque : Les flux existants qui contiennent un noeud Analyse des liens du texte provenant d'une version antérieure à la version 5.0 risquent de ne pas être complètement exécutables tant que les noeuds n'ont pas été mis à jour. Certaines améliorations apportées aux versions ultérieures de IBM SPSS Modeler requièrent le remplacement d'anciens noeuds par leur nouvelle version, à la fois plus déployable et plus puissante.

Il est également possible d'effectuer une traduction automatique de certaines langues. Cette fonction permet d'explorer les documents figurant dans une langue que vous ne parlez pas. Si vous souhaitez utiliser la fonction de traduction, vous devez avoir accès à SDL Software as a Service (SaaS). Pour plus d'informations, voir Paramètres de traduction.

Mise en cache des résultats TLA

Si vous procédez à une mise en cache, les résultats de l'analyse des liens du texte se situent dans le flux. Pour éviter de répéter l'extraction des résultats de l'analyse des liens du texte à chaque exécution du flux, sélectionnez le noeud analyse des liens du texte, puis les options de menu suivantes **Edition > noeud >> Cache > Activer**. Lors de l'exécution suivante du flux, la sortie est mise en cache dans le noeud. L'icône du noeud affiche une petite image représentant un « document » qui passe de la couleur blanche à la couleur verte lorsque le cache est rempli. Le cache est conservé pendant toute la durée de la session. Pour conserver le cache plus longtemps (après fermeture et réouverture du flux), sélectionnez le noeud, puis les options de menu suivantes : **Modifier > noeud > Cache > Enregistrer le cache**. Lors de l'ouverture suivante du flux, vous pouvez recharger le cache enregistré plutôt que d'exécuter à nouveau la traduction.

Vous pouvez également enregistrer ou activer un cache de noeud en cliquant avec le bouton droit de la souris sur le noeud et en choisissant **Cache** dans le menu contextuel.

Utilisation du noeud Analyse des liens du texte dans un flux

Le noeud Analyse des liens du texte sert à accéder aux données et à extraire des concepts dans un flux. Vous pouvez utiliser n'importe quel noeud source pour accéder aux données.

Exemple : le noeud Fichier statistiques avec le noeud Analyse des liens du texte

L'exemple suivant indique le mode d'utilisation du noeud Analyse des liens du texte.

Figure 15. Exemple : le noeud Fichier statistiques avec le noeud Analyse des liens du texte

1. **Noeud Fichier statistiques (onglet Données).** Nous avons tout d'abord ajouté ce noeud au flux pour indiquer l'emplacement de stockage du texte.

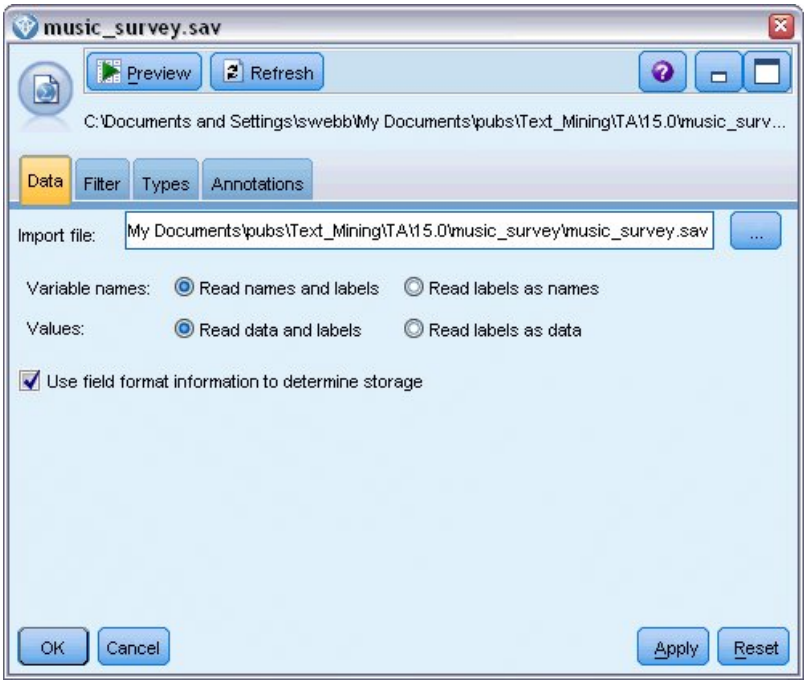


Figure 16. Boîte de dialogue Noeud Fichier statistiques (onglet Données).

2. **Noeud Analyse des liens du texte (onglet Champs).** Nous avons ensuite relié ce noeud au flux afin d'extraire les concepts en vue de l'affichage ou de la modélisation en aval. Nous avons indiqué le champ d'ID et le nom du champ de texte contenant les données, ainsi que d'autres paramètres.

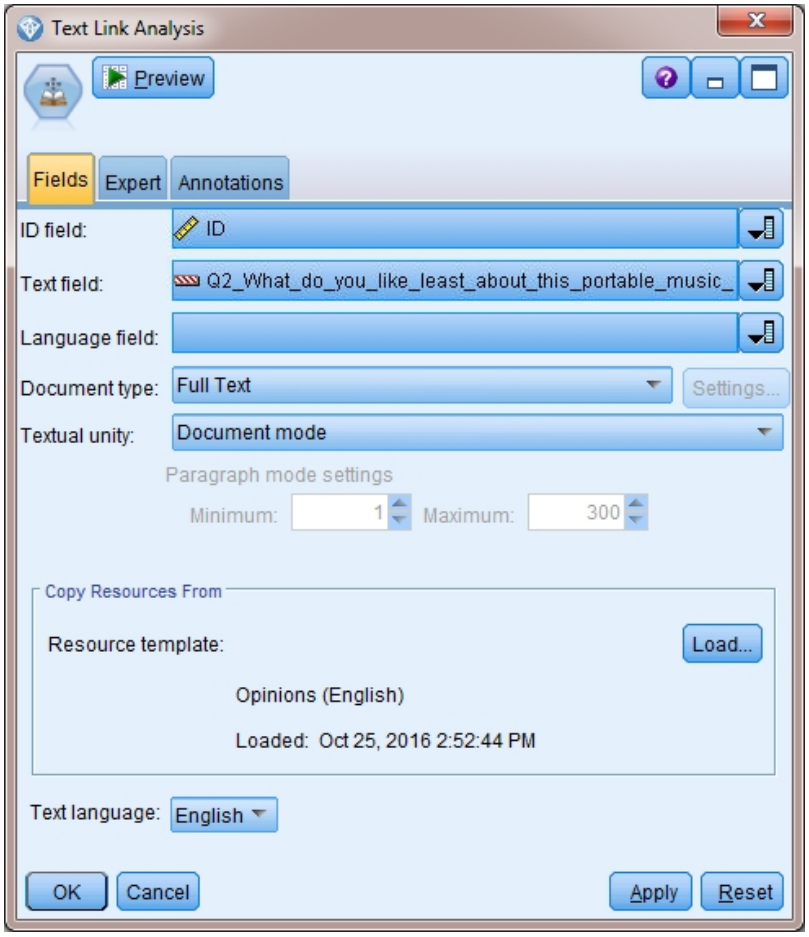
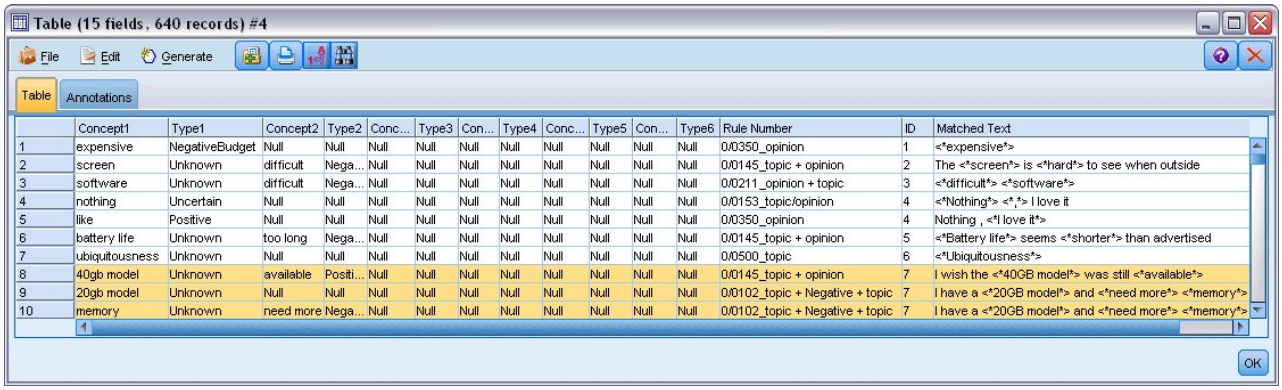


Figure 17. Boîte de dialogue Noeud Analyse des liens du texte (onglet Champs)

3. **Noeud Table.** Enfin, nous avons ajouté un noeud Table pour afficher les concepts extraits des documents texte. Dans la Table de résultats présentée, vous pouvez visualiser les résultats de patrons TLA trouvés dans les données après l'exécution de ce flux avec un noeud analyse des liens du texte. Certains résultats indiquent que seul un concept/type a été mis en correspondance. D'autres résultats sont plus complexes et contiennent plusieurs types et concepts. En outre, après l'exécution des données via le noeud analyse des liens du texte et l'extraction des concepts, plusieurs aspects des données ont changé. Les données d'origine de notre exemple contenaient 8 champs et 405 enregistrements. Après exécution du noeud analyse des liens du texte, elles comportent 15 champs et 640 enregistrements. Il existe désormais une ligne pour chaque résultat de patron TLA trouvé. Ainsi, la ligne `ID 7` s'est transformée en trois lignes car trois résultats de motifs TLA ont été extraits. Pour fusionner les données de sortie dans vos données d'origine, vous pouvez utiliser un noeud Fusionner.



	Concept1	Type1	Concept2	Type2	Conc...	Type3	Con...	Type4	Conc...	Type5	Con...	Type6	Rule Number	ID	Matched Text
1	expensive	NegativeBudget	Null	Null	Null	Null	Null	Null	Null	Null	Null	Null	0/0350_opinion	1	<*expensive*>
2	screen	Unknown	difficult	Nega...	Null	Null	Null	Null	Null	Null	Null	Null	0/0145_topic + opinion	2	The <*screen*> is <*hard*> to see when outside
3	software	Unknown	difficult	Nega...	Null	Null	Null	Null	Null	Null	Null	Null	0/0211_opinion + topic	3	<*difficult*> <*software*>
4	nothing	Uncertain	Null	Null	Null	Null	Null	Null	Null	Null	Null	Null	0/0153_topic/opinion	4	<*Nothing*> <*,*> I love it
5	like	Positive	Null	Null	Null	Null	Null	Null	Null	Null	Null	Null	0/0350_opinion	4	Nothing , <*I love it*>
6	battery life	Unknown	too long	Nega...	Null	Null	Null	Null	Null	Null	Null	Null	0/0145_topic + opinion	5	<*Battery life*> seems <*shorter*> than advertised
7	ubiquitousness	Unknown	Null	Null	Null	Null	Null	Null	Null	Null	Null	Null	0/0500_topic	6	<*Ubiquitousness*>
8	40gb model	Unknown	available	Positi...	Null	Null	Null	Null	Null	Null	Null	Null	0/0145_topic + opinion	7	I wish the <*40GB model*> was still <*available*>
9	20gb model	Unknown	Null	Null	Null	Null	Null	Null	Null	Null	Null	Null	0/0102_topic + Negative + topic	7	I have a <*20GB model*> and <*need more*> <*memory*>
10	memory	Unknown	need more	Nega...	Null	Null	Null	Null	Null	Null	Null	Null	0/0102_topic + Negative + topic	7	I have a <*20GB model*> and <*need more*> <*memory*>

Figure 18. Noeud de sortie Table

Chapitre 5. Navigation dans le texte source externe

Noeud Afficheur de fichiers

Lorsque vous explorez une collection de documents, vous pouvez spécifier les noms de chemin d'accès complet aux fichiers directement dans vos noeuds de modélisation Text Mining. Cependant, lors de la sortie d'un noeud Table, vous ne pouvez afficher que le nom de chemin complet d'un document au lieu du texte qu'il contient. Le noeud afficheur de fichiers peut être utilisé comme un analogue du noeud Table et vous permet d'accéder au texte réel de chaque document sans avoir à les fusionner en un fichier unique.

Le noeud afficheur de fichiers peut vous permettre de mieux comprendre les résultats issus de l'extraction de texte ; en effet, il vous fournit un accès au texte source, ou au texte non traduit, dont ont été extraits les concepts, et qui serait autrement inaccessible dans le flux. Ce noeud est ajouté au flux après un noeud liste fichiers afin d'obtenir une liste des liens vers l'ensemble des fichiers.

Ce noeud donne comme résultat une fenêtre affichant tous les documents qui ont été lus et utilisés pour extraire les concepts. L'une des icônes de la barre d'outils de cette fenêtre vous permet de lancer le rapport dans un navigateur externe répertoriant le nom des documents sous forme de liens hypertexte. Cliquez sur un lien pour ouvrir le document correspondant. Pour plus d'informations, voir «Utilisation du noeud afficheur de fichiers», à la page 56.

Vous pouvez trouver ce noeud dans l'onglet IBM SPSS Modeler Text Analytics de la palette de noeuds en bas de la fenêtre IBM SPSS Modeler. Pour plus d'informations, voir «Noeuds IBM SPSS Modeler Text Analytics», à la page 8.

Remarque : Lorsque vous travaillez en mode client-serveur et que les noeuds Afficheur de fichiers font partie du flux, les collections de documents doivent être stockées dans un référentiel de serveur Web sur le serveur. Le noeud de sortie Text Mining génère la liste des documents stockés dans le répertoire du serveur Web ; les paramètres de sécurité du serveur Web gèrent donc les droits d'accès à ces documents.

Paramètres du noeud afficheur de fichiers

Vous pouvez spécifier les paramètres suivants pour le noeud afficheur de fichiers.

Champ de document. Sélectionnez le champ de vos données qui contient le nom et le chemin complets des documents à afficher.

Titre de la page HTML générée. Créez un titre qui doit apparaître en haut de la page contenant la liste des documents.

Utilisation du noeud afficheur de fichiers

L'exemple suivant indique le mode d'utilisation du noeud visualiseur de fichiers.

Exemple : noeud Liste fichiers et noeud visualiseur de fichiers

Figure 19. Flux illustrant l'utilisation d'un noeud afficheur de fichiers

1. **Noeud Liste fichiers (onglet Paramètres).** Nous avons tout d'abord ajouté le noeud afin d'indiquer l'emplacement des documents.

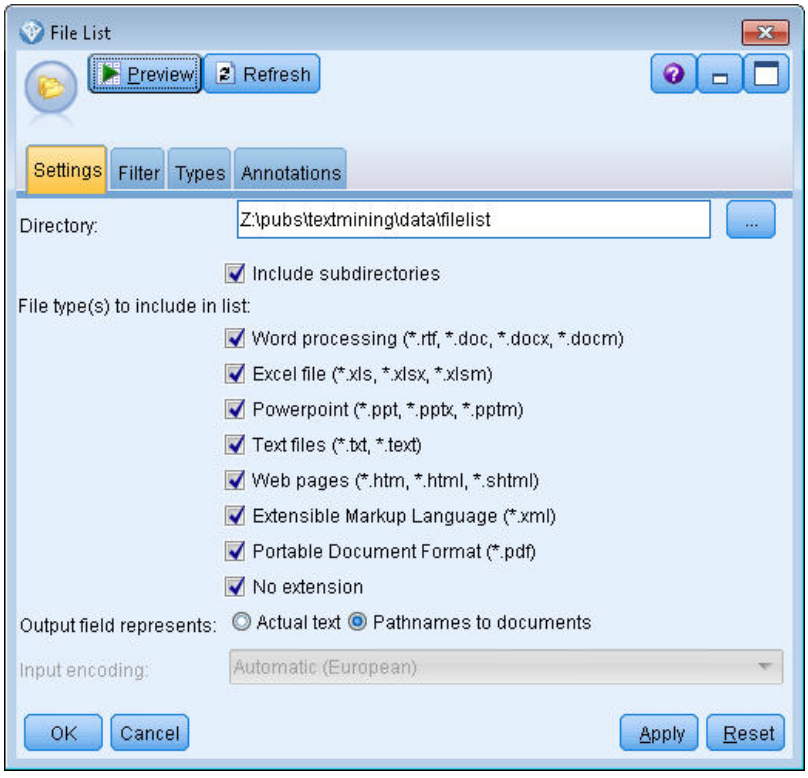


Figure 20. Boîte de dialogue Noeud Liste fichiers : onglet Paramètres

2. **Noeud visualiseur de fichiers (onglet Paramètres).** Nous avons ensuite relié le noeud visualiseur de fichiers pour produire une liste HTML des documents.

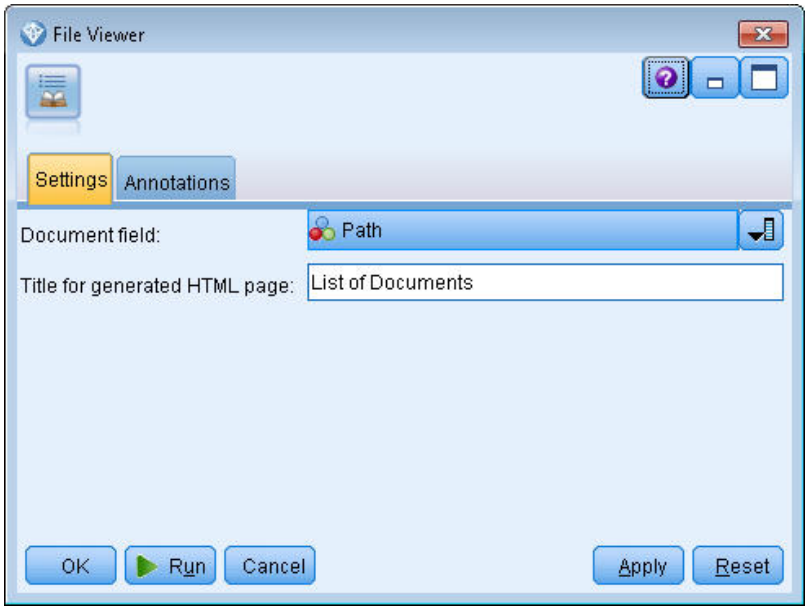


Figure 21. Boîte de dialogue Noeud visualiseur de fichiers : onglet Paramètres

3. **Boîte de dialogue Sortie Afficheur de fichiers.** Nous avons exécuté le flux qui sort la liste de documents dans une nouvelle fenêtre.

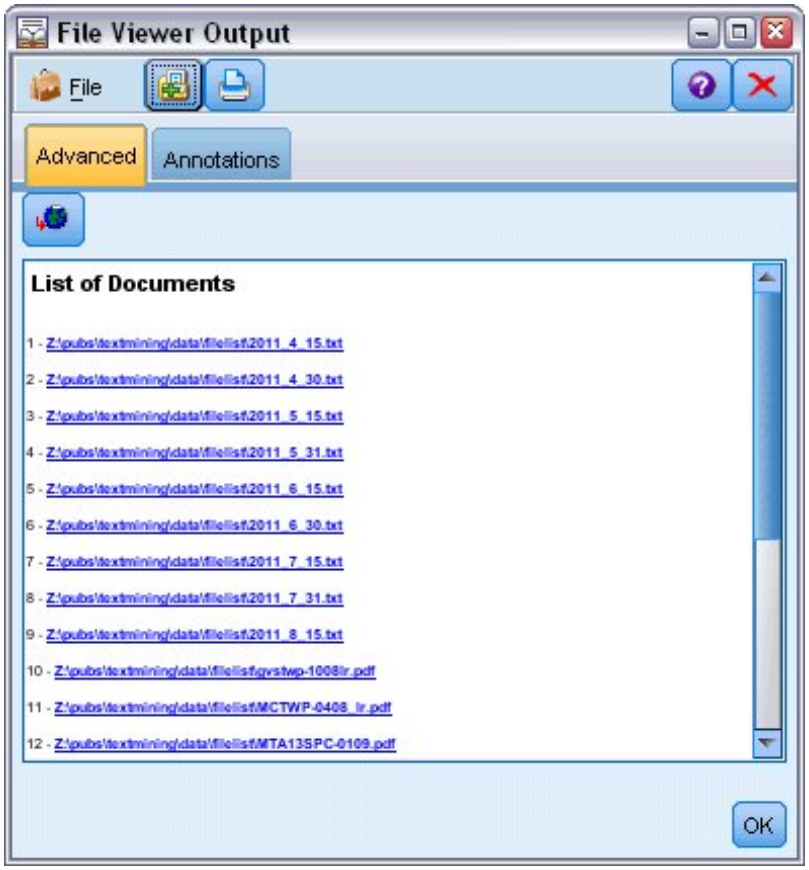


Figure 22. Sortie Afficheur de fichiers

4. Pour voir les documents, nous avons cliqué sur le bouton de la barre d'outils représentant un globe avec une flèche rouge. Une liste de liens hypertexte de document est alors apparue dans notre navigateur.

Chapitre 6. Propriétés des noeuds pour la génération de scripts

IBM SPSS Modeler dispose d'un langage de génération de scripts qui vous permet d'exécuter des flux à partir de la ligne de commande. Ici, vous pouvez en savoir plus sur les propriétés des noeuds spécifiques à chacun des noeuds fournis avec IBM SPSS Modeler Text Analytics. Pour plus d'informations sur l'ensemble standard des noeuds fournis avec IBM SPSS Modeler, reportez-vous au Guide de génération des scripts et d'automatisation.

Noeud liste fichiers : filelistnode

Vous pouvez utiliser les propriétés du tableau ci-dessous pour la génération de scripts. Le noeud lui-même est appelé `filelistnode`.

Tableau 7. propriétés de génération de scripts du noeud liste fichiers

Propriétés de génération de scripts	**Type de données**
`path`	*chaîne*
`recurse`	*indicateur*
`word_processing`	*indicateur*
`excel_file`	*indicateur*
`powerpoint_file`	*indicateur*
`text_file`	*indicateur*
`web_page`	*indicateur*
`xml_file`	*indicateur*
`pdf_file`	*indicateur*
`no_extension`	*indicateur*

Remarque : le paramètre de création de liste n'est plus disponible ; les scripts contenant cette option seront automatiquement convertis en sortie de type fichier.

Noeud Flux de nouvelles : webfeednode

Vous pouvez utiliser les propriétés du tableau ci-dessous pour la génération de scripts. Le noeud lui-même est appelé `webfeednode`.

Tableau 8. Propriétés de génération de scripts du noeud Flux de nouvelles

Propriétés de génération de scripts	**Type de données**	**Description de la propriété**
`urls`	*chaîne1 chaîne2 ...chaînen*	Chaque URL est spécifiée dans la structure de la liste. La liste des URL utilise le séparateur "\n"
`recent_entries`	*indicateur*	
`limit_entries`	*entier*	Nombre des entrées les plus récentes à lire (par URL).
`use_previous`	*indicateur*	Pour enregistrer et réutiliser le cache des flux de nouvelles.
`use_previous_label`	*chaîne*	Nom du cache des fils de nouvelles enregistré.

Tableau 8. Propriétés de génération de scripts du noeud Flux de nouvelles (suite)

Propriétés de génération de scripts	Type de données	Description de la propriété
`start_record`	*chaîne*	Balise de début non RSS.
`url` *n* `.title`	*chaîne*	Pour chaque URL de la liste, vous devez également définir une propriété ici. La première propriété sera `url1.title`, où le chiffre correspond à sa position dans la liste des URL. Ceci est la balise de début contenant le titre du contenu.
`url` *n* `.short_description`	*chaîne*	Identique à `url` *n* `.title`.
`url` *n* `.description`	*chaîne*	Identique à `url` *n* `.title`.
`url` *n* `.authors`	*chaîne*	Identique à `url` *n* `.title`.
`url` *n* `.contributors`	*chaîne*	Identique à `url` *n* `.title`.
`url` *n* `.published_date`	*chaîne*	Identique à `url` *n* `.title`.
`url` *n* `.modified_date`	*chaîne*	Identique à `url` *n* `.title`.
`html_alg`	`Aucun` `HTMLCleaner`	Méthode de filtrage du contenu.
`discard_lines`	*indicateur*	Ignorer les lignes courtes. Utilisé avec `min_words`
`min_words`	*entier*	Nombre minimal de mots.
`discard_words`	*indicateur*	Ignorer les lignes courtes. Utilisé avec `min_avg_len`.
`min_avg_len`	*entier*	
`discard_scw`	*indicateur*	Ignorer les lignes contenant de nombreux mots à caractère unique. Utilisé avec `max_scw`
`max_scw`	*entier*	Proportion maximum en pourcentage de mots à caractère unique dans une ligne
`discard_tags`	*indicateur*	Ignorer les lignes contenant des balises spécifiques.
`tags`	*chaîne*	Les caractères spéciaux doivent être échappés avec une barre oblique inverse (\).
`discard_spec_words`	*indicateur*	Ignorer les lignes contenant des chaînes spécifiques
`words`	*chaîne*	Les caractères spéciaux doivent être échappés avec une barre oblique inverse (\).

Noeud Langue : languageidentifier

Vous pouvez utiliser les propriétés du tableau ci-dessous pour la génération de scripts. Le noeud lui-même est appelé `languageidentifier`.

Tableau 9. Propriétés de génération de scripts du noeud Langue

Propriétés de génération de scripts	Type de données	Description de la propriété
`text`	*zone*	
`language_field_name`	*chaîne*	Nom de champ généré comme sortie.

`unidentified_language_value`	`Non défini`
	`Pris en charge`
	`Personnalisé`
		Valeur par défaut à utiliser si la langue ne peut pas être identifiée.

Tableau 9. Propriétés de génération de scripts du noeud Langue (suite)

Propriétés de génération de scripts	Type de données	Description de la propriété
unidentified_language_supported	en de es fr it ja nl pt	Code ISO. Disponible uniquement si unidentified_language_value est Pris en charge.
unidentified_language_custom	*chaîne*	Disponible uniquement si unidentified_language_value est Personnalisé.

Noeud Text Mining : TextMiningWorkbench

Vous pouvez utiliser les paramètres suivants pour définir ou mettre à jour un noeud via la génération de scripts. Le noeud lui-même est appelé TextMiningWorkbench.

Important : Il est impossible d'indiquer un autre modèle de ressources via la génération de scripts. Si vous pensez avoir besoin d'un modèle, vous devez le sélectionner dans la boîte de dialogue Noeud.

Tableau 10. propriétés de génération de scripts de noeuds modélisation Text Mining

Propriétés de génération de scripts	Type de données	Description de la propriété
text	*zone*	
method	ReadText ReadPath	
docType	*entier*	Avec des valeurs possibles (0,1,2), où 0 = Texte intégral, 1 = Texte structuré et 2 = XML
codage	Automatique "UTF-8" "UTF-16" "ISO-8859-1" "US-ASCII" "CP850" "EUC-JP" "SHIFT-JIS" "ISO2022-JP"	Les valeurs contenant des caractères spéciaux (du type "UTF-8") doivent être mises entre guillemets afin de ne pas être confondues avec un opérateur mathématique.
unity	*entier*	Avec des valeurs possibles (0,1), où 0 = Paragraphe, 1 = Document
para_min	*entier*	
para_max	*entier*	
mtag	*chaîne*	Contient tous les paramètres mtag (de la boîte de dialogue Paramètres pour les fichiers XML).
mclef	*chaîne*	Contient tous les paramètres mclef (de la boîte de dialogue Paramètres pour les fichiers texte structuré).
partition	*zone*	
custom_field	*indicateur*	Indique si un champ de partitionnement sera spécifié ou non.
use_model_name	*indicateur*	

Tableau 10. propriétés de génération de scripts de noeuds modélisation Text Mining (suite)

Propriétés de génération de scripts	Type de données	Description de la propriété
model_name	*chaîne*	
use_partitioned_data	*indicateur*	Si un champ de partition est défini, seules les données d'apprentissage sont utilisées pour la création du modèle.
model_output_type	Interactif Model	Résultats Interactive dans un modèle de catégorie. Résultats Model dans un modèle de concept.
use_interactive_info	*indicateur*	Pour la création interactive dans une session uniquement.
reuse_extraction_results	*indicateur*	Pour la création interactive dans une session uniquement.
interactive_view	Catégories TLA Clusters	Pour la création interactive dans une session uniquement.
extract_top	*entier*	Ce paramètre est utilisé lorsque model_type = Concept
use_check_top	*indicateur*	
check_top	*entier*	
use_uncheck_top	*indicateur*	
uncheck_top	*entier*	
langue	de en es fr it ja nl pt	
frequency_limit	*entier*	Abandonné dans la version 14.0.
concept_count_limit	*entier*	Limiter l'extraction aux concepts ayant une fréquence globale supérieure à cette valeur.
fix_punctuation	*indicateur*	
fix_spelling	*indicateur*	
spelling_limit	*entier*	
extract_uniterm	*indicateur*	
extract_nonlinguistic	*indicateur*	
upper_case	*indicateur*	
group_names	*indicateur*	
permutation	*entier*	Nombre maximum de mots utiles soumis à une permutation pour le regroupement (la valeur par défaut est 3).

Nugget de modèle Text Mining : TMWBModelApplier

Vous pouvez utiliser les propriétés du tableau ci-dessous pour la génération de scripts. Le nugget lui-même est appelé `TMWBModelApplier`.

Tableau 11. Propriétés de nugget de modèle Text Mining

Propriétés de génération de scripts	Type de données	Description de la propriété
`scoring_mode`	`Champs` `Enregistrements`	
`field_values`	`Indicateurs` `Nombres`	Cette option n'est pas disponible dans les nuggets de modèles de catégories. Pour `Indicateurs`, valeur `TRUE` ou `FALSE`
`true_value`	*chaîne*	Avec les `Indicateurs`, définissez la valeur true.
`false_value`	*chaîne*	Avec les `Indicateurs`, définissez la valeur false.
`extension_concept`	*chaîne*	Spécifiez l'extension du nom de champ. Les noms de champ générés reprennent le nom de concept et l'extension. Spécifiez où placez cette extension à l'aide de la valeur `add_as`.
`extension_category`	*chaîne*	Extension nom de champ. Vous pouvez choisir de spécifier une extension préfixe/suffixe pour le nom de champ ou vous pouvez choisir d'utiliser les codes de catégories. Les noms de champ générés reprennent le nom de catégorie et l'extension. Spécifiez où placez cette extension à l'aide de la valeur `add_as`.
`add_as`	`Suffixe` `Préfixe`	
`fix_punctuation`	*indicateur*	
`excluded_subcategories_descriptors`	`RollUpToParent` `Ignorer`	Pour les modèles de catégories uniquement. Si une sous-catégorie n'est pas sélectionnée. Cette option permet de déterminer comment seront manipulés les descripteurs appartenant aux sous-catégories qui n'étaient pas sélectionnées pour le scoring. Il y a deux options. • `Ignorer`. Avec l'option Exclure complètement ses descripteurs de l'évaluation, les descripteurs des sous-catégories qui ne sont pas cochées (qui ne sont pas sélectionnées) seront ignorés et ne seront pas utilisés pendant le scoring. • `RollUpToParent`. Avec l'option Agréger les descripteurs avec ceux des catégories parents, les descripteurs des sous-catégories qui ne sont pas cochées (qui ne sont pas sélectionnées) seront utilisés comme descripteurs pour la catégorie parent (la catégorie au-dessus de cette sous-catégorie). S'il existe plusieurs niveaux de sous-catégories et que celles-ci ne sont pas sélectionnées, les descripteurs seront reportés sous la première catégorie parent disponible.
`check_model`	*indicateur*	Abandonné dans la version 14
`text`	*zone*	
`method`	`ReadText` `ReadPath`	

Tableau 11. Propriétés de nugget de modèle Text Mining (suite)

Propriétés de génération de scripts	Type de données	Description de la propriété
docType	*entier*	Avec des valeurs possibles (0,1,2), où 0 = Texte intégral, 1 = Texte structuré et 2 = XML
codage	Automatique "UTF-8" "UTF-16" "ISO-8859-1" "US-ASCII" "CP850" "EUC-JP" "SHIFT-JIS" "ISO2022-JP"	Les valeurs contenant des caractères spéciaux (du type "UTF-8") doivent être mises entre guillemets afin de ne pas être confondues avec un opérateur mathématique.
langue	de en es fr it ja nl pt	

Noeud Analyse des liens du texte : textlinkanalysis

Vous pouvez utiliser les paramètres du tableau suivant pour définir ou mettre à jour un noeud via la génération de scripts. Le noeud lui-même est appelé textlinkanalysis.

Important : Il est impossible de spécifier un modèle de ressources via la génération de scripts. Pour sélectionner un modèle, utilisez la boîte de dialogue du noeud.

Tableau 12. Propriétés du noeud analyse des liens du texte (TLA)

Propriétés de génération de scripts	Type de données	Description de la propriété
id_field	*zone*	
text	*zone*	
method	ReadText ReadPath	
docType	*entier*	Avec des valeurs possibles (0,1,2), où 0 = Texte intégral, 1 = Texte structuré et 2 = XML
codage	Automatique "UTF-8" "UTF-16" "ISO-8859-1" "US-ASCII" "CP850" "EUC-JP" "SHIFT-JIS" "ISO2022-JP"	Les valeurs contenant des caractères spéciaux (du type "UTF-8") doivent être mises entre guillemets afin de ne pas être confondues avec un opérateur mathématique.
unity	*entier*	Avec des valeurs possibles (0,1), où 0 = Paragraphe, 1 = Document
para_min	*entier*	
para_max	*entier*	
mtag	*chaîne*	Contient tous les paramètres mtag (de la boîte de dialogue Paramètres pour les fichiers XML).

Tableau 12. Propriétés du noeud analyse des liens du texte (TLA) (suite)

Propriétés de génération de scripts	Type de données	Description de la propriété
mclef	*chaîne*	Contient tous les paramètres mclef (de la boîte de dialogue Paramètres pour les fichiers texte structuré).
langue	de en es fr it ja nl pt	
concept_count_limit	*entier*	Limiter l'extraction aux concepts ayant une fréquence globale supérieure à cette valeur.
fix_punctuation	*indicateur*	
fix_spelling	*indicateur*	
spelling_limit	*entier*	
extract_uniterm	*indicateur*	
extract_nonlinguistic	*indicateur*	
upper_case	*indicateur*	
group_names	*indicateur*	
permutation	*entier*	Nombre maximum de mots utiles soumis à une permutation pour le regroupement (la valeur par défaut est 3).

Chapitre 7. Mode Plan de travail interactif

A partir d'un noeud modélisation de Text Mining, vous pouvez choisir de lancer une session de plan de travail interactif au cours de l'exécution du flux. Dans cette session, vous pouvez extraire des concepts clés de vos données textuelles, créer des catégories et explorer des clusters et des motifs d'analyse des liens du texte, et vous pouvez également générer des modèles de catégorie. Ce chapitre fournit une présentation générale de l'interface du plan de travail et des principaux éléments avec lesquels vous serez amené à travailler, parmi lesquels :

- **Résultats d'extraction.** Une fois l'extraction effectuée, les résultats extraits correspondent aux principaux mots et groupes de mots identifiés et extraits des données textuelles, également appelés *concepts*. Ces concepts sont regroupés en *types*. En utilisant ces concepts et types, vous pouvez explorer vos données et créer des catégories. Vous pouvez les gérer dans la vue **Catégories et concepts**.
- **Catégories.** Vous pouvez utiliser des descripteurs (résultats d'extraction, motifs et règles, par exemple) en tant que définition pour créer manuellement ou automatiquement un ensemble de catégories auxquelles des documents et des enregistrements sont affectés selon qu'ils contiennent ou non une partie de la définition de catégorie. Vous pouvez les gérer dans la vue **Catégories et concepts**.
- **Clusters.** Les *Clusters* représentent un regroupement de concepts. Des liens indiquant l'existence d'une relation entre ces concepts ont été établis. Ces concepts sont regroupés à l'aide d'un algorithme complexe qui s'appuie notamment sur la fréquence à laquelle deux concepts apparaissent ensemble par rapport à la fréquence à laquelle ils apparaissent séparément. Vous pouvez les gérer dans la vue **Clusters**. Vous pouvez également ajouter les concepts qui constituent un cluster à des catégories.
- **Motifs Analyse des liens du texte.** Si vous avez des règles de motifs d'analyse des liens du texte dans les ressources linguistiques ou si vous utilisez un modèle de ressources qui possède déjà certaines règles de TLA, vous pouvez alors extraire des motifs des données textuelles. Ces motifs peuvent vous permettre de découvrir des relations intéressantes entre les concepts figurant dans vos données. Vous pouvez également utiliser des motifs comme descripteurs dans vos catégories. Vous pouvez les gérer dans la vue **Analyse des liens du texte**.
- **Ressources linguistiques.** Le processus d'extraction s'appuie sur un ensemble de paramètres et de définitions linguistiques pour gérer la façon dont le texte est extrait et géré. Vous pouvez gérer ces paramètres et définitions sous la forme de modèles et de bibliothèques dans la vue **Editeur de ressources**.

Problèmes potentiels du plan de travail interactif

- Les sessions multiples de plan de travail interactif peuvent ralentir le système. SPSS Modeler Text Analytics et SPSS Modeler partagent un même moteur d'exécution Java lorsqu'une session de plan de travail interactif est lancée. En fonction du nombre de sessions de plan de travail interactif que vous appelez lors d'une session SPSS Modeler, la mémoire système peut ralentir l'application, même si vous ouvrez et fermez une même session. Cet effet peut être particulièrement prononcé si vous traitez un volume important de données ou que votre machine dispose d'une quantité de mémoire vive inférieure à celle de 4 Go recommandée. Si vous constatez que votre machine est lente à répondre, sauvegardez tout votre travail, arrêtez SPSS Modeler, puis relancez l'application. L'exécution de SPSS Modeler Text Analytics sur une machine dont la quantité de mémoire est inférieure à celle recommandée, en particulier si vous traitez un volume important de données ou que les traitements sont longs, peut entraîner un arrêt de Java pour mémoire insuffisante. Il est fortement conseillé d'ajouter de la mémoire pour atteindre ou dépasser la quantité de mémoire recommandée (ou d'utiliser le serveur SPSS Modeler Text Analytics) si vous traitez un volume important de données.
- Le client SPSS Modeler peut manquer de mémoire si plusieurs sessions de plan de travail interactif SPSS Modeler Text Analytics sont exécutées sans redémarrer l'application. Surveillez la quantité de mémoire utilisée sur la ligne d'état et, si celle-ci vient à manquer, fermez, puis rouvrez le client SPSS Modeler.

Vue Catégories et concepts

L'interface de l'application comporte plusieurs vues. La vue Catégories et concepts correspond à la fenêtre dans laquelle vous pouvez créer et explorer des catégories, mais également explorer et modifier les résultats d'extraction. Le terme *Catégories* désignent un groupe d'idées et de motifs étroitement liés auxquels des documents et des enregistrements sont affectés lors d'un processus de scoring. Alors que les *concepts* se rapportent au niveau le plus basique des résultats d'extraction disponibles à utiliser en tant que blocs de construction pour vos catégories, appelés les descripteurs.

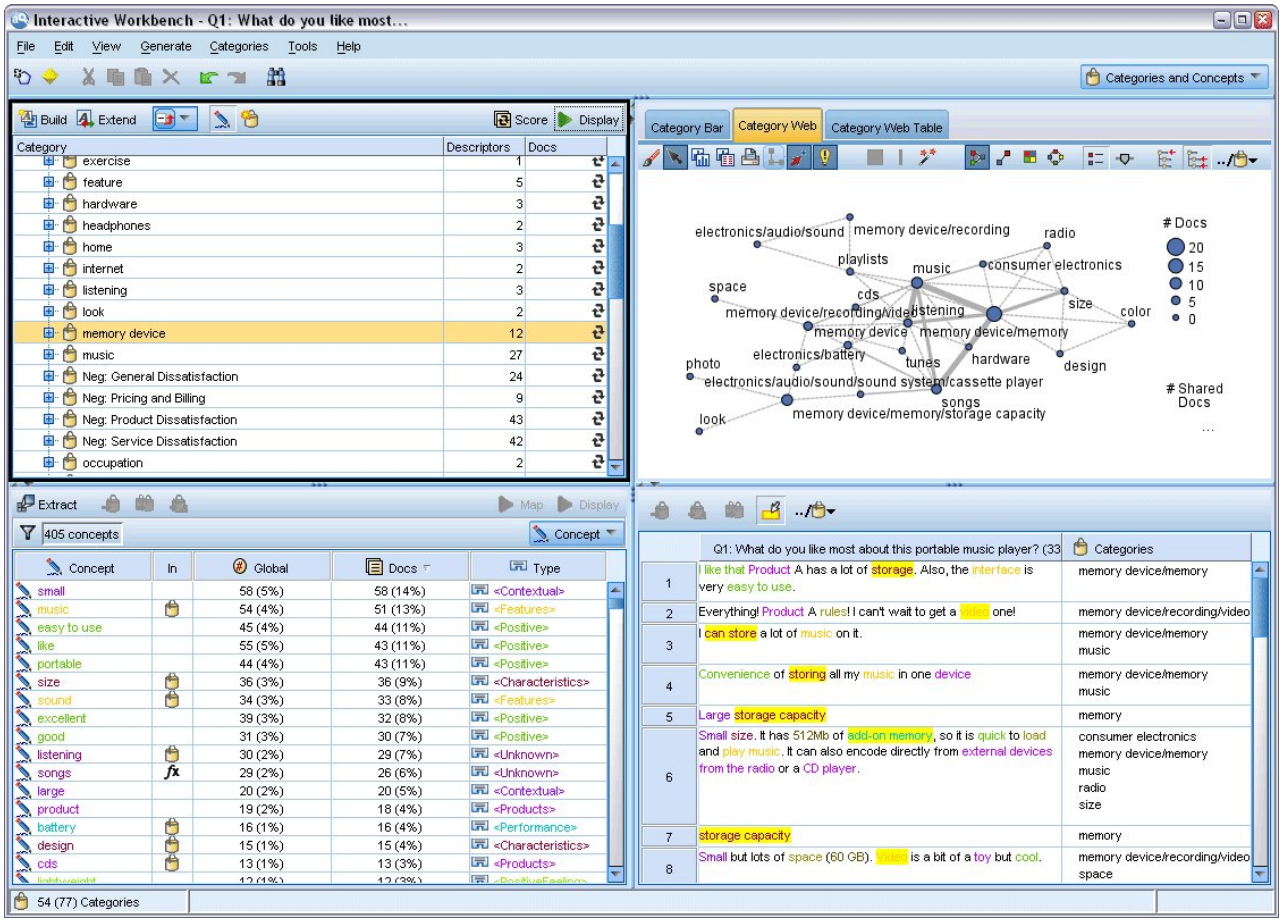


Figure 23. Vue Catégories et concepts

La vue Catégories et concepts comporte quatre sous-fenêtres ; vous pouvez masquer ou afficher chacun d'entre eux en sélectionnant son nom dans le menu Vue. Pour plus d'informations, voir Chapitre 9, «Catégorisation des données textuelles», à la page 95.

Sous-fenêtre Catégories

Cette zone, située dans l'angle supérieur gauche, représente un tableau dans lequel vous pouvez gérer les catégories que vous créez. Une fois les concepts et types extraits de vos données textuelles, vous pouvez commencer à générer des catégories en appliquant des techniques (réseaux sémantiques et inclusion de concepts, par exemple) ou en procédant de façon manuelle. Lorsque vous double-cliquez sur le nom d'une catégorie, la boîte de dialogue Définitions de catégorie s'ouvre. Tous les descripteurs qui composent la définition de cette catégorie (concepts, types, règles) y figurent. Pour plus d'informations, voir Chapitre 9, «Catégorisation des données textuelles», à la page 95. L'ensemble des techniques automatiques ne sont pas disponibles pour toutes les langues.

Lorsque vous sélectionnez une ligne dans cette sous-fenêtre, vous pouvez afficher les informations concernant les descripteurs ou les documents/enregistrements correspondants dans les sous-fenêtres Données et Visualisation.

Sous-fenêtre Résultats d'extraction

Cette zone, situé dans l'angle inférieur gauche, présente les résultats de l'extraction. Lorsque vous exécutez une extraction, le moteur du programme d'extraction parcourt les données textuelles, identifie les concepts pertinents et affecte un type à chaque concept. Les *concepts* correspondent à des mots ou groupes de mots extraits à partir des données textuelles. Les *types* correspondent à des regroupements sémantiques de concepts. Une fois l'extraction terminée, les concepts et les types apparaissent avec un codage couleur dans la sous-fenêtre Résultats d'extraction. Pour plus d'informations, voir «Résultats d'extraction : concepts et types», à la page 81.

Vous pouvez voir l'ensemble des termes sous-jacents pour un concept en passant la souris sur le nom du concept. En procédant ainsi, une info-bulle apparaît indiquant le nom du concept et plusieurs lignes de termes qui sont groupés sous ce concept. Ces termes sous-jacents incluent les synonymes définis dans les ressources linguistiques (que ceux-ci aient été rencontrés ou non dans le texte) ainsi que tous les termes extraits au pluriel/singulier, les termes permutés, les termes provenant du regroupement flou, etc. Vous pouvez également copier ces termes ou voir l'ensemble complet des termes sous-jacents en cliquant avec le bouton droit sur le nom du concept et en choisissant l'option du menu contextuel.

Text Mining est un processus itératif au cours duquel les résultats de l'extraction sont passés en revue en fonction du contexte des données textuelles. Ils sont ensuite affinés afin de générer de nouveaux résultats avant d'être réévalués. Vous pouvez affiner les résultats de l'extraction en modifiant les ressources linguistiques. Vous pouvez procéder en partie à ce réglage à partir de la sous-fenêtre Résultats d'extraction ou de la sous-fenêtre de données directement, ou bien directement dans la vue Editeur de ressources. Pour plus d'informations, voir «Vue Editeur de ressources», à la page 74.

Remarque : Si le nombre de résultats est plus important que ce que peut contenir la sous-fenêtre visible, vous pouvez utiliser les commandes de la partie inférieure de la sous-fenêtre pour vous déplacer vers l'avant ou vers l'arrière dans les résultats ou entrer un numéro de page pour y accéder.

Sous-fenêtre Visualisation

Cette zone, située dans l'angle supérieur droit, présente, selon plusieurs perspectives, les éléments communs apparaissant dans la catégorisation des documents/enregistrements. Chaque graphique ou diagramme fournit des informations similaires, mais les présente d'une façon différente ou avec un niveau de détail différent. Vous pouvez vous appuyer sur ces graphiques et diagrammes pour analyser les résultats de la catégorisation, et affiner les catégories ou générer des rapports. Par exemple, un graphique peut révéler des catégories trop similaires (lorsqu'elles ont en commun plus de 75 % de leurs enregistrements, par exemple) ou trop différentes. Le contenu d'un graphique ou d'un diagramme dépend des éléments sélectionnés dans les autres sous-fenêtres. Pour plus d'informations, voir «Graphiques et diagrammes de catégorie», à la page 155.

Sous-fenêtre Données

La sous-fenêtre Données est situé dans l'angle inférieur droit. Cette sous-fenêtre présente un tableau contenant les documents ou les enregistrements correspondant à une sélection dans une autre zone de la vue. En fonction des éléments sélectionnés, seul le texte correspondant apparaît dans la sous-fenêtre Données. Une fois votre sélection effectuée, cliquez sur un bouton **Afficher** pour remplir la sous-fenêtre de données à l'aide du texte correspondant.

Si une autre sous-fenêtre contient une sélection, les documents ou enregistrements correspondants représentent les concepts mis en surbrillance en couleur pour vous permettre de les repérer plus facilement dans le texte. Vous pouvez également pointer sur un élément faisant l'objet d'un codage

couleur pour afficher une info-bulle contenant le nom du concept sous lequel il a été extrait et le type auquel il a été affecté. Pour plus d'informations, voir «La sous-fenêtre Données», à la page 104.

Recherche dans la vue Catégories et concepts

Il peut s'avérer nécessaire de localiser rapidement des informations dans une section particulière. Avec la barre d'outils Rechercher, vous pouvez entrer la chaîne à rechercher et définir un autre critère de recherche comme la sensibilité à la casse ou la direction de la recherche. Puis vous pouvez choisir la sous-fenêtre dans laquelle effectuer la recherche.

Pour utiliser la fonction de recherche

1. Dans la vue Catégories et concepts, choisissez **Edition > Rechercher** dans les menus. La barre d'outils Rechercher apparaît au-dessus de la sous-fenêtre Catégories et des sous-fenêtres Visualisation.
2. Saisissez la chaîne de mots que vous recherchez dans la zone de texte. Vous pouvez contrôler la casse, les correspondances partielles et le sens de la recherche à l'aide des boutons de la barre d'outils.
3. Dans la barre d'outils, cliquez sur le nom de la sous-fenêtre dans laquelle effectuer la recherche. Si une correspondance est détectée, le texte est mis en surbrillance dans la fenêtre.
4. Pour rechercher la correspondance suivante, cliquez de nouveau sur le nom de la sous-fenêtre.

Vue Clusters

Dans la vue Clusters, vous pouvez créer et explorer les résultats de cluster trouvés dans vos données textuelles. Les *clusters* constituent des regroupements de concepts générés par des algorithmes de classification qui se fondent sur la fréquence d'apparition des concepts et sur la fréquence à laquelle ils apparaissent ensemble. Les clusters ont pour objectif de regrouper les concepts apparaissant ensemble, alors que les catégories ont pour objectif de regrouper les documents ou les enregistrements en fonction des correspondances existant entre le texte et les descripteurs (concept, règles, motifs) pour chaque catégorie.

Plus les concepts figurant dans un cluster apparaissent ensemble, moins ils apparaissent avec d'autres concepts et plus le cluster permet d'identifier des relations intéressantes entre les concepts. Deux concepts font l'objet d'une co-occurrence lorsqu'ils apparaissent tous les deux (ou que l'un de leurs synonymes ou termes apparaît) dans le même document ou enregistrement. Pour plus d'informations, voir Chapitre 10, «Analyse des clusters», à la page 143.

Vous pouvez créer des clusters et les explorer dans un ensemble de diagrammes et de graphiques susceptibles de révéler les relations existant entre des concepts, découverte qui prendrait autrement beaucoup trop de temps. Alors que vous ne pouvez pas ajouter de clusters entiers à des catégories, vous pouvez ajouter les concepts qui figurent dans un cluster à une catégorie à l'aide de la boîte de dialogue Définitions du cluster. Pour plus d'informations, voir «Définitions de cluster», à la page 147.

Vous pouvez modifier les paramètres de classification non supervisée de manière à orienter les résultats. Pour plus d'informations, voir «Création de clusters», à la page 144.

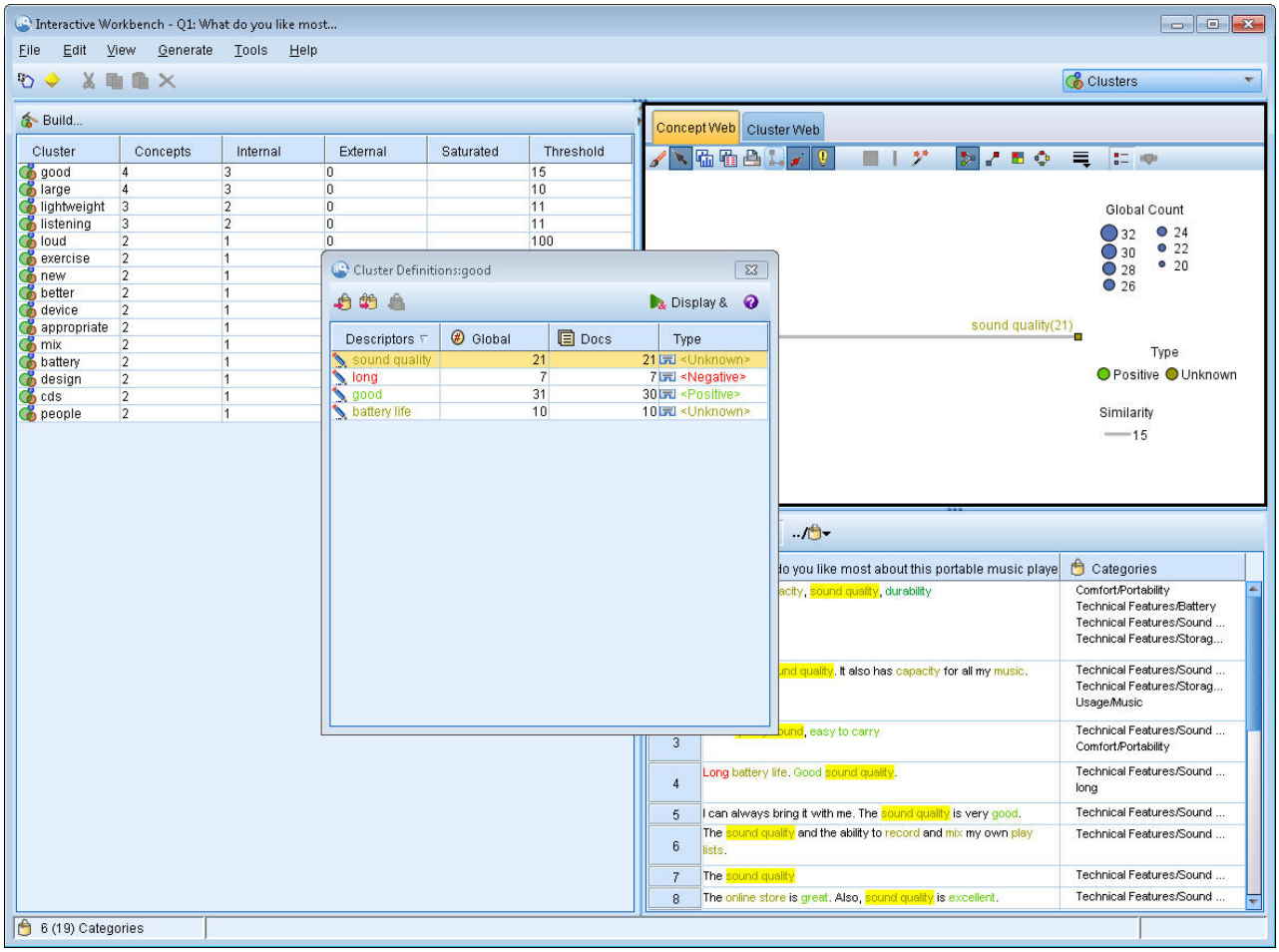


Figure 24. Vue Clusters

La vue Clusters est organisée en trois sous-fenêtres ; vous pouvez masquer ou afficher chacun d'entre eux en sélectionnant son nom dans le menu Vue. En règle générale, seules les sous-fenêtres Clusters et Visualisation sont affichées.

Sous-fenêtre Clusters

Cette sous-fenêtre, située à gauche, présente les clusters trouvés dans les données textuelles. Pour générer des résultats de classification, cliquez sur le bouton **Créer**. Les clusters sont générés par un algorithme de classification non supervisée qui tente d'identifier les concepts qui apparaissent fréquemment ensemble.

A chaque nouvelle extraction, les résultats de cluster sont effacés et vous devez générer une nouvelle fois les clusters pour obtenir les tous derniers résultats. Lorsque vous générez les clusters, vous pouvez modifier certains paramètres (nombre maximal de clusters à créer, nombre maximal de concepts pouvant exister ou nombre maximal de liens avec des concepts extérieurs, par exemple). Pour plus d'informations, voir «Exploration des Clusters», à la page 147.

Sous-fenêtre Visualisation

Située dans l'angle supérieur droit, cette sous-fenêtre offre deux perspectives de groupement : un graphique Relations par concept et un graphique Relations par cluster. Si cette sous-fenêtre est masquée, vous pouvez y accéder à partir du menu Vue (**Affichage > Visualisation**). En fonction des éléments sélectionnés dans la sous-fenêtre Clusters, vous pouvez afficher les interactions correspondantes entre les clusters ou à l'intérieur des clusters. Les résultats sont présentés sous plusieurs formes :

- **Relations par concept.** Graphique Relations représentant l'ensemble des concepts des clusters sélectionnés, ainsi que les concepts liés en dehors du cluster.
- **Relations par cluster.** Graphique Relations représentant les liens existant entre les clusters sélectionnés et d'autres clusters, ainsi que les liens entre ces autres clusters.

Remarque : Pour pouvoir afficher un graphique Relations par cluster, vous devez avoir créé des clusters présentant des liens externes. Les liens externes sont des liens entre des paires de concepts dans des clusters distincts (un concept au sein d'un cluster et un concept en dehors, dans un autre cluster). Pour plus d'informations, voir «Graphiques Cluster», à la page 157.

Sous-fenêtre Données

La sous-fenêtre Données, située dans l'angle inférieur droit, est masquée par défaut. Vous ne pouvez pas afficher la sous-fenêtre Données depuis la sous-fenêtre Clusters étant donné que ces clusters couvrent plusieurs documents/enregistrements, ce qui rend les résultats sans intérêt. Vous pouvez toutefois consulter les données correspondant à une sélection dans la boîte de dialogue Définitions du cluster. En fonction des éléments sélectionnés dans cette boîte de dialogue, seul le texte correspondant figure dans la sous-fenêtre Données. Une fois les éléments sélectionnés, cliquez sur le bouton **Afficher &** pour remplir la sous-fenêtre de données avec les documents ou les enregistrements contenant l'ensemble des concepts.

Les documents ou enregistrements correspondants indiquent les concepts en surbrillance colorée pour vous aider à les identifier facilement dans le texte. Vous pouvez également pointer sur un élément faisant l'objet d'un codage couleur pour afficher le concept sous lequel il a été extrait et le type auquel il a été affecté. La sous-fenêtre Données peut comporter plusieurs colonnes ; la colonne correspondant au champ de texte est toujours affichée. Elle porte le nom du champ de texte utilisé lors de l'extraction ou le nom d'un document lorsque les données textuelles figurent dans plusieurs fichiers différents. D'autres colonnes sont disponibles. Pour plus d'informations, voir «La sous-fenêtre Données», à la page 104.

Vue Analyse des liens du texte

Dans la vue Analyse des liens du texte, vous pouvez créer et explorer les motifs d'analyse des liens du texte trouvés dans vos données textuelles. L'analyse des liens du texte est une technologie de mise en correspondance de motifs qui vous permet de définir des règles de TLA et de les comparer aux concepts extraits et aux relations trouvées dans le texte.

Les motifs s'avèrent particulièrement utiles lorsque vous tentez de découvrir des relations entre des concepts ou des opinions sur un sujet donné. En voici quelques exemples : extraction d'opinions sur des produits à partir de données d'enquête, extraction de relations génomiques à partir de rapports de recherche médicale ou extraction de relations entre des personnes ou des lieux à partir de renseignements.

Une fois les motifs TLA extraits, vous pouvez les explorer dans la sous-fenêtre Données ou Visualisation et même les ajouter à des catégories dans la vue Catégories et concepts. Pour pouvoir extraire les résultats de l'analyse des liens du texte, des règles d'analyse des liens du texte doivent être définies dans le modèle de ressources ou dans les bibliothèques que vous utilisez. Pour plus d'informations, voir Chapitre 18, «A propos des règles des liens du texte», à la page 213.

Si vous extrayez des résultats de motifs d'analyse des liens du texte, les résultats s'affichent dans cette vue. Sinon, vous devrez utiliser le bouton **Extraire** et choisir l'option pour activer l'extraction des motifs .

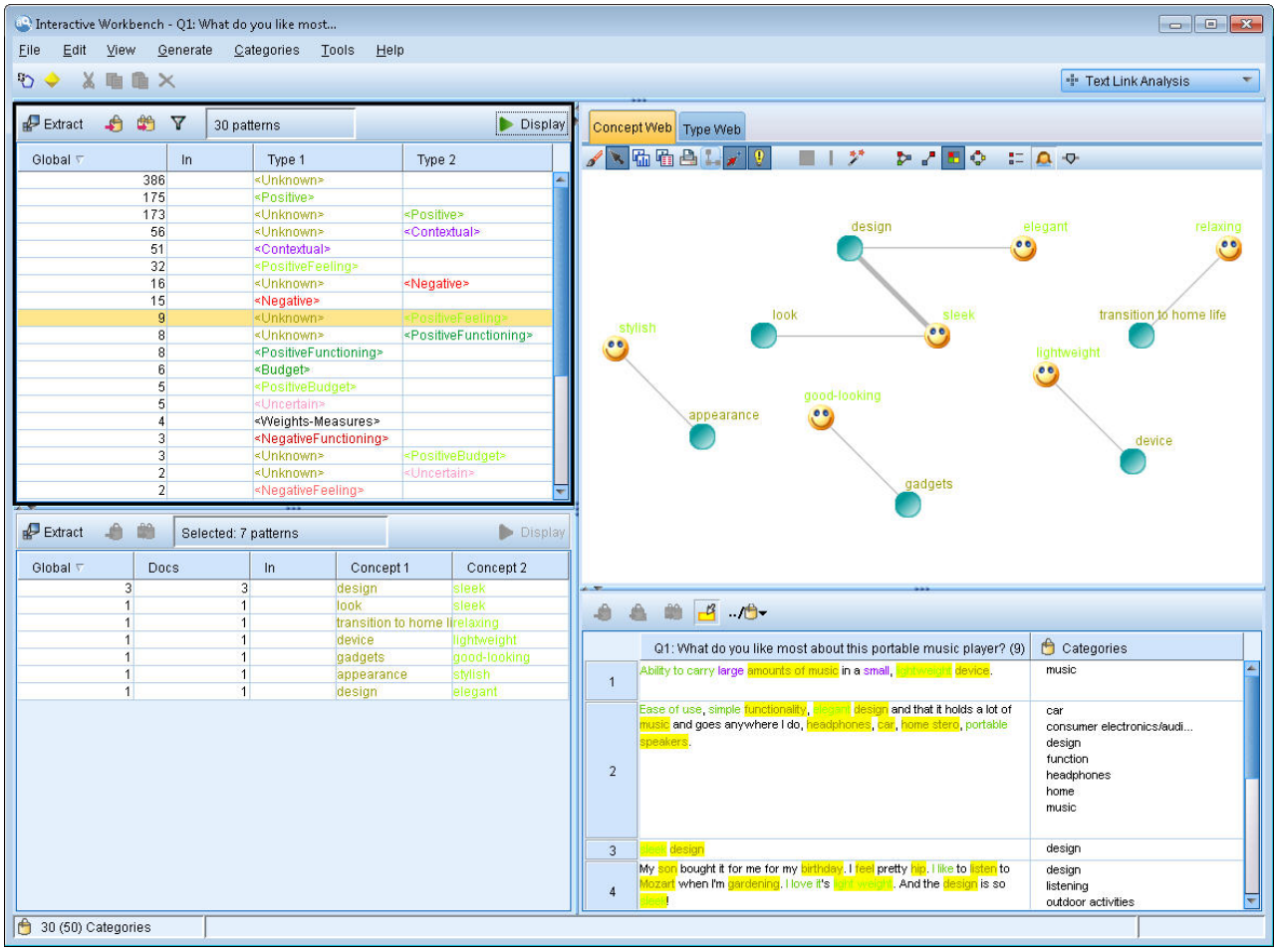


Figure 25. Vue Analyse des liens du texte

La vue Analyse des liens du texte est organisée en quatre sous-fenêtres ; vous pouvez masquer ou afficher chacun d'entre eux en sélectionnant son nom dans le menu Vue. Pour plus d'informations, voir Chapitre 11, «Exploration de l'analyse des liens du texte», à la page 149.

Sous-fenêtres Motifs de type et Motifs de concept

Dans ces deux sous-fenêtres interconnectées situées à gauche, vous pouvez explorer et sélectionner des résultats de motifs TLA. Les motifs comportent jusqu'à six types ou six concepts. La règle de motifs d'analyse des liens du texte définie dans les ressources linguistiques détermine la complexité des résultats de motifs. Pour plus d'informations, voir Chapitre 18, «A propos des règles des liens du texte», à la page 213.

Les résultats de motifs sont d'abord regroupés au niveau du type, puis ils sont divisés en motifs de concept. Pour cette raison, il existe deux sous-fenêtres de résultats différentes : Motifs de type (en haut à gauche) et Motifs de concept (en bas à gauche).

- **Motifs de type.** La sous-fenêtre Motifs de type présente les motifs extraits comportant au moins deux types associés correspondant à une règle de motifs TLA. Les motifs de type se présentent sous la forme `<Organization> + <Location> + <Positive>`, ce qui permet d'obtenir un commentaire positif concernant une organisation situé dans une location particulière.
- **Motifs de concept.** La sous-fenêtre Motifs de concept présente les motifs extraits au niveau du concept pour tous les motifs de type actuellement sélectionnés dans la sous-fenêtre Motifs de type située au-dessus. Les motifs de concept suivent une structure de type `hôtel + paris + merveilleux`.

Comme avec les résultats d'extraction dans la vue Catégories et concepts, vous pouvez vérifier les résultats ici. Si vous souhaitez affiner les types et concepts qui constituent ces motifs, procédez aux modifications dans la sous-fenêtre Résultats d'extraction de la vue Catégories et concepts ou directement dans l'éditeur de ressources, puis exécutez une nouvelle extraction des motifs.

Sous-fenêtre Visualisation

Cette sous-fenêtre, située dans l'angle supérieur droit de la vue Analyse des liens du texte, représente un graphique Relations des motifs sélectionnés sous forme de motifs de type ou de motifs de concept. Si cette sous-fenêtre est masquée, vous pouvez y accéder à partir du menu Vue (**Affichage** > **Visualisation**). En fonction des éléments sélectionnés dans d'autres sous-fenêtres, vous pouvez afficher les interactions correspondantes entre les documents/enregistrements et les motifs.

Les résultats sont présentés sous plusieurs formes :

- **Graphique de concept.** Ce graphique présente tous les concepts figurant dans les motifs sélectionnés. Dans un graphique de concept, l'épaisseur des lignes et la taille des noeuds (lorsque les icônes de type ne sont pas représentées) indiquent le nombre d'occurrences globales figurant dans la table sélectionnée.
- **Graphique de type.** Ce graphique présente tous les types figurant dans les motifs sélectionnés. Dans un graphique de type, l'épaisseur des lignes et la taille des noeuds (lorsque les icônes de type ne sont pas représentées) indiquent le nombre d'occurrences globales figurant dans la table sélectionnée. Les noeuds sont représentés soit par une couleur de type, soit par une icône.

Pour plus d'informations, voir «Graphiques Analyse des liens du texte», à la page 158.

Sous-fenêtre Données

La sous-fenêtre Données est situé dans l'angle inférieur droit. Cette sous-fenêtre présente un tableau contenant les documents ou les enregistrements correspondant à une sélection dans une autre zone de la vue. En fonction des éléments sélectionnés, seul le texte correspondant apparaît dans la sous-fenêtre Données. Une fois votre sélection effectuée, cliquez sur un bouton **Afficher** pour remplir la sous-fenêtre de données à l'aide du texte correspondant.

Si une autre sous-fenêtre contient une sélection, les documents ou enregistrements correspondants représentent les concepts mis en surbrillance en couleur pour vous permettre de les repérer plus facilement dans le texte. Vous pouvez également pointer sur un élément faisant l'objet d'un codage couleur pour afficher une info-bulle contenant le nom du concept sous lequel il a été extrait et le type auquel il a été affecté. Pour plus d'informations, voir «La sous-fenêtre Données», à la page 104.

Vue Editeur de ressources

IBM SPSS Modeler Text Analytics utilise un robuste moteur d'extraction pour capturer rapidement et avec précision des concepts-clés de données texte. Ce moteur s'appuie essentiellement sur les ressources linguistiques pour déterminer la quantité de données textuelles non structurées à analyser et à interpréter.

La vue Editeur de ressources permet de visualiser et d'affiner les ressources linguistiques utilisées pour extraire des concepts, les regrouper sous des types, reconnaître des motifs dans les données texte, etc. IBM SPSS Modeler Text Analytics offre plusieurs modèles de ressources préconfigurés. Pour certaines langues, vous pouvez également utiliser les ressources dans un pack d'analyse de texte. Pour plus d'informations, voir «Utilisation des packs d'analyse de texte», à la page 136.

Comme ces ressources risquent de ne pas toujours être parfaitement adaptées au contexte de vos données, vous pouvez créer, modifier et gérer vos propres ressources pour un contexte ou un domaine particulier dans l'Editeur de ressources. Pour plus d'informations, voir Chapitre 15, «Utilisation des bibliothèques», à la page 177.

Pour simplifier le processus d'affinage des ressources linguistiques, vous pouvez effectuer les tâches communes liées aux dictionnaires à partir de la vue Catégories et concepts via les menus contextuels des sous-fenêtres Résultats d'extraction et Données. Pour plus d'informations, voir la rubrique «Affinage des résultats d'extraction», à la page 89.

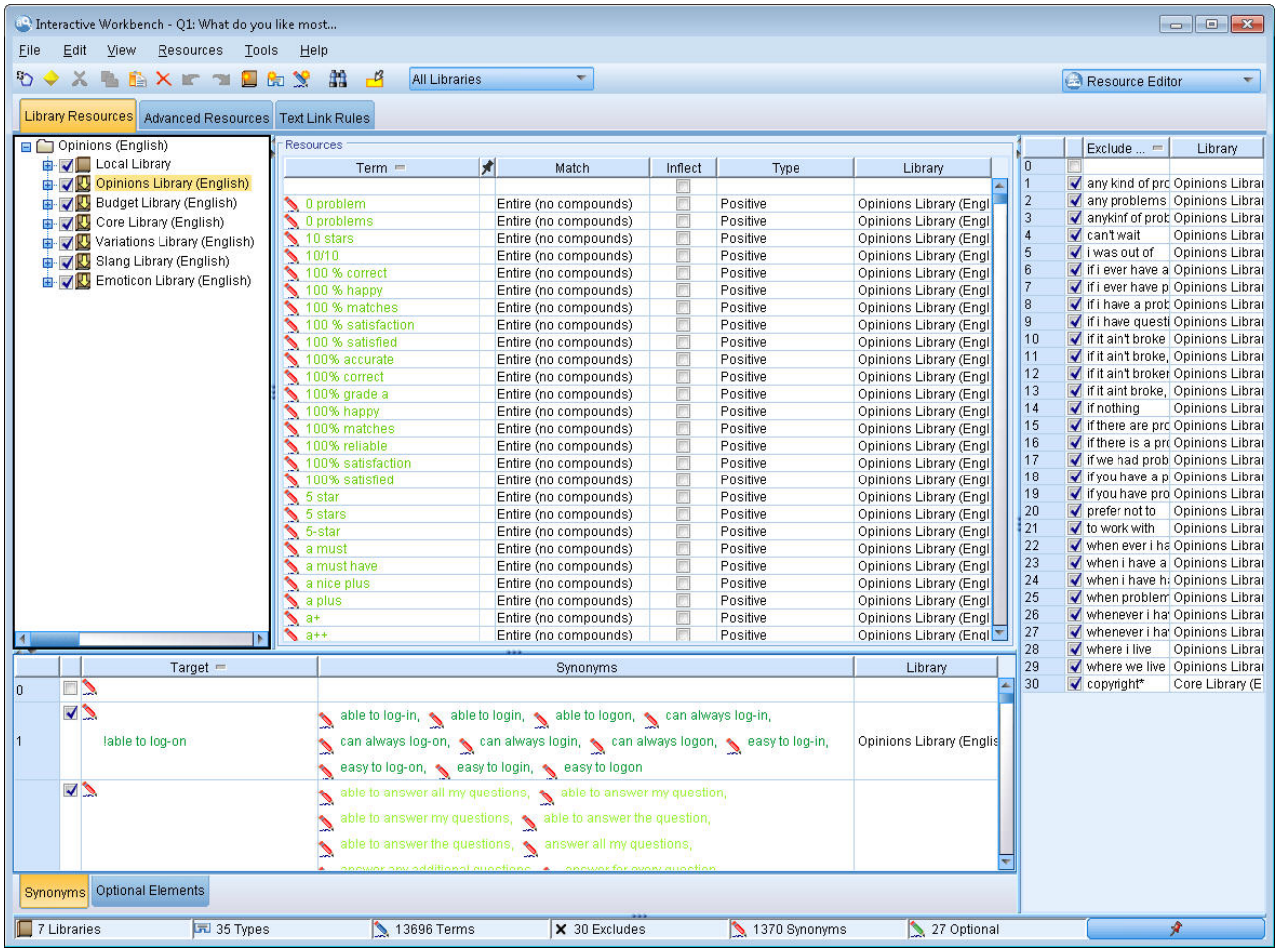


Figure 26. Vue Editeur de ressources

Les opérations que vous effectuez dans la vue Editeur de ressources concernent la gestion et l'adaptation des ressources linguistiques. Ces ressources sont stockées sous la forme de modèles et de bibliothèques. La vue Editeur de ressources est divisée en quatre sous-fenêtres : Arborescence de la bibliothèque, Dictionnaire de types, Dictionnaire des substitutions et Dictionnaire d'exclusions.

Remarque : Pour plus d'informations, voir «Interface de l'éditeur», à la page 168.

Définition des options

Vous pouvez définir les options générales de IBM SPSS Modeler Text Analytics dans la boîte de dialogue Options. Cette boîte de dialogue comporte les onglets suivants :

- **Session.** Cet onglet contient les options générales et les séparateurs.
- **Affichage.** Cet onglet contient les options des couleurs utilisées dans l'interface.
- **Sons.** Cet onglet contient les options de son.

Pour éditer les options

1. A partir des menus, sélectionnez **Outils> Options**. La boîte de dialogue Options s'ouvre.

2. Sélectionnez l'onglet contenant les informations à modifier.
3. Changez les options souhaitées.
4. Cliquez sur **OK** pour enregistrer les modifications.

Options : onglet Session

Cet onglet permet de définir certains paramètres de base.

Affichage du panneau Données et du graphique des catégories. Ces options affectent la manière dont les données sont présentées dans la sous-fenêtre de données et dans la sous-fenêtre Visualisation dans la vue Catégories et concepts.

- **Afficher la limite du panneau de données et des relations de catégorie.** Cette option permet de définir le nombre maximal de documents à afficher ou à utiliser pour remplir les sous-fenêtres de données ou les graphiques et les diagrammes de la vue Catégories et concepts.
- **Afficher les catégories des documents/enregistrements au moment de l'affichage.** Si cette option est sélectionnée, les documents ou les enregistrements font l'objet d'un scoring à chaque fois que vous cliquez sur Afficher de sorte que toutes les catégories auxquelles ils appartiennent peuvent être affichées dans la colonne Catégorie de la sous-fenêtre de données, ainsi que dans les graphiques de catégorie. Dans certains cas, notamment avec des ensembles de données importants, vous pouvez désactiver cette option pour que les données et les graphiques s'affichent plus rapidement.

Ajouter à la catégorie à partir du panneau de données. Ces options affectent les éléments ajoutés aux catégories lors de l'ajout de documents et d'enregistrements de la sous-fenêtre de données.

- **Dans la vue Catégories et concepts, copier.** Ajouter un document ou un enregistrement de la sous-fenêtre de données dans cette vue copiera soit **uniquement les concepts** soit à la fois **les concepts et les motifs**.
- **Dans la vue Analyse des liens du texte, copier.** Ajouter un document ou un enregistrement de la sous-fenêtre de données dans cette vue copiera soit **uniquement les motifs** soit à la fois **les concepts et les motifs**.

Délimiteur de l'éditeur de ressource. Sélectionnez le caractère à utiliser en tant que séparateur lors de la saisie d'éléments, tels que des concepts, des synonymes et des éléments optionnels dans la vue Editeur de ressources.

Options : onglet Afficher

Dans cet onglet, vous pouvez éditer les options qui affectent la présentation globale de l'application et les couleurs utilisées pour différencier les éléments.

Remarque : pour modifier l'impression générale du produit et lui donner une apparence classique ou ressemblant à une version antérieure, ouvrez la boîte de dialogue Options utilisateur dans le menu Outils de la fenêtre principale IBM SPSS Modeler.

Couleurs personnalisées. Modifiez la couleur des éléments apparaissant à l'écran. Vous pouvez modifier la couleur de chacun des éléments du tableau. Pour indiquer une couleur personnalisée, cliquez sur la zone Couleur à droite de l'élément à modifier et choisissez une couleur dans la liste déroulante des couleurs.

- **Texte non extrait.** Données textuelles, qui n'ont pas encore été extraites, visibles dans la sous-fenêtre Données.
- **Mettre en évidence l'arrière-plan.** Couleur d'arrière-plan de sélection de texte utilisée lors de la sélection d'éléments dans les sous-fenêtres ou de texte dans la sous-fenêtre Données.
- **Arrière-plan d'extraction requise.** Couleur d'arrière-plan des sous-fenêtres Résultats d'extraction, Motifs et Clusters indiquant que des modifications ont été apportées aux bibliothèques et qu'une extraction est nécessaire.

- **Arrière-plan de commentaires de catégorie.** Couleur d'arrière-plan de catégorie apparaissant à l'issue d'une opération.
- **Type par défaut.** Couleur par défaut des types et des concepts apparaissant dans la sous-fenêtre de données et la sous-fenêtre Résultats d'extraction. Cette couleur sera appliquée à tous les types personnalisés que vous créerez dans l'éditeur de ressources. Vous pouvez remplacer cette couleur par défaut pour les dictionnaires de types personnalisés en éditant leurs propriétés dans l'Editeur de ressources. Pour plus d'informations, voir «Création de types», à la page 189.
- **Table rayée 1.** Première des deux couleurs utilisées alternativement pour la table dans la boîte de dialogue Modifier les concepts forcés, afin de différencier chaque ensemble de lignes.
- **Table rayée 2.** Deuxième des deux couleurs utilisées alternativement pour la table dans la boîte de dialogue Modifier les concepts forcés, afin de différencier chaque ensemble de lignes.

Remarque : Si vous cliquez sur le bouton **Rétablir les valeurs par défaut**, les valeurs d'origine de toutes les options de cette boîte de dialogue sont restaurées.

Options : onglet Sons

Cet onglet permet d'éditer les options concernant les sons. Dans Evénements sonores, vous pouvez indiquer un son à utiliser pour vous avertir lorsqu'un événement se produit. Un grand nombre de sons sont disponibles. Utilisez le bouton des points de suspension (**...**) pour explorer les sons et en sélectionner un. Les fichiers *.wav* utilisés pour créer des sons pour IBM SPSS Modeler Text Analytics sont stockés dans le sous-répertoire */media* du répertoire d'installation. Si vous ne souhaitez pas que les sons soient joués, sélectionnez **Désactiver tous les sons**. Les sons sont désactivés par défaut.

Remarque : Si vous cliquez sur le bouton **Rétablir les valeurs par défaut**, les valeurs d'origine de toutes les options de cette boîte de dialogue sont restaurées.

Paramètres d'aide d'Microsoft Internet Explorer

Paramètres Microsoft Internet Explorer

La plupart des fonctions d'aide de cette application utilisent des techniques basées sur Microsoft Internet Explorer. Certaines versions d'Internet Explorer (y compris celle fournie avec Microsoft Windows XP, Service Pack 2) bloquent par défaut ce qu'elles considèrent comme du "contenu actif" dans les fenêtres Internet Explorer de votre ordinateur local. Ce paramètre par défaut risque de bloquer du contenu dans les fonctions d'aide. Pour visualiser tout le contenu de l'Aide, modifiez le comportement par défaut d'Internet Explorer.

1. Dans les menus Internet Explorer, choisissez :

   Outils > **Options Internet**
2. Cliquez sur l'onglet **Avancé**.
3. Accédez à la section **Sécurité**.
4. Sélectionnez **Autoriser l'exécution du contenu actif dans les fichiers de mon ordinateur**.

Génération de nuggets de modèle et de noeuds modélisation

Lors d'une session interactive, vous pouvez utiliser le travail que vous avez effectué pour générer les éléments suivants :

- **Un noeud modélisation Text Mining.** Un noeud modélisation généré à partir d'une session de plan de travail interactif est un noeud Text Mining dont les paramètres et les options reflètent ceux enregistrés dans la session interactive ouverte. Cela peut être utile lorsque vous ne disposez plus du noeud Text Mining d'origine ou que vous souhaitez créer une version. Pour plus d'informations, voir Chapitre 3, «Text Mining pour les concepts et les catégories», à la page 19.

- **Un nugget de modèle de catégories.** Un nugget de modèle créé à partir d'une session de plan de travail interactif est un nugget de modèle de catégories. Vous devez disposer d'au moins une catégorie dans la vue Catégories et concepts pour pouvoir générer un modèle de catégories. Pour plus d'informations, voir «Nugget de Text Mining : Modèle de catégorie», à la page 40.

Pour générer un noeud modélisation Text Mining

1. Dans les menus, sélectionnez **Générer > Générer le noeud modélisation**. Un noeud modélisation Text Mining est ajouté à l'espace de travail en utilisant l'ensemble des paramètres actuellement définis dans la session interactive. Le nom du noeud est indiqué après le champ de texte.

Pour générer un nugget de modèle de catégories

1. Dans les menus, sélectionnez **Générer > Générer le modèle**. Un nugget de modèle est généré directement dans la palette Modèle avec le nom par défaut.

Mise à jour des noeuds modélisation et enregistrement

Lors d'une session interactive, nous vous conseillons de mettre à jour le noeud modélisation de temps en temps afin d'enregistrer vos modifications. Vous devez également mettre à jour le noeud modélisation chaque fois que vous avez terminé de travailler dans la session de plan de travail interactif et que vous souhaitez enregistrer votre travail. Lorsque vous mettez à jour le noeud modélisation, le contenu de la session interactive est enregistré dans le noeud Text Mining à l'origine d'une session plan de travail interactif. Cela n'a pas pour effet de fermer la fenêtre de sortie.

Important ! Cette mise à jour n'enregistrera pas votre flux. Pour enregistrer votre flux, rendez-vous dans la fenêtre principale de IBM SPSS Modeler après la mise à jour du noeud modélisation.

Pour mettre à jour un noeud modélisation

1. Dans les menus, sélectionnez **Fichier > Mettre à jour le noeud modélisation**. Le noeud modélisation est mis à jour avec les paramètres d'extraction et de création, ainsi qu'avec les options et les catégories que vous avez créées.

Fermeture et fin de sessions

Lorsque vous avez terminé de travailler dans votre session, vous pouvez la quitter de trois manières différentes :

- **Enregistrer.** Cette option permet d'enregistrer au préalable votre travail dans le noeud modélisation d'origine pour les sessions futures, et de publier les bibliothèques pour les réutiliser dans d'autres sessions. Pour plus d'informations, voir «Partage de bibliothèques», à la page 182. Une fois que vous avez sauvegardé votre travail, la fenêtre de session se ferme et la session est supprimée du gestionnaire des sorties dans la fenêtreIBM SPSS Modeler.
- **Quitter.** Cette option supprime tout travail non enregistré, ferme la fenêtre de la session et supprime la session du gestionnaire de sortie de la fenêtre IBM SPSS Modeler. Pour libérer de la mémoire, nous vous conseillons d'enregistrer tout travail important et de quitter la session.
- **Fermer.** Cette option n'enregistre pas ni ne supprime votre travail. Elle ferme la fenêtre de la session, mais la session reste active. Vous pouvez rouvrir la fenêtre de la session en sélectionnant cette session dans le gestionnaire de sortie de la fenêtre IBM SPSS Modeler.

Pour fermer une session interactive

1. Dans les menus, sélectionnez **Fichier > Fermer**.

Accessibilité via le clavier

L'interface du plan de travail interactif propose des raccourcis clavier afin de rendre les fonctionnalités du produit plus accessibles. Vous pouvez appuyer simultanément sur la touche Alt + la touche appropriée pour activer les menus de la fenêtre (par exemple, Alt+F pour accéder au menu Fichier) ou sur la touche Tab pour passer d'une commande à l'autre dans une boîte de dialogue. Dans cette section, nous étudierons les raccourcis clavier, qui proposent une autre manière de naviguer. Il existe d'autres raccourcis clavier pour l'interface IBM SPSS Modeler.

Tableau 13. Raccourcis clavier génériques

Touche de raccourci	Fonction
Ctrl+1	Afficher le premier onglet d'une sous-fenêtre à onglets.
Ctrl+2	Afficher le second onglet d'une sous-fenêtre à onglets.
Ctrl+A	Sélectionner tous les éléments de la sous-fenêtre active.
Ctrl+C	Copier le texte sélectionné dans le Presse-papiers.
Ctrl+E	Lancer une extraction dans les vues Catégories et concepts et Analyse des liens du texte.
Ctrl+F	Afficher la barre d'outils Rechercher dans l'Editeur de ressources/Editeur de modèle, si elle ne l'est pas déjà, et la rendre active.
Ctrl+I	Dans la vue Catégories et concepts, lancer la boîte de dialogue Définitions de catégorie pour la catégorie sélectionnée. Dans la vue Clusters, lancer la boîte de dialogue Définitions du cluster pour le cluster sélectionné.
Ctrl+R	Ouvrir la boîte de dialogue Ajouter des termes dans l'Editeur de ressources/Editeur de modèle.
Ctrl+T	Ouvrir la boîte de dialogue Propriétés de type afin de créer un type dans l'Editeur de ressources/Editeur de modèle.
Ctrl+V	Coller le contenu du presse-papiers.
Ctrl+X	Couper les éléments sélectionnés dans l'Editeur de ressources/Editeur de modèle.
Ctrl+Y	Rétablir la dernière action effectuée dans la vue.
Ctrl+Z	Annuler la dernière action effectuée dans la vue.
F1	Afficher l'Aide ou, dans une boîte de dialogue, afficher l'aide contextuelle d'un élément.
F2	Activer ou désactiver le mode Edition dans les cellules du tableau.
F6	Activer les différents sous-fenêtres principales dans la vue active.
F8	Activer les barres de fractionnement de la sous-fenêtre afin de le redimensionner.
F10	Développer le menu Fichier principal.
Flèche vers le haut, flèche vers le bas	Redimensionner la sous-fenêtre verticalement lorsque la barre de fractionnement est sélectionnée.
Flèche vers la gauche, flèche vers la droite	Redimensionner la sous-fenêtre horizontalement lorsque la barre de fractionnement est sélectionnée.
Origine, Fin	Redimensionner les sous-fenêtres à une taille minimale ou maximale lorsque la barre de fractionnement est sélectionnée.
Tabulation	Passer à l'élément suivant dans la fenêtre, la sous-fenêtre ou la boîte de dialogue.
Maj+F10	Afficher le menu contextuel d'un élément.
Maj+Tab	Passer à l'élément précédent dans la fenêtre ou la boîte de dialogue.
Maj+flèche	Sélectionner les caractères dans le champ Editer en mode Edition (F2).
Ctrl+Tab	Activer la zone principale suivante dans la fenêtre.
Maj+Ctrl+Tab	Activer la zone principale précédente dans la fenêtre.

Raccourcis pour les boîtes de dialogue

Plusieurs touches de raccourci et de lecteur d'écran peuvent être utiles lorsque vous utilisez les boîtes de dialogue. Lorsque vous ouvrez une boîte de dialogue, vous pouvez appuyer sur la touche de tabulation pour activer la première commande et lancer le lecteur d'écran. La liste exhaustive des raccourcis clavier et de lecteur d'écran spéciaux est fournie dans le tableau suivant.

Tableau 14. Raccourcis pour boîtes de dialogue

Touche de raccourci	**Fonction**
Tabulation	Passer à l'élément suivant dans la fenêtre ou la boîte de dialogue.
Ctrl+Tab	Passer d'une zone de texte à l'élément suivant.
Maj+Tab	Passer à l'élément précédent dans la fenêtre ou la boîte de dialogue.
Maj+Ctrl+Tab	Passer d'une zone de texte à l'élément précédent.
barre d'espacement	Sélectionner la commande ou le bouton actif.
Echap	Annuler les modifications et fermer la boîte de dialogue.
Enter	Valider les modifications et fermer la boîte de dialogue (équivalent au bouton OK). Si vous êtes dans une zone de texte, vous devez tout d'abord appuyer sur Ctrl+Tab pour la quitter.

Chapitre 8. Extraction de concepts et de types

Chaque fois que vous exécutez un flux qui lance le plan de travail interactif, une extraction des données textuelles de ce flux est automatiquement effectuée. Le résultat final de cette extraction correspond à un ensemble de concepts, de types, voire de motifs lorsque les ressources linguistiques contiennent des motifs d'analyse de liens du texte (TLA). Vous pouvez afficher et utiliser les concepts et les types dans la sous-fenêtre Résultats d'extraction. Pour plus d'informations, voir «Fonctionnement de l'extraction», à la page 5.

Si vous souhaitez affiner vos résultats d'extraction, vous pouvez modifier les ressources linguistiques et procéder à une réextraction. Pour plus d'informations, voir «Affinage des résultats d'extraction», à la page 89. Le processus d'extraction se base sur les ressources et les paramètres indiqués dans la boîte de dialogue Extraire pour déterminer la manière d'extraire et d'organiser les résultats. Vous pouvez utiliser les résultats d'extraction pour définir la plupart, voire l'ensemble, de vos définitions de catégorie.

Résultats d'extraction : concepts et types

Le processus d'extraction analyse les données textuelles, puis identifie les concepts pertinents, les extrait et les affecte à des types. Une fois l'extraction terminée, les résultats apparaissent dans la sous-fenêtre Résultats d'extraction, dans l'angle inférieur gauche de la vue Catégories et concepts. La première fois que vous lancez une session, c'est le modèle de ressources linguistiques que vous avez sélectionné dans le noeud qui sert à extraire et à organiser ces concepts et ces types.

Remarque : Si le nombre de résultats est plus important que ce que peut contenir la sous-fenêtre visible, vous pouvez utiliser les commandes de la partie inférieure de la sous-fenêtre pour vous déplacer vers l'avant ou vers l'arrière dans les résultats ou entrer un numéro de page pour y accéder.

Les concepts, les types et les motifs TLA extraits sont désignés collectivement par le terme **résultats d'extraction** et servent de descripteurs ou blocs de construction pour vos catégories. Vous pouvez également utiliser des concepts, des types et des motifs dans vos règles de catégorie. De plus, les techniques automatiques s'appuient sur des concepts et des types pour créer des catégories.

Text mining est un processus itératif au cours duquel les résultats de l'extraction sont passés en revue en fonction du contexte des données textuelles. Ils sont ensuite affinés afin de générer de nouveaux résultats avant d'être réévalués. Après l'extraction, passez en revue les résultats et apportez les modifications que vous estimez nécessaires en modifiant les ressources linguistiques. Vous pouvez affiner en partie les ressources, directement à partir de la sous-fenêtre Résultats d'extraction, de la sous-fenêtre Données, de la boîte de dialogue Définitions de catégorie, ou de la boîte de dialogue Définitions de cluster. Pour plus d'informations, voir «Affinage des résultats d'extraction», à la page 89. Vous pouvez également procéder à cette opération directement dans la vue Editeur de ressources. Pour plus d'informations, voir la rubrique «Vue Editeur de ressources», à la page 74.

Après cette optimisation, vous pouvez procéder à une nouvelle extraction pour voir les nouveaux résultats. En affinant vos résultats d'extraction dès le départ, vous êtes assuré d'obtenir à chaque nouvelle extraction des résultats identiques et parfaitement adaptés au contexte des données dans vos définitions de catégorie. De cette manière, l'attribution des documents/ enregistrements à vos définitions de catégorie est plus précise et plus à même d'être répétée.

Concepts

Lors du processus d'extraction, les données textuelles sont analysées pour découvrir des mots isolés intéressants ou pertinents (par exemple, `élection` ou `paix`) et des groupes de mots (par exemple, `élection présidentielle`, `élection du président` ou `traités de paix`). Ces mots et groupes de mots

sont appelés des *termes*. En utilisant les ressources linguistiques, les termes pertinents sont extraits et les termes similaires sont regroupés sous un terme principal appelé **concept**.

Vous pouvez voir l'ensemble des termes sous-jacents pour un concept en passant la souris sur le nom du concept. En procédant ainsi, une info-bulle apparaît indiquant le nom du concept et plusieurs lignes de termes qui sont groupés sous ce concept. Ces termes sous-jacents incluent les synonymes définis dans les ressources linguistiques (que ceux-ci aient été rencontrés ou non dans le texte) ainsi que tous les termes extraits au pluriel/singulier, les termes permutés, les termes provenant du regroupement flou, etc. Vous pouvez également copier ces termes ou voir l'ensemble complet des termes sous-jacents en cliquant avec le bouton droit sur le nom du concept et en choisissant l'option du menu contextuel.

Par défaut, les concepts apparaissent en minuscules et sont triés par ordre décroissant en fonction du nombre de documents . Quand les concepts sont extraits, un type leur est affecté pour regrouper les concepts similaires. Ils apparaissent sous différents codes de couleurs en fonction de ce type. Le choix des couleurs s'effectue sous les propriétés du type, dans l'Editeur de ressources. Pour plus d'informations, voir «Dictionnaires de types», à la page 187.

Chaque fois qu'un concept, type ou motif est utilisé dans une définition de catégorie, une icône apparaît dans la colonne **In** .

Types

Les **types** correspondent à des regroupements sémantiques de concepts. Quand les concepts sont extraits, un type leur est affecté pour regrouper les concepts similaires. Plusieurs types intégrés sont livrés avec IBM SPSS Modeler Text Analytics, tels que `<Location>`, `<Organization>`, `<Person>`, `<Positive>`, `<Negative>`, etc. Par exemple, le type `<Location>` regroupe des mots clés géographiques et des lieux. Ce type est affecté à des concepts tels que `chicago`, `paris` et `tokyo`. Pour la plupart des langues, les concepts qui sont extraits du texte mais ne figurent dans aucun dictionnaire de types sont automatiquement typés `<Unknown>`,Pour plus d'informations, voir «Types intégrés», à la page 188.

Lorsque vous sélectionnez la vue Type, les types extraits apparaissent par défaut en ordre décroissant, classés par fréquence globale. Vous pouvez également remarquer que les types font l'objet d'un codage couleur pour être plus faciles à distinguer. Les couleurs font partie des propriétés du type. Pour plus d'informations, voir «Création de types», à la page 189. Vous pouvez également créer vos propres types.

Motifs

Vous pouvez également extraire des motifs de vos données textuelles. Toutefois, vous devez disposer d'une bibliothèque qui contient des règles de patrons TLA (analyse des liens du texte) dans l'Editeur de ressources. Vous devez choisir d'extraire ces motifs dans le paramètre de noeud IBM SPSS Modeler Text Analytics ou dans la boîte de dialogue Extraire à l'aide de l'option **Extraction avec analyse des liens du texte**. Pour plus d'informations, voir Chapitre 11, «Exploration de l'analyse des liens du texte», à la page 149.

Extraction de données

Chaque fois qu'une extraction est nécessaire, la sous-fenêtre Résultats d'extraction devient jaune et le message **Appuyer sur le bouton Extraire pour extraire les concepts** apparaît sous la barre d'outils de cette sous-fenêtre.

Vous devez procéder à une extraction si vous ne disposez pas encore de résultats d'extraction, si vous avez apporté des modifications aux ressources linguistiques et devez mettre à jour les résultats d'extraction, ou si vous avez réouvert une session où vous n'aviez pas sauvegardé les résultats d'extraction (**Outils** > **Options**).

Remarque : Si vous modifiez le noeud source de votre flux après la mise en cache des résultats d'extraction à l'aide de l'option **Utiliser le travail d'une session...**, vous devrez exécuter une nouvelle extraction une fois la session de plan de travail interactif démarrée, pour obtenir les résultats d'extraction mis à jour.

Lorsque vous effectuez une extraction, un indicateur de progression apparaît pour vous fournir des informations sur l'état de l'extraction. Pendant ce temps, le moteur du programme d'extraction analyse toutes les données textuelles et identifie les termes et motifs pertinents puis les extrait et les attribue à un type. Ensuite, le moteur tente de regrouper les synonymes sous un terme principal, appelé un concept. Une fois le processus terminé, les concepts, les types et les motifs obtenus apparaissent dans la sous-fenêtre Résultats d'extraction.

Le résultat du processus d'extraction correspond à un ensemble de concepts, de types et de motifs d'analyse des liens du texte (TLA) si la fonction a été activée. Vous pouvez afficher et utiliser ces concepts et ces types dans la sous-fenêtre Résultats d'extraction de la vue Catégories et concepts. Si vous avez extrait des motifs TLA, vous pouvez les visualiser dans la vue Analyse des liens du texte.

Remarque : Il existe une relation entre la taille du jeu de données et le temps requis pour procéder à l'extraction. Vous pouvez toujours envisager d'insérer un noeud Echantillonner en amont ou d'optimiser la configuration de votre machine.

Pour extraire des données

1. A partir des menus, sélectionnez **Outils** > **Extraire**. Vous pouvez également cliquer sur le bouton **Extraire** de la barre d'outils.
2. Si vous choisissez de toujours afficher la boîte de dialogue Paramètres d'extraction, elle apparaît pour vous permettre d'apporter des modifications. Vous trouverez plus loin dans cette rubrique des descriptions de chaque paramètre.
3. Cliquez sur **Extraire** pour lancer le processus d'extraction. Dès le début de l'extraction, une boîte de dialogue indique la progression du processus. Après l'extraction, les résultats apparaissent dans la sous-fenêtre Résultats d'extraction. Par défaut, les concepts apparaissent en minuscules et sont triés par ordre décroissant en fonction du nombre de documents .

Vous pouvez examiner les résultats à l'aide des options de la barre d'outils pour trier les résultats différemment, les filtrer, ou encore afficher une autre vue (concepts ou types). Vous pouvez aussi redéfinir vos résultats d'extraction en travaillant avec les ressources linguistiques. Pour plus d'informations, voir la rubrique «Affinage des résultats d'extraction», à la page 89.

Problèmes potentiels d'extraction

Les sessions multiples de plan de travail interactif peuvent ralentir le système. SPSS Modeler Text Analytics et SPSS Modeler partagent un même moteur d'exécution Java lorsqu'une session de plan de travail interactif est lancée. En fonction du nombre de sessions de plan de travail interactif que vous appelez lors d'une session SPSS Modeler, même si vous ouvrez et fermez une même session, la mémoire système peut ralentir l'application. Cet effet peut être particulièrement prononcé si vous traitez un volume important de données ou que votre machine dispose d'une quantité de mémoire vive inférieure à celle de 4 Go recommandée. Si vous constatez que votre machine est lente à répondre, sauvegardez tout votre travail, arrêtez SPSS Modeler, puis relancez l'application. L'exécution de SPSS Modeler Text Analytics sur une machine dont la quantité de mémoire est inférieure à celle recommandée, en particulier si vous traitez un volume important de données ou que les traitements sont longs, peut entraîner un arrêt de Java pour mémoire insuffisante. Il est fortement conseillé d'ajouter de la mémoire pour atteindre ou dépasser la quantité de mémoire recommandée (ou d'utiliser le serveur SPSS Modeler Text Analytics) si vous traitez un volume important de données.

Pour le texte en néerlandais, en anglais, en français, en allemand, en italien, en portugais, et en espagnol

La boîte de dialogue Paramètres d'extraction contient des options d'extraction élémentaires.

Activer l'extraction d'analyse des liens du texte. Indique que vous souhaitez extraire les patrons TLA de vos données textuelles. L'option suppose également que vous disposez de règles de patrons TLA dans l'une de vos bibliothèques de l'éditeur de ressources. Cette option risque d'augmenter considérablement la durée de l'extraction. Pour plus d'informations, voir Chapitre 11, «Exploration de l'analyse des liens du texte», à la page 149.

Adapter les erreurs de ponctuation. Cette option normalise temporairement le texte contenant des erreurs de ponctuation (par exemple, usage incorrect) au cours de l'extraction pour améliorer l'extraction des concepts. Cette option s'avère extrêmement utile lorsque le texte est court et de qualité médiocre (réponses ouvertes, messages électroniques, données CRM, etc.) ou qu'il contient de nombreuses abréviations.

Traitement des fautes de frappe. Nombre de caractères minimum requis : [n] Cette option applique une technique de regroupement flou qui permet de regrouper les mots mal orthographiés ou d'orthographe similaire sous un seul concept. L'algorithme de regroupement flou retire temporairement toutes les voyelles (à l'exception de la première) et les consonnes doubles/triples des mots extraits pour comparer ces derniers et voir s'ils sont identiques, afin par exemple que `modélisation` et `modéllisation` soient regroupés. Néanmoins, si chaque type est affecté à un type différent et que vous excluez le type `<Unknown>`, le regroupement flou n'est pas appliqué.

Vous pouvez également définir le nombre minimum de caractères *racines* requis avant d'utiliser le regroupement flou. Le nombre de caractères racine d'un terme est calculé en ajoutant l'ensemble des caractères et en soustrayant les caractères formant des suffixes inflexionnels et, dans le cas des termes apparaissant sous la forme de mots composés, les déterminants et les prépositions. Par exemple, le terme `exercices` sera considéré comme contenant 8 caractères racines dans la forme “exercice,” la lettre *s* de fin étant une inflexion (marque du pluriel). De même, `sauce soja` contient 9 caractères racines (“sauce soja”) et `usine de voitures` en contient 12 (“usine voiture”). Cette méthode de calcul est uniquement utilisée pour vérifier si le regroupement flou doit être appliqué, mais n'influe pas sur la mise en correspondance des mots.

Remarque : Si vous constatez plus tard que certains mots sont regroupés par erreur, vous pouvez exclure des paires de mots en les déclarant explicitement dans la section **Regroupement flou : Exceptions** de l'onglet Ressources avancées. Pour plus d'informations, voir «Regroupement flou», à la page 203.

Enlever les expressions unitermes Cette option extrait les mots uniques (unitermes) dans la mesure où le mot ne fait pas déjà partie d'un mot composé et si c'est un nom ou une catégorie grammaticale non reconnue.

Extraire les entités non linguistiques Cette option extrait les entités non linguistiques, telles que les numéros de téléphone, numéros de sécurité sociale, heures, dates, devises, chiffres, pourcentages, adresses électroniques, adresses HTTP, etc. Vous pouvez inclure ou exclure certains types d'entités non linguistiques dans la section **Entités non linguistiques : Configuration** de l'onglet Ressources avancées. Désactivez les entités dont vous n'avez pas besoin pour éviter au moteur d'extraction un temps de traitement inutile. Pour plus d'informations, voir «Configuration», à la page 208.

Algorithme des majuscules Cette option extrait les termes simples et composés ne figurant pas dans les dictionnaires intégrés, si la première lettre est une majuscule. Cette option offre une bonne manière d'extraire la plupart des noms propres.

Regrouper si possible les noms de personnes partiels et complets Cette option regroupe les noms qui apparaissent différemment dans le texte. Cette fonction est utile car les noms sont souvent cités dans leur forme complète au début du texte puis uniquement en version abrégée. Cette option essaye de faire correspondre tout uniterme ayant le type `<Unknown>` avec le dernier mot de tout terme composé entré comme `<Person>`. Par exemple, si *martin* est trouvé et initialement saisi comme `<Unknown>`, le moteur

d'extraction vérifie si un terme composé de type <Person> contient *martin* comme dernier mot, tel que *pierre martin*. Cette option ne s'applique pas aux prénoms, car la plupart ne sont jamais extraits comme unitermes.

Taille maximale pour la permutation des mots utiles Cette option indique le nombre maximal de mots utiles pouvant être présents lorsque s'applique la technique de permutation. Cette technique de permutation regroupe les expressions similaires qui ne diffèrent les unes des autres que par la présence de mots utiles (par exemple, de et la), quelle que soit leur inflexion. Par exemple, si vous avez paramétré cette valeur sur au moins deux mots et si les deux expressions `représentants d'entreprise` et `représentants de l'entreprise` ont été extraites. Dans ce cas, les deux termes extraits sont regroupés dans la liste de concepts finale car les deux termes sont considérés comme étant les mêmes lorsque `de l'` est ignoré.

Utiliser la dérivation des termes (grouper par composants multitermes) Pour le traitement de Big Data, sélectionnez cette option pour regrouper les multitermes à l'aide de règles de dérivation.

Option de l'index pour la carte de concept Indique que vous souhaitez créer l'index de la carte au moment de l'extraction afin que les cartes de concept puissent être rapidement tracées plus tard. Pour modifier les paramètres de l'index, cliquez sur **Paramètres**. Pour plus d'informations, voir «Génération d'un index de relations de concept», à la page 89.

Toujours afficher cette boîte de dialogue avant le début d'une extraction Spécifiez si vous souhaitez que la boîte de dialogue Paramètres d'extraction apparaisse à chaque fois que vous effectuez une extraction, si vous ne voulez pas qu'elle s'affiche sauf si vous l'ouvrez via le menu Outils ou si vous voulez qu'on vous demande à chaque fois que vous effectuez une extraction si vous souhaitez modifier les paramètres d'extraction.

Filtrage des résultats d'extraction

Lorsque vous travaillez sur des ensembles de données très volumineux, le processus d'extraction peut renvoyer des millions de résultats. Pour de nombreux utilisateurs, cette quantité peut compliquer l'examen des résultats. Par conséquent, afin de mettre en évidence les résultats les plus intéressants, vous pouvez les filtrer à l'aide de la boîte de dialogue Filtrer dans la sous-fenêtre Résultats d'extraction.

N'oubliez pas que tous les paramètres de cette boîte de dialogue Filtrer sont utilisés pour filtrer les résultats d'extraction disponibles pour les catégories.

Filtrer par fréquence Vous pouvez appliquer un filtre afin de n'afficher que les résultats présentant une certaine valeur de fréquence globale ou de documents.

- La **fréquence globale** est le nombre total d'apparitions d'un concept dans l'ensemble de documents ou d'enregistrements. Cette valeur apparaît dans la colonne **Global**.
- La **fréquence de documents** est le nombre total de documents ou d'enregistrements dans lesquels un concept apparaît. Cette valeur est visible dans la colonne **Docs**.

Par exemple, si le concept `otan` est apparu 800 fois dans 500 enregistrements, nous en déduisons qu'il présente une fréquence globale de 800 et une fréquence de document de 500.

Et par type Vous pouvez appliquer un filtre qui n'affiche que les résultats appartenant à certains types. Vous pouvez choisir tous les types ou uniquement des types spécifiques.

Et par texte correspondant Vous pouvez également appliquer un filtre n'affichant que les résultats correspondant à la règle que vous définissez ici. Entrez l'ensemble de caractères devant être renvoyés dans le champ **Texte correspondant** et sélectionnez la condition dans laquelle il faut appliquer la correspondance.

Tableau 15. Conditions de correspondance de texte

Condition	Description
Contient	Le texte est mis en correspondance si la chaîne apparaît n'importe où. (Option par défaut)
Commence par	Le texte est seulement mis en correspondance si le concept ou le type commence par le texte entré.
Se termine par	Le texte est seulement mis en correspondance si le concept ou le type se termine par le texte entré.
Correspondance totale	Toute la chaîne doit concorder avec le nom du concept ou du type.

Résultats affichés dans la sous-fenêtre Résultats d'extraction

Voici quelques exemples de la manière dont les résultats peuvent s'afficher, en anglais, dans la barre d'outils de la sous-fenêtre Résultats d'extraction en fonction des filtres.

Tableau 16. Exemples d'informations sur les filtres

Informations sur les filtres	Description
	La barre d'outils indique le nombre de résultats. Comme aucun filtre de correspondance de texte n'est défini et que le maximum n'est pas atteint, aucune icône supplémentaire n'apparaît.
	La barre d'outils indique le nombre de résultats en fonction du nombre maximal spécifié dans le filtre, 300 dans ce cas de figure. La présence d'une icône violette indique que le nombre maximal de concepts est atteint. Placez le curseur sur l'icône pour obtenir plus d'informations.
	La barre d'outils montre que les résultats ont été limités à l'aide d'un filtre de correspondance de texte. Cela est indiqué par l'icône représentant une loupe.

Pour filtrer les résultats

1. Dans les menus, sélectionnez **Outils** > **Filtrer**. La boîte de dialogue Filtrer s'ouvre.
2. Sélectionnez et affinez les filtres à utiliser.
3. Cliquez sur **OK** pour appliquer les filtres et visualiser les nouveaux résultats dans la sous-fenêtre Résultats d'extraction.

Exploration des cartes de concept

Vous pouvez créer une carte de concepts pour explorer les interrelations entre les concepts. En sélectionnant un concept unique et en cliquant sur **Carte**, une fenêtre de carte de concept s'ouvre et vous permet d'explorer l'ensemble des concepts associés au concept sélectionné. Vous pouvez filtrer les concepts à afficher en modifiant les paramètres tels que les types à inclure, les types de relations à rechercher, etc.

Important : Avant de pouvoir créer une carte, il convient de générer un index. L'opération peut prendre plusieurs minutes. Cependant, une fois que vous avez généré l'index, vous n'avez pas à le générer de nouveau jusqu'à ce que vous procédiez à une nouvelle extraction. Si vous souhaitez que l'index soit généré automatiquement à chaque extraction, sélectionnez cette option dans les paramètres d'extraction. Pour plus d'informations, voir «Extraction de données», à la page 82.

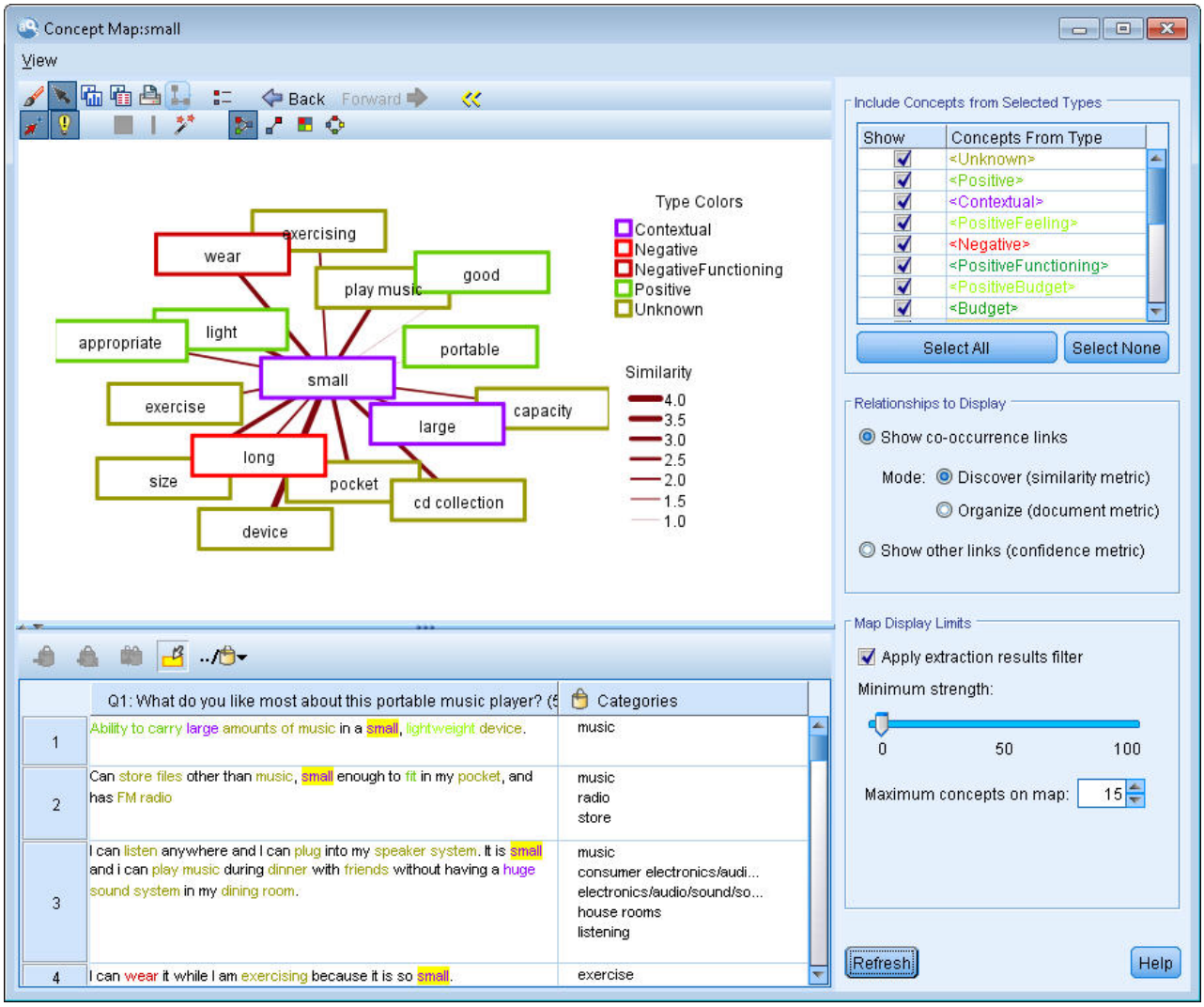


Figure 27. Une carte de concepts pour le concept sélectionné

Pour visualiser une carte de concepts

1. Dans la sous-fenêtre Résultats d'extraction, sélectionnez un concept unique.
2. Dans la barre d'outils de cette sous-fenêtre, cliquez sur le bouton **Carte**. Si l'index de la carte a déjà été généré, la carte de concept s'ouvre dans une boîte de dialogue distincte. Si l'index de la carte n'a pas été généré ou qu'il est obsolète, l'index doit être reconstruit. Ce processus peut durer plusieurs minutes.
3. Cliquez sur la carte pour l'explorer. Si vous double-cliquez sur un concept lié, la carte se redessine automatiquement et vous montre les concepts liés pour le concept sur lequel vous venez de double-cliquer.
4. La barre d'outils supérieure propose quelques outils de base pour la carte tels que le retour à une carte précédente, des liens de filtrage en fonction de la puissance des relations ainsi que l'ouverture de la boîte de dialogue du filtre pour contrôler les types de concepts qui apparaissent ainsi que les types de relations à représenter. Une seconde ligne dans la barre d'outils contient les outils d'édition de graphiques. Pour plus d'informations, voir «Utilisation des palettes et des barres d'outils de graphiques», à la page 159.
5. Si les types de liens trouvés ne vous donnent pas satisfaction, consultez de nouveau les paramètres de cette carte qui se trouvent à droite de la carte.

Paramètres de carte : Inclure les concepts des types sélectionnés

Seuls les concepts appartenant aux types sélectionnés dans le tableau apparaissent sur la carte. Pour masquer les concepts d'un certain type, désélectionnez ce type dans le tableau.

Paramètres de carte : Relations à afficher

Afficher les liens de co-occurrence Si vous souhaitez afficher des liens de co-occurrence, choisissez ce mode. Ce mode influe sur la façon dont la puissance des liens est calculée.

- *Découvrir (mesure de similarité)*. Avec cette mesure, la puissance du lien est calculée à l'aide d'un calcul plus complexe qui prend en compte la fréquence à laquelle deux concepts apparaissent séparément et celle à laquelle ils apparaissent ensemble. Une valeur de puissance élevée que deux concepts ont tendance à apparaître plus souvent ensemble que séparément. Avec la formule suivante, les valeurs à virgule flottante sont converties en entiers.

$$\text{similarity coefficient} = \frac{(C_{IJ})^2}{(C_I \times C_J)}$$

Figure 28. Formule du coefficient de similarité

  Dans cette formule, C_I est le nombre de documents ou d'enregistrements dans lequel apparaît le concept I.

  C_J est le nombre de documents ou d'enregistrements dans lequel apparaît le concept J.

  C_{IJ} est le nombre de documents ou d'enregistrements dans lequel la paire de concepts I et J est co-occurrente dans l'ensemble de documents.
- *Organiser (mesure de document)*. La puissance de ces liens avec cette mesure est déterminée par le nombre brut de co-occurrences. En général, plus deux concepts sont fréquents, plus ils ont de chances d'apparaître ensemble. Une valeur de puissance élevée que deux concepts apparaissent souvent ensemble.

Afficher les autres liens (mesure de confiance). Vous pouvez choisir d'autres liens à afficher ; ils peuvent être sémantiques, de dérivation (morphologiques) ou d'inclusion (syntaxiques) et sont associés au nombre d'étapes supprimées d'un concept par rapport au concept auquel il est associé. Ils peuvent également vous aider à affiner des ressources (synonymie) ou à lever des ambiguïtés. Pour plus de détails sur les méthodes de regroupement, voir «Paramètres linguistiques avancés», à la page 108

Remarque : Aucun de ces éléments ne sera affiché s'ils n'ont pas été sélectionnés lors de la génération de l'index ou si aucune relation n'est détectée. Pour plus d'informations, voir «Génération d'un index de relations de concept», à la page 89.

Paramètres de carte : Limites d'affichage de la carte

Appliquer le filtre des résultats d'extraction. Si vous ne souhaitez pas utiliser tous les concepts, vous pouvez utiliser le filtre de la sous-fenêtre des résultats d'extraction pour limiter l'affichage. Sélectionnez ensuite cette option et IBM SPSS Modeler Text Analytics recherchera les concepts associés à l'aide de l'ensemble de filtres. Pour plus d'informations, voir «Filtrage des résultats d'extraction», à la page 85.

Force minimale. Définissez ici la force minimale des liens. Tous les concepts associés ayant une force de relation inférieure à cette limite n'apparaîtront pas sur la carte.

Nombre de concepts maximal sur la carte. Définissez le nombre maximal de relations à afficher sur la carte.

Génération d'un index de relations de concept

Avant de pouvoir créer une carte, il convient de générer un index de relations de concept. Chaque fois que vous créez une carte de concept, IBM SPSS Modeler Text Analytics fait référence à cet index. Vous pouvez choisir les relations à indexer en sélectionnant les techniques dans cette boîte de dialogue

Techniques de regroupement. Choisissez une ou plusieurs techniques. Pour des descriptions brèves de chacune de ces techniques, voir «A propos des techniques linguistiques», à la page 110 Toutes les techniques ne sont pas disponibles dans toutes les langues.

Empêcher l'appariement de concepts spécifiques. Cochez cette case pour arrêter le processus de regroupement ou d'appariement de deux concepts dans les résultats. Pour créer ou gérer des paires de concepts, cliquez sur **Gérer les paires**. Pour plus d'informations, voir «Gestion des paires d'exceptions de liens», à la page 110.

La création de l'index peut durer plusieurs minutes. Cependant, une fois l'index généré, il est inutile de le générer de nouveau jusqu'à la prochaine extraction ou si vous souhaitez modifier des paramètres pour inclure plus de relations. Si vous souhaitez générer un index à chaque extraction, vous pouvez sélectionner une option dans les paramètres d'extraction. Pour plus d'informations, voir «Extraction de données», à la page 82.

Affinage des résultats d'extraction

L'extraction est un processus itératif selon lequel vous pouvez extraire des résultats, les examiner, les modifier, puis procéder à une nouvelle extraction pour les mettre à jour. Etant donné que l'exactitude et la continuité sont essentielles à la réussite de l'exploration de texte et de la catégorisation, le fait d'affiner les résultats de l'extraction dès le départ garantit qu'à chaque nouvelle extraction, vous obtenez exactement les mêmes résultats dans vos définitions de catégorie. De cette manière, l'attribution des documents et des enregistrements à vos catégories est plus précise et plus répétable.

Les résultats de l'extraction font office de blocs de construction pour les catégories. Lorsque vous créez des catégories à l'aide de ces résultats d'extraction, les documents et les enregistrements sont automatiquement attribués aux catégories s'ils contiennent du texte qui correspond à un ou plusieurs descripteurs de catégorie. Bien que vous puissiez commencer la catégorisation avant d'affiner les ressources linguistiques, il est utile d'examiner les résultats de l'extraction au moins une fois au préalable.

Durant l'examen des résultats, vous pouvez trouver des éléments pour lesquels vous souhaitez que le moteur du programme d'extraction se comporte de manière différente. Prenons les exemples suivants :

- **Synonymes non reconnus.** Supposons que vous trouviez plusieurs concepts qui, selon vous, sont synonymes, par exemple `intelligent`, `astucieux`, `brillant` et `ingénieux`, et qu'ils apparaissent tous en tant que concepts individuels dans les résultats d'extraction. Vous pouvez créer une définition de synonyme dans laquelle les concepts `astucieux`, `brillant` et `ingénieux` sont regroupés sous le concept cible `intelligent`. Cette action regroupe ainsi tous les concepts avec `intelligent` et l'effectif de fréquences global est alors supérieur. Pour plus d'informations, voir «Ajout de synonymes», à la page 90.
- **Concepts mal définis.** Supposons que les concepts de vos résultats d'extraction apparaissent sous un type et vous souhaitez qu'ils soient affectés à un autre type. Dans un autre exemple, imaginez que vous trouviez 15 concepts relatifs aux légumes dans vos résultats d'extraction et que vous souhaitiez tous les ajouter à un nouveau type appelé `<Légume>`. Pour la plupart des langues, les concepts qui sont extraits du texte mais ne figurent dans aucun dictionnaire de types sont automatiquement typés `<Unknown>`, Vous pouvez ajouter des concepts aux types. Pour plus d'informations, voir «Ajout de concepts à des types», à la page 91.
- **Concepts non pertinents.** Supposons que vous trouviez un concept qui a été extrait à une fréquence très importante, c'est-à-dire qu'il figure dans de nombreux enregistrements ou documents. Toutefois, vous considérez que ce concept n'est pas pertinent pour l'analyse. Vous pouvez l'exclure de l'extraction. Pour plus d'informations, voir «Exclusion de concepts de l'extraction», à la page 92.

- **Correspondances incorrectes.** Supposons que lors de l'examen des enregistrements ou des documents qui contiennent un certain concept, vous découvriez que deux mots ont été regroupés de façon incorrecte, par exemple `faculté` et `facilité`. Cette correspondance peut être due à un algorithme interne, connu sous le nom de regroupement flou, qui ignore temporairement les consonnes et voyelles doubles ou triples, de façon à regrouper les fautes d'orthographe courantes. Vous pouvez ajouter ces mots à une liste de paires de mots qui ne doivent pas être regroupés. Pour plus d'informations, voir «Regroupement flou», à la page 203.
- **Concepts non extraits.** Supposons que vous remarquiez, au moment de l'examen du texte du document ou de l'enregistrement, que quelques mots ou expressions n'ont pas été extraits alors que vous vous attendiez à ce qu'ils le soient. Il s'agit souvent de verbes ou d'adjectifs qui ne vous intéressent pas. Cependant, il arrive parfois que vous souhaitiez utiliser un mot ou une expression qui n'a pas été extrait comme faisant partie d'une définition de catégorie. Pour extraire le concept, vous pouvez imposer un terme dans un dictionnaire de types. Pour plus d'informations, voir «Extraction de mots imposée», à la page 93.

La plupart des modifications peuvent être effectuées directement à partir de la sous-fenêtre Résultats d'extraction , de la sous-fenêtre Données, de la boîte de dialogue Définitions de catégorie, ou de la boîte de dialogue Définitions de cluster en sélectionnant un ou plusieurs éléments et en cliquant avec le bouton droit de la souris pour accéder aux menus contextuels.

Une fois les modifications apportées, la couleur d'arrière-plan de la sous-fenêtre change pour indiquer que vous devez procéder à une nouvelle extraction pour visualiser ces modifications. Pour plus d'informations, voir «Extraction de données», à la page 82. Si vous travaillez avec des jeux de données plus importants, il peut être plus efficace de procéder à une nouvelle extraction après plusieurs modifications.

Remarque : Vous pouvez visualiser l'ensemble des ressources linguistiques éditables utilisées pour produire les résultats de l'extraction dans la vue Editeur de ressources (Vue > Editeur de ressources). Dans cette vue, les ressources apparaissent sous la forme de bibliothèques et de dictionnaires. Vous pouvez personnaliser les concepts et les types directement au sein des bibliothèques et des dictionnaires. Pour plus d'informations, voir Chapitre 15, «Utilisation des bibliothèques», à la page 177.

Ajout de synonymes

Les *synonymes* sont des mots ayant le même sens. Les synonymes sont souvent utilisés pour regrouper des termes et leurs abréviations, ou pour réunir les mots fréquemment mal orthographiés sous la version correcte du mot. Grâce à l'utilisation de synonymes, la fréquence du concept cible est plus élevée, ce qui facilite largement la recherche d'informations similaires qui se présentent sous différentes formes dans vos données textuelles.

Les modèles et bibliothèques de ressources linguistiques fournis avec le produit contiennent de nombreux synonymes prédéfinis. Néanmoins, si vous découvrez des synonymes non reconnus, vous pouvez les définir de façon à ce qu'ils soient reconnus lors de la prochaine extraction.

La première étape consiste à décider du concept cible, ou concept principal. Le *concept cible* est l'expression ou le mot sous lequel vous souhaitez regrouper tous les termes synonymes dans les résultats finaux. Au cours de l'extraction, les synonymes sont regroupés sous ce concept cible. La deuxième étape consiste à identifier tous les synonymes de ce concept. Le concept cible vient remplacer tous les synonymes dans l'extraction finale. Pour être un synonyme, un terme doit être extrait. En revanche, il n'est pas nécessaire que le concept cible soit extrait pour que la substitution se produise. Par exemple, si vous souhaitez que le terme `astucieux` soit remplacé par `intelligent`, alors `astucieux` est le synonyme et `intelligent` est le concept cible.

Si vous créez une définition de synonyme, un nouveau concept cible est ajouté au dictionnaire. Vous devez ensuite ajouter des synonymes à ce concept cible. Lorsque vous créez ou éditez des synonymes, ces modifications sont enregistrées dans les dictionnaires de synonymes de l'Editeur de ressources. Pour

visualiser la totalité du contenu de ces dictionnaires de synonymes ou pour apporter un nombre important de modifications, travaillez plutôt directement dans l'Editeur de ressources. Pour plus d'informations, voir «Dictionnaires de substitutions/synonymes», à la page 195.

Les nouveaux synonymes sont automatiquement stockés dans la première bibliothèque répertoriée dans l'arborescence de bibliothèques d'Editeur de ressources ; par défaut il s'agit de la **bibliothèque locale**.

Remarque : Si vous recherchez une définition de synonyme et que vous ne la trouvez ni via les menus contextuels ni directement dans l'Editeur de ressources, une correspondance a peut-être été obtenue à partir d'une technique interne de regroupement flou. Pour plus d'informations, voir «Regroupement flou» , à la page 203.

Pour créer un synonyme

1. Dans la sous-fenêtre Résultats d'extraction, la sous-fenêtre Données, la boîte de dialogue Définitions de catégorie, ou la boîte de dialogue Définitions de cluster, sélectionnez le ou les concepts pour lesquels vous voulez créer un synonyme.
2. Dans les menus, choisissez **Edition** > **Ajouter au synonyme** > **Nouveau**. La boîte de dialogue Créer un synonyme apparaît.
3. Entrez un concept cible dans la zone de texte Cible. Il s'agit du concept sous lequel tous les synonymes seront regroupés.
4. Si vous souhaitez ajouter davantage de synonymes, entrez-les dans la zone de liste Synonymes. Utilisez le séparateur global pour séparer chaque terme synonyme. Pour plus d'informations, voir la rubrique «Options : onglet Session», à la page 76.
5. Cliquez sur **OK** pour appliquer les modifications. La boîte de dialogue se ferme et la couleur d'arrière-plan de la sous-fenêtre Résultats d'extraction change, indiquant que vous devez procéder à une nouvelle extraction pour visualiser les modifications. Si vous prévoyez plusieurs modifications, apportez-les avant de procéder à une nouvelle extraction.

Pour ajouter un synonyme

1. Dans la sous-fenêtre Résultats d'extraction, la sous-fenêtre Données, la boîte de dialogue Définitions de catégorie, ou la boîte de dialogue Définitions de cluster, sélectionnez le ou les concepts que vous voulez ajouter à la définition d'un synonyme existant.
2. Dans les menus, choisissez **Edition** > **Ajouter au synonyme** . Le menu affiche un ensemble de synonymes, plaçant en début de liste les synonymes créés le plus récemment. Sélectionnez le nom du synonyme auquel vous souhaitez ajouter les concepts sélectionnés. Si vous trouvez le synonyme recherché, sélectionnez-le ; les concepts sélectionnés sont ajoutés à cette définition de synonyme. Si vous ne le trouvez pas, sélectionnez **Plus** pour afficher la boîte de dialogue Tous les synonymes.
3. Dans la boîte de dialogue Tous les synonymes, vous pouvez trier la liste selon l'ordre de tri naturel (ordre de création), ou selon l'ordre croissant ou décroissant. Sélectionnez le nom du synonyme auquel vous souhaitez ajouter les concepts sélectionnés et cliquez sur **OK**. La boîte de dialogue se ferme et les concepts sont ajoutés à la définition de synonyme.

Ajout de concepts à des types

Au cours d'une extraction, les concepts extraits sont affectés à des types dans le but de regrouper les termes ayant quelque chose en commun. IBM SPSS Modeler Text Analytics est livré avec de nombreux types intégrés. Pour plus d'informations, voir «Types intégrés», à la page 188. Pour la plupart des langues, les concepts qui sont extraits du texte mais ne figurent dans aucun dictionnaire de types sont automatiquement typés `<Unknown>`,

Lors de l'analyse de vos résultats, vous pouvez découvrir que certains concepts apparaissent dans un type que vous voulez attribuer à un autre ou qu'un groupe de mots appartient en fait à un nouveau type propre. Dans ce cas, vous pouvez réaffecter les concepts à un autre type ou créer un nouveau type.

Par exemple, supposons que vous travaillez avec des données d'enquête liées aux automobiles et que vous souhaitez procéder à une catégorisation en vous centrant sur différentes parties des véhicules. Vous pouvez créer un type appelé `<Tableau de bord>` pour regrouper tous les concepts liés aux indicateurs et boutons situés sur le tableau de bord des véhicules. Ensuite, vous pouvez attribuer des concepts tels que `jauge de carburant`, `chauffage`, `radio` et `compteur kilométrique` à ce nouveau type.

Autre exemple : supposons que vous travaillez avec des données d'enquête liées aux universités et que l'extraction a défini `Jean Moulin` (l'université) en tant que type `<Person>` plutôt qu'en tant que type `<Organization>`. Dans ce cas, vous pouvez ajouter ce concept au type `<Organization>`.

Lorsque vous créez un type ou ajoutez des concepts à une liste de termes d'un type, ces modifications sont enregistrées dans des dictionnaires de types dans les bibliothèques de ressources linguistiques de l'Editeur de ressources. Pour visualiser le contenu de ces bibliothèques ou pour apporter un nombre important de modifications, travaillez plutôt directement dans l'Editeur de ressources. Pour plus d'informations, voir «Ajout de termes», à la page 190.

Pour ajouter un concept à un type

1. Dans la sous-fenêtre Résultats d'extraction, la sous-fenêtre Données, la boîte de dialogue Définitions de catégorie, ou la boîte de dialogue Définitions de cluster, sélectionnez le ou les concepts que vous voulez ajouter à la définition d'un type existant.
2. Cliquez avec le bouton droit de la souris pour ouvrir le menu contextuel.
3. Dans les menus, choisissez **Edition** > **Ajouter au type** . Le menu affiche un ensemble de types, plaçant en début de liste les types créés le plus récemment. Sélectionnez le nom du type auquel vous souhaitez ajouter les concepts sélectionnés. Si vous trouvez le nom du type recherché, sélectionnez-le ; les concepts sélectionnés sont ajoutés à ce type. Si vous ne le trouvez pas, sélectionnez **Plus** pour afficher la boîte de dialogue Tous les types.
4. Dans la boîte de dialogue Tous les types, vous pouvez trier la liste selon l'ordre de tri naturel (ordre de création) ou selon l'ordre croissant ou décroissant. Sélectionnez le nom du type auquel vous souhaitez ajouter les concepts sélectionnés et cliquez sur **OK**. La boîte de dialogue se ferme et les concepts sont ajoutés au type en tant que termes.

Pour créer un type

1. Dans la sous-fenêtre Résultats d'extraction, la sous-fenêtre Données, la boîte de dialogue Définitions de catégorie, ou la boîte de dialogue Définitions de cluster, sélectionnez les concepts pour lesquels vous voulez créer un type.
2. Dans les menus, choisissez **Edition** > **Ajouter au type** > **Nouveau**. La boîte de dialogue Propriétés de type s'ouvre.
3. Dans la zone de texte Nom, entrez le nom de ce nouveau type et apportez éventuellement des modifications aux autres champs. Pour plus d'informations, voir «Création de types», à la page 189.
4. Cliquez sur **OK** pour appliquer les modifications. La boîte de dialogue se ferme et la couleur d'arrière-plan de la sous-fenêtre Résultats d'extraction change, indiquant que vous devez procéder à une nouvelle extraction pour visualiser les modifications. Si vous prévoyez plusieurs modifications, apportez-les avant de procéder à une nouvelle extraction.

Exclusion de concepts de l'extraction

Lors de l'examen des résultats, vous pouvez éventuellement découvrir des concepts dont vous ne souhaitiez pas l'extraction ou l'utilisation par une technique de génération de catégorie automatisée. Dans certains cas, ces concepts présentent un effectif de fréquences très élevé et ne sont pas du tout pertinents pour votre analyse. Vous pouvez dès lors marquer un concept à exclure de l'extraction finale. En règle générale, les concepts que vous entrez dans cette liste sont des mots ou des expressions de liaison utilisés pour la continuité du texte, mais qui ne lui apportent rien d'important et risquent d'encombrer les résultats de l'extraction. En ajoutant ces concepts au dictionnaire d'exclusions, vous êtes assuré qu'ils ne seront jamais extraits.

Le processus d'exclusion implique que toutes les variantes du concept exclu disparaissent des résultats lors de la prochaine extraction. Si ce concept apparaît déjà en tant que descripteur dans une catégorie, il restera dans la catégorie avec un effectif nul après la nouvelle extraction.

Lorsque vous excluez, ces modifications sont enregistrées dans un dictionnaire d'exclusions dans l'Editeur de ressources. Pour visualiser toutes les définitions d'exclusion et les éditer directement, travaillez plutôt directement dans l'Editeur de ressources. Pour plus d'informations, voir «Dictionnaires d'exclusions», à la page 198.

Pour exclure des concepts

1. Dans la sous-fenêtre Résultats d'extraction, la sous-fenêtre Données, la boîte de dialogue Définitions de catégorie, ou la boîte de dialogue Définitions de cluster, sélectionnez le ou les concepts que vous voulez exclure de l'extraction.
2. Cliquez avec le bouton droit de la souris pour ouvrir le menu contextuel.
3. Sélectionnez **Exclure de l'extraction**. Le concept est ajouté au dictionnaire d'exclusions dans l'Editeur de ressources et la couleur d'arrière-plan de la sous-fenêtre Résultats d'extraction change, indiquant que vous devez procéder à une nouvelle extraction pour visualiser les modifications. Si vous prévoyez plusieurs modifications, apportez-les avant de procéder à une nouvelle extraction.

Remarque : Tout mot exclu est automatiquement stocké dans la première bibliothèque listée dans l'arborescence de bibliothèques d'Editeur de ressources ; par défaut, il s'agit de la **bibliothèque locale**.

Extraction de mots imposée

Lors de l'examen des données textuelles dans la sous-fenêtre Données après l'extraction, vous pouvez découvrir que certains mots ou expressions n'ont pas été extraits. Il s'agit souvent de verbes ou d'adjectifs qui ne vous intéressent pas. Cependant, il arrive parfois que vous souhaitiez utiliser un mot ou une expression qui n'a pas été extrait comme faisant partie d'une définition de catégorie.

Si vous souhaitez que ces mots et expressions soient extraits, vous pouvez imposer un terme dans une bibliothèque de types. Pour plus d'informations, voir «Ajout des termes forcés», à la page 193.

Important ! Le marquage d'un terme dans un dictionnaire comme étant imposé n'est pas infaillible. En effet, même si vous avez explicitement ajouté un terme à un dictionnaire, il arrive qu'il n'apparaisse pas dans la sous-fenêtre Résultats d'extraction après la nouvelle extraction ou qu'il apparaisse bien, mais pas exactement tel que vous l'avez déclaré. Bien que cet événement soit rare, il peut avoir lieu lorsqu'un mot ou une expression a déjà été extrait dans le cadre d'une expression plus longue. Pour éviter cela, appliquez l'option de mise en correspondance **Entier (pas de composés)** à ce terme dans le dictionnaire de types. Pour plus d'informations, voir «Ajout de termes», à la page 190.

Chapitre 9. Catégorisation des données textuelles

La vue Catégories et concepts permet de créer des **catégories** qui représentent essentiellement des concepts de niveau supérieur ou des rubriques pour capturer les principales idées, les connaissances et les attitudes exprimées dans le texte.

En ce qui concerne l'édition de IBM SPSS Modeler Text Analytics 14, les catégories peuvent également avoir une structure hiérarchique, ce qui signifie qu'elles peuvent contenir des sous-catégories et que ces sous-catégories peuvent elles-même contenir des sous-catégories, et ainsi de suite. Vous pouvez importer des structures de catégories prédéfinies, nommées auparavant plans de codage, avec des catégories hiérarchiques ou bien créer ces catégories hiérarchiques à l'intérieur du produit.

En effet, les catégories hiérarchiques vous permettent de créer une structure en arborescence contenant une ou plusieurs sous-catégories afin d'y regrouper plus clairement les éléments par groupe de concepts ou de rubriques. Un exemple simple peut être lié aux activités de loisir, par exemple, poser une question telle que *Quelle activité aimeriez-vous faire si vous aviez plus de temps ?* Vous pouvez avoir des catégories principales telles que *sports*, *travaux manuels*, *pêche*, etc., puis des niveaux inférieurs, par exemple, sous *sports*, des sous-catégories *jeux de balle*, *sports nautiques*, etc.

Les catégories sont constituées d'un ensemble de descripteurs, tels que des *concepts*, des *types*, des *motifs* et des *règles de catégorie*. Ensemble, ces descripteurs permettent d'identifier si un document ou un enregistrement appartient ou non à une catégorie. Le texte d'un document ou d'un enregistrement peut être analysé afin de déterminer s'il correspond à un descripteur. Si une correspondance est détectée, le document/l'enregistrement est attribué à cette catégorie. Ce processus est appelé **catégorisation**.

Vous pouvez créer des catégories, les utiliser et les explorer à l'aide des données affichées dans les quatre sous-fenêtres de la vue Catégories et concepts, que vous pouvez masquer ou afficher en sélectionnant leur nom dans le menu Vue.

- **Sous-fenêtre Catégories.** Dans cette sous-fenêtre, créez et gérez vos catégories. Pour plus d'informations, voir «La sous-fenêtre Catégories», à la page 96 for more information.
- **Sous-fenêtre Résultats d'extraction.** Dans cette sous-fenêtre, explorez et utilisez les concepts et les types extraits. Pour plus d'informations, voir la rubrique «Résultats d'extraction : concepts et types», à la page 81.
- **Sous-fenêtre Visualisation.** Dans cette sous-fenêtre, explorez visuellement vos catégories et analysez la manière dont elles interagissent. Pour plus d'informations, voir la rubrique «Graphiques et diagrammes de catégorie», à la page 155.
- **Sous-fenêtre Données.** Dans cette sous-fenêtre, explorez et passez en revue le texte contenu dans les documents et les enregistrements qui correspondent aux sélections effectuées. Pour plus d'informations, voir la rubrique «La sous-fenêtre Données», à la page 104.

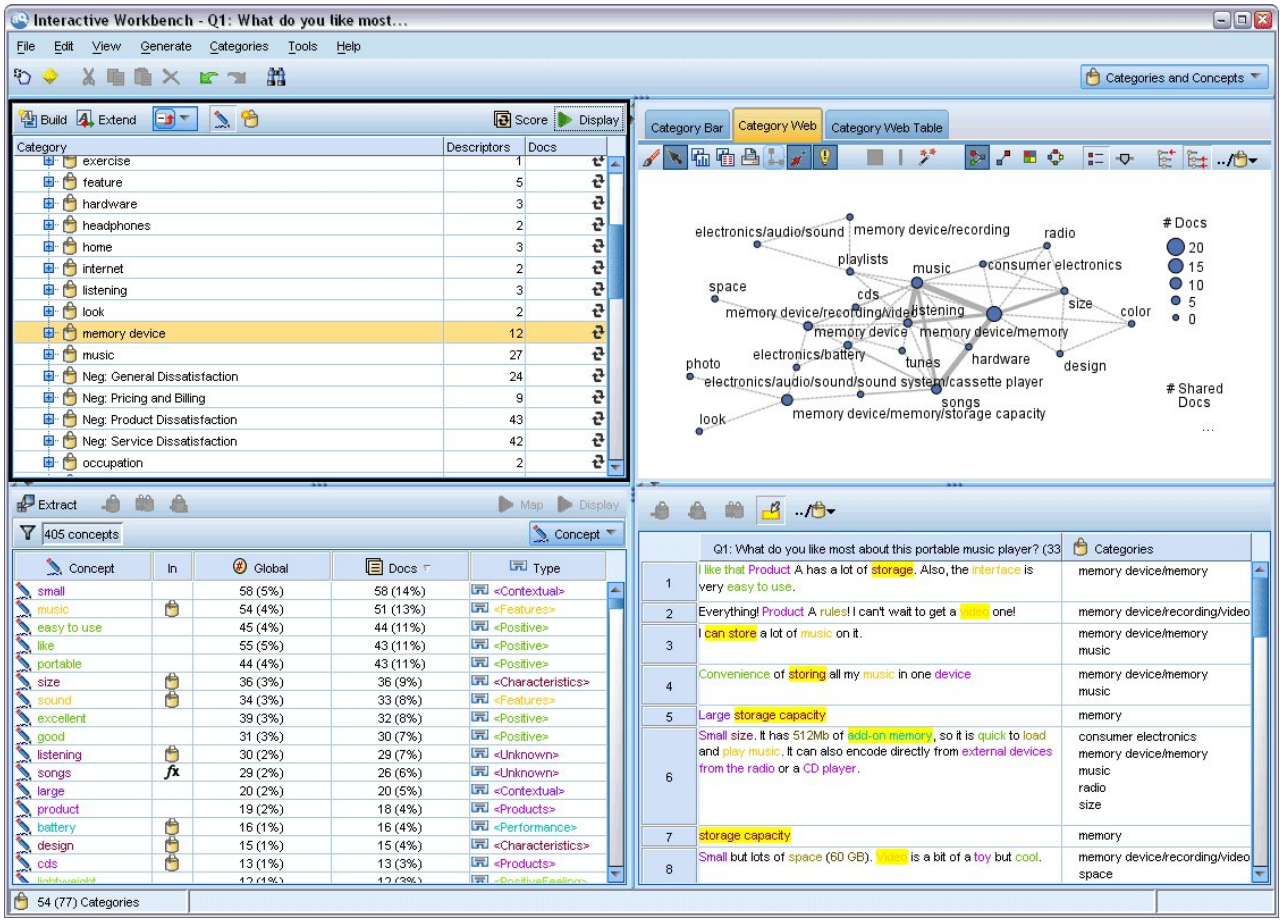


Figure 29. Vue Catégories et concepts

Vous pouvez commencer avec un ensemble de catégories provenant d'un pack d'analyse de texte (TAP), ou importer depuis un fichier de catégories prédéfinies, vous pouvez également avoir besoin de créer votre propre ensemble. Les catégories peuvent être créées automatiquement à l'aide des techniques fiables et automatisées qui utilisent les résultats d'extraction (concepts, types et motifs) pour générer des catégories et leurs descripteurs. Les catégories peuvent aussi être créées manuellement en utilisant des informations supplémentaires que vous pouvez avoir au sujet des données. Toutefois, vous pouvez uniquement créer des catégories manuellement ou les affiner via le plan de travail interactif. Pour plus d'informations, voir «Noeud Text Mining : onglet Modèle», à la page 24. Vous pouvez créer des définitions de catégorie manuellement en faisant glisser et en déposant les résultats d'extraction dans les catégories. Vous pouvez enrichir ces catégories ou une catégorie vide par l'ajout de règles de catégorie à une catégorie, par l'utilisation de vos propres catégories prédéfinies ou par une combinaison.

Chaque technique convient à un type de données et à certaines situations. Cependant, il est souvent judicieux de combiner plusieurs méthodes dans la même analyse afin de capturer le maximum de documents ou d'enregistrements. En cours de catégorisation, vous pouvez également envisager d'apporter d'autres modifications aux ressources linguistiques.

La sous-fenêtre Catégories

La sous-fenêtre Catégories est la zone dans laquelle vous pouvez créer et gérer vos catégories. Elle est située dans l'angle supérieur gauche de la vue Catégories et concepts. Après avoir extrait les concepts et les types de vos données textuelles, vous pouvez générer des catégories automatiquement à l'aide de technique telles que l'inclusion de concept, la co-occurrence, etc. ou vous pouvez les créer manuellement. Pour plus d'informations, voir «Génération de catégories», à la page 106.

Chaque fois qu'une catégorie est créée ou mise à jour, les documents ou enregistrements peuvent être évalués en cliquant sur le bouton **Score** pour déterminer si un texte correspond à un descripteur dans une catégorie donnée. Si une correspondance est détectée, le document ou l'enregistrement est attribué à cette catégorie. Le résultat final est que la plupart, voire l'intégralité, des documents ou des enregistrements est affectée à des catégories en fonction des descripteurs des catégories.

Remarque : Si le nombre de catégories est plus important que ce que peut contenir la sous-fenêtre visible, vous pouvez utiliser les commandes de la partie inférieure de la sous-fenêtre pour vous déplacer vers l'avant ou vers l'arrière dans les catégories ou entrer un numéro de page pour y accéder.

Table d'arborescence des catégories

La table d'arborescence de cette sous-fenêtre présente l'ensemble de catégories, des sous-catégories et des descripteurs. L'arborescence possède également plusieurs colonnes présentant des informations pour chaque élément de l'arborescence. Les colonnes pouvant être affichées sont les suivantes :

- **Code** Répertorie la valeur de code de chaque catégorie. Par défaut, cette colonne est masquée. Vous pouvez l'afficher à l'aide duu menu **Vue** > **Sous-fenêtre Catégories**.
- **Catégorie.** Contient l'arborescence des catégories avec le nom de la catégorie et des sous-catégories. De plus si vous cliquez sur l'icône de la barre d'outils des descripteurs, l'ensemble des descripteurs sera aussi affiché.
- **Descripteurs.** Fournit le nombre de descripteurs qui la définissent. Ce nombre ne comprend pas le nombre de descripteurs dans les sous-catégories. Il n'est pas indiqué lorsque le nom de descripteur est indiqué dans la colonne **Catégories**. Vous pouvez afficher ou masquer les descripteurs eux-mêmes dans l'arborescence à l'aide du menu **Vue** > **Sous-fenêtre Catégories** > **Tous les descripteurs**.
- **Docs** Après le scoring, cette colonne affiche le nombre de documents ou d'enregistrements qui sont catégorisés dans une catégorie et toutes ses sous-catégories. Si 5 enregistrements correspondent à votre catégorie principale en fonction de ses descripteurs, et que 7 autres enregistrements correspondent à une sous-catégorie en fonction de ses descripteurs, le nombre total de documents pour la catégorie principale correspond à la somme des deux, soit 12 dans cet exemple. Toutefois, si un même enregistrement est en correspondance avec la catégorie principale et la sous-catégorie, le nombre total est 11.

S'il n'existe aucune catégorie, le tableau contient tout de même deux lignes. La ligne supérieure, appelée **Tous les documents**, représente le nombre total de documents ou d'enregistrements. Une seconde ligne, **Sans catégorie**, représente le nombre de documents/d'enregistrements devant être catégorisés.

Pour chaque catégorie présente dans la sous-fenêtre, une icône représentant un petit seau jaune précède le nom de la catégorie. Lorsque vous double-cliquez sur le nom d'une catégorie , ou choisissez **Vue > Définitions de catégorie** dans les menus, la boîte de dialogue Définitions de catégorie s'ouvre et affiche tous ses éléments (appelés *descripteurs*, qui composent la définition de cette catégorie (concepts, types, règles). Pour plus d'informations, voir «A propos des catégories», à la page 103. Par défaut, la table d'arborescence de catégories n'affiche pas les descripteurs dans les catégories. Si vous souhaitez voir les descripteurs directement dans l'arborescence plutôt que dans la boîte de dialogue Définitions de catégories, cliquez sur le bouton bascule à l'aide de l'icône du stylo dans la barre d'outils. Lorsque le bouton bascule est sélectionné, vous pouvez développer votre arborescence pour afficher également les descripteurs.

Scoring des catégories

La colonne **Documents**. dans la table d'arborescence des catégories affiche le nombre de documents ou d'enregistrements qui sont catégorisés dans cette catégorie spécifique. Si les nombres sont obsolètes ou n'ont pas été calculés, une icône apparaît dans cette colonne. Vous pouvez cliquer sur **Score** dans la barre d'outil de la sous-fenêtre pour recalculer le nombre de documents. Gardez à l'esprit que le processus de scoring peut prendre un certain temps lorsque vous utilisez des ensembles de données volumineux.

Sélection des catégories dans l'arborescence

Lorsque vous effectuez des sélections dans l'arborescence, vous ne pouvez sélectionner que des catégories Frère, c'est-à-dire, si vous sélectionnez des catégories de niveau supérieur, vous ne pouvez pas également sélectionner une sous-catégorie. Ou si vous sélectionnez 2 sous-catégories d'une catégorie donnée, vous ne pouvez pas sélectionner en même temps une sous-catégorie d'une autre catégorie. La sélection d'une catégorie discontinue provoquera la perte de votre sélection précédente.

Affichage dans les sous-fenêtres Données et Visualisation

Lorsque vous sélectionnez une ligne dans le tableau, vous pouvez cliquer sur le bouton **Afficher** pour que les sous-fenêtres Visualisation et Données soient actualisées avec les informations correspondant à votre sélection. Si une sous-fenêtre n'est pas visible, le fait de cliquer sur **Afficher** l'ouvre.

Réglage de vos catégories

La catégorisation peut ne pas générer des résultats parfaits pour vos données lors de votre première tentative, et il se peut que vous souhaitiez supprimer des catégories ou les combiner avec d'autres catégories. Vous pouvez également vous apercevoir, en consultant les résultats de l'extraction, que certaines catégories n'ayant pas été créées vous seraient utiles. Dans ce cas, vous pouvez apporter des modifications manuelles aux résultats afin de les adapter à votre contexte. Pour plus d'informations, voir la rubrique «Edition et affinage des catégories», à la page 138.

Stratégies et méthodes de création de catégories

Si vous n'avez pas encore effectué d'extraction ou que vos résultats d'extraction ne sont pas à jour, l'utilisation de l'une des techniques de génération ou d'extension de catégorie vous invitera automatiquement à effectuer une extraction. Après avoir appliqué une technique, les concepts et les types qui ont été regroupés dans une catégorie restent disponibles et pourront être classés par le biais d'autres techniques. Cela signifie que vous pouvez voir un concept dans plusieurs catégories sauf si vous choisissez de ne pas les réutiliser.

Afin de vous aider à créer les catégories les plus pertinentes, veuillez examiner les points suivants :

- **Méthodes de création de catégories**
- **Stratégies de création de catégories**
- **Conseils pour la création de catégories**

Méthodes de création de catégories

Puisque chaque ensemble de données est unique, le nombre de méthodes de création de catégories et l'ordre dans lequel vous les appliquez peut varier. De plus, dans la mesure où vos objectifs d'exploration de texte peuvent être différents d'un ensemble de données à un autre, vous pouvez tester les différentes méthodes afin de déterminer celle qui génère les meilleurs résultats pour les données textuelles en question. Aucune technique automatique ne permet de catégoriser parfaitement vos données ; nous vous recommandons de rechercher et d'appliquer la ou les techniques qui sont les mieux adaptées à vos données.

Outre l'utilisation de packs d'analyse de texte (TAP,*.*tap*) avec des ensembles de catégories prédéfinis, vous pouvez également catégoriser vos réponses à l'aide de toute combinaison des méthodes suivantes :

- **Techniques de création automatique.** Plusieurs options de catégorie basées sur la linguistique et la fréquence sont disponibles pour créer automatiquement des catégories. Pour plus d'informations, voir «Génération de catégories», à la page 106.
- **Techniques d'extension automatique.** Plusieurs techniques linguistiques sont disponibles pour étendre les catégories existantes en ajoutant et en améliorant les descripteurs afin qu'ils capturent davantage d'enregistrements. Pour plus d'informations, voir «Extension de catégories», à la page 116.

- **Techniques manuelles.** Il existe plusieurs méthodes manuelles, telles que le glisser-déposer. Pour plus d'informations, voir «Création manuelle de catégories», à la page 120.

Stratégies de création de catégories

La liste de stratégies suivante n'est en aucun cas exhaustive mais elle peut vous donner une idée de l'approche de la création de vos catégories.

- Lorsque vous définissez le noeud Text Mining, sélectionnez un ensemble de catégories à partir d'un pack d'analyse de texte (TAP), de manière à commencer les analyses avec des catégories préconfigurées. Il est possible que ces catégories suffisent à l'analyse de votre texte depuis le début. Cependant, si vous souhaitez ajouter d'autres catégories, vous pouvez modifier les paramètres de création de catégories (**Catégories > Configurer les paramètres**). Ouvrez la boîte de dialogue **Paramètres avancés : Linguistique**, choisissez l'option d'entrée de la catégorie**Résultats d'extraction non utilisés** et générez les catégories supplémentaires.
- Lorsque vous définissez le noeud, sélectionnez un ensemble de catégories à partir d'un TAPdans la vue Catégories et concepts du plan de travail interactif. Ensuite, faites glisser et déposer les concepts ou motifs inutilisés dans les catégories qui vous semblent appropriées. Ensuite, étendez les catégories existantes que vous venez de modifier (**Catégories > Etendre les catégories**) pour obtenir davantage de descripteurs associés aux descripteurs de catégorie existants.
- Créez des catégories automatiquement à l'aide des paramètres linguistiques avancés (**Catégories > Créer des catégories**). Ensuite, affinez manuellement les catégories en supprimant des descripteurs, des catégories ou en fusionnant des catégories similaires jusqu'à ce que vous soyez satisfait des catégories résultantes. De plus, si vous créez des catégories **sans** utiliser l'option **Généraliser avec des caractères génériques lorsque cela est possible**, vous pouvez également essayer de simplifier automatiquement les catégories à l'aide de la fonctionnalité Etendre les catégories, en activant l'option **Généraliser**.
- Importez un fichier de catégories prédéfinies avec des noms de catégories très descriptifs et/ou des annotations. Si vous avez, dans un premier temps, procédé à l'importation **sans** avoir sélectionné l'option qui permet d'importer ou de générer des descripteurs à partir de noms de catégories, vous pouvez le faire ultérieurement dans la boîte de dialogue Etendre les catégories, en sélectionnant l'option **Etendre les catégories vides avec des descripteurs générés à partir du nom de la catégorie**. Puis, étendez ces catégories une deuxième fois, mais utilisez alors les techniques de regroupement.
- Créez manuellement un premier ensemble de catégories en triant les concepts ou motifs de concept par fréquence puis en faisant glisser et déposer les plus intéressants dans la sous-fenêtre Catégories. Une fois que vous avez cet ensemble initial de catégories, utilisez la fonction Etendre (**Catégories > Etendre les catégories**) pour développer et affiner toutes les catégories sélectionnées afin qu'elles comprennent les autres descripteurs associés et correspondent ainsi à plus d'enregistrements.

Après avoir appliqué ces techniques, nous vous recommandons de passer en revue les catégories résultantes et d'utiliser des techniques manuelles pour effectuer des changements mineurs, supprimer toute mauvaise réaffectation ou ajouter les enregistrements ou mots ayant été oubliés. En outre, puisque l'utilisation de techniques différentes peut être à l'origine de catégories redondantes, vous avez la possibilité de fusionner ou de supprimer des catégories si nécessaire. Pour plus d'informations, voir la rubrique «Edition et affinage des catégories», à la page 138.

Conseils pour la création de catégories

Afin de vous aider à créer de meilleures catégories, vous pouvez examiner certains conseils qui peuvent vous aider à prendre des décisions à propos de votre approche.

Conseils concernant le rapport Catégorie-Document

Il est rare que les catégories auxquelles les documents et les enregistrements sont affectés s'excluent mutuellement dans l'analyse de texte qualitative, et ce, pour au moins deux raisons :

- En premier lieu, il est généralement admis que plus un document ou un enregistrement texte est long, plus les idées et opinions qu'il exprime sont diverses. Ainsi, un document ou un enregistrement a plus de chances de se voir attribuer plusieurs catégories.
- En second lieu, il existe souvent plusieurs manières de regrouper ou d'interpréter des documents ou des enregistrements texte qui ne sont pas logiquement distincts. Dans le cas d'une enquête comportant une question ouverte sur les opinions politiques de la personne interrogée, nous pourrions créer des catégories, telles que *libéral* et *conservateur*, ou *Républicain* et *Démocrate*, ainsi que des catégories plus spécifiques, telles que *libéral pour l'aspect social*, *conservateur pour l'aspect fiscal* et ainsi de suite. Il n'est pas nécessaire que ces catégories s'excluent mutuellement, ni qu'elles soient exhaustives.

Conseils sur le nombre de catégories à créer

La création d'une catégorie doit découler directement des données, et représenter un type d'information intéressante par rapport à celles-ci. En général, le nombre de catégories que vous pouvez créer n'est pas limité. En revanche, si vous créez trop de catégories, vous aurez des difficultés à les gérer. Deux principes sont applicables :

- **Fréquence de catégorie.** Pour qu'une catégorie soit utile, elle doit contenir une quantité minimale de documents ou d'enregistrements. Un ou deux documents peuvent contenir une information particulière, mais s'il s'agit d'un ou deux documents sur mille, cette information n'est pas assez fréquemment citée pour qu'elle soit utile et significative dans ce contexte.
- **Complexité.** Plus vous créez de catégories, plus vous avez d'informations à examiner et à résumer après analyse. Toutefois, un trop grand nombre de catégories risque de créer une trop grande complexité sans apporter de détails utiles.

Malheureusement, aucune règle ne détermine à partir de quel nombre les catégories sont jugées trop nombreuses, ni ne définit le nombre minimal d'enregistrements par catégorie. C'est à vous d'effectuer ces déterminations, en fonction des exigences de votre situation.

Nous pouvons toutefois vous fournir nos conseils et vous indiquer où commencer. Bien que le nombre de catégories ne doive pas être trop élevé, il est préférable d'en avoir trop que pas assez lors des premières étapes de l'analyse. Il est plus simple de regrouper des catégories relativement similaires que de diviser des observations en catégories nouvelles. La stratégie consistant à travailler d'un plus grand nombre à un moins grand nombre de catégories est donc généralement la meilleure. Etant donné la nature itérative de l'exploration de texte et la facilité avec laquelle il peut être exécuté avec ce logiciel, il est acceptable de créer plus de catégories au départ.

Choix des meilleurs descripteurs

Les informations suivantes vous fourniront quelques conseils afin de choisir ou de créer les meilleurs descripteurs (concepts, types, motifs TLA et règles de catégorie) pour vos catégories. Les descripteurs sont les blocs de construction des catégories. Lorsqu'une partie ou l'ensemble du texte d'un document ou d'un enregistrement correspond à un descripteur, le document ou l'enregistrement est mis en correspondance avec la catégorie.

Un descripteur n'est mis en correspondance avec des documents ou des enregistrements que s'il contient ou correspond à un concept ou à un motif extrait. Par conséquent, utilisez les concepts, les types, les motifs et les règles de catégorie de la manière indiquée dans les paragraphes suivants.

Les concepts représentant un ensemble de termes sous-jacents (en plus de se représenter eux-mêmes) qui peuvent comprendre un ensemble vaste de termes allant des formes singulier/pluriel, aux synonymes, ou aux variations orthographiques, seul le concept lui-même doit être utilisé en tant que descripteur ou comme partie d'un descripteur. Pour en savoir plus sur les termes sous-jacents de chaque concept donné, cliquez sur le nom du concept dans la sous-fenêtre Résultats d'extraction de la vue Catégories et concepts. Lorsque vous placez la souris sur le nom d'un concept, une info-bulle apparaît et affiche tous les termes sous-jacents trouvés dans votre texte lors de la dernière extraction. Les concepts ne possèdent pas tous des termes sous-jacents. Par exemple, si `voiture` et `véhicule` sont des synonymes mais que

voiture est extrait comme concept et véhicule comme terme sous-jacent, alors vous devez utiliser voiture comme descripteur, puisqu'il permettra de mettre en correspondance le document ou les enregistrements qui contiennent également le terme véhicule.

Concepts et types utilisés comme descripteurs

Vous devez utiliser un concept comme descripteur lorsque vous souhaitez trouver tous les documents ou enregistrements contenant ce concept (ou ses termes sous-jacents). Dans ce cas, l'utilisation d'une règle de catégorie plus complexe n'est pas nécessaire puisque le nom du concept exact est suffisant. Souvenez-vous que lorsque vous utilisez des ressources qui extraient des opinions, il se peut que les concepts changent lors de l'extraction des motifs TLA, afin de capturer le sens exact de la phrase (reportez-vous à l'exemple de la section qui suit concernant les motifs TLA).

Par exemple, un résultat de sondage indiquant les fruits préférés de chaque personne, par exemple "*les pommes et les ananas sont les meilleurs*", peut entraîner l'extraction de pommes et de ananas. En ajoutant le concept pommes en tant que descripteur à votre catégorie, toutes les réponses contenant le concept pommes (ou ses termes sous-jacents) sont mises en correspondance avec cette catégorie.

Toutefois, si vous voulez juste connaître les réponses qui mentionnent le terme *pommes* de quelque manière que ce soit, vous pouvez écrire une règle de catégorie, comme par exemple * pommes *, pour capturer toutes les réponses contenant des concepts tels que pommes, jus de pommes, ou tarte aux pommes.

Vous pouvez également capturer tous les documents ou enregistrements qui contiennent des concepts de même type en utilisant un type directement en tant que descripteur comme par exemple <Fruit>. Remarque : vous ne pouvez pas utiliser la fonction * avec les types.

Pour plus d'informations, voir la rubrique «Résultats d'extraction : concepts et types», à la page 81.

Motifs d'analyse des liens du texte (TLA) utilisés comme descripteurs

Utilisez un résultat de motif TLA comme descripteur si vous souhaitez capturer des idées plus nuancées ou plus fines. Lorsque le texte est analysé lors de l'extraction TLA, le texte (document ou enregistrement) est traité phrase par phrase ou clause par clause, plutôt que dans son ensemble. En prenant en considération l'ensemble des parties d'une phrase, l'analyse des liens du texte peut identifier les opinions, les relations entre deux éléments, ou une négation, par exemple, et comprendre ainsi le sens exact de la phrase. Vous pouvez utiliser des motifs de concepts ou des motifs de type en tant que descripteurs. Pour plus d'informations, voir la rubrique «Motifs de type et Motifs de concept», à la page 151.

Prenons par exemple le texte "*la pièce n'est pas propre*", les concepts pièce et propre seront extraits. Toutefois, si l'extraction TLA avait été activée dans les paramètres de l'extraction, elle aurait pu détecter que propre est utilisé de manière négative et correspond en fait au concept pas propre, qui est aussi synonyme du concept sale. Vous pouvez voir dans cet exemple que l'utilisation du concept propre seul, en tant que descripteur sera mis en correspondance avec ce texte, mais capturera aussi d'autres documents ou enregistrements qui mentionnent la propreté. Par conséquent, il peut être plus judicieux d'utiliser le motif de concept TLA sale comme concept de sortie, car il sera mis en correspondance avec ce texte et sera plus approprié en tant que descripteur.

Règles métier de catégories utilisées comme descripteurs

Les règles de catégorie sont des instructions qui classifient automatiquement les documents ou les enregistrements en une catégorie basée sur une expression logique à l'aide de concepts, de types et de motifs extraits ou d'opérateurs booléens. Par exemple, vous pouvez créer une expression qui signifie *inclure tous les enregistrements contenant le concept extrait* ambassade *mais ne pas inclure* argentine *dans cette catégorie*.

Vous pouvez écrire et utiliser des règles de catégories en tant que descripteurs dans vos catégories, afin d'exprimer des idées différentes, à l'aide des booléens `&`, `|` et `!()`. Pour plus d'information sur la syntaxe de ces règles et sur la manière de les écrire et de les éditer, voir «Utilisation des règles de catégorie», à la page 121.

- Utilisez une règle de catégorie avec l'opérateur booléen `&` (AND) pour trouver des documents ou des enregistrements dans lesquels se trouvent 2 concepts ou plus. Les concepts connectés par des opérateurs `&` ne doivent pas obligatoirement se trouver dans la même phrase, ils peuvent apparaître n'importe où dans le même document ou enregistrement pour être mis en correspondance avec la catégorie. Par exemple, si vous créez la règle de catégorie suivante : `nourriture & bon marché` comme descripteur, elle correspondra à un enregistrement comportant le texte "*la nourriture était très chère, mais les chambres étaient bon marché*" bien que le terme `nourriture` ne soit pas le nom auquel s'applique l'adjectif `bon marché`, et ce car le texte contient à la fois les termes `nourriture` et `bon marché`.
- Utilisez une règle de catégorie avec l'opérateur booléen `!()` (NOT) en tant que descripteur pour trouver des documents ou des enregistrements dans lesquels se trouvent certains termes mais pas d'autres. Ceci peut permettre de ne pas regrouper des informations qui semblent être reliées en fonction des mots mais pas en fonction du contexte. Par exemple, si vous créez la règle de catégorie `<Organization> & !(ibm)` comme descripteur, elle renvoie le texte *la société SPSS Inc. a été créée en 1967* et non pas le texte *la société de logiciels a été acquise par IBM.*.
- Utilisez une règle de catégorie avec l'opérateur booléen `|` (OR) en tant que descripteur pour trouver des documents ou des enregistrements contenant un ou plusieurs concepts ou types. Par exemple, si vous créez la règle de catégorie suivante : `(personnel|équipe|employés|collaborateurs) & mauvais` comme descripteur, elle correspondra à tout document ou enregistrement dans lequel l'un de ces noms de trouvent en même temps que le concept de `mauvais`.
- Utilisez les types dans les règles de catégorie afin de les rendre plus génériques et plus déployables. Par exemple, si vous travaillez sur des données hôtelières, il se peut que vous soyez intéressé par l'opinion des clients sur le personnel de l'hôtel. Des termes associés peuvent inclure des mots comme réceptionniste, serveur, serveuse, réception, etc. Dans ce cas, vous pouvez créer un nouveau type nommé `<PersonnelHôtel>` et lui ajouter tous les termes précédents. Alors qu'il est possible de créer une règle de catégorie pour chaque type d'employés tels que`[* serveuse* & aimable]`, `[* réception* & amical]`, `[* réceptionniste * & accueillant]`, vous pouvez également créer une règle de catégorie unique, plus générale à l'aide du type `<PersonnelHôtel>` pour capturer toutes les réponses contenant des opinions positives sur le personnel de l'hôtel sous la forme de `[<PersonnelHôtel> & <Positif>]`.

Remarque : vous pouvez utiliser à la fois `+` et `&` dans les règles de catégories lorsque vous y incluez des motifs TLA. Pour plus d'informations, voir «Utilisation des motifs TLA dans les règles de catégorie», à la page 123.

Exemple des différences de mise en correspondance lors de l'utilisation des concepts, des motifs TLA ou des règles de catégories comme descripteurs.

L'exemple suivant montre la manière dont l'utilisation des concepts, des règles de catégorie ou des motifs TLA en tant que descripteurs affecte la classification des documents ou des enregistrements en catégories. Considérons les 5 enregistrements suivants :

- A : *"équipe de service parfaite, nourriture excellente, chambres propres et confortables."*
- B : *"équipe de service lamentable, mais chambres propres."*
- C : *"Chambres propres et confortables."*
- D : *"Ma chambre n'était pas propre."*
- E : *"Propre."*

Dans la mesure où les enregistrements contiennent le mot *propre* et que vous souhaitez capturer cette information, vous pouvez créer un des descripteurs indiqué dans le tableau ci-dessous. En fonction de la nature des informations que vous essayez de capturer, vous pouvez utiliser un type de descripteur et voir les résultats produits.

Tableau 17. Mise en correspondance des exemples d'enregistrements avec les descripteurs.

Descripteur	A	B	C	D	E	Explication
`propre`	*concordance*	*concordance*	*concordance*	*concordance*	*concordance*	Le descripteur est un concept extrait. Chaque enregistrement contient le concept `propre`, y compris l'enregistrement D pour lequel il n'y a pas de règle TLA indiquant que "*pas propre*" `sale`.
`propre + .`	-	-	-	-	*concordance*	Le descripteur est un motif TLA qui représente `propre` lui-même. Ne correspond qu'à l'enregistrement dans lequel `propre` a été extrait sans concept associé lors de l'extraction TLA.
`[propre]`	*concordance*	*concordance*	*concordance*	-	*concordance*	Le descripteur est une règle de catégorie qui cherche une règle TLA contenant `propre` seul ou associé à autre chose. Correspond à tous les enregistrements dans lesquels une sortie TLA contenant `propre` a été trouvée, que `propre` soit lié à un autre concept, comme par exemple `chambre`, ou non et indépendamment de sa position dans la phrase.

A propos des catégories

Les **catégories** font référence à un groupe de concepts, d'opinions ou d'attitudes étroitement associés. Pour être utile, une catégorie doit également être décrite par un libellé ou une phrase courte évoquant l'essentiel de sa signification.

Par exemple, si vous analysez des réponses de consommateurs à une enquête sur une nouvelle lessive, vous pouvez créer une catégorie libellée *odeur* qui contient toutes les réponses décrivant l'odeur de ce produit. Mais cette catégorie ne fera pas la différence entre ceux qui ont aimé l'odeur et ceux qui ne l'ont pas aimée. IBM SPSS Modeler Text Analytics permettant d'extraire des opinions lorsque les ressources appropriées sont utilisées, vous pouvez créer deux autres catégories pour identifier les consommateurs qui ont *aimé l'odeur* et ceux qui *n'ont pas aimé l'odeur*.

Vous pouvez créer et travailler avec vos propres catégories dans la sous-fenêtre Catégories, dans la partie supérieure gauche de la fenêtre Vue Catégories et concepts . Chaque catégorie est définie par un ou plusieurs descripteurs. Les **descripteurs** sont des concepts, des types, des motifs et des règles de catégorie, utilisés pour définir une catégorie.

Si vous souhaitez voir les descripteurs qui forment une catégorie donnée, vous pouvez cliquer sur l'icône du crayon dans la barre d'outils de la sous-fenêtre Catégories puis développer l'arborescence pour afficher les descripteurs. Vous pouvez également sélectionner la catégorie et ouvrir la boîte de dialogue Définitions de catégorie (**Vue > Définitions de catégorie**).

Lorsque vous générez des catégories automatiquement en utilisant des méthodes telles que l'inclusion de concept, les concepts et les types sont utilisés en tant que descripteurs pour la création des catégories. Si vous extrayez des motifs TLA, v ous pouvez également ajouter tout ou partie de ces motifs en tant que descripteurs de catégorie. Pour plus d'informations, voir Chapitre 11, «Exploration de l'analyse des liens du texte», à la page 149. Si vous créez des clusters, vous pouvez ajouter les concepts d'un cluster à des

catégories nouvelles ou existantes. Enfin, vous pouvez créer manuellement des règles de catégorie à utiliser en tant que descripteurs dans vos catégories. Pour plus d'informations, voir «Utilisation des règles de catégorie», à la page 121.

Propriétés de la catégorie

En plus des descripteurs, les catégories ont également des propriétés que vous pouvez éditer afin de renommer les catégories, d'ajouter un libellé ou une annotation.

Les propriétés suivantes sont disponibles :

- **Nom.** Ce nom apparaît dans l'arborescence par défaut. Lorsqu'une catégorie est créée à l'aide d'une technique automatique, un nom lui est attribué automatiquement.
- **Libellé.** L'utilisation de libellés permet de créer des descriptions de catégorie plus significatives en vue de les utiliser dans d'autres produits, ou dans d'autres tableaux ou graphiques. Si vous sélectionnez l'option permettant d'afficher le libellé, celui-ci est alors utilisé dans l'interface pour désigner la catégorie.
- **Code.** Le numéro de code correspond à la valeur de code de cette catégorie. .
- **Annotation.** Vous pouvez ajouter une description courte pour chaque catégorie dans ce champ. Lorsqu'une catégorie est générée par la boîte de dialogue Créer des catégories, une note est ajoutée automatiquement à cette annotation. Vous pouvez aussi ajouter un échantillon de texte à une annotation directement dans la sous-fenêtre Données en sélectionnant le texte et en sélectionnant **Catégories > Ajouter à l'annotation** dans le menu.

La sous-fenêtre Données

Lorsque vous créez des catégories, vous pouvez parfois souhaiter examiner certaines des données textuelles utilisées. Par exemple, si vous créez une catégorie dans laquelle 640 documents sont catégorisés, vous pouvez souhaiter consulter une partie ou l'intégralité de ces documents afin de découvrir le texte qui était en réalité rédigé. Vous pouvez consulter les enregistrements ou les documents dans la sous-fenêtre Données, située dans l'angle inférieur droit. S'il n'apparaît pas par défaut, sélectionnez **Vue> Sous-fenêtres > Données** dans les menus.

La sous-fenêtre de données affiche une ligne par document ou enregistrement correspondant à la sélection dans la sous-fenêtre Catégories, dans la sous-fenêtre Résultats d'extraction ou dans la boîte de dialogue Définitions de catégorie jusqu'à une certaine limite d'affichage. Par défaut, le nombre de document oud'enregistrements affichés dans la sous-fenêtre de données est limitée pour vous permettre de consulter vos données plus rapidement. Cependant, vous pouvez modifier cette limite dans la boîte de dialogue Options. Si vous traitez un nombre important de fichiers, la vitesse d'affichage peut être améliorée en désactivant l'option d'affichage des catégories. Pour plus d'informations, voir «Options : onglet Session», à la page 76.

Remarque : Si le nombre d'enregistrements est plus important que ce que peut contenir la sous-fenêtre visible, vous pouvez utiliser les commandes de la partie inférieure de la sous-fenêtre pour vous déplacer vers l'avant ou vers l'arrière dans les enregistrements ou entrer un numéro de page pour y accéder.

Affichage et actualisation de la sous-fenêtre Données

L'affichage de la sous-fenêtre Données n'est pas automatiquement actualisée car en présence d'ensembles de données volumineux, l'opération prendrait trop de temps. Par conséquent, chaque fois que vous effectuez une sélection dans une autre sous-fenêtre de cette vue ou dans la boîte de dialogue Définitions de catégorie, cliquez sur **Afficher** pour actualiser le contenu de la sous-fenêtre Données.

Documents texte ou enregistrements

Si vos données textuelles sont sous la forme d'enregistrements et que le texte est relativement bref, le champ de texte de la sous-fenêtre Données affiche les informations dans leur intégralité. Cependant, si vous utilisez des enregistrements et de grands ensembles de données, la colonne du champ de texte affiche une petite partie du texte et ouvre une sous-fenêtre Aperçu du texte à droite qui permet de consulter une plus grande partie du texte de l'enregistrement sélectionné dans la table, voire son intégralité. Si vos données textuelles se présentent sous la forme de documents, la sous-fenêtre Données affiche le nom de fichier du document. Lorsque vous sélectionnez un document, la sous-fenêtre Aperçu du texte s'ouvre et affiche le texte du document sélectionné.

Couleurs et mise en évidence

Chaque fois que vous affichez des données, des concepts et des descripteurs trouvés dans ces documents ou enregistrements, ils apparaissent en couleur pour vous permettre de les identifier facilement dans le texte. Le code couleur correspond aux types auxquels les concepts appartiennent. Vous pouvez également pointer sur un élément faisant l'objet d'un codage couleur pour afficher le concept sous lequel il a été extrait et le type auquel il a été affecté. Tout texte n'ayant pas été extrait apparaît en noir. En règle générale, ces mots non extraits sont souvent des connecteurs (*et* ou *avec*), des pronoms (*me* ou *ils*), et des verbes (*être*, *avoir* ou *prendre*).

Colonnes de la sous-fenêtre Données

Alors que la colonne de champ de texte est toujours visible, il est possible d'afficher également d'autres colonnes. Pour afficher d'autres colonnes, cliquez sur **Affichage** > **Panneau Données** dans les menus, puis sélectionnez la colonne que vous souhaitez afficher dans le panneau de données. Les colonnes pouvant être affichées sont les suivantes :

- **"Nom du champ de texte" (#)/Documents** Ajoute une colonne pour les données textuelles à partir desquelles des concepts et des types ont été extraits. Si vos données sont contenues dans des documents, la colonne est appelée Documents, et seul le nom de fichier du document ou son chemin complet est visible. Pour examiner le texte de ces documents, vous devez consulter la sous-fenêtre Aperçu du texte. Le nombre de lignes de la sous-fenêtre Données est indiqué entre parenthèses après le nom de cette colonne. Il peut arriver que les documents ou les enregistrements ne soient pas tous affichés en raison d'une limite définie dans la boîte de dialogue Options pour optimiser la vitesse de chargement. Si la limite est atteinte, le nombre sera suivi de **- Max**. Pour plus d'informations, voir «Options : onglet Session», à la page 76.
- **Catégories** Répertorie chacune des catégories à laquelle appartient un enregistrement. Lorsque cette colonne est affichée, l'actualisation de la sous-fenêtre Données peut prendre plus de temps afin d'afficher les informations les plus récentes.
- **Rang de pertinence** Donne un rang pour chaque enregistrement dans une seule catégorie. Ce rang montre dans quelle mesure l'enregistrement correspond à la catégorie par rapport aux autres enregistrements dans cette catégorie. Sélectionnez une catégorie dans la sous-fenêtre Catégories (sous-fenêtre supérieure gauche) pour voir le rang. Pour plus d'informations, voir «Pertinence des catégories».
- **Nombre de catégories** Répertorie le nombre de catégories auxquelles appartient un enregistrement.

Pertinence des catégories

Pour vous aider à créer de meilleures catégories, vous pouvez examiner la pertinence des documents ou des enregistrements dans chaque catégorie ainsi que la pertinence de toutes les catégories auxquelles appartient un document ou unenregistrement.

Pertinence d'une catégorie pour un enregistrement

Quand un document ou un enregistrement s'affiche dans la sous-fenêtre Données, toutes les catégories auxquelles il appartient sont répertoriées dans la colonne Catégories. Quand un document ou un enregistrement appartient à plusieurs catégories, les catégories dans cette colonne s'affichent dans l'ordre

de la correspondance la plus pertinente à la moins pertinente. La première catégorie est considérée comme correspondant le mieux à ce document ou à cet enregistrement. Pour plus d'informations, voir la rubrique «La sous-fenêtre Données», à la page 104.

Pertinence d'un enregistrement pour une catégorie

Quand vous sélectionnez une catégorie, vous pouvez examiner la pertinence de chacun de ses enregistrements dans la colonne Rang de pertinence dans la sous-fenêtre Données. Ce rang de pertinence indique à quel point le document ou l'enregistrement correspond à la catégorie sélectionnée par rapport aux autres enregistrements de cette catégorie. Pour voir le rang des enregistrements pour une seule catégorie, sélectionnez celle-ci dans la sous-fenêtre Catégories (sous-fenêtre supérieure gauche) et le rang du document ou de l'enregistrement s'affiche dans la colonne. Cette colonne n'est pas visible par défaut mais vous pouvez choisir de l'afficher. Pour plus d'informations, voir la rubrique «La sous-fenêtre Données», à la page 104.

Plus le numéro de rang de l'enregistrement est petit, plus cet enregistrement est pertinent par rapport à la catégorie sélectionnée, la valeur 1 représentant la meilleure correspondance. Si plusieurs enregistrements ont la même pertinence, chacun est affiché avec le même rang suivi d'un signe égal (=) pour montrer qu'ils ont la même pertinence. Par exemple, vous pouvez avoir les rangs suivants 1=, 1=, 3, 4, etc., ce qui signifie qu'il existe deux enregistrements considérés de manière égale comme étant les meilleures correspondances pour cette catégorie.

Astuce : vous pouvez ajouter le texte de l'enregistrement le plus pertinent à l'annotation de catégorie afin de fournir une meilleure description de cette catégorie. Ajoutez le texte directement à partir de la sous-fenêtre Données en sélectionnant le texte et en choisissant **Catégories > Ajouter à l'annotation** dans le menu.

Génération de catégories

Bien que vous puissiez avoir des catégories d'un pack d'analyse de texte, vous pouvez aussi créer automatiquement des catégories à l'aide de diverses techniques linguistiques et de fréquence. Dans la boîte de dialogue Créer des paramètres de catégorie, vous pouvez appliquer les techniques linguistiques et de fréquence automatiques pour créer des catégories à partir de concepts ou de patrons de concept.

En général, les catégories peuvent être constituées de différents types de descripteurs (types, concepts, motifs TLA, règles de catégorie). Quand vous créez des catégories à l'aide des techniques de génération de catégorie automatiques, les catégories créées sont nommées selon un concept ou un motif de concept (en fonction de l'entrée que vous sélectionnez) et chacune d'elle contient un ensemble de descripteurs. Ces descripteurs peuvent avoir la forme de règles de catégorie ou de concepts et comprennent tous les concepts associés découverts par les techniques.

Après avoir généré les catégories, vous pouvez en apprendre beaucoup à leur sujet en les examinant dans la sous-fenêtre Catégories ou en les explorant à travers les graphiques et les diagrammes. Vous pouvez utiliser des techniques manuelles pour effectuer des changements mineurs, supprimer toute mauvaise réaffectation, ou ajouter les enregistrements ou mots ayant été oubliés. Après avoir appliqué une technique, les concepts, les types et les motifs qui ont été regroupés dans une catégorie peuvent encore être classés par le biais d'autres techniques. En outre, puisque l'utilisation de techniques différentes peut également être à l'origine de catégories redondantes ou inappropriées, vous avez la possibilité de fusionner ou de supprimer des catégories. Pour plus d'informations, voir la rubrique «Edition et affinage des catégories», à la page 138.

Important ! Dans les éditions précédentes, les règles de cooccurence et de synonymes étaient entre crochets. Dans l'édition actuelle, les crochets indiquent désormais un résultat de motifs d'analyse des liens du texte. Les règles de co-occurrence et de synonymes sont maintenant entourées de parenthèses, comme dans `(enceintes acoustiques|enceintes)`.

Pour créer des catégories

1. Dans les menus, sélectionnez **Catégories > Créer des catégories**. Un message apparaît, sauf si vous avez choisi de ne pas recevoir d'invite.
2. Choisissez si vous voulez créer maintenant ou éditer d'abord les paramètres.

- Cliquez sur **Créer Maintenant** pour commencer à générer des catégories à l'aide des paramètres actuels. Les paramètres sélectionnés par défaut sont souvent suffisants pour commencer le processus de catégorisation. Le processus de génération de catégories commence et une boîte de dialogue de progression apparaît.
- Cliquez sur **Editer** pour examiner et modifier les paramètres de création.

Remarque : Vous pouvez afficher jusqu'à 10 000 catégories. Un avertissement s'affiche lorsque ce nombre est atteint ou dépassé. Si cela se produit, modifiez les options Créer ou développer des catégories afin de réduire le nombre de catégories créées.

Entrées

Les catégories sont créées à partir de descripteurs dérivés de motifs de type ou de types. Dans le tableau, vous pouvez sélectionner les types ou patrons individuels à inclure dans le processus de création de catégories.

Motifs de type. Si vous sélectionnez les motifs de type, des catégories sont créées à partir de motifs plutôt qu'à partir de types et de concepts uniquement. De cette manière, tous les enregistrements ou documents contenant un motif de concept appartenant au motif de type sélectionné sont classés en catégories. Ainsi, si vous sélectionnez le motif de type `<Budget>` et `<Positive>` dans le tableau, des catégories telles que `coût & <Positive>` ou `taux & excellent` peuvent être créées.

Lorsque des motifs de type sont utilisés comme entrées pour la génération de catégories automatiques, les techniques identifient parfois plusieurs façons de former la structure des catégories. Techniquement, il n'existe pas une bonne façon de créer des catégories, mais une structure peut être plus adaptée à votre analyse qu'une autre. Pour aider à personnaliser la sortie dans ce cas, vous pouvez choisir un type préféré. Toutes les catégories de niveau supérieur créées proviendront d'un concept du type sélectionné ici (et pas d'un autre type). Chaque sous-catégorie contiendra un motif des liens du texte de ce type. Choisissez ce type dans la zone **Structurer les catégories par type de motif :** ; la table sera mise à jour et contiendra uniquement les motifs applicables contenant le type sélectionné. La plupart du temps, le type `<Unknown>` sera présélectionné. Ainsi, tous les motifs contenant le type `<Unknown>` seront sélectionnés. Le tableau affiche les types dans l'ordre décroissant en commençant par celui contenant le plus grand nombre d'enregistrements ou de documents (effectifs des **documents**).

Types. Si vous sélectionnez les types, les catégories seront créées à partir des concepts appartenant aux types sélectionnés. Ainsi, si vous sélectionnez le type `<Budget>` dans le tableau, des catégories telles que `coût` ou `prix` peuvent être créées car `coût` et `prix` sont des concepts attribués au type `<Budget>`.

Par défaut, seuls les types qui capturent le plus d'enregistrements ou de documents sont sélectionnés. Cette présélection vous permet de considérer rapidement les types les plus intéressants et d'éviter de créer des catégories sans intérêt. Le tableau affiche les types dans l'ordre décroissant en commençant par celui contenant le plus grand nombre d'enregistrements ou de documents (effectifs des **documents**). Les types de la bibliothèque `Opinions` sont désélectionnés par défaut dans le tableau des types.

Les entrées que vous choisissez ont une influence sur les catégories obtenues. Lorsque vous choisissez d'utiliser Types comme entrée, vous pouvez plus facilement voir les concepts associés. Par exemple, si vous créez des catégories en utilisant Types comme entrée, vous pourrez obtenir une catégorie `Fruit` avec des concepts comme `ananas`, `poire`, `agrumes`, `orange` etc. Si vous choisissez Motifs de type comme entrée et que vous sélectionnez le motif `<Unknown> + <Positive>`, par exemple, alors vous pouvez obtenir une catégorie `fruit + <Positive>` avec une ou deux sortes de fruits comme `fruit + savoureux` et `pomme + bonne`. Ce deuxième résultat n'affiche que 2 motifs de concept car les autres occurrences de fruit ne sont

pas nécessairement considérés comme positives. Et bien que cela puisse être suffisant pour vos donnés textuelles actuelles, dans les enquêtes longitudinales où différents ensembles de documents sont utilisés, il est conseillé d'ajouter manuellement d'autres descripteurs comme `agrume + positif` ou d'utiliser des types. La simple utilisation de types comme entrée vous permettra de trouver tous les fruits possibles.

Techniques

Puisque chaque ensemble de données est unique, le nombre de méthodes et l'ordre dans lequel vous les appliquez peut varier. Dans la mesure où vos objectifs d'exploration de texte peuvent être différents d'un ensemble de données à un autre, vous pouvez tester les différentes techniques afin de déterminer celle qui génère les meilleurs résultats pour les données textuelles en question.

Vous n'avez pas besoin d'être un expert de ces paramètres pour les utiliser. Par défaut, les paramètres les plus communs et moyens sont déjà sélectionnés. C'est pourquoi vous pouvez contourner la boîte de dialogue Paramètres avancés et créer directement vos catégories. De même, si vous effectuez des modifications ici, vous n'avez pas besoin de revenir à la boîte de dialogue Paramètres à chaque fois car les derniers paramètres sont toujours conservés.

Sélectionnez les techniques linguistiques ou de fréquence et cliquez sur le bouton Paramètres avancés pour afficher les paramètres des techniques sélectionnées. Aucune technique automatique ne permet de catégoriser parfaitement vos données ; nous vous recommandons de rechercher et d'appliquer la ou les techniques qui sont les mieux adaptées à vos données.Vous ne pouvez pas créer de catégories en utilisant en même temps les techniques de fréquence et linguistique.

- **Techniques linguistiques avancées**. Pour plus d'informations, voir «Paramètres linguistiques avancés».
- **Techniques de fréquence avancées**. Pour plus d'informations, voir «Paramètres de fréquence avancés», à la page 115.

Paramètres linguistiques avancés

Quand vous créez des catégories, vous pouvez choisir parmi diverses techniques de génération de catégorie linguistiques avancées, dont la *dérivation des racines de concept*, l'*inclusion de concepts*, les *réseaux sémantiques* (uniquement en anglais) et les *règles de co-occurrence*. Ces techniques peuvent être utilisées individuellement ou conjointement pour créer des catégories.

Notez que puisque chaque ensemble de données est unique, le nombre de méthodes et l'ordre dans lequel vous les appliquez peut varier. Dans la mesure où vos objectifs d'exploration de texte peuvent être différents d'un ensemble de données à un autre, vous pouvez tester les différentes techniques afin de déterminer celle qui génère les meilleurs résultats pour les données textuelles en question. Aucune technique automatique ne permet de catégoriser parfaitement vos données ; nous vous recommandons de rechercher et d'appliquer la ou les techniques qui sont les mieux adaptées à vos données.

Les zones et champs suivants sont disponibles dans la boîte de dialogue Paramètres avancés :
Linguistique :

Fichiers d'entrée et de sortie

Entrée de la catégorie Sélectionnez à partir de quoi les catégories seront créées :

- **Résultats d'extraction non utilisés**. Cette option permet aux catégories d'être créées à partir des résultats d'extraction qui ne sont pas encore utilisés dans une catégorie existante. Ceci réduit la tendance des enregistrements à correspondre à plusieurs catégories et limite le nombre de catégories produites.
- **Tous les résultats d'extraction**. Cette option permet aux catégories d'êtres créées à l'aide de tous les résultats d'extraction. Ceci est particulièrement utile quand aucune ou peu de catégories existent déjà.

Sortie de la catégorie Sélectionnez la structure générale des catégories qui seront créées :

- **Hiérarchique avec des sous-catégories**. Cette option active la création de sous-catégories et de sous-sous-catégories. Vous pouvez définir la profondeur de vos catégories en choisissant le nombre de niveaux maximum (champ **Nombre maximum de niveaux créés**) pouvant être créés. Si vous choisissez 3, les catégories peuvent contenir des sous-catégories et ces sous-catégories peuvent également contenir des sous-catégories.
- **Catégories plates (un seul niveau)**. Cette option n'active qu'un seul niveau de catégories à créer, ce qui signifie qu'aucune sous-catégorie ne sera créée.

Techniques de regroupement

Chaque technique disponible convient à certains types de données et de situation. Cependant, il est souvent judicieux de combiner des techniques dans la même analyse afin de capturer le maximum de documents ou d'enregistrements. Aussi est-il possible de voir un concept figurer dans plusieurs catégories ou de rencontrer des catégories redondantes.

Inclusion de concept. Cette technique crée des catégories en regroupant les concepts multitermes (mots composés) selon qu'ils contiennent ou non des mots qui sont des sous-ensembles ou des super-ensembles d'un mot dans l'autre. Par exemple, le concept `siège` serait regroupé avec `siège de sécurité`, `siège couchette` et `commande de siège éjectable`. Pour plus d'informations, voir «Inclusion de concepts», à la page 112.

Réseau sémantique. Cette technique commence par identifier les sens possibles de chaque concept à partir de son index complet de relations existant entre les mots, puis crée des catégories en regroupant les concepts associés. Cette technique est plus performante lorsque les concepts sont connus dans le réseau sémantique et qu'ils ne sont pas trop ambigus. Son efficacité est cependant amoindrie lorsque le texte contient des termes spécialisés dont le réseau n'a pas connaissance. Par exemple, le concept `pomme granny smith` pourrait être regroupé avec `pomme gala` et `pomme golden` car il s'agit de soeurs de la granny smith. Pour donner un autre exemple, le concept `animal` pourrait être regroupé avec `chat` et `kangourou` car il s'agit d'hyponymes d'`animal`. Cette technique est uniquement disponible pour les textes en anglais dans cette édition. Pour plus d'informations, voir «Réseaux sémantiques», à la page 113.

Remarque : L'option **Distance de recherche maximale** n'est disponible que si vous sélectionnez **Réseau sémantique**.

Distance de recherche maximale Sélectionnez jusqu'où vous souhaitez que les techniques effectuent la recherche avant de créer des catégories. Plus cette valeur est faible, moins les résultats seront nombreux ; toutefois, ils seront plus précis et liés ou associés entre eux de manière significative. Plus cette valeur est élevée, plus les résultats seront nombreux ; toutefois ils seront moins précis et moins fiables. Bien que cette option soit généralement appliquée à toutes les techniques, son effet est maximal sur les co-occurrences et les réseaux sémantiques.

Empêcher l'appariement de concepts spécifiques. Cochez cette case pour arrêter le processus de regroupement ou d'appariement de deux concepts dans les résultats. Pour créer ou gérer des paires de concepts, cliquez sur **Gérer les paires**. Pour plus d'informations, voir «Gestion des paires d'exceptions de liens», à la page 110.

Généraliser avec des caractères génériques lorsque cela est possible Sélectionnez cette option pour permettre au produit de créer des règles génériques dans les catégories à l'aide du caractère générique astérisque. Par exemple, au lieu de produire plusieurs descripteurs tels que `[tarte aux pommes + .]` et `[tarte aux fraises + .]`, l'utilisation de caractères génériques peut donner `[tarte * + .]`. Si vous généralisez avec des caractères génériques, vous obtenez souvent exactement le même nombre d'enregistrements ou de documents que précédemment. Toutefois, cette option a l'avantage de réduire le nombre de descripteurs de catégorie et de les simplifier. De plus, cette option augmente les possibilités de catégoriser davantage d'enregistrements ou de documents en utilisant ces catégories sur de nouvelles données textuelles (par exemple dans les enquêtes longitudinales/par vagues).

Autres options de génération de catégories

En plus de sélectionner les techniques de regroupement à appliquer, vous pouvez éditer plusieurs autres options de création comme suit :

Nombre maximum de catégories de niveau supérieur créées. Cette option sert à limiter le nombre de catégories pouvant être générées lorsque vous cliquez sur le bouton Créer des catégories. Dans certains cas, vous pouvez obtenir de meilleurs résultats si vous réglez cette valeur élevée puis supprimez les catégories sans intérêt.

Nombre minimum de descripteurs et/ou de sous-catégories par catégorie. Utilisez cette option pour définir le nombre minimum de descripteurs et de sous-catégories qu'une catégorie doit contenir pour être créée. Cette option permet de limiter la création de catégories qui ne capturent pas un nombre assez important d'enregistrements ou de documents.

Permettre aux descripteurs d'apparaître dans plus d'une catégorie Lorsqu'elle est sélectionnée, cette option permet aux descripteurs d'être utilisés dans plusieurs des catégories qui seront créées ensuite. Cette option est généralement sélectionnée car les éléments entrent fréquemment ou "naturellement" dans deux catégories ou plus, et le fait de leur donner cette possibilité permet d'obtenir des catégories d'une plus grande qualité. Si vous ne sélectionnez pas cette option, vous réduisez le chevauchement d'enregistrements dans plusieurs catégories et en fonction du type de données que vous avez, ceci peut être souhaitable. Toutefois, avec la plupart des types de données, le fait de limiter les descripteurs à une seule catégorie entraîne une perte de la qualité ou de la diversité des catégories. Par exemple, imaginons que vous avez un concept `fabricant sièges-auto`. Avec cette option, ce concept peut apparaître dans une catégorie basée sur le texte `sièges-auto` et dans une autre basée sur `fabricant`. Mais si cette option n'est pas sélectionnée, bien que vous puissiez quand même obtenir les deux catégories, le concept `fabricant sièges-auto` n'apparaîtra comme descripteur que dans la catégorie à laquelle il correspond le mieux en fonction de plusieurs facteurs, notamment du nombre d'enregistrements dans lesquels `sièges-auto` et `fabricant` apparaissent respectivement.

Résoudre les noms de catégories en double en Choisissez la manière de manipuler les nouvelles catégories ou sous-catégories dont le nom sera identique dans des catégories existantes. Vous pouvez fusionner les nouvelles catégories ou sous-catégories (et leur descripteurs) avec les catégories existantes qui portent le même nom. Vous pouvez également choisir d'ignorer la création de ces catégories si un nom en double se rencontre dans les catégories existantes.

Gestion des paires d'exceptions de liens

Au cours de la génération de catégorie, de la classification et du regroupement de concepts, les algorithmes internes regroupent les mots par associations connues. Pour ne pas associer deux concepts par paires, ou pour ne pas les lier, vous pouvez activer cette fonctionnalité dans la boîte de dialogue **Paramètres avancés de création de catégories**, la boîte de dialogue **Créer des clusters** et la boîte de dialogue **Paramètres de l'index de la carte de concepts**, puis cliquer sur le bouton **Gérer les paires**.

Dans la boîte de dialogue **Gérer les exceptions de liens**, vous pouvez ajouter, modifier ou supprimer les paires de concepts. Tapez une paire par ligne. Si vous entrez les paires ici, cela évitera la génération d'association par paires lors de la création ou de l'extension des catégories, de la classification, ou du regroupement de concepts. Saisissez les mots exactement comme vous les voulez, par exemple la version accentuée d'un mot n'est pas la même que la version non accentuée.

Par exemple, si vous voulez vous assurer que `foudre` et `coup de foudre` ne soient pas regroupés, vous pouvez ajouter la paire dans une ligne séparée du tableau :

A propos des techniques linguistiques

Quand vous créez ou développez des catégories, vous pouvez choisir parmi diverses techniques de génération de catégorie linguistiques avancées, dont la *dérivation des racines de concept*, l'*inclusion de*

concepts, les *réseaux sémantiques* (uniquement en anglais) et les *règles de co-occurrence*. Ces techniques peuvent être utilisées individuellement ou conjointement pour créer des catégories.

Vous n'avez pas besoin d'être un expert de ces paramètres pour les utiliser. Par défaut, les paramètres les plus communs et moyens sont déjà sélectionnés. Vous pouvez contourner la boîte de dialogue Paramètres avancés et créer ou étendre directement vos catégories. De même, si vous effectuez des modifications ici, vous n'avez pas besoin de revenir à la boîte de dialogue Paramètres à chaque fois car les derniers paramètres sont toujours conservés.

Néanmoins, notez que puisque chaque ensemble de données est unique, le nombre de méthodes et l'ordre dans lequel vous les appliquez peut varier. Dans la mesure où vos objectifs d'exploration de texte peuvent être différents d'un ensemble de données à un autre, vous pouvez tester les différentes techniques afin de déterminer celle qui génère les meilleurs résultats pour les données textuelles en question. Aucune technique automatique ne permet de catégoriser parfaitement vos données ; nous vous recommandons de rechercher et d'appliquer la ou les techniques qui sont les mieux adaptées à vos données.

Les principales techniques linguistiques automatiques pour la génération de catégories sont :

- **Dérivation des racines de concept.** Cette technique crée des catégories en sélectionnant un concept et en recherchant les autres concepts y étant associés. Pour cela, elle analyse si certains composants du concept sont liés d'un point de vue morphologique. Pour plus d'informations, voir «Dérivation des racines de concept».
- **Inclusion de concept.** Cette technique crée des catégories en sélectionnant un concept et en recherchant les autres concepts le contenant. Pour plus d'informations, voir «Inclusion de concepts», à la page 112.
- **Réseau sémantique.** Cette technique commence par identifier les sens possibles de chaque concept à partir de son index complet de relations existant entre les mots, puis crée des catégories en regroupant les concepts associés. Pour plus d'informations, voir «Réseaux sémantiques», à la page 113. Cette option est disponible uniquement pour le texte en anglais.
- **Co-occurrence.** Cette technique crée des règles de co-occurrence qui peuvent être utilisées pour créer une nouvelle catégorie, étendre une catégorie ou comme entrée pour une autre technique de catégorie. Pour plus d'informations, voir «Règles de co-occurrence», à la page 114.

Dérivation des racines de concept

La technique de dérivation des racines de concept crée des catégories en sélectionnant un concept et en recherchant les autres concepts y étant associés. Pour cela, elle analyse si certains composants du concept sont liés d'un point de vue morphologique. Un composant est un mot. La technique tente de regrouper les concepts en observant les terminaisons (suffixes) de chaque composant d'un concept et en recherchant les autres concepts ayant pu être créés à partir d'eux. Ainsi, lorsque des mots dérivent d'autres mots, il est vraisemblable qu'ils partagent la même signification ou qu'ils s'en rapprochent. Des règles internes, propres à chaque langue, identifient les terminaisons. Par exemple, le concept `opportunités d'avancer` serait regroupé avec les concepts `opportunité d'avancement` et `opportunité d'un avancement`.

Vous pouvez utiliser la dérivation des racines de concept sur n'importe quel type de texte. Utilisée seule, elle renvoie un nombre assez peu élevé de catégories et chacune d'entre elles contient généralement peu de concepts. Les concepts de chaque catégorie sont synonymes ou associés dans le cadre de la situation. Cet algorithme peut s'avérer utile, même si vous générez les catégories manuellement ; les synonymes qu'il repère peuvent être synonymes des concepts qui vous intéressent particulièrement.

Remarque : Vous pouvez éviter le regroupement des concepts en les spécifiant de manière explicite. Pour plus d'informations, voir «Gestion des paires d'exceptions de liens», à la page 110.

Fractionnement des termes en composants et suppression de l'inflexion

Lorsque vous appliquez les techniques de dérivation des racines de concept ou d'inclusion de concepts, les termes sont tout d'abord fractionnés en composants (mots), puis l'inflexion des composants est

supprimée. Lorsque vous appliquez une technique, les concepts et leurs termes associés sont chargés et fractionnés en composants en fonction de leurs séparateurs, tels que les espaces, les traits d'union et les apostrophes. Par exemple, le terme `administrateur système` est fractionné en composants comme suit : `{administrateur, système}`.

Cependant, certaines parties du terme d'origine peuvent ne pas être utilisées. C'est ce que l'on appelle des mots vides. Voici certains des composants pouvant être ignorés en français : `un`, `une`, `et`, `par`, `pour`, `de`, `du`, `des`, `dans`, `ou`, `le`, `à` et `avec`.

Par exemple, le terme `analyse des données` est constitué de l'ensemble de composants `{données, analyse}` et tant les mots `de` et `le` sont considérés comme pouvant être ignorés. Par ailleurs, l'ordre des mots dans un ensemble de composants n'est pas significatif. Ainsi les deux termes suivants sont équivalents : `prêt de voiture de luxe`, `voiture de luxe en prêt`, puisqu'ils ont tous les deux le même ensemble de composants `{voiture, prêt, luxe}`. Chaque fois que deux termes sont considérés comme étant équivalents, les concepts correspondants sont fusionnés pour constituer un nouveau concept qui référence l'ensemble des termes.

De plus, puisque les composants d'un terme peuvent être infléchies, des règles propres à la langue sont appliquées en interne pour identifier les termes équivalents, quelle que soit la variation occasionnée par leur flexion (formes au pluriel, par exemple). De cette manière, les termes `niveau de fiabilité` et `niveaux de fiabilité` peuvent être identifiés comme étant équivalents, puisque la forme au singulier sans inflexion est `niveau`.

Fonctionnement de la dérivation des racines de concept

Après que les termes ont été fractionnés en composants et que leur inflexion a été supprimée (voir la section précédente), l'algorithme de dérivation des racines de concept analyse les terminaisons, ou suffixes, des composants pour retrouver leur radical, puis regroupe les concepts dont les radicaux sont identiques ou similaires. Les terminaisons sont identifiées à l'aide d'un ensemble de règles de dérivation linguistique propres à la langue du texte. Par exemple, une règle de dérivation concernant un texte rédigé en langue française détermine qu'un composant de concept présentant le suffixe `ique` peut être dérivé d'un concept ayant le même radical, mais se terminant par le suffixe `ie`. Grâce à cette règle (et à la suppression de l'inflexion), l'algorithme peut regrouper les concepts `étude épidémiologique` et `études épidémiologiques`.

Etant donné que les termes ont déjà été fractionnés en composants et que les composants pouvant être ignorés (par exemple, `en` et `d'`) ont été identifiés, l'algorithme de dérivation des racines de concept est également capable de regrouper le concept `étude d'épidémiologie` with `étude épidémiologique`.

Les règles de dérivation d'un ensemble de composants ont été choisies de sorte que la plupart des concepts regroupés par cet algorithme soient synonymes, ce qui est le cas de `étude d'épidémiologies`, `étude épidémiologique` et `étude en épidémiologie`. Pour augmenter le degré d'exhaustivité, certaines règles de dérivation laissent l'algorithme regrouper des concepts qui sont associés dans le cadre de la situation. Par exemple, l'algorithme peut regrouper des concepts tels que `réparer les voitures` et `réparation de voitures`.

Inclusion de concepts

La technique d'inclusion de concepts crée des catégories en prenant un concept et, à l'aide d'algorithmes de série lexicale, identifie les concepts inclus dans d'autres concepts. Ainsi, lorsque les mots d'un concept constituent un sous-ensemble d'un autre concept, ces algorithmes reflètent une relation sémantique sous-jacente. L'inclusion est une technique puissante qui peut être utilisée avec n'importe quel type de texte.

Elle fonctionne bien en conjonction avec des réseaux sémantiques, mais elle peut être utilisée séparément. L'inclusion de concepts peut aussi fournir de meilleurs résultats lorsque les documents ou les enregistrements contiennent de nombreux termes appartenant à un domaine spécifique. Ceci se confirme

particulièrement lorsque vous avez préalablement affiné les dictionnaires de telle sorte que les termes spéciaux soient extraits et regroupés de façon appropriée (avec leurs synonymes).

Fonctionnement de l'inclusion de concept

Avant l'application de l'algorithme d'inclusion de concept, les termes sont fragmentés en composants et leur inflexion est supprimée. Pour plus d'informations, voir «Dérivation des racines de concept», à la page 111. Ensuite, l'algorithme d'inclusion de concept analyse les ensembles de composants. Pour chacun d'entre eux, l'algorithme recherche un ensemble de composants qui soit un sous-ensemble du premier.

Par exemple, si vous disposez du concept `déjeuner diététique`, qui comporte l'ensemble de composants `{déjeuner, diététique}`, et le concept `déjeuner`, qui comporte l'ensemble de composants `{déjeuner}`, l'algorithme en conclurait que `déjeuner diététique` est un type de `déjeuner` et les regrouperait.

Pour citer un exemple plus étendu, si le concept `siège` figure dans la sous-fenêtre Résultats d'extraction et que vous appliquez cet algorithme, les concepts tels que `siège de sécurité`, `siège en cuir`, `siège couchette`, `commande de siège éjectable`, `siège-auto coque` et `instructions pour siège-auto` seraient également regroupés dans cette catégorie.

Puisque les termes sont déjà fractionnés en composants et que les composants pouvant être ignorés (par exemple, `en` et `d'`) ont été identifiés, l'algorithme d'inclusion de concepts reconnaît que le concept `cours d'espagnol avancé` inclut le concept `cours en espagnol`.

Remarque : vous pouvez éviter que les concepts soient regroupés ensemble en les spécifiant de manière explicite. Pour plus d'informations, voir «Gestion des paires d'exceptions de liens», à la page 110.

Réseaux sémantiques

Dans cette édition, la technique de réseau sémantique n'est disponible que pour les textes rédigés en anglais.

Cette technique crée des catégories à l'aide d'un réseau intégré de relations entre les mots. De ce fait, cette technique peut produire d'excellents résultats lorsque les termes sont précis et ne sont pas trop ambigus. En revanche, ne vous attendez pas à identifier de nombreux liens entre des concepts hautement techniques/spécialisés. En travaillant avec des concepts de ce type, il est probable que les techniques d'inclusion de concepts et de dérivation des racines de concept s'avèrent plus utiles.

Fonctionnement du réseau sémantique

La technique de réseau sémantique vise à exploiter les relations connues existant entre les mots pour créer des catégories de synonymes ou d'hyponymes. Un **hyponyme** est un concept considéré comme étant un type de second concept, de façon à ce qu'une relation hiérarchique, également appelée relation ISA, soit établie. Par exemple, si `animal` est un concept, `chat` et `kangourou` sont des hyponymes d'`animal`, puisque ce sont des types d'animaux.

Outre les relations de synonyme et d'hyponyme, la technique du réseau sémantique examine également les liens partie-tout existant entre les concepts du type `<Location>`. Par exemple, cette technique regroupe les concepts `normandie`, `provence` et `france` en une seule catégorie, car la Normandie et la Provence sont des parties de la France.

Les réseaux sémantiques commencent par identifier les sens possibles de chaque concept du réseau. Lorsque des concepts sont identifiés comme étant synonymes ou hyponymes, ils sont regroupés dans une seule catégorie. Par exemple, une catégorie pourra regrouper les concepts `pomme`, `pomme sucrée` et `granny smith`, car le réseau sémantique contient les informations suivantes : 1) `pomme sucrée` est un synonyme de `pomme` et 2) `granny smith` est une sorte de `pomme` (hyponyme de `pomme`).

Considérés individuellement, de nombreux concepts, en particulier les termes univoques, sont ambigus. Par exemple, le concept buffet peut désigner un type de repas ou un meuble. Si l'ensemble des concepts inclut repas, meuble et buffet, l'algorithme est forcé de choisir entre regrouper buffet avec repas ou avec meuble. Sachez que, dans certains cas, les choix effectués par l'algorithme peuvent ne pas être appropriés dans le contexte d'un ensemble particulier d'enregistrements ou de documents.

La technique du réseau sémantique peut offrir de bien meilleures performances que l'inclusion de concepts avec certains types de données. Alors que le réseau sémantique et l'inclusion de concepts reconnaissent que pinceau rouge est une sorte de pinceau, seul le réseau sémantique reconnaît que rouleau est également une sorte depinceau.

Les réseaux sémantiques peuvent être utilisés conjointement avec les autres techniques. Par exemple, imaginons que vous ayez sélectionné les techniques de réseau sémantique et d'inclusion, et que le réseau sémantique ait regroupé le concept professeur avec le concept tuteur (car un tuteur est un type de professeur). L'algorithme d'inclusion peut regrouper le concept tuteur diplômé avec tuteur ; les deux algorithmes sont alors à même de générer une catégorie de sortie contenant les trois concepts tuteur, tuteur diplômé et professeur.

Options du réseau sémantique

Il existe de nombreux paramètres supplémentaires pouvant être intéressants avec cette technique.

- Changez la **distance de recherche maximale**. Sélectionnez jusqu'où vous souhaitez que les techniques effectuent la recherche avant de créer des catégories. Plus cette valeur est faible, moins les résultats seront nombreux ; toutefois, ils seront plus précis et liés ou associés entre eux de manière significative. Plus cette valeur est élevée, plus les résultats seront nombreux ; toutefois ils seront moins précis et moins fiables.

Par exemple, en fonction de la distance, l'algorithme recherche de pain au chocolat jusqu'à viennoiserie (son parent), puis petit pain (grand-parent) et vers le haut jusqu'à pain.

En réduisant la distance de recherche, cette technique produit des catégories plus petites qui peuvent faciliter le travail lorsque les catégories produites sont trop grandes ou regroupent trop d'éléments.

Important ! Il est recommandé, lorsque vous utilisez cette technique, de ne pas appliquer l'option **Nombre de caractères minimum requis** (qui figure dans l'onglet Expert du noeud ou dans la boîte de dialogue Extraire) pour les regroupements flous, car certains mauvais regroupements peuvent avoir un impact très négatif sur les résultats.

Règles de co-occurrence

Les règles de co-occurrence vous permettent d'identifier et de regrouper les concepts étroitement liés au sein de l'ensemble de documents ou d'enregistrements. Ainsi, lorsque des concepts apparaissent souvent ensemble dans des documents et des enregistrements, la co-occurrence reflète une relation sous-jacente qui a vraisemblablement de l'intérêt dans vos définitions de catégories. Cette technique crée des règles de co-occurrence qui peuvent être utilisées pour créer une nouvelle catégorie, étendre une catégorie ou comme entrée pour une autre technique de catégorie. Deux concepts co-existent fortement s'ils apparaissent fréquemment ensemble dans un ensemble d'enregistrements et rarement séparément dans les autres enregistrements. Cette technique génère de bons résultats avec de plus grands ensembles de données contenant au moins plusieurs centaines de documents ou d'enregistrements.

Par exemple, si de nombreux enregistrements contiennent les mots prix et disponibilité, ces concepts peuvent être regroupés dans une règle de co-occurrence, (prix & disponible). Autre exemple : si les concepts beurre, confiture et croissant apparaissent plus souvent ensemble que seuls, ils seront regroupés dans une règle de co-occurrence de concept (beurre & confiture & croissant).

Important ! Dans les éditions précédentes, les règles de cooccurence et de synonymes étaient entre crochets. Dans l'édition actuelle, les crochets indiquent désormais un résultat de motifs d'analyse des liens du texte. Les règles de co-occurrence et de synonymes sont maintenant entourées de parenthèses, comme dans `(enceintes acoustiques|enceintes)`.

Fonctionnement des règles de co-occurrence

Cette technique analyse les documents ou les enregistrements à la recherche d'au moins deux concepts apparaissant souvent ensemble. Deux concepts co-existent fortement s'ils apparaissent fréquemment ensemble dans un ensemble de documents ou d'enregistrements et rarement séparément dans les autres documents ou enregistrements.

Une règle de catégorie se crée dès que le système identifie des concepts co-occurrents. Ces règles comportent au moins deux concepts reliés par l'opérateur booléen &. Il s'agit d'instructions logiques qui classent automatiquement un document ou un enregistrement dans une catégorie, si l'ensemble des concepts qu'elles régissent sont co-occurrents dans ces documents ou enregistrements.

Options des règles de co-occurrence

Si vous utilisez la technique des règles de co-occurrence, vous pouvez régler plusieurs paramètres ayant une influence sur les règles obtenues :

- Changez la **distance de recherche maximale**. Sélectionnez la distance de recherche de co-occurrences. Si vous augmentez la distance de recherche, la valeur de similarité minimum requise pour chaque co-occurrence diminue et par conséquent, de nombreuses règles de co-occurrence peuvent être produites, mais celles avec une valeur de similarité basse auront généralement peu d'importance. Lorsque vous diminuez la distance de recherche, la valeur de similarité requise minimum augmente et par conséquent, moins de règles de co-occurrences sont produites mais elles ont tendance à être plus importantes (plus fortes).
- **Nombre minimum de documents.** Le nombre minimum d'enregistrements ou de documents qui doivent contenir une paire de concepts donnée pour qu'elle soit considérée comme une co-occurrence. Plus cette option a une valeur basse, plus il est facile de trouver des co-occurrences. Augmenter la valeur produit des co-occurrences moins nombreuses mais plus importantes. Par exemple, imaginons que les concepts « pomme » et « poire » se trouvent ensemble dans 2 enregistrements (et qu'aucun des deux ne se trouve dans d'autres enregistrements). Avec **Nombre minimum de documents.** définis à 2 (valeur par défaut), la technique de co-occurrence crée une règle de catégorie (pomme et poire). Si la valeur passe à 3, la règle ne sera plus créée.

Remarque : avec de petits ensemble de données (< 1000 réponses), vous risquez de ne pas obtenir de co-occurrence avec les paramètres par défaut. Si c'est le cas, essayez d'augmenter la valeur de la distance de recherche.

Remarque : vous pouvez éviter que les concepts soient regroupés ensemble en les spécifiant de manière explicite. Pour plus d'informations, voir «Gestion des paires d'exceptions de liens», à la page 110.

Paramètres de fréquence avancés

Vous pouvez créer des catégories en fonction d'une technique de fréquence simple et mécanique. Avec cette technique, vous pouvez créer une catégorie pour chaque élément (type, concept ou motif) situé en amont des effectifs des enregistrements ou des documents. Vous pouvez également créer une catégorie regroupant tous les termes moins fréquents. Les effectifs désignent le nombre d'enregistrements ou de documents contenant le concept extrait (et tous ses synonymes), le type ou le motif dans la question par rapport au nombre total d'occurrences d'un concept, d'un type ou d'un motif dans l'ensemble du texte.

Le regroupement d'éléments fréquents peut produire des résultats intéressants, car il peut indiquer une réponse courante ou importante. Elle est très efficace si elle a été exécutée sur les résultats d'extraction inutilisés après que d'autres techniques ont été appliquées. Une autre application consiste à exécuter cette

technique immédiatement après l'extraction lorsqu'aucune autre catégorie n'existe, éditer les résultats pour supprimer les catégories sans intérêt, puis étendre ces catégories afin de leur faire correspondre toujours plus d'enregistrements ou de documents. Pour plus d'informations, voir «Extension de catégories».

A la place de cette technique, vous pouvez également trier les concepts et les motifs de concept par nombre décroissant des enregistrements ou des documents dans la sous-fenêtre Résultats d'extraction puis en faisant glisser et en déposant les premiers dans la sous-fenêtre Catégories pour créer les catégories correspondantes.

Les zones suivantes sont disponibles dans la boîte de dialogue Paramètres avancés : Fréquences :

Générer des descripteurs de catégories à. Sélectionnez le type d'entrée pour les descripteurs. Pour plus d'informations, voir «Génération de catégories», à la page 106.

- **Niveau de concept**. Le fait de sélectionner cette option signifie que les fréquences de concepts ou de motifs de concept seront utilisées. Les concepts sont utilisés si les types ont été sélectionnés comme entrée pour la génération de catégorie et les motifs de concept sont utilisés si les motifs de type ont été sélectionnés. En général, appliquer cette technique au niveau du concept produira des résultats plus précis car les concepts et motifs de concept représentent un niveau de mesure moins élevé.
- **Niveau de type**. Le fait de sélectionner cette option signifie que les fréquences de types ou de motifs de type seront utilisées. Les types sont utilisés si les types ont été sélectionnés comme entrée pour la génération de catégorie et les motifs de type sont utilisés si les motifs de type ont été sélectionnés. Cette technique appliquée au niveau du type permet d'obtenir une vue rapide relative au genre d'informations présentes.

Effectif minimal docs pour que des éléments aient leur propre catégorie. Cette option vous permet de créer des catégories à partir des éléments fréquents. Cette option limite les résultats aux seules catégories qui contiennent un descripteur que l'on rencontre dans au moins X enregistrements ou X documents, où X est la valeur à saisir pour cette option.

Regrouper tous les éléments restants dans une catégorie nommée. Cette option vous permet de regrouper tous les concepts ou types peu fréquents en une seule catégorie 'fourre-tout' portant le nom de votre choix. Par défaut, cette catégorie se nomme *Autre*.

Entrée de la catégorie. Sélectionnez le groupe auquel appliquer ces techniques :

- **Résultats d'extraction non utilisés**. Cette option permet aux catégories d'être créées à partir des résultats d'extraction qui ne sont pas encore utilisés dans une catégorie existante. Ceci réduit la tendance des enregistrements à correspondre à plusieurs catégories et limite le nombre de catégories produites.
- **Tous les résultats d'extraction**. Cette option permet aux catégories d'êtres créées à l'aide de tous les résultats d'extraction. Ceci est particulièrement utile quand aucune ou peu de catégories existent déjà.

Résoudre les noms de catégories en double en. Choisissez la manière de manipuler les nouvelles catégories ou sous-catégories dont le nom sera identique dans des catégories existantes. Vous pouvez fusionner les nouvelles catégories ou sous-catégories (et leur descripteurs) avec les catégories existantes qui portent le même nom. Vous pouvez également choisir d'ignorer la création de ces catégories si un nom en double se rencontre dans les catégories existantes.

Extension de catégories

L'extension est un processus au cours duquel des descripteurs sont ajoutés ou améliorés automatiquement pour « agrandir » les catégories existantes. L'objectif est de produire une meilleure catégorie qui capture les enregistrements ou documents associés qui n'ont pas été attribués à cette catégorie à l'origine.

Les techniques de regroupement automatiques que vous sélectionnez tenteront d'identifier les concepts, motifs TLA et règles de catégorie associées aux descripteurs de catégorie existants. Ces nouveaux concepts, motifs et règles de catégorie sont ensuite ajoutés comme nouveaux descripteurs ou ajoutés aux descripteurs existants. Les techniques de regroupement pour l'extension incluent la *dérivation des racines de concept*, l'*inclusion de concepts*, les *réseaux sémantiques* (uniquement en anglais) et les *règles de co-occurrence*. La méthode **Etendre les catégories vides avec des descripteurs générés à partir du nom de la catégorie** génère des descripteurs à l'aide des mots dans les noms de catégories. Ainsi, plus les noms de catégories sont descriptifs, meilleurs sont les résultats.

Remarque : Les techniques de fréquence ne sont pas disponibles pour l'extension de catégories.

L'extension est une excellente manière d'améliorer vos catégories de manière interactive. Voici quelques exemples de cas où vous pouvez étendre une catégorie :

- Après avoir glissé/déposé les motifs de concept pour créer des catégories dans la sous-fenêtre Catégories
- Après avoir créé des catégories manuellement et avoir ajouté des règles de catégorie et des descripteurs simples
- Après avoir importé un fichier de catégorie prédéfini où les catégories portent des noms très descriptifs
- Après avoir affiné les catégories provenant du TAP que vous avez choisi

Vous pouvez étendre plusieurs fois une catégorie. Par exemple, si vous avez importé un fichier de catégorie prédéfinie avec des noms très descriptifs, vous pouvez effectuer l'extension avec l'option **Etendre les catégories vides avec des descripteurs créés à partir du nom de la catégorie** pour obtenir un premier ensemble de descripteurs puis étendre de nouveau ces catégories. Néanmoins, dans d'autres cas, l'extension multiple peut entraîner une catégorie trop générique si les descripteurs sont de plus en plus étendus. Comme les techniques de création et d'extension de regroupement utilisent des algorithmes sous-jacents similaires, l'extension directement après la génération de catégories ne produira probablement pas de résultats plus intéressants.

Conseil :

- Si vous tentez une extension et ne voulez pas utiliser les résultats, vous pouvez toujours annuler l'opération (**Edition > Annuler**) immédiatement après l'extension.
- L'extension peut produire au moins deux règles de catégorie dans une catégorie qui correspondent exactement au même ensemble de documents, les règles étant créées indépendamment pendant le processus. Si besoin est, vous pouvez afficher les catégories et supprimer les redondances en modifiant manuellement la description des catégories. Pour plus d'informations, voir «Modification des descripteurs de catégorie», à la page 139.

Pour étendre des catégories

1. Dans la sous-fenêtre Catégories, sélectionnez les catégories à étendre.
2. Dans les menus, sélectionnez **Catégories > Etendre des catégories**. Un message apparaît, sauf si vous avez choisi de ne pas recevoir d'invite.
3. Choisissez si vous voulez créer maintenant ou éditer d'abord les paramètres.

- Cliquez sur **Etendre** pour commencer à étendre des catégories à l'aide des paramètres actuels. Le processus commence et une boîte de dialogue de progression apparaît.
- Cliquez sur **Editer** pour examiner et modifier les paramètres.

Après la tentative d'extension, toutes les catégories pour lesquelles de nouveaux descripteurs ont été trouvés sont signalées par le mot **Etendue** dans la sous-fenêtre Catégories, afin que vous puissiez les identifier rapidement. Le texte Etendue reste affiché jusqu'à ce que vous étendiez à nouveau la catégorie, que vous la modifiez d'une autre manière, ou la supprimiez via le menu contextuel.

Remarque : Vous pouvez afficher jusqu'à 10 000 catégories. Un avertissement s'affiche lorsque ce nombre est atteint ou dépassé. Si cela se produit, modifiez les options Créer ou développer des catégories afin de réduire le nombre de catégories créées.

Chaque technique disponible lors de la création ou l'extension de catégories convient à certains types de données et de situations. Cependant, il est souvent judicieux de combiner plusieurs techniques dans la même analyse afin de capturer le maximum de documents ou d'enregistrements. Dans le plan de travail interactif, les concepts et les types qui ont été regroupés dans une catégorie restent disponibles et pourront être classés par le biais d'autres techniques. Aussi est-il possible de voir un concept figurer dans plusieurs catégories ou de rencontrer des catégories redondantes.

Les zones et champs suivants sont disponibles dans la boîte de dialogue Etendre les catégories : paramètres :

Etendre avec. Sélectionnez l'entrée à utiliser pour étendre les catégories :

- **Résultats d'extraction non utilisés**. Cette option permet aux catégories d'être créées à partir des résultats d'extraction qui ne sont pas encore utilisés dans une catégorie existante. Ceci réduit la tendance des enregistrements à correspondre à plusieurs catégories et limite le nombre de catégories produites.
- **Tous les résultats d'extraction**. Cette option permet aux catégories d'êtres créées à l'aide de tous les résultats d'extraction. Ceci est particulièrement utile quand aucune ou peu de catégories existent déjà.

Techniques de regroupement

Pour des descriptions brèves de chacune de ces techniques, voir «Paramètres linguistiques avancés», à la page 108. Ces techniques comprennent :

- **Dérivation des racines de concept**
- **Réseau sémantique** (texte anglais uniquement et non utilisé si l'option Généraliser uniquement est sélectionnée.)
- **Inclusion de concept**
- **Co-occurrence** et sous-option **Nombre minimal de documents**.

Plusieurs types sont définitivement exclus de la technique de réseau sémantique car ils ne renverront pas de résultats pertinents. Ils comprennent `<Positive>`, `<Negative>`, `<IP>`, d'autres types non-linguistiques, etc.

Distance de recherche maximale Sélectionnez jusqu'où vous souhaitez que les techniques effectuent la recherche avant de créer des catégories. Plus cette valeur est faible, moins les résultats seront nombreux ; toutefois, ils seront plus précis et liés ou associés entre eux de manière significative. Plus cette valeur est élevée, plus les résultats seront nombreux ; toutefois ils seront moins précis et moins fiables. Bien que cette option soit généralement appliquée à toutes les techniques, son effet est maximal sur les co-occurrences et les réseaux sémantiques.

Empêcher l'appariement de concepts spécifiques. Cochez cette case pour arrêter le processus de regroupement ou d'appariement de deux concepts dans les résultats. Pour créer ou gérer des paires de concepts, cliquez sur **Gérer les paires**. Pour plus d'informations, voir «Gestion des paires d'exceptions de liens», à la page 110.

Si possible : Choisissez d'étendre uniquement, de généraliser les descripteurs à l'aide de caractères génériques, ou les deux.

- **Etendre et généraliser.** Cette option permet d'étendre les catégories sélectionnées et de généraliser les descripteurs. Lorsque vous choisissez de généraliser, le produit crée des règles de catégorie génériques dans les catégories à l'aide du caractère générique astérisque. Par exemple, au lieu de produire plusieurs descripteurs tels que `[tarte aux pommes + .]` et `[tarte aux fraises + .]`, l'utilisation de

caractères génériques peut donner [tarte * + .]. Si vous généralisez avec des caractères génériques, vous obtenez souvent exactement le même nombre d'enregistrements ou de documents que précédemment. Toutefois, cette option a l'avantage de réduire le nombre de descripteurs de catégorie et de les simplifier. De plus, cette option augmente les possibilités de catégoriser davantage d'enregistrements ou de documents en utilisant ces catégories sur de nouvelles données textuelles (par exemple dans les enquêtes longitudinales/par vagues).

- **Etendre uniquement**. Cette option étend vos catégories sans généralisation. Il peut être utile de choisir d'abord l'option **Etendre uniquement** pour les catégories créées manuellement puis d'étendre de nouveau ces mêmes catégories à l'aide de l'option **Etendre et généraliser**.
- **Généraliser uniquement.** Cette option généralisera les descripteurs sans étendre vos catégories de quelque autre façon.

  Remarque : Si vous sélectionnez cette option, l'option **Réseau sémantique** est désactivée, car elle est disponible uniquement lorsqu'une description doit être étendue.

Autres options d'extension de catégories

En plus de sélectionner les techniques de regroupement à appliquer, vous pouvez éditer les options suivantes :

Nombre maximal d'éléments par lequel étendre un descripteur. Lors de l'extension d'un descripteur avec des éléments (concepts, types et autres expressions), définissez le nombre maximum d'éléments pouvant être ajoutés à un seul descripteur. Si vous fixez cette limite à 10, un maximum de 10 éléments supplémentaires peuvent être ajoutés à un descripteur existant. Si plus de 10 éléments doivent être ajoutés, les techniques arrêtent d'ajouter de nouveaux éléments après l'ajout du dixième. Ceci peut raccourcir la liste de descripteur mais ne garantit pas que les éléments les plus intéressants aient été utilisés en premier. Vous préférez peut-être réduire la taille de l'extension sans pénaliser la qualité en utilisant l'option **Généraliser avec des caractères génériques lorsque cela est possible**. Cette option s'applique uniquement aux descripteurs qui contiennent un opérateur booléen& (AND) ou ! (NOT).

Etendre également les sous-catégories. Cette option étendra également toutes les sous-catégories sous les catégories sélectionnées.

Etendre les catégories vides avec des descripteurs générés à partir du nom de la catégorie. Cette méthode s'applique uniquement aux catégories vides qui ont 0 descripteurs. Si une catégorie contient déjà des descripteurs, elle ne sera pas étendue de cette manière. Cette option tente de créer automatiquement des descripteurs pour chaque catégorie en fonction des mots constituant le nom de la catégorie. Le nom de catégorie est analysé pour voir si les mots du nom correspondent à des concepts extraits. Si un concept est reconnu, il est utilisé pour trouver les motifs de concept correspondants et ceux-ci sont utilisés pour générer des descripteurs pour la catégorie. Cette option produit les meilleurs résultats lorsque les noms de catégories sont à la fois longs et descriptifs. C'est une méthode rapide pour générer des descripteurs de catégories, qui à leur tour permettent à la catégorie de capturer les enregistrements contenant ces descripteurs. Cette option est particulièrement utile quand vous importez des catégories d'ailleurs ou quand vous créez des catégories manuellement avec de longs noms descriptifs.

Générer des descripteurs comme. Cette option s'applique uniquement si l'option précédente est sélectionnée.

- **Concepts.** Choisissez cette option pour produire les résultats des descripteurs sous la forme de concepts, qu'ils aient été extraits ou non du texte source.
- **Motifs.** Choisissez cette option pour produire les résultats des descripteurs sous la forme de motifs, que les résultats des motifs ou tout autre motif aient été extraits ou non.

Création manuelle de catégories

En plus de créer des catégories à l'aide des techniques de génération de catégories automatiques que sont l'éditeur de règles, vous pouvez également créer des catégories manuellement. Les méthodes manuelles suivantes existent :

- Création d'une catégorie vide dans laquelle vous ajouterez des éléments un par un. Pour plus d'informations, voir «Création de catégories ou attribution d'un nouveau nom aux catégories».
- Déplacement de termes, de types et de motifs dans la sous-fenêtre des catégories. Pour plus d'informations, voir «Création de catégories par la méthode Glisser-déposer».

Création de catégories ou attribution d'un nouveau nom aux catégories

Vous pouvez créer des catégories vides en vue d'y ajouter des concepts et des types. Vous pouvez également renommer vos catégories.

Pour créer une catégorie vide

1. Accédez à la sous-fenêtre Catégories.
2. Dans les menus, choisissez **Catégories > Créer une catégorie vide**. La boîte de dialogue Propriétés de catégorie apparaît.
3. Entrez un nom dans le champ Nom pour cette catégorie.
4. Cliquez sur **OK** pour valider ce nom et fermer la boîte de dialogue. La boîte de dialogue se ferme et un nouveau nom de catégorie apparaît dans la sous-fenêtre.

Vous pouvez maintenant commencer à ajouter des éléments à cette catégorie. Pour plus d'informations, voir «Ajout de descripteurs aux catégories», à la page 139.

Pour renommer une catégorie

1. Sélectionnez une catégorie, puis choisissez **Catégories > Renommer la catégorie**. La boîte de dialogue Propriétés de catégorie apparaît.
2. Entrez un nouveau nom dans le champ Nom pour cette catégorie.
3. Cliquez sur **OK** pour valider ce nom et fermer la boîte de dialogue. La boîte de dialogue se ferme et un nouveau nom de catégorie apparaît dans la sous-fenêtre.

Création de catégories par la méthode Glisser-déposer

La technique du glisser-déposer est manuelle et n'est pas basée sur des algorithmes. Vous pouvez créer des catégories dans la sous-fenêtre Catégories en glissant-déposant :

- des concepts, des types ou des motifs extraits de la sous-fenêtre Résultats d'extraction vers la sous-fenêtre Catégories.
- des concepts extraits de la sous-fenêtre Données vers la sous-fenêtre Catégories.
- des lignes entières de la sous-fenêtre Données vers la sous-fenêtre Catégories. Une catégorie sera créée avec tous les concepts et les motifs extraits contenus dans cette ligne.

Remarque : La sous-fenêtre Résultats d'extraction prend en charge la sélection multiple afin de faciliter le glisser-déposer d'éléments multiples.

Important ! Vous ne pouvez pas faire glisser et déposer des concepts provenant de la sous-fenêtre Données et qui n'ont pas été extraits du texte. Si vous souhaitez forcer l'extraction d'un concept trouvé dans les données, vous devez l'ajouter à un type. Puis réexécutez l'extraction. Les nouveaux résultats de l'extraction contiendront le concept que vous venez d'ajouter. Vous pouvez l'utiliser dans votre catégorie. Pour plus d'informations, voir «Ajout de concepts à des types», à la page 91.

Pour créer des catégories avec la méthode de glisser-déposer :

1. Dans la sous-fenêtre Résultats d'extraction ou la sous-fenêtre Données, sélectionnez un ou plusieurs concepts, motifs, types, enregistrements ou enregistrements partiels.
2. Tout en maintenant le bouton de la souris appuyé, faites glisser l'élément vers une catégorie existante ou vers la zone de la sous-fenêtre pour créer une nouvelle catégorie.
3. Lorsque vous atteignez la zone où vous souhaitez déposer cet élément, relâchez le bouton de la souris. Cet élément est ajouté à la sous-fenêtre Catégories. Les catégories modifiées apparaissent avec une couleur d'arrière-plan spécifique. Cette couleur s'appelle **l'arrière-plan des commentaires de catégorie**. Pour plus d'informations, voir la rubrique «Définition des options», à la page 75.

Remarque : la catégorie résultante est nommée automatiquement. Si vous le souhaitez, vous pouvez modifier ce nom. Pour plus d'informations, voir «Création de catégories ou attribution d'un nouveau nom aux catégories», à la page 120.

Si vous souhaitez voir quels enregistrements sont affectés à une catégorie, sélectionnez cette catégorie dans la sous-fenêtre Catégories. La sous-fenêtre des données est automatiquement actualisée et affiche tous les enregistrements pour cette catégorie.

Utilisation des règles de catégorie

Vous pouvez créer des catégories de différentes façons. Une de ces façons est de définir des règles de catégorie permettant d'exprimer des idées. Les règles de catégorie sont des instructions qui classifient automatiquement les documents ou les enregistrements en une catégorie basée sur une expression logique à l'aide de concepts, de types et de motifs extraits ou d'opérateurs booléens. Par exemple, vous pouvez créer une expression qui signifie *inclure tous les enregistrements contenant le concept extrait* `ambassade` *mais ne pas inclure* `argentine` *dans cette catégorie*.

Certaines règles de catégorie sont produites automatiquement lors de la génération de catégories à l'aide des techniques de regroupement telles que la *co-occurrence* et la *dérivation des racines de concept* (**Catégories > >Créer des paramètres > Paramètres avancés : Linguistique**) ; vous pouvez également créer manuellement des règles de catégorie dans l'éditeur de règle à l'aide de votre compréhension des données et du contexte. Chaque règle est rattachée à une seule catégorie afin que chaque document ou enregistrement correspondant à la règle soit ensuite scoré dans cette catégorie.

Les règles de catégorie permettent d'améliorer la qualité et la productivité de vos résultats de Text Mining et d'autres analyses quantitatives en vous aidant à classer les réponses dans des catégories plus spécifiques. Votre expérience et vos connaissances du marché peuvent vous offrir une compréhension particulière de vos données et du contexte. Cette compréhension peut vous permettre de traduire vos connaissances en règles de catégorie afin de classer vos documents ou vos enregistrements de manière plus efficace et plus précise en combinant des éléments extraits avec la logique booléenne.

La possibilité de créer ces règles améliore la précision, l'efficacité et la productivité du codage, tout en vous permettant d'incorporer vos connaissances du marché à la technologie d'extraction de produit.

Remarque : vous trouverez des exemples de règles correspondant à du texte à la rubrique «Exemples de règles de catégorie», à la page 127

Syntaxe des règles de catégorie

Certaines règles de catégorie sont produites automatiquement lors de la génération de catégories à l'aide des techniques de regroupement telles que la *co-occurrence* et la *dérivation des racines de concept* (**Catégories > >Créer des paramètres > Paramètres avancés : Linguistique**) ; vous pouvez également créer manuellement des règles de catégorie dans l'éditeur de règle. Chaque règle est le descripteur d'une seule catégorie afin que chaque document ou enregistrement correspondant à la règle soit automatiquement scoré dans cette catégorie.

Remarque : vous trouverez des exemples de règles correspondant à du texte à la rubrique «Exemples de règles de catégorie», à la page 127

Lorsque vous créez ou modifiez une règle, elle doit être ouverte dans la sous-fenêtre de l'éditeur de règles. Vous pouvez ajouter des concepts, des types ou des motifs et utiliser des caractères génériques pour étendre les correspondances. Lorsque vous utilisez des concepts, des types ou des motifs extraits, vous pouvez bénéficier de la recherche de tous les concepts associés.

Important ! Pour éviter les erreurs courantes, nous vous recommandons de glisser-déposer directement les concepts de la sous-fenêtre Résultats d'extraction, des sous-fenêtres Analyse des liens du texte ou de la sous-fenêtre Données dans l'éditeur de règles ou de les ajouter à l'aide des menus contextuels lorsque cela est possible.

Lorsque les concepts, les types et les motifs sont reconnus, une icône apparaît à côté du texte.

Tableau 18. Icônes d'extraction

Icône	Description
	Concept extrait
	Type extrait
	Motif extrait

Syntaxe et opérateurs de règle

Le tableau suivant contient les caractères avec lesquels vous définissez la syntaxe de votre règle. Utilisez ces caractères avec les concepts, les types et les motifs pour créer votre règle.

Tableau 19. Syntaxe prise en charge

Caractère	Description

`&`	Le "et" booléen. Par exemple, `a & b` contient à la fois `a` *et* `b` comme dans :
	`- invasion & états unis`
	`- jeux olympiques & 2016`
	`- bonne & pomme`

`|`	Le "ou" booléen est inclusif ce qui signifie que si un ou tous les éléments sont trouvés, une correspondance est établie. Par exemple, `a | b` contient soit `a` *soit* `b` comme dans :
	`- attaque | france`
	`- copropriété | appartement`

`!()`	Le "pas" booléen. Par exemple, `!(a)` ne contient pas `a`, comme dans : `!(good & hotel)`, `assassination & !(austria)`, or `!(gold) & !(copper)`
`*`	Un caractère générique représente toute expression, d'un caractère unique à un mot entier selon la façon dont il est utilisé. Pour plus d'informations, voir «Utilisation de caractères génériques dans les règles de catégorie», à la page 125.
`()`	Un délimiteur d'expression. Toute expression entre parenthèses est évaluée en premier.
`+`	Le connecteur du motif utilisé pour former un motif avec un ordre spécifique. Vous devez utiliser les crochets lorsqu'ils sont disponibles. Pour plus d'informations, voir «Utilisation des motifs TLA dans les règles de catégorie», à la page 123.

Tableau 19. Syntaxe prise en charge (suite)

Caractère	Description
[]	Le délimiteur de motifs est nécessaire si vous cherchez à effectuer des correspondances en fonction d'un motif TLA extrait à l'intérieur d'une règle de catégorie. Le contenu entre crochets fait référence aux motifs TLA et ne correspondra jamais à des concepts ou des types basés sur une simple co-occurrence. Si vous n'avez pas extrait ce motif TLA, aucune correspondance ne sera possible. Pour plus d'informations, voir «Utilisation des motifs TLA dans les règles de catégorie». N'utilisez pas les crochets si vous voulez établir des correspondances avec des concepts et des types et non avec des motifs. *Remarque* : dans les versions précédentes, les règles de synonymes et de co-occurrence générées par les techniques de création de catégorie étaient placées entre crochets. Dans toutes les nouvelles versions, les crochets indiquent désormais la présence d'un motif TLA. Les règles produites par une technique de co-occurrence et des synonymes sont maintenant entourées de parenthèses, comme dans `(enceintes acoustiques|enceintes)`.

Les opérateurs `&` et `|` sont commutatifs ; par conséquent `a & b = b & a` et `a | b = b | a`.

Echappement de caractères avec une barre oblique inverse

Si un concept contient un caractère qui est également un caractère de syntaxe, vous devez placer une barre oblique inverse devant ce caractère afin que la règle soit correctement interprétée. La barre oblique inverse (\) permet d'ignorer des caractères qui autrement auraient une signification particulière. Lorsque vous le faites glisser dans l'éditeur, la barre oblique inverse est automatiquement ajoutée.

Les caractères de syntaxe de règle suivants doivent être précédés d'une barre oblique inverse si vous souhaitez qu'ils soient traités tels qu'ils sont plutôt que comme une syntaxe de règle :
`& ! | + < > ( ) [ ] *`

Par exemple, le concept `r&d` contenant l'opérateur "et" `&`), la barre oblique inverse est requise lorsque cette chaîne est saisie dans l'éditeur de règle, comme dans : `r\&d`.

Utilisation des motifs TLA dans les règles de catégorie

Les motifs d'analyse des liens du texte peuvent être précisément définis en règles de catégorie afin d'obtenir des résultats encore plus précis et contextuels. Lorsque vous définissez un motif dans une règle de catégorie, vous ignorez les résultats d'extraction de concept les plus simples et vous utilisez uniquement les documents et les enregistrements correspondants basés sur les résultats des motifs d'analyse des liens du texte extraits.

Important ! Afin de mettre en correspondance des documents à l'aide de motifs TLA dans vos règles de catégorie, vous devez avoir effectué une extraction avec l'analyse des liens du texte activée. La règle de catégorie recherchera les correspondances trouvées pendant l'extraction. Si vous n'avez pas choisi d'explorer les résultats TLA dans l'onglet Modèle de votre noeud Text Mining, vous pouvez choisir d'activer l'extraction TLA dans les paramètres d'extraction pendant la session interactive puis effectuer une nouvelle extraction. Pour plus d'informations, voir «Extraction de données», à la page 82.

Délimitation par crochets. Un motif TLA doit être entouré de crochets `[ ]` si vous l'utilisez à l'intérieur d'une règle de catégorie. Le délimiteur de motifs est nécessaire si vous cherchez à effectuer des correspondances en fonction d'un motif TLA extrait. Puisque les catégories peuvent contenir des types, des concepts ou des motifs, les crochets expliquent à la règle que le contenu entre crochets fait référence au motif TLA extrait. Si vous n'avez pas extrait ce motif TLA, aucune correspondance ne sera possible. Si vous voyez un motif sans crochets, comme `pomme + bonne` dans la sous-fenêtre Catégories, cela signifie que le motif a été ajouté directement à la catégorie en-dehors de l'éditeur de règle de catégorie. Par exemple, si vous ajoutez un motif de concept directement à la catégorie à partir de la vue Analyse des

liens du texte , il n'apparaît pas entre crochets. Mais, lorsqu'un motif est utilisé à l'intérieur d'une règle de catégorie, vous devez placer le motif entre crochets dans la règle de catégorie, comme dans `[banane + !(bon)]`.

Utilisation du signe + dans les motifs. Dans IBM SPSS Modeler Text Analytics, vous pouvez avoir des motifs ayant jusqu'à 6 parties (ou propriétés). Pour indiquer que l'ordre est important, utilisez le signe + pour connecter chaque élément, comme dans `[entreprise1 + a acheté + entreprise2]`. Ici l'ordre est important car il modifierait laquelle des deux entreprises est l'entreprise acquéreuse. L'ordre n'est pas déterminé par la structure de la phrase mais par la structure de la sortie du motif TLA. Par exemple, imaginons le texte "*J'aime Paris*" et que vous souhaitiez extraire cette idée, alors le motif a de fortes chances d'être `[paris + aime]` or `[<Emplacement> + <Positif>]` plutôt que `[<Positif> + <Emplacement>]` car les ressources d'opinion par défaut placent généralement les opinions en deuxième position dans les motifs à deux parties. Afin d'éviter les problèmes, il peut donc être nécessaire d'utiliser directement le motif comme descripteur dans votre catégorie. Mais si vous avez besoin d'utiliser un patron dans une instruction plus complexe, faites particulièrement attention à l'ordre des éléments dans les patrons présentés dans la vue Analyse des liens du texte car l'ordre est très important pour trouver des correspondances.

Par exemple, si vous avez les deux textes suivants : "J'*aime l'ananas*" et "Je déteste *l'ananas*. Néanmoins, j'*aime* les fraises". L'expression `aime & l'ananas` correspondrait aux deux textes car il s'agit de l'expression d'un concept et non pas d'une règle des liens du texte (elle n'est pas entre crochets). L'expression `ananas + aime` correspond uniquement à "J'*aime l'ananas*" car dans le deuxième texte, le terme *aime* est associé à *fraises* à la place.

Regroupement des motifs. Vous pouvez simplifier les règles avec vos propres motifs. Imaginons que vous souhaitez capturer les trois expressions suivantes, `poivre de cayenne + aime`, `poivre blanc + aime`, et`poivre + aime`. Vous pouvez les regrouper en une seule règle de catégorie comme `[poivre * & aime]`. Si vous avez une autre expression `poivre noir + bon`, vous pouvez regrouper les quatre en une règle comme `[poivre * + <Positive>]`.

Ordre dans les motifs. Afin de mieux organiser les résultats, les règles d'analyse des liens du texte fournies dans les modèles installés avec votre produits tentent de générer des motifs de base dans le même ordre quel que soit l'ordre des mots dans une phrase. Par exemple, si vous avez un enregistrement contenant le texte "*Bonne présentations.*" et un autre enregistrement contenant le texte "*la présentation était bonne*", les deux textes sont pris en compte par la même règle et sortent dans le même ordre que `présentation + bonne` dans les résultats du motif de concept plutôt que `présentation + bonne` et également `bonne + présentation`. Et dans les motifs à deux propriétés comme dans ceux de cet exemple, les concepts affectés aux types dans la bibliothèque Opinions seront présentés en dernier dans les résultats par défaut comme dans `pomme + mauvaise`.

Tableau 20. Syntaxe de motifs et utilisation booléenne

Expression	Correspond à un document ou à un enregistrement qui
`[ ]`	Contient tous les motifs TLA. Le délimiteur de motifs est nécessaire *dans les règles de catégorie* si vous cherchez à effectuer des correspondances en fonction d'un motif TLA extrait. Le contenu entre crochets fait référence aux motifs TLA et pas à des concepts ou types simples. Si vous n'avez pas extrait ce motif TLA, aucune correspondance ne sera possible. Pour créer une règle n'incluant aucun motif, utilisez `!([ ])`.
`[a]`	Contient un motif dont au moins un des éléments est `a` quelle que soit sa position dans le motif. Par exemple, `[affaire]` peut correspondre à `[affaire + bonne]` ou seulement à `[deal + .]`

Tableau 20. Syntaxe de motifs et utilisation booléenne (suite)

Expression	Correspond à un document ou à un enregistrement qui
`[a + b]`	Contient un motif de concept. Par exemple, `[affaire + bonne]`. *Remarque* : si vous souhaitez seulement capturer ce motif sans ajouter d'autre élément, nous vous recommandons d'ajouter le motif directement à votre catégorie plutôt que d'en faire une règle.
`[a + b + c]`	Contient un motif de concept. Le signe + indique que l'ordre des éléments correspondants est important. Par exemple, `[entreprise1 + a acheté + entreprise2]`.
`[<A> + <B>]`	Contient tout motif avec le type `<A>` comme première propriété et le type `<B>` comme deuxième propriété et il y a exactement deux propriétés. Le signe + indique que l'ordre des éléments correspondants est important. Par exemple, `[<Budget> + <Negative>]`. *Remarque* : si vous souhaitez seulement capturer ce motif sans ajouter d'autre élément, nous vous recommandons d'ajouter le motif directement à votre catégorie plutôt que d'en faire une règle.
`[<A> & <B>]`	Contient n'importe quel motif de type `<A>` et de type `<B>`. Par exemple, `[<Budget> & <Négatif>]`. Ce motif TLA ne sera jamais extrait, mais lorsqu'il est écrit ainsi, il est vraiment égal à `[<Budget> + <Négatif>]|[<Négatif> + <Budget>]`. L'ordre des éléments mis en correspondance n'est pas important. De plus, le motif peut contenir d'autres éléments mais il doit au moins contenir `<Budget>` et `<Négatif>`.
`[a + .]`	Contient un motif où `a` est le seul concept et où les autres propriétés de ce motif sont vides. Par exemple, `[affaire + .]` correspond au motif de concept dans lequel le seul résultat est le concept `affaire`. Si vous avez ajouté le concept `affaire` comme descripteur de catégorie, vous obtiendrez tous les enregistrements avec le concept affaire, y compris les instructions positives sur une affaire. Mais, l'utilisation de `[affaire + .]` ne renverra que les résultats de motifs d'enregistrements qui représentent `affaire` et aucune autre relation ou opinion et ne renverra pas `affaire + fantastique`. *Remarque* : si vous souhaitez seulement capturer ce motif sans ajouter d'autre élément, nous vous recommandons d'ajouter le motif directement à votre catégorie plutôt que d'en faire une règle.
`[<A> + <>]`	Contient un motif où `<A>` est le seul type. Par exemple, `[<Budget> + <>]` correspond au motif dans lequel le seul résultat est un concept de type `<Budget>`. *Remarque* : vous pouvez utiliser `<>` pour dénoter un type vide uniquement en le plaçant après le symbole + dans le motif type comme `[<Budget> + <>]` mais pas `[prix + <>]`. *Remarque* : si vous souhaitez seulement capturer ce motif sans ajouter d'autre élément, nous vous recommandons d'ajouter le motif directement à votre catégorie plutôt que d'en faire une règle.
`[a + !(b)]`	Contient au moins un motif incluant le concept `a` mais pas le concept `b`. Doit inclure au moins un motif. Par exemple, `[price + !(high)]` ou pour les types, `[!(<Fruits>|<Légumes>) + <Positif>]`
`!([<A> & <B>])`	Ne contient pas de motifs spécifiques. Par exemple, `!([<Budget> & <Negative>])`.

Remarque : vous trouverez des exemples de règles correspondant à du texte à la rubrique «Exemples de règles de catégorie», à la page 127

Utilisation de caractères génériques dans les règles de catégorie

Les caractères génériques peuvent être ajoutés dans les règles afin d'étendre les capacités de mise en correspondance. Le caractère générique * (astérisque) peut être placé avant et/ou après un mot pour indiquer la façon dont les concepts peuvent être mis en correspondance. Deux types de caractères génériques sont disponibles :

- **Les caractères génériques d'affixe**. Ces caractères génériques sont placés en suffixe ou en préfixe sans espace entre la chaîne et l'astérisque. Par exemple, `opérat*` pourrait renvoyer *opérateur*, *opération*, *opérations*, *opérationnel*, *opérationnelles*, etc.
- **Les caractères génériques nominaux**. Ces caractères génériques sont placés en suffixe ou en préfixe d'un concept avec un espace entre le concept et l'astérisque. Par exemple, `opération *` pourrait renvoyer *opération*, *opération chirurgicale*, *opération mathématique*, etc. De plus, un caractère générique nominal peut être utilisé avec un caractère générique affixe, comme par exemple `* opérat* *`, qui pourrait renvoyer *opération*, *opération chirurgicale*, *opérateur téléphonique*, *post opératoire*, etc. Comme cet exemple l'illustre, nous vous recommandons d'utiliser les caractères génériques avec prudence afin que la recherche ne soit pas trop large et ne capture pas de correspondances indésirables.

Exceptions !

- Un caractère générique ne peut jamais être utilisé seul. Par exemple, `(pomme | * )` n'est pas accepté.
- Un caractère générique ne peut jamais être utilisé pour mettre en correspondance des noms de type. `<Negative*>`ne sera pas mis en correspondance avec des noms de type.
- Vous ne pouvez pas éviter que certains types soient mis en correspondance avec des concepts à l'aide de caractères génériques. Le type auquel le concept est affecté est automatiquement utilisé.
- Un caractère générique ne peut pas être placé au milieu d'une séquence de mots, qu'il soit ajouté à la fin d'un mot (ouv* compte) ou qu'il soit utilisé seul (ouvrir * compte). Vous ne pouvez pas non plus utiliser les caractères génériques dans les noms de type. Par exemple, `mot* mot`, tel que `pot* recette`, ne correspondra pas à une recette de potage ni à rien d'autre. Toutefois, `pomme* *` correspondra à *pomme de terre*, *pomme vapeur*, *pomme* etc. Dans un autre exemple, `mot * mot`, comme `gâteau * pomme`, ne correspondra pas à *gâteau cannelle pomme* ni à rien d'autre car l'astérisque apparaît entre deux mots. Cependant, `pomme *` correspondra à *pomme de terre*, *pomme*, *pomme vapeur* etc.

Tableau 21. Utilisation des caractères génériques

Expression	Correspond à un document ou à un enregistrement qui
`*fixe`	Contient un concept qui se termine par une lettre mais qui peut contenir n'importe quel nombre de lettres comme préfixe. Par exemple, `*fixe` se termine par les lettres *fixe* mais peut prendre un préfixe tel que : - `fixe` - `préfixe` - `suffixe`
`fixe*`	Contient un concept qui commence par des lettres mais qui peut contenir n'importe quel nombre de lettres comme suffixe. Par exemple, `fixe*` commence par les lettres *fixe* mais peut prendre ou non un suffixe : - `fixe` - `fixette` - `fixement` Par exemple, `pomme* & !(poire* | coing)`, qui contient un concept commençant par les lettres `pomme` et non un concept commençant par les lettres *poire* ou le concept `coing`, ne correspondrait pas à : `pomme & coing` mais correspondrait à : - `pomme d'api` - `pomme & orange`
`*vers*`	Contient un concept qui comprend les lettres `vers`, mais qui peut avoir n'importe quel nombre de lettres comme préfixe, suffixe ou les deux. Par exemple : `*vers*` pourrait correspondre à : - `vers` - `revers` - `verseau`

Tableau 21. Utilisation des caractères génériques (suite)

Expression	Correspond à un document ou à un enregistrement qui

`* terme`	Contient un concept qui comprend le mot `terme` mais peut être un composé avec un autre mot placé devant. Par exemple, `* terme` pourrait renvoyer :
	`- terme`
	`- long terme`
	`- court terme`
	
	Par exemple, `[* livraison + <Negative>]` contient un concept qui se termine par le mot `argent` et contient un type `<Negative>` et pourrait renvoyer les motifs de concept suivants :
	`- livraison de module + lent`
	`- livraison le lendemain + retard`

`voiture *`	Contient un concept qui comprend le mot `voiture` mais qui peut être un composé suivi d'un autre mot. Par exemple, `voiture *` pourrait renvoyer :
	`- voiture`
	`- voiture de location`
	`- voiture de fonction`

`* pomme *`	Contient un concept qui peut commencer par n'importe quel mot suivi du mot `pomme` et suivi d'un autre mot. `*` signifie 0 ou n caractères, donc cela correspond aussi au mot `pomme`. Par exemple, `* pomme *` pourrait renvoyer :
	`-  gelée de pomme`
	`- gâteau à la pomme granny smith`
	`- pomme au four`
	`- fixe`
	
	Par exemple, `[* réservation* * + <Positive>]` qui contient un concept avec le mot `réservation` (quel que soit l'endroit où il se trouve dans le concept) en première position et un type `<Positive>` en deuxième position, pourrait renvoyer les motifs de concept :
	`- réservation en ligne + bon`
	`- système de réservation + bon`

Remarque : vous trouverez des exemples de règles correspondant à du texte à la rubrique «Exemples de règles de catégorie»

Exemples de règles de catégorie

Pour mieux comprendre en quoi les correspondances entre les règles et les enregistrements sont différentes selon la syntaxe utilisée pour les exprimer, étudiez l'exemple suivant.

Exemples d'enregistrements

Imaginons deux enregistrements :

- **Enregistrement A** : *"lorsque j'ai vérifié mon porte-monnaie, il manquait 5 dollars."*
- **Enregistrement B**: *"on a trouvé 5$ sur l'aire de pique-nique, mais il manquait la couverture."*

Les deux tableaux suivants montrent ce qui peut être extrait pour les concepts et les types ainsi que les motifs de concept et les motifs de type.

Concepts et types extraits de l'exemple

Tableau 22. Exemple de concepts et de types extraits

Concept extrait	Concepts saisis comme
`porte-monnaie`	`<Inconnu>`
`manquer`	`<Négatif>`

Tableau 22. Exemple de concepts et de types extraits (suite)

Concept extrait	Concepts saisis comme
5$	<Devise>
couverture	<Inconnu>
aire de pique-nique	<Inconnu>

Motifs TLA extraits de l'exemple

Tableau 23. Exemple de sortie de motif TLA extraite

Motifs de concepts extraits	Motifs de types extraits	De l'enregistrement
aire de pique-nique + .	<Inconnu> + <>	Enregistrement B
porte-monnaie + .	<Inconnu> + <>	Enregistrement A
couverture + manquer	<Inconnu> + <Négatif>	Enregistrement B
5 dollars + .	<Devise> + <>	Enregistrement B
5 dollars + manquer	<Devise> + <Négatif>	Enregistrement A

Comment mettre en correspondance les règles de catégorie possibles

Le tableau suivant contient une syntaxe qui pourrait être entrée dans l'éditeur de règle de catégorie. Toutes les règles qui se trouvent ici ne fonctionnent pas et toutes ne correspondent pas aux mêmes enregistrements. Regardez la façon dont les différentes syntaxes ont un impact sur les enregistrements mis en correspondance.

Tableau 24. Règles concernant les échantillons

Syntaxe de règle	Résultat
5 dollars & manquer	Correspond à la fois à A et à B car ils contiennent tous les deux le concept extrait manquer et le concept extrait 5$. Cela revient à : (5 dollars & manquer)
manquer & 5 dollars	Correspond à la fois à A et à B car ils contiennent tous les deux le concept extrait manquer et le concept extrait 5$. Cela revient à : (manquer & 5 dollars)
manquer & <Devise>	Correspond à la fois à A et à B car ils contiennent tous les deux le concept extrait manquer et un concept correspondant au type <Devise>. Cela revient à : (manquer & <Devise>)
<Devise> & manquer	Correspond à la fois à A et à B car ils contiennent tous les deux le concept extrait manquer et un concept correspondant au type <Devise>. Cela revient à : (<Devise> & manquer)
[5 dollars + manquer]	Correspond à A mais **pas** à B car l'enregistrement B n'a pas produit de sortie de motif B contenant 5 dollars + manquer (voir le tableau précédent). Cela revient à la sortie de motif TLA : 5 dollars + manquer
[manquer + 5 dollars]	Ne correspond ni à l'enregistrement A ni au B car aucun motif TLA extrait (voir tableau précédent) ne correspond à l'ordre exprimé ici avec manquer en première position. Cela revient à la sortie de motif TLA : 5 dollars + manquer

Tableau 24. Règles concernant les échantillons (suite)

Syntaxe de règle	Résultat
`[manquer & 5 dollars]`	Correspondance de A mais **non** de B car aucun canevas TLA correspondant n'a été extrait de l'enregistrement B. L'utilisation du caractère `&` indique que l'ordre n'a pas d'importance dans la mise en correspondance ; par conséquent, cette règle recherche une correspondance avec le canevas `[manquer + USD5]` ou `[USD5 + manquer]`. Seul `[5 dollars + manquer]` de l'enregistrement A possède une correspondance.
`[manquer + <Devise>]`	Ne correspond ni à l'enregistrement A ni à l'enregistrement B car aucun motif TLA extrait ne correspondait à cet ordre. Cela n'a pas d'équivalent car une sortie TLA n'est basée que sur les termes (`5 dollars + manquer`) ou sur les types (`<Devise> + <Négatif>`), mais ne mélange pas les types et les concepts.
`[<Devise> + <Négatif>]`	Correspond à l'enregistrement A mais pas à l'enregistrement B car aucun patron TLA n'a été extrait de l'enregistrement B. Cela équivaut à la sortie TLA : `<Devise> + <Négatif>`
`[<Négatif> + <Devise>]`	Ne correspond ni à l'enregistrement A ni à l'enregistrement B car aucun motif TLA extrait ne correspondait à cet ordre. Par défaut, dans le modèle `Opinions` lorsqu'une *rubrique* est trouvée avec une *opinion*, cette *rubrique* (`<Devise>`) occupe la première position et *opinion* (`<Négatif>`) occupe la deuxième.

Création de règles de catégorie

Lorsque vous créez ou modifiez une règle de catégorie, elle doit être ouverte dans la sous-fenêtre de l'éditeur de règles. Vous pouvez ajouter des concepts, des types ou des motifs et utiliser des caractères génériques pour étendre les correspondances. Lorsque vous utilisez des concepts, des types ou des motifs reconnus, vous pouvez bénéficier de la recherche de tous les concepts associés. Par exemple, lorsque vous utilisez un concept, tous ses termes associés, ses formes plurielles et ses synonymes sont également associés à cette règle. De même, lorsque vous utilisez un type, tous ses concepts sont également capturés par la règle.

Vous pouvez ouvrir l'éditeur de règles en modifiant une règle existante ou en faisant un clic droit sur le nom de la catégorie et en choisissant **Créer une règle**.

Vous pouvez utiliser les menus contextuels, la méthode glisser-déposer ou saisir manuellement les concepts, les types et les motifs dans l'éditeur. Combinez-les avec des opérateurs booléens (`&`, `!()`, `|`) et des crochets pour former des expressions de règles. Pour éviter les erreurs communes, nous vous recommandons de déplacer les concepts en les faisant glisser directement de la sous-fenêtre Résultats d'extraction ou de la sous-fenêtre Données vers l'éditeur de règle. Soyez attentif à la syntaxe des règles afin d'éviter les erreurs. Pour plus d'informations, voir «Syntaxe des règles de catégorie», à la page 121.

Remarque : vous trouverez des exemples de règles correspondant à du texte à la rubrique «Exemples de règles de catégorie», à la page 127.

Pour créer une règle

1. Si vous n'avez pas encore extrait de données ou que votre extraction a expiré, faites-le maintenant. Pour plus d'informations, voir «Extraction de données», à la page 82.

   Remarque : si vous filtrez une extraction de sorte que plus aucun concept ne soit visible, un message d'erreur s'affiche lorsque vous tentez de créer ou d'éditer une règle de catégorie. Vous devez alors modifier le filtre d'extraction de sorte que des concepts soient disponibles.

2. Dans la sous-fenêtre Catégories, sélectionnez la catégorie dans laquelle vous voulez ajouter votre règle.
3. Dans les menus, choisissez **Catégories > Créer une règle**. La sous-fenêtre de l'éditeur des règles de catégorie apparaît dans la fenêtre.
4. Dans le champ Nom de règle, entrez un nom pour votre règle. Si vous ne donnez pas de nom à la règle, l'expression sera automatiquement utilisée comme nom. Vous pourrez renommer cette règle ultérieurement.
5. Dans le plus grand champ de texte de l'expression, vous pouvez effectuer les opérations suivantes :
   - Saisir directement le texte dans le champ ou le faire glisser depuis unz autre sous-fenêtre. Utilisez uniquement des concepts, types et motifs extraits. Par exemple, si vous entrez le mot `chats`, mais que seule la forme au singulier, `chat`, apparaît dans la sous-fenêtre Résultats d'extraction, l'éditeur ne sera pas en mesure de reconnaître `chats`. Dans ce cas, la forme singulière pourra inclure automatiquement la forme plurielle, ou vous pourrez utiliser un caractère générique. Pour plus d'informations, voir «Syntaxe des règles de catégorie», à la page 121.
   - Sélectionner les concepts, types ou motifs à ajouter aux règles et utiliser les menus.
   - Ajouter des opérateurs booléens pour relier les éléments de votre règle. Utilisez les boutons de la barre d'outils pour ajouter l'opérateur booléen "and" **&**, "or" **|** ou "not" **!()**, des parenthèses **()** et des crochets de motifs **[]** dans votre règle.
6. Cliquez sur le bouton **Tester la règle** pour vérifier que votre règle est bien constituée. Pour plus d'informations, voir «Syntaxe des règles de catégorie», à la page 121. Le nombre de documents ou d'enregistrements trouvés apparaît entre parenthèses en regard du texte **Résultats de test**. Les éléments de la règle qui ont été reconnus ou les messages d'erreur éventuels apparaissent à droite de ce texte. Si le graphique à côté du type, du motif ou du concept apparaît avec un point d'interrogation rouge, ceci indique que l'élément ne correspond à aucune extraction connue. S'il n'y a pas de correspondance, la règle ne trouvera aucun enregistrement.
7. Pour tester une partie de votre règle, sélectionnez-la, puis cliquez sur **Tester la sélection**.
8. En cas de problème, apportez toutes les modifications nécessaires et testez à nouveau la règle.
9. Lorsque vous avez terminé, cliquez sur **Enregistrer & Fermer** pour enregistrer à nouveau la règle et fermer l'éditeur. Le nouveau nom de la règle apparaît dans la catégorie.

Modification et suppression des règles

Vous pouvez éditer à tout moment une règle que vous avez créée et enregistrée. Pour plus d'informations, voir «Syntaxe des règles de catégorie», à la page 121.

Si une règle ne vous est plus utile, vous pouvez la supprimer.

Pour modifier des règles

1. Dans le tableau Descripteurs de la boîte de dialogue Définitions de catégorie, sélectionnez la règle.
2. Dans les menus, choisissez **Catégories > Editer la règle** ou double-cliquez sur le nom de la règle. L'éditeur s'ouvre, avec la règle sélectionnée.
3. Modifiez la règle à l'aide des résultats de l'extraction et des boutons de la barre d'outils.
4. Testez à nouveau la règle pour vous assurer qu'elle renvoie les résultats attendus.
5. Cliquez sur **Enregistrer & Fermer** pour enregistrer à nouveau la règle et fermer l'éditeur.

Pour supprimer une règle

1. Dans le tableau Descripteurs de la boîte de dialogue Définitions de catégorie, sélectionnez la règle.
2. Dans les menus, sélectionnez **Edition > Supprimer**. La règle est supprimée de la catégorie.

Import et export de catégories prédéfinies

Si vous avez vos propres catégories stockées dans un fichier Microsoft Excel (*.xls, *.xlsx), vous pouvez les importer dans IBM SPSS Modeler Text Analytics .

Vous pouvez également exporter des catégories depuis une session de plan de travail interactif ouverte vers un fichier Microsoft Excel (*.xls, *.xlsx). Lorsque vous exportez vos catégories, vous pouvez choisir d'inclure ou d'exclure des informations supplémentaires telles que les descripteurs et les scores. Pour plus d'informations, voir «Exportation de catégories», à la page 135.

Si vos catégories prédéfinies n'ont pas de code ou que vous souhaitez de nouveaux codes, vous pouvez générer automatiquement un nouvel ensemble de codes pour l'ensemble de catégories, dans la sous-fenêtre des catégories en choisissant **Catégories > Gestion des catégories > Générer automatiquement des codes** à partir des menus. Tous les codes existants seront supprimés et renumérotés automatiquement.

Importation de catégories prédéfinies

Vous pouvez importer des catégories prédéfinies dans IBM SPSS Modeler Text Analytics . Avant l'importation, assurez-vous que le fichier de la catégorie prédéfinie est un fichier Microsoft Excel (*.xls, *.xlsx) et qu'il est structuré dans un format pris en charge. Vous pouvez également choisir de laisser le produit détecter le format pour vous. Les formats suivants sont pris en charge :

- **Format liste à plat** : pour plus d'informations, voir «Format liste plate», à la page 132.
- **Format compact** : pour plus d'informations, voir «Format compact», à la page 133.
- **Format indenté** : pour plus d'informations, voir «Format indenté», à la page 134.

Pour importer des catégories prédéfinies

1. A partir des menus du plan de travail interactif, sélectionnez **Catégories > Gérer les catégories > Importer les catégories prédéfinies**. Un assistant d'importation des catégories prédéfinies apparaît.
2. Dans la liste déroulante Rechercher dans, sélectionnez l'unité et le dossier dans lesquels se trouve le fichier.
3. Sélectionnez le fichier dans la liste. Le nom du fichier apparaît dans la zone de texte Nom de fichier.
4. Sélectionnez dans la liste, la feuille de calcul contenant les catégories prédéfinies. Le nom de la feuille de calcul apparaît dans le champ Feuille de calcul.
5. Pour accéder aux options de format des données, cliquez sur **Suivant**.
6. Choisissez le format de votre fichier, ou choisissez l'option permettant au produit de détecter automatiquement le format. La détection automatique est la plus efficace sur les formats les plus courants.
   - **Format liste à plat** : pour plus d'informations, voir «Format liste plate», à la page 132.
   - **Format compact** : pour plus d'informations, voir «Format compact», à la page 133.
   - **Format indenté** : pour plus d'informations, voir «Format indenté», à la page 134.
7. Pour accéder aux options d'importation supplémentaires, cliquez sur **Suivant**. Si vous avez choisi la détection automatique du format, vous serez orienté directement vers l'étape finale.
8. Si une ou plusieurs lignes contiennent des titres de colonnes ou d'autres types d'information, choisissez le numéro de la ligne à partir de laquelle vous voulez commencer l'importation dans l'option **Commencer l'import à la ligne**. Par exemple : si vos noms de catégories commencent à la ligne 7, vous devez entrer le chiffre 7 dans cette option pour importer correctement votre fichier.
9. Si votre fichier contient des codes de catégories, choisissez l'option **Contient des codes de catégories**. Ainsi, il sera plus facile à l'assistant de reconnaître correctement vos données.
10. Consultez les cellules de codage couleur et la légende afin de vous assurer que les données ont été correctement identifiées. Toute erreur détectée dans le fichier est affichée en rouge et référencée sous le tableau d'aperçu de format. Si le format sélectionné n'est pas le bon, revenez en arrière et

choisissez-en un autre. Si vous devez apporter des corrections à votre fichier, effectuez vos changements et redémarrez l'assistant en sélectionnant de nouveau ce fichier. Vous devez corriger toutes les erreurs avant de fermer l'assistant.

11. Pour voir l'ensemble des catégories et sous-catégories qui seront importées et pour définir la manière de créer des descripteurs pour ces catégories, cliquez sur **Suivant**.
12. Consultez l'ensemble des catégories qui seront importées dans le tableau. Si les mots clés que vous avez définis comme descripteurs n'apparaissent pas, il se peut qu'ils n'aient pas été reconnus durant l'importation. Assurez-vous qu'ils ont été correctement préfixés et qu'ils apparaissent dans la bonne cellule.
13. Choisissez la manière dont vous voulez gérer les catégories pré-existantes dans votre session.
    - **Remplacer toutes les catégories existantes**. Cette option purge toutes les catégories existantes et les catégories nouvellement importées sont alors utilisées seules à leur place.
    - **Ajouter aux catégories existantes**. Cette option vous permettra d'importer les catégories et de les fusionner avec les catégories déjà existantes. En ajoutant des catégories à des catégories déjà existantes, vous devez déterminer comment vous voulez que les doublons soient traités. L'option **Fusionner** permet de fusionner une catégorie importée dans une catégorie existante portant le même nom. L'option **Exclure de l'importation** permet d'empêcher l'importation d'une catégorie lorsqu'il en existe déjà une du même nom.
14. **Importer les mots clés en tant que descripteurs** est une option de descripteur, qui importe les mots-clés identifiés dans vos données en tant que descripteurs pour la catégorie associée.
15. **Etendre les catégories à partir des descripteurs** est une option générant des descripteurs à partir de mots qui représentent le nom de la catégorie ou de la sous-catégorie, et/ou à partir de mots qui forment l'annotation. Si ces mots correspondent aux résultats extraits, alors ils seront ajoutés en tant que descripteurs de la catégorie. Cette option produit de meilleurs résultats lorsque les noms de catégories sont à la fois longs et descriptifs. C'est une méthode rapide pour générer des descripteurs de catégories, qui à leur tour permettent à la catégorie de capturer les enregistrements contenant ces descripteurs.
    - Le champ **A partir de** vous permet de choisir depuis quel texte les descripteurs seront choisis, les noms ou les catégories ou sous-catégories, les mots dans les annotations, ou les deux.
    - Le champ **En tant que** vous permet de créer ces descripteurs sous la forme de concepts ou de motifs TLA. Si l'extraction TLA n'a pas eu lieu, les options de **motifs** de cet assistant sont désactivés.
16. Pour importer les catégories prédéfinies dans la sous-fenêtre Catégories, cliquez sur **Terminer**.

Format liste plate

Dans le format liste plate, il y a un niveau de catégories unique sans hiérarchie, ce qui signifie qu'il n'y a pas de sous-catégories ni de sous-réseaux. Les noms de catégories sont dans une seule colonne.

Les informations suivantes peuvent être contenues dans un fichier de ce format :

- La colonne des **codes** optionnels contient des valeurs numériques qui identifient chaque catégorie de façon unique. Si vous précisez que les fichiers de données contiennent des codes (option **Contient des codes de catégories** dans l'étape **Paramètres de contenu**), alors une colonne contenant des codes uniques pour chaque catégorie doit exister dans la cellule située à gauche du nom de la catégorie. Si vos données ne contiennent pas de codes, mais que vous voulez créer des codes plus tard, vous pourrez toujours le faire en utilisant l'option **Catégories > Gestion des catégories > Générer automatiquement des codes**.
- Une colonne *obligatoire* **nom des catégories** contient tous les noms de catégories. Cette colonne est nécessaire pour importer en utilisant ce format.
- Des **annotations** optionnelles sont disponibles dans la cellule située à droite du nom de catégorie. Cette annotation consiste en un texte décrivant vos catégories ou sous-catégories.
- Des **mots clés** optionnels peuvent être importés en tant que descripteurs de catégories. Afin qu'ils soient reconnus, ces mots clés doivent être placés dans la cellule situé directement sous le nom de la catégorie/sous-catégorie associée. La liste des mots clés doit également être préfixée d'un caractère de

soulignement (_) par exemple « _armesàfeu, armes / revolvers ». La cellule de mot clé peut contenir un ou plusieurs mots pour décrire chaque catégorie. Ces mots seront importés en tant que descripteurs, ou ignorés, en fonction de ce que vous aurez spécifié durant la dernière étape de l'assistant. Plus tard, les descripteurs sont comparés aux résultats extraits du texte. Si une correspondance est détectée, le document ou l'enregistrement est attribué à la catégorie contenant ce descripteur.

Tableau 25. Format liste plate avec codes, mots clés et annotations

Colonne A	**Colonne B**	**Colonne C**
Code de catégorie (*facultatif*)	Nom de catégorie	Annotation
	Liste Descripteur/Mot clé (*facultatif*)	

Format compact

Ce format compact est similaire au format liste plate, mis à part le fait qu'il est utilisé avec les catégories hiérarchiques. Par conséquent, une colonne de niveau de code est nécessaire afin de déterminer le niveau hiérarchique de chaque catégorie et sous-catégorie.

Les informations suivantes peuvent être contenues dans un fichier de ce format :

- Une colonne *obligatoire* de **niveau de code** contient les nombres indiquant la position hiérarchique des informations ultérieures pour cette ligne. Par exemple, si les valeurs 1, 2 ou 3 sont spécifiées et que vous avez à la fois des catégories et des sous-catégories, alors 1 correspond aux catégories, 2 aux sous-catégories et 3 aux sous-sous-catégories. Si vous avez uniquement des catégories et des sous-catégories, 1 correspond aux catégories et 2 correspond aux sous-catégories. Et ainsi de suite, jusqu'à la profondeur de catégories souhaitée.
- La colonne optionnelle des **codes** contient des valeurs numériques qui identifient chaque catégorie de façon unique. Si vous précisez que les fichiers de données contiennent des codes (option **Contient des codes de catégories** dans l'étape **Paramètres de contenu**), alors une colonne contenant des codes uniques pour chaque catégorie doit exister dans la cellule située à gauche du nom de la catégorie. Si vos données ne contiennent pas de codes, mais que vous voulez créer des codes plus tard, vous pourrez toujours le faire en utilisant l'option **Catégories > Gestion des catégories > Générer automatiquement des codes**.
- Une colonne *obligatoire* **nom des catégories** contient tous les noms de catégories et de sous-catégories. Cette colonne est nécessaire pour importer en utilisant ce format.
- Des **annotations** optionnelles sont disponibles dans la cellule située à droite du nom de catégorie. Cette annotation consiste en un texte décrivant vos catégories ou sous-catégories.
- Des **mots clés** optionnels peuvent être importés en tant que descripteurs de catégories. Afin qu'ils soient reconnus, ces mots clés doivent être placés dans la cellule situé directement sous le nom de la catégorie/sous-catégorie associée. La liste des mots clés doit également être préfixée d'un caractère de soulignement (_) par exemple « _armesàfeu, armes / revolvers ». La cellule de mot clé peut contenir un ou plusieurs mots pour décrire chaque catégorie. Ces mots seront importés en tant que descripteurs, ou ignorés, en fonction de ce que vous aurez spécifié durant la dernière étape de l'assistant. Plus tard, les descripteurs sont comparés aux résultats extraits du texte. Si une correspondance est détectée, le document ou l'enregistrement est attribué à la catégorie contenant ce descripteur.

Tableau 26. Exemple de format compact avec codes

Colonne A	**Colonne B**	**Colonne C**
Niveau de code hiérarchique	Code de catégorie (*facultatif*)	Nom de catégorie
Niveau de code hiérarchique	Code de sous-catégorie (*facultatif*)	Nom de sous-catégorie

Tableau 27. Exemple de format compact sans codes

Colonne A	**Colonne B**
Niveau de code hiérarchique	Nom de catégorie

Tableau 27. Exemple de format compact sans codes (suite)

Colonne A	Colonne B
Niveau de code hiérarchique	Nom de sous-catégorie

Format indenté

Dans le format de fichier indenté, le contenu est organisé de façon hiérarchique ; c'est-à-dire qu'il contient des catégories et un ou plusieurs niveaux de sous-catégories. De plus, sa structure est indentée pour refléter cette hiérarchie. Chaque ligne du fichier contient soit une catégorie, soit une sous-catégorie. Les sous-catégories sont indentées à partir des catégories, et toute sous-sous-catégorie est indentée à partir des sous-catégories, etc. Vous pouvez créer manuellement cette structure dans Microsoft Excel ou utiliser une structure exportée à partir d'un autre produit et enregistrée sous un format Microsoft Excel.

- Les **Codes et les noms de catégories de niveau supérieur** occupent respectivement les colonnes A et B. Ou, si aucun code n'est présent, alors le nom de catégorie occupe la colonne A.
- Les **codes de sous-catégories et les noms de sous-catégories** occupent respectivement les colonnes B et C. Ou, si aucun code n'est présent, alors le nom de sous-catégorie occupe la colonne B. La sous-catégorie est un membre d'une catégorie. Vous ne pouvez pas avoir de sous-catégories si vous n'avez pas de catégories.

Tableau 28. Structure indentée avec des codes

Colonne A	Colonne B	Colonne C	Colonne D
Code de catégorie (*facultatif*)	Nom de catégorie		
	Code de sous-catégorie (*facultatif*)	Nom de sous-catégorie	
		Code de sous-sous-catégorie (*facultatif*)	Nom de sous-sous-catégorie

Tableau 29. Structure indentée sans codes

Colonne A	Colonne B	Colonne C
Nom de catégorie		
	Nom de sous-catégorie	
		Nom de sous-sous-catégorie

Les informations suivantes peuvent être contenues dans un fichier de ce format :

- Les **codes** optionnels doivent être des valeurs identifiant de manière unique chaque catégorie ou sous-catégorie. Si vous précisez que les fichiers de données contiennent des codes (option **Contient des codes de catégories** dans l'étape **Paramètres de contenu**), alors un code unique pour chaque catégorie ou sous-catégorie doit exister dans la cellule située à gauche du nom de la catégorie/sous-catégorie. Si vos données ne contiennent pas de codes, mais que vous voulez créer des codes plus tard, vous pourrez toujours le faire en utilisant l'option **Catégories > Gestion des catégories > Générer automatiquement des codes**.
- Un **nom** est *obligatoire* pour chaque catégorie et sous-catégorie. Les sous-catégories doivent être indentées des catégories d'une cellule vers la droite et sur une ligne différente.
- Des **annotations** optionnelles sont disponibles dans la cellule située à droite du nom de catégorie. Cette annotation consiste en un texte décrivant vos catégories ou sous-catégories.
- Des **mots clés** optionnels peuvent être importés en tant que descripteurs de catégories. Afin qu'ils soient reconnus, ces mots clés doivent être placés dans la cellule situé directement sous le nom de la catégorie/sous-catégorie associée. La liste des mots clés doit également être préfixée d'un caractère de soulignement (_) par exemple « `_armesàfeu, armes / revolvers` ». La cellule de mot clé peut contenir un ou plusieurs mots pour décrire chaque catégorie. Ces mots seront importés en tant que descripteurs,

ou ignorés, en fonction de ce que vous aurez spécifié durant la dernière étape de l'assistant. Plus tard, les descripteurs sont comparés aux résultats extraits du texte. Si une correspondance est détectée, le document ou l'enregistrement est attribué à la catégorie contenant ce descripteur.

Important ! Si vous utilisez un code à un niveau, vous devez inclure un code pour chaque catégorie et sous-catégorie. Si cela n'est pas le cas, votre importation échouera.

Exportation de catégories

Vous pouvez également exporter des catégories depuis une session de plan de travail interactif vers un fichier Microsoft Excel (**.xls*, **.xlsx*). Les données qui seront exportées viennent essentiellement du contenu actuel de la sous-fenêtre des catégories ou des propriétés des catégories. Il est donc recommandé de scorer de nouveau, si vous prévoyez d'exporter également la valeur de score de **Docs.**.

Tableau 30. Options d'exportation de catégories

Toujours exporté...	Exporté de manière optionnelle...
• S'ils sont présents, les codes de catégories • Noms de catégories (et de sous-catégories) • S'ils sont présents, les niveaux de codes (Format *Plat/Compact*) • En-têtes de colonnes (Format *Plat/Compact*)	• **Docs.** scores • Annotations de catégories • Noms de descripteurs • Nombres de descripteurs

Important ! Lorsque vous exportez des descripteurs, ils sont convertis en chaînes de texte et un caractère de soulignement leur est ajouté comme préfixe. Si vous effectuez de nouveau une importation dans ce produit, la capacité de distinction entre les descripteurs qui sont des motifs, des règles de catégorie ou des concepts bruts est perdue. Si vous souhaitez réutiliser ces catégories dans ce produit, nous vous recommandons fortement de créer un fichier de pack d'analyse de texte (TAP), car ce format permet de conserver tous les descripteurs tels qu'ils sont définis, ainsi que toutes vos catégories, vos codes et les ressources linguistiques utilisées. Les fichiers TAP peuvent être utilisés à la fois dans IBM SPSS Modeler Text Analytics et dans IBM SPSS Text Analytics for Surveys . Pour plus d'informations, voir «Utilisation des packs d'analyse de texte», à la page 136.

Pour exporter des catégories prédéfinies

1. A partir des menus du plan de travail interactif, sélectionnez **Catégories > Gérer les catégories > Exporter des catégories**. Un assistant d'exportation des catégories apparaît.
2. Sélectionnez un emplacement et saisissez le nom du fichier à exporter.
3. Entrez le nom du fichier de sortie dans la zone de texte Nom du Fichier.
4. Pour choisir le format selon lequel les données de catégories seront exportées, cliquez sur **Suivant**.
5. Choisissez le format :
   - **Format de liste plat/compact** : pour plus d'informations, voir «Format liste plate», à la page 132. Une liste à plat ne contient pas de sous-catégorie. Pour plus d'informations, voir «Format compact», à la page 133. Une liste compacte contient des catégories hiérarchiques.
   - **Format indenté** : pour plus d'informations, voir «Format indenté», à la page 134.
6. Pour commencer à choisir le contenu à exporter et afficher les données proposées, cliquez sur **Suivant**.
7. Consulter le contenu du fichier exporté.
8. Sélectionnez ou désélectionnez les paramètres de contenu supplémentaire à exporter tels que les **annotations** ou les **noms de descripteur**.
9. Pour exporter les catégories, cliquez sur **Terminer**.

Utilisation des packs d'analyse de texte

Un pack d'analyse de texte, également appelé TAP (Text Analysis Package), sert en tant que modèle de catégorisations de réponses texte. Un TAP permet de catégoriser des données texte facilement et avec un minimum d'intervention de votre part, car il contient les ensembles de catégories préconfigurés et les ressources linguistiques nécessaires au codage rapide et automatique d'un grand nombre d'enregistrements. Les ressources linguistiques permettent d'analyser et d'exploiter les données textuelles pour en extraire des concepts clés. En fonction des concepts clés et des motifs trouvés dans le texte, les enregistrements peuvent être catégorisés dans l'ensemble de catégories sélectionné dans le TAP. Vous pouvez créer votre propre TAP ou en mettre un à jour.

Un TAP est composé des éléments suivants :

- **Ensemble(s) de catégories**. Un ensemble de catégories est principalement composé de catégories prédéfinies, de codes de catégorie, de descripteurs pour chaque catégorie et enfin d'un nom pour l'ensemble en son entier. Les descripteurs sont des éléments linguistiques (concepts, types, motifs et règles) comme le terme *cher* ou le motif *bon prix*. Les descripteurs sont utilisés pour définir une catégorie. Ainsi, lorsque le texte correspond à un descripteur de catégorie, le document ou l'enregistrement est placé dans cette catégorie.
- **Ressources linguistiques**. Les ressources linguistiques sont un ensemble de bibliothèques et de ressources avancées qui permettent d'extraire des concepts et motifs clés. Ces concepts et motifs d'extraction sont utilisés comme descripteurs qui permettent aux enregistrements d'être placés dans une catégorie de l'ensemble de catégories.

Vous pouvez créer votre propre TAP, en mettre un à jour ou charger des packs d'analyse de texte.

Une fois que vous avez sélectionné le TAP et choisi un ensemble de catégories, SPSS Modeler Text Analytics peut extraire et catégoriser vos enregistrements.

Remarque : Un TAP peut être créé et utilisé indifféremment entre IBM SPSS Text Analytics for Surveys et SPSS Modeler Text Analytics . Notez cependant que le scoring à partir de règles peut être différente dans SPSS Modeler Text Analytics suivant que vous chargiez directement un packs d'analyse de texte de SPSS Modeler Text Analytics ou que vous en chargiez un d'IBM SPSS Text Analytics for Surveys . Nous vous recommandons d'utiliser des packs d'analyse de texte créés dans SPSS Modeler Text Analytics ; en effet, ceux créés dans IBM SPSS Text Analytics for Surveys peuvent l'être à l'aide d'une version différente des ressources linguistiques.

Création des packs d'analyse de texte

Lorsque vous avez une session avec au moins une catégorie et des ressources, vous pouvez créer un pack d'analyse de texte (TAP) à partir du contenu de la session de plan de travail interactif. L'ensemble de catégories et de descripteurs (concepts, types, règles ou résultats de motifs TLA) peut être transformé en TAP utilisant toutes les ressources linguistiques ouvertes dans l'éditeur de ressources.

Vous pouvez voir la langue pour laquelle les ressource ont été créées. La langue est définie dans l'onglet Ressources avancées de l'Editeur de modèle ou de l'Editeur de ressources.

Pour créer un pack d'analyse de texte

1. Dans les menus, choisissez **Fichier > Packs d'analyse de texte > Créer un pack**. La boîte de dialogue Créer un pack apparaît.
2. Accédez au répertoire dans lequel vous avez enregistré le TAP. Par défaut, les TAP sont sauvegardés dans le sous-répertoire `\TAP` du répertoire d'installation du produit.
3. Entrez un nom pour le TAP dans le champ **Nom du fichier**.
4. Saisissez un libellé dans le champ **Libellé du pack**. Lorsque vous saisissez un nom de fichier, ce nom apparaît automatiquement comme libellé mais vous pouvez la modifier.

5. Pour exclure un ensemble de catégories du TAP, désélectionnez la case **Inclure**. Ainsi, il ne sera pas ajouté à votre pack. Par défaut, un ensemble de catégories par question est inclus dans le TAP. Le TAP doit contenir au moins un ensemble de catégories.
6. Renommer des ensembles de catégories. La colonne **Nouvel ensemble de catégories** contient des noms génériques par défaut qui sont générés en ajoutant le préfixe `Cat_` au nom de la variable de texte. Un simple clic dans la cellule permet de modifier ce nom. Appuyer sur Entrée ou cliquer à un autre droit permet d'appliquer la modification du nom. Si vous renommez un ensemble de catégories, le nom est modifié dans le TAP, mais le nom de la variable qui apparaît dans la session reste le même.
7. Vous pouvez changer l'ordre des ensembles de catégories avec les flèches situées à droite du tableau des ensembles de catégories.
8. Cliquez sur **Enregistrer** pour créer le pack d'analyse de texte. La boîte de dialogue se ferme.

Chargement des packs d'analyse de texte

Lors de la configuration d'un noeud modélisation Text Mining, vous devez spécifier les ressources à utiliser pendant l'extraction. Plutôt que de choisir un modèle de ressources, vous pouvez sélectionner un TAP (pack d'analyse de texte) pour copier non seulement les ressources mais également un ensemble de catégories dans le noeud.

Les TAP sont plus intéressants lors de la création d'un modèle de catégories interactif car vous pouvez utiliser l'ensemble de catégories comme point de départ pour la catégorisation. Lorsque vous exécutez le flux, la session de plan de travail interactif est ouverte et cet ensemble de catégories apparaît dans la sous-fenêtre Catégories. Ainsi, vous déterminez immédiatement le score de vos documents et de vos enregistrements à l'aide de ces catégories puis vous continuez à affiner, créer et étendre ces catégories jusqu'à ce que vous soyez satisfait. Pour plus d'informations, voir «Stratégies et méthodes de création de catégories», à la page 98.

A partir de la version 14, vous pouvez aussi afficher la langue pour laquelle les ressources de ce TAP ont été définies lorsque vous cliquez sur **Charger** et que vous choisissez le TAP.

Pour charger un pack d'analyse de texte

1. Modifiez le noeud modélisation Text Mining.
2. Dans l'onglet Modèle, choisissez *Pack d'analyse de texte* dans la section **Copier les ressources depuis**.
3. Cliquez sur **Charger**. La boîte de dialogue Charger un pack d'analyse de texte s'ouvre.
4. Recherchez l'emplacement du TAP contenant les ressources et l'ensemble de catégories à copier dans le noeud. Par défaut, les TAP sont sauvegardés dans le sous-répertoire `\TAP` du répertoire d'installation du produit.
5. Entrez un nom pour le TAP dans le champ **Nom du fichier**. Ce libellé apparaît automatiquement.
6. Sélectionnez l'ensemble de catégories à utiliser. Il s'agit de l'ensemble de catégories qui apparaîtra dans la session de plan de travail interactif. Vous pouvez ensuite modifier et améliorer ces catégories manuellement ou en utilisant les options Créer ou étendre des catégories.
7. Cliquez sur **Charger** pour copier le contenu du pack d'analyse de texte dans le noeud. La boîte de dialogue se ferme. Lorsqu'un TAP est chargé, une copie de ce TAP est copiée dans le noeud ; ainsi, toutes les modifications effectuées sur les ressources et les catégories ne seront pas reflétées dans le TAP sauf si vous le mettez à jour et le rechargez de manière explicite.

Mise à jour des Packs d'analyse de texte

Si vous apportez des améliorations à un ensemble de catégories, à des ressources linguistiques ou que vous créez un tout nouvel ensemble de catégories, vous pouvez mettre à jour un pack d'analyse de texte (TAP) pour que ces améliorations soient plus faciles à réutiliser ultérieurement. Pour ce faire, vous devez ouvrir la session contenant les informations que vous voulez intégrer au TAP. Lorsque vous effectuez une

mise à jour, vous pouvez choisir d'ajouter des ensembles de catégories, de remplacer des ressources, de changer le libellé du pack ou de renommer/réorganiser les ensembles.

Pour mettre à jour un pack d'analyse de texte

1. Dans les menus, choisissez **Fichier > Packs d'analyse de texte > Mise à jour d'un pack**. La boîte de dialogue Mettre à jour le pack apparaît.
2. Accédez au répertoire contenant le pack d'analyse de texte à mettre à jour.
3. Entrez un nom pour le TAP dans le champ **Nom du fichier**.
4. Pour remplacer les ressources linguistiques dans le TAP par celles de la session en cours, sélectionnez l'option **Remplacer les ressources de ce pack par celles de la session ouverte**. Généralement, il est utile de mettre à jour les ressources linguistiques car elles ont servi à extraire les concepts et motifs clés utilisés pour créer les définitions de catégories. Utiliser les ressources linguistiques les plus récentes permet d'obtenir de meilleurs résultats lors de la catégorisation de vos enregistrements. Si vous ne sélectionnez pas cette option, les ressources linguistiques déjà contenues dans le pack sont conservées en l'état.
5. Pour modifier uniquement les ressources linguistiques, vérifiez que l'option **Remplacer les ressources de ce pack par celles de la session ouverte** est bien sélectionnée puis sélectionnez uniquement les ensembles de catégories actuels qui se trouvaient dans le TAP.
6. Pour inclure de nouveaux ensembles de catégories depuis la session dans le TAP, sélectionnez la case à cocher pour chaque ensemble à ajouter. Vous pouvez ajouter un, plusieurs ou aucun des ensembles de catégories.
7. Pour supprimer des ensembles de catégories du TAP, désélectionnez la case **Inclure** correspondante. Vous pouvez choisir de supprimer un ensemble de catégories qui se trouvait déjà dans le TAP car vous en ajoutez un ayant été amélioré. Pour ce faire, désélectionnez la case **Inclure** de l'ensemble de catégories correspondant dans la colonne Ensemble de catégories actuel. Le TAP doit contenir au moins un ensemble de catégories.
8. Si nécessaire, modifiez les noms des ensembles de catégories. Un simple clic dans la cellule permet de modifier ce nom. Appuyer sur entrée ou cliquer à un autre droit permet d'appliquer la modification du nom. Si vous renommez un ensemble de catégories, le nom est modifié dans le TAP, mais le nom de la variable qui apparaît dans la session reste le même. Si deux ensembles de catégories ont le même nom, ces noms apparaissent en rouge tant que le doublon n'a pas été corrigé.
9. Pour créer un nouveau pack en intégrant le contenu de la session au contenu du TAP sélectionné, cliquez sur **Enregistrer en tant que Nouveau**. La boîte de dialogue Enregistrer comme pack d'analyse de texte apparaît. Suivez les instructions suivantes.
10. Cliquez sur **Mettre à jour** pour enregistrer les modifications effectuées dans le TAP sélectionné.

Pour enregistrer un pack d'analyse de texte

1. Accédez au répertoire dans lequel vous avez enregistré le fichier du TAP. Par défaut, les fichiers TAP sont enregistrés dans le sous-répertoire TAP du répertoire d'installation.
2. Entrez un nom pour le fichier TAP dans le champ Nom du fichier.
3. Saisissez un libellé dans le champ Libellé du pack. Le nom de fichier choisi est automatiquement utilisé comme libellé. Mais vous pouvez renommer ce libellé. Un libellé est nécessaire.
4. Cliquez sur **Enregistrer** pour créer le pack.

Edition et affinage des catégories

Une fois les catégories créées, vous souhaiterez certainement les analyser et procéder à des ajustements. Outre le réglage des ressources linguistiques, vous devez procéder à l'examen de vos catégories en recherchant des moyens de combiner ou de nettoyer leurs définitions. Vérifiez également certains documents ou enregistrements catégorisés. Vous pouvez également examiner les documents ou les enregistrements d'une catégorie et y opérer des ajustements, de sorte que les catégories puissent collecter les nuances et les distinctions.

Vous pouvez utiliser les techniques intégrées de création de catégories automatisées pour créer vos catégories. Cependant, vous voudrez probablement apporter quelques modifications à ces catégories. Lorsque vous utilisez une ou plusieurs techniques, un certain nombre de nouvelles catégories apparaissent dans la fenêtre. Vous pouvez alors examiner les données d'une catégorie et les ajuster jusqu'à ce que la définition vous convienne. Pour plus d'informations, voir «A propos des catégories», à la page 103.

Voici quelques options pour affiner vos catégories, la plupart étant décrites dans les pages suivantes :

Ajout de descripteurs aux catégories

Après avoir utilisé des techniques automatisées, il est possible que vous disposiez encore de résultats d'extraction qui n'ont été utilisés dans aucune des définitions de catégorie. Consultez cette liste dans la sous-fenêtre des Résultats de l'extraction. Si vous trouvez des éléments que vous souhaitez déplacer vers une catégorie, vous pouvez les ajouter dans une catégorie existante ou nouvelle.

Pour ajouter un concept ou un type à une catégorie

1. Dans les sous-fenêtres Résultats d'extraction et Données, sélectionnez les éléments que vous souhaitez ajouter à une catégorie nouvelle ou existante.
2. Dans les menus, sélectionnez **Catégories > Ajouter à la catégorie**. La boîte de dialogue affiche l'ensemble des catégories. Sélectionnez la catégorie à laquelle vous voulez ajouter les éléments sélectionnés. Si vous voulez ajouter les éléments à une nouvelle catégorie, sélectionnez **Nouvelle catégorie**. Une nouvelle catégorie apparaît dans la sous-fenêtre Catégories, sous le nom du premier élément sélectionné.

Modification des descripteurs de catégorie

Dès que vous avez créé des catégories, vous pouvez ouvrir chacune d'entre elles pour visualiser l'ensemble des descripteurs qui constituent sa définition. Dans la boîte de dialogue Définitions de catégorie, vous pouvez apporter un certain nombre de modifications à vos descripteurs de catégorie. De plus, si des catégories apparaissent dans l'arborescence des catégories, vous pouvez également les utiliser.

Pour éditer une catégorie

1. Sélectionnez la catégorie à éditer dans la sous-fenêtre Catégories.
2. Dans les menus, sélectionnez **Vue > Définitions de catégorie**. La boîte de dialogue Définitions de catégorie apparaît.
3. Sélectionnez le descripteur que vous souhaitez éditer, puis cliquez sur le bouton correspondant dans la barre d'outils.

Le tableau suivant décrit chaque bouton de la barre d'outils que vous pouvez utiliser pour éditer des définitions de catégorie.

Tableau 31. Boutons de la barre d'outils et descriptions.

Icônes	Description
	Supprime les descripteurs sélectionnés de la catégorie.
	Déplace les descripteurs sélectionnés vers une catégorie existante ou nouvelle.
	Déplace les descripteurs sélectionnés vers une catégorie sous la forme d'une règle de catégorie &. Pour plus d'informations, voir «Utilisation des règles de catégorie», à la page 121.

Tableau 31. Boutons de la barre d'outils et descriptions (suite).

Icônes	Description
	Déplace chacun des descripteurs sélectionnés dans une nouvelle catégorie qui lui est propre.
 **Afficher**	Met à jour l'affichage des sous-fenêtres Données et Visualisation en fonction des descripteurs sélectionnés.

Déplacement de catégories

Si vous voulez placer une catégorie dans une autre catégorie existante ou déplacer des descripteurs dans une autre catégorie, vous pouvez les déplacer.

Pour déplacer une catégorie

1. Dans la sous-fenêtre Catégories, sélectionnez les catégories à déplacer dans une autre catégorie.
2. Dans les menus, choisissez **Catégories > Déplacer vers la catégorie**. Le menu présente un ensemble de catégories, la plus récente figurant en haut de la liste. Sélectionnez le nom de la catégorie vers laquelle vous voulez déplacer les concepts sélectionnés.

- Si vous voyez le nom que vous recherchez, sélectionnez-le pour l'ajouter à cette catégorie.
- S'il n'apparaît pas, sélectionnez **Plus** pour afficher la boîte de dialogue Toutes les catégories et sélectionnez la catégorie dans la liste.

Aplatissement des catégories

Lorsque vous avez une structure de catégories hiérarchique avec des catégories et des sous-catégories, vous pouvez aplatir votre structure. Lorsque vous aplatissez une catégorie, tous les descripteurs des sous-catégories de cette catégorie sont déplacés vers la catégorie sélectionnée et les sous-catégories désormais vides sont supprimées. Ainsi, tous les documents qui étaient utilisés pour la mise en correspondance des sous-catégories sont désormais catégorisés dans la catégorie sélectionnée.

Pour aplatir une catégorie

1. Dans la sous-fenêtre Catégories, sélectionnez une catégorie (de niveau supérieur ou une sous-catégorie) que vous souhaitez aplatir.
2. Dans les menus, sélectionnez **Catégories > Aplatir des catégories**. Les sous-catégories sont supprimées et les descripteurs sont fusionnés dans la catégorie sélectionnée.

Fusion ou combinaison de catégories

Si vous souhaitez combiner deux catégories ou plus dans une nouvelle catégorie, vous pouvez les fusionner. Quand vous fusionnez des catégories, une nouvelle catégorie est créée avec un nom générique. Tous les concepts, types et motifs utilisés dans les descripteurs de catégorie sont déplacés dans cette nouvelle catégorie. Par la suite, vous pouvez renommer cette catégorie en éditant ses propriétés.

Pour fusionner tout ou partie d'une catégorie

1. Dans la sous-fenêtre Catégories, sélectionnez les éléments que vous souhaitez fusionner.
2. Dans les menus, sélectionnez **Catégories > Fusionner les catégories**. La boîte de dialogue Propriétés des catégories s'affiche et vous permet d'entrer un nom pour la catégorie nouvellement créée. Les catégories sélectionnées sont fusionnées dans la nouvelle catégorie en tant que sous-catégories.

Suppression de catégories

Si vous ne souhaitez pas conserver une catégorie, vous pouvez la supprimer.

Pour supprimer une catégorie

1. Dans la sous-fenêtre Catégories, sélectionnez la ou les catégories à supprimer.
2. Dans les menus, sélectionnez **Edition > Supprimer**.

Chapitre 10. Analyse des clusters

Vous pouvez créer et explorer des clusters de concept dans la vue Clusters (**Vue** > **Clusters**). Un *cluster* est un regroupement de concepts liés, généré par des algorithmes de classification non supervisée qui se fondent sur la fréquence d'apparition de ces concepts dans l'ensemble de documents/enregistrements et la fréquence d'apparition conjointe des concepts dans le même document (ou *co-occurrence*). Chaque concept d'un cluster est co-occurrent avec au moins un autre concept du cluster. Les clusters ont pour objectif de regrouper les concepts apparaissant ensemble, alors que les catégories ont pour objectif de regrouper les documents ou les enregistrements en fonction des correspondances existant entre le texte et les descripteurs (concept, règles, motifs) pour chaque catégorie.

Un cluster adéquat est un cluster présentant des concepts fortement liés et fréquemment cooccurrents, ainsi que dotés de peu de liens vers des concepts d'autres clusters. Lorsque vous travaillez avec des ensembles de données volumineux, cette technique peut aboutir à des temps de traitement beaucoup plus longs.

La classification non supervisée est un processus qui commence par l'analyse d'un ensemble de concepts et la recherche des concepts qui sont souvent cooccurrents dans les documents. Deux concepts cooccurrents dans un document sont considérés comme formant une paire de concepts. Ensuite, le processus de classification non supervisée évalue la *valeur de similarité* de chaque paire de concepts en comparant le nombre de documents dans lequel la paire apparaît au nombre de documents dans lequel chaque concept apparaît. Pour plus d'informations, voir «Calcul des valeurs du lien de similarité», à la page 146.

En dernier lieu, le processus de classification non supervisée regroupe les concepts similaires en clusters, par agrégation, et prend en compte la valeur de leurs liens et les paramètres définis dans la boîte de dialogue Créer des clusters. Par "agrégation", nous entendons l'ajout de concepts ou la fusion des clusters plus petits en un cluster plus grand jusqu'à ce que le cluster soit saturé. Un cluster est *saturé* lorsqu'une fusion supplémentaire de concepts ou de petits clusters entraînerait le dépassement des paramètres de la boîte de dialogue Créer des clusters (nombre de concepts, de liens internes ou de liens externes). Un cluster prend le nom du concept qui, en son sein, présente le nombre global le plus élevé de liens vers d'autres concepts du même cluster.

Finalement, toutes les paires de concepts ne se retrouvent pas dans le même cluster puisqu'il peut exister un lien plus fort dans un autre cluster ou la saturation peut empêcher la fusion des clusters dans lesquels elles apparaissent. C'est la raison pour laquelle il existe à la fois des liens internes et des liens externes.

- Les *liens internes* sont des liens entre des paires de concepts au sein d'un cluster. Dans un cluster, tous les concepts ne sont pas liés les uns aux autres. Toutefois, chaque concept est au moins lié à un autre concept de son cluster.
- Les *liens externes* sont des liens entre des paires de concepts dans des clusters distincts (un concept au sein d'un cluster et un concept en dehors, dans un autre cluster).

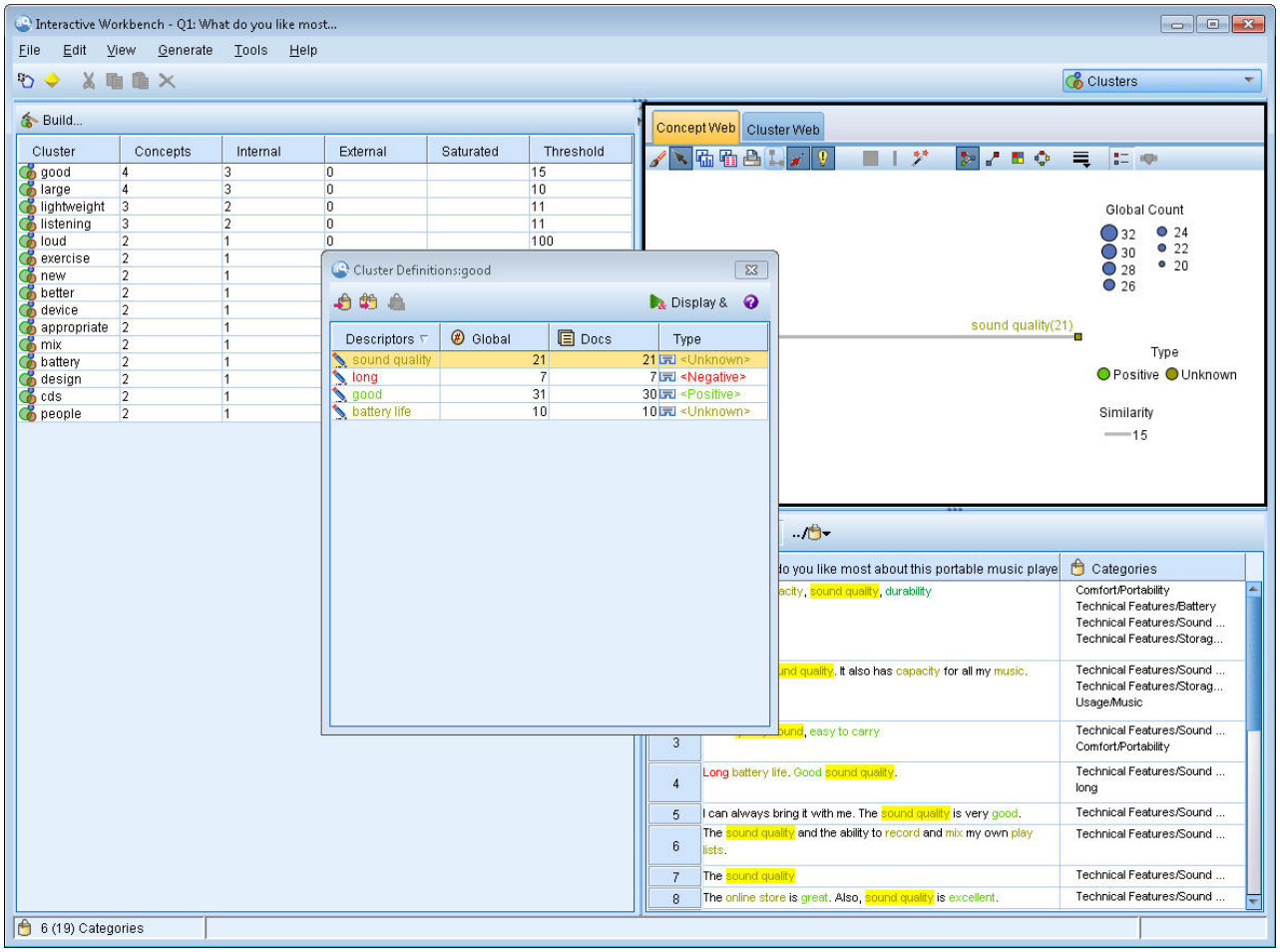


Figure 30. Vue Clusters

La vue Clusters est organisée en trois sous-fenêtres (vous pouvez masquer ou afficher chaque sous-fenêtre en sélectionnant son nom dans le menu Vue) :

- **Sous-fenêtre Clusters** Vous pouvez créer et gérer vos clusters dans cette sous-fenêtre. Pour plus d'informations, voir «Exploration des Clusters», à la page 147.
- **Sous-fenêtre Visualisation** Vous pouvez explorer visuellement vos clusters et analyser la manière dont ils interagissent dans cette sous-fenêtre. Pour plus d'informations, voir «Graphiques Cluster», à la page 157.
- **Sous-fenêtre Données** Vous pouvez explorer et passer en revue le texte contenu dans les documents et enregistrements qui correspondent aux sélections effectuées dans la boîte de dialogue Définitions du cluster. Pour plus d'informations, voir «Définitions de cluster», à la page 147.

Création de clusters

Lorsque vous accédez pour la première fois à la vue Clusters, aucun cluster n'est visible. Vous pouvez créer des clusters via les menus (**Outils > Créer des clusters**) ou en cliquant sur le bouton **Créer** de la barre d'outils. Cette action ouvre la boîte de dialogue Créer des clusters, dans laquelle vous pouvez définir les paramètres et les limites pour la création des clusters.

Remarque : lorsque les résultats de l'extraction ne correspondent plus aux ressources, cette sous-fenêtre devient jaune, tout comme la sous-fenêtre Résultats d'extraction. Vous pouvez procéder à une nouvelle extraction pour obtenir les derniers résultats d'extraction et la couleur jaune disparaîtra. Toutefois, à

chaque nouvelle extraction, la sous-fenêtre Clusters est effacée et vous devez recréer vos clusters. De même, les clusters ne sont pas enregistrés d'une session à l'autre.

Les zones et champs suivants sont disponibles dans la boîte de dialogue Créer des clusters :

Entrées

Tableau Entrées Les clusters sont créés à partir de descripteurs dérivés de certains types. Dans le tableau, vous pouvez sélectionner les types à inclure dans le processus de création. Les types qui capturent le plus d'enregistrements ou de documents sont présélectionnés par défaut.

Concepts à classifier : Sélectionnez la méthode pour choisir les concepts que vous voulez utiliser pour la classification. En réduisant le nombre de concepts, vous pouvez accélérer le processus de classification non supervisée. Vous pouvez effectuer une classification à l'aide d'un certain nombre de meilleurs concepts, d'un pourcentage de meilleurs concepts ou en utilisant tous les concepts :

- **Numéro basé sur le nombre de documents** Lorsque vous sélectionnez **Plus grand nombre de concepts**, entrez le nombre de concepts à prendre en compte pour la classification. Les concepts choisis le sont en fonction des plus grands effectifs de documents. Il s'agit du nombre de documents ou d'enregistrements dans lesquels le concept apparaît. La valeur maximale est de 150 000.
- **Pourcentage basé sur le nombre de documents** Lorsque vous sélectionnez **Plus grand pourcentage de concepts**, entrez le pourcentage de concepts à prendre en compte pour la classification. Les concepts choisis le sont en fonction du pourcentage de concepts qui présentent les plus grands effectifs de documents.

Limites de sortie

Nombre maximal de clusters à créer Cette valeur correspond au nombre maximal de clusters à générer et à afficher dans la sous-fenêtre Clusters. Durant le processus de classification non supervisée, les clusters saturés sont présentés avant les clusters non saturés et, par conséquent, de nombreux clusters obtenus sont saturés. De façon à visualiser davantage de clusters non saturés, vous pouvez régler ce paramètre sur une valeur supérieure au nombre de clusters saturés.

Nombre maximal de concepts dans un cluster Cette valeur correspond au nombre maximal de concepts que peut contenir un cluster.

Nombre minimal de concepts dans un cluster Cette valeur correspond au nombre minimal de concepts que peut contenir un cluster.

Nombre maximal de liens internes Cette valeur correspond au nombre maximal de liens internes que peut contenir un cluster. Les liens internes sont des liens entre des paires de concepts au sein d'un cluster.

Nombre maximal de liens externes Cette valeur correspond au nombre maximal de liens vers des concepts situés en dehors du cluster. Les liens externes sont des liens entre des paires de concepts dans des clusters distinctes.

Valeur de lien minimale Cette valeur correspond à la plus petite valeur de lien acceptée pour prendre en considération une paire de concepts dans le cadre de la classification non supervisée. La valeur du lien est calculée à l'aide d'une formule de similarité. Pour plus d'informations, voir «Calcul des valeurs du lien de similarité», à la page 146.

Empêcher l'appariement de concepts spécifiques. Cochez cette case pour arrêter le processus de regroupement ou d'appariement de deux concepts dans les résultats. Pour créer ou gérer des paires de concepts, cliquez sur **Gérer les paires**. Pour plus d'informations, voir «Gestion des paires d'exceptions de liens», à la page 110.

Calcul des valeurs du lien de similarité

La seule connaissance du nombre de documents dans lequel deux concepts sont cooccurrents n'indique pas à elle seule leur degré de similarité. Dans ce cas, il peut être utile de connaître la valeur de similarité. La valeur du lien de similarité est mesurée par le biais de la comparaison entre les effectifs des documents de co-occurrence et les effectifs des documents pour chaque concept de la relation. Lors du calcul de la similarité, l'unité de mesure est le nombre de documents (effectifs des documents) dans lesquels apparaît un concept ou une paire de concepts. Un concept ou une paire de concepts sont "détectés" dans un document s'ils apparaissent *au moins* une fois dans ce dernier. Vous pouvez choisir de représenter graphiquement la valeur du lien de similarité par l'épaisseur de ligne du graphique de concept.

L'algorithme révèle les relations les plus fortes, ce qui signifie que la tendance des concepts à apparaître ensemble dans les données textuelles est largement supérieure à leur tendance à apparaître de façon indépendante. En interne, l'algorithme produit un coefficient de similarité qui s'étend de 0 à 1, où la valeur 1 signifie que les deux concepts apparaissent toujours ensemble et jamais séparément. Le résultat du coefficient de similarité est ensuite multiplié par 100 et arrondi au nombre entier le plus proche. Le coefficient de similarité est calculé à l'aide de la formule présentée dans la figure suivante.

$$\text{similarity coefficient} = \frac{(C_{IJ})^2}{(C_I \times C_J)}$$

Figure 31. Formule du coefficient de similarité

Où :

- C_I est le nombre de documents ou d'enregistrements dans lequel apparaît le concept I.
- C_J est le nombre de documents ou d'enregistrements dans lequel apparaît le concept J.
- C_{IJ} est le nombre de documents ou d'enregistrements dans lequel la paire de concepts I et J est cooccurrente dans l'ensemble de documents.

Par exemple, supposons que vous disposiez de 5 000 documents. Supposons également que I et J soient des concepts extraits et IJ, une paire de concepts co-occurents de I et de J. Le tableau suivant propose deux scénarios démontrant la manière dont le coefficient et la valeur du lien sont calculés.

Tableau 32. Exemple de fréquence des concepts

Concept/paire	**Scénario A**	**Scénario B**
Concept : I	Présence dans 20 documents	Présence dans 30 documents
Concept : J	Présence dans 20 documents	Présence dans 60 documents
Paire de concepts : IJ	Co-occurrence dans 20 documents	Co-occurrence dans 20 documents
Coefficient de similarité	1	0,22222
Valeur du lien de similarité	100	22

Dans le scénario A, les concepts I et J ainsi que la paire IJ apparaissent dans 20 documents, ce qui produit un coefficient de similarité de 1. Cela qui signifie que les concepts apparaissent toujours ensemble. La valeur du lien de similarité pour cette paire est de 100.

Dans le scénario B, le concept I apparaît dans 30 documents et le concept J dans 60 documents. En revanche, la paire IJ est seulement présente dans 20 documents. Par conséquent, le coefficient de similarité est de 0,22222. La valeur du lien de similarité pour cette paire est arrondie à 22.

Exploration des Clusters

Après avoir créé des clusters, vous pouvez visualiser un ensemble de résultats dans la sous-fenêtre Clusters. Pour chaque cluster, le tableau présente les informations suivantes :

- **Cluster.** Il s'agit du nom du cluster. Les clusters sont nommés d'après le concept présentant le plus grand nombre de liens internes.
- **Concepts.** Il s'agit du nombre de concepts dans le cluster. Pour plus d'informations, voir «Définitions de cluster».
- **Interne.** Il s'agit du nombre de liens internes dans le cluster. Les liens internes sont des liens entre des paires de concepts au sein d'un cluster.
- **Externe.** Il s'agit du nombre de liens externes dans le cluster. Les liens externes représentent des liens entre des paires de concepts lorsqu'un concept se situe dans un cluster et l'autre dans un autre cluster.
- **Sat.** La présence de ce symbole indique que ce cluster aurait pu être plus grand, mais que dans ce cas, une ou plusieurs limites auraient été dépassées ; par conséquent, le processus de classification non supervisée a pris fin pour ce cluster, lequel est considéré alors comme étant *saturé*. A la fin du processus de classification non supervisée, les clusters saturées sont présentées avant les clusters non saturés et, par conséquent, de nombreux clusters obtenues sont saturées. De façon à visualiser davantage de clusters non saturés, vous pouvez modifier le paramètre **Nombre maximal de clusters à créer** sur une valeur supérieure au nombre de clusters saturés ou diminuer le paramètre **Valeur de lien minimale**. Pour plus d'informations, voir «Création de clusters», à la page 144.
- **Seuil.** Pour toutes les paires de concepts cooccurrents dans le cluster, il s'agit de la valeur la plus faible du lien de similarité de tout le cluster. Pour plus d'informations, voir «Calcul des valeurs du lien de similarité», à la page 146. Un cluster avec une valeur de seuil élevée signifie que les concepts qu'il regroupe ont une similarité globale élevée et sont liés de manière plus forte que les concepts d'un cluster avec une valeur de seuil moins élevée.

Pour en savoir plus sur un cluster donné, sélectionnez-le et la sous-fenêtre de visualisation à droite présente alors deux graphiques permettant d'explorer les clusters. Pour plus d'informations, voir «Graphiques Cluster», à la page 157. Vous pouvez également couper/coller le contenu d'un tableau dans une autre application.

Lorsque les résultats de l'extraction ne correspondent plus aux ressources, cette sous-fenêtre devient jaune, tout comme la sous-fenêtre Résultats d'extraction. Vous pouvez procéder à une nouvelle extraction pour obtenir les derniers résultats d'extraction et la couleur jaune disparaîtra. Toutefois, à chaque nouvelle extraction, la sous-fenêtre Clusters est effacée et vous devez recréer vos clusters. De même, les clusters ne sont pas enregistrés d'une session à l'autre.

Définitions de cluster

Vous pouvez visualiser tous les concepts d'un cluster en le sélectionnant dans la sous-fenêtre Clusters et en ouvrant la boîte de dialogue Définitions du cluster (**Vue > Définition du cluster**).

Tous les concepts du cluster sélectionné apparaissent dans la boîte de dialogue Définitions du cluster. Si vous sélectionnez un ou plusieurs concepts dans la boîte de dialogue Définitions du cluster et que vous cliquez sur **Afficher &**, la sous-fenêtre Données affiche tous les enregistrements ou documents dans lesquels *tous les concepts sélectionnés apparaissent ensemble*. Toutefois, la sous-fenêtre Données n'affiche pas d'enregistrements ou de documents texte lorsque vous sélectionnez un cluster dans la sous-fenêtre Cluster. Pour plus d'informations sur la sous-fenêtre Données, voir dans.

La sélection de concepts dans cette boîte de dialogue modifie également le graphique Relations par concept. Pour plus d'informations, voir «Graphiques Cluster», à la page 157. De même, lorsque vous sélectionnez un ou plusieurs concepts dans la boîte de dialogue Définitions du cluster, la sous-fenêtre Données affiche tous les liens internes et externes de ces concepts.

Descriptions de colonne

Des icônes s'affichent pour vous permettre d'identifier facilement chaque descripteur.

Tableau 33. Colonnes et icônes de descripteur

Colonnes	Description
Descripteurs	Nom du concept.
 **Global**	Affiche le nombre de fois que ce descripteur apparaît dans l'ensemble de données, également connu sous le nom de fréquence globale.
 **Docs**	Affiche le nombre de documents ou d'enregistrements dans lesquels ce descripteur apparaît, également connu sous le nom de fréquence du document.
Type	Affiche le ou les types auxquels le descripteur appartient. Si le descripteur est une règle de catégorie, aucun nom de type n'est indiqué dans cette colonne.

Actions de la barre d'outils

Dans cette boîte de dialogue, vous pouvez également sélectionner un ou plusieurs concepts à utiliser dans une catégorie. Il existe plusieurs façons de procéder mais il est plus intéressant de sélectionner des concepts cooccurrents dans un cluster et de les ajouter en tant que règle de catégorie. Pour plus d'informations, voir «Règles de co-occurrence», à la page 114. Vous pouvez utiliser les boutons de la barre d'outils pour ajouter les concepts aux catégories.

Tableau 34. Boutons de la barre d'outils pour l'ajout de concepts aux catégories

Icônes	Description
	Ajoute les concepts sélectionnés à une nouvelle catégorie ou à une catégorie existante.
	Ajoute les concepts sélectionnés sous la forme d'une règle de catégorie & à une nouvelle catégorie ou à une catégorie existante. Pour plus d'informations, voir «Utilisation des règles de catégorie», à la page 121.
	Ajoute chacun des concepts sélectionnés sous la forme d'une nouvelle catégorie qui lui est propre.
	Met à jour l'affichage des sous-fenêtres Données et Visualisation en fonction des descripteurs sélectionnés.

Remarque : vous pouvez également utiliser les menus contextuels pour ajouter des concepts à un type, en tant que synonymes ou éléments à exclure.

Chapitre 11. Exploration de l'analyse des liens du texte

Dans la vue Analyse des liens du texte (TLA), vous pouvez explorer des résultats de motifs d'analyse des liens du texte. L'analyse des liens du texte est une technologie de mise en correspondance de motifs qui vous permet de définir des règles de motifs et de les comparer aux concepts extraits et aux relations trouvées dans le texte.

Par exemple, l'extraction d'idées concernant une organisation peut ne présenter qu'un intérêt relatif pour vous. Grâce à l'analyse des liens du texte, vous pouvez en savoir plus sur les liens entre cette organisation et d'autres organisations ou sur les personnes au sein d'une organisation. Vous pouvez également utiliser l'analyse TLA pour extraire des opinions sur des produits ou, pour certaines langues, les relations entre des gènes.

Lorsque vous avez extrait des résultats de motifs TLA, vous pouvez les consulter dans les sous-fenêtres Motifs de type et Motifs de concept de la vue Analyse des liens du texte. Pour plus d'informations, voir «Motifs de type et Motifs de concept», à la page 151. Vous pouvez les examiner de manière plus détaillée dans les sous-fenêtres Données ou Visualisation de cette vue. Mais surtout, vous pouvez les ajouter aux catégories.

Si vous n'avez pas choisi cette action, vous pouvez cliquer sur **Extraire** et choisir **Extraction avec analyse des liens du texte** dans la boîte de dialogue Paramètres d'extraction. Pour plus d'informations, voir «Extraction des résultats de patrons TLA», à la page 150.

Pour pouvoir extraire les résultats de motifs de l'analyse des liens du texte, des règles de motifs d'analyse des liens du texte doivent être définies dans le modèle de ressources ou dans les bibliothèques que vous utilisez. Vous pouvez utiliser les motifs TLA dans certains modèles de ressources fournis avec IBM SPSS Modeler Text Analytics. Le genre de relations et de motifs que vous pouvez extraire dépend entièrement des règles TLA définies dans vos ressources. Vous pouvez définir vos propres règles TLA. Les motifs sont constitués de macros, listes de mots et intervalles de mots pour former une requête booléenne, ou règle, qui est comparée à votre texte d'entrée. Pour plus d'informations, voir Chapitre 18, «A propos des règles des liens du texte», à la page 213.

Lorsqu'une règle de patron TLA correspond au texte, il est possible d'extraire ce texte sous la forme d'un motif et de le restructurer sous la forme de données de sortie. Les résultats sont alors visibles dans les sous-fenêtres de la vue Analyse des liens du texte. Chaque sous-fenêtre peut être masquée ou affichée en sélectionnant son nom dans le menu Vue :

- **Sous-fenêtres Motifs de type et Motifs de concept.** Ces deux sous-fenêtres permettent de créer et d'explorer des motifs. Pour plus d'informations, voir «Motifs de type et Motifs de concept», à la page 151.
- **Sous-fenêtre Visualisation.** Cette sous-fenêtre permet d'explorer visuellement la façon dont les concepts et les types des motifs interagissent. Pour plus d'informations, voir «Graphiques Analyse des liens du texte», à la page 158.
- **Sous-fenêtre Données.** Vous pouvez explorer et passer en revue le texte contenu dans les documents et enregistrements qui correspondent aux sélections effectuées dans une autre sous-fenêtre. Pour plus d'informations, voir «Sous-fenêtre Données», à la page 153.

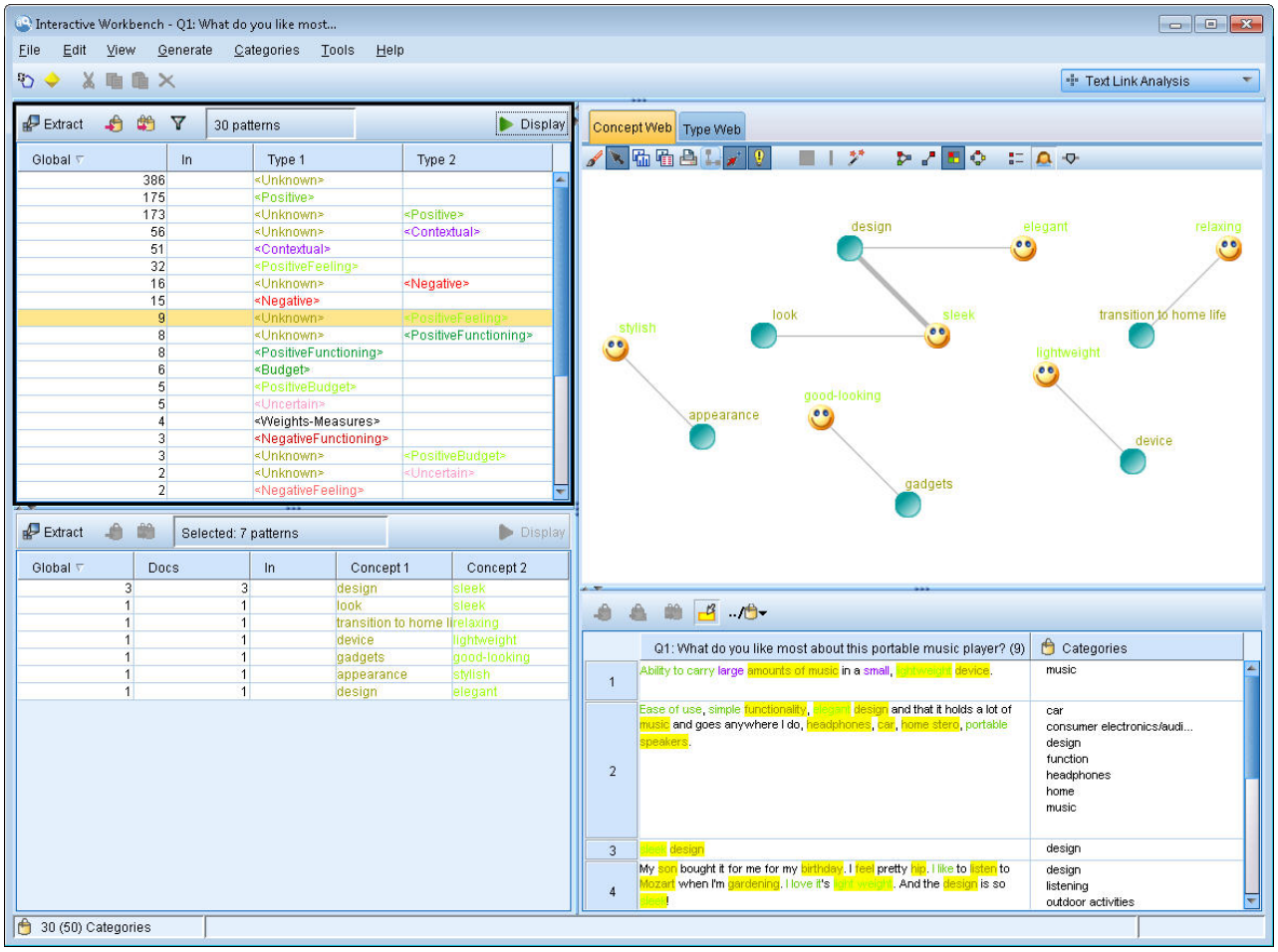


Figure 32. Vue Analyse des liens du texte

Extraction des résultats de patrons TLA

Le résultat du processus d'extraction correspond à un ensemble de concepts, de types et de motifs d'analyse des liens du texte (TLA) si la fonction a été activée. Si vous avez extrait des motifs TLA, vous pouvez les visualiser dans la vue Analyse des liens du texte. Lorsque les résultats de l'extraction ne sont pas synchronisés avec les ressources, les sous-fenêtres de motifs deviennent jaunes et indiquent ainsi qu'une nouvelle extraction produirait des résultats différents.

Vous devez choisir d'extraire ces motifs dans le paramètre de noeud ou dans la boîte de dialogue Extraire à l'aide de l'option **Extraction avec analyse des liens du texte**. Pour plus d'informations, voir «Extraction de données», à la page 82.

Remarque : il existe une relation entre la taille du jeu de données et le temps requis pour procéder à l'extraction. Reportez-vous aux instructions d'installation pour obtenir des statistiques et des recommandations sur les performances. Vous pouvez toujours envisager d'insérer un noeud échantillon en amont ou d'optimiser la configuration de votre machine.

Pour extraire des données

1. A partir des menus, sélectionnez **Outils> Extraire**. Vous pouvez également cliquer sur le bouton **Extraire** de la barre d'outils.

2. Modifiez les options à utiliser. Gardez à l'esprit que l'option **Extraction avec analyse des liens du texte** doit être sélectionnée dans cet onglet et que votre modèle doit comporter des règles TLA pour que des résultats de patrons TLA puissent être extraits. Pour plus d'informations, voir «Extraction de données», à la page 82.
3. Cliquez sur **Extraire** pour lancer le processus d'extraction.

Dès le début de l'extraction, une boîte de dialogue indique la progression du processus. Si vous voulez annuler l'extraction, cliquez sur **Annuler**. Lorsque l'extraction est terminée, la boîte de dialogue se ferme et les résultats apparaissent dans la sous-fenêtre. Pour plus d'informations, voir «Motifs de type et Motifs de concept».

Motifs de type et Motifs de concept

Les motifs sont constitués de deux éléments : une combinaison de concepts et de types. Les motifs s'avèrent particulièrement utiles lorsque vous tentez de découvrir des opinions sur un sujet donné ou des relations entre des concepts. Par exemple, l'intérêt de l'extraction du nom du produit de votre concurrent peut être limité. Dans ce cas, vous pouvez examiner les motifs extraits pour éventuellement découvrir des exemples de document ou d'enregistrement contenant du texte indiquant que le produit est bon, mauvais ou cher.

Un motif peut inclure jusqu'à six types ou six concepts. C'est la raison pour laquelle les lignes des deux sous-fenêtres de motifs contiennent jusqu'à six propriétés, ou positions. Chaque propriété correspond à la position spécifique d'un élément dans la règle de patron TLA telle qu'elle est définie dans les ressources linguistiques. Dans le plan de travail interactif, si une propriété ne contient pas de valeur, elle n'apparaît pas dans le tableau. Par exemple, si les résultats de motifs les plus longs ne contiennent pas plus de quatre propriétés, les deux dernières n'apparaissent pas. Pour plus d'informations, voir Chapitre 18, «A propos des règles des liens du texte», à la page 213.

Lorsque vous extrayez des résultats de motifs, ils sont d'abord regroupés au niveau du type, puis divisés en motifs de concept. Pour cette raison, il existe deux sous-fenêtres de résultats différentes : **Motifs de type** (en haut à gauche) et **Motifs de concept** (en bas à gauche). Pour afficher tous les motifs de concepts retournés, sélectionnez tous les motifs de types. La sous-fenêtre inférieure des concepts affichera alors tous les motifs de concepts jusqu'à la valeur de rang maximal (tel que défini dans la boîte de dialogue Filtrer).

Motifs de type Cette sous-fenêtre présente les résultats de motifs comportant au moins deux types associés correspondant à une règle de motif TLA. Les motifs de type se présentent sous la forme `<Organization>` + `<Location>` + `<Positive>`, ce qui permet d'obtenir un commentaire positif concernant une organisation situé dans une location particulière. La syntaxe est la suivante :

```
<Type1> + <Type2> + <Type3> + <Type4> + <Type5> + <Type6>
```

Motifs de concept Cette sous-fenêtre présente les résultats de motifs au niveau du concept pour tous les motifs de type actuellement sélectionnés dans la sous-fenêtre Motifs de type située au-dessus. Les motifs de concept suivent une structure de type `hôtel` + `paris` + `merveilleux`. La syntaxe est la suivante :

```
concept1 + concept2 + concept3 + concept4 + concept5 + concept6
```

Lorsque les résultats de motifs utilisent moins de six propriétés (valeur maximale), seul le nombre de propriétés (ou colonnes) nécessaire apparaît. Toute propriété vide trouvée entre deux propriétés complétées est supprimée de sorte que le motif `<Type1>+<>+<Type2>+<>+<>+<>` soit représenté par `<Type1>+<Type3>`. Pour un motif de concept, cela équivaut à `concept1+.+concept2` (où `.` représente une valeur nulle).

Comme avec les résultats d'extraction dans la vue Catégories et concepts, vous pouvez vérifier les résultats ici. Si vous souhaitez affiner les types et concepts qui constituent ces motifs, procédez aux modifications dans la sous-fenêtre Résultats d'extraction de la vue Catégories et concepts ou directement

dans l'éditeur de ressources, puis exécutez une nouvelle extraction des motifs. Chaque fois qu'un concept, type ou motif est utilisé tel quel dans une définition de catégorie ou comme partie de règle, une icône de catégorie ou de règle apparaît dans la colonne **In** du tableau des motifs ou des résultats d'extraction.

Remarque : Si le nombre de résultats est plus important que ce que peut contenir la sous-fenêtre visible, vous pouvez utiliser les commandes de la partie inférieure de la sous-fenêtre pour vous déplacer vers l'avant ou vers l'arrière dans les résultats ou entrer un numéro de page pour y accéder.

Filtrage des résultats TLA

Lorsque vous travaillez sur des ensembles de données très volumineux, le processus d'extraction peut renvoyer des millions de résultats. Pour de nombreux utilisateurs, cette quantité peut compliquer l'examen des résultats. Vous avez cependant la possibilité de filtrer ces résultats, afin de mettre en évidence les plus intéressants. Pour limiter les motifs à afficher, vous pouvez modifier les paramètres dans la boîte de dialogue Filtrer. Tous ces paramètres sont utilisés ensemble.

Dans la vue TLA, la boîte de dialogue Filtrer contient les zones et les champs suivants.

Filtrer par fréquence Vous pouvez appliquer un filtre afin de n'afficher que les résultats présentant une certaine valeur de fréquence globale ou de documents.

- La **fréquence globale** est le nombre total d'apparitions d'un motif dans l'ensemble de documents ou d'enregistrements. Cette valeur apparaît dans la colonne **Global**.
- La **fréquence de documents** est le nombre total de documents ou d'enregistrements dans lesquels un motif apparaît. Elle est indiquée dans la colonne **Docs**.

Par exemple, si un motif apparaît 300 fois dans 500 enregistrements, nous déclarons que ce motif a une fréquence globale de 300 et une fréquence de documents de 500.

Et par texte correspondant Vous pouvez également appliquer un filtre n'affichant que les résultats correspondant à la règle que vous définissez ici. Entrez l'ensemble de caractères qui doit être mis en correspondance dans le champ **Texte correspondant**, puis indiquez si la recherche de ce texte doit porter sur les noms de concept ou de type (en identifiant le numéro de propriété) ou si elle doit porter sur l'ensemble. Ensuite, sélectionnez la condition à laquelle appliquer la correspondance (il n'est pas nécessaire d'utiliser des chevrons pour marquer le début ou la fin d'un nom de type). Sélectionnez **Et** ou **Ou** dans la liste déroulante de sorte que la règle corresponde aux deux instructions ou à une seule, puis définissez la seconde instruction de correspondance du texte de la même manière que la première.

Tableau 35. Conditions de correspondance de texte

Condition	Description
Contient	Texte mis en correspondance si la chaîne apparaît n'importe où. (Option par défaut)
Commence par	Le texte est seulement mis en correspondance si le concept ou le type commence par le texte entré.
Se termine par	Le texte est seulement mis en correspondance si le concept ou le type se termine par le texte entré.
Correspondance exacte	Toute la chaîne doit concorder avec le nom du concept ou du type.

Résultats affichés dans la sous-fenêtre de motifs

Supposons que vous utilisez une version du logiciel en anglais ; voici quelques exemples de la manière dont les résultats peuvent s'afficher sur la barre d'outils de la sous-fenêtre Motifs en fonction des filtres.

Figure 33. Résultats du filtre, exemple 1

Dans cet exemple, la barre d'outils montre que le nombre de motifs renvoyé est limité en raison du rang maximal spécifié dans le filtre. La présence d'une icône violette indique que le nombre maximal de motifs est atteint. Placez le curseur sur l'icône pour obtenir plus d'informations. Reportez-vous ci-dessus à l'explication du filtre **Et par rang**.

Figure 34. Résultats du filtre, exemple 2

Dans cet exemple, la barre d'outils montre que les résultats ont été limités par un filtre de correspondance de texte (voir l'icône loupe). Vous pouvez pointer sur l'icône pour visualiser la correspondance de texte.

Pour filtrer les résultats

1. Dans les menus, sélectionnez **Outils** > **Filtrer**. La boîte de dialogue Filtrer s'ouvre.
2. Sélectionnez et affinez les filtres à utiliser.
3. Cliquez sur **OK** pour appliquer les filtres et visualiser les nouveaux résultats.

Sous-fenêtre Données

Lorsque vous extrayez et explorez des motifs d'analyse des liens du texte, vous pouvez passer en revue certaines des données avec lesquelles vous travaillez. Par exemple, vous pouvez visualiser les enregistrements réels dans lesquels un groupe de motifs a été découvert. Vous pouvez consulter les enregistrements ou les documents dans la sous-fenêtre Données, située dans l'angle inférieur droit. S'il n'apparaît pas par défaut, sélectionnez **Vue** > **Sous-fenêtres** > **Données** dans les menus.

La sous-fenêtre Données présente une ligne par document ou enregistrement correspondant à une sélection dans la vue, jusqu'à une certaine limite d'affichage. Par défaut, le nombre de documents ou d'enregistrements affichés dans la sous-fenêtre Données est limité pour vous permettre de consulter vos données plus rapidement. Cependant, vous pouvez modifier cette limite dans la boîte de dialogue Options. Pour plus d'informations, voir «Options : onglet Session», à la page 76.

Remarque : Si le nombre de résultats est plus important que ce que peut contenir la sous-fenêtre visible, vous pouvez utiliser les commandes de la partie inférieure de la sous-fenêtre pour vous déplacer vers l'avant ou vers l'arrière dans les résultats ou entrer un numéro de page pour y accéder.

Affichage et actualisation de la sous-fenêtre Données

L'affichage de la sous-fenêtre Données n'est pas automatiquement actualisée car en présence d'ensembles de données volumineux, l'opération prendrait trop de temps. Par conséquent, lorsque vous sélectionnez des motifs de type ou de concept dans cette vue, vous pouvez cliquer sur **Afficher** pour actualiser le contenu de la sous-fenêtre Données.

Documents texte ou enregistrements

Si vos données textuelles sont sous la forme d'enregistrements et que le texte est relativement bref, le champ de texte de la sous-fenêtre Données affiche les informations dans leur intégralité. Cependant, si vous utilisez des enregistrements et de grands ensembles de données, la colonne du champ de texte affiche une petite partie du texte et ouvre une sous-fenêtre Aperçu du texte à droite qui permet de consulter une plus grande partie du texte de l'enregistrement sélectionné dans la table, voire son intégralité. Si vos données textuelles se présentent sous la forme de documents, la sous-fenêtre Données

affiche le nom de fichier du document. Lorsque vous sélectionnez un document, la sous-fenêtre Aperçu du texte s'ouvre et affiche le texte du document sélectionné.

Couleurs et mise en évidence

Chaque fois que vous affichez des données, des concepts et des descripteurs trouvés dans ces documents ou enregistrements, ils apparaissent en couleur pour vous permettre de les identifier facilement dans le texte. Le code couleur correspond aux types auxquels les concepts appartiennent. Vous pouvez également pointer sur un élément faisant l'objet d'un codage couleur pour afficher le concept sous lequel il a été extrait et le type auquel il a été affecté. Tout texte n'ayant pas été extrait apparaît en noir. En règle générale, ces mots non extraits sont souvent des connecteurs (*et* ou *avec*), des pronoms (*me* ou *ils*), et des verbes (*être*, *avoir* ou *prendre*).

Colonnes de la sous-fenêtre Données

Alors que la colonne de champ de texte est toujours visible, il est possible d'afficher également d'autres colonnes. Pour afficher d'autres colonnes, cliquez sur **Affichage** > **Panneau Données** dans les menus, puis sélectionnez la colonne que vous souhaitez afficher dans le panneau de données. Les colonnes pouvant être affichées sont les suivantes :

- **"Nom du champ de texte" (#)/Documents** Ajoute une colonne pour les données textuelles à partir desquelles des concepts et des types ont été extraits. Si vos données sont contenues dans des documents, la colonne est appelée Documents, et seul le nom de fichier du document ou son chemin complet est visible. Pour examiner le texte de ces documents, vous devez consulter la sous-fenêtre Aperçu du texte. Le nombre de lignes de la sous-fenêtre Données est indiqué entre parenthèses après le nom de cette colonne. Il peut arriver que les documents ou les enregistrements ne soient pas tous affichés en raison d'une limite définie dans la boîte de dialogue Options pour optimiser la vitesse de chargement. Si la limite est atteinte, le nombre sera suivi de **- Max**. Pour plus d'informations, voir «Options : onglet Session», à la page 76.
- **Catégories** Répertorie chacune des catégories à laquelle appartient un enregistrement. Lorsque cette colonne est affichée, l'actualisation de la sous-fenêtre Données peut prendre plus de temps afin d'afficher les informations les plus récentes.
- **Rang de pertinence** Donne un rang pour chaque enregistrement dans une seule catégorie. Ce rang montre dans quelle mesure l'enregistrement correspond à la catégorie par rapport aux autres enregistrements dans cette catégorie. Sélectionnez une catégorie dans la sous-fenêtre Catégories (sous-fenêtre supérieure gauche) pour voir le rang. Pour plus d'informations, voir «Pertinence des catégories», à la page 105.
- **Nombre de catégories** Répertorie le nombre de catégories auxquelles appartient un enregistrement.

Chapitre 12. Visualisation des graphiques

La vue Catégories et concepts, la vue Clusters et la vue Analyse des liens du texte présentent toutes une sous-fenêtre de visualisation dans l'angle supérieur droit de la fenêtre. Vous pouvez utiliser cette sous-fenêtre pour explorer visuellement les données. Les graphiques et diagrammes suivants sont disponibles.

- **Vue Catégories et concepts.** Cette vue contient trois diagrammes et graphiques : *Barre Catégorie*, *Relations de catégorie* et *Tableau des relations de catégorie*. Dans cette vue, les graphiques sont mis à jour uniquement lorsque vous cliquez sur **Afficher**. Pour plus d'informations, voir «Graphiques et diagrammes de catégorie».
- **Vue Clusters.** Cette vue contient deux graphique Relations : *Relations par concept* et *Relations par cluster*. Pour plus d'informations, voir «Graphiques Cluster», à la page 157.
- **Vue Analyse des liens du texte.** Cette vue contient deux graphique Relations : *Relations par concept* et *Relations par type*. Pour plus d'informations, voir «Graphiques Analyse des liens du texte», à la page 158.

Pour plus d'informations sur les palettes et les barres d'outils générales utilisées pour l'édition des graphiques, voir la section correspondante dans l'aide en ligne ou dans le fichier *ModelerSPOnodes.pdf*, disponible avec le téléchargement de votre produit.

Graphiques et diagrammes de catégorie

Lorsque vous générez vos catégories, prenez le temps d'examiner les définitions de catégories, les documents ou enregistrements qu'elles contiennent, et la manière dont les catégories se chevauchent. La sous-fenêtre de visualisation offre plusieurs perspectives sur les catégories. Elle est située dans l'angle supérieur droit de la vue Catégories et concepts . Si elle est masquée, vous pouvez y accéder à partir du menu Vue (**Vue > Sous-fenêtres > Visualisation**).

Dans cette vue, la sous-fenêtre de visualisation offre trois perspectives sur les communités dans une catégorisation de document ou d'enregistrement . Vous pouvez vous appuyer sur les graphiques et diagrammes de cette sous-fenêtre pour analyser les résultats de la catégorisation, et affiner les catégories ou générer des rapports. Par exemple, lorsque vous affinez les catégories, vous pouvez utiliser cette sous-fenêtre pour examiner les définitions de catégories et vous rendre compte que certaines d'entre elles sont trop similaires (c'est-à-dire qu'elles partagent plus de 75% de leurs documents ou enregistrements) ou trop différentes. Si deux catégories sont trop similaires, vous pouvez décider de les combiner. Vous pouvez également décider d'affiner les définitions des catégories en supprimant certains descripteurs d'une catégorie.

En fonction des éléments sélectionnés dans la sous-fenêtre Résultats d'extraction, de la sous-fenêtre Catégories ou de la boîte de dialogue Définitions de catégories, vous pouvez afficher les interactions correspondantes entre les documents/enregistrements et les catégories dans chacun des onglets. Chacun présente des informations similaires, mais d'une façon différente ou avec un niveau de détail différent. Toutefois, de façon à actualiser un graphique pour la sélection actuelle, cliquez sur **Afficher** dans la barre d'outils de la sous-fenêtre ou dans la boîte de dialogue dans laquelle vous avez effectué votre sélection.

La sous-fenêtre Visualisation de la vue Catégories et Concepts propose les graphiques et diagrammes suivants :

- **Diagramme à barres de catégorie.** Une table et un diagramme à barres présentent le chevauchement entre les documents/enregistrements correspondant à votre sélection et aux catégories associées. Le diagramme à barres présente également le rapport entre lesdocuments/enregistrements dans les catégories et le nombre total de documents/enregistrements. Pour plus d'informations, voir «Graphique à Barres Catégorie», à la page 156.

- **Graphique Relations de catégorie.** Ce graphique présente le chevauchement de documents/enregistrements pour les catégories auxquelles les documents/enregistrements appartiennent en fonction de la sélection dans les autres sous-fenêtres. Pour plus d'informations, voir «Graphique Relations de catégorie».
- **Tableau des relations de catégorie.** Ce tableau présente les mêmes informations que l'onglet Relations de catégorie, mais sous forme de tableau. Il contient trois colonnes qu'il est possible de trier en cliquant sur leur en-tête. Pour plus d'informations, voir «Tableau des relations de catégorie», à la page 157.

Pour plus d'informations, voir Chapitre 9, «Catégorisation des données textuelles», à la page 95.

Graphique à Barres Catégorie

Cet onglet affiche une table et un diagramme à barres présentant le chevauchement entre les documents/enregistrements correspondant à votre sélection et aux catégories associées. Le diagramme à barres présente également le rapport entre lesdocuments/enregistrements dans les catégories et le nombre total de documents ou d'enregistrements . Vous ne pouvez pas modifier la présentation de ce diagramme. Toutefois, vous pouvez trier les colonnes en cliquant sur leur en-tête.

Le diagramme contient les colonnes suivantes :

- **Catégorie.** Cette colonne présente le nom des catégories de votre sélection. Par défaut, la catégorie la plus courante de votre sélection est répertoriée en premier.
- **Barre.** Cette colonne présente, de manière visuelle, le rapport entre les documents ou enregistrements d'une catégorie donnée et le nombre total de documents ou enregistrements.
- **Sélection %.** Cette colonne présente un pourcentage basé sur le rapport entre le nombre total de documents ou enregistrements pour une catégorie et le nombre total de documents ou enregistrements représentés dans la sélection.
- **Docs.** Cette colonne présente le nombre de documents ou d'enregistrements dans la sélection pour la catégorie donnée.

Graphique Relations de catégorie

Cet onglet affiche un graphique Relations de catégorie. Les relations présentent le chevauchement de documents/enregistrements pour les catégories auxquelles les documents/enregistrements appartiennent en fonction de la sélection dans les autres sous-fenêtres. S'il existe des libellés de catégorie, celles-ci apparaissent dans le graphique. Vous pouvez sélectionner la présentation du graphique (réseau, cercle, orientée ou grille) à l'aide des boutons de la barre d'outils de cette sous-fenêtre.

Dans les relations, chaque noeud représente une catégorie. A l'aide de la souris, vous pouvez sélectionner et déplacer les noeuds au sein de la sous-fenêtre. La taille du noeud représente la taille relative basée sur le nombre de documents ou d'enregistrements pour cette catégorie dans votre sélection. L'épaisseur et la couleur de la ligne entre deux catégories représentent le nombre de documents ou d'enregistrements communs qu'elles contiennent. Si vous pointez sur un noeud en mode d'interaction, une info-bulle affiche le nom (ou le libellé) de la catégorie et le nombre global de documents ou enregistrements dans la catégorie.

Remarque : Par défaut, le mode d'interaction est activé pour les graphiques sur lesquels vous pouvez déplacer des noeuds. Toutefois, vous pouvez passer en mode d'édition pour modifier la présentation de vos graphiques, notamment les couleurs, polices et légendes. Pour plus d'informations, voir «Utilisation des palettes et des barres d'outils de graphiques», à la page 159.

Si vous copiez les données de graphique, à l'aide du bouton **Copier les données de la visualisation** et que vous les collez dans un tableur ou un éditeur de texte, vous pouvez constater que ces données reçoivent des en-têtes de colonnes, tels que V1, V2, etc. (jusqu'à V7). Ces colonnes contiennent les informations suivantes :

- **V1, V2** Ces valeurs correspondent aux coordonnées d'écran (X et Y, respectivement).

- **V3, V5** Indique le concept de la catégorie.
- **Taille, V6** Indique le nombre de documents dans lesquels les concepts ont été trouvés.
- **V7** Actuellement inutilisé.

Tableau des relations de catégorie

Cet onglet présente les mêmes informations que l'onglet Relations de catégorie, mais sous forme de tableau. Le tableau contient trois colonnes qu'il est possible de trier en cliquant sur leur en-tête :

- **Comptage.** Cette colonne présente le nombre de documents ou d'enregistrements que les deux catégories partagent ou ont en commun.
- **Catégorie 1.** Cette colonne présente le nom de la première catégorie suivie du nombre total (entre parenthèses) de documents ou d'enregistrements qu'elle contient.
- **Catégorie 2.** Cette colonne présente le nom de la deuxième catégorie suivie du nombre total (entre parenthèses) de documents ou d'enregistrements qu'elle contient.

Graphiques Cluster

Après avoir créé vos clusters, vous pouvez les explorer visuellement dans les graphiques Relations de la sous-fenêtre Visualisation. La sous-fenêtre visualisation offre deux perspectives de groupement : un graphique Relations par concept et un graphique Relations par cluster. Il est possible d'utiliser les graphiques Relations de cette sous-fenêtre pour analyser les résultats de classification non supervisée et découvrir certains concepts et règles que vous pouvez ajouter à vos catégories. La sous-fenêtre Visualisation est située dans l'angle supérieur droit de la vue Clusters. S'il est masqué, vous pouvez y accéder à partir du menu Vue (**Vue > Sous-fenêtres > Visualisation**). En sélectionnant un cluster dans la sous-fenêtre Cluster, vous pouvez afficher automatiquement les graphiques correspondants dans la sous-fenêtre Visualisation.

Remarque : par défaut, les graphiques sont en mode interactif/sélection dans lequel vous pouvez déplacer des noeuds. Toutefois, vous pouvez modifier la présentation de vos graphiques en mode d'édition, notamment les couleurs, polices et légendes. Pour plus d'informations, voir «Utilisation des palettes et des barres d'outils de graphiques», à la page 159.

La vue Clusters inclut deux graphiques Relations.

- **Graphique Relations par concept**. Ce graphique présente tous les concepts des clusters sélectionnés, ainsi que des concepts liés en dehors du cluster. Ce graphique peut vous aider à visualiser la façon dont les concepts sont liés au sein d'un cluster, ainsi que d'éventuels liens externes. Pour plus d'informations, voir «Graphique Relations par concept».
- **Graphique Relations par cluster**. Ce graphique présente sous la forme de lignes en pointillé les clusters sélectionnés avec tous les liens externes entre eux. Pour plus d'informations, voir «Graphique Relations par cluster», à la page 158.

Pour plus d'informations, voir Chapitre 10, «Analyse des clusters», à la page 143.

Graphique Relations par concept

Cet onglet affiche un graphique Relations représentant tous les concepts des clusters sélectionnés, ainsi que les concepts liés en dehors du cluster. Ce graphique peut vous aider à visualiser la façon dont les concepts sont liés au sein d'un cluster, ainsi que d'éventuels liens externes. Dans un cluster, chaque concept est représenté sous la forme d'un noeud avec un code de couleur en fonction de la couleur de type. Pour plus d'informations, voir «Création de types», à la page 189.

Les liens internes entre les concepts d'un cluster sont tracés et l'épaisseur des lignes représentant chaque lien est directement liée aux effectifs des documents pour la co-occurrence de chaque paire de concepts

ou à la valeur du lien de similarité, selon votre choix dans la barre d'outils du graphique. Les liens externes entre les concepts d'un cluster et les concepts situés en dehors du cluster sont également représentés.

Si des concepts sont sélectionnés dans la boîte de dialogue Définitions du cluster, le graphique Relations par concept affiche ces concepts et tout lien interne et externe associé à ces concepts. Les liens entre d'autres concepts, qui n'incluent pas l'un des concepts sélectionnés, n'apparaissent pas sur le graphique.

Remarque : Par défaut, les graphiques sont en mode interactif/sélection dans lequel vous pouvez déplacer des noeuds. Toutefois, vous pouvez modifier la présentation de vos graphiques en mode d'édition, notamment les couleurs, polices et légendes. Pour plus d'informations, voir «Utilisation des palettes et des barres d'outils de graphiques», à la page 159.

Si vous copiez les données de graphique, à l'aide du bouton **Copier les données de la visualisation** et que vous les collez dans un tableur ou un éditeur de texte, vous pouvez constater que ces données reçoivent des en-têtes de colonnes, tels que V1, V2, etc. (jusqu'à V7). Ces colonnes contiennent les informations suivantes :

- **V1, V2** Ces valeurs correspondent aux coordonnées d'écran (X et Y, respectivement).
- **V3, V6** Indique le type de concept.
- **V4, V5** Indique le libellé du concept.
- **V7** Actuellement inutilisé.

Graphique Relations par cluster

Cet onglet affiche un graphique Relations représentant les clusters sélectionnés. Les liens externes entre les clusters sélectionnés, ainsi que les liens entre d'autres clusters apparaissent tous sous la forme de lignes en pointillé. Dans un graphique Relations par cluster, chaque noeud représente la totalité d'un cluster et l'épaisseur des lignes tracées entre eux représente le nombre de liens externes entre deux clusters.

Important ! Pour pouvoir afficher un graphique Relations par cluster, vous devez avoir créé des clusters présentant des liens externes. Les liens externes sont des liens entre des paires de concepts dans des clusters distincts (un concept au sein d'un cluster et un concept en dehors, dans un autre cluster).

Par exemple, prenons deux clusters. `Cluster A` a trois concepts : `A1`, `A2` et `A3`. `Cluster B` a deux concepts : `B1` et `B2`. Les concepts suivants sont liés : `A1-A2`, `A1-A3`, `A2-B1` (External), `A2-B2` (External), `A1-B2` (externe) et `B1-B2`. En d'autres termes, dans le graphique Relations par cluster, l'épaisseur de ligne représente les trois liens externes.

Remarque : par défaut, les graphiques sont en mode interactif/sélection dans lequel vous pouvez déplacer des noeuds. Toutefois, vous pouvez modifier la présentation de vos graphiques en mode d'édition, notamment les couleurs, polices et légendes. Pour plus d'informations, voir «Utilisation des palettes et des barres d'outils de graphiques», à la page 159.

Graphiques Analyse des liens du texte

Après avoir extrait vos motifs Analyse des liens du texte (TLA), vous pouvez les explorer visuellement dans les graphiques Relations de la sous-fenêtre Visualisation. La sous-fenêtre visualisation offre deux perspectives de motifs TLA : un graphique Relations par concept et un graphique Relations par cluster. Il est possible d'utiliser les graphiques Relations de cette sous-fenêtre pour représenter visuellement des motifs. La sous-fenêtre Visualisation est située dans l'angle supérieur droit de la vue Analyse des liens du texte. S'il est masqué, vous pouvez y accéder à partir du menu Vue (**Vue > Sous-fenêtres > Visualisation**). En l'absence de sélection, la zone de graphique est vide.

Remarque : par défaut, les graphiques sont en mode interactif/sélection dans lequel vous pouvez déplacer des noeuds. Toutefois, vous pouvez modifier la présentation de vos graphiques en mode d'édition, notamment les couleurs, polices et légendes. Pour plus d'informations, voir «Utilisation des palettes et des barres d'outils de graphiques».

La vue Analyse des liens du texte inclut deux graphiques Relations.

- **Graphique Relations par concept**. Ce graphique présente tous les concepts figurant dans les motifs sélectionnés. Dans un graphique de concept, l'épaisseur des lignes et la taille des noeuds (lorsque les icônes de type ne sont pas représentées) indiquent le nombre d'occurrences globales figurant dans la table sélectionnée. Pour plus d'informations, voir «Graphique Relations par concept».
- **Graphique Relations par type**. Ce graphique présente tous les types figurant dans les motifs sélectionnés. Dans un graphique de type, l'épaisseur des lignes et la taille des noeuds (lorsque les icônes de type ne sont pas représentées) indiquent le nombre d'occurrences globales figurant dans la table sélectionnée. Les noeuds sont représentés soit par une couleur de type, soit par une icône. Pour plus d'informations, voir «Graphique Relations par type».

Pour plus d'informations, voir Chapitre 11, «Exploration de l'analyse des liens du texte», à la page 149.

Graphique Relations par concept

Ce graphique Relations présente tous les concepts représentés dans la sélection actuelle. Par exemple, si vous avez sélectionné un motif de type avec trois patrons de concept correspondants, le graphique affiche trois ensembles de concepts liés. L'épaisseur de ligne et la taille des noeuds d'un graphique de concept représentent les effectifs de fréquences globaux. Le graphique représente visuellement des informations identiques aux éléments sélectionnés dans les sous-fenêtres de motifs. Le type de chaque concept est présenté par une couleur ou une icône, selon ce que vous avez sélectionné dans la barre d'outils du graphique. Pour plus d'informations, voir «Utilisation des palettes et des barres d'outils de graphiques».

Graphique Relations par type

Ce graphique Relations représente chaque motif de type pour la sélection actuelle. Par exemple, si vous avez sélectionné deux patrons de concept, le graphique affiche un noeud par type dans les motifs sélectionnés et les liens entre eux détectés dans le même motif. L'épaisseur de ligne et la taille des noeuds représentent les effectifs de fréquences globaux pour l'ensemble. Le graphique représente visuellement des informations identiques aux éléments sélectionnés dans les sous-fenêtres de motifs. Outre les noms de type qui apparaissent dans le graphique, les types sont également identifiés grâce à leur couleur ou à une icône de type, selon ce que vous avez sélectionné dans la barre d'outils du graphique. Pour plus d'informations, voir «Utilisation des palettes et des barres d'outils de graphiques».

Utilisation des palettes et des barres d'outils de graphiques

Pour chaque graphique, une barre d'outils fournit un accès rapide à certaines palettes courantes avec lesquelles vous pouvez effectuer un certain nombre d'actions sur vos graphiques. Chaque vue (Catégories et concepts, Clusters et Analyse des liens du texte) présente une barre d'outils légèrement différente. Vous pouvez choisir entre le mode de vue *Sélection/Interaction* ou le mode de vue *Edition*.

Tandis que le mode d'interaction vous permet d'explorer de manière analytique les données et les valeurs représentées par la visualisation, le mode d'édition vous permet de modifier la mise en forme et l'aspect de la visualisation. Vous pouvez par exemple modifier les polices et les couleurs pour respecter le guide de style de votre organisation. Pour sélectionner ce mode, sélectionnez **Vue** > **Sous-fenêtre Visualisation** > **Mode d'édition** dans les menus (ou cliquez sur l'icône de la barre d'outils).

En mode d'édition, il existe plusieurs barres d'outils permettant de modifier l'aspect de la présentation des visualisations. Si vous n'utilisez pas certaines barres d'outils, vous pouvez les masquer afin

d'augmenter l'espace disponible dans la boîte de dialogue dans laquelle le graphique est affiché. Pour sélectionner ou désélectionner des barres d'outils, cliquez sur le nom de la barre d'outils ou de la palette souhaitée dans le menu Vue.

Pour plus d'informations sur les palettes et les barres d'outils générales utilisées pour l'édition des graphiques, voir la section correspondante dans l'aide en ligne ou dans le fichier *ModelerSPOnodes.pdf*, disponible avec le téléchargement de votre produit.

Tableau 36. Boutons de la barre d'outil de Text Analytics.

Bouton/Liste	Description
	Active le mode d'édition. Basculez vers le mode d'édition pour modifier l'apparence du graphique, par exemple pour agrandir la police, modifier les couleurs pour respecter le guide de style de l'entreprise, ou supprimer des libellés et des légendes.
	Active le mode d'interaction. Par défaut, le mode de sélection/d'interaction est activé, ce qui signifie que vous pouvez déplacer et faire glisser des noeuds dans le graphique et pointer sur des objets du graphique pour obtenir des informations supplémentaires via une info-bulle.
	Sélectionnez un type de vue des relations pour les graphiques dans les vues Catégories et concepts et Analyse des liens du texte. • **Présentation Cercle** Présentation générale qu'il est possible d'appliquer à n'importe quel graphique. Pour créer un graphique, elle suppose que les liens ne sont pas orientés et traite tous les noeuds de la même façon. Les noeuds sont placés uniquement autour du périmètre d'un cercle. • **Présentation Réseau** Présentation générale qu'il est possible d'appliquer à n'importe quel graphique. Pour créer un graphique, elle suppose que les liens ne sont pas orientés et traite tous les noeuds de la même façon. Les noeuds sont placés librement au sein de la présentation. • **Présentation Orientée** Il convient d'utiliser cette présentation uniquement pour les graphiques orientés. Cette présentation produit des structures ressemblant à un arbre, des noeuds racine aux noeuds feuille, et fournit une organisation par couleurs. Les données hiérarchiques s'affichent très bien dans cette présentation. • **Présentation Grille** Présentation générale qu'il est possible d'appliquer à n'importe quel graphique. Pour créer un graphique, elle suppose que les liens ne sont pas orientés et traite tous les noeuds de la même façon. Les noeuds sont placés uniquement aux points de grille au sein de l'espace.
	Représentation de la taille des liens. Choisissez ce que l'épaisseur de la ligne représente dans le graphique. S'applique uniquement à la vue Clusters. Le graphique Relations par cluster affiche uniquement le nombre de liens externes entre les clusters. Vous avez le choix entre : • **Similarité** L'épaisseur indique le nombre de liens externes entre deux clusters • **Co-occurrence** L'épaisseur indique le nombre de documents dans lesquels une co-occurrence de descripteurs a lieu.
	Bouton bascule qui, lorsqu'il est pressé, affiche la légende. Lorsque le bouton est désactivé, la légende n'apparaît pas.
	Bouton bascule qui, lorsqu'il est activé, affiche les icônes de type dans le graphique plutôt que les couleurs de type. S'applique uniquement à la vue Analyse des liens du texte.
	Bouton bascule qui, lorsqu'il est pressé, affiche le curseur des liens sous le graphique. Vous pouvez filtrer les résultats en faisant glisser la flèche.
	Il affichera le graphique pour le niveau de catégories sélectionné le plus élevé plutôt que celui de leurs sous-catégories.

Tableau 36. Boutons de la barre d'outil de Text Analytics (suite).

Bouton/Liste	Description
	Il affichera le graphique pour le niveau de catégories sélectionné le plus bas.
	Cette option contrôle le mode d'affichage des noms de sous-catégories dans les résultats. • **Chemin d'accès complet à la catégorie** Cette option va générer le nom de la catégorie et le chemin complet aux catégories parents le cas échéant en utilisant des barres obliques pour séparer les noms de catégories des noms de sous-catégories. • **Chemin d'accès court à la catégorie** Cette option va générer seulement le nom de la catégorie mais utilise des point de suspension pour afficher le nombre de catégories parent pour la catégorie en question. • **Catégorie de niveau le plus bas** Cette option va générer seulement le nom de la catégorie sans afficher le chemin complet ou les catégories parent.

Chapitre 13. Editeur de ressources de session

IBM SPSS Modeler Text Analytics capture et extrait rapidement et avec précision des concepts-clés de données texte. Ce processus d'extraction repose principalement sur les ressources linguistiques afin de déterminer la façon d'extraire des informations des données textuelles. Par défaut, les ressources proviennent de modèles de ressources.

IBM SPSS Modeler Text Analytics est fourni avec des **modèles de ressources** spécialisés qui contiennent des ressources linguistiques et non linguistiques sous forme de bibliothèques et de ressources avancées, pour vous permettre de définir le traitement et l'extraction des données. Pour plus d'informations, voir Chapitre 14, «Modèles et ressources», à la page 167.

Dans la boîte de dialogue de noeud, vous pouvez charger une copie des ressources du modèle vers le noeud. Dans une session de plan de travail interactif, vous pouvez personnaliser ces ressources en fonction des données du noeud si vous le désirez. Au cours d'une session de plan de travail interactif, vous pouvez utiliser vos ressources dans la vue de Editeur de ressources. Lorsque vous lancez une session interactive, une extraction est effectuée en utilisant les ressources chargées dans la boîte de dialogue du noeud, si vous n'avez pas placé les données et les résultats d'extraction dans le noeud.

Modification des ressources dans l'éditeur de ressources

L'Editeur de ressources permet d'accéder aux ressources utilisées pour produire les résultats de l'extraction (concepts, types et motifs) pour une session de plan de travail interactif. Cet éditeur est très similaire à l'Editeur de modèle, sauf que dans l'Editeur de ressources vous modifiez les ressources de la session. Une fois que vous avez terminé d'utiliser les ressources et toute autre tâche, vous pouvez mettre à jour le noeud modélisation pour enregistrer ce travail pour pouvoir le restaurer dans une session de plan de travail interactif suivant. Pour plus d'informations, voir «Mise à jour des noeuds modélisation et enregistrement», à la page 78.

Si vous voulez travailler directement dans les modèles utilisés pour charger les ressources dans les noeuds, il est recommandé d'utiliser l'Editeur de modèle. La plupart des tâches que vous pouvez exécuter dans l'Editeur de ressources s'exécutent de la même manière dans l'Editeur de modèle, à savoir :

- **Utilisation de bibliothèques** Pour plus d'informations, voir Chapitre 15, «Utilisation des bibliothèques» , à la page 177.
- **Création de dictionnaires de types** Pour plus d'informations, voir «Création de types», à la page 189.
- **Ajout de termes à la déclaration** Pour plus d'informations, voir «Ajout de termes», à la page 190.
- **Création de synonymes**. Pour plus d'informations, voir «Définition de synonymes», à la page 196.
- **Import et export des modèles** Pour plus d'informations, voir «Import et export des modèles», à la page 175.
- **Publication de bibliothèques**. Pour plus d'informations, voir «Publication de bibliothèques», à la page 183.

Pour le texte en néerlandais, en anglais, en français, en allemand, en italien, en portugais, et en espagnol

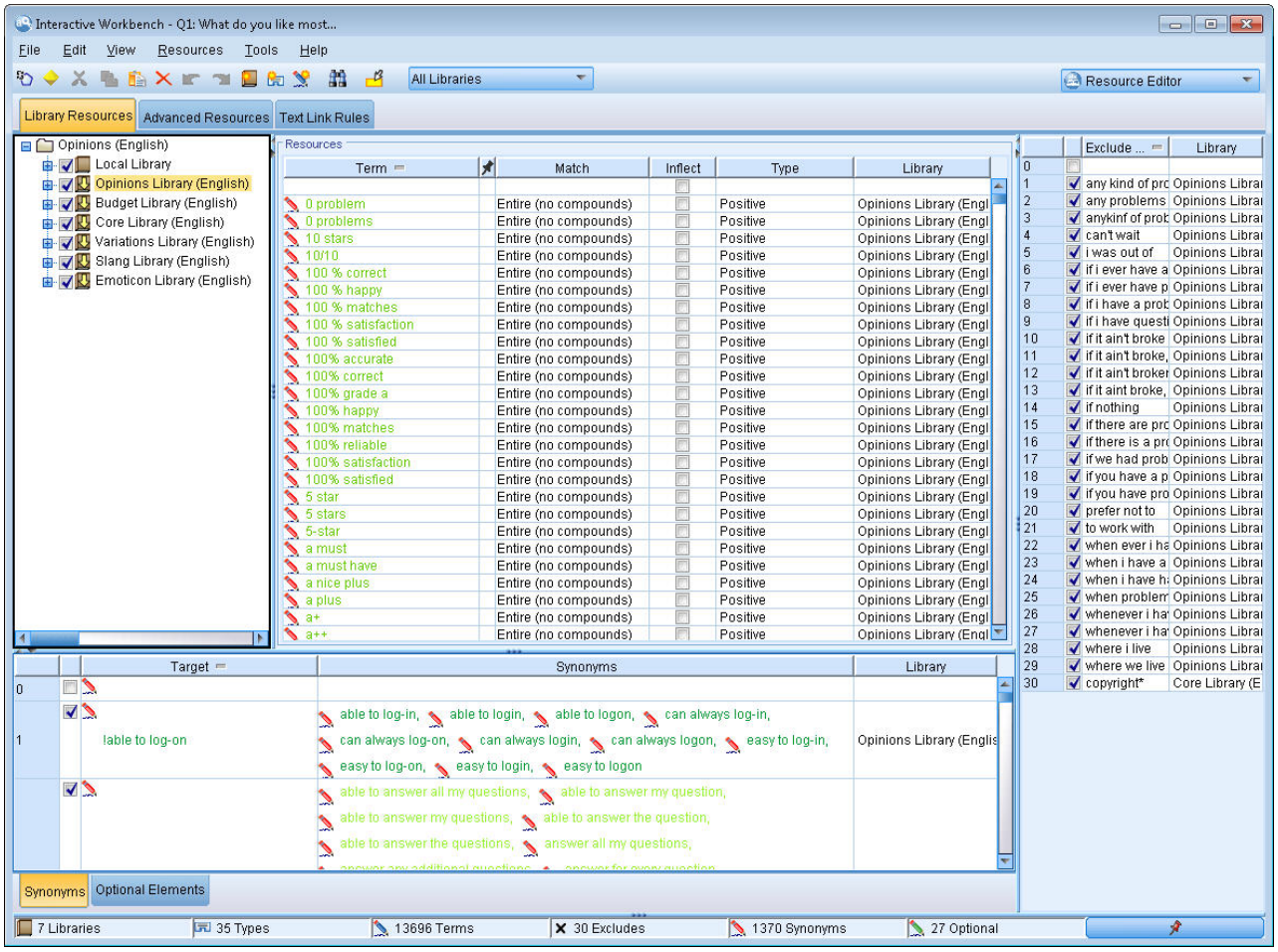


Figure 35. Vue Editeur de ressources

Création et mise à jour de modèles

Chaque fois que vous apportez des modifications à vos ressources et que vous souhaitez les réutiliser ultérieurement, enregistrez-les en tant que modèle. Lorsque vous effectuez cette opération, vous pouvez choisir d'enregistrer vos ressources en utilisant le nom d'un modèle existant ou en indiquant un nouveau nom. Ensuite, chaque fois que vous chargerez ce modèle, vous obtiendrez les mêmes ressources. Pour plus d'informations, voir la rubrique «Copie des ressources à partir de modèles et de TAP», à la page 26.

Remarque : vous pouvez également publier et partager vos bibliothèques. Pour plus d'informations, voir «Partage de bibliothèques», à la page 182.

Pour créer (ou mettre à jour) un modèle

1. Dans les menus de la vue Editeur de ressources, choisissez **Ressources > Créer un modèle de ressource**. La boîte de dialogue Créer un modèle de ressource apparaît.
2. Entrez un nouveau nom dans le champ Nom du modèle, si vous créez un modèle. Sélectionnez un modèle dans le tableau si vous souhaitez remplacer un modèle existant par les ressources chargées.
3. Cliquez sur **Enregistrer** pour créer le modèle.

Important ! Etant donné que les modèles sont chargés lorsque vous les sélectionnez dans le noeud et non pas lorsque le flux est exécuté, veillez à recharger le modèle de ressources dans tous les autres noeuds

dans lesquels il est utilisé pour obtenir les dernières modifications apportées. Pour plus d'informations, voir «Mise à jour des ressources d'un noeud après le chargement», à la page 173.

Changement des modèles de ressources

Si vous voulez remplacer les ressources chargées dans la session par une copie de celles d'un autre modèle, vous pouvez changer ces modèles. Cette action remplacera toutes les ressources chargées dans la session. Si vous remplacez des ressources pour disposer de règles de motifs Analyse des liens du texte (TLA) prédéfinies, veillez à sélectionner un modèle qui les contient dans la colonne TLA.

Remplacer des ressources est particulièrement utile lorsque vous souhaitez restaurer le travail d'une session (catégories, motifs et ressources) tout en chargeant une copie mise à jour des ressources à partir d'un modèle pour ne pas perdre le reste du travail de votre session.Vous pouvez sélectionner le modèle dont vous souhaitez copier le contenu dans l'Editeur de ressources puis cliquer sur **OK**. Ceci remplace les ressources que vous avez dans cette session. Veillez à mettre à jour le noeud modélisation à la fin de la session de manière à retrouver les modifications apportées lors de la prochaine session de plan de travail interactif.

Remarque : Si vous changez de modèle au cours d'une session interactive, le nom du modèle listé dans le noeud reste celui du dernier modèle chargé et copié. Pour pouvoir tirer parti de ces ressources ou du travail d'une autre session, mettez à jour le noeud modélisation avant de quitter la session et sélectionnez l'option **Utiliser le travail d'une session** dans le noeud. Pour plus d'informations, voir «Mise à jour des noeuds modélisation et enregistrement», à la page 78.

Pour changer de modèle

1. Dans les menus de la vue Editeur de ressources, choisissez **Ressources > Changer de modèles de ressource**. La boîte de dialogue Changer de modèle apparaît.
2. Sélectionnez le modèle à utiliser parmi ceux répertoriés dans le tableau.
3. Cliquez sur **OK** pour abandonner les ressources chargées et charger une copie de celles contenues dans le modèle sélectionné à la place. Si vous avez apporté des modifications à vos ressources et souhaitez enregistrer vos bibliothèques pour un usage ultérieur, vous pouvez les publier, les mettre à jour et les partager avant de passer à un autre modèle. Pour plus d'informations, voir «Partage de bibliothèques», à la page 182.

Chapitre 14. Modèles et ressources

IBM SPSS Modeler Text Analytics capture et extrait rapidement et avec précision des concepts-clés de données texte. Ce processus d'extraction repose principalement sur les ressources linguistiques afin de déterminer la façon d'extraire des informations des données textuelles. Pour plus d'informations, voir «Fonctionnement de l'extraction», à la page 5. Vous pouvez affiner ces ressources dans la vue Editeur de ressources.

Lorsque vous installez le logiciel, vous obtenez aussi un ensemble de ressources spécialisées. Ces ressources fournies vous permettent de bénéficier d'années de recherche et d'un réglage pour des langues et applications spécifiques. Etant donné que les ressources fournies ne sont pas toujours parfaitement adaptées au contexte de vos données, vous pouvez modifier ces modèles de ressources, ou même créer et utiliser des bibliothèques personnalisées adaptées aux données de votre entreprise. Ces ressources se présentent sous des formes variées et peuvent être utilisées dans votre session. Les ressources se trouvent aux endroits suivants :

- **Modèles de ressources**. Les modèles sont constitués d'un ensemble de bibliothèques, de types et de ressources avancées qui forment un ensemble spécialisé de ressources adapté à un domaine ou contexte particulier comme les opinions sur des produits.
- **Packs d'analyse de texte (TAP)**. En plus des ressources stockées dans un modèle, les TAP regroupent également un ou plusieurs ensembles de catégories générés à l'aide de ces ressources afin que les catégories et les ressources soient stockées ensemble et puissent être réutilisées. Pour plus d'informations, voir «Utilisation des packs d'analyse de texte», à la page 136.
- **Bibliothèques**. Les bibliothèques sont utilisées comme blocs de construction pour les TAP et les modèles. Elles peuvent également être ajoutées individuellement aux ressources dans votre session. Chaque bibliothèque est composée de plusieurs déclarations utilisées pour définir et gérer les types, les synonymes et les exclusions. Alors que les bibliothèques sont également fournies individuellement, elles sont prégroupées dans les modèles et les TAP. Pour plus d'informations, voir Chapitre 15, «Utilisation des bibliothèques», à la page 177.

Remarque : au cours de l'extraction, certaines ressources internes compilées sont également utilisées. Ces ressources compilées contiennent de nombreuses définitions qui complètent les types de la bibliothèque principale. Ces ressources compilées ne peuvent pas être éditées.

L'Editeur de ressources permet d'accéder aux ressources utilisées pour produire les résultats de l'extraction (concepts, types et motifs). Vous pouvez effectuer de nombreuses tâches dans l'Editeur de ressources, dont :

- **Utilisation de bibliothèques** Pour plus d'informations, voir Chapitre 15, «Utilisation des bibliothèques», à la page 177.
- **Création de dictionnaires de types** Pour plus d'informations, voir «Création de types», à la page 189.
- **Ajout de termes à la déclaration** Pour plus d'informations, voir «Ajout de termes», à la page 190.
- **Création de synonymes**. Pour plus d'informations, voir «Définition de synonymes», à la page 196.
- **Mise à jour des ressources dans les TAP**. Pour plus d'informations, voir «Mise à jour des Packs d'analyse de texte», à la page 137.
- **Création de modèles**. Pour plus d'informations, voir «Création et mise à jour de modèles», à la page 164.
- **Import et export des modèles** Pour plus d'informations, voir «Import et export des modèles», à la page 175.
- **Publication de bibliothèques**. Pour plus d'informations, voir «Publication de bibliothèques», à la page 183.

Editeur de modèle et éditeur de ressources

Vous pouvez utiliser et modifier les modèles, les bibliothèques et les ressources principalement de deux manières. Vous pouvez travailler sur les ressources linguistiques dans l'Editeur de modèle ou l'Editeur de ressources.

Editeur de modèle

L'Editeur de modèle vous permet de créer et modifier des modèles de ressources sans session de plan de travail interactif et indépendamment d'un noeud ou flux spécifique. Vous pouvez utiliser cet éditeur pour créer ou modifier des modèles de ressources avant de les charger dans le noeud d'analyse des liens du texte et le noeud modélisation Text Mining.

L'Editeur de modèle est accessible via la barre d'outils principaleIBM SPSS Modeler à partir du menu **Outils > Editeur de modèle Text Analytics**.

Editeur de ressources

L'Editeur de ressources qui est accessible dans une session de plan de travail interactif, permet d'utiliser les ressources dans le contexte d'un noeud ou d'un ensemble de données. Lorsque vous ajoutez un noeud modélisation Text Mining à un flux, vous pouvez charger une copie du contenu du modèle de ressources ou une copie d'un pack d'analyse de texte (ensembles de catégories *et* ressources) pour contrôler l'extraction de texte pour l'opération de Text Mining. Lorsque vous lancez une session de plan de travail interactif, vous pouvez créer des catégories, extraire des motifs d'analyse des liens du texte et créer des modèles de catégorie, mais également ajuster les ressources des données de la session dans une vue de l'Editeur de ressources intégrée. Pour plus d'informations, voir «Modification des ressources dans l'éditeur de ressources», à la page 163.

Lorsque vous utilisez les ressources dans une session de plan de travail interactif, les modifications s'appliquent uniquement à la session. Pour sauvegarder votre travail (ressources, catégories, motifs, etc.) et accéder à une autre session, vous devez mettre à jour le noeud modélisation. Pour plus d'informations, voir «Mise à jour des noeuds modélisation et enregistrement», à la page 78.

Si vous voulez enregistrer les modifications dans le modèle d'origine, dont le contenu a été copié vers le noeud modélisation, pour pouvoir charger le modèle mis à jour dans d'autres noeuds, vous pouvez créer un modèle depuis les ressources. Pour plus d'informations, voir «Création et mise à jour de modèles», à la page 164.

Interface de l'éditeur

Les opérations que vous effectuez dans l'Editeur de modèle ou l'Editeur de ressources concernent la gestion et l'adaptation des ressources linguistiques. Ces ressources sont stockées sous la forme de modèles et de bibliothèques. Pour plus d'informations, voir «Dictionnaires de types», à la page 187.

Onglet Ressources de bibliothèque

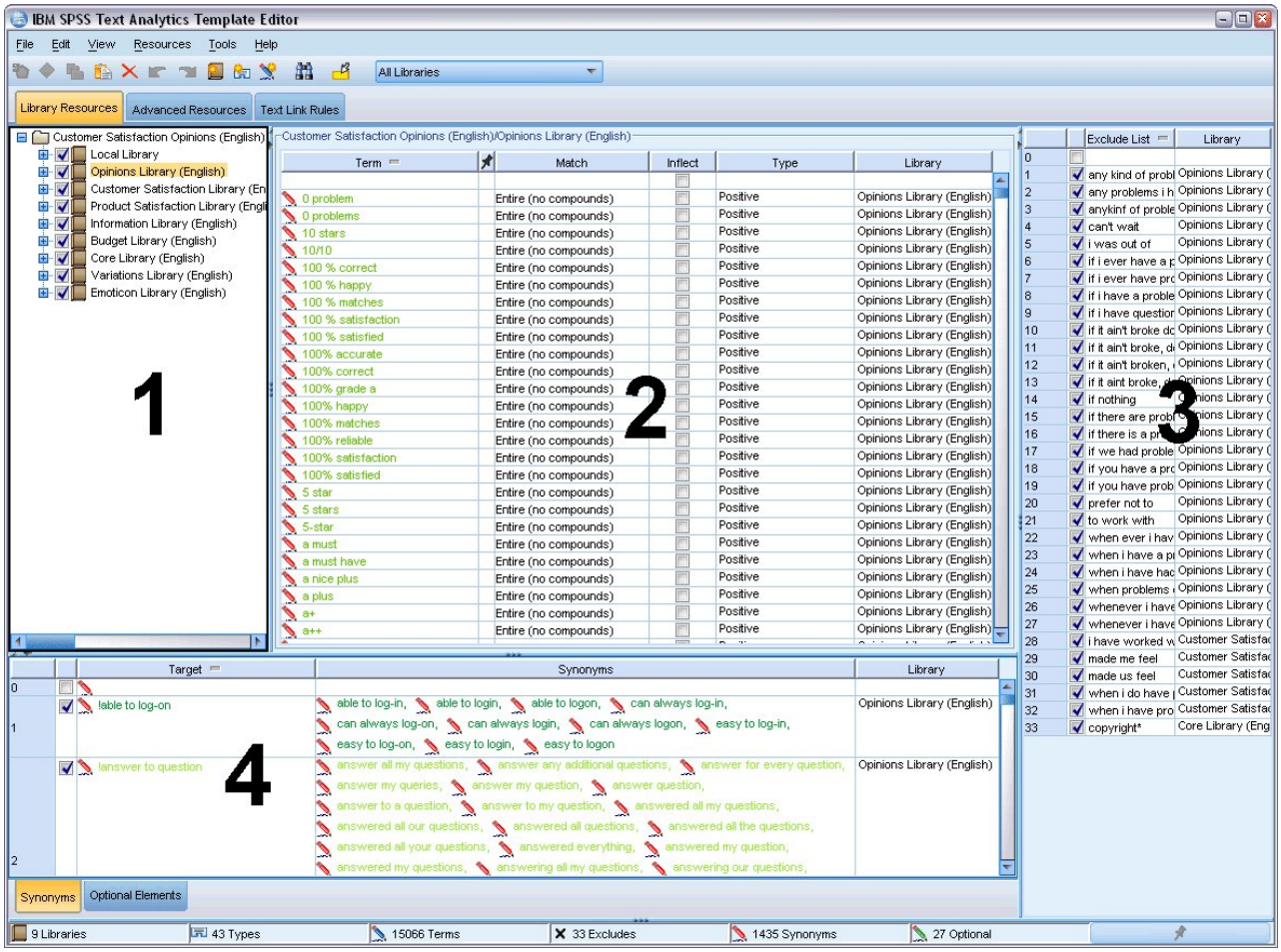


Figure 36. Editeur de modèle de Text Mining

L'interface comporte quatre parties :

1. Sous-fenêtre Arborescence des bibliothèques. Situé dans l'angle supérieur gauche, ce plan présente l'arborescence des bibliothèques. Vous pouvez activer et désactiver les bibliothèques figurant dans cette arborescence. Vous pouvez également filtrer les vues des autres sous-fenêtres en sélectionnant une bibliothèque dans l'arborescence. Vous pouvez effectuer plusieurs opérations dans cette arborescence à l'aide des menus contextuels. Lorsque vous développez une bibliothèque figurant dans l'arborescence, vous pouvez consulter l'ensemble des types qu'elle comporte. Vous pouvez également filtrer cette liste avec le menu **Vue** si vous souhaitez vous concentrer sur une bibliothèque spécifique.

2. Sous-fenêtre Liste des termes des dictionnaires de types. Cette sous-fenêtre, située à droite de l'arborescence des bibliothèques, affiche les listes de termes des dictionnaires de types correspondant aux bibliothèques sélectionnées dans l'arborescence des bibliothèques. Un **dictionnaire de types** est un ensemble de termes devant être regroupés sous un même libellé ou sous un même nom de type. Lorsque le moteur du programme d'extraction lit vos données textuelles, il compare les mots trouvés dans le texte aux termes figurant dans les dictionnaires de types. Si un concept extrait figure sous la forme d'un terme dans un dictionnaire de types, le nom du type en question lui est alors affecté. Vous pouvez considérer le dictionnaire de types comme étant un dictionnaire distinct qui regroupe les termes présentant des points communs. Par exemple, le type `<Location>` qui figure dans Core Library comporte des concepts tels que `new orleans`, `great britain` et `new york`. Ces termes désignent tous des lieux géographiques. Une bibliothèque peut comporter un ou plusieurs dictionnaires de types. Pour plus d'informations, voir «Dictionnaires de types», à la page 187.

3. Sous-fenêtre Dictionnaire d'exclusions. Cette sous-fenêtre, située sur le côté droit, affiche la collection de termes qui seront exclus des résultats finaux de l'extraction. Les termes apparaissant dans ce dictionnaire d'exclusions n'apparaissent pas dans la sous-fenêtre Résultats d'extraction. Les termes exclus peuvent être stockés dans la bibliothèque de votre choix. Toutefois, la sous-fenêtre Dictionnaire de substitutions affiche tous les contenus de toutes les bibliothèques visibles dans l'arborescence des bibliothèques. Pour plus d'informations, voir «Dictionnaires d'exclusions», à la page 198.

4. Sous-fenêtre Dictionnaire des substitutions. Cette sous-fenêtre, située en bas à gauche, affiche les synonymes et les éléments optionnels, chacun dans leur propre onglet. Les synonymes et les éléments optionnels permettent de grouper les termes similaires sous un concept principal ou cible dans les résultats d'extraction finaux. Ce dictionnaire peut comporter des synonymes connus, des synonymes et des éléments définis par l'utilisateur, ainsi que les fautes d'orthographe courantes associées à leur correction. Les définitions des synonymes et les éléments optionnels peuvent être stockés dans la bibliothèque de votre choix. Mais, la sous-fenêtre Dictionnaire de substitutions affiche tous les contenus de toutes les bibliothèques visibles dans l'arborescence des bibliothèques. Pendant que cette sous-fenêtre affiche tous les synonymes et éléments facultatifs de toutes les bibliothèques, les substitutions de toutes les bibliothèques de l'arborescence apparaissent ensemble dans cette sous-fenêtre. Une bibliothèque ne peut comporter qu'un seul dictionnaire de substitutions. Pour plus d'informations, voir «Dictionnaires de substitutions/synonymes», à la page 195.

Remarques :

- pour procéder à un filtrage visant à n'afficher que les informations propres à une seule bibliothèque, vous pouvez modifier la vue de la bibliothèque à l'aide de la liste déroulante figurant dans la barre d'outils. Elle comporte une entrée de niveau supérieur appelée **Toutes les bibliothèques**, ainsi qu'une autre entrée supplémentaire pour chacune des bibliothèques. Pour plus d'informations, voir «Affichage des bibliothèques», à la page 180.

Onglet Ressources avancées

Les ressources avancées sont disponibles dans le deuxième onglet de la vue de l'éditeur. Vous pouvez consulter et modifier les ressources avancées dans cet onglet. Pour plus d'informations, voir Chapitre 17, «A propos des ressources avancées», à la page 201.

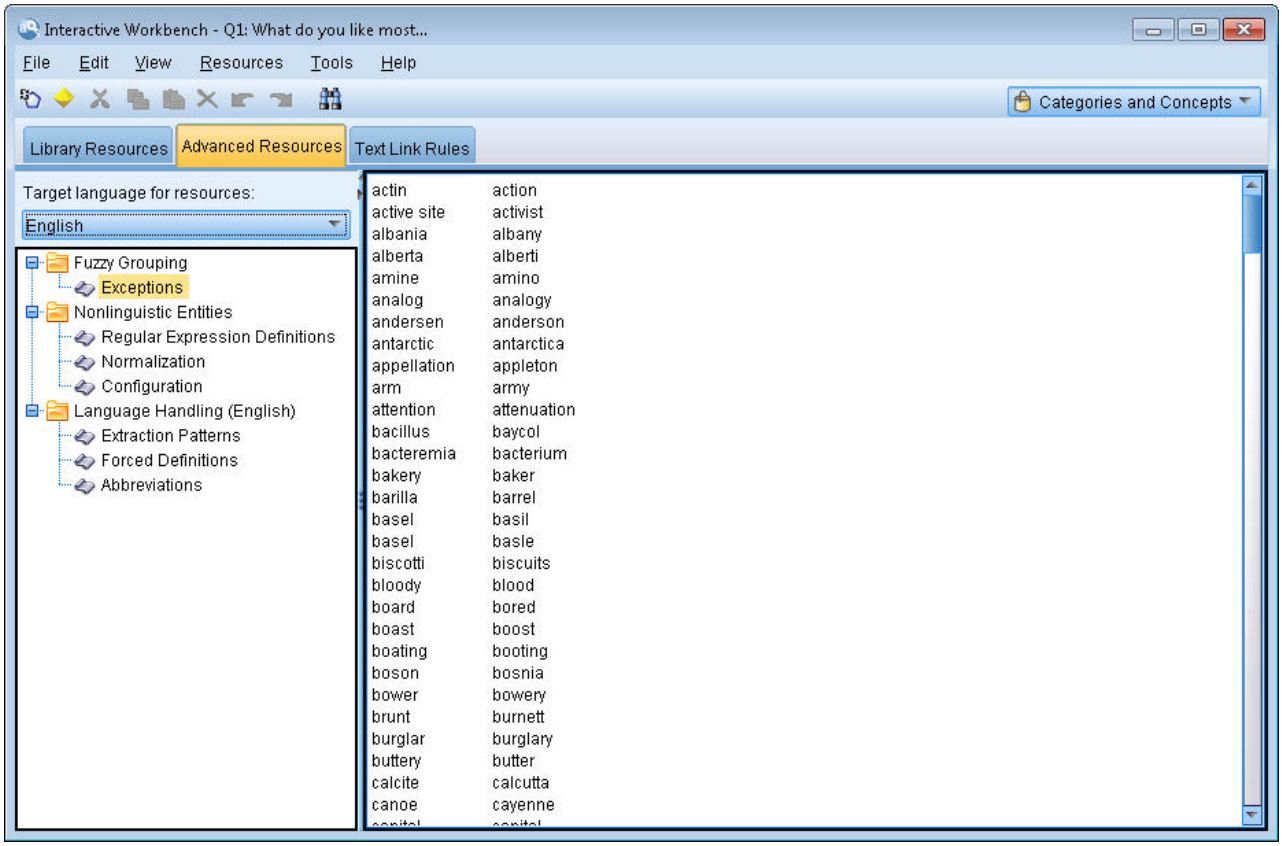


Figure 37. Editeur de modèles de Text Mining - Onglet Ressources avancées

Onglet Règles des liens du texte

Depuis la version 14, les règles d'analyse des liens du texte peuvent être modifiées dans leur propre onglet de la vue de l'éditeur. Vous pouvez utiliser l'éditeur de règles, créer vos propres règles et même effectuer des simulations pour connaître l'influence de vos règles sur les résultats TLA. Pour plus d'informations, voir Chapitre 18, «A propos des règles des liens du texte», à la page 213.

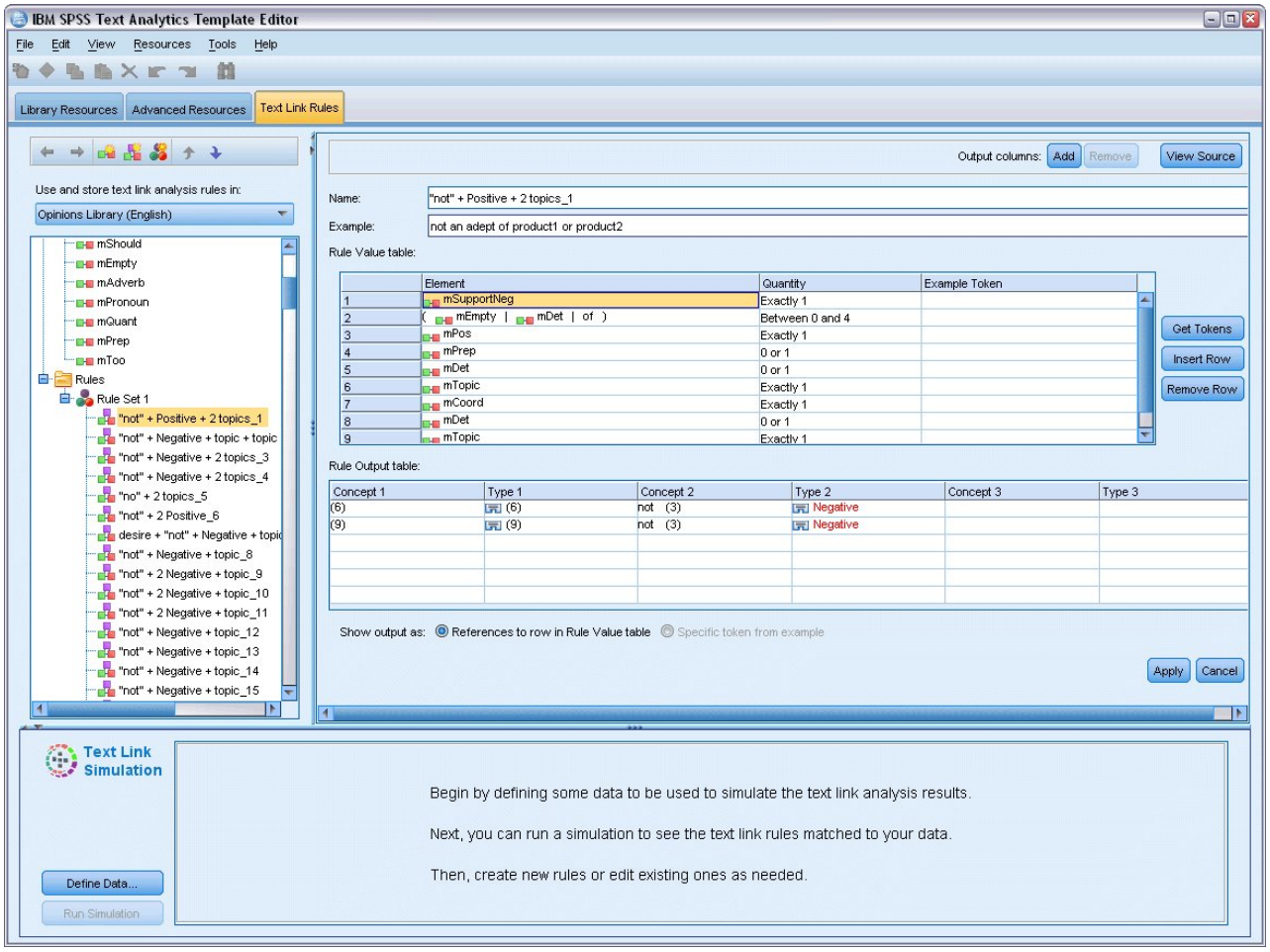


Figure 38. Editeur de modèles de Text Mining - Onglet Règles des liens du texte

Ouverture des modèles

Lorsque vous lancez l'Editeur de modèle, un message vous demande d'ouvrir un modèle. Vous pouvez également ouvrir un modèle depuis le menu Fichier. Si vous voulez un modèle contenant des règles d'analyse des liens du texte (TLA), sélectionnez un modèle ayant une icône dans la colonne TLA. La langue pour laquelle le modèle est créé figure dans la colonne Langue.

Si vous voulez importer un modèle qui ne figure pas dans le tableau ou exporter un modèle, vous pouvez utiliser les boutons de la boîte de dialogue Ouvrir un modèle. Pour plus d'informations, voir «Import et export des modèles», à la page 175.

Pour ouvrir un modèle

1. Dans les menus de l'Editeur de modèle, choisissez **Fichier > Ouvrir un modèle de ressources**. La boîte de dialogue Ouvrir un modèle de ressources apparaît.
2. Sélectionnez le modèle à utiliser parmi ceux répertoriés dans le tableau.
3. Cliquez sur **OK** pour ouvrir ce modèle. Si un modèle est ouvert dans l'éditeur, cliquez sur OK pour abandonner ce modèle et afficher le modèle que vous avez sélectionné. Si vous avez apporté des modifications à vos ressources et souhaitez enregistrer vos bibliothèques pour un usage ultérieur, vous pouvez les publier, les mettre à jour et les partager avant d'ouvrir un autre modèle. Pour plus d'informations, voir «Partage de bibliothèques», à la page 182.

Enregistrement des modèles

Dans l'Editeur de modèle, vous pouvez enregistrer les modifications d'un modèle. Vous pouvez choisir d'enregistrer vos ressources en utilisant le nom d'un modèle existant ou en indiquant un nouveau nom.

Si vous modifiez un modèle que vous avez chargé dans un noeud précédemment, vous devez recharger le contenu du modèle dans le noeud pour disposer des toutes dernières modifications. Pour plus d'informations, voir «Copie des ressources à partir de modèles et de TAP», à la page 26.

Ou bien, si vous utilisez l'option **Utiliser les informations interactives enregistrées** dans l'onglet Modèle du noeud Text Mining, ce qui implique que vous utilisez les ressources d'une session de plan de travail interactif précédente, vous devez utiliser les ressources de ce modèle dans la session interactive. Pour plus d'informations, voir «Changement des modèles de ressources», à la page 165.

Remarque : vous pouvez également publier et partager vos bibliothèques. Pour plus d'informations, voir «Partage de bibliothèques», à la page 182.

Pour enregistrer un modèle

1. Dans les menus de l'Editeur de modèle, choisissez **Fichier > Enregistrer un modèle de ressources**. La boîte de dialogue Enregistrer un modèle de ressources s'ouvre.
2. Insérez un nouveau nom dans le champ Nom du modèle, si vous souhaitez enregistrer ce modèle en tant que nouveau modèle. Sélectionnez un modèle dans le tableau si vous souhaitez remplacer un modèle existant par les ressources chargées.
3. Si vous le souhaitez, entrez une description pour afficher un commentaire ou une annotation dans le tableau.
4. Cliquez sur **Enregistrer** pour enregistrer le modèle.

Important ! Parce que les ressources des modèles ou des TAP sont chargées/copiées dans le noeud, vous devez mettre à jour ces ressources en les rechargeant si vous effectuez des modifications sur un modèle et que vous souhaitez retrouver ces modifications dans le flux existant. Pour plus d'informations, voir «Mise à jour des ressources d'un noeud après le chargement».

Mise à jour des ressources d'un noeud après le chargement

Par défaut, lorsque vous ajoutez un noeud à un flux, un ensemble de ressources d'un modèle par défaut est chargé et incorporé à votre noeud. Et si vous modifiez des modèles ou utilisez un TAP, lorsque vous les chargez, une copie de ces ressources remplace alors les ressources. Parce que les modèles et les TAP ne sont pas directement liés au noeud, les modifications effectuées sur un modèle ou un TAP ne sont pas automatiquement disponibles dans un noeud préexistant. Afin de pouvoir utiliser ces modifications, vous devez mettre à jour les ressources dans ce noeud. Les ressources peuvent être mises à jour de deux manières.

Méthode 1 : Rechargement des ressources dans l'onglet Modèle

Pour mettre à jour les ressources dans le noeud en utilisant un modèle ou un TAP nouveau ou mis à jour, vous pouvez le recharger dans l'onglet Modèle du noeud. En le rechargeant, vous remplacez la copie des ressources dans le noeud par une copie plus récente. Par souci pratique, l'heure et la date de mise à jour apparaissent dans l'onglet Modèle avec le nom du modèle d'origine. Pour plus d'informations, voir «Copie des ressources à partir de modèles et de TAP», à la page 26.

Toutefois, si vous utilisez les données d'une session interactive dans un noeud modélisation Text Mining et avez sélectionné l'option**Utiliser le travail d'une session** dans l'onglet Modèle, le travail et les ressources de la session sauvegardée sont utilisés et le bouton **Charger** est désactivé. Il est désactivé, car au cours d'une session de plan de travail interactif, vous avez choisi l'option **Mettre à jour le noeud modélisation** et vous avez conservé les catégories, les ressources et un autre travail de session. Dans ce

cas, si vous voulez utiliser ou mettre à jour ces ressources, vous pouvez essayer d'utiliser la méthode suivante de changement des ressources dans l'Editeur de ressources.

Méthode 2 : Changement de ressources dans l'Editeur de ressources

Lorsque vous voulez utiliser des ressources différentes au cours d'une session interactive, vous pouvez changer ces ressources en utilisant la boîte de dialogue Changer de modèle. Cela est particulièrement pratique si vous voulez réutiliser un travail de catégorie existant et remplacer les ressources. Dans ce cas, vous pouvez sélectionner l'option **Utiliser le travail d'une session** dans l'onglet Modèle d'un noeud modélisation Text Mining. Cette action désactive l'option de rechargement d'un modèle en utilisant la boîte de dialogue du noeud et conserve les paramètres et modifications effectuées pendant votre session. Vous pouvez alors lancer la session de plan de travail interactif en exécutant le flux et en remplaçant les ressources dans l'Editeur de ressources. Pour plus d'informations, voir «Changement des modèles de ressources», à la page 165.

Pour pouvoir conserver le travail d'une session dans les sessions suivantes, y compris les ressources, vous devez mettre à jour le noeud modélisation sans la session de plan de travail interactif pour que les ressources (et les autres données) soient enregistrées dans le noeud. Pour plus d'informations, voir «Mise à jour des noeuds modélisation et enregistrement», à la page 78.

Remarque : si vous changez de modèle au cours d'une session interactive, le nom du modèle listé dans le noeud reste celui du dernier modèle chargé et copié. Pour pouvoir tirer parti de ces ressources ou du travail d'une autre session, mettez à jour le noeud modélisation avant de quitter la session.

Gestion des modèles

Il existe également certaines tâches de gestion de base que vous pouvez effectuer occasionnellement sur vos modèles telles que l'attribution de nouveau nom, l'import et l'export, ou la suppression des modèles obsolètes. Ces tâches sont réalisées dans la boîte de dialogue Gérer les modèles. L'import et l'export des modèles permettent de les partager avec d'autres utilisateurs. Pour plus d'informations, voir «Import et export des modèles», à la page 175.

Remarque : il est impossible de renommer ou de supprimer les modèles installés ou livrés avec le produit. Si vous voulez renommer, vous pouvez ouvrir le modèle installé et en créer un nouveau avec le nom de votre choix. Vous pouvez supprimer vos modèles personnalisés ; néanmoins, si vous essayez de supprimer un modèle fourni, il sera réinitialisé à la version installée à l'origine.

Pour renommer un modèle

1. Dans les menus, sélectionnez **Ressources > Gérer les modèles de ressources**. La boîte de dialogue Gérer les modèles apparaît.
2. Sélectionnez le modèle à renommer et cliquez sur **Renommer**. La zone de nom devient un champ modifiable dans le tableau.
3. Tapez un nouveau nom et appuyez sur la touche Entrée. Un message de confirmation apparaît.
4. Si le nouveau nom vous satisfait, cliquez sur **Oui**. Sinon, cliquez sur **Non**.

Pour supprimer un modèle

1. Dans les menus, sélectionnez **Ressources > Gérer les modèles de ressources**. La boîte de dialogue Gérer les modèles apparaît.
2. Dans la boîte de dialogue Gérer les modèles, sélectionnez le modèle à supprimer.
3. Cliquez sur **Supprimer**. Un message de confirmation apparaît.
4. Cliquez sur **Oui** pour supprimer ou sur **Non** pour annuler la demande de suppression. Si vous cliquez sur **Oui**, le modèle est supprimé.

Import et export des modèles

Vous pouvez partager les modèles avec d'autres utilisateurs ou ordinateurs en les important et en les exportant. Les modèles sont stockés dans une base de données interne, mais ils peuvent être exportés sous forme de fichiers *.*lrt* sur votre disque dur.

Etant donné que dans certains cas, vous voulez importer et exporter des modèles, il existe des boîtes de dialogue qui offrent ces fonctions.

- Ouvrez la boîte de dialogue dans l'Editeur de modèle
- Boîte de dialogue Charger des ressources dans le noeud modélisation Text Mining et le noeud analyse des liens du texte.
- Boîte de dialogue Gérer les modèles dans l'Editeur de modèle et l'Editeur de ressources.

Pour importer un modèle

1. Dans la boîte de dialogue, cliquez sur **Importer**. La boîte de dialogue Importer un modèle apparaît.
2. Sélectionnez le fichier du modèle de ressources (*.*lrt*) à importer et cliquez sur **Importer**. Enregistrez le modèle importé en lui attribuant un nouveau nom ou en remplaçant le modèle existant. La boîte de dialogue se ferme et le modèle apparaît dans le tableau.

Pour exporter un modèle

1. Dans la boîte de dialogue, sélectionnez le modèle à exporter et cliquez sur **Exporter**. La boîte de dialogue Sélectionner un répertoire apparaît.
2. Sélectionnez le répertoire vers lequel vous souhaitez exporter et cliquez sur **Exporter**. La boîte de dialogue se ferme et le modèle exporté porte l'extension *.*lrt*.

Sortie de l'Editeur de modèle

Lorsque vous avez terminé de travailler dans l'Editeur de modèle, vous pouvez enregistrer votre travail et quitter l'éditeur.

Pour quitter l'Editeur de modèle

1. Dans les menus, sélectionnez **Fichier > Fermer**. La boîte de dialogue Enregistrer et fermer s'ouvre.
2. Sélectionnez **Enregistrer les modifications dans un modèle** pour enregistrer le modèle ouvert avant de fermer l'éditeur.
3. Sélectionnez **Publier les bibliothèques** pour publier les bibliothèques dans le modèle ouvert avant de fermer l'éditeur. Si vous sélectionnez cette option, un message vous demande de sélectionner les bibliothèques à publier. Pour plus d'informations, voir «Publication de bibliothèques», à la page 183.

Sauvegarde des ressources

Il se peut que vous ayez besoin, de temps à autre, de sauvegarder vos ressources par mesure de sécurité.

Important ! Lorsque vous procédez à une restauration, le contenu intégral de vos ressources est supprimé et seul le contenu du fichier de sauvegarde est accessible dans le produit. Ceci inclut tout travail en cours.

Remarque : Vous pouvez uniquement effectuer une sauvegarde et une restauration vers la version principale de votre logiciel. Par exemple, si vous effectuez une sauvegarde à partir de la version 15, vous ne pouvez pas effectuer une sauvegarde vers la version 16.

Pour sauvegarder les ressource

1. Dans les menus, sélectionnez **Ressources > Outils de sauvegarde > Sauvegarder les ressources**. La boîte de dialogue Sauvegarder apparaît.

2. Nommez votre fichier de sauvegarde et cliquez sur **Enregistrer**. La boîte de dialogue se ferme et le fichier de sauvegarde est créé.

Pour restaurer les ressources

1. Dans les menus, sélectionnez **Ressources > Outils de sauvegarde > Restaurer les ressources**. Une alerte vous avertit que la restauration va écraser le contenu actuel de votre base de données.
2. Cliquez sur **Oui** pour continuer. La boîte de dialogue apparaît.
3. Sélectionnez le fichier de sauvegarde à restaurer et cliquez sur **Ouvrir**. La boîte de dialogue se ferme et les ressources sont restaurées dans l'application.

Importation des fichiers de ressources

Si vous avez effectué des modifications directement dans les fichiers de ressources en-dehors de ce produit, vous pouvez les importer dans une bibliothèque donnée, en sélectionnant cette bibliothèque et en procédant à l'importation. Lorsque vous importez un répertoire, vous pouvez également importer l'ensemble des fichiers pris en charge dans une bibliothèque ouverte spécifique. Vous ne pouvez importer que les fichiers `*.txt`.

Chaque fichier importé doit contenir seulement une entrée par fichier, et si le contenu est structuré comme :

- Une liste de mots ou d'expressions (une par ligne). Le fichier est importé comme liste de termes d'un dictionnaire de types, où le dictionnaire de types prend le nom du fichier sans l'extension.
- Une liste d'entrées telles que `terme1 <TAB> terme2`, est ensuite importée en tant que liste de synonymes, où `terme1` est l'ensemble du terme sous-jacent et `terme2`, le terme cible.

Pour importer un fichier de ressources unique

1. Dans les menus, choisissez **Ressources > Importer des fichiers > Importer un fichier**. La boîte de dialogue Importer un fichier apparaît.
2. Sélectionnez le fichier que vous souhaitez importer et cliquez sur **Importer**. Un format interne est appliqué au contenu du fichier qui est ensuite ajouté à votre bibliothèque.

Pour importer l'ensemble des fichiers d'un répertoire

1. Dans les menus, choisissez **Ressources > Importer des fichiers > Importer un répertoire entier**. La boîte de dialogue Importer un répertoire apparaît.
2. Dans la liste **Importer**, sélectionnez la bibliothèque dans laquelle vous souhaitez importer l'ensemble des fichiers de ressources. Si vous sélectionnez l'option **Par défaut**, une bibliothèque est créée. Elle porte le nom du répertoire.
3. Sélectionnez le répertoire à partir duquel les fichiers doivent être importés. Les sous-répertoires ne seront pas lus.
4. Cliquez sur **Importer**. La boîte de dialogue se ferme et le contenu des fichiers de ressources importés apparaît dans l'éditeur sous forme de dictionnaires et de fichiers de ressources avancés.

Chapitre 15. Utilisation des bibliothèques

Les ressources utilisées par le moteur d'extraction pour extraire et regrouper les termes des données textuelles contiennent une ou plusieurs bibliothèques. Vous pouvez voir l'ensemble des bibliothèques dans l'arborescence située dans la partie supérieure gauche de l'Editeur de modèle et de l' Editeur de ressources. Les bibliothèques sont composées de trois sortes de dictionnaires : de type, de substitution et d'exclusion. Pour plus d'informations, voir Chapitre 16, «A propos des dictionnaires de bibliothèque», à la page 187.

Le modèle de ressources ou les ressources du TAP choisi comprennent plusieurs bibliothèques afin de vous permettre de procéder immédiatement à l'extraction des concepts de vos données textuelles. Cependant, vous pouvez également créer vos propres bibliothèques et les publier pour pouvoir les réutiliser. Pour plus d'informations, voir «Publication de bibliothèques», à la page 183.

Supposons, par exemple, que vous utilisez fréquemment des données textuelles relatives au secteur automobile. Après avoir analysé vos données, vous décidez de créer des ressources personnalisées pour gérer le vocabulaire propre à ce secteur d'activité. Avec l'Editeur de modèle, vous pouvez créer un modèle et, dans ce modèle, une bibliothèque pour extraire et y intégrer des termes propres au secteur automobile. Vous serez probablement amené à réutiliser ultérieurement les informations de cette bibliothèque, vous pouvez donc publier celle-ci dans un référentiel central, accessible à partir de la boîte de dialogue **Gérer les bibliothèques** ; ainsi, vous pourrez la réutiliser dans d'autres sessions de flux .

Supposons que vous souhaitez également regrouper les termes propres à différents sous-secteurs, tels que les dispositifs électroniques, les moteurs, les systèmes de refroidissement, voire un fabricant ou un marché particulier. Vous pouvez créer une bibliothèque pour chaque groupe, puis publier ces bibliothèques de manière à ce qu'elles puissent être utilisées avec plusieurs ensembles de données textuelles. De cette manière, vous pouvez ajouter les bibliothèques qui correspondent le mieux au contenu de vos données textuelles.

Remarque : des ressources supplémentaires peuvent être configurées et gérées dans l'onglet Ressources avancées. Certaines s'appliquent à toutes les bibliothèques et gèrent les entités non linguistiques, les exceptions de regroupement flou, etc. De plus, vous pouvez éditer les règles de motifs TLA, spécifiques à chaque bibliothèque, à partir de l'onglet Règles des liens du texte. Pour plus d'informations, voir Chapitre 17, «A propos des ressources avancées», à la page 201.

Bibliothèques fournies

Par défaut, plusieurs bibliothèques sont installées avec IBM SPSS Modeler Text Analytics. Vous pouvez utiliser ces bibliothèques préformatées pour accéder à des milliers de termes et synonymes prédéfinis, ainsi qu'à plusieurs types différents. Ces bibliothèques fournies sont affinées pour être utilisées dans plusieurs domaines différents et sont disponibles dans plusieurs langues.

De nombreuses bibliothèques existent mais les plus couramment utilisées sont les suivantes :

- **Bibliothèque locale.** Utilisée pour stocker les dictionnaires définis par l'utilisateur. Il s'agit d'une bibliothèque vide ajoutée par défaut à toutes les ressources. Elle contient également un dictionnaire de types vide. Cette bibliothèque est particulièrement utile lorsque vous apportez des modifications ou des améliorations directement dans les ressources (ajout d'un mot à un type, par exemple) à partir de la vue Catégories et concepts, Clusters ou Analyse des liens du texte . Dans ce cas, ces modifications et ces améliorations sont automatiquement stockées dans la première bibliothèque répertoriée dans l'arborescence de bibliothèques d' Editeur de ressources ; par défaut, il s'agit de la *bibliothèque locale*. Vous ne pouvez pas publier cette bibliothèque qui est spécifique aux données de session . Pour publier son contenu, renommez tout d'abord la bibliothèque.

- **Bibliothèque principale.** Utilisée dans la plupart des cas, puisqu'elle comprend les cinq types de base intégrés représentant les personnes, les lieux, les organisations, les produits et les éléments inconnus. Bien que seuls quelques termes soient répertoriés dans l'une des dictionnaires de types, les types représentés dans la bibliothèque principale viennent en fait compléter les types fiables détectés dans les ressources internes, compilées et livrées avec votre produit de Text Mining. Ces ressources internes et compilées contiennent des milliers de termes pour chaque type. Pour cette raison, même si vous ne voyez pas un terme dans une liste de termes des dictionnaires de types, il peut quand même être extrait et saisi avec un type Core. Cela explique la façon dont les noms comme *George* peuvent être extraits et recevoir le type `<Person>` alors que seul *John* apparaît dans le dictionnaire de types `<Person>` de la bibliothèque principale. De même, si vous n'incluez pas la bibliothèque principale, il se peut que vous aperceviez toujours ces types dans vos résultats d'extraction puisque les ressources compilées qui les contiennent sont toujours utilisées par le moteur du programme d'extraction.
- **Bibliothèque d'opinions.** Utilisé le plus souvent pour extraire des opinions et des sentiments des données textuelles. Cette bibliothèque inclut des milliers de mots représentant des attitudes, des qualificatifs ou des préférences qui, utilisés avec d'autres termes, expriment une opinion sur un sujet. Cette bibliothèque contient de nombreux types intégrés, des synonymes et des exclusions. Elle comprend également un grand nombre de règles de motifs utilisées pour l'analyse des liens du texte. Pour pouvoir utiliser les règles d'analyse des liens du texte dans cette bibliothèque et les résultats de motifs qu'elles génèrent, cette bibliothèque doit être spécifiée dans l'onglet Règles des liens du texte. Pour plus d'informations, voir Chapitre 18, «A propos des règles des liens du texte», à la page 213.
- **Bibliothèque Budget.** Utilisée pour extraire les termes faisant référence au coût d'un objet ou d'un service. La bibliothèque comprend plusieurs mots et expressions représentant des adjectifs, des qualificatifs et des jugements concernant le prix ou la qualité d'un objet ou d'un service.
- **Bibliothèque Variations.** Utilisée pour inclure les observations dans lesquelles certaines variations de langue requièrent des définitions de synonyme pour être correctement regroupées. Cette bibliothèque ne comporte que des définitions de synonyme.

Bien que certaines bibliothèques fournies en dehors des modèles ressemblent au contenu de certains modèles, les modèles ont été spécifiquement adaptés à certaines applications et contiennent des ressources avancées supplémentaires. Nous vous recommandons d'essayer d'utiliser un modèle conçu pour le genre de données textuelles que vous utilisez et d'effectuer les modifications sur ces ressources plutôt que de simplement ajouter des bibliothèques individuelles à un modèle plus générique.

Des ressources compilées sont également fournies avec IBM SPSS Modeler Text Analytics. Elles sont systématiquement utilisées au cours du processus d'extraction et contiennent un grand nombre de définitions complémentaires dans les dictionnaires de types intégrés aux bibliothèques par défaut. Etant donné que ces ressources sont compilées, il est impossible de les visualiser ou de les modifier. Vous pouvez toutefois forcer n'importe quel autre dictionnaire à accepter un terme classé par type par les ressources compilées. Pour plus d'informations, voir «Ajout des termes forcés», à la page 193.

Création de bibliothèques

Vous pouvez créer autant de bibliothèques que vous le souhaitez. Après avoir créé une nouvelle bibliothèque, vous pouvez commencer à y créer des dictionnaires de types et insérer des termes, des synonymes et des exclusions.

Pour créer une bibliothèque

1. A partir des menus, sélectionnez **Ressources > Nouvelle bibliothèque**. La boîte de dialogue Propriétés de bibliothèque s'ouvre.
2. Nommez la bibliothèque dans le champ Nom.
3. Si vous le souhaitez, insérez un commentaire dans le champ Annotation.
4. Cliquez sur **Publier** si vous souhaitez publier cette bibliothèque maintenant, sans y insérer quoi que ce soit. Pour plus d'informations, voir «Partage de bibliothèques», à la page 182. Vous pouvez également la publier ultérieurement.

5. Cliquez sur **OK** pour créer la bibliothèque. La boîte de dialogue se ferme et la bibliothèque apparaît dans l'arborescence. Si vous développez la bibliothèque dans l'arborescence, vous vous apercevez qu'un dictionnaire de types vide y a été inclus automatiquement. Vous pouvez ajouter des termes immédiatement. Pour plus d'informations, voir «Ajout de termes», à la page 190.

Ajout de bibliothèques publiques

Si vous souhaitez réutiliser une bibliothèque de données de session , vous pouvez l'ajouter aux ressources en cours tant qu'il s'agit d'une bibliothèque publique. Une *bibliothèque publique* est une bibliothèque qui a été publiée. Pour plus d'informations, voir «Publication de bibliothèques», à la page 183.

Lorsque vous ajoutez une bibliothèque publique, une copie *locale* est imbriquée dans les données de session . Il est possible d'apporter des modifications à cette bibliothèque, cependant, vous êtes tenu de publier à nouveau sa version publique si vous souhaitez partager les modifications effectuées.

Lorsque vous ajoutez une bibliothèque publique, une boîte de dialogue Résoudre les conflits peut apparaître s'il existe des conflits entre les termes et les types d'une bibliothèque et les autres bibliothèques locales. Il est nécessaire de résoudre ces conflits ou d'accepter les résolutions proposées afin de pouvoir réaliser cette opération. Pour plus d'informations, voir «Résolution des conflits», à la page 184.

Remarque : Si vous mettez toujours à jour vos bibliothèques lorsque vous lancez une session de plan de travail interactif ou que vous les publiez lorsque vous fermez une session, vous avez moins de risques d'avoir des bibliothèques non synchronisées. Pour plus d'informations, voir «Partage de bibliothèques», à la page 182.

Pour ajouter une bibliothèque

1. A partir des menus, sélectionnez **Ressources > Ajouter une bibliothèque**. La boîte de dialogue Ajouter une bibliothèque apparaît.
2. Sélectionnez la ou les bibliothèques dans la liste.
3. Cliquez sur **Ajouter**. En cas de conflits entre les bibliothèques qui viennent d'être ajoutées et les anciennes, vous êtes invité à vérifier les résolutions de conflit ou à les modifier avant de réaliser l'opération. Pour plus d'informations, voir «Résolution des conflits», à la page 184.

Recherche de termes et de types

Vous pouvez rechercher dans les nombreux sous-fenêtres de l'éditeur à l'aide de la fonction Rechercher. Dans l'éditeur, sélectionnez **Edition > Rechercher** dans les menus. La barre d'outils Rechercher apparaît. Utilisez cette barre d'outils pour rechercher une occurrence à la fois. En cliquant à nouveau sur **Rechercher**, vous pouvez rechercher les occurrences suivantes du terme recherché.

Lors de la recherche, l'éditeur effectue la recherche uniquement dans les bibliothèques répertoriées dans la liste déroulante de la barre d'outils Rechercher. Si l'option **Toutes les bibliothèques** est sélectionnée, le programme effectue la recherche dans tous les supports dans l'éditeur.

Lorsque vous démarrez une recherche, elle débute dans la zone active. La recherche continue dans chaque section, formant une boucle jusqu'à revenir à la cellule active. Il est possible d'inverser l'ordre de la recherche en utilisant les flèches de direction. Vous pouvez également choisir si votre recherche est sensible ou non à la casse.

Pour rechercher des chaînes dans la vue

1. Dans les menus, sélectionnez **Edition** > **Rechercher**. La barre d'outils Rechercher apparaît.
2. Insérez la chaîne de caractères que vous souhaitez rechercher.
3. Cliquez sur le bouton **Rechercher** pour commencer la recherche. La prochaine occurrence du terme ou du type est ensuite mise en surbrillance.

4. Cliquez à nouveau sur le bouton pour passer d'une occurrence à une autre.

Utilisation d'un astérisque dans les termes

L'utilisation d'un astérisque (*) dans les termes est particulièrement utile dans le cas d'une langue agglutinante qui crée de nouveaux mots en en regroupant d'autres sans espaces. Par exemple, le mot allemand *Übernachtungspreis* est constitué des mots *Übernachtung* + *s* + *Preis*.

Par exemple, si vous recherchez `preis*` dans les termes de type `Budget`, les concepts extraits, tels que `preiserhöhung`, seront affichés. De la même manière, `*preis` correspondra à `Übernachtung` et `*preis*`, à `Übernachtungspreiserhöhung`.

Affichage des bibliothèques

Vous pouvez afficher le contenu d'une bibliothèque donnée ou celui de toutes les bibliothèques. Cela peut s'avérer utile lorsque vous travaillez avec plusieurs bibliothèques ou lorsque vous souhaitez passer en revue le contenu d'une bibliothèque spécifique avant sa publication. Le fait de modifier la vue n'a un impact que sur ce que vous voyez dans l'onglet Ressources de bibliothèque mais cette modification n'empêche pas l'utilisation des bibliothèques au cours de l'extraction. Pour plus d'informations, voir «Désactivation des bibliothèques locales», à la page 181.

La vue par défaut est **Toutes les bibliothèques**. Cette option affiche toutes les bibliothèques de l'arborescence et leur contenu dans d'autres sous-fenêtres. Vous pouvez modifier cette sélection en utilisant la liste déroulante de la barre d'outils ou via une sélection de menus (**Vue > Bibliothèques**). Si une seule bibliothèque est affichée, tous les éléments des autres bibliothèques disparaissent de la vue mais continuent d'être lus pendant l'extraction.

Pour afficher la vue Bibliothèque

1. Dans les menus de l'onglet Ressources de bibliothèque, choisissez **Vue > Bibliothèques**. Un menu comprenant toutes les bibliothèques locales apparaît.
2. Sélectionnez la bibliothèque que vous souhaitez afficher ou sélectionnez l'option **Toutes les bibliothèques** pour afficher le contenu de toutes les bibliothèques. Le contenu de la vue est filtré en fonction de votre sélection.

Gestion des bibliothèques locales

Les bibliothèques locales sont internes à votre session de plan de travail interactif ou intégrées à un modèle, contrairement aux bibliothèques publiques. Pour plus d'informations, voir «Gestion des bibliothèques publiques», à la page 181. Renommer, désactiver ou supprimer une bibliothèque locale sont des tâches de gestion courante que vous pourrez être amené à effectuer.

Attribution d'un nouveau nom à une bibliothèque locale

Vous pouvez renommer les bibliothèques locales. Lorsque vous renommez une bibliothèque locale, vous la dissociez de la version publique éventuelle. Cela signifie que les modifications suivantes ne peuvent plus être partagées avec la version publique. Vous pouvez republier cette bibliothèque locale sous son nouveau nom. Cela signifie également que vous ne pouvez pas incorporer les modifications apportées à cette version locale dans la version publique originale.

Remarque : il est impossible de renommer une bibliothèque publique.

1. Dans les menus, sélectionnez **Edition> Propriétés de bibliothèque**. La boîte de dialogue Propriétés de bibliothèque apparaît.

Pour renommer une bibliothèque locale

1. Dans l'arborescence, sélectionnez la bibliothèque à renommer.

2. Attribuez un nouveau nom à la bibliothèque dans le champ Nom.
3. Cliquez sur **OK** pour accepter le nouveau nom de la bibliothèque. La boîte de dialogue se ferme et son nom apparaît dans l'arborescence.

Désactivation des bibliothèques locales

Si vous souhaitez exclure temporairement une bibliothèque du processus d'extraction, désélectionnez la case à cocher à gauche du nom de la bibliothèque dans l'arborescence. Vous indiquez ainsi que vous souhaitez conserver la bibliothèque, mais en ignorer le contenu pendant la recherche de conflits et le processus d'extraction.

Pour désactiver une bibliothèque

1. Dans l'arborescence de bibliothèques, sélectionnez la bibliothèque à désactiver.
2. Cliquez sur la barre d'espacement. La coche disparaît de la case figurant à gauche du nom.

Suppression des bibliothèques locales

Vous pouvez supprimer une bibliothèque sans supprimer la version publique de la bibliothèque et inversement. La suppression d'une bibliothèque locale entraîne la destruction de son contenu uniquement pour cette session , mais pas pour les autres sessions , ni pour la version publique. Pour plus d'informations, voir «Gestion des bibliothèques publiques».

Pour supprimer une bibliothèque locale

1. Dans l'arborescence, sélectionnez la bibliothèque à supprimer.
2. Dans les menus, choisissez **Edition > Supprimer** pour supprimer la bibliothèque. La bibliothèque est supprimée.
3. Si vous n'avez jamais publié cette bibliothèque auparavant, un message vous demande si vous souhaitez supprimer ou conserver la bibliothèque. Cliquez sur **Supprimer** pour poursuivre ou sur **Conserver** pour conserver cette bibliothèque.

Remarque : il doit toujours rester au moins une bibliothèque.

Gestion des bibliothèques publiques

Afin de réutiliser les bibliothèques locales, vous pouvez les publier et ensuite les utiliser à partir de la boîte de dialogue Gérer les bibliothèques (**Ressources > Gérer les bibliothèques**). Pour plus d'informations, voir «Partage de bibliothèques», à la page 182. L'importation, l'exportation et la suppression figurent parmi les tâches élémentaires de gestion des bibliothèques publiques. Il est impossible de renommer une bibliothèque publique.

Importation de bibliothèques publiques

1. Dans la boîte de dialogue Gérer les bibliothèques, cliquez sur **Importer**. La boîte de dialogue Importer une bibliothèque apparaît.
2. Sélectionnez le fichier de bibliothèque (**.lib*) à importer et, si vous souhaitez également ajouter cette bibliothèque localement, sélectionnez l'option **Ajouter une bibliothèque au projet actuel**.
3. Cliquez sur **Importer**. La boîte de dialogue se ferme. Si une bibliothèque publique de même nom existe déjà, vous êtes invité à renommer la bibliothèque en cours d'importation ou à remplacer la bibliothèque publique en cours.

Exportation de bibliothèques publiques

Vous pouvez exporter les bibliothèques publiques au format *.lib* afin de les partager.

1. Dans la boîte de dialogue Gérer les bibliothèques, sélectionnez dans la liste la bibliothèque à exporter.
2. Cliquez sur **Exporter**. La boîte de dialogue Sélectionner un répertoire apparaît.

3. Sélectionnez le répertoire vers lequel vous souhaitez exporter et cliquez sur **Exporter**. La boîte de dialogue disparaît et le fichier de bibliothèque (*.*lib*) est exporté.

Suppression de bibliothèques publiques

Vous pouvez supprimer une bibliothèque locale sans supprimer la version publique de la bibliothèque et inversement. Toutefois, une bibliothèque supprimée à partir de cette boîte de dialogue ne peut plus être ajoutée à une ressource de session tant qu'une version locale n'a pas été à nouveau publiée.

Si vous supprimez une bibliothèque qui a été installée avec le produit, la version originale installée est restaurée.

1. Dans la boîte de dialogue Gérer les bibliothèques, sélectionnez la bibliothèque à supprimer. Vous pouvez trier la liste en cliquant sur l'en-tête approprié.
2. Cliquez sur **Supprimer**. IBM SPSS Modeler Text Analytics vérifie si la version locale de la bibliothèque est identique à la version publique. Si c'est le cas, la bibliothèque est supprimée sans message d'alerte. Si les versions de la bibliothèque diffèrent, un message d'alerte apparaît pour vous demander si vous souhaitez conserver ou supprimer la version publique.

Partage de bibliothèques

Les bibliothèques vous permettent de travailler avec des ressources facilement partageables entre plusieurs sessions de plan de travail interactif . Les bibliothèques existent en deux états ou deux versions. Les bibliothèques modifiables dans l'éditeur et faisant partie d'une session de plan de travail interactif sont appelées des **bibliothèques locales**. Lorsque vous travaillez dansune session de plan de travail interactif , vous pouvez faire de nombreux changements dans une bibliothèque, par exemple *Vegetables* (légumes). Si vos changements peuvent s'avérer utiles avec d'autres données, vous pouvez rendre ces ressources disponibles en créant une **version publique** de la bibliothèque *Vegetables* (Légumes). Une bibliothèque publique, comme l'indique sa dénomination, est accessible à partir de toutes les ressources de session de plan de travail interactif.

Vous pouvez visualiser les bibliothèques publiques dans la boîte de dialogue Gérer les bibliothèques. Une fois que cette version de bibliothèque publique existe, vous pouvez l'ajouter aux ressources dans d'autres contextes de manière à ce que ces ressources linguistiques personnalisées puissent être partagées.

Les bibliothèques fournies sont initialement des bibliothèques publiques. Il est possible de modifier les ressources de ces bibliothèques avant de créer une version publique. Les nouvelles versions sont alors accessibles à partir d'une autre session de plan de travail interactif.

Lorsque vous utilisez vos bibliothèques et y apportez des modifications, les différentes versions ne sont plus synchronisées. Dans certains cas, une version locale peut être plus récente que la version publique et, dans d'autres cas, la version publique peut être plus récente que la version locale. Il est également possible que la version publique ou locale ne contiennent pas les mêmes informations, si la version publique est mise à jour à partir d'une autre session de plan de travail interactif. Si les versions de vos bibliothèques ne sont plus synchronisées, vous pouvez rétablir la synchronisation. La synchronisation des versions de bibliothèque consiste à publier une nouvelle fois et/ou mettre à jour les bibliothèques locales.

Lorsque vous lancez ou fermez une session de plan de travail interactif , vous êtes invité à synchroniser les bibliothèques devant être mises à jour ou republiées. De plus, vous pouvez facilement identifier l'état de synchronisation de votre bibliothèque locale à l'aide de l'icône qui apparaît à côté du nom de la bibliothèque dans l'arborescence ou en affichant la boîte de dialogue Propriétés de bibliothèque. Vous pouvez également effectuer cette opération à tout moment en sélectionnant les options de menu. Le tableau ci-dessous décrit les cinq états possibles et leurs icônes associées.

Tableau 37. Etats de synchronisation des bibliothèques locales.

Icône	Description de l'état des bibliothèques locales
	Non publiée - La bibliothèque locale n'a jamais été publiée.
	Synchronisée — Les versions locale et publique de la bibliothèque sont identiques. Cet état s'applique également à la *bibliothèque locale*, qui ne peut pas être publiée car elle est conçue pour contenir uniquement des ressources spécifiques à une session .
	Obsolète — La version publique de la bibliothèque est plus récente que la version locale. Vous pouvez mettre à jour votre version locale avec les modifications.
	Plus récente — La version locale de la bibliothèque est plus récente que la version publique. Vous pouvez republier votre version locale pour que la version publique soit actualisée.
	Non synchronisées — Les versions locale et publique d'une bibliothèque contiennent toutes deux des modifications que l'autre n'a pas. Vous devez choisir soit de mettre à jour votre bibliothèque locale, soit de la publier. Si vous la mettez à jour, vous perdez les modifications que vous avez effectuées depuis votre dernière mise à jour ou publication. Si vous optez pour la publication, vous supprimez les changements effectués dans la version publique.

Remarque : si vous mettez toujours à jour vos bibliothèques lorsque vous lancez une session de plan de travail interactif ou que vous les publiez lorsque vous fermez une session, vous avez moins de risques d'avoir des bibliothèques non synchronisées.

Vous pouvez republier une bibliothèque chaque fois que vous jugez que les modifications apportées seront utiles aux autres flux pouvant également contenir cette bibliothèque. Dans ce cas, vous devrez mettre à jour la version locale de ces flux. De cette manière, vous pouvez créer des flux pour chaque contexte ou domaine s'appliquant à vos données en créant de nouvelles bibliothèques et/ou en ajoutant des bibliothèques publiques à vos ressources.

Lorsqu'une version publique de bibliothèque est partagée, les risques de différence entre les versions locale et publique sont plus importants. Lorsque vous lancez ou fermez et publiez à partir d'une session de plan de travail interactif, ou que vous ouvrez ou fermez un modèle à partir de l'Editeur de modèle , un message s'affiche pour vous inviter à publier et/ou mettre à jour des bibliothèques dont les versions ne sont pas synchronisées avec celles de la boîte de dialogue Gérer les bibliothèques. Si la version publique de la bibliothèque est plus récente que la version locale, une boîte de dialogue vous demande si vous souhaitez procéder à une mise à jour. Vous pouvez choisir de conserver la version locale en l'état au lieu de la mettre à jour pour qu'elle contienne les changements de la version publique. Vous pouvez également décider de la mettre à jour pour qu'elle reflète les changements insérés dans la version publique.

Publication de bibliothèques

Si vous n'avez jamais publié de bibliothèques, sachez que la publication consiste à créer une copie publique de votre bibliothèque locale dans la base de données. Si vous republiez une bibliothèque, le contenu de la bibliothèque locale remplace le contenu de la version publique existante. Après la republication, vous pouvez mettre à jour cette bibliothèque dans les autres sessions de flux de sorte que leurs versions locales soient synchronisées avec la version publique. Même si vous pouvez publier une bibliothèque, la session contient toujours une version locale.

Important ! Si vous apportez des modifications à votre bibliothèque locale et si, dans le même temps, la version publique de la bibliothèque est également modifiée, votre bibliothèque est considérée comme désynchronisée. Nous vous recommandons de commencer par mettre à jour la version locale avec les modifications de la version publique, d'effectuer tous les changements nécessaires et ensuite, de publier à

nouveau la version locale pour que les deux versions soient identiques. Si vous effectuez des changements et que vous commencez par une publication, vous remplacerez les changements apportés à la version publique.

Pour publier des bibliothèques locales dans la base de données

1. A partir des menus, sélectionnez **Ressources > Publier les bibliothèques**. La boîte de dialogue Publier les bibliothèques apparaît. Toutes les bibliothèques devant être publiées sont sélectionnées par défaut.
2. Cochez la case située à gauche de chaque bibliothèque à publier ou republier.
3. Cliquez sur **Publier** pour publier les bibliothèques dans la base de données Gérer les bibliothèques.

Mise à jour des bibliothèques

Lorsque vous lancez ou fermez une session de plan de travail interactif, vous pouvez mettre à jour ou publier des bibliothèques qui ne sont plus synchronisées avec les versions publiques. Si la version publique de la bibliothèque est plus récente que la version locale, une boîte de dialogue vous demande si vous souhaitez mettre à jour la bibliothèque. Vous pouvez choisir soit de conserver la version locale au lieu de la mettre à jour avec le contenu de la version publique, soit de remplacer la version locale par la publique. Si la version publique d'une bibliothèque est plus récente que la version locale, vous pouvez mettre à jour la version locale pour synchroniser son contenu avec celui de la version publique. La mise à jour consiste à incorporer dans votre version locale les changements trouvés dans la version publique.

Remarque : si vous mettez toujours à jour vos bibliothèques lorsque vous lancez une session de plan de travail interactif ou que vous les publiez lorsque vous fermez une session , vous avez moins de risques d'avoir des bibliothèques non synchronisées. Pour plus d'informations, voir «Partage de bibliothèques», à la page 182.

Pour mettre à jour les bibliothèques locales

1. A partir des menus, sélectionnez **Ressources > Mettre à jour les bibliothèques**. La boîte de dialogue Mettre à jour les bibliothèques apparaît. Toutes les bibliothèques devant être mises à jour sont sélectionnées par défaut.
2. Cochez la case située à gauche de chaque bibliothèque à publier ou republier.
3. Cliquez sur **Mettre à jour** pour mettre à jour les bibliothèques locales.

Résolution des conflits

Conflits entre bibliothèque locale et bibliothèque publique

Lorsque vous démarrez une session de flux, IBM SPSS Modeler Text Analytics compare les bibliothèques locales et celles listées dans la boîte de dialogue Gérer les bibliothèques. Si une ou plusieurs bibliothèques locales de votre session ne sont pas synchronisées avec les versions publiées, la boîte de dialogue Avertissement de synchronisation de bibliothèques s'ouvre. Vous pouvez choisir parmi les options suivantes afin de sélectionner les versions des bibliothèques que vous souhaitez utiliser :

- **Toutes les bibliothèques locales vers le fichier du projet.** Cette option conserve toutes vos bibliothèques en l'état. Vous pouvez toujours les republier ou les mettre à jour ultérieurement.
- **Toutes les bibliothèques publiées sur cet ordinateur.** Cette option remplace les bibliothèques locales affichées par les versions de la base de données.
- **Toutes les bibliothèques les plus récentes.** Cette option remplace toutes les anciennes bibliothèques locales par les versions publiques plus récentes contenues dans la base de données.
- **Autre.** Cette option permet de sélectionner manuellement les versions de votre choix en les sélectionnant dans le tableau.

Conflits de termes forcés

Lorsque vous ajoutez une bibliothèque publique ou mettez à jour une bibliothèque locale, vous pouvez détecter des conflits et des doublons entre les termes et les types de cette bibliothèque et ceux des autres bibliothèques stockées dans vos ressources. Si cela se produit, vous êtes invité à vérifier, dans la boîte de dialogue Editer les termes forcés, les résolutions de conflit proposées ou à les modifier avant de réaliser l'opération. Pour plus d'informations, voir «Ajout des termes forcés», à la page 193.

La boîte de dialogue Editer les termes forcés contient toutes les paires de termes et types contradictoires. L'alternance des couleurs de l'arrière-plan permet de distinguer chaque paire contradictoire. Ces couleurs peuvent être modifiées dans la boîte de dialogue Options. Pour plus d'informations, voir la rubrique «Options : onglet Afficher», à la page 76. La boîte de dialogue Editer les termes forcés contient deux onglets :

- **Doublons.** Cet onglet contient les doublons rencontrés dans les bibliothèques. Si une icône représentant une punaise apparaît après un terme, cela signifie que l'occurrence du terme a été forcée. Si une icône X apparaît en noir, cela signifie que cette occurrence du terme sera ignorée pendant l'extraction, car elle a été imposée ailleurs.
- **Défini par l'utilisateur.** Cet onglet contient la liste de tous les termes qui ont été forcés manuellement dans la sous-fenêtre des termes du dictionnaire de types et non pas via les conflits.

Remarque : la boîte de dialogue Editer les termes forcés s'ouvre après l'ajout ou la mise à jour d'une bibliothèque. Si vous quittez la boîte de dialogue, vous n'annulez pas la mise à jour ou l'ajout de la bibliothèque.

Pour résoudre des conflits

1. Dans la boîte de dialogue Editer les termes forcés, sélectionnez le bouton radio de la colonne Utiliser du terme que vous souhaitez forcer.
2. Lorsque vous avez terminé, cliquez sur **OK** pour appliquer les termes forcés et fermer la boîte de dialogue. Si vous cliquez sur **Annuler**, vous annulerez les changements effectués dans cette boîte de dialogue.

Chapitre 16. A propos des dictionnaires de bibliothèque

Les ressources servant à extraire des données textuelles sont stockées sous la forme de modèles et de bibliothèques. Une bibliothèque peut être constituée de trois dictionnaires.

- Le **dictionnaire de types** comprend un ensemble de termes regroupés sous un même libellé ou sous un même nom de type. Lorsque le moteur d'extraction lit vos données textuelles, il compare les mots rencontrés dans le texte aux termes définis dans vos dictionnaires de types. Au cours de l'extraction, les formes infléchies des termes et synonymes d'un type sont regroupées sous un terme cible nommé concept. Les concepts extraits sont affectés au dictionnaire de types dans lequel ils apparaissent en tant que termes. La gestion des dictionnaires de types s'effectue dans les sous-fenêtres supérieures gauche et centrales de l'éditeur de l'arborescence de bibliothèque et de termes. Pour plus d'informations, voir «Dictionnaires de types».
- Le **dictionnaire de substitutions** contient un ensemble de mots définis comme synonymes ou comme éléments optionnels qui sont utilisés pour regrouper des termes similaires sous un seul terme cible, appelé concept dans les résultats d'extraction finaux. Vous pouvez gérer vos dictionnaires de substitutions dans la sous-fenêtre inférieure gauche de l'éditeur à l'aide des onglets Synonymes et Optionnels. Pour plus d'informations, voir «Dictionnaires de substitutions/synonymes», à la page 195.
- Le **dictionnaire d'exclusions** comprend un ensemble de termes et de types qui seront supprimés des résultats finaux de l'extraction. Vous pouvez gérer vos dictionnaires d'exclusion dans la sous-fenêtre la plus à droite de l'éditeur. Pour plus d'informations, voir «Dictionnaires d'exclusions», à la page 198.

Pour plus d'informations, voir Chapitre 15, «Utilisation des bibliothèques», à la page 177.

Dictionnaires de types

Un *dictionnaire de types* est constituée d'un nom de type, d'un libellé et d'une liste de termes. La gestion des dictionnaires de types s'effectue dans les sous-fenêtres supérieures gauche et centrales de l'onglet Ressources de bibliothèque dans l'éditeur. Vous pouvez accéder à cette vue avec **Vue > Editeur de ressources** dans les menus lors d'une session de plan de travail interactif. Autrement, vous pouvez modifier les dictionnaires d'un modèle spécifique dans l'Editeur de modèle.

Lorsque le moteur du programme d'extraction lit vos données textuelles, il compare les mots rencontrés dans le texte aux termes définis dans vos dictionnaires de types. Les termes sont des mots ou des expressions dans les dictionnaires de types dans vos ressources linguistiques.

Lorsqu'un mot correspond à un terme, il est attribué au nom de type de ce terme. Lorsque les ressources sont lues au cours de l'extraction, les termes trouvés dans le texte traversent ensuite plusieurs étapes de traitement avant de devenir des concepts dans la sous-fenêtre Résultats d'extraction. Si plusieurs termes appartenant au même dictionnaire de types sont déterminés comme étant des synonymes par le moteur du programme d'extraction, ils sont regroupés sous le terme le plus fréquent et nommés un *concept* dans la sous-fenêtre Résultats d'extraction. Par exemple, si les termes `question` et `requête` peuvent apparaître dans le nom de concept `question` à la fin.

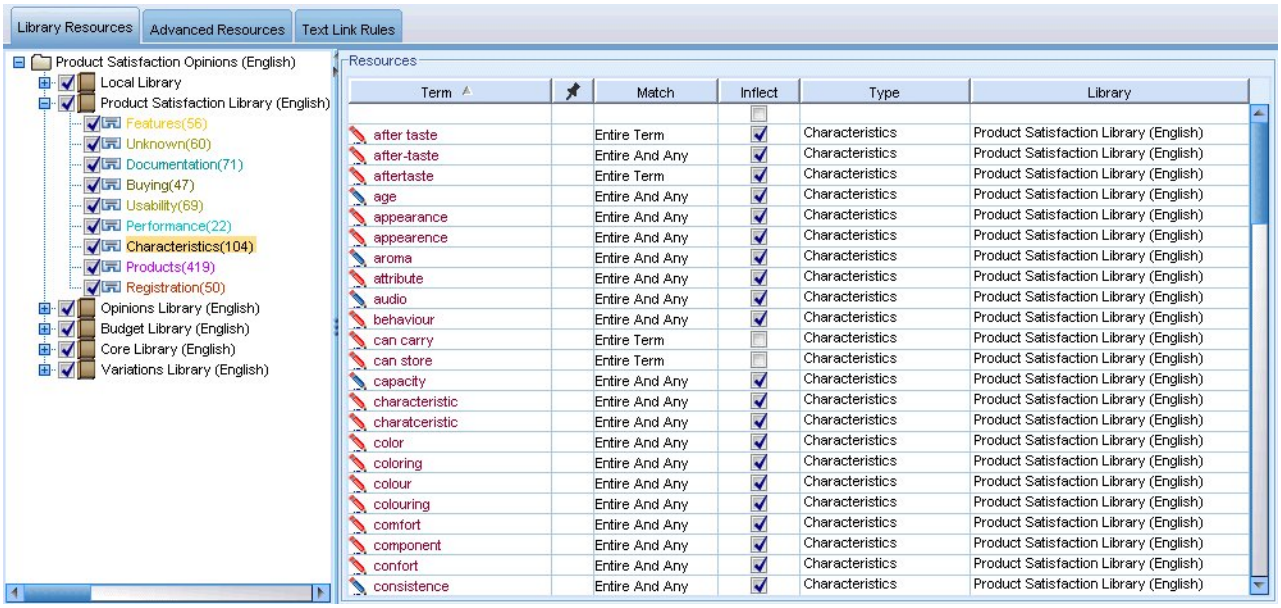


Figure 39. Arborescence de bibliothèques et sous-fenêtre de termes

La liste des dictionnaires de types est affichée dans la sous-fenêtre qui contient l'arborescence de bibliothèques, à gauche. Le contenu de chaque dictionnaire de types apparaît dans la sous-fenêtre centrale. Les dictionnaires de types offrent bien plus qu'une simple liste de termes. La manière dont les mots de vos données textuelles sont mis en correspondance avec les concepts affectés aux dictionnaires de types est déterminée par l'option de mise en correspondance. Une **option de mise en correspondance** définit le positionnement d'un terme par rapport à une expression ou un mot candidat dans les données textuelles. Pour plus d'informations, voir «Ajout de termes», à la page 190.

Par ailleurs, vous pouvez étendre les termes de votre dictionnaire de types en indiquant si vous souhaitez générer et ajouter automatiquement des formes infléchies aux termes du dictionnaire. En générant les formes infléchies, vous ajoutez automatiquement les formes plurielles de termes au singulier, les formes au singulier de termes au pluriel et les adjectifs dans le dictionnaire de types. Pour plus d'informations, voir «Ajout de termes», à la page 190.

Remarque : Pour la plupart des langues, les concepts qui sont extraits du texte mais ne figurent dans aucun dictionnaire de types sont automatiquement typés `<Unknown>`,

Utilisation d'un astérisque dans les termes

L'utilisation d'un astérisque (*) dans les termes est particulièrement utile dans le cas d'une langue agglutinante qui crée de nouveaux mots en en regroupant d'autres sans espaces. Par exemple, le mot allemand *Übernachtungspreis* est constitué des mots *Übernachtung* + *s* + *Preis*.

Par exemple, si vous recherchez `preis*` dans les termes de type `Budget`, les concepts extraits, tels que `preiserhöhung`, seront affichés. De la même manière, `*preis` correspondra à `Übernachtung` et `*preis*`, à `Übernachtungspreiserhöhung`.

Types intégrés

IBM SPSS Modeler Text Analytics est livré avec un ensemble de ressources linguistiques sous forme de bibliothèques et de ressources compilées. Les bibliothèques fournies contiennent un ensemble de dictionnaires de types intégrés parmi lesquelles `<Location>`, `<Organization>`, `<Person>` et `<Product>`.

Ces dictionnaires de types sont utilisés par le moteur du programme d'extraction pour affecter des types aux concepts qu'il extrait (par exemple, le type `<Location>` est affecté au concept `paris`). Bien qu'un grand

nombre de termes aient été définis dans les dictionnaires de types intégrés, ceux-ci ne couvrent pas toutes les possibilités. Par conséquent, vous pouvez les compléter ou en créer qui vous soient propres. Pour obtenir la description du contenu d'un dictionnaire de types fourni en particulier, reportez-vous à l'annotation dans la boîte de dialogue Propriétés de type. Sélectionnez le type dans l'arborescence et choisissez **Edition > Propriétés** dans le menu contextuel.

Remarque :

En complément des bibliothèques fournies, les ressources compilées (également utilisées par le moteur d'extraction) contiennent un grand nombre de définitions complémentaires aux dictionnaires de types intégrés, mais leur contenu n'est pas visible dans le produit. Vous pouvez toutefois forcer n'importe quel autre dictionnaire à accepter un terme classé par type par les déclarations compilées. Pour plus d'informations, voir «Ajout des termes forcés», à la page 193.

Création de types

Vous pouvez créer des dictionnaires de types pour faciliter le regroupement de termes similaires. Quand des termes qui figurent dans ce dictionnaire sont découverts au cours du processus d'extraction, ils sont affectés à ce nom de type et extraits sous un nom de concept. Lorsque vous créez une bibliothèque, une bibliothèque de types vide est toujours incluse, afin de vous permettre de commencer immédiatement à entrer des termes.

Si vous analysez un texte concernant des aliments et souhaitez regrouper les termes relatifs aux légumes, vous pouvez créer votre propre dictionnaire de types `<Légumes>`. Vous pouvez ensuite y ajouter des termes tels que `carotte`, `brocoli` et `épinard` si vous estimez qu'il s'agit de termes importants qui vont apparaître dans le texte. Ensuite, au cours de l'extraction, si l'un de ces termes est identifié, il est extrait en tant que concept et affecté au type `<Vegetables>` (Légumes).

Il n'est pas nécessaire de définir toutes les formes d'un mot ou d'une expression, car vous pouvez choisir de générer toutes les formes infléchies des termes. Lorsque vous choisissez cette option, le moteur d'extraction reconnaît automatiquement les formes au singulier ou au pluriel des termes parmi les autres formes et les associe à ce type. Cette option s'avère particulièrement utile quand votre type contient principalement des noms, dans la mesure où il est peu probable que vous cherchiez à obtenir les formes infléchies de verbes ou d'adjectifs.

La boîte de dialogue Propriétés de type contient les champs suivants.

Nom. Nom que vous attribuez au dictionnaire de types que vous créez. Nous vous recommandons de ne pas utiliser d'espaces dans les noms de types, en particulier si deux noms de types ou plus commencent par le même mot.

Remarque : Il existe certaines contraintes par rapport aux noms de types et à l'utilisation de symboles. Par exemple, vous ne pouvez pas utiliser les symboles tels que "@" ou "!" dans le nom.

Mise en correspondance par défaut. L'attribut Mise en correspondance par défaut indique au moteur d'extraction la manière dont il doit mettre en correspondance ce terme et les données textuelles. Dès que vous ajoutez un terme à ce dictionnaire de types, cet attribut de mise en correspondance lui est associé automatiquement. Vous avez toujours la possibilité de modifier manuellement le choix de correspondance dans la liste des termes. Les options incluent : **Terme entier**, **Démarrer**, **Terminer**, **Tout**, **Début ou fin**, **Entier et début**, **Entier et fin**, **Entier et (début ou fin)** et **Entier (pas de composés)**. Pour plus d'informations, voir «Ajout de termes», à la page 190.

Ajouter à. Ce champ indique la bibliothèque dans laquelle vous allez créer votre nouveau dictionnaire de types.

Générer les formes infléchies par défaut. Cette option indique au moteur d'extraction d'utiliser la fonction de morphologie grammaticale pour collecter et regrouper les formes similaires des termes que vous ajoutez à ce dictionnaire (le singulier et le pluriel d'un terme, par exemple). Cette option est particulièrement utile lorsque votre type contient surtout des noms. Quand vous sélectionnez cette option, elle est appliquée par défaut à tous les nouveaux termes ajoutés au type, mais vous pouvez la modifier manuellement dans la liste.

Couleur. Ce champ vous permet de distinguer les résultats de ce type dans l'interface. Si vous sélectionnez **Par défaut**, la couleur par défaut du type est également utilisée pour ce dictionnaire de types. La couleur par défaut se configure dans la boîte de dialogue des options. Pour plus d'informations, voir la rubrique «Options : onglet Afficher», à la page 76. Si vous sélectionnez **Personnalisé**, choisissez une couleur à partir de la liste déroulante.

Annotation. Ce champ est facultatif et sert à entrer des commentaires ou des descriptions.

Pour créer un dictionnaire de types

1. Sélectionnez la bibliothèque dans laquelle vous souhaitez créer un dictionnaire de types.
2. A partir des menus, sélectionnez **Outils > Nouveau type**. La boîte de dialogue Propriétés de type s'ouvre.
3. Entrez le nom de votre dictionnaire de types dans la zone de texte **Nom** et choisissez les options souhaitées.
4. Cliquez sur **OK** pour créer le dictionnaire de types. Le nouveau type apparaît dans l'arborescence de bibliothèques et s'affiche dans la sous-fenêtre centrale. Vous pouvez commencer immédiatement à ajouter des termes. Pour plus d'informations, voir «Ajout de termes».

Remarque : Ces instructions montrent comment effectuer des modifications dans la vue de l'Editeur de ressources ou dans l'Editeur de modèle. Vous pouvez également effectuer ces modifications directement dans la sous-fenêtre Résultats d'extraction , la sous-fenêtre Données, la sous-fenêtre Catégories ou la boîte de dialogue Définitions de cluster dans les autres vues. Pour plus d'informations, voir «Affinage des résultats d'extraction», à la page 89.

Ajout de termes

L'arborescence de bibliothèques contient les bibliothèques. Elle peut être développée pour présenter les dictionnaires de types qu'elles contiennent. Dans la sous-fenêtre centrale, une liste affiche les termes de la bibliothèque ou du dictionnaire de types sélectionné, en fonction de l'élément sélectionné dans l'arborescence.

Dans Editeur de ressources, vous pouvez ajouter des termes directement à un dictionnaire de types : soit directement dans la sous-fenêtre de termes, soit par l'intermédiaire de la boîte de dialogue Ajouter des nouveaux termes. Les termes que vous ajoutez peuvent être des mots simples ou composés. La ligne vide, située en haut de la liste, vous permet d'ajouter de nouveaux termes à tout moment.

Remarque : Ces instructions montrent comment effectuer des modifications dans la vue de l'Editeur de ressources ou dans l'Editeur de modèle. Vous pouvez également effectuer ces modifications directement dans la sous-fenêtre Résultats d'extraction , la sous-fenêtre Données, la sous-fenêtre Catégories ou la boîte de dialogue Définitions de cluster dans les autres vues. Pour plus d'informations, voir «Affinage des résultats d'extraction», à la page 89.

Colonne Terme

Dans cette colonne, entrez des mots simples ou composés dans la cellule. La couleur d'affichage du terme dépend de la couleur du type dans lequel le terme est enregistré ou a été imposé. Vous pouvez changer les couleurs des types dans la boîte de dialogue Propriétés de type. Pour plus d'informations, voir «Création de types», à la page 189.

Colonne Forcer

Dans cette colonne, sélectionnez cette cellule pour afficher une icône de punaise qui indique au moteur d'extraction d'ignorer les autres occurrences de ce même terme dans les autres bibliothèques. Pour plus d'informations, voir «Ajout des termes forcés», à la page 193.

Colonne Correspondance

Dans cette colonne, sélectionnez une option de mise en correspondance pour indiquer au moteur d'extraction comment il doit faire correspondre ce terme aux données textuelles. Consultez le tableau pour voir des exemples. Pour changer la valeur par défaut, éditez les propriétés du type. Pour plus d'informations, voir «Création de types», à la page 189. Dans les menus, sélectionnez **Edition > Modifier la correspondance**. Ci-dessous figurent les options de mise en correspondance de base, car des combinaisons de celles-ci sont également possibles :

- **Démarrer**. Si le terme dans le dictionnaire correspond au premier mot d'un concept extrait du texte, ce type est attribué. Par exemple, si vous entrez `tarte`, le terme `tarte aux pommes` est renvoyé.
- **Fin**. Si le terme dans le dictionnaire correspond au dernier mot d'un concept extrait du texte, ce type est attribué. Par exemple, si vous entrez `tarte`, `moule à tarte` est renvoyé.
- **Tout**. Si le terme dans le dictionnaire correspond à tout mot d'un concept extrait du texte, ce type est attribué. Par exemple, si vous entrez `tarte`, l'option **Tout** classe `tarte aux pommes`, `moule à tarte` et `moule pour tarte aux pommes` sous un même type.
- **Terme entier**. Ce type est attribué si l'intégralité du concept extrait du texte correspond exactement avec le terme du dictionnaire. L'ajout d'un terme comme **Terme entier**, **Entier et début**, **Entier et fin**, **Entier et n'importe** ou **Entier (pas de composés)** force l'extraction d'un terme.

  De plus, comme le type `<Person>` extrait uniquement les noms en deux parties, comme *edith piaf* ou *mohandas gandhi*, vous voudrez peut-être ajouter explicitement les prénoms à ce dictionnaire de types si vous essayez d'extraire un prénom lorsqu'aucun nom de famille n'est mentionné. Par exemple, si vous voulez extraire toutes les instances du nom *edith*, vous devez ajouter `edith` au type `<Person>` à l'aide des options **Terme entier** ou **Entier et début**.
- **Entier (pas de composés)**. Si le concept complet extrait du texte correspond au terme exact dans le dictionnaire, ce type est attribué et l'extraction s'arrête afin d'empêcher l'extraction d'associer le terme aux composés plus longs. Par exemple, si vous saisissez `pomme`, l'option **Entier (pas de composé)** produira `pomme` et n'extraira pas le composé `jus de pomme`, à moins que ce ne soit forcé ailleurs.

Dans le tableau suivant, on considère que le terme `pomme` est dans un dictionnaire type. En fonction de l'option de mise en correspondance, ce tableau indique les concepts qui seraient extraits et entrés s'ils étaient trouvés dans le texte.

Tableau 38. Exemples de correspondance

Options de mise en correspondance pour le terme :  **pomme**	**Concepts extraits**			
	pomme	**tarte** *pommes*	*pommes* **mûres**	*tarte* **tarte** *pommes*
Terme entier	✔			
Début		✔		

Tableau 38. Exemples de correspondance (suite)

Options de mise en correspondance pour le terme :  **pomme**	**Concepts extraits**			
	pomme	**tarte** *pommes*	*pommes* **mûres**	*tarte* **tarte** *pommes*
Fin			✔	
Début ou fin		✔	✔	
Entier et début	✔	✔		
Entier et fin	✔		✔	
Entier et (début ou fin)	✔	✔	✔	
Tout		✔	✔	✔
Entier et Tout	✔	✔	✔	✔
Entier (pas de composés)	✔	*jamais extrait*	*jamais extrait*	*jamais extrait*

Colonne Inflexion

Dans cette colonne, sélectionnez si le moteur du programme d'extraction doit générer les formes infléchies de ce terme au cours de l'extraction afin qu'elles soient regroupées. La valeur par défaut de cette colonne est définie dans les propriétés du type, mais vous pouvez modifier cette option au cas par cas directement dans la colonne. Dans les menus, sélectionnez **Edition > Modifier l'inflexion**.

Colonne Type

Dans cette colonne, sélectionnez un dictionnaire de types dans la liste déroulante. La liste de types est filtrée en fonction de votre sélection dans l'arborescence de la bibliothèque. Le premier type de la liste est toujours celui sélectionné par défaut dans l'arborescence de la bibliothèque. Dans les menus, sélectionnez **Edition > Modifier le type**.

Colonne Bibliothèque

Dans cette colonne, la bibliothèque dans laquelle votre terme est stocké s'affiche. Pour transférer un terme dans un autre type de l'arborescence, faites-le glisser et déposez-le sur l'autre bibliothèque.

Pour ajouter un terme unique à un dictionnaire de types

1. Dans l'arborescence de la bibliothèque, sélectionnez le dictionnaire de types auquelle vous voulez ajouter le terme.
2. Dans la liste de termes de la sous-fenêtre centrale, entrez votre terme dans la première cellule vide disponible et définissez les options souhaitées pour ce terme.

Pour ajouter plusieurs termes à un dictionnaire de types

1. Dans l'arborescence de la bibliothèque, sélectionnez le dictionnaire de types auquel vous voulez ajouter des termes.
2. A partir des menus, sélectionnez **Outils> Nouveaux termes**. La boîte de dialogue Ajouter des nouveaux termes s'ouvre.
3. Entrez les termes que vous souhaitez ajouter au dictionnaire de types sélectionné. Pour ce faire, saisissez les termes au clavier, ou copiez et collez un ensemble de termes. Si vous entrez plusieurs termes, séparez-les au moyen du séparateur défini dans la boîte de dialogue Options ou ajoutez un terme par nouvelle ligne. Pour plus d'informations, voir la rubrique «Définition des options», à la page 75.
4. Cliquez sur **OK** pour ajouter les termes au dictionnaire. L'option de mise en correspondance est définie automatiquement sur la valeur par défaut de cette bibliothèque de types. La boîte de dialogue se ferme et les nouveaux termes apparaissent dans le dictionnaire.

Ajout des termes forcés

Si vous souhaitez affecter un terme à un type particulier, vous pouvez l'ajouter au dictionnaire de types correspondant. Toutefois, si plusieurs termes ont le même nom, le moteur d'extraction doit savoir quel type utiliser. Par conséquent, vous êtes invité à sélectionner le type à utiliser. Cette opération est appelée *ajout forcé* d'un terme dans un type. Cette option est particulièrement utile lorsque vous remplacez le type attribué d'un dictionnaire compilé (interne, non modifiable). En général, nous recommandons d'éviter les termes doubles.

L'ajout des termes forcés ne *supprime* pas les autres occurrences du terme, mais elles sont ignorées par le moteur d'extraction. Vous pouvez ensuite désigner l'occurrence qui doit être utilisée en activant ou en désactivant l'ajout des termes forcés. Vous pouvez également imposer un terme dans un dictionnaire de types lorsque vous ajoutez une bibliothèque publique ou la mettez à jour.

Vous pouvez voir les termes qui sont imposés ou ignorés dans la colonne Forcer, en deuxième position dans la sous-fenêtre des termes. Si une icône de punaise apparaît, cela signifie que cette occurrence du terme a été imposée. Si une icône X apparaît en noir, cela signifie que cette occurrence du terme sera ignorée pendant l'extraction, car elle a été imposée ailleurs. Par ailleurs, quand vous imposez un terme, celui-ci apparaît dans la couleur du type dans lequel il a été imposé. Cela signifie que si vous avez imposé dans `Type 1` un terme présent à la fois dans `Type 1` et `Type 2`, il apparaît dans la fenêtre dans la couleur définie pour `Type 1`.

Vous pouvez double-cliquer sur l'icône pour modifier le statut. Si le terme apparaît ailleurs, la boîte de dialogue Résoudre les conflits qui apparaît vous permet de sélectionner l'occurrence à utiliser.

Renommage de types

Vous pouvez renommer un dictionnaire de types ou configurer ses autres paramètres en éditant les propriétés du type.

Important : Nous vous recommandons de ne pas utiliser d'espaces dans les noms de types, en particulier si deux noms de types ou plus commencent par le même mot. Nous vous recommandons également de ne pas renommer les types dans les bibliothèques Principale ou Opinions ni de modifier les attributs de correspondance par défaut.

Pour renommer un type

1. Dans l'arborescence de bibliothèques, sélectionnez le dictionnaire de types que vous souhaitez renommer.
2. Cliquez avec le bouton droit de la souris, puis choisissez **Propriétés de type** dans le menu contextuel. La boîte de dialogue Propriétés de type s'ouvre.
3. Entrez le nouveau nom de votre dictionnaire de types dans la zone de texte Nom.
4. Cliquez sur **OK** pour accepter le nouveau nom. Le nouveau nom du type apparaît dans l'arborescence de bibliothèques.

Déplacement de types

Vous pouvez faire glisser un dictionnaire de types vers un autre emplacement d'une même bibliothèque ou vers une autre bibliothèque de l'arborescence.

Pour réorganiser un type au sein d'une même bibliothèque

1. Dans l'arborescence de bibliothèques, sélectionnez le dictionnaire de types à déplacer.
2. Dans les menus, choisissez **Edition > Monter d'un niveau** pour remonter le dictionnaire de types d'une position dans l'arborescence de la bibliothèque ou **Edition > Descendre d'un niveau** pour le descendre d'une position.

Pour déplacer un type vers une autre bibliothèque

1. Dans l'arborescence de bibliothèques, sélectionnez le dictionnaire de types à déplacer.
2. Cliquez avec le bouton droit de la souris, puis choisissez **Propriétés de type** dans le menu contextuel. La boîte de dialogue Propriétés de type s'ouvre. (Vous pouvez également faire glisser et déposer le type dans une autre bibliothèque).
3. Dans la zone de liste Ajouter à, sélectionnez la bibliothèque dans laquelle vous voulez déplacer le dictionnaire de types.
4. Cliquez sur **OK**. La boîte de dialogue se referme et le type figure maintenant dans la bibliothèque sélectionnée.

Désactivation et suppression de types

Si vous souhaitez supprimer temporairement un dictionnaire de types, vous pouvez la désactiver en décochant la case située à gauche de son nom dans l'arborescence de bibliothèques. Vous indiquez ainsi que vous souhaitez conserver le dictionnaire dans votre bibliothèque, mais en ignorer le contenu pendant la recherche de conflits et le processus d'extraction.

Vous pouvez aussi supprimer définitivement des dictionnaires de types d'une bibliothèque.

Pour désactiver un dictionnaire de types

1. Dans l'arborescence de bibliothèques, sélectionnez le dictionnaire de types à désactiver.
2. Cliquez sur la barre d'espacement. La coche disparaît de la case figurant à gauche du nom du type.

Pour supprimer un dictionnaire de types

1. Dans l'arborescence de bibliothèques, sélectionnez le dictionnaire de types à supprimer.
2. Dans les menus, choisissez **Edition > Supprimer** pour supprimer le dictionnaire de types.

Dictionnaires de substitutions/synonymes

Un *dictionnaire de substitutions* est un ensemble de termes qui permet de regrouper des termes similaires sous un terme cible. Les dictionnaires de substitutions sont gérés dans la sous-fenêtre du bas de l'onglet Ressources de bibliothèque. Vous pouvez accéder à cette vue avec **Vue > Editeur de ressources** dans les menus lors d'une session de plan de travail interactif. Autrement, vous pouvez modifier les dictionnaires d'un modèle spécifique dans l'Editeur de modèle.

Vous pouvez définir deux formes de substitution dans ce dictionnaire : *synonymes* et *éléments facultatifs*. Vous pouvez cliquer sur les onglets de cette sous-fenêtre pour passer de l'un à l'autre.

Après avoir exécuté une extraction sur vos données textuelles, vous pouvez trouver plusieurs concepts qui sont des synonymes ou des formes infléchies d'autres concepts. En identifiant les éléments optionnels et les synonymes, vous pouvez forcer le moteur du programme d'extraction à les faire correspondre à un terme cible unique.

La substitution à l'aide de synonymes et d'éléments optionnels réduit le nombre de concepts dans la sous-fenêtre Résultats d'extraction en les combinant pour former des concepts plus importants et représentatifs avec une fréquence Doc. plus élevée.

Synonymes

Les synonymes sont des mots ayant le même sens. Vous pouvez également utiliser des synonymes pour regrouper des termes et leurs abréviations, ou pour réunir les mots fréquemment mal orthographiés sous la version correcte du mot. Vous pouvez définir ces synonymes dans l'onglet Synonymes.

Une définition de synonyme est constituée de deux parties. La première est un terme **cible**, qui est le terme sous lequel vous souhaitez que le moteur du programme d'extraction regroupe tous les termes synonymes. Ce terme cible deviendra probablement le concept apparaissant dans la sous-fenêtre Résultats d'extraction, sauf s'il est utilisé comme synonyme d'un autre terme cible ou s'il est exclu. La deuxième est la liste de synonymes qui sera regroupée sous le terme cible.

Par exemple, si vous voulez que `automobile` soit remplacé par `véhicule`, `automobile` sera le synonyme et `véhicule` le terme cible.

Vous pouvez entrer n'importe quel mot dans la colonne **Synonyme**, mais si le mot n'est pas trouvé au cours de l'extraction et si le terme avait une option de mise en correspondance de `Entier`, aucune substitution ne peut être effectuée. En revanche, il n'est pas nécessaire que le terme cible soit extrait pour que les synonymes soient regroupés sous ce terme.

Eléments facultatifs

Les éléments optionnels désignent les mots optionnels d'un terme composé qui peuvent être ignorés pendant l'extraction afin de conserver un regroupement de termes similaires, même s'ils apparaissent légèrement différents dans le texte. Les éléments optionnels sont des mots simples dont la suppression d'un terme composé peut créer une correspondance avec un autre terme. Ces mots simples peuvent apparaître à n'importe quel endroit du terme composé, à savoir au début, au milieu ou à la fin. Vous pouvez définir les éléments optionnels dans l'onglet Optionnels.

Par exemple, pour regrouper les termes `ibm` et `ibm corp`, vous devez déclarer que `corp` doit ici être traité comme élément optionnel. Dans un autre exemple, si vous signalez que le terme `accès` est un élément optionnel et que l'extraction renvoie `vitesse d'accès internet` et `vitesse internet`, ils sont regroupés sous le terme qui revient le plus fréquemment.

Définition de synonymes

Dans l'onglet Synonymes, vous pouvez entrer une définition de synonyme dans la ligne vide qui figure en haut du tableau. Commencez par définir le terme cible et ses synonymes. Vous pouvez également sélectionner la bibliothèque dans laquelle vous souhaitez stocker la définition. Toutes les occurrences des synonymes sont ensuite regroupées sous le terme cible lors de l'extraction finale. Pour plus d'informations, voir «Ajout de termes», à la page 190.

Par exemple, si vos données texte comprennent beaucoup d'informations de communication, vous pourriez avoir ces termes : `téléphone portable`, `téléphone sans fil` et `téléphone cellulaire`. Vous pouvez alors décider d'indiquer que `cellulaire` et `mobile` sont synonymes de `portable`. Si vous définissez ces synonymes, chaque occurrence extraite de `téléphone cellulaire` et `téléphone mobile` sera considérée comme le même terme que `téléphone portable` et apparaîtra avec elle dans la liste des termes.

Quand vous créez vos dictionnaires de types, vous pouvez entrer un terme puis penser à trois ou quatre synonymes de ce terme. Dans ce cas, vous pouvez saisir tous les termes puis votre terme cible dans le dictionnaire de substitution avant de faire glisser les synonymes.

La substitution de synonymes est également appliquée aux formes infléchies (le pluriel, par exemple) du synonyme. En fonction du contexte, il est conseillé de définir des limites régissant le processus de substitution des termes. Des caractères sont alors utilisés pour restreindre le traitement des synonymes :

- **Point d'exclamation (!).** Lorsque le point d'exclamation précède directement le synonyme `!synonyme`, cela indique qu'aucune forme infléchie du synonyme ne sera substituée par le terme cible. Néanmoins, un point d'exclamation placé directement avant le terme cible `!target-term` signale qu'aucune partie du mot composé cible ou les variantes ne doit faire l'objet d'une substitution supplémentaire.
- **Astérisque (*).** Un astérisque placé directement après un synonyme, comme `synonyme*`, signifie que vous souhaitez remplacer ce mot par le terme cible. Par exemple, si vous définissez `manage*` comme synonyme et `management` comme terme cible, le mot `managers` est remplacé par le terme cible `management`. Vous pouvez également ajouter un espace et un astérisque après le mot (`synonyme *`) comme `internet *`. Si vous définissez le terme cible `internet` et les synonymes `internet * *` et `web *`, les termes `internet haut débit` et `web indépendant` sont alors remplacés par `internet`. Vous ne pouvez pas utiliser le caractère générique de l'astérisque au début d'un mot ou d'une chaîne dans ce dictionnaire.
- **Caret (^).** Un caret et un espace placés avant le synonyme, comme `^ synonyme`, indiquent que le regroupement de synonymes ne s'applique que lorsque le terme commence par le synonyme. Par exemple, si vous définissez `^ salaire` comme synonyme et `revenus` comme cible et que ces deux termes sont extraits, ils sont regroupés sous le terme `revenus`. En revanche, si « `bulletin de salaire` » et « `revenus` » sont extraits, ils ne sont pas réunis car « `bulletin de salaire` » ne commence pas par « `salaire` ». Un espace doit séparer ce symbole du synonyme.
- **Symbole dollar ($).** Un espace et un symbole dollar placés après le synonyme (`synonyme $`, par exemple) signalent que le regroupement de synonymes ne s'applique que lorsque le synonyme figure à la fin du terme. Par exemple, si vous définissez le synonyme `capital $` et le terme cible `argent` et que ces deux termes sont extraits, ils sont tous deux regroupés sous le terme `argent`. En revanche, si `capital monétaire` et `argent` sont extraits, ils ne sont pas réunis car `capital monétaire` ne se termine pas par `capital`. Un espace doit séparer ce symbole du synonyme.
- **Caret (^) et symbole dollar ($).** Si le caret et le symbole dollar sont utilisés ensemble, comme dans `^ synonyme $`, un terme correspond au synonyme uniquement s'il s'agit d'une correspondance exacte. Cela signifie qu'aucun mot ne peut apparaître avant ou après le synonyme dans le terme extrait pour que le regroupement s'effectue. Par exemple, vous pouvez définir `^ van $` comme synonyme et `fourgon` comme cible pour que `van` soit regroupé avec `fourgon`, mais pas `ludwig van beethoven`. Par ailleurs, dès lors que vous définissez un synonyme à l'aide des symboles caret et dollar et que ce mot apparaît n'importe où dans le texte source, le synonyme est automatiquement extrait.

Pour ajouter une entrée de synonyme

1. Affichez la sous-fenêtre de substitution et cliquez sur l'onglet **Synonymes** dans l'angle inférieur gauche.
2. Dans la ligne vide en haut du tableau, entrez le terme cible dans la colonne Cible. Le terme cible que vous entrez apparaît en couleur. Cette couleur représente le type dans lequel le terme apparaît ou a été imposé, le cas échéant. Si le terme apparaît en noir, il n'est alors inclus dans aucun dictionnaire de types.
3. Cliquez sur la deuxième cellule à droite de la cible et entrez l'ensemble de synonymes. Séparez chaque entrée à l'aide du séparateur global défini dans la boîte de dialogue Options. Pour plus d'informations, voir la rubrique «Définition des options», à la page 75. Les termes que vous entrez apparaissent en couleur. Cette couleur représente le type dans lequel le terme apparaît. Si le terme apparaît en noir, il n'est alors inclus dans aucun dictionnaire de types.
4. Cliquez sur la dernière cellule pour sélectionner la bibliothèque dans laquelle vous voulez stocker cette définition de synonyme.

Remarque : Ces instructions montrent comment effectuer des modifications dans la vue de l'Editeur de ressources ou dans l'Editeur de modèle. Vous pouvez également effectuer ces modifications directement dans la sous-fenêtre Résultats d'extraction , la sous-fenêtre Données, la sous-fenêtre Catégories ou la boîte de dialogue Définitions de cluster dans les autres vues. Pour plus d'informations, voir «Affinage des résultats d'extraction», à la page 89.

Définition des éléments optionnels

Dans l'onglet Optionnels, vous pouvez définir des éléments optionnels pour n'importe quelle bibliothèque. Ces entrées sont regroupées pour chaque bibliothèque. Dès que vous ajoutez une bibliothèque dans l'arborescence de bibliothèques, une ligne vide d'élément optionnel est ajoutée dans l'onglet Optionnels.

Toutes les entrées sont transformées automatiquement en mots en minuscules. Le moteur d'extraction met indifféremment en correspondance les mots du texte en majuscules et en minuscules.

Remarque : Les termes sont délimités à l'aide des séparateurs définis dans la boîte de dialogue Options. Pour plus d'informations, voir la rubrique «Définition des options», à la page 75. Si le séparateur fait partie intégrante de l'élément optionnel entré, vous devez le faire précéder d'une barre oblique inverse.

Pour ajouter une entrée

1. La sous-fenêtre de substitution étant affiché, cliquez sur l'onglet Optionnels dans l'angle inférieur gauche de l'éditeur.
2. Cliquez dans la cellule de la colonne Eléments optionnels de la bibliothèque à laquelle vous souhaitez ajouter cette entrée.
3. Entrez l'élément optionnel. Séparez chaque entrée à l'aide du séparateur global défini dans la boîte de dialogue Options. Pour plus d'informations, voir la rubrique «Définition des options», à la page 75.

Désactivation et suppression de substitutions

Vous pouvez supprimer une entrée de façon temporaire. Pour cela, vous devez la désactiver dans votre dictionnaire. Lorsque vous désactivez une entrée, celle-ci est ignorée au cours des extractions.

Vous pouvez aussi supprimer toutes les entrées obsolètes de votre dictionnaire de substitutions.

Pour désactiver une entrée

1. Dans votre dictionnaire, sélectionnez l'entrée à désactiver.
2. Cliquez sur la barre d'espacement. La coche disparaît de la case figurant à gauche de l'entrée.

Remarque : vous pouvez également décocher la case figurant à gauche de l'entrée pour la désactiver.

Pour supprimer une entrée de synonyme

1. Dans votre dictionnaire, sélectionnez l'entrée à supprimer.
2. Dans le menu, sélectionnez **Edition > Supprimer** ou appuyez sur la touche **Suppr** sur votre clavier. L'entrée ne figure plus dans le dictionnaire.

Pour supprimer une entrée d'élément optionnel

1. Dans votre dictionnaire, double-cliquez sur l'entrée à supprimer.
2. Supprimez manuellement le terme.
3. Appuyez sur Entrée pour appliquer la modification.

Dictionnaires d'exclusions

Un *dictionnaire d'exclusions* est une liste de mots, de phrases ou de chaînes partielles. Tout terme correspondant à ou contenant une entrée dans le dictionnaire d'exclusions sera ignoré ou exclu de l'extraction. La gestion des dictionnaires d'exclusions s'effectue dans la sous-fenêtre de droite de l'éditeur. En règle générale, vous entrez dans cette liste les mots ou expressions de liaison utilisés pour la continuité du texte, mais qui ne lui apportent rien d'important et risquent d'encombrer les résultats de l'extraction. En ajoutant ces termes au dictionnaire d'exclusions, vous êtes assuré qu'ils ne seront jamais extraits.

La gestion des dictionnaires d'exclusion s'effectue dans la sous-fenêtre supérieure droit de l'onglet Ressources de bibliothèque dans l'éditeur. Vous pouvez accéder à cette vue avec **Vue > Editeur de ressources** dans les menus lors d'une session de plan de travail interactif. Autrement, vous pouvez modifier les dictionnaires d'un modèle spécifique dans l'Editeur de modèle.

Vous pouvez entrer un mot, une expression ou une chaîne partielle dans la ligne vide figurant dans le haut du tableau du dictionnaire d'exclusions. Vous pouvez ajouter des chaînes de caractères à votre dictionnaire d'exclusions sous la forme d'un ou de plusieurs mots, voire de mots partiels à l'aide du caractère générique astérisque. Les entrées déclarées dans le dictionnaire d'exclusions servent à empêcher l'extraction de ces concepts. Si une entrée est également déclarée ailleurs dans l'interface, par exemple dans un dictionnaire de types, elle apparaît barrée dans les autres dictionnaires, ce qui indique qu'elle est actuellement exclue. Il n'est pas nécessaire que cette chaîne apparaisse dans les données textuelles ou qu'elle fasse partie d'un dictionnaire de types pour être appliquée.

Remarque : Si vous ajoutez un concept au dictionnaire d'exclusions qui agit également en tant que cible dans une entrée de synonyme, la cible et tous ses synonymes seront également exclus. Pour plus d'informations, voir «Définition de synonymes», à la page 196.

Utilisation de caractères génériques (*)

pouvez utiliser le caractère générique pour indiquer que vous voulez exclure l'entrée à traiter en tant que chaîne partielle. Tous les mots détectés par le moteur du programme d'extraction, et qui commencent ou finissent par l'une des chaînes entrées ici, sont exclus de l'extraction finale. Toutefois, il est interdit d'utiliser les caractères génériques dans deux cas de figure :

- Tiret (-) précédé d'un astérisque, comme *-
- Apostrophe (') précédée d'un astérisque, tel que *'*s*

Tableau 39. Exemples d'entrées d'exclusion

Entrée	Exemple	Résultats
mot	*suivant*	Aucun concept (ou ses termes) n'est extrait s'il contient le mot *suivant*.
expression	*par exemple*	Aucun concept (ou ses termes) n'est extrait s'il contient l'expression *par exemple*.
partiel	*copyright**	Exclut tous les concepts (ou leurs termes) concordants ou contenant des variations du mot *climat*, tels que `climatique`, `climatiseur`, `climatisation` ou `climatologue`.

Tableau 39. Exemples d'entrées d'exclusion (suite)

Entrée	Exemple	Résultats
partiel	**piste*	Exclut tous les concepts (ou leurs termes) concordants ou contenant des variations du mot *piste*, tels que `trappiste`, `monotypiste`, `copiste`, `endoscopiste`, `alpiste` ou `avant-piste`.

Pour ajouter des entrées

- Dans la ligne vide qui figure dans le haut du tableau, entrez un terme. Le terme que vous entrez apparaît en couleur. Cette couleur représente le type dans lequel le terme apparaît. Si le terme apparaît en noir, il n'est alors inclus dans aucun dictionnaire de types.

Pour désactiver des entrées

Vous pouvez supprimer temporairement une entrée en la désactivant dans votre dictionnaire d'exclusions. Lorsque vous désactivez une entrée, celle-ci est ignorée au cours des extractions.

1. Dans votre dictionnaire d'exclusions, sélectionnez l'entrée à désactiver.
2. Cliquez sur la barre d'espacement. La coche disparaît de la case figurant à gauche de l'entrée.

Remarque : Vous pouvez également décocher la case figurant à gauche de l'entrée pour la désactiver.

Pour supprimer des entrées

Vous pouvez supprimer les entrées inutiles de votre dictionnaire d'exclusions.

1. Dans votre dictionnaire d'exclusions, sélectionnez l'entrée à supprimer.
2. Dans les menus, sélectionnez **Edition > Supprimer**. L'entrée ne figure plus dans le dictionnaire.

Chapitre 17. A propos des ressources avancées

Outre les dictionnaires de types, d'exclusions et de substitutions, vous pouvez également utiliser divers paramètres de ressources avancées tels que les paramètres de regroupements flous ou les définitions de type non linguistiques. Vous pouvez accéder à ces ressources à partir de l'onglet Ressources avancées dans la vue de l' Editeur de modèle ou de l'Editeur de ressources.

Dans l'onglet Ressources avancées, vous pouvez modifier les informations suivantes :

- **Langue cible pour les ressources.** Permet de sélectionner la langue pour laquelle les ressources seront créées et affinées. Pour plus d'informations, voir «Langue cible pour les ressources», à la page 203.
- **Regroupement flou (Exceptions).** Permet d'exclure des paires de mots de l'algorithme de regroupement flou (correction des fautes d'orthographe). Pour plus d'informations, voir «Regroupement flou», à la page 203.
- **Entités non linguistiques.** Permet d'activer et de désactiver les entités non linguistiques pouvant être extraites, ainsi que les expressions régulières et les règles de normalisation qui sont appliquées lors de leur extraction. Pour plus d'informations, voir «Entités non linguistiques», à la page 204.
- **Traitement des langues.** Permet de déclarer les méthodes spéciales de structuration des phrases (motifs d'extraction et définitions forcées) et d'utilisation des abréviations pour la langue sélectionnée. Pour plus d'informations, voir «Traitement des langues», à la page 209.

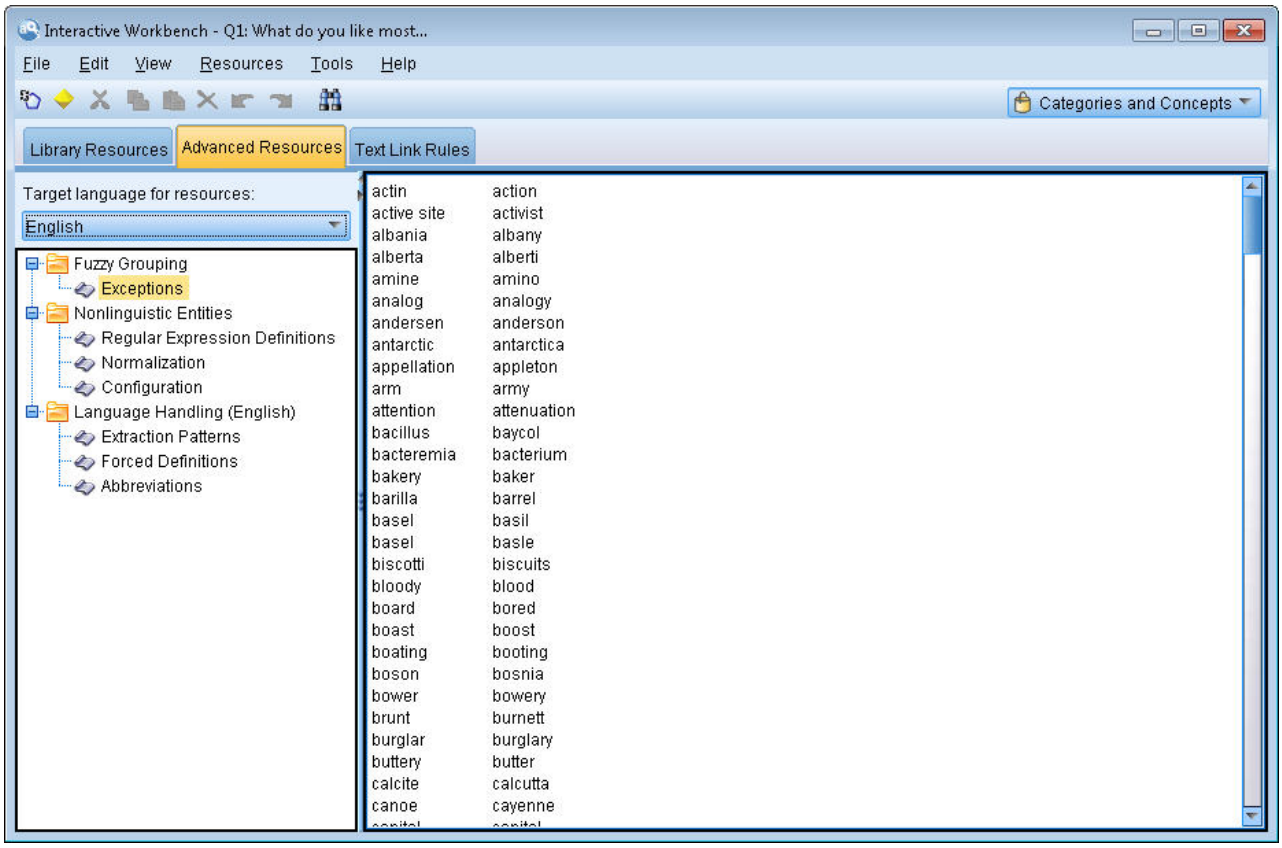


Figure 40. Editeur de modèles de Text Mining - Onglet Ressources avancées

Remarque : Vous pouvez utiliser la barre d'outils de recherche/remplacement pour rechercher rapidement une information et apporter des modifications globales à une section. Pour plus d'informations, voir «Remplacement».

Pour modifier les ressources avancées

1. Localisez et sélectionnez la section de ressources dans laquelle vous souhaitez effectuer la modification. Le contenu apparaît dans la sous-fenêtre de droite.
2. Si nécessaire, à l'aide du menu ou des boutons de la barre d'outils, coupez, copiez ou collez des éléments.
3. Editez les fichiers que vous souhaitez modifier en utilisant les règles de formatage de la section. Les modifications sont enregistrées dès que vous les apportez. Utilisez la flèche d'annulation ou de rétablissement de la barre d'outils pour rétablir vos précédentes modifications.

Recherche

Il peut s'avérer nécessaire de localiser rapidement des informations dans une section particulière. Par exemple, si vous procédez à l'analyse des liens du texte, vous disposez de centaines de macros et définitions de motifs. La fonction Rechercher vous permet de rechercher rapidement une règle spécifique. Pour rechercher des informations dans une section, utilisez la barre d'outils Rechercher.

Pour utiliser la fonction de recherche

1. Localisez et sélectionnez la section de ressources dans laquelle vous souhaitez effectuer la recherche. Son contenu apparaît dans la sous-fenêtre de droite de l'éditeur.
2. Dans les menus, sélectionnez **Edition > Rechercher**. La barre d'outils Rechercher apparaît dans la partie supérieure droite de la boîte de dialogue Editer les ressources avancées.
3. Saisissez la chaîne de mots que vous recherchez dans la zone de texte. Vous pouvez contrôler la casse, les correspondances partielles et le sens de la recherche à l'aide des boutons de la barre d'outils.
4. Cliquez sur **Rechercher** pour lancer la recherche. Si une correspondance est détectée, le texte est mis en surbrillance dans la fenêtre.
5. Cliquez de nouveau sur **Rechercher** pour rechercher la correspondance suivante.

Remarque : Dans l'onglet Règles des liens du texte, l'option Rechercher est disponible uniquement lorsque vous visualisez le code source.

Remplacement

Dans certains cas, il peut s'avérer nécessaire d'effectuer des mises à jour globales à vos ressources avancées. La fonction Remplacer peut vous aider à effectuer des mises à jour homogènes sur vos informations.

Pour utiliser la fonction de remplacement

1. Localisez et sélectionnez la section de ressources dans laquelle vous souhaitez effectuer la recherche et le remplacement. Son contenu apparaît dans la sous-fenêtre de droite de l'éditeur.
2. Dans les menus, sélectionnez **Edition > Remplacer**. La boîte de dialogue Remplacer apparaît.
3. Dans la zone de texte **Rechercher**, saisissez la chaîne de mots que vous souhaitez rechercher.
4. Dans la zone de texte **Remplacer par**, saisissez la chaîne que vous souhaitez utiliser à la place du texte trouvé.
5. Sélectionnez **Rechercher uniquement les mots entiers** si vous ne souhaitez rechercher ou remplacer que des mots complets.
6. Sélectionnez **Respecter la casse** si vous ne souhaitez rechercher ou remplacer que les mots dont la casse correspond exactement.

7. Cliquez sur **Suivant** pour rechercher une correspondance. Si une correspondance est détectée, le texte est mis en surbrillance dans la fenêtre. Si vous ne voulez pas remplacer cette correspondance, cliquez de nouveau sur **Suivant** jusqu'à ce que vous ayez trouvé une correspondance que vous souhaitez remplacer.
8. Cliquez sur **Remplacer** pour remplacer la correspondance sélectionnée.
9. Cliquez sur **Remplacer** pour remplacer toutes les correspondances de la section. Un message indique le nombre de remplacements effectués.
10. Lorsque les remplacements sont terminés, cliquez sur **Fermer**. La boîte de dialogue se ferme.

Remarque : Si vous effectuez un remplacement par erreur, vous pouvez l'annuler en fermant la boîte de dialogue et en choisissant **Edition > Annuler** à partir des menus. Vous devez suivre cette procédure pour chaque modification à annuler.

Langue cible pour les ressources

Les ressources sont créées pour une langue de texte spécifique. La langue pour laquelle ces ressources sont adaptées est définie dans l'onglet Ressources avancées. Si nécessaire, vous pouvez passer à une autre langue en sélectionnant cette langue dans la zone de liste déroulante **Langue cible pour les ressources**. En outre, la langue répertoriée ici sera la langue utilisée pour tout pack d'analyse de texte que vous créez avec ces ressources.

Important : Vous aurez rarement besoin de modifier la langue de vos ressources. Si vous la modifiez, cela peut générer des problèmes, puisque dans ce cas vos ressources ne correspondraient plus à la langue d'extraction. Toutefois, vous pouvez avoir besoin de modifier une langue si vous projetez d'utiliser l'option `TOUTES` les langues lors de l'extraction, dans le cas où votre texte comporte plusieurs langues. Lorsque vous changez de langue, vous accédez, par exemple, aux ressources de traitement pour les motifs d'extraction, les abréviations et les définitions forcées qui correspondent à la deuxième langue à laquelle vous vous intéressez. Toutefois, gardez présent à l'esprit le fait que, avant de publier ou de sauvegarder les modifications de ressources que vous avez apportées ou avant d'exécuter une nouvelle extraction, vous devez rétablir la langue principale avec laquelle vous travaillez pour l'extraction.

Regroupement flou

Dans le noeud Text Mining et dans Paramètres d'extraction, si vous sélectionnez **Traitement des fautes de frappe. Nombre de caractères minimum requis :**, vous avez activé l'algorithme de regroupement flou.

Le regroupement flou permet de regrouper les mots dont l'orthographe est souvent incorrecte ou proche en supprimant temporairement toutes les voyelles (à l'exception de la première voyelle) et les consonnes doubles ou triples des mots extraits, et en les comparant ensuite afin de déterminer s'ils sont identiques. Pendant le processus d'extraction, la fonction de regroupement flou est appliquée aux termes extraits et les résultats sont comparés afin de déterminer s'il existe des correspondances. Si tel est le cas, les termes initiaux sont regroupés dans la liste d'extraction finale. Ils sont placés sous le terme qui compte le plus d'occurrences dans les données.

Remarque : Si les deux termes comparés sont affectés à des types différents, à l'exception du type `<Unknown>`, la méthode de regroupement flou n'est pas appliquée à cette paire. Autrement dit, les termes doivent appartenir au même type ou au type `<Unknown>` afin d'appliquer la méthode.

Si vous avez activé cette fonction et remarqué que deux mots orthographiés de façon similaire étaient regroupés de façon erronée, vous pouvez les exclure du regroupement flou. Pour cela, entrez les paires dont la mise en correspondance est incorrecte dans la section Exceptions de l'onglet Ressources avancées. Pour plus d'informations, voir Chapitre 17, «A propos des ressources avancées», à la page 201.

L'exemple suivant montre le fonctionnement du regroupement flou. Si le regroupement flou est activé, ces mots apparaissent similaires et sont mis en correspondance comme suit :

```
couleur  -> colr                paysage -> pasg
couleurs -> colr                passage  -> pasg

modelisation  -> modlstn           fourniture -> forntr
rondelle -> rondl           conjoncture -> conjtr
```

Dans l'exemple précédent, vous souhaiterez vraisemblablement que les termes `paysage` et `passage` ne soient pas regroupés. Par conséquent, vous pouvez les entrer dans la section Exceptions comme suit :

```
paysage      passage
```

Important : Dans certains cas, les exceptions de regroupement flou n'empêchent pas 2 mots d'être mis en paire car certaines règles de synonymes s'appliquent. Dans ce cas, vous pouvez essayer d'entrer les synonymes en utilisant le caractère générique ! (point d'exclamation) pour empêcher les mots de devenir des synonymes dans la sortie. Pour plus d'informations, voir «Définition de synonymes», à la page 196.

Règles de formatage pour le Regroupement flou (Exceptions)

- Définissez une seule paire d'exception par ligne.
- Utilisez des mots simples ou composés.
- N'utilisez que des minuscules dans les mots. Les mots en majuscules seront ignorés.
- Utilisez le caractère `TAB` pour séparer les deux mots de chaque paire.

Entités non linguistiques

Lorsque vous travaillez avec certains types de données, vous pouvez souhaiter extraire des dates, des numéros de Sécurité sociale, des pourcentages ou toute autre entité non linguistique. Ces entités sont déclarées explicitement dans le fichier de configuration, dans lequel vous pouvez les activer ou les désactiver. Pour plus d'informations, voir «Configuration», à la page 208. Pour optimiser la sortie du moteur d'extraction, l'entrée de traitement non linguistique est normalisé dans des entités en fonction de formats prédéfinis. Pour plus d'informations, voir «Méthode de normalisation», à la page 207.

Remarque : Vous pouvez activer ou désactiver l'extraction d'entités non linguistiques dans les paramètres d'extraction.

Entités non linguistiques disponibles

Les entités non linguistiques du tableau suivant peuvent être extraites. Le nom de type est entre parenthèses.

Tableau 40. Entités non linguistiques pouvant être extraites

Adresses	(<Adresse>)
Acides aminés	(<Acide aminé>)
Devises	(<Devise>)
Dates	(<Date>)
Délai	(<Délai>)
Chiffres	(<Chiffre>)
Adresses électroniques	(<adresse électronique>)
Adresses HTTP/URL	(<url>)
Adresse IP	(<IP>
Organisations	(<Organisation>)
Pourcentages	(<Pourcentage>)
Produits	(<Produit>)

Tableau 40. Entités non linguistiques pouvant être extraites (suite)

Protéines	(<Gene>)
Numéros de téléphone	(<NuméroTéléphone>)
Heures	(<Heure>)
Numéro de sécurité sociale (Etats-Unis)	(<NuméroSécuritéSociale>)
Poids et mesures	(<Poids-Mesures>)

Nettoyage du texte pour traitement

Avant l'extraction des entités non linguistiques, le texte d'entrée est nettoyé. Durant cette étape, les modifications temporaires suivantes sont effectuées afin que les entités non linguistiques puissent être identifiées et extraites ainsi :

- Toute séquence de deux espaces ou plus est remplacée par un seul espace.
- Les tabulations sont remplacées par un espace.
- Les caractères de fin de ligne uniques ou les caractères de séquence sont remplacés par un espace, tandis que les séquences de fin de ligne multiples sont marquées comme la fin d'un paragraphe. Une fin de ligne peut être indiquée par des retours chariot (CR) et des sauts de ligne (LF) ou les deux à la fois.
- Les balises HTML et XML sont temporairement supprimées et ignorées.

Expressions régulières

Lors de l'extraction d'entités non linguistiques, vous pouvez modifier ou ajouter des définitions d'expressions régulières à utiliser pour identifier des expressions régulières. Pour cela, rendez-vous dans la section **Expressions régulières** de l'onglet Ressources avancées. Pour plus d'informations, voir Chapitre 17, «A propos des ressources avancées», à la page 201.

Le fichier est divisé en plusieurs sections distinctes. La première section s'intitule `[macros]`. Une autre section peut venir s'y ajouter, pour chaque entité non linguistique. Vous pouvez ajouter des sections à ce fichier. Dans chaque section, les règles sont numérotées (*regexp1*, *regexp2*, etc.). Ces règles doivent être numérotées séquentiellement de 1 à *n*. Toute rupture dans la numérotation entraîne la suspension du traitement de ce fichier.

Dans certains cas, une entité varie en fonction de la langue. Une entité est considérée comme variant en fonction de la langue si son paramètre de langue a une valeur autre que 0 dans le fichier de configuration. Pour plus d'informations, voir «Configuration», à la page 208. Lorsqu'une entité varie en fonction de la langue, celle-ci doit être utilisée comme préfixe de nom de section, par exemple `[english/PhoneNumber]`. Si l'entité `PhoneNumber` adopte la valeur de langue 2, cette section contient des règles applicables uniquement aux numéros de téléphone anglais.

Important ! Si vous modifiez ce fichier (ou un autre) dans l'éditeur et que le moteur du programme d'extraction ne fonctionne plus comme vous le souhaitez, sélectionnez l'option **Rétablir les valeurs d'origine** dans la barre d'outils pour rétablir le fichier d'origine livré. Ce fichier requiert un certain niveau de connaissance des expressions standard. Pour obtenir une assistance dans ce domaine, contactez IBM Corp..

Caractères spéciaux. [] {} () \ * + ? | ^ $

Tous les caractères équivalent à eux-mêmes sauf les caractères spéciaux suivants, utilisés dans des expressions : `.[{()\*+?|^$` Pour utiliser ces caractères pour eux-mêmes, faites-les précéder d'une barre oblique inverse (\).

Par exemple, si vous essayez d'extraire des adresses Web, le point est très important pour l'entité, par conséquent, vous devez le faire précéder d'une barre oblique inverse :

```
www\.[a-z]+\.[a-z]+
```

Opérateurs de répétition et quantificateurs ? + * {}

Pour rendre les définitions plus flexibles, vous pouvez utiliser plusieurs caractères génériques qui sont standard dans les expressions régulières. Il s'agit de `* ? +`

- *L'astérisque* * indique qu'il y a *zéro occurrence ou plus* de la chaîne précédente. Par exemple, `ab*c` correspond à "*ac*", "*abc*", "*abbbc*", etc.
- *Le signe plus* + indique qu'il y a *une ou plusieurs occurrences* de la chaîne précédente. Par exemple, `ab+c` correspond à "*abc*", "*abbc*", "*abbbc*", mais pas à "*ac*".
- *Point d'interrogation ?* indique qu'il y a *zéro ou une occurrence* de la chaîne précédente. Par exemple, `modell?ing` correspond à "*modeling*" et à "*modeling*".
- *Limiter la répétition avec des crochets {}* indique les limites de la répétition. Par exemple,

  `[0-9]{n}` correspond à un chiffre répété exactement *n* fois. Par exemple, `[0-9]{4}` correspond à "*1998*", mais ni à "*33*" ni à "*19983*".

  `[0-9]{n,}` correspond à un chiffre répété *n fois ou plus*. Par exemple, `[0-9]{3,}` correspond à "*199*" ou "*1998*", mais pas à "*19*".

  `[0-9]{n,m}` correspond à un chiffre répété entre *n et m fois inclus*. Par exemple, `[0-9]{3,5}` correspond à "*199*", "*1998*" ou "*19983*", mais pas à "*19*" ni à "*199835*".

Espaces facultatifs et traits d'union

Dans certains cas, vous voulez inclure un espace facultatif dans une définition. Par exemple, si vous vouliez extraire des devises telles que « *pesos uruguayens* », « *peso uruguayen* », « *pesos Uruguay* », « *peso Uruguay* », « *pesos* » ou « *peso* », vous devriez tenir compte du fait qu'il peut y avoir deux mots séparés par un espace. Dans ce cas, cette définition devrait être écrite sous la forme `?pesos?( uruguayens| Uruguay )`. Puisque *uruguayen* ou *Uruguay* sont précédés par un espace lorsqu'ils sont utilisés avec *pesos/peso*, l'espace facultatif doit être défini à l'intérieur de la séquence facultative `( uruguayens| Uruguay)`. Si l'espace n'est pas inclus dans la séquence, comme dans `(uruguayan|uruguay)? pesos?` par exemple, "*pesos*" ou "*peso*" ne seront pas pris en compte, car l'espace est requis.

Si vous cherchez des séries d'éléments incluant des traits d'union (-) dans une liste, le trait d'union doit être défini en dernier. Par exemple, si vous cherchez une virgule (,) ou un trait d'union (-), utilisez `[,-]` et non pas `[-,]`.

Ordre des chaînes dans les listes et les macros

Vous devez toujours définir la séquence la plus longue avant une plus courte, sinon la plus longue ne sera jamais lue puisque la correspondance se fera sur la plus courte. Par exemple, si vous recherchez les chaînes "*billion*" ou "*bill*", "*billion*" doit être défini avant "*bill*". Ainsi, par exemple `(janvier|janv)` et non `(janv|janvier)`. Cela s'applique aussi aux macros, puisque les macros sont des listes de chaînes.

Ordre des règles dans la section des définitions

Définissez une seule règle par ligne. Dans chaque section, les règles sont numérotées (*regexp1*, *regexp2*, etc.). Ces règles doivent être numérotées séquentiellement de 1 à *n*. Toute rupture dans la numérotation entraîne la suspension du traitement de ce fichier. Pour désactiver une entrée, placez le symbole # au début de la ligne utilisée pour définir l'expression régulière. Pour activer une entrée, supprimez le caractère # en début de ligne.

Dans chaque section, les règles les plus spécifiques doivent être définies avant les plus générales afin de garantir un traitement correct. Par exemple, si vous recherchez une date au format "*mois année*" et au format "*mois*", la règle "*mois année*" doit être définie avant la règle "*mois*". Voici comment elle doit être définie :

```
#@# January 1932
regexp1=$(MONTH),? [0-9]{4}

#@# January
regexp2=$(MONTH)
```

et non

```
#@# January
regexp1=$(MONTH)

#@# January 1932
regexp2=$(MONTH),? [0-9]{4}
```

Utilisation des macros dans les règles

Chaque fois qu'une séquence spécifique est utilisée dans plusieurs règles, vous pouvez utiliser une macro. Ensuite, si vous devez modifier la définition de cette séquence, vous aurez à la modifier une seule fois et non dans toutes les règles y faisant référence. Par exemple, en supposant que vous ayez la macro suivante :

```
MONTH=((january|february|march|april|june|july|august|september|october|
november|december)|(jan|feb|mar|apr|may|jun|jul|aug|sep|oct|nov|dec)(\.)?)
```

Lorsque vous faites référence au nom de la macro, utilisez la syntaxe `$()`, par exemple : `regexp1=$(MONTH)`

Toutes les macros doivent être définies dans la section `[macros]`.

Méthode de normalisation

Lors de l'extraction d'entités non linguistiques, les entités rencontrées sont normalisées pour regrouper les entités semblables en fonction de formats prédéfinis. Par exemple, les symboles de devise et leur équivalent en toutes lettres sont considérés comme étant identiques. Les entrées de normalisation sont stockées dans la section **Normalisation** de l'onglet Ressources avancées. Pour plus d'informations, voir Chapitre 17, «A propos des ressources avancées», à la page 201. Le fichier est divisé en plusieurs sections distinctes.

Important ! Ce fichier s'adresse à des utilisateurs avertis uniquement. Il est peu probable que vous ayez à modifier ce fichier. Pour obtenir une assistance dans ce domaine, contactez IBM Corp..

Règles de formatage pour normalisation

- Ajoutez uniquement une entrée de normalisation par ligne.
- Respectez bien les sections de ce fichier. Vous ne pouvez pas ajouter de nouvelles sections.
- Pour désactiver une entrée, entrez le symbole # au début de la ligne concernée. Pour activer une entrée, supprimez le caractère # en début de ligne.

Dates au format anglais dans la normalisation

Par défaut, les dates dans un modèle en anglais sont formatées dans le style anglais-américain, c'es-à-dire, mois, jour, année. Pour modifier ce format en jour, mois année, désactivez la ligne "format:US" (en ajoutant un caractère # en début de ligne) et activez la ligne "format:UK" (en supprimant le caractère # en début de ligne).

Configuration

Vous pouvez activer et désactiver les types d'entité non linguistique que vous souhaitez extraire dans le fichier de configuration des entités non linguistiques. Désactivez les entités dont vous n'avez pas besoin pour diminuer le temps de traitement nécessaire. Pour cela, rendez-vous dans la section **Configuration** de l'onglet Ressources avancées. Pour plus d'informations, voir Chapitre 17, «A propos des ressources avancées», à la page 201. Si l'extraction non linguistique est activée, le moteur d'extraction lit ce fichier de configuration au cours du processus d'extraction et détermine quels types d'entités non linguistiques doivent être extraites.

La syntaxe de ce fichier est la suivante :

#name<TAB>*Language*<TAB>*Code*

Tableau 41. Syntaxe du fichier de configuration.

Libellé de colonne	Description
#name	Libellé selon lequel les entités non linguistiques seront référencées dans les deux autres fichiers nécessaires à l'extraction d'entités non linguistiques. Les mots utilisés ici distinguent les majuscules des minuscules.
Language	Langue des documents . Il est préférable de sélectionner une langue précise. Toutefois, vous pouvez également choisir l'option **N'importe**. Les options possibles sont les suivantes : 0 = N'importe quelle langue utilisée lorsqu'une règle regexp n'est pas spécifique à une langue et peut être utilisé dans plusieurs modèles avec différentes langues, par exemple une adresse IP/URL/email ; 1 = français ; 2 = anglais ; 4 = allemand ; 5 = espagnol ; 6 = néerlandais ; 8 = portugais ; 10 = italien.
Code	Code de catégories grammaticales. La plupart des entités prennent la valeur “s”. Les valeurs possibles sont : s = mot vide ; a = adjectif ; n = nom. Lorsque cette option est activée, les entités non linguistiques sont d'abord extraites, puis les motifs d'extraction sont appliqués afin d'identifier le rôle dans un contexte plus large. Par exemple, la valeur "a" est affectée aux pourcentages. Imaginons que la valeur 30 % soit extraite en tant qu'entité non linguistique. Elle sera identifiée comme étant un adjectif. Ensuite, si le texte contient "30% d'augmentation de salaire," l'entité non linguistique “30%” correspond au motif de partie du discours “ann” (adjectif nom nom).

Ordre dans la définition des entités

L'ordre dans lequel les entités sont déclarées dans ce fichier est important car il affecte leur mode d'extraction. Ces entrées sont appliquées dans l'ordre dans lequel elles sont répertoriées. Toute modification de l'ordre se répercute sur les résultats. Les entités non linguistiques les plus spécifiques doivent être définies avant les plus générales.

Par exemple, l'entité non liguistique “Acide aminé” est définie par :

```
regexp1=($(AA)-?$(NUM))
```

où $(AA) correspond à “(ala|arg|asn|asp|cys|gln|glu|gly|his|ile|leu|lys|met|phe|pro|ser)”, qui sont des séquences de 3 lettres correspondant à des acides aminés particuliers.

D'autre part, l'entité non linguistique « Gène » est plus générale et elle est définie par :

```
regexp1=p[0-9]{2,3}
regexp2=[a-z]{2,4}-?[0-9]{1,3}-?[r]
regexp3=[a-z]{2,4}-?[0-9]{1,3}-?p?
```

Si « Gène » est défini avant « Acide aminé » dans la section Configuration, alors « Acide aminé » ne sera jamais mis en correspondance, puisque regexp3 de « Gène » sera toujours mis en correspondance en premier.

Règles de formatage pour la configuration

- Utilisez le caractère TAB pour séparer les entrées d'une colonne.
- Ne supprimez aucune ligne.
- Respectez la syntaxe indiquée dans le tableau précédent.
- Pour désactiver une entrée, entrez le symbole # au début de la ligne concernée. Pour activer une entité, supprimez le caractère # en début de ligne.

Traitement des langues

Chaque langue moderne exprime des idées, structure des phrases et utilise des abréviations d'une façon particulière. Dans la section Gestion des langues, vous pouvez éditer les motifs d'extraction, forcer des définitions pour ces motifs et indiquer des abréviations pour la langue que vous avez sélectionnée dans la liste déroulante Langue.

- Motifs d'extraction
- Définitions forcées
- Abréviations

Motifs d'extraction

Lors de l'extraction d'informations depuis vos documents, le moteur d'extraction applique un ensemble de motifs d'extraction de parties du discours à une "pile" de mots du texte afin d'identifier les termes (mots ou expressions) candidats à l'extraction. Vous pouvez ajouter ou modifier les motifs d'extraction.

Les parties du discours incluent les éléments grammaticaux tels que substantifs, adjectifs, participes passés, déterminants, prépositions, coordonnants, prénoms, sigles et particules. Une série de ces éléments constitue un motif d'extraction. Dans les produits de Text Mining IBM Corp., chaque catégorie grammaticale est représentée par un caractère unique afin de faciliter la définition des motifs. Par exemple, un adjectif est représenté par la lettre *a* en minuscule. L'ensemble des codes pris en charge apparaît par défaut au début de chaque section de motifs d'extraction, avec des exemples illustrant chaque motif et chaque code.

Règles de formatage des motifs d'extraction

- Un motif par ligne.
- Utilisez le caractère # en début de ligne pour désactiver un motif.

L'ordre dans lequel vous répertoriez les motifs d'extraction est très important car une séquence de mots donnée est lue une seule fois par le moteur du programme d'extraction et est affectée au premier motif d'extraction pour lequel le moteur détecte une correspondance.

Codes de partie du discours pris en charge

Vous trouverez ci-après un tableau de tous les codes de partie du discours pris en charges qui sont définis dans le dictionnaire anglais compilé.

Toutes les parties du discours utilisées dans un modèle particulier sont répertoriées au début de **Ressources avancées** > **Motifs d'extraction**.

La principale différence entre les modèles Ressources de base et Opinions est la suivante : lorsque des déterminants ("d") et des prépositions ("c") minimaux sont utilisés dans le modèle Ressources de base, leurs équivalents étendus ("e" et "r") sont utilisés dans le modèle Opinions. En outre, dans le modèle Opinions, tous les termes avec des parties du discours "a" et "Q" ne sont traités que comme "Q." "0," "1" et "2" ont une utilisation limitée dans tous les modèles Opinions. Voir **Ressources avancées** > **Traitement des langues (anglais)** > **Définitions forcées et Motifs d'extraction**.

D'autres modèles anglais peuvent utiliser des parties du discours non répertoriées dans le dictionnaire (par exemple, "w" et "W", dans le modèle Market Intelligence). Toutefois, ces parties du discours sont alors affectées à des termes spécifiques sous **Ressources avancées** > **Définitions forcées**.

Tableau 42. Codes de partie du discours pris en charge

Code	Signification	Exemple
a	adjectif	abdominal, bleu...
A	inutilisé	inutilisé
b	adverbe	fréquemment, souvent, très, ...
B	inutilisé	inutilisé
c	préposition	"de"
C	code interne pour les mots mal orthographiés	
d	déterminant	"le"
D	inutilisé	inutilisé
e	étendu	déterminant le, un, mon, votre...
E	inutilisé	inutilisé
f	prénom	Jean, Marie...
F	inutilisé	inutilisé
g	inutilisé	inutilisé
G	adjectif de nationalité	français, américain...
h	inutilisé	inutilisé
H	inutilisé	inutilisé
i	inutilisé	inutilisé
I	inutilisé	inutilisé
j	inutilisé	inutilisé
J	inutilisé	inutilisé
k	inutilisé	inutilisé
K	inutilisé	inutilisé
l	inutilisé	inutilisé
L	inutilisé	inutilisé
m	nom ou inconnu	chien, ibm
M	inutilisé	inutilisé
n	nom	chien
N	inutilisé	inutilisé
o	coordination	"et", "&"
O	inutilisé	inutilisé
p	participe passé	abandonné, personnalisé...
P	inutilisé	inutilisé
q	inutilisé	inutilisé
Q	qualificateur	cher, petit, correct, ...
r	préposition étendue	de, parmi, contre, à partir de...
R	inutilisé	inutilisé

Tableau 42. Codes de partie du discours pris en charge (suite)

Code	Signification	Exemple
s	mot vide	tout mot que nous ne souhaitons pas extraire
S	inutilisé	inutilisé
t	titre	mme, m., capitaine, brig., ...
T	adjectifs techniques	ultra-violet... (tous les "T" sont également des "a")
u	inconnu par définition, absent du dictionnaire	
U	inutilisé	inutilisé
v	verbe	manger, mange, mangea, mangeant, ...
V	verbe à l'infinitif	manger, ...
w	inutilisé	inutilisé
W	inutilisé	inutilisé
x	auxiliaire	être
X	inutilisé	inutilisé
y	particule	von, di, de, ... (permet d'extraire des noms de personne : John von Doe)
Y	inutilisé	inutilisé
z	inutilisé	inutilisé
Z	inutilisé	inutilisé
0	adverbe d'opinion	Uniquement dans Opinions. Voir **Ressources avancées** > **Traitement des langues (anglais)** > **Définitions forcées**.
1	"vers" dans opinions	Voir **Ressources avancées** > **Traitement des langues (anglais)** > **Définitions forcées**.
2	qualificatif spécifique	Uniquement dans Opinions. Voir **Ressources avancées** > **Traitement des langues (anglais)** > **Définitions forcées**.
3	inutilisé	inutilisé
4	inutilisé	inutilisé
5	inutilisé	inutilisé
6	inutilisé	inutilisé
7	inutilisé	inutilisé
8	inutilisé	inutilisé
9	inutilisé	inutilisé

Définitions forcées

Lors de l'extraction d'informations depuis vos documents , le moteur d'extraction analyse le texte et identifie la partie du discours de chaque mot rencontré. Dans certains cas, un mot peut présenter différents rôles en fonction du contexte. Pour imposer une catégorie grammaticale particulière à un mot ou pour exclure complètement le mot du traitement, vous pouvez utilisez la section **Définition forcée** dans l'onglet Ressources avancées. Pour plus d'informations, voir Chapitre 17, «A propos des ressources avancées», à la page 201.

Pour imposer une catégorie grammaticale pour un mot donné, vous devez ajouter une ligne à cette section à l'aide de la syntaxe suivante :

```
term:code
```

Tableau 43. Description de la syntaxe.

Entrée	Description
`term`	Un nom de terme.
`code`	Code à caractère unique représentant la catégorie grammaticale. Vous pouvez répertorier jusqu'à six codes de catégorie grammaticale différents par expression uniterme. En outre, vous pouvez arrêter l'extraction d'un mot en mots ou phrases composés en utilisant le code `s` (en minuscule), par exemple : `additional:s`.

Règles de formatage applicables aux définitions forcées.

- Une ligne par mot.
- Les termes ne peuvent pas contenir le signe « deux-points ».
- Utilisez le code `s` (lettre minuscule) en tant que code de catégorie grammaticale pour arrêter définitivement l'extraction d'un mot.
- Utilisez jusqu'à six codes de catégorie grammaticale par ligne. Les codes de catégories grammaticales pris en charge sont affichés dans la section Motifs d'extraction. Pour plus d'informations, voir «Motifs d'extraction», à la page 209.
- Utilisez l'astérisque (*) en tant que caractère générique à la fin d'une chaîne pour les correspondances partielles. Par exemple, si vous saisissez `add*:s`, des mots tels que `add`, `addition`, `additionnel`, `addenda` et `additif` ne sont jamais extraits en tant que mot ou dans le cadre d'un mot composé. Toutefois, si une correspondance de mot est explicitement déclarée en tant que terme dans un dictionnaire compilé ou dans les définitions forcées, le mot sera extrait. Par exemple, si vous saisissez à la fois `add*:s` et `addenda:n`, le mot `addenda` sera extrait s'il est détecté dans le texte.

Abréviations

Lorsque le moteur du programme d'extraction traite le texte, il interprète généralement tout point rencontré comme marquant la fin d'une phrase. La plupart du temps, cela s'avère correct ; toutefois, cette gestion des signes de ponctuation que sont les points ne s'applique pas lorsque le texte contient des abréviations.

Si vous extrayez des termes de votre texte et que vous vous apercevez que certaines abréviations ont été mal gérées, vous devez déclarer ces dernières de manière explicite dans cette section.

Remarque : si l'abréviation apparaît déjà dans une définition de synonyme ou est définie en tant que terme dans un dictionnaire type, il est inutile de l'ajouter ici.

Règles de formatage applicables aux abréviations

- Définissez uniquement une abréviation par ligne.

Chapitre 18. A propos des règles des liens du texte

L'analyse des liens du texte (TLA) est une technologie de mise en correspondance de motifs, utilisée pour extraire les relations trouvées dans votre texte à l'aide d'un ensemble de règles. Lorsque l'analyse des liens du texte est activée pour l'extraction, les données du texte sont évaluées en fonction de ces règles. Lorsqu'une correspondance est trouvée, le motif de l'analyse des liens du texte est extrait et présenté. Ces règles sont définies dans l'onglet Règles des liens du texte.

Par exemple, extraire des informations sur une organisation peut ne pas représenter assez d'intérêt pour vous, mais grâce à l'analyse des liens du texte, vous pouvez également vous familiariser avec les liens existant entre différentes organisations ou avec les personnes associées à cette organisation. La TLA peut également permettre d'extraire des opinions sur différents sujets comme le ressenti des gens face à un produit ou une expérience.

Pour bénéficier de l'analyse TLA, vous devez disposer de ressources contenant des règles de liens du texte (TLA). Lorsque vous sélectionnez un modèle, vous pouvez voir les modèles qui ont des règles TLA selon qu'ils ont ou non une icône dans la colonne TLA.

Les motifs d'analyse des liens du texte sont extraits des données textuelles pendant la phase de mise en correspondance des motifs du processus d'extraction. Pendant cette phase, les règles sont comparées aux données textuelles et lorsqu'une correspondance est trouvée, ces informations sont extraites en tant que motif. Il est possible que vous exigiez davantage de l'analyse des liens du texte ou que vous souhaitiez modifier les critères des correspondances. Dans ces cas, vous pouvez affiner les règles pour les adapter à vos besoins spécifiques. Pour cela, utilisez l'onglet Règles des liens du texte.

Remarque : la prise en charge des variables a cessé dans la version 13. Utilisez les macros à la place. Pour plus d'informations, voir «Utilisation des macros», à la page 218.

Utilisation des règles des liens du texte

Vous pouvez éditer et créer des règles directement dans l'onglet Règles des liens du texte de la vue Editeur de modèle ou Editeur de ressources. Pour obtenir un aperçu de la façon dont les règles peuvent correspondre au texte, vous pouvez effectuer une simulation dans cet onglet. Pendant la simulation, une extraction est effectuée sur l'échantillon de données de simulation uniquement et les règles des liens du texte sont appliquées pour savoir si des motifs correspondants existent. Toute règle qui correspond au texte apparaît ensuite dans la sous-fenêtre de simulation. En fonction de ces correspondances, vous pouvez choisir d'éditer les règles et des macros pour modifier les critères de correspondance du texte.

Contrairement à d'autres ressources avancées, les règles TLA sont spécifiques à la bibliothèque ; c'est pourquoi vous pouvez uniquement utiliser les règles TLA d'une seule bibliothèque à la fois. A partir de l'Editeur de modèle ou de l' Editeur de ressources, accédez à l'onglet **Règles des liens du texte**. Dans cet onglet, vous pouvez spécifier la bibliothèque de votre modèle qui contient les règles TLA que vous souhaitez utiliser ou modifier. C'est pourquoi nous vous conseillons fortement de stocker toutes vos règles dans une seule bibliothèque, à moins que vous ne le souhaitiez pas pour une raison particulière.

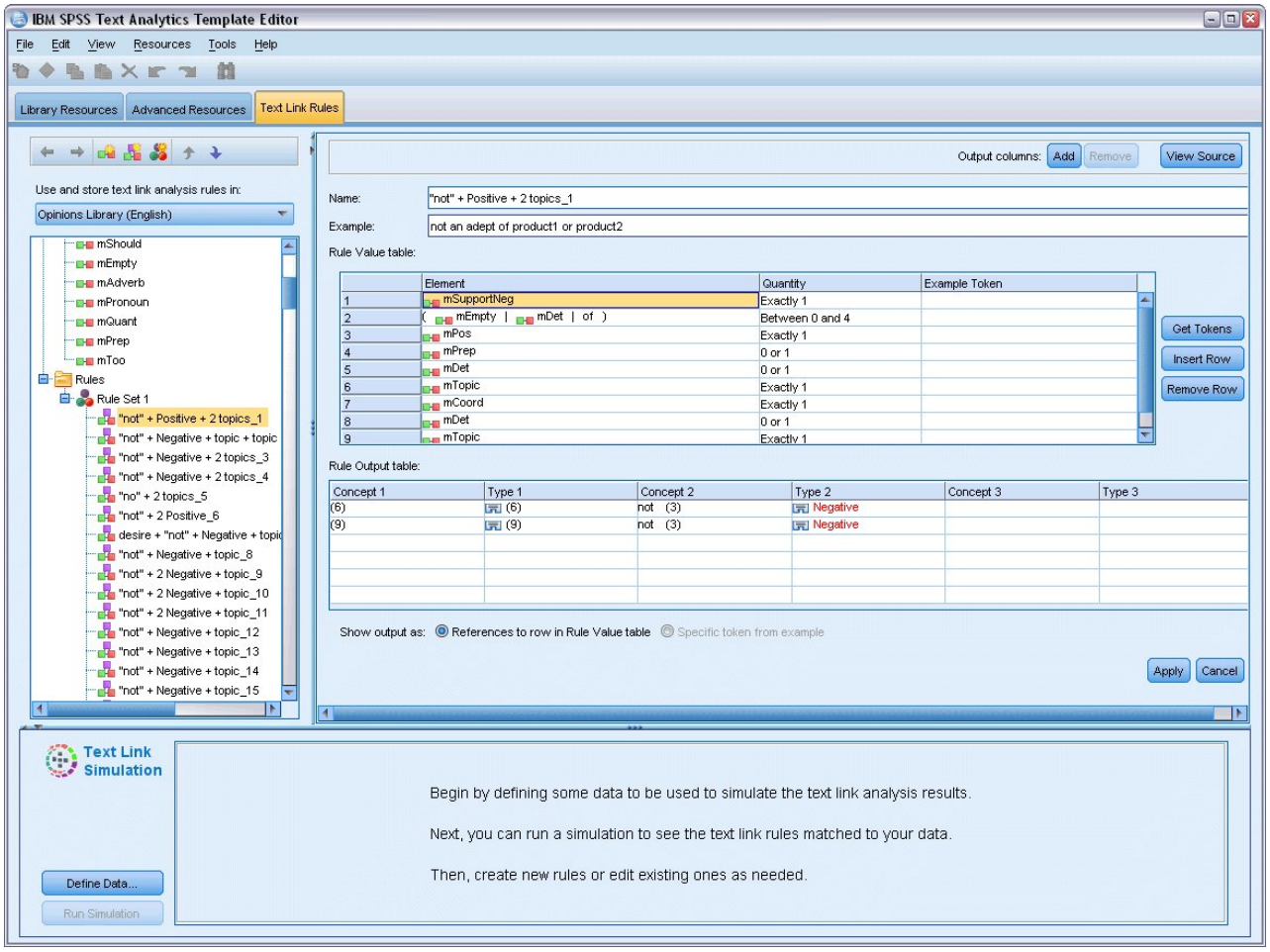


Figure 41. Onglet Règles des liens du texte

Où commencer

Il existe plusieurs façons de commencer à travailler dans l'éditeur de l'onglet Règles des liens du texte :

- Commencez par simuler des résultats avec un échantillon de texte et éditez ou créez des règles de correspondance en fonction de la façon dont l'ensemble de règles actuel extrait les motifs des données de simulation.
- Créez une nouvelle règle à partir de notes ou modifiez une règle existante.
- Travaillez directement dans la vue des sources.

Quand éditer ou créer des règles

Bien que les règles d'analyse des liens du texte fournies avec chaque modèle soient généralement adaptées à l'extraction de nombreuses relations de votre texte, qu'elles soient simples ou compliquées, il se peut que vous souhaitiez parfois modifier ces règles ou en créer de nouvelles. Par exemple :

- Pour capturer une idée ou une relation qui n'est pas extraite avec des règles existantes en créant une nouvelle règle ou une nouvelle macro.
- Pour modifier le comportement par défaut d'un type que vous avez ajouté aux ressources. Ceci nécessite généralement la modification d'une macro telle que `mTopic` ou `mNonLingEntities`. Pour plus d'informations, voir «Macros spéciales : mTopic, mNonLingEntities, SEP», à la page 221.

- Pour ajouter de nouveaux types aux macros et règles d'analyse des liens du texte existantes. Par exemple, si vous trouvez que le type <Organization> est trop large, vous pouvez créer de nouveaux types pour les organisations dans différents secteurs du marché tels que <Pharmaceuticals>, <Car Manufacturing>, <Finance>, etc. Dans ce cas, vous devez modifier les règles d'analyse des liens du texte et/ou créer une macro pour prendre en compte ces nouveaux types et les traiter en conséquence.
- Pour ajouter des types à une règle d'analyse des liens du texte existante. Par exemple, supposons que vous ayez une règle qui capture le texte suivant Pierre Martin appelle Martine Dupond, et vous voulez que cette règle, qui capture les communications téléphoniques, capture également les échanges par email. Vous pouvez ajouter à la règle le type d'entité non linguistique pour l'email, de sorte que le texte johndoe@ibm.com envoie un email à janedoe@ibm.com soit également capturé.
- Pour modifier légèrement une règle existante au lieu d'en créer une nouvelle. Par exemple, supposons que vous avez une règle correspondant au texte suivant xyz est très bien, et vous voulez que cette règle capture également xyz est très très bien.

Simulation des résultats d'analyse des liens du texte

Afin de définir plus facilement de nouvelles règles d'analyse des liens du texte ou de mieux comprendre la façon dont certaines phrases sont mises en correspondance pendant l'analyse des liens du texte, il est souvent utile de prendre un extrait de texte et d'effectuer une simulation. Pendant la simulation, une extraction est effectuée sur l'échantillon de données de simulation à l'aide de l'ensemble des ressources linguistiques actuelles et des paramètres d'extraction actuels. L'objectif est d'obtenir les résultats simulés et d'utiliser ces résultats pour améliorer vos règles, en créer de nouvelles ou mieux comprendre les mises en correspondance. Pour chaque extrait de texte (phrase, mot ou clause selon le contexte), le résultat de la simulation affiche l'ensemble des chaînes de caractères et présente les règles TLA ayant trouvé un motif dans ce texte. Une **chaîne de caractères**est définie comme tout mot ou toute phrase identifié/e pendant l'extraction.

Contrairement à d'autres ressources avancées, les règles TLA sont spécifiques à la bibliothèque ; c'est pourquoi vous pouvez uniquement utiliser les règles TLA d'une seule bibliothèque à la fois. A partir de l'Editeur de modèle ou de l' Editeur de ressources, accédez à l'onglet **Règles des liens du texte**. Dans cet onglet, vous pouvez spécifier la bibliothèque de votre modèle qui contient les règles TLA que vous souhaitez utiliser ou modifier. C'est pourquoi nous vous conseillons fortement de stocker toutes vos règles dans une seule bibliothèque, à moins que vous ne le souhaitiez pas pour une raison particulière.

Important ! Si vous utilisez un fichier de données, nous vous conseillons fortement de vous assurer que le texte qu'il contient est court afin de minimiser le temps de traitement. L'objectif d'une simulation est de voir comment une partie de texte est interprétée et de comprendre de quelle manière les règles correspondent à ce texte. Ces informations vont vous aider à rédiger et à modifier vos règles. Utilisez le noeud analyse des liens du texte ou exécutez un flux avec une session interactive, en ayant activé l'extraction TLA afin d'obtenir des résultats sur un ensemble plus complet de données. Cette simulation a pour seuls objectifs de tester et de créer des règles.

Définition des données pour la simulation

Pour obtenir un aperçu de la façon dont les règles peuvent correspondre au texte, vous pouvez effectuer une simulation à l'aide d'un échantillon de données. La première étape est de définir ces données.

Définition des données

1. Cliquez sur **Définir les données** dans la sous-fenêtre de simulation au bas de l'onglet **Règles des liens du texte**. Sinon, dans le cas où aucune donnée n'a été définie auparavant, choisissez **Outils** > **Exécuter la simulation** dans le menu. L'assistant Données de simulation s'ouvre.
2. Spécifiez le type de données en sélectionnant une des options suivantes :
   - **Coller ou saisir directement les données** Une zone de texte vous permet de coller du texte depuis le presse-papiers ou de saisir manuellement le texte à traiter. Vous pouvez entrer une phrase par

ligne ou utiliser la ponctuation (virgules ou points) pour séparer les phrases. Une fois le texte saisi, vous pouvez commencer la simulation en cliquant sur **Exécuter la simulation**.

- **Spécifier une source de données de fichier** Cette option indique que vous souhaitez utiliser un fichier qui contient du texte. Cliquez sur **Suivant** pour accéder à l'étape de l'assistant dans laquelle vous pouvez définir le fichier à traiter. Une fois le fichier sélectionné, vous pouvez commencer la simulation en cliquant sur **Exécuter la simulation**. Les types de fichier suivants sont pris en charge : `.txt` et `.text`. Le fichier de données choisi est lu 'tel quel' pendant la simulation. Le fichier entier est traité de la même façon que si vous aviez connecté un noeud Liste de fichiers à un noeud Text Mining.

Important : Si vous utilisez un fichier de données, nous vous conseillons fortement de vous assurer que le texte qu'il contient est court afin de minimiser le temps de traitement. L'objectif d'une simulation est de voir comment une partie de texte est interprétée et de comprendre de quelle manière les règles correspondent à ce texte. Ces informations vont vous aider à rédiger et à modifier vos règles. Utilisez le noeud analyse des liens du texte ou exécutez un flux avec une session interactive, en ayant activé l'extraction TLA afin d'obtenir des résultats sur un ensemble plus complet de données. Cette simulation a pour seuls objectifs de tester et de créer des règles.

3. Pour démarrer le processus de simulation, cliquez sur **Exécuter la simulation**. Une boîte de dialogue de progression apparaît. Si vous vous trouvez dans une session interactive, les paramètres d'extraction utilisés pendant la simulation sont ceux qui sont actuellement sélectionnés dans la session interactive (voir **Outils** > **Paramètres d'extraction** dans la vue Concepts et Catégories) Si vous vous trouvez dans l'Editeur de modèle, les paramètres d'extraction utilisés durant la simulation sont les paramètres d'extraction par défaut ; ce sont les mêmes que ceux affichés dans l'onglet Expert du noeud analyse des liens du texte. Pour plus d'informations, voir «Comprendre les résultats de la simulation».

Comprendre les résultats de la simulation

Pour obtenir un aperçu de la façon dont les règles peuvent correspondre au texte, vous pouvez effectuer une simulation à l'aide d'un échantillon de données et consulter les résultats. Ensuite, vous pouvez modifier votre ensemble de règles pour mieux l'adapter à vos données. Lorsque le processus d'extraction et de simulation est terminé, les résultats de la simulation s'afficheront.

Pour chaque “phrase” identifiée au cours de l'extraction, plusieurs informations vous sont fournies : la "phrase" exacte, la répartition des chaînes de caractères dans ce texte, et les règles de correspondance de texte de cette phrase. Par "**phrase**", nous entendons un mot, une phrase ou une clause en fonction de la façon dont l'extracteur a découpé le texte en parties lisibles.

Un **jeton** est défini comme tout mot ou toute expression identifié(e) pendant l'extraction. Par exemple, dans la phrase *Mon oncle vit à New York*, les jetons suivants peuvent être définis pendant l'extraction : *mon*, *oncle*, *vit*, *à* et *new york*. De plus, *oncle* peut être extrait comme concept et entré comme `<Unknown>`, et *new york* peut également être extrait comme concept et entré comme `<Location>`. Tous les concepts sont des jetons mais tous les jetons ne sont pas des concepts. Les jetons peuvent également être des macros, des chaînes littérales et des intervalles de mots. Seuls les mots et les expressions typés peuvent être des concepts.

Lorsque vous travaillez dans la session interactive ou dans l'éditeur de ressources, vous travaillez au niveau du concept. Les règles TLA sont plus précises et les chaînes de caractères individuelles d'une phrase peuvent s'utiliser dans la définition d'une règle si elles n'ont jamais été extraites ou dotées d'un type. Pouvoir utiliser des chaînes de caractères qui ne sont pas des concepts donne encore plus de flexibilité aux règles pour capturer des relations complexes dans votre texte.

Si vous avez plusieurs phrases dans vos données de simulation, vous pouvez avancer et reculer dans les résultats en cliquant sur **Suivant** et sur **Précédent**.

Dans ce genre de cas où une phrase ne correspond à aucune règle TLA dans la bibliothèque sélectionnée (voir le nom de la bibliothèque au-dessus de l'arborescence dans cet onglet), les résultats sont considérés

comme étant sans correspondance et les boutons **Prochains éléments sans correspondance** et **Eléments sans correspondance précédents** sont activés pour vous dire qu'il existe du texte pour lequel aucune règle n'a trouvé de correspondance et pour vous permettre d'accéder rapidement à ces instances.

Après avoir créé de nouvelles règles, avoir édité vos règles ou modifié vos ressources ou paramètres d'extraction, il est possible que vous souhaitiez effectuer une nouvelle simulation. Pour effectuer une nouvelle simulation, cliquez sur **Exécuter une simulation** dans la sous-fenêtre de simulation et les même données d'entrée seront réutilisées.

Les champs et tableaux suivants apparaissent dans les résultats de simulation :

Texte d'entrée. La phrase proprement dite, identifiée par le processus d'extraction se trouvant dans les données de simulation que vous avez définies dans l'assistant. Par phrase, nous entendons un mot, une phrase ou une clause en fonction de la façon dont l'extracteur a découpé le texte en parties lisibles.

Vue Système. Un ensemble de chaînes de caractères que le processus d'extraction a identifié.

- **Chaîne de caractères du texte d'entrée.** Chaque chaîne de caractères trouvée dans le texte d'entrée. Les chaînes de caractères ont été définies précédemment dans cette rubrique.
- **Entré comme.** Si une chaîne de caractères a été identifiée comme concept et dotée d'un type, alors le nom du type associé (tel que `<Unknown>`, `<Person>`, `<Location>`) apparaît dans cette colonne.
- **Macro correspondante.** Si une chaîne de caractères correspond à une macro existante, le nom de la macro associée apparaît dans cette colonne.

Règles mises en correspondance avec le texte d'entrée. Ce tableau présente les règles TLA mises en correspondance avec le texte d'entrée. Pour chaque règle mise en correspondance, vous voyez apparaître le nom de la règle dans la colonne **Sortie de règle** et les valeurs de sortie associées pour cette règle (Paires Concept + Type). Vous pouvez double-cliquer sur le nom de la règle correspondante pour ouvrir la règle dans la sous-fenêtre de l'éditeur au-dessus de la sous-fenêtre de simulation.

Bouton **Générer une règle**. Si vous cliquez sur ce bouton dans la sous-fenêtre de simulation, une nouvelle règle s'ouvre dans la sous-fenêtre de l'éditeur de règle au-dessus de la sous-fenêtre de simulation. Il utilise le texte d'entrée comme exemple. De même, toutes les chaînes de caractères dotées d'un type ou mises en correspondance avec une macro pendant la simulation sont automatiquement insérées en tant qu'éléments de colonne dans la **table Valeurs de règle**. Si une chaîne de caractères a été dotée d'un type *et* mise en correspondance avec une macro, la valeur de la macro est celle qui sera utilisée dans la règle afin de la simplifier. Par exemple, la phrase "*J'aime la pizza*" peut être tapée au cours de la simulation en tant que `<Unknown>` et mise en correspondance avec la macro `mTopic` si vous utilisez les ressources d'anglais de base. Dans ce cas, `mTopic` sera utilisée comme l'élément de la règle générée. Pour plus d'informations, voir «Utilisation des règles des liens du texte», à la page 221.

Navigation parmi les règles et les macros de l'arborescence

Lorsqu'une analyse des liens du texte est effectuée pendant l'extraction, les règles des liens du texte stockées dans la bibliothèque sélectionnée dans l'onglet **Règles des liens du texte** seront utilisées.

Contrairement à d'autres ressources avancées, les règles TLA sont spécifiques à la bibliothèque ; c'est pourquoi vous pouvez uniquement utiliser les règles TLA d'une seule bibliothèque à la fois. A partir de l'Editeur de modèle ou de l' Editeur de ressources, accédez à l'onglet **Règles des liens du texte**. Dans cet onglet, vous pouvez spécifier la bibliothèque de votre modèle qui contient les règles TLA que vous souhaitez utiliser ou modifier. C'est pourquoi nous vous conseillons vivement de stocker toutes vos règles dans une seule bibliothèque, à moins que vous ne le souhaitiez pas pour une raison particulière.

Vous pouvez spécifier dans quelle bibliothèque vous souhaitez travailler dans l'onglet Règles des liens du texte en sélectionnant cette bibliothèque dans la liste déroulante **Utiliser et stocker les règles d'analyse des liens du texte dans :** de cet onglet. Lorsqu'une analyse des liens du texte est effectuée pendant

l'extraction, les règles des liens du texte stockées dans la bibliothèque sélectionnée dans l'onglet **Règles des liens du texte** seront utilisées. Ainsi, si vous définissez des règles des liens du texte (règles TLA) dans plusieurs bibliothèques, seule la première bibliothèque dans laquelle les règles TLA sont trouvées sera utilisée pour l'analyse des liens du texte. C'est pourquoi nous vous conseillons fortement de stocker toutes vos règles dans une seule bibliothèque, à moins que vous ne le souhaitiez pas pour une raison particulière.

Lorsque vous sélectionnez une macro ou une règle dans l'arborescence, son contenu apparaît dans la sous-fenêtre de l'éditeur à droite. Si vous faites un clic droit sur un des éléments de l'arborescence, un menu contextuel vous présente les autres tâches possibles, comme :

- Créer une nouvelle macro dans l'arborescence et l'ouvrir dans l'éditeur à droite.
- Créer une nouvelle règle dans l'arborescence et l'ouvrir dans l'éditeur à droite.
- Créer un nouvel ensemble de règles dans l'arborescence.
- Couper, copier et coller des éléments pour simplifier l'édition.
- Supprimer des macros, des règles et des ensembles de règles pour les effacer des ressources.
- Désactiver des macros, des règles et des ensembles de règles pour indiquer qu'ils doivent être ignorés pendant le traitement.
- Faire monter ou descendre les règles d'un niveau pour changer l'ordre de traitement.

Avertissements dans l'arbre

Les avertissements apparaissent avec un triangle jaune dans l'arbre et vous informent d'un problème potentiel. Placez le curseur de la souris sur la macro ou la règle ayant un problème pour afficher l'explication dans l'info-bulle. Dans la plupart des cas, un message tel que **Avertissement : Pas d'exemple fourni. Entrez un exemple** s'affiche ; vous devez donc entrer un exemple.

S'il manque un exemple ou si l'exemple ne correspond à aucune règle, vous ne pourrez pas utiliser la fonction Obtenir des chaînes de caractères, par conséquent nous vous recommandons de n'entrer qu'un seul exemple par règle.

Lorsque la règle est surlignée en jaune, cela signifie qu'un type ou une macro est inconnu de l'éditeur TLA. Un message tel que **Avertissement : Type ou macro inconnu** s'affiche. Cela vous informe qu'un élément défini par `$quelquechose` dans la vue source (par exemple, `$myType`) ne correspond pas à un type existant dans voter bibliothèque ou n'est pas une macro.

Pour mettre à jour le vérificateur de syntaxe, vous devez passer à une autre règle ou macro ; il est inutile d'effectuer une recompilation. Donc, par exemple, si la règle `A` affiche un avertissement parce qu'il manque un exemple, vous devez entrer un exemple, cliquer sur une règle précédente ou suivante, puis revenir à la règle `A` pour vérifier que tout est correct.

Utilisation des macros

Les macros peuvent simplifier l'apparence des règles d'analyse des liens du texte puisqu'elles vous permettent de regrouper des types, d'autres macros et des chaînes (de mots) littérales à l'aide d'un opérateur `OR` (`|`). L'avantage de l'utilisation des macros est que non seulement vous pouvez réutiliser les macros dans différentes règles d'analyse des liens du texte pour les simplifier, mais cela vous permet également d'effectuer des mises à jour dans une macro au lieu de devoir effectuer des mises à jour à travers toutes vos règles d'analyse des liens du texte. La plupart des règles TLA fournies contiennent des macros prédéfinies. Les macros apparaissent au-dessus de l'arborescence dans la sous-fenêtre le plus à gauche de l'onglet Règles des liens du texte.

Les champs et tableaux suivants apparaissent dans les résultats de simulation :

Nom. Un nom unique identifiant cette macro. Nous vous conseillons d'ajouter le préfixe m minuscule au noms des macros pour les identifier rapidement dans vos règles. Lorsque vous faites manuellement référence aux macros dans les règles (par modification d'une ligne ou dans la vue des sources), vous devez utiliser le préfixe $ afin que le processus d'extraction puisse rechercher ce nom spécifique. Mais si vous déplacez et collez le nom de la macro ou l'ajoutez avec les menus contextuels, le produit le reconnaît immédiatement comme macro et aucun $ n'est ajouté.

Table **Valeur de macro**.

- Plusieurs lignes représentant toutes les valeurs possibles que cette macro peut représenter. Ces valeurs sont sensibles à la casse.
- Ces valeurs peuvent comprendre un type, une chaîne littérale, un intervalle de mots ou une macro ou une combinaison de ces formes. Pour plus d'informations, voir «Eléments pris en charge pour les règles et les macros», à la page 228.
- Pour entrer une valeur pour un élément de macro, cliquez deux fois sur la ligne dans laquelle vous voulez travailler. Une boîte de texte éditable apparaît, dans laquelle vous pouvez entrer une référence de type, une référence de macro, une chaîne littérale ou un intervalle de mots. Sinon, cliquez avec le bouton droit sur la cellule pour afficher un menu contextuel proposant des listes de macros communes, des noms de types et des noms de types non linguistiques. Pour référencer un type ou une macro, faites précéder son nom par le caractère '$' (par exemple, entrez `$mTopic` pour la macro `mTopic`). Lors de la combinaison d'arguments, vous devez utiliser des parenthèses `( )` pour regrouper les arguments et le caractère `|` pour indiquer un booléen `OR`.
- Vous pouvez ajouter ou supprimer des lignes dans la table Valeur de macro à l'aide des boutons à sa droite.
- Saisissez chaque élément dans sa propre ligne. Par exemple, si vous souhaitez créer une macro qui représente une des 3 chaînes littérales comme `suis OR étais OR est`, vous devez saisir chaque chaîne littérale dans une ligne distincte de la vue et votre tableau Macro doit contenir 3 lignes.

Création et édition de macros

Vous pouvez créer de nouvelles macros ou éditer les macros existantes. Suivez les conseils et descriptions de l'éditeur de macro. Pour plus d'informations, voir «Utilisation des macros», à la page 218.

Création de macros

1. A partir des menus, sélectionnez **Outils> Nouvelle macro**. Vous pouvez également cliquer sur l'icône Nouvelle macro dans la barre d'outils en arborescence pour ouvrir une nouvelle macro dans l'éditeur.
2. Saisissez un nom unique et définissez les éléments de valeur de la macro.
3. Cliquez sur **Appliquer** lorsque vous avez terminé de rechercher les erreurs.

Edition des macros

1. Cliquez sur le nom de la macro dans l'arborescence. La macro s'ouvre dans la sous-fenêtre de l'éditeur à droite.
2. Apportez vos modifications.
3. Cliquez sur **Appliquer** lorsque vous avez terminé de rechercher les erreurs.

Désactivation et suppression des macros

Désactivation de macros

Si vous souhaitez qu'une macro soit ignorée pendant le processus, vous pouvez la désactiver. Cette action peut provoquer des avertissements et des erreurs dans les règles qui référencent encore cette macro désactivée. Soyez prudent lors de la suppression et de la désactivation des macros.

1. Cliquez sur le nom de la macro dans l'arborescence. La macro s'ouvre dans la sous-fenêtre de l'éditeur à droite.
2. Cliquez avec le bouton droit de la souris sur le nom.

3. Dans les menus contextuels, choisissez **Désactiver**. L'icône de la macro se grise et la macro ne peut plus être modifiée.

Suppression de macros

Si vous souhaitez effacer une macro, vous pouvez la supprimer. Cette action peut provoquer des erreurs dans les règles qui référencent encore cette macro. Soyez prudent lors de la suppression et de la désactivation des macros.

1. Cliquez sur le nom de la macro dans l'arborescence. La macro s'ouvre dans la sous-fenêtre de l'éditeur à droite.
2. Cliquez avec le bouton droit de la souris sur le nom.
3. Dans les menus contextuels, choisissez **Supprimer**. La macro disparaît de la liste.

Vérification des erreurs, enregistrement et annulation

Application des modifications de macro

Si vous cliquez en-dehors de l'éditeur de macro ou si vous cliquez sur **Appliquer**, la macro est automatiquement analysée pour savoir s'il y a des erreurs. Si une erreur est trouvée, vous devez la résoudre avant de passer à une autre partie de l'application.

Mais si des erreurs moins graves sont détectées, un simple avertissement est donné. Par exemple, si votre macro contient des définitions incomplètes ou non référencées de types ou d'autres macros, un avertissement apparaît. Lorsque vous cliquez sur **Appliquer**, tout avertissement non corrigé fait apparaître une icône d'avertissement à gauche du nom de la macro dans l'arborescence des règles et des macros dans la sous-fenêtre de gauche.

Appliquer une macro ne signifie pas que votre macro est enregistrée définitivement. Lorsque vous appliquez une macro, le processus de validation recherche les erreurs et les avertissements.

Enregistrement des ressources à l'intérieur d'une session de plan de travail interactif

1. Pour enregistrer les modifications effectuées sur vos ressources pendant une session de plan de travail interactif et pouvoir les utiliser à la prochaine utilisation de votre flux, vous devez :
   - mettre à jour votre noeud modélisation pour vous assurer d'obtenir les mêmes ressources à la prochaine exécution de votre flux. Pour plus d'informations, voir «Mise à jour des noeuds modélisation et enregistrement», à la page 78. Vous devez ensuite enregistrer votre flux. Pour ce faire, rendez-vous dans la fenêtre principale de IBM SPSS Modeler après la mise à jour du noeud modélisation.
2. Pour enregistrer les modifications effectuées sur vos ressources pendant une session de plan de travail interactif et pouvoir les utiliser dans un autre flux, vous devez :

- mettre à jour le modèle utilisé ou en créer un nouveau. Pour plus d'informations, voir «Création et mise à jour de modèles», à la page 164. Les modifications apportées au noeud en cours ne seront pas enregistrées (voir l'étape précédente)
- Ou mettre à jour le TAP utilisé. Pour plus d'informations, voir «Mise à jour des Packs d'analyse de texte», à la page 137.

Enregistrement des ressources dans l'Editeur de modèle

1. D'abord, publiez la bibliothèque. Pour plus d'informations, voir «Publication de bibliothèques», à la page 183.
2. Puis enregistrez le modèle avec **Fichier > Enregistrer un modèle de ressources** dans les menus.

Annuler les modifications des macros

1. Si vous voulez ignorer les modifications, cliquez sur **Annuler**.

Macros spéciales : mTopic, mNonLingEntities, SEP

Le modèle Opinions (et les modèles similaires) de même que les modèles Ressources de base sont fournis avec deux macros spéciales nommées `mTopic` et `mNonLingEntities`.

mTopic

Par défaut, la macro `mTopic` regroupe tous le types fournis dans le modèle prévu pour être en connexion avec une opinion, tels que les types de bibliothèque *principale* suivants : `<Person>`, `<Organization>`, `<Location>`, etc., tant que le type n'est pas un type d'opinion (par exemple, `<Negative>` ou `<Positive>`) ou un type défini comme entité non linguistique dans les ressources avancées.

Lorsque vous créez un nouveau type dans un modèle Opinions (ou un modèle similaire), le produit suppose que sauf si ce type est spécifié dans une autre macro ou dans la section des entités non linguistiques de l'onglet Ressources avancées, il sera traité de la même façon que les autres types définis dans la macro `mTopic`.

Considérons que vous créez deux types dans les ressources à partir du modèle Opinions : `<Vegetables>` (Légumes) et `<Fruit>` (Fruits). Sans avoir à effectuer aucune modification, vos nouveaux types sont traités comme des types `mTopic` pour que vous puissiez automatiquement connaître les opinions positives, négatives, neutres et contextuelles sur vos nouveaux types. Au cours de l'extraction, par exemple, la phrase "*J'aime les brocolis mais je déteste le pamplemousse*" produirait les 2 motifs de sortie suivants :

```
broccoli <Vegetables> + aime <Positive>

grapefruit <Fruit> + déteste <Negative>
```

Mais, si vous souhaitez exécuter ces types autrement que pour les autres types dans mTopic, vous pouvez ajouter le nom du type dans une macro existante comme `mPos`, qui regroupe tous les types d'opinions positives, ou créer une nouvelle macro que vous pourrez ensuite référencer dans une ou plusieurs règles.

Important ! Si vous créez un nouveau type tel que `<Vegetables>`, ce nouveau type sera inclus en tant que type dans mTopic. Cependant, ce nom de type ne sera pas explicitement visible dans la définition de la macro.

mNonLingEntities

De même, si vous ajoutez de nouvelles entités non linguistiques dans la section **Entités non linguistiques** de l'onglet Ressources avancées, elles seront automatiquement traitées comme `mNonLingEntities` sauf spécification contraire. Pour plus d'informations, voir «Entités non linguistiques», à la page 204.

SEP

Vous pouvez également utiliser la macro prédéfinie `SEP`, qui correspond au séparateur global défini sur l'ordinateur local et qui est en général un virgule (,).

Utilisation des règles des liens du texte

Une règle d'analyse des liens du texte est une requête booléenne utilisée pour réaliser une mise en correspondance sur une phrase. Les règles d'analyse des liens du texte contiennent les arguments suivants : types, macros, chaînes littérales ou intervalle de mots. Vous devez avoir au moins une règle d'analyse des liens du texte pour extraire les résultats TLA.

Les zones et champs suivants apparaissent dans l'éditeur de l'onglet Règles des liens du texte :

Champ **Nom**. Le nom unique de la règle des liens du texte.

Champ **Exemple**. Vous pouvez éventuellement inclure un exemple de phrase ou de séquence de mots qui serait capturée par cette règle. Nous vous conseillons d'utiliser des exemples. Dans cet éditeur, vous pourrez générer des chaînes de caractères à partir de cet exemple de texte pour savoir quelle est sa correspondance avec la règle et quels en seront les résultats. Un **jeton** est défini comme tout mot ou toute expression identifié(e) pendant l'extraction. Par exemple, dans la phrase *Mon oncle vit à New York*, les jetons suivants peuvent être définis pendant l'extraction : *mon*, *oncle*, *vit*, *à* et *new york*. De plus, *oncle* peut être extrait comme concept et entré comme `<Unknown>`, et *new york* peut également être extrait comme concept et entré comme `<Location>`. Tous les concepts sont des jetons mais tous les jetons ne sont pas des concepts. Les jetons peuvent également être des macros, des chaînes littérales et des intervalles de mots. Seuls les mots et les expressions typés peuvent être des concepts.

Table Valeur de règle. Cette table contient les éléments de la règle qui sont utilisés pour faire correspondre une règle à une phrase. Vous pouvez ajouter ou supprimer des lignes de la table à l'aide des boutons à sa droite. La table est composée de 3 colonnes :

- La colonne **Elément**. Saisissez des valeurs sous la forme d'un type, d'une chaîne littérale, d'un intervalle de mots (`<Toutes les chaînes de caractères>`) ou d'une macro ou d'une combinaison de ces formes. Pour plus d'informations, voir «Eléments pris en charge pour les règles et les macros», à la page 228. Cliquez deux fois sur la cellule de l'élément pour entrer directement l'information. Sinon, cliquez avec le bouton droit sur la cellule pour afficher un menu contextuel proposant des listes de macros communes, des noms de types et des noms de types non linguistiques. Si vous entrez directement l'information dans la cellule, n'oubliez pas de faire précéder le nom de type ou de macro par le caractère '`$`' (par exemple, entrez `$mTopic` pour la macro `mTopic`). L'ordre dans lequel vous créez vos lignes d'élément joue un rôle critique pour la façon dont la règle sera mise en correspondance avec le texte. Lors de la combinaison d'arguments, vous devez utiliser des parenthèses `( )` pour regrouper les arguments et le caractère `|` pour indiquer un booléen `OR`. N'oubliez pas que les valeurs distinguent les majuscules des minuscules.
- La colonne **Quantité**. Elle indique le nombre d'occurrences minimal et maximal d'un élément présent dans un texte pour effectuer une correspondance. Par exemple, si vous souhaitez définir un intervalle ou une série de mots, entre deux autres éléments de 0 à 3 mots, vous pouvez choisir **Entre 0 et 3** dans la liste ou entrer directement les nombres dans la boîte de dialogue. La valeur par défaut est '**Exactement 1**'. Il est possible que dans certains cas vous souhaitiez rendre un élément optionnel. Si c'est le cas, il aura alors une quantité minimale de 0 et une quantité maximale supérieure à 0 (c'est-à-dire 0 ou 1, entre 0 et 2). Veuillez noter que le premier élément d'une règle ne peut pas être optionnel, c'est-à-dire qu'il ne peut pas avoir une quantité de 0.
- La colonne **Exemple de chaîne de caractères**. Si vous cliquez sur **Obtenir des chaînes de caractères**, le programme découpe l'**Exemple** de texte en chaînes de caractères et utilise ces dernières pour remplir cette colonne avec les chaînes de caractères qui correspondent aux éléments définis. Si vous le désirez, vous pouvez également voir ces chaînes de caractères dans la table de sortie.

Table de sortie de règle Chaque ligne de cette table définit la façon dont la sortie des motifs TLA apparaîtra dans les résultats. La sortie de règle peut produire des motifs contenant jusqu'à 6 paires de colonnes Concept/Type, chacune représentant une *propriété*. Par exemple, le motif de type `<Location> + <Positive>` est un motif à deux propriétés c'est-à-dire qu'il est composé de deux paires de colonnes Concept/Type.

Remarque : Les termes de la colonne **Elément** de la **Table de sortie de règle** ou de toute colonne **Concept** de la **Table de sortie de règle** ne peuvent pas commencer par l'un des caractères suivants : `, #, %, ^, *, _, -, :, <, >, /, \ ou ".

Alors que le langage nous donne la liberté d'exprimer les mêmes idées de différentes façons, vous pouvez définir plusieurs règles pour capturer la même idée. Par exemple, le texte « *Paris est un endroit que j'aime* » et le texte « *J'aime vraiment Paris et Florence* » représentent la même idée de base (qu'on aime Paris) mais exprimée de différentes façons. Il faudrait donc deux règles différentes pour les capturer tous

les deux. Mais il est plus simple d'utiliser les résultats de motif si les mêmes idées sont regroupées. C'est pourquoi, tout en ayant deux règles différentes pour capturer ces deux phrases, vous pouvez définir la même sortie pour ces deux règles, par exemple, le motif de type `<Emplacement> + <Positif>`, de façon à ce qu'elle représente les deux textes. Vous pouvez alors voir que les résultats ne suivent pas toujours la structure ou l'ordre des mots trouvés dans le texte d'origine. De plus, ce type de motif peut correspondre à d'autres expressions et produire des motifs de concepts tels que `paris + aime` et `tokyo + aime`.

Pour vous aider à définir plus rapidement les résultats en faisant moins d'erreurs, vous pouvez utiliser le menu contextuel pour choisir l'élément que vous voulez voir apparaître dans les résultats. Vous pouvez également faire glisser et déposer des éléments du tableau Valeur de règle dans les résultats. Par exemple, si vous avez une règle qui contient une référence à la macro `mTopic` dans la ligne 2 du tableau Valeur de règle et que vous souhaitez que cette valeur fasse partie de vos résultats, vous pouvez simplement faire glisser l'élément de `mTopic` et le déposer dans la première paire de colonnes de la table Sortie de règle. Vous remplirez ainsi le concept et le type pour la paire sélectionnée. Ou si vous souhaitez que les résultats commencent par le type défini par le troisième élément (ligne 3) de la table de valeur de règle, faites glisser ce type de la table Valeur de règle jusqu'à la cellule **Type 1** de la table de sortie. La table sera mise à jour et affichera la référence de la ligne entre parenthèses (3).

Vous pouvez également saisir ces références manuellement dans le tableau en double-cliquant sur la cellule de chaque colonne **Concept** devant faire partie des résultats et en entrant le symbole $ suivi du numéro de la ligne, comme `$2` pour faire référence à l'élément défini dans la ligne 2 du tableau Valeur de règle. Lorsque vous entrez manuellement les informations, vous devez également définir la colonne **Type**, entrer le symbole # suivi du numéro de ligne, comme `#2` pour faire référence à l'élément défini dans la ligne 2 de la table Valeur de règle.

De plus, vous pouvez même combiner les méthodes. Imaginons le type `<Positive>` dans la ligne 4 de votre table Valeur de règle. Vous pouvez le faire glisser et le déposer dans la colonne `Type 2` puis double-cliquer sur la cellule dans la colonne `Concept 2` pour entrer ensuite manuellement le mot « *pas* » devant elle. Le titre de la colonne des résultats serait alors `pas (4)` dans le tableau, ou si vous étiez en mode édition ou en mode source `pas $4`. Vous pouvez ensuite faire un clic droit dans la colonne Type 1 et sélectionner, par exemple, la macro appelée `mTopic`. Il pourrait donc en résulter un motif de concept `voiture + mauvais`.

La majorité des règles n'ont qu'une seule ligne de résultats mais parfois, plusieurs résultats sont possibles et souhaités. Dans ce cas, définissez un résultat par ligne dans la table Sortie de règle.

Important : Gardez à l'esprit que les autres opérations de traitement linguistique sont exécutées lors de l'extraction des motifs TLA. Par conséquent, lorsque la sortie lit `t$3\t#3`, cela signifie que le motif va afficher le concept final et le type final pour le troisième élément après l'application de tous les traitements linguistiques (synonymes et autres regroupements).

- **Afficher la sortie comme**. Par défaut, l'option **Références à la ligne dans la table de valeur de règle** est sélectionnée et les résultats utilisent les références numériques de la ligne comme défini dans l'onglet Valeur de règle. Si vous avez auparavant cliqué sur Obtenir des chaînes de caractères et que vous en avez dans la colonne Exemple de chaînes de caractères de la table Valeur de règle, vous pouvez choisir d'afficher les résultats de ces chaînes de caractères spécifiques en choisissant cette option.

Remarque : Si la table de sortie ne contient pas suffisamment de paires concept/type, vous pouvez en ajouter par l'intermédiaire du bouton Ajouter de la barre d'outils de l'éditeur. Si 3 paires apparaissent et que vous cliquez sur Ajouter, 2 colonnes supplémentaires (Concept 4 et Type 4) sont ajoutées à la table. Cela signifie que vous verrez désormais 4 paires dans la table de sortie pour toutes les règles. Vous pouvez également supprimer les paires non utilisées tant qu'aucune autre règle dans l'ensemble de règles de cette bibliothèque n'utilise cette paire.

Exemple de règle

Supposons que vos ressources contiennent la règle d'analyse des liens du texte TLA suivante et que vous avez activé l'extraction des résultats TLA :

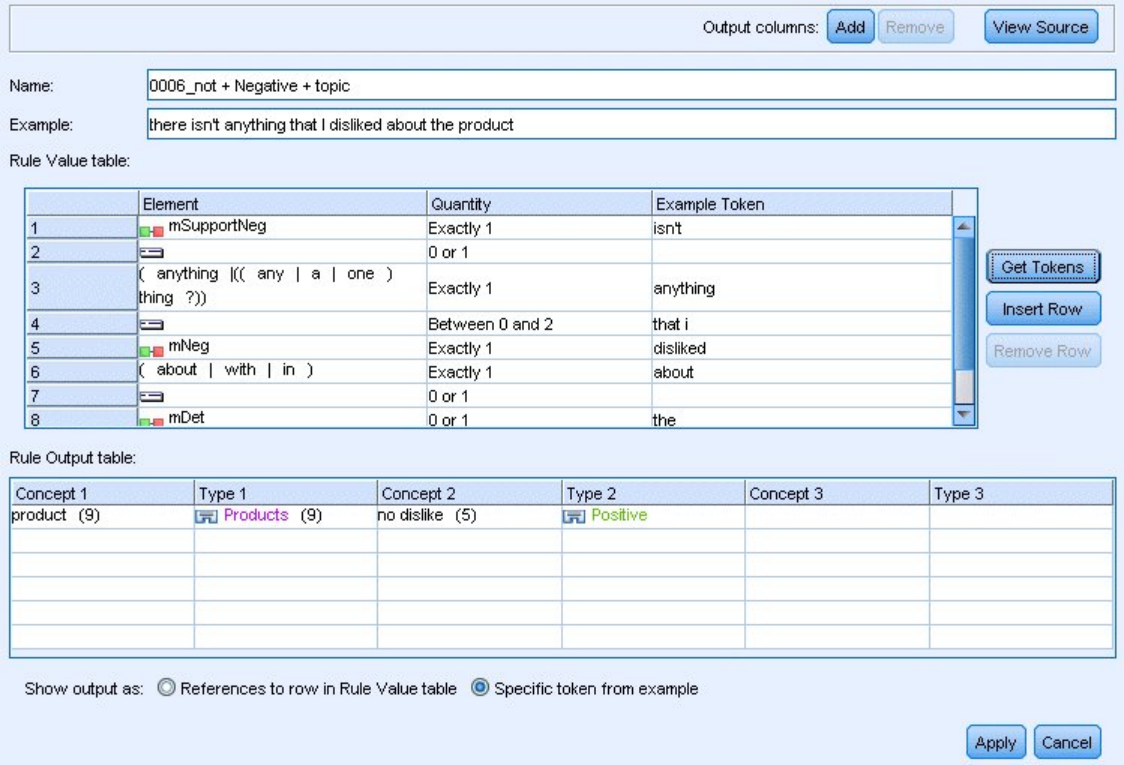


Figure 42. Onglet Règles des liens du texte : Editeur de règles

Lors de l'extraction, le moteur du programme d'extraction lit chaque phrase et essaie de faire correspondre la séquence suivante :

Tableau 44. Exemple de séquence d'extraction

Elément (ligne)	Description des arguments
1	Le concept d'un des types représentés par les macros `mPos` ou `mNeg` ou du type `<Uncertain>`.
2	Un concept entré comme un des types représentés par la macro `mTopic`.
3	Un des mots représentés par la macro `mBe`.
4	Un élément optionnel, 0 ou 1 mot également, également appelé un intervalle de mots ou `<Toutes les chaînes de caractères>`
5	Un concept entré comme un des types représentés par la macro `mTopic`.

La table de sortie montre que tout ce qui est attendu de cette règle est un motif où tout concept ou type correspond à la macro `mTopic` définie dans la ligne 5 de la **table Valeur de règle** + tout concept ou type correspond à la macro `mPos`, `mNeg`, ou `<Uncertain>`, comme défini dans la ligne 1 de la **table Valeur de règle**. Cela peut être `saucisse + aimer` ou `<Unknown> + <Positive>`.

Création et édition des règles

Vous pouvez créer de nouvelles règles ou éditer les règles existantes. Suivez les conseils et descriptions de l'éditeur de règle. Pour plus d'informations, voir «Utilisation des règles des liens du texte», à la page 221.

Création de nouvelles règles

1. Dans les menus, sélectionnez **Outils > Nouvelle règle**. Vous pouvez également cliquer sur l'icône Nouvelle règle dans la barre d'outils en arborescence pour ouvrir une nouvelle règle dans l'éditeur.
2. Saisissez un nom unique et définissez les éléments de valeur de la règle.
3. Cliquez sur **Appliquer** lorsque vous avez terminé de rechercher les erreurs.

Edition de règles

1. Cliquez sur le nom de la règle dans l'arborescence. La règle s'ouvre dans la sous-fenêtre de l'éditeur à droite.
2. Apportez vos modifications.
3. Cliquez sur **Appliquer** lorsque vous avez terminé de rechercher les erreurs.

Désactivation et suppression des règles

Désactivation des règles

Si vous souhaitez qu'une règle soit ignorée pendant le processus, vous pouvez la désactiver. Soyez prudent lors de la suppression et de la désactivation des règles.

1. Cliquez sur le nom de la règle dans l'arborescence. La règle s'ouvre dans la sous-fenêtre de l'éditeur à droite.
2. Cliquez avec le bouton droit de la souris sur le nom.
3. Dans les menus contextuels, choisissez **Désactiver**. L'icône de la règle se grise et la règle ne peut plus être modifiée.

Suppression de règles

Si vous souhaitez effacer une règle, vous pouvez la supprimer. Soyez prudent lors de la suppression et de la désactivation des règles.

1. Cliquez sur le nom de la règle dans l'arborescence. La règle s'ouvre dans la sous-fenêtre de l'éditeur à droite.
2. Cliquez avec le bouton droit de la souris sur le nom.
3. Dans les menus contextuels, choisissez **Supprimer**. La règle disparaît de la liste.

Vérification des erreurs, enregistrement et annulation

Application des modifications des règles

Si vous cliquez en-dehors de l'éditeur de règle ou si vous cliquez sur **Appliquer**, la règle est automatiquement analysée pour savoir s'il y a des erreurs. Si une erreur est trouvée, vous devez la résoudre avant de passer à une autre partie de l'application.

Mais si des erreurs moins graves sont détectées, un simple avertissement est donné. Par exemple, si votre règle contient des définitions incomplètes ou non référencées de types ou de macros, un avertissement apparaît. Lorsque vous cliquez sur **Appliquer**, tout avertissement non corrigé fait apparaître une icône d'avertissement à gauche du nom de la règle dans l'arborescence de la sous-fenêtre de gauche.

Appliquer une règle ne signifie pas que votre règle est enregistrée de manière définitive. Lorsque vous appliquez une macro, le processus de validation recherche les erreurs et les avertissements.

Enregistrement des ressources à l'intérieur d'une session de plan de travail interactif

1. Pour enregistrer les modifications effectuées sur vos ressources pendant une session de plan de travail interactif et pouvoir les utiliser à la prochaine utilisation de votre flux, vous devez :
   - mettre à jour votre noeud modélisation pour vous assurer d'obtenir les mêmes ressources à la prochaine exécution de votre flux. Pour plus d'informations, voir «Mise à jour des noeuds modélisation et enregistrement», à la page 78. Vous devez ensuite enregistrer votre flux. Pour ce faire, rendez-vous dans la fenêtre principale de IBM SPSS Modeler après la mise à jour du noeud modélisation.
2. Pour enregistrer les modifications effectuées sur vos ressources pendant une session de plan de travail interactif et pouvoir les utiliser dans un autre flux, vous devez :
   - mettre à jour le modèle utilisé ou en créer un nouveau. Pour plus d'informations, voir «Création et mise à jour de modèles», à la page 164. Les modifications apportées au noeud en cours ne seront pas enregistrées (voir l'étape précédente)
   - Ou mettre à jour le TAP utilisé. Pour plus d'informations, voir «Mise à jour des Packs d'analyse de texte», à la page 137.

Enregistrement des ressources dans l'Editeur de modèle

1. D'abord, publiez la bibliothèque. Pour plus d'informations, voir «Publication de bibliothèques», à la page 183.
2. Puis enregistrez le modèle avec **Fichier > Enregistrer un modèle de ressources** dans les menus.

Annulation des modifications de règles

1. Si vous voulez ignorer les modifications, cliquez sur **Annuler** dans la sous-fenêtre de l'éditeur.

Ordre de traitement des règles

Lorsque l'analyse des liens du texte est effectuée pendant l'extraction, une « phrase » (clause, mot, expression), sera mise en correspondance avec chaque règle l'une après l'autre, jusqu'à ce qu'une correspondance soit trouvée, ou que toutes les règles aient été épuisées. La position au sein de l'arborescence indique l'ordre dans lequel les règles sont testées. La meilleure façon de faire est d'ordonner vos règles de la plus spécifique à la plus générique. Les règles les plus spécifiques doivent se trouver en haut de l'arborescence. Pour modifier l'ordre d'une règle ou d'un ensemble de règles spécifique, sélectionnez **Monter d'un niveau** ou **Descendre d'un niveau** dans le menu contextuel Arborescence des règles et des macros ou avec les flèches vers le haut et vers le bas de la barre d'outils.

Si vous vous trouvez *dans la vue des sources*, vous ne pouvez pas modifier l'ordre des règles en les déplaçant dans l'éditeur. Plus la règle est haute dans la vue des sources, plus vite elle sera traitée. Nous vous conseillons vivement de réordonner les règles dans l'arborescence uniquement pour éviter les problèmes de copier/coller.

Important ! Dans les versions précédentes de IBM SPSS Modeler Text Analytics, vous deviez avoir un ID de règle numérique unique. A partir de la version 18.1.1, vous pouvez indiquer l'ordre de traitement en déplaçant la règle vers le haut ou vers le bas dans l'arborescence, ou en modifiant sa position dans la vue des sources.

Par exemple, supposons que votre texte contienne les deux phrases suivantes :

J'aime les anchois

J'aime les anchois et les poivrons verts

Supposons également que les deux règles d'analyse des liens du texte suivantes soient définies :

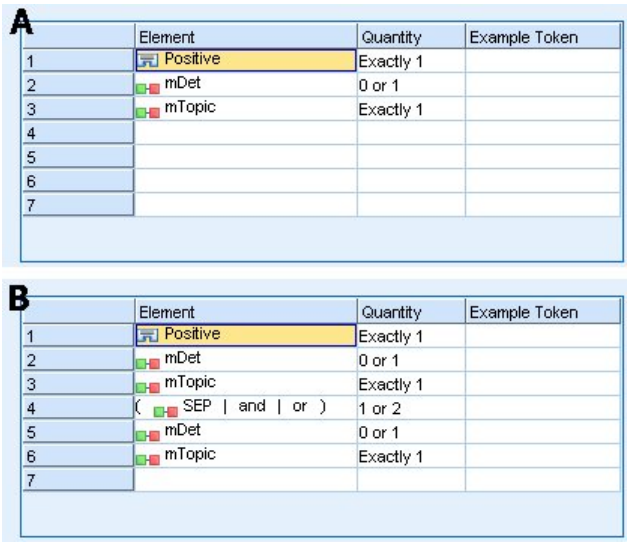


Figure 43. 2 exemples de règles

Dans la vue source, les valeurs de règles peuvent ressembler à ceci :

A : `value = $Positive $mDet? $mTopic`

B : `value = $Positive $mDet? $mTopic ($SEP|and|or){1,2} $mDet? $mTopic`

Si la règle **A** est plus haute dans l'arborescence (plus près du haut) que la règle **B**, alors **A** sera d'abord traitée et la phrase *J'aime les anchois et les poivrons verts* sera d'abord mise en correspondance avec `$Positive $mDet? $mTopic`, et produira une sortie de motif incomplète (`anchois + aimer`) car elle aura été mise en correspondance avec une règle qui ne recherchait pas 2 correspondances `$mTopic`.

Ainsi, pour capturer l'essence réelle du texte, la valeur la plus spécifique, dans ce cas **B** doit être placée plus haut dans l'arborescence que la valeur plus générique, dans ce cas la règle **A**.

Utilisation d'ensembles de règles (traitement en plusieurs étapes)

Un ensemble de règles permet de regrouper logiquement un ensemble de règles dans l'arborescence Règles et macros afin de pouvoir effectuer un traitement en plusieurs étapes. Un ensemble de règles n'a pas de définition en soi autre qu'un nom et permet d'organiser vos règles dans des groupes de manière logique. Dans certains contextes, le texte est trop riche et varié pour être traité en une seule fois. Par exemple, si vous travailler avec des données liées à la sécurité intérieure, le texte peut contenir des liens entre des individus déterminés à partir de contacts (*x a appelé y*), de relations familiales (*x est le beau-frère de y*), d'échange de sommes d'argent (*x a prêté 100 dollars à y*), etc. Dans ce cas, il est utile de créer des ensembles spécialisés de règles d'analyse des liens du texte, dont chacun est ciblé sur un certain type de relation comme celui destiné pour découvrir les contacts, un autre pour découvrir les membres de la famille, etc.

Pour créer un ensemble de règles, sélectionnez “Créer un jeu de règles” dans le menu contextuel Arborescence des règles et des macros ou à partir de la barre d'outils. Vous pouvez ensuite créer de nouvelles règles directement dans un noeud Ensemble de règles de l'arborescence ou déplacer les règles existantes vers un ensemble de règles.

Lorsque vous effectuez une extraction à l'aide de ressources dans lesquelles des règles sont regroupées dans des ensembles de règles, le moteur d'extraction est forcé d'effectuer plusieurs analyses du texte afin de faire correspondre différents motifs dans chaque analyse. De cette façon, une « phrase » peut être

mise en correspondance avec une règle de chaque ensemble de règles, alors que sans ensemble de règles, elle ne peut être mise en correspondance qu'avec une seule règle.

Remarque : vous pouvez ajouter jusqu'à 512 règles par ensemble de règles.

Création de nouveaux ensembles de règles

1. Dans les menus, sélectionnez **Outils > Nouvel ensemble de règles**. Vous pouvez également cliquer sur l'icône Nouvel ensemble de règles dans la barre d'outils de l'arborescence. Un ensemble de règles apparaît dans l'arborescence des règles.
2. Ajoutez ces nouvelles règles à cet ensemble de règles ou déplacez les règles existantes dans l'ensemble.

Désactivation des ensembles de règles

1. Faites un clic droit sur le nom de l'ensemble de règles dans l'arborescence.
2. Dans les menus contextuels, choisissez **Désactiver**. L'icône de l'ensemble de règles se grise et toutes les règles de cet ensemble de règles sont également désactivées et ignorées pendant le traitement.

Supprimer des ensembles de règles

1. Faites un clic droit sur le nom de l'ensemble de règles dans l'arborescence.
2. Dans les menus contextuels, choisissez **Supprimer**. L'ensemble de règles et toutes les règles qu'il contient sont supprimées des ressources.

Eléments pris en charge pour les règles et les macros

Les arguments suivants sont acceptés pour les paramètres de valeur dans les règles d'analyse des liens du texte et les macros :

Macros

Vous pouvez utiliser une macro directement dans une règle d'analyse des liens du texte ou dans une autre macro. Si vous saisissez manuellement le nom d'une macro ou depuis une vue source (en opposition à la sélection du nom d'une macro dans un menu contextuel), vérifiez que le nom comporte un préfixe sous la forme du caractère dollar ($), comme `$mTopic`. Le nom de macro distingue les majuscules des minuscules. Lorsque vous sélectionnez des macros à partir des menus contextuels, vous pouvez choisir parmi toutes les macros définies dans l'onglet Règles des liens du texte en cours.

Types

Vous pouvez utiliser un type directement dans une règle d'analyse des liens du texte ou une macro. Si vous saisissez manuellement le nom d'un type ou depuis une vue des sources (en opposition à la sélection d'un type dans un menu contextuel), vérifiez que le nom du type comporte un préfixe sous la forme du caractère dollar ($), comme `$Person`. Le nom de type distingue les majuscules des minuscules. Si vous utilisez les menus contextuels, vous pouvez choisir parmi tous les types de l'ensemble de ressources utilisé.

Si vous référencez un type non reconnu, vous recevrez un message d'avertissement et la règle aura une icône d'avertissement dans l'arborescence Règles et macros jusqu'à ce que vous le corrigiez.

Chaînes littérales

Pour inclure des informations qui n'ont jamais été extraites, vous pouvez définir une chaîne littérale que le moteur du programme d'extraction va rechercher. Tous les mots ou expressions extraits ont été attribués à un type, c'est pourquoi ils ne peuvent pas être utilisés dans les chaînes littérales. Si vous utilisez un mot qui a été extrait, il sera ignoré, même si son type est `<Unknown>`.

Une chaîne littérale peut être constituée d'un ou de plusieurs mots. Les règles suivantes s'appliquent lors de la définition d'une liste de chaînes littérales :

- Incluez la liste de chaînes entre parenthèses, par exemple (son). Si vous avez le choix entre les chaînes littérales, alors chaque chaîne doit être séparée par l'opérateur OR, comme par exemple `(un|une|le)` ou `(son|sa|ses)`.
- Utilisez des mots uniques ou composés.
- Séparez chaque terme de la liste par une barre verticale (|), qui équivaut à l'opérateur booléen `OR`.
- Entrez les formes au singulier et au pluriel si votre recherche porte sur les deux formes. Les inflexions ne sont pas générées automatiquement.
- Utilisez uniquement des minuscules.
- Pour réutiliser des chaînes littérales, définissez-les en tant que macro, puis utilisez cette macro dans vos autres macros et règles d'analyse des liens du texte.
- Si une chaîne contient des points ou des traits d'union, vous devez les inclure. Par exemple, pour trouver les correspondances de `a.k.a` dans le texte, entrez les points avec les lettres `a.k.a` comme chaîne littérale.

Opérateur d'exclusion

Utilisez `!` comme opérateur d'exclusion pour empêcher toute expression de la négation d'occuper un emplacement particulier. Vous ne pouvez ajouter un opérateur d'exclusion que manuellement avec l'édition de cellule (double-cliquez sur la cellule dans la table Valeur de règle ou Valeur de macro) ou dans la vue source. Par exemple, si vous ajoutez `$mTopic @{0,2} !($Positive) $Budget` à votre règle d'analyse des liens du texte, vous recherchez du texte contenant (1) un terme attribué à l'un des types dans la macro `mTopic`, (2) un intervalle de zéro à deux mots, (3) aucune instance d'un terme attribué au type `<Positive>` et (4) un terme attribué au type `<Budget>`. Ceci peut capturer « *ces voitures ont un prix correct* » mais ignorer « *les magasins proposent d'excellentes réductions* ».

Pour utiliser cet opérateur, vous devez saisir le point d'exclamation et les parenthèses manuellement dans la cellule d'élément en double-cliquant sur la cellule.

Intervalles de mots (<Toutes les chaînes de caractères>)

Un intervalle de mots, également appelé `<Toutes les chaînes de caractères>`, définit une plage numérique de chaînes de caractères pouvant être présente entre deux éléments. Les intervalles de mots sont très utiles lors de la mise en correspondance d'expressions très semblables qui ne diffèrent que très légèrement en raison de la présence de déterminants supplémentaires, de prépositions, d'adjectifs ou autres.

Tableau 45. Exemple des éléments dans une table Valeur de règle sans intervalle de mots

#	Elément
1	 `Inconnu`
2	 `mBeHave`
3	 `Positif`

Remarque : Dans la vue source, la valeur est définie comme suit : `$Inconnu $mBeHave $Positif`

Cette valeur correspondra aux expressions comme "*le personnel de l'hôtel était gentil*", où *personnel de l'hôtel* appartient au type`<Inconnu>`, *était* est sous la macro `mBeHave` et *gentil* est `<Positif>`. Mais elle ne correspondra pas à "*le personnel de l'hôtel était très gentil*".

Tableau 46. Exemple des éléments dans une table Valeur de règle avec un intervalle de mots <Toutes les chaînes de caractères>

#	Elément
1	`Inconnu`
2	`mBeHave`
3	
4	`Positif`

Remarque : Dans la vue source, la valeur est définie comme suit : `$Inconnu $mBeHave @{0,1} $Positif`

Si vous ajoutez un intervalle de mots dans la valeur de règle, celle-ci correspondra à "*le personnel de l'hôtel était gentil*" et à "*le personnel de l'hôtel était très gentil*".

Dans la vue source ou avec l'édition de lignes, la syntaxe d'un intervalle de mots est `@{#,#}`, où `@` signifie un intervalle de mots et où `{#,#}` définit le nombre minimum et maximum de mots acceptés entre l'élément précédent et l'élément suivant. Par exemple, `@{1,3}` signifie qu'une correspondance peut être trouvée entre les deux éléments définis, si ces derniers sont séparés par au moins un mot et par pas plus de trois mots. `@{0,3}` signifie qu'une correspondance peut être trouvée entre les deux éléments définis si 0, 1, 2 ou 3 mots sont présents mais pas plus de trois mots.

Affichage et utilisation en mode source

Pour chaque règle ou macro, l'éditeur TLA génère le code source sous-jacent qui est utilisé par l'extracteur pour mettre en correspondance et produire la sortie TLA. Si vous préférez travailler avec le code lui-même, vous pouvez l'afficher et l'éditer en cliquant sur le bouton "Afficher la source" dans la partie supérieure de l'éditeur. La vue des sources vous amène directement à la règle ou macro sélectionnée et la met en surbrillance. Cependant, nous vous conseillons d'utiliser les sous-fenêtres de l'éditeur afin de réduire les risques d'erreurs.

Lorsque vous avez terminé de visualiser ou de modifier la source, cliquez sur **Quitter la source**. Si vous générez une syntaxe non valide pour une règle, vous devrez résoudre le problème avant de quitter la vue des sources.

Important : Si vous effectuez une modification dans la vue des sources, nous vous recommandons vivement de modifier les règles et les macros une par une. Après avoir modifié une macro, validez les résultats en les extrayant. Si vous êtes satisfait du résultat, nous vous recommandons d'enregistrer le modèle avant d'effectuer une autre modification. Si vous n'êtes pas satisfait du résultat ou si une erreur se produit, rétablissez vos ressources enregistrées.

Macros dans la vue des sources

```
[macro]
name = nom_macro
value = ([nom_type|nom_macro|chaîne_littérale|intervalle_mots])
```

Tableau 47. Entrées de macro

`[macro]`	Chaque macro doit commencer avec la ligne marquée [macro] pour désigner le début d'une macro.
`name`	Le nom de la définition de macro. Chaque nom doit être unique.
`value`	Combinaison de un ou plusieurs types, de chaînes littérales, d'intervalles de mots ou de macros. Pour plus d'informations, voir «Eléments pris en charge pour les règles et les macros», à la page 228. Si vous combinez des arguments, vous devez utiliser des parenthèses ( ) pour regrouper les arguments et le caractère `|` pour indiquer un booléen OR.

En plus des conseils et de la syntaxe que présente la rubrique sur les macros, la vue des sources contient quelques conseils supplémentaires qui ne sont pas obligatoires lors de l'utilisation de la vue de l'éditeur. Les macros doivent également respecter les règles suivantes lors de l'utilisation du mode Source :

- Chaque macro doit commencer avec la ligne marquée `[macro]` pour désigner le début d'une macro.
- Pour désactiver un élément, ajoutez un indicateur de commentaire (#) au début de chaque ligne concernée.

Exemple. Cet exemple définit une macro nommée `mTopic`. La valeur de `mTopic` est la présence d'un terme correspondant à *un* des types suivants : `<Product>`, `<Person>`, `<Location>`, `<Organization>`, `<Budget>` ou `<Unknown>`.

```
[macro]
name=mTopic
value=($Unknown|$Product|$Person|$Location|$Organization|$Budget|$Currency)
```

Règles dans la vue des sources

```
[pattern(ID)]
name = nom_motif
value = [$nom_type|nom_macro|intervalle_mots|chaînes_littérales]
output = $caractère[\t]#caractère[\t]$caractère[\t]#caractère[\t]$caractère[\t]#caractère[\t]
```

Tableau 48. Entrées de règle

`[pattern (<ID>)]`	Indique le début de la règle d'analyse des liens du texte et fournit un ID numérique unique pour déterminer l'ordre de traitement.
`name`	Génère un nom unique pour cette règle d'analyse des liens du texte.
`value`	Fournit la syntaxe et les arguments à mettre en correspondance avec le texte. Pour plus d'informations, voir «Eléments pris en charge pour les règles et les macros», à la page 228.
`output`	Le format de sortie des motifs mis en correspondance résultants trouvés dans le texte. La sortie ne ressemble pas toujours à la position originale exacte des éléments du texte source. De plus, il est possible d'avoir plusieurs lignes de sortie pour une règle d'analyse des liens du texte donnée en plaçant chaque sortie sur une ligne séparée. Syntaxe de la sortie : • Sortie séparée avec le code de tabulation `\t`, tel que `$1\t#1\t$3\t#3` • `$` et un numéro indique le terme trouvé correspondant à l'argument défini dans le paramètre de valeur à cette position. Ainsi, `$1` signifie le terme correspondant au premier argument défini pour la valeur. • `#` et un numéro indique le nom du type de l'élément à cette position. Si un élément est une liste de chaînes littérales, le type `<Unknown>` est attribué. • Une valeur `Null\tNull` ne créera aucun résultat.

En plus des conseils et de la syntaxe que présente la rubrique sur les règles, la vue des sources contient quelques conseils supplémentaires qui ne sont pas obligatoires lors de l'utilisation de la vue de l'éditeur. Les règles doivent également respecter les règles suivantes lors de l'utilisation du mode Source :

- Lorsque deux éléments ou plus sont définis, ils doivent être placés entre parenthèses, qu'ils soient facultatifs ou non (par exemple, `($Negative|$Positive)` *ou* `($mCoord|$SEP)?`). `$SEP` représente une virgule.
- Le premier élément d'une règle d'analyse des liens du texte ne peut pas être un élément facultatif. Par exemple, vous ne pouvez pas commencer par `value = $mTopic?` ou `value = @{0,1}`.
- Vous pouvez associer une quantité (ou un nombre d'instances) à une chaîne de caractères. Cela vous permet de créer une règle qui englobe tous les cas, plutôt que de rédiger une règle distincte par cas. Vous pouvez, par exemple, utiliser la chaîne littérale `($SEP|et)` si vous recherchez une virgule (`,`) ou la conjonction `et`. Si vous étendez cette recherche en ajoutant une quantité de sorte que la chaîne littérale devienne `($SEP|and){1,2}`, les instances "`,`" "`et`" "`, et`" seront prises en compte.
- Les espaces ne sont pas pris en charge entre un nom de macro et les caractères `$` et `?` dans la `valeur` de la règle d'analyse des liens du texte.
- Les espaces ne sont pas pris en charge dans la `sortie` de la règle d'analyse des liens du texte.
- Pour désactiver un élément, ajoutez un indicateur de commentaire (#) au début de chaque ligne concernée.

Exemple. Supposons que vos ressources contiennent la règle d'analyse des liens du texte TLA suivante et que vous avez activé l'extraction des résultats TLA :

```
## Pierre Martin est l'ancien DRH d'IBM en France
[pattern(201)]
name= 1_201
value = $Person ($SEP|$mDet|$mSupport|comme|puis){1,2} @{0,1} $Function
 (de|avec|pour|dans|vers|à) @{0,1} $Organization @{0,2} $Location
output = $1\t#1\t$4\t#4\t$7\t#7\t$9\t#9
```

Lors de l'extraction, le moteur du programme d'extraction lit chaque phrase et essaie de faire correspondre la séquence suivante :

Tableau 49. Exemple de séquence d'extraction

Position	Description des arguments
1	Le nom d'une personne (`$Person`),
2	Un ou deux des éléments suivants : virgule (`$SEP`), déterminant (`$mDet`), verbe auxiliaire (`$mSupport`), les chaînes “`puis`” ou “`comme`”,
3	0 ou 1 mot (`@{0,1}`)
4	Une fonction (`$Function`)
5	Une des chaînes suivantes : “`de`”, “`avec`”, “`pour`”, “`dans`”, “`vers`” ou “`à`”,
6	0 ou 1 mot (`@{0,1}`)
7	Le nom d'une organisation (`$Organization`),
8	0, 1 ou 2 mots (`@{0,2}`)
9	Le nom d'un emplacement (`$Location`),

Cet exemple de règle d'analyse des liens du texte correspondrait à des phrases ou des expressions comme :

Pierre Martin, le DRH d'IBM en France

Pierre Martin est l'ancien DRH d'IBM en France

IBM a nommé Pierre Martin comme DRH d'IBM en France

Cet exemple de règle d'analyse des liens du texte produirait la sortie suivante :

```
Pierre Martin <Person> directeur ressources humaines <Function> ibm <Organization> france <Location>
```

Où :

- `Pierre Martin` est le terme correspondant à `$1` (le premier élément de la règle d'analyse des liens du texte) et `<Person>` est le type de `Pierre Martin` (`#1`),
- `directeur ressources humaines` est le terme correspondant à `$4` (le 4ème élément de la règle d'analyse des liens du texte) et `<Function>` est le type de `directeur ressources humaines` (`#4`),
- `ibm` est le terme correspondant à `$7` (le 7ème élément de la règle d'analyse des liens du texte) et `<Organization>` est le type de `ibm` (`#7`),
- `france` est le terme correspondant à `$9` (le 9ème élément de la règle d'analyse des liens du texte) et `<Location>` est le type de `france` (`#9`)

Ensembles de règles dans la vue des sources

```
[set(<ID>)]
```

Où `[set (<ID>)]` indique le début d'un jeu de règles et fournit un ID numérique unique pour déterminer l'ordre de traitement des jeux.

Exemple. La phrase suivante contient des informations à propos des individus, leur fonction au sein d'une entreprise ainsi que les activités de fusion/acquisition de cette entreprise.

```
Org1 Inc a conclu un accord définitif de fusion avec Org2 Ltd, a déclaré John Doe, PDG de Org2
Ltd.
```

Vous pouvez écrire une règle avec plusieurs sorties pour traiter les différents résultats possibles :

```
## Org1 Inc a conclu un accord définitif de fusion avec Org2, a déclaré
John Doe, PDG de Org2 Ltd.

[pattern(020)]
name=020
value = $Organization @{0,4} $ActionNouns @{0,6}  $mOrg @{1,2}
 $Person @{0,2} $Function @{0,1} $Organization
output = $1\t#1\t$3\t#3\t$5\t#5
output = $7\t#7\t$9\t#9\t$11\t#11
```

qui génère les 2 motifs de sortie suivants :

- `org1 inc<Organisation> + merges with <InstructionActive> + org2 ltd<Organisation>`
- `john doe <Personne> + ceo <Fonction> + org2 ltd<Organisation>`

Important ! Gardez à l'esprit que les autres opérations de traitement linguistique sont exécutées lors de l'extraction des motifs TLA. Dans ce cas, `fusion` est regroupé sous `fusionner` lors de la phase de regroupement des synonymes du processus d'extraction. Et puisque `fusionner` est du type `<ActiveVerb>`, ce nom de type apparaît dans la sortie du motif TLA finale. Par conséquent, lorsque la sortie lit `t$3\t#3`, cela signifie que le motif va afficher le concept final et le type final pour le troisième élément après l'application de tous les traitements linguistiques (synonymes et autres regroupements).

A la place de règles complexes comme la précédente, il est plus facile de gérer et d'utiliser deux règles. La première est spécialisée dans la recherche de fusions/acquisition entre les entreprises :

```
[set(1)]
## Org1 Inc a conclu un accord définitif de fusion avec Org2 Ltd
[pattern(44)]
name=firm + action + firm_0044
value=$mOrg @{0,20} $ActionNouns @{0,6} $mOrg
output(1)=$1\t#1\t$3\t#3\t$5\t#5
```

qui génère org1 inc<Organisation> + merges with <InstructionActive> + org2 ltd <Organisation>

La seconde est spécialisée dans l'individu/fonction/entreprise :

```
[set(2)]
## a déclaré John Doe, PDG de Org2 Ltd
[pattern(52)]
name=individual + role + firm_0007
value=$Person @{0,3} $mFunction (at|of)? ($mOrg|$Media|$Unknown)
output(1)=$1\t#1\t$3\tFunction\t$5\t#5
```

et génère john doe <Personne> + ceo <Fonction> + org2 ltd <Organisation>

Remarques

Ces informations ont été développées pour les produits et services offerts dans le monde.

Le présent document peut contenir des informations ou des références concernant certains produits, logiciels ou services IBM non annoncés dans ce pays. Pour plus de détails, référez-vous aux documents d'annonce disponibles dans votre pays, ou adressez-vous à votre partenaire commercial IBM. Toute référence à un produit, programme ou service IBM n'implique pas que seul ce produit, programme ou service IBM puisse être utilisé. Tout produit, programme ou service fonctionnellement équivalent peut être utilisé s'il n'enfreint aucun droit de propriété intellectuelle d'IBM. Cependant l'utilisateur doit évaluer et vérifier l'utilisation d'un produit, programme ou service non IBM.

IBM peut détenir des brevets ou des demandes de brevet couvrant les produits mentionnés dans le présent document. L'octroi de ce document n'équivaut aucunement à celui d'une licence pour ces brevets. Vous pouvez envoyer par écrit des questions concernant la licence à :

IBM Director of Licensing
IBM Corporation
North Castle Drive, MD-NC119
Armonk, NY 10504-1785
US

Pour toute demande au sujet des licences concernant les jeux de caractères codés sur deux octets (DBCS), contactez le service Propriété intellectuelle IBM de votre pays ou adressez vos questions par écrit à :

Intellectual Property Licensing
Legal and Intellectual Property Law
IBM Japan Ltd.
19-21, Nihonbashi-Hakozakicho, Chuo-ku
Tokyo 103-8510, Japan

LE PRESENT DOCUMENT EST LIVRE "EN L'ETAT". IBM DECLINE TOUTE RESPONSABILITE, EXPLICITE OU IMPLICITE, RELATIVE AUX INFORMATIONS QUI Y SONT CONTENUES, Y COMPRIS EN CE QUI CONCERNE LES GARANTIES DE VALEUR MARCHANDE OU D'ADAPTATION A VOS BESOINS. Certaines juridictions n'autorisent pas l'exclusion de garanties explicites ou implicites lors de certaines transactions, par conséquent, il est possible que cet énoncé ne vous concerne pas.

Ces informations peuvent contenir des erreurs techniques ou des erreurs typographiques. Ce document est mis à jour périodiquement. Chaque nouvelle édition inclut les mises à jour. IBM peut, à tout moment et sans préavis, modifier les produits et logiciels décrits dans ce document.

Les références à des sites Web non IBM contenues dans le présent document sont fournies à titre d'information uniquement et n'impliquent en aucun cas une adhésion aux données qu'ils contiennent. Les éléments figurant sur ces sites Web ne font pas partie des éléments du présent produit IBM et l'utilisation de ces sites relève de votre seule responsabilité.

IBM pourra utiliser ou diffuser, de toute manière qu'elle jugera appropriée et sans aucune obligation à votre égard, tout ou partie des informations qui lui seront fournies.

Les licenciés souhaitant obtenir des informations permettant : (i) l'échange des données entre des logiciels créés de façon indépendante et d'autres logiciels (dont celui-ci), et (ii) l'utilisation mutuelle des données ainsi échangées, doivent adresser leur demande à :

IBM Director of Licensing
IBM Corporation
North Castle Drive, MD-NC119
Armonk, NY 10504-1785
US

Ces informations peuvent être disponibles, soumises à des conditions générales, et dans certains cas payantes.

Le programme sous licence décrit dans le présent document et tous les éléments sous licence disponibles s'y rapportant sont fournis par IBM conformément aux dispositions du Livret Contractuel IBM, des Conditions internationales d'utilisation des Logiciels IBM ou de tout autre contrat équivalent.

Les données de performances et les exemples de client ne sont présentés qu'à des fins d'illustration. Les performances réelles peuvent varier en fonction des configurations et des conditions d'exploitation spécifiques.

Les informations concernant des produits non IBM ont été obtenues auprès des fournisseurs de ces produits, par l'intermédiaire d'annonces publiques ou via d'autres sources disponibles. IBM n'a pas testé ces produits et ne peut confirmer l'exactitude de leurs performances ni leur compatibilité. Elle ne peut recevoir aucune réclamation concernant des produits non IBM. Toute question concernant les performances de produits non IBM doit être adressée aux fournisseurs de ces produits.

Toute instruction relative aux intentions d'IBM pour ses opérations à venir est susceptible d'être modifiée ou annulée sans préavis, et doit être considérée uniquement comme un objectif.

Ces informations contiennent des exemples de données et de rapports utilisés au cours d'opérations quotidiennes standard. Pour les illustrer le mieux possible, ces exemples contiennent des noms d'individus, d'entreprises, de marques et de produits. Toute ressemblance avec des noms de personnes ou de sociétés réelles serait purement fortuite.

Marques

IBM, le logo IBM et ibm.com sont des marques d'International Business Machines dans de nombreux pays. Les autres noms de produits et de services peuvent être des marques d'IBM ou d'autres sociétés. La liste actualisée de toutes les marques d'IBM est disponible sur la page Web "Copyright and trademark information", à l'adresse www.ibm.com/legal/copytrade.shtml.

Intel, le logo Intel, Intel Inside, le logo Intel Inside, Intel Centrino, le logo Intel Centrino, Celeron, Intel Xeon, Intel SpeedStep, Itanium, et Pentium sont des marques commerciales ou des marques déposées de Intel Corporation ou de ses filiales aux Etats-Unis et dans d'autres pays.

Linux est une marque déposée de Linus Torvalds aux Etats-Unis et/ou dans d'autres pays.

Microsoft, Windows, Windows NT et le logo Windows sont des marques commerciales de Microsoft Corporation aux Etats-Unis et/ou dans d'autres pays.

UNIX est une marque déposée de The Open Group aux Etats-Unis et dans d'autres pays.

Java ainsi que tous les logos et toutes les marques incluant Java sont des marques d'Oracle et/ou de ses sociétés affiliées.

Les autres noms de produits et de services peuvent être des marques d'IBM ou d'autres sociétés.

Index

Caractères spéciaux

A

B

C

D

E

O

P

R

S

T

V

Imprimé en France