IBM SPSS Modeler Text Analytics １５ユーザー ガイド

注:この情報とこの情報がサポートする製品を使用する前に、注意 p.421 中の全般的情報をお読みください。

本版は、IBM® SPSS® Modeler Text Analytics 15 およびその後のリリースと変更に、新しい版が別途表示されない限り、適用されます。

アドビ製品の画面コピーは、Adobe Systems Incorporated の承認を得て掲載しています。

Microsoft 製品の画面コピーは、Microsoft Corporation の承認を得て掲載しています。

はじめに

IBM® SPSS® Modeler Text Analytics は強力なテキスト分析機能を提供するものであり、高度な言語テクノロジと自然言語処理 (Natural Language Processing、NLP) を使用して、さまざまな無構造テキスト データを高速で処理し、重要な概念を抽出および整理します。さらに、SPSS Modeler Text Analytics はこれらのコンセプトをカテゴリにグループ化できます。

組織内のおよそ 80% のデータは、テキスト ドキュメントの形式です (例: レポート、Web ページ、電子メール、コール センターのメモ)。テキストは、組織が顧客の動向をより良く理解するための重要な要素です。NLP を組み込むシステムは、結合句などのコンセプトを効率的に抽出できます。さらに、規定となる言語の情報を使用して、キーワードを製品、組織、人物など、意味や状況をに応じて関連グループに分類できます。その結果、情報のニーズに対する関連性を迅速に確認できます。これらの抽出されたコンセプトとカテゴリは、人口統計など既存の構造化されたデータと組み合わせせることができ、さらに IBM® SPSS® Modeler の完全なデータ マイニング ツール パッケージを使ったモデル作成に適用することにより、より適切で焦点を絞った決定を行うことができます。

言語システムは情報の感度が高い、つまり辞書に多くの情報が含まれるほど、結果の品質が高くなります。SPSS Modeler Text Analytics には、キーワードや類義語の辞書、ライブラリ、およびテンプレートなど、一連の言語リソースが付属しています。またこの製品を使用すると、状況に合わせてこれらの言語リソースを開発および調整できます。言語リソースの調整はインタラクティブなプロセスで、正確なコンセプトの取得とカテゴリ化に必要です。CRM およびゲノムなど、特定のドメインのカスタム テンプレート、ライブラリ、辞書も含まれています。

IBM Business Analytics について

IBM Business Analytics ソフトウェアは、意思決定者がビジネス パフォーマンスの改善のために使用可能な完全で整合性があり、正確な情報を提供します。ビジネス インテリジェンス、予測解析、資金業務と戦略マネージメント、そして 分析アプリケーションの包括的ポートフォリオは、現行の業務に明確、迅速で、現実的な洞察をもたらし、将来的な結果を予測する能力を実現します。豊富な産業用ソリューション、証明された実践法、それに専門家によるサービスを組み合せることにより、あらゆる規模の会社組織が、最高の生産性を推進し、信頼できる意志決定を自動化し、そして、よりよい結果を実現させることができます。

このポートフォリオの一部として、IBM SPSS Predictive Analytics ソフトウェアは、会社組織がよりよい業績結果を推進するために、未来の事象を予測し、その洞察に基づき積極的に行動できるようにお役に立ちます。世界中の民間企業、官公庁、学校関係のユーザーが、不正を削減しリスクを緩和しながら、顧客を惹き付け、維持し、そして増やすことができ

る競争上の優位性のために、IBM SPSS テクノロジーに信頼をおいています。IBM SPSS ソフトウェアを日常業務に取り入れることにより、会社組織は先見性のある企業となります - ビジネスのゴールを達成し、計測可能な競争上の優位性を達成するための意志決定を直接的、自動的に行うことが可能となります。詳細な情報、または、営業担当者へのご連絡には、http://www.ibm.com/spssを開いてください。

テクニカル サポート

テクニカル サポートはカスタマーのメンテナンスに使用いただけます。IBM Corp.製品の使用方法、または対応するハードウェア環境へのインストールについてサポートが必要な場合は、テクニカル サポートにご連絡ください。テクニカル サポートのご利用には、http://www.ibm.com/supportのIBM Corp. ウェブ サイトをご覧ください。支援を要請される場合は、事前にユーザー、会社組織、そして、サポート契約を明確にしておいていただくよう、お願いします。

内容

パート I: Text Mining ノード

パート I:
Text Mining ノード

章

1

IBM SPSS Modeler Text Analytics について

IBM® SPSS® Modeler Text Analytics は強力なテキスト分析機能を提供するものであり、高度な言語テクノロジと自然言語処理 (Natural Language Processing、NLP) を使用して、さまざまな無構造テキスト データを高速で処理し、重要な概念を抽出および整理します。さらに、SPSS Modeler Text Analytics はこれらのコンセプトをカテゴリにグループ化できます。

組織内のおよそ 80% のデータは、テキスト ドキュメントの形式です (例: レポート、Web ページ、電子メール、コール センターのメモ)。テキストは、組織が顧客の動向をより良く理解するための重要な要素です。NLP を組み込むシステムは、結合句などのコンセプトを効率的に抽出できます。さらに、規定となる言語の情報を使用して、キーワードを製品、組織、人物など、意味や状況をに応じて関連グループに分類できます。その結果、情報のニーズに対する関連性を迅速に確認できます。これらの抽出されたコンセプトとカテゴリは、人口統計など既存の構造化されたデータと組み合わせせることができ、さらに IBM® SPSS® Modeler の完全なデータ マイニング ツール パッケージを使ったモデル作成に適用することにより、より適切で焦点を絞った決定を行うことができます。

言語システムは情報の感度が高い、つまり辞書に多くの情報が含まれるほど、結果の品質が高くなります。SPSS Modeler Text Analytics には、キーワードや類義語の辞書、ライブラリ、およびテンプレートなど、一連の言語リソースが付属しています。またこの製品を使用すると、状況に合わせてこれらの言語リソースを開発および調整できます。言語リソースの調整はインタラクティブなプロセスで、正確なコンセプトの取得とカテゴリ化に必要です。CRM およびゲノムなど、特定のドメインのカスタム テンプレート、ライブラリ、辞書も含まれています。

展開: 非構造化データのリアルタイムのスコアリングに IBM® SPSS® Modeler Solution Publisher を使用して、テキスト マイニング ストリームを展開できます。これらのストリームを展開する機能により、正常でクローズドループのテキスト マイニングの実装を実現します。たとえば、組織は、予測モデルを適用してマーケティング メッセージの精度をリアルタイムに向上させることにより、受信者または発信者からのメモ帳のメモを分析することができます。

注:SPSS Modeler Solution Publisher で SPSS Modeler Text Analytics を実行するには、ディレクトリ <install_directory>/ext/bin/spss.TMWBServer を $LD_LIBRARY_PATH 環境変数に追加します。

サポート言語の自動翻訳:SPSS Modeler Text Analytics と Language Weaver を組み合わせて、アラビア語、中国語、ペルシア語などのサポート言語のテキストを英語に翻訳できます。翻訳済みテキストのテキスト分析を実行して、これらの結果をソース言語の内容を理解できなかった人々に配布することができます。テキスト マイニングの結果は自動的に対応する外国語テキストにリンクされるため、組織は必要とされるネイティブ スピーカーを分析の最も重要な結果にのみ、集中させることができます。Language Weaver では、20 人年の高度な翻訳リサーチをかけて作成された統計翻訳アルゴリズムを使用した、自動言語翻訳を提供しています。

IBM SPSS Modeler Text Analytics バージョン 15 へのアップグレード

PASW Text Analytics または Text Mining for Clementin の前バージョンからのアップグレード

IBM® SPSS® Modeler Text Analytics バージョンを 15インストールする前に、新しいバージョンでも使用したいすべての TAP、テンプレート、ライブラリを現行のバージョンから保存してエクスポートする必要があります。これらのファイルは、最新バージョンのインストール時に削除または上書きされることのないディレクトリに保存することをお奨めします。

SPSS Modeler Text Analytics の最新バージョンをインストールした後に、保存した TAP ファイルをロードし、保存したライブラリーを追加し、そして、保存したテンプレートをインポートしロードして、最新バージョンでも使用することができます。

重要!最初に必要なファイルを保存やエクスポートしないで現行のバージョンをアンインストールすると、前バージョンで使用した TAP、テンプレート、パブリック ライブラリは失われ、SPSS Modeler Text Analytics バージョン 15で使用することはできません。

テキスト マイニングについて

現在、顧客の電子メール、コール センターのメモ、自由記述式のアンケート回答、ニュース フィード、Web フォームなど、非構造化または半構造のフォーマットの情報量が増加しています。この情報過多によって、多くの組織に「この情報をどのように収集、検証そして活用するのか」という問題をもたらします。

テキスト マイニングとは、テキスト形式の素材のコレクションを分析するプロセスで、作者がこれらのコンセプトの表現に使用した正確な単語またはキーワードを知らなくても、主要キーワードをキャプチャし、隠れた関連性や傾向を明らかにします。テキスト マイニングと情報検索は全く異な

りますが、これらが混同される場合があります。正確な情報検索および保存は大きな問題ですが、高品質のコンテンツ、情報に含まれる高品質な内容、用語集、および関連性の抽出および管理は非常に重要なプロセスです。

テキスト マイニングおよびデータ マイニング

テキストの各項目について、言語学的テキスト マイニングによりコンセプトのインデックス、およびこれらのコンセプトについての情報を返します。この抜き出された、構造化された情報は、その他のデータ ソースと組み合わせて、次のような質問を処理することができます。

- いっしょに出現するのはどのコンセプトですか ?
- コンセプトがリンクしているのは何ですか ?
- 抽出した情報から作成できる高レベルのカテゴリは何ですか ?
- コンセプトまたはカテゴリから予測するのは何ですか ?
- コンセプトまたはカテゴリからどのように動作を予測しますか ?

テキスト マイニングとデータ マイニングを組み合わせると、構造化データまたは非構造化データだけで行うよりも、すぐれた洞察が可能です。この処理は通常、次のステップに従って行われます。

1. **マイニングするテキストを特定する :** マイニングするテキストを準備します。テキストが複数のファイルにある場合、ファイルを 1 つの場所に保存します。データベースについては、テキストを入力するフィールドを決定します。
2. **テキストをマイニングして構造化データを抽出する :** テキスト マイニング アルゴリズムをソース テキストに適用します。
3. **コンセプト モデルおよびカテゴリ モデルを作成する :** 主要キーワードを特定またはカテゴリを作成します。非構造化データから返されるコンセプト数は通常大きくなります。スコアリングに最適なコンセプトおよびカテゴリを特定します。
4. **構造化データを分析する :** クラスタリング、分類、予測モデル作成など、従来のデータ マイニング手法を採用して、コンセプト間の関連性を検出します。抽出されたコンセプトを他の構造化データを結合し、コンセプトに基づいて今後の行動を予測します。

テキスト分析およびカテゴリ化

定性的分析の形式であるテキスト分析では、テキストからの役立つ情報を抽出し、このテキスト内の主要キーワードを適切な数のカテゴリにグループ化します。テキスト分析はすべての種類および長さのテキストに実行できますが、分析へのアプローチは若干異なります。

比較的短いレコードまたはドキュメントは、それほど複雑でなく、通常不明確な単語や回答があまり含まれていないため、最も容易にカテゴリ化されます。たとえば、短い自由記述式のアンケートで好きな休日の過ごし方を 3 つ挙げるよう質問した場合、ビーチに行く、国立公園に行く、または何もしないなどの多くの回答が見られることが予想される場合があります。一方、比較的長い自由記述式のアンケートの回答は、特に回答者が高学歴で意欲があり、またアンケートを記入するのに十分な時間がある場合、非常に複雑で長くなることがあります。アンケートで政治に関する考えを尋ねる、または政治に関するブログ フィードがある場合、あらゆる種類の問題および立場について、長いコメントがいくつかあると予想されることがあります。

非常に短い時間で長いテキスト ソースから主要キーワードを抽出して洞察に満ちたカテゴリを作成する機能は、IBM® SPSS® Modeler Text Analytics を使用するうえでの重要な利点です。この利点は、自動化された言語学的手法と統計的手法を組み合わせて得られるもので、テキスト分析プロセスの段階ごとに最も信頼できる結果を生成します。

言語処理および自然言語処理

すべての構造のないテキスト データの管理における主な問題は、コンピュータが理解できるようなテキストを作成するための標準的な規則がないという点です。言語、すなわち意味はすべてのドキュメントおよびすべてのテキストの部分で異なります。そのような非構造化データを正確に取得し構成する唯一の方法は、言語を分析してその意味を明らかにすることです。非構造化情報からコンセプトを抽出するには、いくつかの自動化されたアプローチがあります。これらのアプローチは、言語学的アプローチと非言語学的アプローチの 2 種類に分けられます。

いくつかの組織が、統計およびニューラル ネットワークに基づく自動化された非言語学的ソリューションを採用してきました。これらのソリューションでは、コンピュータ技術を駆使して、人間が読み込むよりはるかに迅速に主要キーワードをスキャンおよびカテゴリ化できます。しかし、こうしたソリューションの精度は非常に低くなります。多くの統計的システムでは、単語が出現する回数をただカウントし、関連するコンセプトへの統計的近接性を計算するだけです。これにより関連性の低い多くの結果、またはノイズを生み出し、また見つけるべき、ノイズのない結果を見逃すことになります。

限られた精度を補うために、いくつかのソリューションで複雑な非言語的規則を組み込み、関連性のある結果および関連性のない結果とを区別します。これを、規則に基づくテキスト マイニングといいます。

一方、言語学に基づくテキスト マイニングでは、人間の言語をコンピュータによる支援で分析する自然言語処理 (NLP) の原則をテキストの単語、句、構文論、または構造に適用します。NLP を組み込むシステムは、結合句などのコンセプトを効率的に抽出できます。さらに、規定とな

る言語の情報を使用して、キーワードを製品、組織、人物など、意味や状況をに応じて関連グループに分類できます。

言語学に基づくテキスト マイニングでは、さまざまな単語の形式が類似した意味を持っていることを認識し、文の構造を分析してテキストを理解するための枠組みを提供することによって、人間と同じようにテキスト マッチで意味を検出します。このアプローチでは、統計的システムの速度およびコストの効率の点を利用し、人間の手をほとんど必要とせず、精度がはるかに高くなります。

日本語以外のすべての言語テキストでの抽出プロセス時における統計的アプローチと言語学的アプローチとの違いを説明するために、reproduction of documents (**ドキュメントの複製**)についての質問に対する回答について考えてみましょう。統計的ソリューションおよび言語学的ソリューションのいずれも、reproduction (**複製**)という単語を展開して、copy (**コピー**)やduplication (**重複**)などの類義語を含めるようにする必要があります。展開しない場合、関連情報が見落とされてしまいます。ただし、統計的ソリューションによって、こうした種類の同義語集、同じ意味を持つ他のキーワードを検索使用する場合、birth (**誕生**)というキーワードも加わり、関連しない多くの結果を生成する場合があります。言語について理解することによって、定義というより信頼できるアプローチによって、言語学的テキスト マイニングを実行してテキストの曖昧さに切り込みます。

注:日本語テキスト展開はIBM® SPSS® Modeler Premiumで利用可能です。

日本語テキストにおける抽出プロセス時の統計的アプローチと言語学的アプローチとの差を、例として**沈む**という単語を使用して説明することができます。この単語を使用した、**日が沈む**や、**気分が沈む**という表現があります。統計的アプローチのみを使用すると、**日**、**気分**、および**沈む**という単語がそれぞれ個別に抽出されます。ただし、言語学的手法に基づく感性分析を使用すると、 **日**、**気分**、および **沈む**が抽出されるだけでなく、**気分が沈む**も抽出され、タイプ <**悪い - 悲しみ全般**> に割り当てられます。感性分析で言語学に基づいた手法を使用すると、より意味のある表現を抽出できます。感性の分析およびキャプチャを行い、定義というより信頼できるアプローチによって、言語学的テキスト マイニングを実行してテキストの曖昧さに切り込みます。

抽出プロセスがどうなっているのかを理解しておくと、言語リソース（ライブラリ、タイプ、類義語など）を微調整する際に役に立ちます。抽出プロセスのステップには以下のものがあります。

- ソース データの標準フォーマットへの変換
- 候補のキーワードの特定
- 類義語の等価クラスおよび統合の特定
- タイプの割り当て
- 二次分析によるインデックスの付与、および必要に応じてパターン マッチ

手順 1: ソース データの標準フォーマットへの変換

最初のステップでは、後続の分析に利用できるように、インポートしたデータを決まった形式に変換します。この変換は内部的に実行され、元のデータは変更されません。

手順 2: 候補のキーワードの特定

言語学的抽出において、候補となるキーワードを特定する際の言語リソースの役割を理解しておくのは大切なことです。言語リソースは、抽出が実行されるごとに使用されます。言語リソースは、テンプレート、ライブラリ、およびコンパイル済み辞書の形式で保存されています。ライブラリには、語のリスト、関係性、また抽出の実行や調整に使用される情報が含まれています。基幹辞書は表示・編集ができません。ただし、残りのリソースをテンプレート エディタ で、またはインタラクティブ ワークベンチ セッションの場合はリソース エディタ で編集できます。

コンパイル済み辞書は、SPSS Modeler Text Analytics の抽出エンジンの主要な、内部コンポーネントです。これらのリソースには、品詞コード (名詞、動詞、形容詞など) を含む基本形のリストを収めた一般辞書が含まれています。また、リソースには、<**地名**>、<**組織**>、または<**人名**>に多くの抽出されたキーワードを割り当てるために使用する、予約済みのビルトインのタイプも含まれています。詳細は、 A 付録 p.412 日本語テキストで使用できるタイプ を参照してください。注:日本語テキスト展開はSPSS Modeler Premiumで利用可能です。

これらコンパイル済み辞書のほか、製品にはいくつかのライブラリが付属し、それらを使用して、コンパイル済み辞書のタイプ定義およびコンセプト定義を補い、また類義語を提供することができます。これらのライブラリ、および作成したユーザー指定のライブラリは、いくつかの辞書で構成されています。これらには、キーワード辞書、類義語辞書、および不要語辞書が含まれています。詳細は、 A 付録 p.407 日本語テキストのリソースの編集 を参照してください。注:日本語テキスト展開はSPSS Modeler Premiumで利用可能です。

データがインポートおよび変換されると、抽出エンジンは抽出の候補のキーワードの特定を開始します。候補となるキーワードとは、テキスト内の概念を特定するのに使用される語や、語の集まりのことです。テキストを処理しているとき、単語 (**ユニターム**) および複合語 (**マルチターム**) は、品詞パターン抽出を使用して特定されます。たとえば、品詞パターンが「<**地名**> + <**名詞**>」のマルチターム**青森りんご**は、2 つの部分に分けられます。そして、候補の感性キーワードは、感性テキスト リンク分析を使用して特定されます。

たとえば、次のような日本語のテキストがあるとします。**写真が新鮮で良かった**。この場合、抽出エンジンは、感性テキスト リンク規則のいずれかを使用して (**品物**) + **が** + **良い** を合致させた後、感性タイプ **良い - 褒め・賞賛** を割り当てます。

注:前述のコンパイル済み一般辞書にあるキーワードは、ユニタームとして重要でないまたは言語学的にあいまいであるすべての単語を示します。これらの単語は、ユニタームを特定するときに不要語に追加されます、ただし、それらは、品詞を決定またはより長い候補の複合語（マルチターム）を参照している場合に再評価されます。

手順 3: 類義語の等価クラスおよび統合の特定

候補のユニタームおよびマルチタームが特定された後、正規化辞書を使用して、等価クラスを特定します。等価クラスは、ある語句の基本形、すなわち同じ語句の２つの表現を１つの形で表わしたものです。句を等価クラスに割り当てる目的は、たとえば、side effect と **副作用** を別のコンセプトとして扱わないようにすることです。等価クラスのどのコンセプトを使用するか、つまり、side effect または**副作用**のどちらを主要キーワードとして使用するかを判断するために、抽出エンジンは、次の規則を順に適用します。

- ライブラリのユーザー指定の形式。
- コンパイル済みリソースで定義されている最も頻度の高い形式。

手順 4: タイプの割り当て

次に、抽出されたコンセプトにタイプを割り当てます。タイプは、コンセプトの意味上のグループ化です。基幹辞書ならびにライブラリの両方がこのステップで使用されます。タイプには、上位レベルのコンセプト、肯定的な単語および否定的な単語、人名、地名、組織名などが含まれます。詳細は、 17 章 p.332 キーワード辞書 を参照してください。

日本語リソースには、タイプの異なるセットがあります。詳細は、 A 付録 p.412 日本語テキストで使用できるタイプ を参照してください。注:日本語テキスト展開はSPSS Modeler Premiumで利用可能です。

言語学的なシステムは、知識に依存します。つまり、辞書に含まれている情報が多いほど、より高い品質の結果が得られます。類義語の定義など、辞書の内容の変更は、そのまま結果の改善につながります。これは、通常、対話的な処理で、正確なコンセプトの検索に不可欠です。NLP は SPSS Modeler Text Analytics の主要な要素です。

抽出の方法

回答の主要キーワードの抽出時、IBM® SPSS® Modeler Text Analytics は言語学に基づくテキスト分析に依存します。このアプローチを用いると統計に基づくシステムがもたらすようなスピードと費用対効果が得られます。また人の手を介することがほとんどないので、極めて高い精度が得られます。言語学に基づくテキスト分析は、自然言語処理、あるいは計量言語学と呼ばれる研究分野に基づいています。

重要! 日本語テキストの場合、抽出プロセスは異なる手順に従って行われます。詳細は、 A 付録 p.401 抽出の方法 を参照してください。注:日本語テキスト展開はIBM® SPSS® Modeler Premiumで利用可能です。

抽出プロセスがどうなっているのかを理解しておくと、言語リソース（ライブラリ、タイプ、類義語など）を微調整する際に役に立ちます。抽出プロセスのステップには以下のものがあります。

- ソース データの標準フォーマットへの変換
- 候補のキーワードの特定
- 類義語の等価クラスおよび統合の特定
- タイプの割り当て
- インデックス化
- パターンおよびイベント抽出の合致

手順 1: ソース データの標準フォーマットへの変換

最初のステップでは、後続の分析に利用できるように、インポートしたデータを決まった形式に変換します。この変換は内部的に実行され、元のデータは変更されません。

手順 2: 候補のキーワードの特定

言語学的抽出において、候補となるキーワードを特定する際の言語リソースの役割を理解しておくのは大切なことです。言語リソースは、抽出が実行されるごとに使用されます。言語リソースは、テンプレート、ライブラリ、およびコンパイル済み辞書の形式で保存されています。ライブラリには、語のリスト、関係性、また抽出の実行や調整に使用される情報が含まれています。基幹辞書は表示・編集ができません。ただし、残りのリソース（テンプレート）をテンプレート エディタ で、またはインタラクティブワークベンチ セッションの場合はリソース エディタ で編集できます。

コンパイル済み辞書は、IBM® SPSS® Modeler Text Analytics の抽出エンジンの主要な、内部コンポーネントです。これらのリソースには、品詞コード（名詞、動詞、形容詞、副詞、分詞、限定詞、接続詞、前置詞）を含む基本形のリストを収めた一般辞書が含まれています。また、リソースには、<**地名**>、<**組織名**>、または<**人名**>に多くの抽出されたキーワードを割り当てるために使用する、予約済みのビルトインのタイプも含まれています。 詳細は、 17 章 p.334 ビルトインのタイプ を参照してください。

これらコンパイル済み辞書のほか、製品にはいくつかのライブラリが付属し、それらを使用して、コンパイル済み辞書のタイプ定義およびコンセプト定義を補い、またその他のタイプや類義語を提供することができます。これらのライブラリ、および作成したユーザー指定のライブラリは、いくつかの辞書で構成されています。これらには、キーワード辞書、類義語辞書（類義語およびオプションの要素）、および不要語辞書が含まれています。 詳細は、 16 章 p.316 ライブラリの使用 を参照してください。

データがインポートおよび変換されると、抽出エンジンは抽出の候補のキーワードの特定を開始します。候補となるキーワードとは、テキスト内の概念を特定するのに使用される語や、語の集まりのことです。テキストの処理中、コンパイル済み辞書にない単語（**ユニターム**）は、抽出の候補のキーワードとして見なされます。候補の複合語（**マルチターム**）は、品詞パターン抽出を使用して特定されます。たとえば、品詞パターンが「形容詞、名詞」のマルチタームsports car（**スポーツ カー**）は、2 つの部分に分けられます。品詞パターンが「形容詞、形容詞、名詞」のマルチタームfast sports car（**高速スポーツ カー**）は、3 つの部分に分けられます。

注:前述のコンパイル済み一般辞書にあるキーワードは、ユニタームとして重要でないまたは言語学的にあいまいであるすべての単語を示します。これらの単語は、ユニタームを特定するときに不要語に追加されます、ただし、それらは、品詞を決定またはより長い候補の複合語（マルチターム）を参照している場合に再評価されます。

最後に、特殊なアルゴリズムを使用して、役職などの大文字の文字列を処理し、これらの特殊なパターンを抽出できるようにします。

手順 3: 類義語の等価クラスおよび統合の特定

候補のユニタームおよびマルチタームが特定された後、一連のアルゴリズムを使用して、ユニタームやマルチタームを比較し、等価クラスを特定します。等価クラスは、ある語句の基本形、すなわち同じ語句の２つの表現を１つの形で表わしたものです。句を等価クラスに割り当てる目的は、たとえば、president of the company（**会社の社長**）および company president（**会社社長**）を別のコンセプトとして扱わないようにすることです。等価クラスのどのコンセプトを使用するか、つまり、president of the company（**会社の社長**）または company president（**会社社長**）のどちらを主要キーワードとして使用するかを判断するために、抽出エンジンは、次の規則を順に適用します。

- ライブラリのユーザー指定の形式。
- テキスト全体で最も出現頻度の高い形式。
- テキスト全体で最も短い形式（通常、基本型に該当）。

手順 4: タイプの割り当て

次に、抽出されたコンセプトにタイプを割り当てます。タイプは、コンセプトの意味上のグループ化です。基幹辞書ならびにライブラリの両方がこのステップで使用されます。タイプには、上位レベルのコンセプト、肯定的な単語および否定的な単語、人名、地名、組織名などが含まれます。ユーザーがタイプを定義して追加することもできます。 詳細は、 17 章 p. 332 キーワード辞書 を参照してください。

手順 5: インデックスの付与

レコードまたはドキュメントのセット全体に、テキストの位置と各等価クラスの代表キーワードの間にポインタを確定してインデックスを付けます。候補のコンセプトの活用形インスタンスはすべて、候補の基本型としてインデックスが付けられます。基本形ごとに全体の出現頻度が計算されます。

手順 6: パターンおよびイベント抽出の合致

IBM SPSS Modeler Text Analytics は、タイプやコンセプトだけでなく、それらの関係性も見つけることができます。この製品ではいくつかのアルゴリズムおよびライブラリを使用でき、またタイプおよびコンセプトの間の関係性パターンを抽出する機能が用意されてます。製品に対する反応などの特定の意見、または政治的グループやゲノムのリンクなど、人々またはオブジェクトの間の関係性リンクを探す場合に特に役立ちます。

カテゴリ化の方法

IBM® SPSS® Modeler Text Analytics でカテゴリモデルを作成する場合、いくつかの手法から選択して、カテゴリを作成できます。すべてのデータセットが一意であるため、手法の数やそれらを適用する順序は、変わる場合があります。結果の解釈が、他の人とは異なる場合があるため、テキスト データにとってどの手法が最良の結果を生み出すか、それぞれの手法を検証する必要があります。SPSS Modeler Text Analytics では、カテゴリをさらに検証し、調整できるワークベンチ セッションでカテゴリ モデルを作成できます。

このガイドの場合、**カテゴリの作成**は、カテゴリ定義の生成および、1 つまたは複数のビルトインの手法を使用した分類を指し、また**カテゴリ化**は、スコアリング、またはラベル付け、一意の識別子（名前/ID/値）を各レコードまたはドキュメントのカテゴリ定義に割り当てるプロセスのことを指します。

カテゴリ作成時、抽出されたコンセプトおよびタイプはカテゴリの構築ブロックとして使用されます。カテゴリを作成すると、カテゴリ定義の要素に一致するテキストが含まれる場合、レコードおよびドキュメントが自動的にカテゴリに割り当てられます。

SPSS Modeler Text Analytics には、自動カテゴリ作成手法がいくつか用意されており、ドキュメントまたはレコードを迅速にカテゴリ化することができます。

グループ化手法

使用できるそれぞれの手法は、特定の種類のデータおよび状況に適していますが、同じ分析で手法を組み合わせて、全範囲のドキュメントまたはレコードをキャプチャすると役に立つ場合があります。 複数のカテゴリのコンセプトを表示したり、重複するカテゴリを見つけることができます。

派生関係のコンセプトの語幹: コンセプト コンポーネントが形態的に関連するか、または語幹を共有するかどうかを分析するとき、コンセプトを取得し、そのコンセプトに関連するその他のコンセプトを検索することによって、カテゴリを作成します。 この手法は、生成された各カテゴリのコンセプトが類義語または意味の上で密接に関連しているため、類義語の複合語コンセプトを特定するのに非常に役立ちます。 長さの異なるデータを処理し、コンパクトなカテゴリをより少なく生成します。 たとえば、コンセプト opportunities to advance は、コンセプト opportunity for advancement および advancement opportunity とグループ化されます。 詳細は、 10 章 p. 204 派生関係のコンセプトの語幹 を参照してください。このオプションは、日本語テキストに対しては使用できません。

セマンティック ネットワーク: 各コンセプトの考えられえる意味を、単語の関係の拡張インデックスから特定することによって開始し、関連するコンセプトをグループ化することによってカテゴリを作成します。 この手法は、コンセプトがセマンティック ネットワークに認識され、あまり曖昧でない場合に最も適しています。 テキストに、ネットワークが認識していない特殊な用語または専門用語が含まれている場合はあまり役に立ちません。 たとえば、コンセプト granny smith apple は、granny smith と横の関係があるため、gala apple および winesap apple とグループ化されます。 またあるいは、コンセプト animal（**動物**）は、その下位語である cat（**ネコ**）およびkangaroo（**カンガルー**）とグループ化されます。 このリリースでは、英語テキストにのみ使用できます。 詳細は、 10 章 p. 207 セマンティック ネットワーク を参照してください。

内包関係のコンセプト: この手法では、一方の共通の文字列である単語を含むかどうかに基づき、マルチタームのコンセプト（複合語）をグループ化することによってカテゴリを作成します。 たとえば、コンセプト seat（**シート**）は、コンセプト safety seat（**セーフティ シート**）、seat belt（**シート ベルト**）、および seat belt buckle（**シート ベルトのバックル**）とグループ化されます。 詳細は、 10 章 p. 206 内包関係のコンセプト を参照してください。

共起: この手法では、テキスト内の共起関係のコンセプトからカテゴリを作成します。 ドキュメントおよびレコードでコンセプトまたはコンセプトパターンがいっしょに出現することが多いとき、共起関係のコンセプトはおそらくカテゴリ定義の値のものである基底となる関連を反映します。 単語が頻繁に共起する場合、共起規則が作成され、新しいサブカテゴリのカテゴリの記述子として使用できます。 たとえば、多くのレコードに単語 price（**価格**）および availability（**有効性**）が含まれている場合（ただし、一方を含み、もう一方を含まないレコードはほとんどない）、これらのコンセプトを共起規則（price & available）にグループ化し、たとえば

カテゴリ price のサブカテゴリに割り当てることができます。詳細は、10 章 p.209 共起規則 を参照してください。

- **最小ドキュメント数:**共起関係のコンセプトの重要性を判断できるようにするため、カテゴリの記述子として使用されるよう、指定の共起関係のコンセプトを含む必要のあるレコードまたはドキュメントの最小数を定義します。

IBM SPSS Modeler Text Analytics ノード

IBM® SPSS® Modeler に付属する多くの標準ノードとともに、テキスト マイニング ノードを使用して、テキスト分析の精度をストリームに組み込むこともできます。IBM® SPSS® Modeler Text Analytics では、それを実行するためのテキスト マイニング ノードがいくつか用意されています。これらのノードは、ノード パレットの SPSS Modeler Text Analytics タブに保存されています。

図 1-1
ノード パレットの IBM SPSS Modeler Text Analytics タブ

図 1-2
すべての IBM SPSS Modeler Text Analytics ノードへのズーム

次のノードが含まれます。

- **ファイル リスト入力ノード**で、テキスト マイニング プロセスへの入力として、ドキュメント名のリストを生成します。これは、テキストがデータベースや他の構造化ファイルではなく、外部ドキュメントに存在する場合に役立ちます。ftf@fCf< fŠfXfg 詳細は、 2 章 p.15 ファイル リスト ノード を参照してください。
- **Web フィード入力ノード**を使用して、RSS または HTML 形式のブログまたはニュース フィードなど、Web フィードからテキストを読み取り、このデータをテキスト マイニング プロセスで使用できます。Web フィード入力ノードは、フィードの各レコードに 1 つまたは複数のフィールドを出力しますが、後続のテキスト マイニング ノードでは、これを入力として選択できます。 詳細は、 2 章 p.21 Web フィード ノード を参照してください。
- **テキスト マイニング ノード**では、言語的手法を使用して、主要なコンセプトをテキストから抽出します。これらのコンセプトおよびそのほかのデータを使用してカテゴリを作成することができ、既知のパターンに基づいてコンセプト間の関係および関連を特製する機能（テキスト リ

ンク分析）を用意しています。ノードを使用して、テキスト データの内容を検討、またはコンセプト モデルまたはカテゴリ モデルのいずれかを作成できます。コンセプトおよびカテゴリは、人口統計などの既存の構造化されたデータを組み合わせることができ、モデル作成に適用することができます。 詳細は、 3 章 p. 35 テキスト マイニング モデル作成ノード を参照してください。

- **テキスト リンク分析ノード**は、コンセプトを抽出し、またテキスト内の既知のパターンに基づいて、コンセプト間の関係を特定します。パターンを抽出して、これらのコンセプトに関連付けられた意見または識別子のほか、コンセプト間の関係を見出すことができます。テキスト リンク分析ノードを使用して、より直接的にテキストからパターンを特定および抽出し、パターンの結果をストリーム内のデータセットに追加できます。ただし、テキスト マイニングモデル作成ノードのインタラクティブ ワークベンチ セッションを使用して TLA を実行することもできます。 詳細は、 4 章 p. 89 テキスト リンク分析ノード を参照してください。
- **翻訳ノード**を使用して、モデル作成を目的として、アラビア語、中国語、ペルシア語などのサポート言語から、英語または他の言語にテキストを翻訳できます。このノードで、そのままではサポートされない 2 バイト言語のドキュメントの翻訳が可能になります。さらに、アナリストが該当する言語を話すことができないとしても、そのようなドキュメントからコンセプトを抽出できるようになります。同じ機能がテキストモデル作成のノードから呼び出されますが、独立した翻訳ノードを使用することで、複数のノード内の翻訳をキャッシュに入れて再使用することができます。 詳細は、 5 章 p. 103 翻訳ノード を参照してください。
- 外部ドキュメントからテキストをマイニングする場合、**テキスト マイニング出力ノード**が、コンセプトが抽出されたドキュメントへのリンクを含む HTML ページを生成するのに使用できます。 詳細は、 6 章 p. 112 ファイル ビューア ノード を参照してください。

アプリケーション

一般的に、定期的に多くの分量のドキュメントを確認して、より詳細に検討するために主要な要素を特定する必要がある人々は、IBM® SPSS® Modeler Text Analytics を使用すると多くの利点があります。

特定のアプリケーションには、次のような機能があります。

- **科学的および医学的リサーチ:** 特許報告書、雑誌の記事、計画書の発行物など、二次リサーチの使用を検証します。以前は知られていなかった（例えば特定の製品に関連した医者など）関連性を識別します。薬品の開発プロセスにかかる時間を最小化します。遺伝子調査の補助として使用します。

- **投資リサーチ:** 毎日のアナリスト レポート、ニュース記事、企業のプレス リリースを確認して、主要な戦略ポイントまたは市場シフトを特定します。こうした情報のトレンド分析により、一定の期間にわたって、企業または業界の緊急の問題または機会について明らかにします。
- **不正検出:** 銀行および医療費の不正を使用して、以上を検出し、多数のテキストから警告を検出します。
- **市場リサーチ:** 市場リサーチの段階で使用し、自由記述式アンケートの回答の主要なトピックを特定します。
- **ブログおよび Web フィード分析:** 新しいフィード、ブログなどの主要なキーワードを使用して、モデルを検証および作成します。
- **CRM:** 電子メール、取引、調査など、すべての顧客との接点からのデータを使用し、モデルを作成します。

章

2

ソース テキストの読み取り

テキスト マイニングのデータは、データベースや行や列にデータを表示する四角形の形式など IBM® SPSS® Modeler で使用される標準的な形式、またはこの構造に準拠しない Microsoft Word、Adobe PDF、または HTML などのドキュメント形式です。

- Microsoft Word、Microsoft Excel、Microsoft PowerPoint、さらに Adobe PDF、XML、HTMLなど、標準データ構造に従っていないドキュメントのテキストを読み取るために、ファイル リスト ノードを使用して、ドキュメントまたはフォルダのリストをテキスト マイニングへの入力として生成できます。 詳細は、 p.15 ファイル リスト ノード を参照してください。
- RSS または HTML 形式のブログまたはニュース フィードなど、Web フィードからテキストを読み取るために、Web フィード ノードを使用して Web フィード データをテキスト マイニング プロセスの入力用に書式設定できます。 詳細は、 p.21 Web フィード ノード を参照してください。
- 顧客のコメント用に 1 つまたは複数のテキスト フィールドを含むデータベースなど、SPSS Modeler で使用する標準データ形式のテキストを読み取るために、SPSS Modeler にネイティブの標準入力ノードを使用できます。詳細は、SPSS Modeler ノードのマニュアルを参照してください。

ファイル リスト ノード

Microsoft Word、Microsoft Excel、Microsoft PowerPoint、さらにAdobe PDF、XML、HTMLなどの形式で保存された構造のないドキュメントのテキストを読み取るために、ファイル リスト ノードを使用して、ドキュメントまたはフォルダのリストをテキスト マイニングへの入力として生成できます。構造のないテキスト ドキュメントは、IBM® SPSS® Modeler で使用される他のデータのようにフィールドやレコード（行および列）で表すことができないため、この方法が必要になります。ファイル リスト ノードは、テキスト マイニング パレットにあります。

ファイル リスト ノードは、入力ノードとして機能しますが、実際のデータを読み取らず、指定したルートの下のドキュメントまたはディレクトリの名前を読み込み、これらをリストとして生成します。出力は、各ファイルの 1 つのレコードを表示した単一のフィールドで、後続のテキスト マイニング ノードでは、これを入力として選択できます。

このノードは、SPSS Modeler ウィンドウの下部にあるノード パレットの IBM® SPSS® Modeler Text Analytics タブにあります。 詳細は、 1 章 p.12 IBM SPSS Modeler Text Analytics ノード を参照してください。

重要!コンピュータのローカル エンコードに含まれていない文字を含むディレクトリ名およびファイル名は、サポートされていません。ファイル リスト ノードを含むストリームを実行しようとすると、これらの文字を含むファイル名またはディレクトリ名によって、ストリームの実行が失敗する場合があります。これは、フランス語のロケールで日本語のファイル名を使用するなど、外国語のディレクトリ名またはファイル名で起こる場合があります。

RTF の処理: RTF ファイルを処理するには、フィルタが必要になります。Microsoft の Web サイトから RTF フィルタをダウンロードして、手動で登録できます。

Adobe PDF 処理。Adobe PDF からテキストを抽出するためには、Adobe Reader バージョン 9 を、SPSS Modeler Text Analytics と IBM® SPSS® Modeler Text Analytics Server が稼働しているコンピュータにインストールする必要があります。

- **注**:Adobe Reader 10 以降にアップグレードしてはなりません。必要なフィルターが含まれていません。
- Adobe Reader をバージョン 9 にアップグレードすると、(概ね 1,000 を超えるような) 大量の Adobe PDF ドキュメント で作業する場合に発生するフィルタのメモリー リークによる処理エラーを回避できます。32 ビットまたは 64 ビットの Microsoft Windows OS で Adobe PDF ドキュメントを処理する場合、32 ビットシステムの場合はAdobe Reader バージョン 9.x に、64 ビットシステムの場合は、Adobe PDF iFilter 9 にアップグレードしてください。いずれもAdobe の Web サイトで入手できます。
- Adobe は、Adobe Reader 8.x 以降使用するフィルタリング ソフトウェアを変更しました。以前のAdobe PDF ファイルは、正しく読めなかったり、文字化けが発生する場合があります。これは Adobe サイドの問題であり、SPSS Modeler Text Analytics では制御できません。
- Adobe PDF の [ドキュメントのプロパティ] ダイアログの [セキュリティ] タブで、「コンテンツのコピーまたは拡張」の Adobe PDF のセキュリティ制限を「許可しない」に設定すると、そのドキュメントをフィルタリングして製品に読み込むことができなくなります。
- Adobe PDF ファイルは、Microsoft Windows 以外のプラットフォームでは処理できません。
- Adobe での制限により、イメージ ベースのAdobe PDF ファイルからのテキスト抽出はできません。

Microsoft Office 処理。

- Microsoft Office 2007 で導入されている Microsoft Word、Microsoft Excel、Microsoft PowerPoint の新しい形式を処理するには、SPSS Modeler Text Analytics Server が稼動しているコンピュータ（ローカルまたはリモート）に Microsoft Office 2007 をインストールするか、新しい Microsoft Office 2007 のフィルタ パック（Microsoft のサイトで入手可）をインストールする必要があります。
- Microsoft Office で作成されたファイルは、Microsoft Windows 以外のプラットフォームでは処理できません。

ローカル データのサポート: リモートの SPSS Modeler Text Analytics Server に接続し、ファイル リスト ノードを含むストリームがある場合、データが SPSS Modeler Text Analytics Server と同じコンピュータ上にあるか、サーバー コンピュータにファイル リスト ノードのソース　データが保存されているフォルダへのアクセス権限が割り当てられている必要があります。

ファイル リスト ノード:[設定] タブ

[設定] タブで、このノードに必要なディレクトリ、ファイルの拡張子、出力を定義できます。

注:テキスト マイニング抽出は、Microsoft Windows 以外のプラットフォームでは Microsoft Office や Adobe PDF のファイルを処理できません。ただし、XML、HTML またはテキスト ファイルは常に処理可能です。

コンピュータのローカル エンコードに含まれていない文字を含むディレクトリ名およびファイル名は、サポートされていません。ファイル リスト ノードを含むストリームを実行しようとすると、これらの文字を含むファイル名またはディレクトリ名によって、ストリームの実行が失敗する場合があります。これは、フランス語のロケールで日本語のファイル名を使用するなど、外国語のディレクトリ名またはファイル名で起こる場合があります。

図 2-1
[ファイル リスト ノード] ダイアログ ボックス[設定] タブ

ディレクトリ: 一覧表示するドキュメントを含むルート フォルダを指定します。

- **サブディレクトリを含む :** サブディレクトリもスキャンする場合に指定します。

リストに含めるファイル形式: 使用するファイル形式および拡張子を選択または選択解除できます。ファイルの拡張子の選択を解除すると、その拡張子を持つファイルは無視されます。次の拡張子でフィルタリングできます。

- .rtf、.doc、.docx、.docm
- .xls、.xlsx、.xlsm
- .ppt、.pptx、.pptm
- .txt、.text
- .htm、.html、.shtml
- .xml
- .pdf
- .$

注: 詳細は、 p. 15 ファイル リスト ノード を参照してください。

重要! バージョン 14 以降、「ディレクトリのリスト」オプションは使用できる、出力だけがファイルのリストとなります。

ファイル リスト ノード:その他のタブ

IBM® SPSS® Modeler ノードの［データ型］タブは、［注釈］タブ同様、標準タブです。

テキスト マイニングでのファイル リスト ノードの使用

テキスト データが、Microsoft Word、Microsoft Excel、Microsoft PowerPoint、Adobe PDF、XML、HTML などの形式の、構造のない外部ドキュメント内にある場合、ファイル リスト ノードを使用します。このノードを使用して、ドキュメントまたはフォルダのリストを、テキスト マイニング プロセス（後続のテキスト マイニング ノードまたはテキスト リンク分析ノード）への入力として生成します。

ファイル リスト ノードを使用する場合、テキスト マイニング ノードまたはテキスト リンク分析ノードの［テキスト］フィールドが **［ドキュメントへのパス名］** を示すように指定してください。選択したフィールドが、マイニングする実際のテキストを示すのではなく、テキストのあるドキュメントへのパスを示すようにします。

次の例では、ファイル リスト ノードをテキスト マイニング ノードに接続して、外部ドキュメント内にあるテキストを指定します。

図 2-2
ストリームの例:テキスト マイニング ノードに接続したファイル リスト ノード

- **ファイル リスト ノード（［設定］タブ）:** まず、このノードをストリームに追加して、テキスト ドキュメントが保存されている場所を指定しました。テキスト マイニングを実行するすべてのドキュメントを含むディレクトリを選択しました。

図 2-3
[ファイル リスト ノード] ダイアログ ボックス[設定] タブ

- **テキスト マイニング ノード ([フィールド] タブ):** 次に、テキスト マイニング ノードを ファイル リスト ノードに追加して接続しました。このノードで、入力形式、リソース テンプレート、および出力形式を定義しました。ファイル リストノードから作成されたフィールド名を選択し、テキスト フィールドが [**ドキュメントへのパス名**] やその他の設定を表示するオプションを選択しました。詳細は、 3 章 p. 54 ストリーム内のテキスト マイニング ノードの使用 を参照してください。

図 2-4
[テキスト マイニング ノード] ダイアログ ボックス:[フィールド] タブ

テキスト マイニング ノード使用の詳細は、「3 章」を参照してください。

Web フィード ノード

Web フィード ノードを使用して、Web フィードのテキスト データをテキスト マイニング プロセス向けに準備することができます。このノードは、次の 2 つの形式で Web フィードを受け入れます。

- **RSS 形式:** RSS は、Web コンテンツ向けに表示化された、単純な XML ベースの形式です。この形式の URL は、組織化されたニュース ソースやブログなどのリンクした記事のセットがあるページを示します。RSS は標準化された形式であるため、リンクした記事は自動的に特定され、データ ストリームの個別のレコードとして扱われます。フィルタリン

グ手法をテキストに適用しない限り、フィードの重要なテキスト データおよびレコードを特定するために、さらなる入力は必要ありません。

- **HTML 形式:** [入力] タブで、1 つまたは複数の URL を HTML ページに定義できます。[レコード] タブで、レコードの開始タグを定義し、対象の内容を区切るタグを指定して、これらのタグを選択した出力フィールド（説明、タイトル、更新日など）に割り当てます。 詳細は、p.24 Web フィード ノード:[レコード] タブ を参照してください。

重要! プロキシ サーバーを経由して Web の情報を取得しようとしている場合、IBM® SPSS® Modeler Text Analytics Client および Server の net.properties ファイルでプロキシ サーバーを有効にする必要があります。このファイル内の説明に従ってください。これは、Web フィード ノードを使用して Web にアクセスする場合または Language Weaver ライセンスを取得する場合に適用されます。これらの場合、接続が を経由するためです。クライアントの場合、デフォルトでは、ファイルは C:¥Program Files¥IBM¥SPSS¥Modeler¥15¥jre¥lib¥net.properties にあります。サーバーの場合、デフォルトでは、ファイルは C:¥Program Files¥IBM¥SPSS¥Modeler¥15ext¥bin¥spss.TMWBServer¥jre¥lib¥net.properties にあります。

このノードの出力は、レコードの説明に使用するフィールドのセットです。[**説明**] フィールドは、多くのテキスト コンテンツが含まれているため、最も一般的に使用されます。ただし、レコードの短い説明（[**短い説明**] フィールド）またはレコードのタイトル（[**タイトル**] フィールド）など、他のフィールドにも関心がある場合があります。出力フィールドのいずれかを、後続のテキスト マイニング ノードの入力として選択できます。

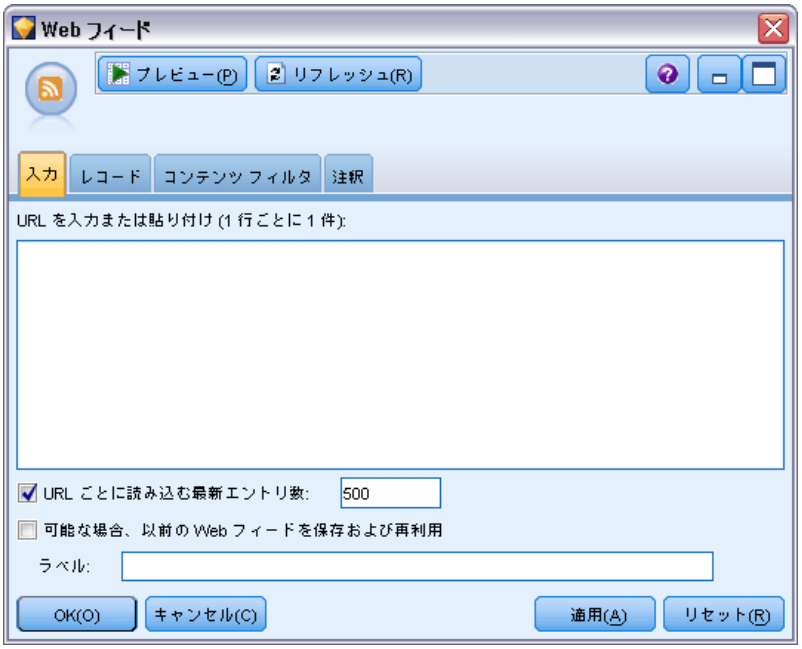
このノードは、IBM® SPSS® Modeler ウィンドウの下部にあるノード パレットの SPSS Modeler Text Analytics タブにあります。 詳細は、 1 章 p.12 IBM SPSS Modeler Text Analytics ノード を参照してください。

Web フィード ノード:[入力] タブ

[入力] タブを使用して、1 つまたは複数の Web アドレスまたは URL を指定し、テキスト データをキャプチャします。テキスト マイニングのコンテキストで、テキスト データを含むフィードの URL を指定できます。

重要! 非 RSS データを扱う場合、WebQL® などの Web スクラッピング ツールを使用して、コンテンツが異なる入力ノードを使用するツールから出力を収集して参照するよう自動化することをお勧めします。

図 2-5
Web フィード ノード ダイアログ ボックス:[入力] タブ



設定できるパラメータを次に示します。

URL を入力または貼り付け: 1 つまたは複数の URL を入力または貼り付けることができます。複数の URL を入力する場合、1 行ごとに 1 つの URL だけが入力し、Enter/Return キーを使用して、行を区切ります。ファイルへの完全な URL パスを入力します。フィードを示すこれらの URL は次の 2 つの形式のいずれかとなります。

- RSS 形式: RSS は、Web コンテンツ向けに表示化された、単純な XML ベースの形式です。この形式の URL は、組織化されたニュース ソースやブログなどのリンクした記事のセットがあるページを示します。RSS は標準化された形式であるため、リンクした記事は自動的に特定され、データ ストリームの個別のレコードとして扱われます。フィルタリング手法をテキストに適用しない限り、フィードの重要なテキスト データおよびレコードを特定するために、さらなる入力は必要ありません。
- HTML 形式: [入力] タブで、1 つまたは複数の URL を HTML ページに定義できます。[レコード] タブで、レコードの開始タグを定義し、対象の内容を区切るタグを指定して、これらのタグを選択した出力フィールド（説明、タイトル、更新日など）に割り当てます。非 RSS データを扱う場合、WebQL などの Web スクラッピング ツールを使用して、コンテンツが異なる入力ノードを使用するツールから出力を収集して参照するよう自動化することをお勧めします。 詳細は、 p.24 Web フィード ノード:[レコード] タブ を参照してください。

URL ごとに読み込む最新エントリ数: フィード内にある最初のレコードから始まるフィールドに表示された各 URL に読み込む最大レコード数を指定します。テキストの量は、テキストマイニング ノードまたはテキスト リンク分析ノード下流の抽出の処理速度に影響を与えます。

可能な場合、以前の Web フィードを保存および再利用: このオプションで、Web フィードをスキャンし、処理された結果をキャッシュします。そして、後続のストリームの実行後、指定されたフィードの内容が変わらない場合、またはフィードにアクセスできない場合（インターネットの機能停止など)、キャッシュされたバージョンを使用して、処理時間を短縮します。これらのフィードで見つかった新しいコンテンツは、次回ノードを実行するときにキャッシュされます。

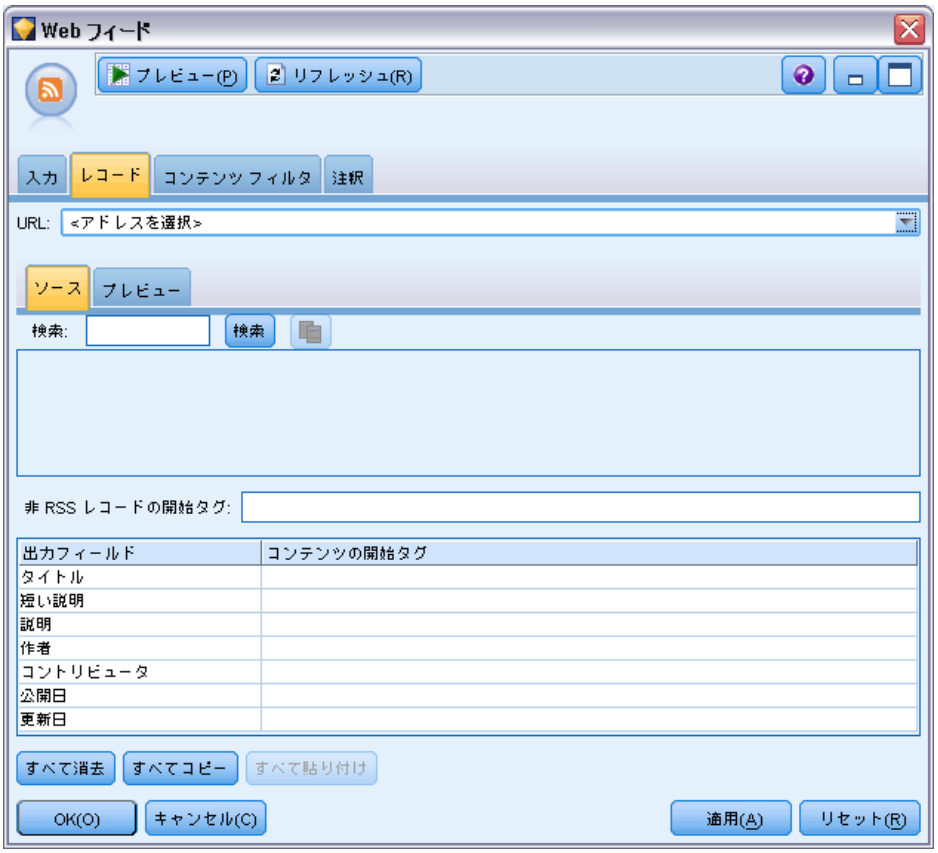
- **ラベル:[可能な場合、以前の Web フィードを保存および再利用]** を選択した場合、その結果のラベル名を指定する必要があります。このラベルを使用して、サーバーのキャッシュされたフィードを説明します。ラベルが指定されていない場合、またはラベルが認識されない場合、再利用はできません。これらの Web フィードのキャッシュを、IBM® SPSS® Text Analytics Administration Console のセッションテーブルで管理できます。詳細は、『SPSS Text Analytics Administration Console ユーザー ガイド』を参照してください。

Web フィード ノード:[レコード] タブ

[レコード] タブを使用して、新しいレコードの開始点、各レコードに関するその他の関連情報を特定し、非 RSS フィードのテキスト コンテンツを指定します。非 RSS フィード (HTML) に複数のレコード内にあるテキストが含まれていることがわかっている場合、レコードの開始タグをここで指定する必要があります。指定しない場合、テキストは 1 つのレコードとして扱われます。RSS フィードは標準化され、このタブでタグの指定は必要ありませんが、[プレビュー] タブで内容をプレビューできます。

重要! 非 RSS データを扱う場合、WebQL® などの Web スクラッピング ツールを使用して、コンテンツが異なる入力ノードを使用するツールから出力を収集して参照するよう自動化することをお勧めします。

図 2-6
Web フィード ノード ダイアログ ボックス:[レコード] タブ



URL: このドロップダウン リストには、[入力] タブで入力された URL のリストが表示されます。HTML 形式および RSS 形式のフィードが表示されます。URL アドレスが長すぎてドロップダウン リストにすべてを表示できない場合、自動的に省略記号を使用して途中で省略したテキストが表示されます (例: http://www.spss.com/example/start-of-address...rest-of-address/path.htm)。

- **HTML 形式のフィード**で、フィードに複数のレコード (エントリ) がある場合、テーブルに表示されたフィールドに対応するデータを含む HTML タグを定義できます。たとえば、新しいレコードが開始したことを示す開始タグ、更新日のタグ、または作成者の名前を定義できます。
- **RSS 形式のフィード**の場合、RSS は標準化された形式であるため、タグの入力は要求されません。ただし、必要に応じて [プレビュー] タグでサンプルの結果を表示できます。認識されたすべての RSS フィードの前に、RSS のロゴ イメージが表示されます。

[ソース] タブ: HTML フィードのソース コードを表示できます。このコードは編集できません。[検索] フィールドを試用して、このページの特定のタグまたは情報を検索できます。それらは、下のテーブルにコピーして貼り付けることができます。[検索] フィールドでは大文字および小文字の区別はされず、また文字列の一部に一致します。

[プレビュー] タブ: Web フィード ノードでレコードがどのように読み取られるかをプレビューできます。[プレビュー] タブ下のテーブルで JTML タグを定義して、レコードがどのように読み取られるかを変更できるため、このオプションは HTML フィードを使用する場合に特に役立ちます。

非 RSS レコードの開始タグ. このオプションは、非 RSS フィードにのみ適用されます。HTML ここで指定します。非 RSS フィードの開始タグを定義しない場合、ページ全体が 1 つのレコードとして扱われ、コンテンツ全体が [説明] フィールドで出力、そしてノードの実行日が [更新日] および [公開日] の両方に使用されます。

フィールド テーブル: このオプションは、非 RSS フィードにのみ適用されます。このテーブルで、事前定義された出力フィールドの開始タグを入力して、テキスト コンテンツを特定の出力フィールドに分割することができます。開始タグのみを入力します。HTML を解析し、テーブルの内容を HTML 内のタグ名および属性に一致させることによって、すべての合致が行われます。下部のボタンを使用して、定義したタグをコピーし、他のフィードにそのタグを再利用します。

テーブル 2-1
非 RSS フィード (HTML 形式) に使用できる出力フィールド

出力フィールド名	期待されるタグの内容
タイトル	レコードのタイトルを区切るタグ。（オプション）
短い説明	短い説明またはラベルを区切るタグ。（オプション）
説明	メインのテキストを区切るタグ。空白のままにすると、このフィールドには <body> タグ（レコードが 1 つある場合）他のすべての内容または現在のレコード内の内容（レコードの区切り文字が指定されている場合）が入力されます。
Author	レコードの作成者を区切るタグ。（オプション）
コントリビュータ	コントリビュータの名前を区切るタグ。（オプション）
公開日	テキストが公開された日付を区切るタグ。空白のままにすると、このフィールドにはノードがデータを読み取った日付が指定されます。
更新日	テキストが更新された日付を区切るタグ。空白のままにすると、このフィールドにはノードがデータを読み取った日付が指定されます。

タグをテーブルに入力すると、このタグを完全一致ではなく一致すべき最小タグとして使用してフィードをスキャンします。つまり、[タイトル] フィールドに と入力した場合、指定した属性を持つタグ（ ）など、フィードの タグに一致し、

がルート タグ（ ）に等しくなり、属性を含むデリバティブが［タイトル］出力フィールドにその内容を使用します。ルート タグを入力すると、さらに詳細な属性も含まれます。

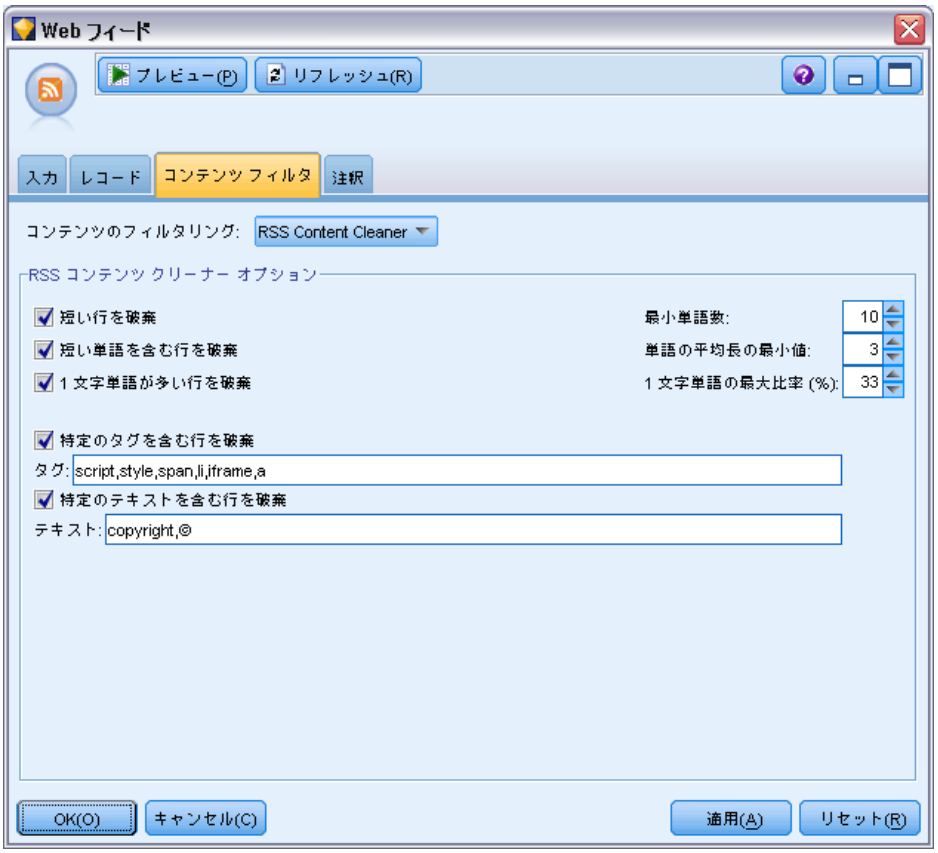
テーブル 2-2
出力フィールドのテキストの特定に使用される HTML タグの例

入力タグ	一致タグ	その他の一致タグ	一致しないタグ
			その他のタグ
<p class="auth">	<p class="auth">	<p color="black" class="auth" id="85643">	<p color="black">

Web フィード ノード:[コンテンツ フィルタ] タブ

［コンテンツ フィルタ］タブを使用して、フィルタ手法を RSS フィード コンテンツに適用します。このタブは、HTML フィードには適用されません。フィードに多くのテキストがヘッダー、フッター、メニュー、広告などの形式で含まれている場合、フィルタリングが必要な場合があります。このタブを使用して、不要な HTML タグ、JavaScript、短い単語または行をコンテンツから除外することができます。

図 2-7
Web フィード ノード ダイアログ ボックス:[コンテンツ フィルタ] タブ



コンテンツのフィルタリング:クリーニング手法を適用したくない場合、[なし] を選択します。適用する場合は、[RSS コンテンツ クリーナー] を選択します。

[RSS コンテンツ クリーナー オプション]:[RSS コンテンツ クリーナー] を選択すると、特定の基準に基づいて、行を破棄することができます。行は、<p> および <li> などの HTML タグによって区切られます。ただし、、<b>、および <font> などのインライン タグは使用されません。
 タグは改行として処理されます。

- **短い行を破棄**:ここで定義する**最小単語数**を含まない行が無視されます。
- **短い単語を含む行を破棄**:ここで定義する**単語の平均長の最小値**より長さが短い行が無視されます。
- **1 文字単語が多い行を破棄**:ここで定義する特定の**1 文字単語の比率**より小さい行が無視されます。

- **特定のタグを含む行を破棄**:フィールドで指定されたタグのいずれかを含む行のテキストが無視されます。
- **特定のテキストを含む行を破棄**:フィールドで指定されたテキストのいずれかを含む行が無視されます。

テキスト マイニングでの Web フィード ノードの使用

Web フィード ノードを使用して、インターネットの Web フィードのテキスト データをテキスト マイニング プロセス向けに準備することができます。このノードは、HTML 形式または RSS 形式で Web フィードを受け入れます。これらのフィードは、テキスト マイニングプロセス（後続のテキスト マイニング ノードまたはテキスト リンク分析ノード）の入力として機能します。

Web フィード ノードを使用する場合、[テキスト] フィールドがテキスト マイニング ノードまたはテキスト リンク分析ノードの**実際のテキスト**を示すよう指定して、これらのフィードが各記事またはブログのエントリに直接リンクするようにする必要があります。

重要! プロキシ サーバーを経由して Web の情報を取得しようとしている場合、IBM® SPSS® Modeler Text Analytics Client および Server の net.properties ファイルでプロキシ サーバーを有効にする必要があります。このファイル内の説明に従ってください。これは、Web フィード ノードを使用して Web にアクセスする場合または Language Weaver ライセンスを取得する場合に適用されます。これらの場合、接続が を経由するためです。クライアントの場合、デフォルトでは、ファイルは C:¥Program Files¥IBM¥SPSS¥Modeler¥15¥jre¥lib¥net.properties にあります。サーバーの場合、デフォルトでは、ファイルは C:¥Program Files¥IBM¥SPSS¥Modeler¥15ext¥bin¥spss.TMWBServer¥jre¥lib¥net.properties にあります。

例:テキストマイニング モデル作成ノードを使用した Web フィード ノード (RSS フィード)

次の例では、Web フィード ノードをテキスト マイニング ノードに接続して、RSS フィードのテキスト データをテキスト マイニング プロセスに提供します。

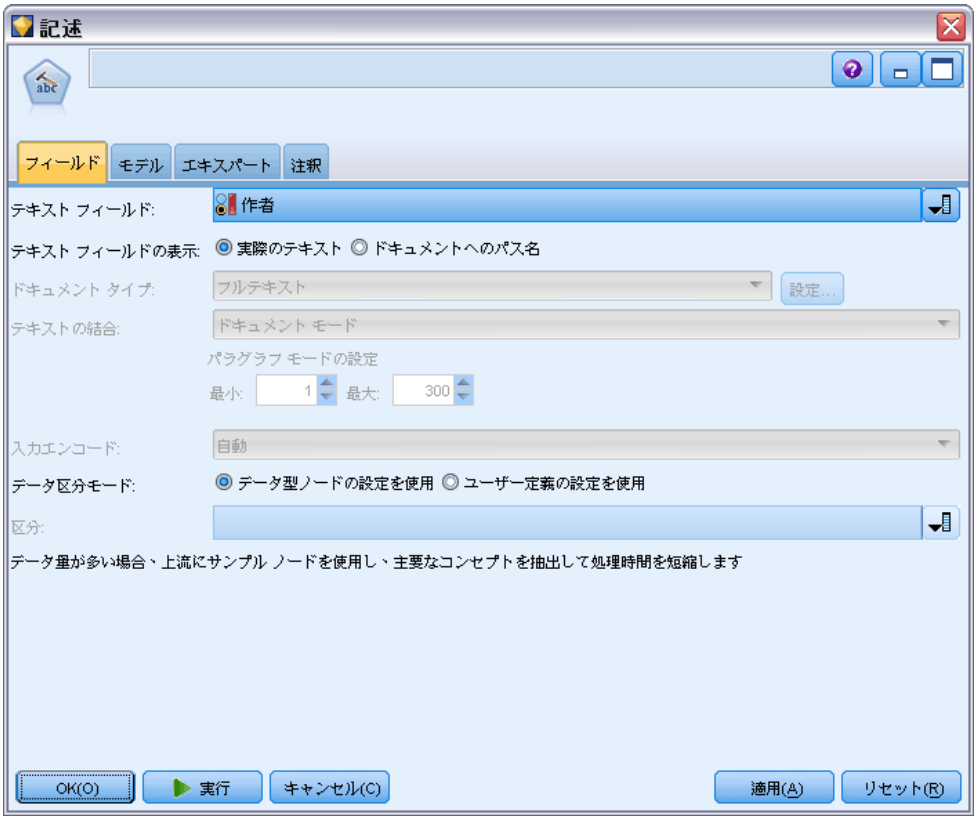
図 2-8
ストリームの例:テキストマイニング モデル作成ノードを使用した Web フィード ノード

- **Web フィード ノード ([入力] タブ):** まず、このノードをストリームに追加して、フィード コンテンツの場所を指定し、コンテンツの構造を検証しました。最初のタブで、URL を RSS フィードに指定しました。この例は RSS フィード向けであるため、形式は既に定義されており、[レコード] タブで変更を行う必要はありません。オプションのコンテンツ フィルタリング アルゴリズムを RSS フィードに使用できますが、この場合は適用されませんでした。

図 2-9
Web フィード ノード ダイアログ ボックス:[入力] タブ

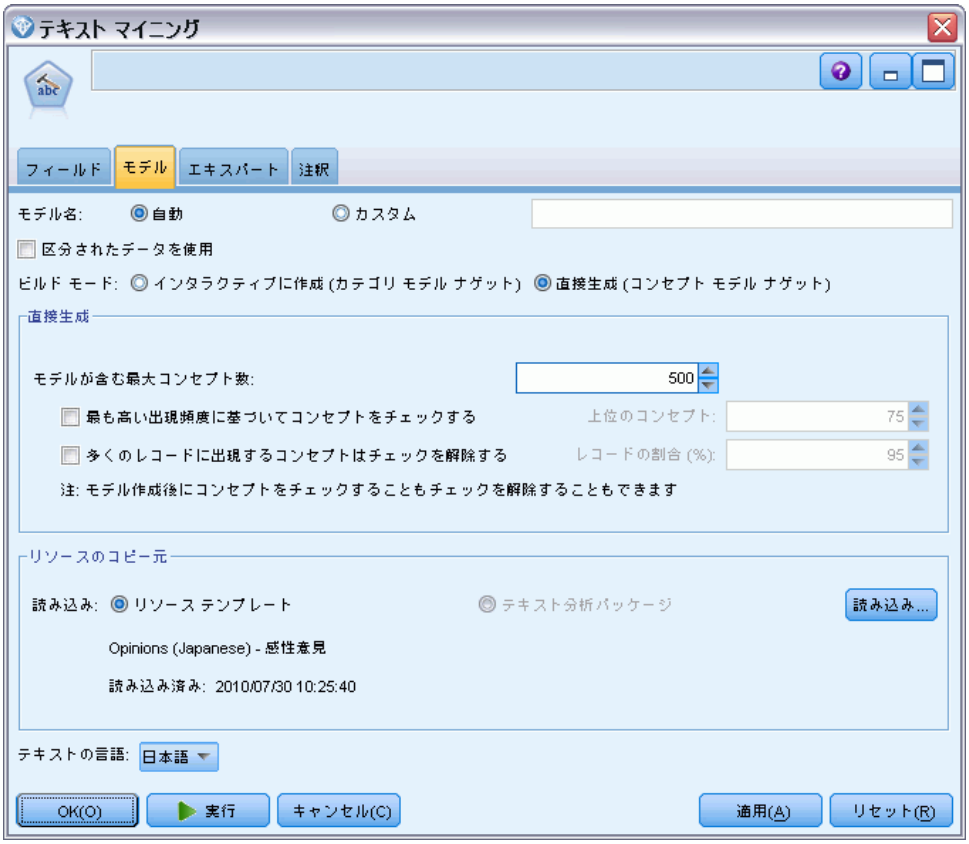
- **テキスト マイニング ノード ([フィールド] タブ):** 次に、テキスト マイニング ノードを Web フィード ノードに追加して接続しました。このタブで、Web フィード ノードによってテキスト フィールド出力を定義しました。この場合、[説明] フィールドを使用する必要がありました。また、[テキスト] フィールドが**実際のテキスト**やその他の設定を示すオプションを選択しました。

図 2-10
[テキスト マイニング ノード] ダイアログ ボックス:[フィールド] タブ



- **テキスト マイニング ノード ([モデル] タブ):** 次に［モデル］タブで、ビルド モードとリソースを選択しました。この例では、デフォルトのリソース テンプレートを使用して、このノードから直接コンセプト モデルを作成するよう選択しました。

図 2-11
テキスト マイニング ノード:[モデル] タブ



テキスト マイニング ノード使用の詳細は、「3 章」を参照してください。

章 3

コンセプトおよびカテゴリのマイニング

テキスト マイニング モデル作成ノードを使用して、次の 2 つのマイニング モデル ナゲットのうちいずれかを生成します。

- コンセプト モデル ナゲット は構造のあるテキスト データまたは構造のないテキスト データの目立ったコンセプトを明らかにし、テキスト データから抽出します。
- カテゴリ モデル ナゲット はドキュメントおよびレコードをスコアリングし、抽出したコンセプト（およびパターン）で構成されたカテゴリに割り当てます。

モデル ナゲットから抽出されたコンセプト、パターン、カテゴリを人口統計など既存の構造化されたデータと組み合わせ、IBM® SPSS® Modeler のツールの完全パッケージを使用して適用し、より適切で焦点を絞った決定を行うことができます。たとえば、顧客が頻繁にログイン問題をオンラインアカウント管理タスクの達成に対する主な障害として頻繁に一覧化する場合、「ログイン問題」をモデルに組み込むことが必要な場合があります。

また、テキスト モデル作成ノードは SPSS Modeler で完全に統合されており、PredictiveCallCenter のようなアプリケーションでの構造のないデータをリアルタイムにスコアリングするために、IBM® SPSS® Modeler Solution Publisher を使用してテキスト マイニング ストリームを展開できます。これらのストリームを展開する機能により、正常でクローズドループのテキスト マイニングの実装を実現します。たとえば、組織は、予測モデルを適用してマーケティング メッセージの精度をリアルタイムに向上させることにより、受信者または発信者からのメモ帳のメモを分析することができます。ストリームでテキスト マイニング モデルの結果を使用すると、予測データ モデルの精度が向上します。

注:SPSS Modeler Solution Publisher で IBM® SPSS® Modeler Text Analytics を実行するには、ディレクトリ <install_directory>/ext/bin/spss.TMWBServer を $LD_LIBRARY_PATH 環境変数に追加します。

SPSS Modeler Text Analytics では、抽出されたコンセプトおよびカテゴリを参照します。探索を目的とする作業およびモデル作成の間、より情報に基づく決定を行うことができるため、コンセプトおよびカテゴリの意味を理解することは重要です。

コンセプトおよびコンセプト モデル ナゲット

抽出プロセスで、テキスト データをスキャンして分析し、テキスト内の関心のあるまたは関連する単語（**選挙**または**平和**など）や語句（**大統領選挙**、**大統領の選挙**、または**平和条約**など）を特定します。これらの単語や句を、まとめて「キーワード」と呼びます。言語リソースを使用して、関連キーワードを抽出し、類似したキーワードを**コンセプト**と呼ばれる代表語でグループ化します。

このように、コンセプトはテキストおよび使用している言語リソースのセットによって、複数の基本キーワードを示すことができます。たとえば、従業員の満足調査を行い、コンセプト**給料**が抽出されたとします。また、**給料**に関連するレコードを参照し、**給料**が常にテキスト内にあるのではなく、特定のレコードにキーワード **賃金**、**報酬**、および**給与**のような類似した単語が含まれているとします。これらのキーワードは、抽出エンジンがキーワードを類似した単語として認識し、処理規則または言語リソースに基づいて類義語であると判断するため、**コスト**という名でグループ化されます。この場合、これらのキーワードのいずれかを含むドキュメントまたはレコードは、単語**給料**を含む場合と同じように扱われます。

コンセプトに基づいてグループ化されるキーワードを確認したい場合、インタラクティブ ワークベンチでコンセプトを検索したり、コンセプト モデルに示される類義語を確認することができます。 詳細は、 p. 66 コンセプト モデルの基本キーワード を参照してください。

コンセプト モデル ナゲットは、コンセプト（類義語またはグループ化されたキーワードを含む）を含むレコードまたはドキュメントの特定に使用できる一連のコンセプトで構成されています。コンセプト モデルは次の 2 つの方法で使用できます。1 つ目の方法は、元のソース テキストで見つかったコンセプトを検証および分析、または関心のあるドキュメントをすばやく特定することです。2 つ目の方法は、このモデルを新しいテキスト レコードまたはドキュメントに適用し、コール センターのメモ帳データから主要キーワードをリアルタイムに発見するなど、新しいドキュメント/レコードの同じ主要キーワードをすばやく特定することです。

詳細は、 p. 61 テキスト マイニング モデル ナゲット：コンセプト モデル を参照してください。

カテゴリおよびカテゴリ モデル ナゲット

テキスト内の主要なキーワード、情報、属性をキャプチャする高いレベルのコンセプトまたはトピックを示す**カテゴリ**を作成できます。カテゴリは、コンセプト、タイプおよび規則などの一連の記述子で構成されています。また、これらの記述子を共に使用して、レコードまたはドキュメントが指定されたカテゴリに属するかどうかを特定します。ドキュメントまたはレコードをスキャンして、テキストが記述子に合致するかどうかを確認する

ことができます。合致が見つかった場合は、ドキュメント/レコードはそのカテゴリに割り当てられます。このプロセスを、**カテゴリ化**といいます。

製品の自動的手法の頑健なセットを使用して、データに関する詳細な洞察を手動で使用し、またはそれらを組み合わせてカテゴリを自動的に作成することができます。このノードの［モデル］タブを使用してテキスト分析パッケージから事前に作成された一連のカテゴリを読み込むこともできます。カテゴリの手動作成またはカテゴリの調整は、インタラクティブ ワークベンチでのみ実行できます。 詳細は、 p.41 テキスト マイニング ノード:[モデル] タブ を参照してください。

カテゴリ モデル ナゲットは、一連のカテゴリとその記述子で構成されています。モデルを使用して、各ドキュメント/レコードのテキストに基づいて、一連のドキュメントまたはレコードをカテゴライズできます。各ドキュメントまたレコードが読み取られ、記述子の合致が見つかった各カテゴリに割り当てられます。このように、ドキュメントまたはレコードを複数のカテゴリに割り当てることができます。カテゴリ モデル ナゲットを使用して、自由記述式アンケートの回答またはブログのエントリなどの不可欠なキーワードを確認することができます。

詳細は、 p.77 テキスト マイニング モデル ナゲット:カテゴリ モデル を参照してください。

テキスト マイニング モデル作成ノード

テキスト マイニングモデル作成ノードは、言語学的手法および出現頻度に基づく手法を使用して、テキストから主要キーワードを抽出し、これらのコンセプトおよびその他のデータでカテゴリを作成します。ノードを使用して、テキスト データの内容を検討、またはコンセプト モデル ナゲットまたはカテゴリ モデル ナゲットのいずれかを作成できます。このモデル作成ノードを実行すると、内部の言語学的抽出エンジンは、自然言語処理手法を使用して、コンセプト、パターンまたはカテゴリを抽出して構成します。

テキスト マイニング ノードを実行し、**[直接生成]**を使用して、コンセプト モデル ナゲットまたはカテゴリ モデル ナゲットを自動的に作成できます。また、コンセプトを抽出、カテゴリを作成、および言語リソースを調整するだけでなく、テキスト リンク分析を実行してクラスタを検証できる**[インタラクティブに作成]**モードを使用して、より実践的な探索的アプローチを使用できます。 詳細は、 p.41 テキスト マイニング ノード:[モデル] タブ を参照してください。

このノードは、IBM® SPSS® Modeler ウィンドウの下部にあるノード パレットの IBM® SPSS® Modeler Text Analytics タブにあります。 詳細は、1 章 p.12 IBM SPSS Modeler Text Analytics ノード を参照してください。

要件: テキスト マイニング モデル作成ノードは、Web フィード ノード、ファイル リスト ノード、または標準的な入力ノードのいずれかからテキスト データを受け入れます。このノードは SPSS Modeler Text Analytics

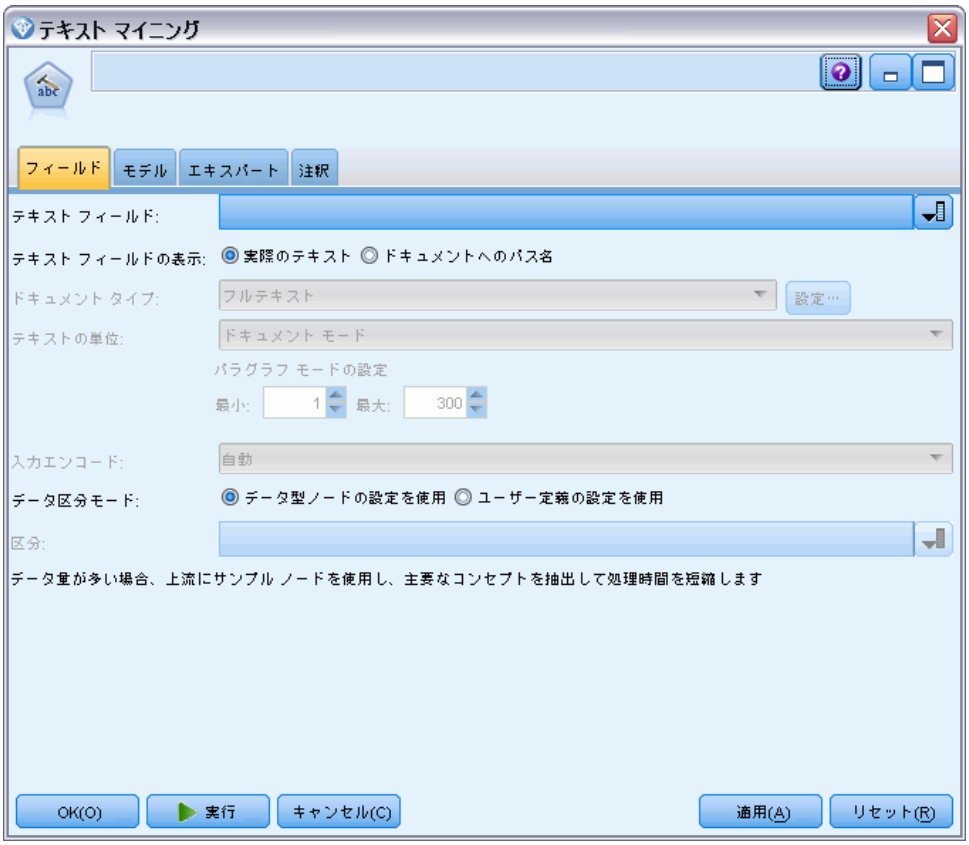
と共にインストールされており、SPSS Modeler Text Analytics パレット上で使用できます。

注:このノードは、以前のバージョンの Text Mining for Clementine に付属していた、すべてのユーザー向けのテキスト抽出ノードおよび日本語ユーザー向けの古いテキスト マイニング ノードに代わるものです。これらのノードまたはモデル ナゲットを使用する古いストリームを使用する場合、新しいテキスト マイニング ノードを使用してストリームを再作成する必要があります。注:日本語テキスト展開はIBM® SPSS® Modeler Premiumで利用可能です。

テキスト マイニング ノード:[フィールド] タブ

[フィールド] タブを使用して、コンセプトを抽出するデータのフィールド設定を指定します。大きいデータセットを扱う場合は、処理時間を短くするために、このノードから上流でサンプル ノードを使用するようにしてください。 詳細は、 p.54 時間を削減する上流のサンプリング を参照してください。

図 3-1
[テキスト マイニング モデル作成ノード] ダイアログ ボックス:[フィールド] タブ



設定できるパラメータを次に示します。

テキスト フィールド: マイニングするテキスト、ドキュメントのパス名、またはドキュメントへのディレクトリ パス名が入力されたフィールドを選択します。このフィールドはデータ ソースによって異なります。

テキスト フィールドの表示: これまでの結果で指定されたテキスト フィールドに何が入力されているかを示します。選択されるのは次のとおりです。

- **実際のテキスト:** コンセプトが抽出される正確なテキストをフィールドに入力する場合、このオプションを指定します。
- **ドキュメントへのパス名:** テキスト ドキュメントの場所へ1 つまたは複数のパス名をフィールドに入力する場合、このオプションを選択します。

ドキュメント タイプ: テキスト フィールドが [ドキュメントへのパス名] を示すよう指定した場合にのみ使用できます。ドキュメント タイプは、テキストの構造を指定します。次に示すタイプの 1 つを選択します。

- **フル テキスト**: 多くのドキュメントまたはテキスト ソースに使用します。テキストのセット全体をスキャンして抽出します。他のオプションとは異なり、このオプションに追加設定はありません。
- **構造のあるテキスト**: 参考文献形式、特許、特定および分析できる通常の構造を含むファイルに使用します。このドキュメントタイプを使用して、抽出プロセスのすべてまたは一部をスキップします。キーワードの区切り文字を定義、タイプを割り当て、出現頻度の最小値を指定できます。このオプションを選択する場合、[**設定**] ボタンをクリックして、[ドキュメント設定] ダイアログ ボックスの [**構造のあるテキストの書式設定**] 領域にテキストの区切り文字を入力します。詳細は、 p. 39［フィールド］タブのドキュメント設定 を参照してください。
- **XML テキスト**: 抽出するテキストを含む XML タグを指定します。他のタグはすべて無視されます。このオプションを選択する場合、[**設定**] ボタンをクリックして、[ドキュメント設定] ダイアログ ボックスの [**XMLテキストの書式設定**] 領域で、抽出プロセスで読み込まれるテキストを含む XML 要素を明示的に指定します。 詳細は、 p. 39［フィールド］タブのドキュメント設定 を参照してください。

テキストの単位: テキスト フィールドが [**ドキュメントへのパス名**] を示すよう指定し、ドキュメント タイプに [**フル テキスト**] を選択した場合にのみ使用できます。次の実行モードを選択します。

- **ドキュメント モード**: 通信社からの記事など、短く意味的に同質のドキュメントに使用します。
- **パラグラフ モード**: Web ページおよびタグのないドキュメントに使用します。抽出プロセスでは、内部タグやシンタックスなどの特徴を利用して、ドキュメントを意味的に分割します。このモードを選択すると、パラグラフごとにスコアリングが適用されます。そのため、**リンゴ**および**オレンジ**が同じパラグラフで見つかった場合にのみ、たとえば規則 **リンゴ & オレンジ** が当てはまります。

パラグラフ モードの設定: テキスト フィールドが [**ドキュメントへのパス名**] を示すよう指定し、テキストの単位オプションを [**パラグラフ モード**] に設定した場合にのみ使用できます。抽出で使用する文字の閾値を指定します。実際のサイズは、最も近いピリオドに丸められます。ドキュメント コレクションのテキストから作成される単語の関連性を典型とするには、抽出サイズが小さすぎないよう指定します。

- **最小**: 抽出で使用する文字の最小数を指定します。
- **最大**: 抽出で使用する文字の最大数を指定します。

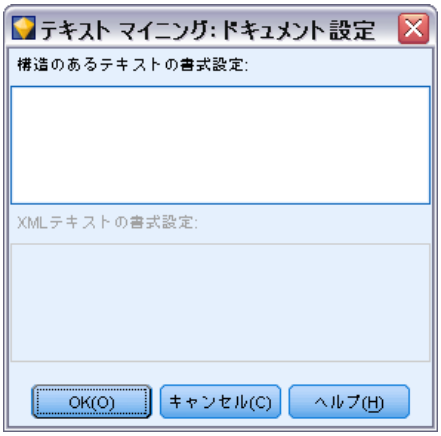
入力エンコード: テキスト フィールドが [**ドキュメントへのパス名**] を示すよう指定した場合にのみ使用できます。デフォルトのテキスト エンコードを指定します。日本語以外のすべての言語について、指定された、または認識されたエンコードから ISO-8859-1 への変換が行われます。そのため、別のエンコードが指定されている場合であっても、抽出エンジンは処理

前にテキストを ISO-8859-1 に変換します。ISO-8859-1 エンコード定義に一致しない文字は、スペースに変換されます。日本語テキストの場合、SHIFT_JIS、EUC_JP、UTF-8、または ISO-2022-JP のエンコード オプションから 1 つ選択できます。

データ区分モード: データ区分モードを使用して、データ型ノードの設定に基づいて区分するか、別のデータ区分を選択するかを選択します。データ区分によって、データを学習サンプルおよび検定サンプルに分割します。

[フィールド] タブのドキュメント設定

図 3-2
[ドキュメント設定] ダイアログ ボックス



構造のあるテキストの書式設定

構造のあるデータがある、またはテキストの処理方法に規則を強制したいために抽出プロセスの全部または一部をスキップしたい場合、[ドキュメント設定] ダイアログ ボックスの **[構造のあるテキストの書式設定]** セクションで **[構造のあるテキスト]** ドキュメント タイプのオプションを使用し、テキストを含むフィールドまたはタグを宣言します。抽出されたキーワードは、派生したフィールドまたはタグ（下位タグ）に含まれるテキストからのみ派生します。宣言されていないフィールドまたはタグは無視されます。

言語処理が必要ではなく、また言語学的抽出エンジンが明示的な宣言に置き換えることができる場合があります。キーワードのフィールドがセミコロン（;）やカンマ（,）のような区切り文字で区切られている参考文献ファイルでは、2 つの区切り文字の間の文字列を抽出すれば十分です。そのため、すべての抽出プロセスを中止し、代わりに特別な処理規則を定義して、キーワードの区切り文字を宣言し、タイプを抽出テキストに割り当てるか、抽出に最小出現頻度を設定します。

構造のあるテキストの要素を宣言する場合、次の規則を使用します。

- 1 行ごとに宣言できるのは、1 つのフィールド、タグ、または要素だけです。それらはデータ内にある必要はありません。
- 宣言では、大文字小文字を区別します。
- <title id="1234"> のような属性のあるタグを宣言し、すべての変異形、またはこの場合すべての ID を追加したい場合、属性および終わりの山カッコ（>）を除いたタグ（<title）を追加します。
- フィールド名またはタグ名の後にコロンを追加して、構造のあるテキストを示します。このコロンは、author: または <place>: のように、フィールドまたはタグの直後、そして区切り文字、タイプ、または出現頻度値の前に追加してください。
- 複数のキーワードがフィールドまたはタグに含まれ、また区切り文字を使用して各キーワードを指定することを示すには、author:, または <section>:; のように、コロンの後に区切り文字を指定します。
- タイプをタグの内容に割り当てるには、author:,Person または <place>:;Location のように、コロンおよび区切り文字の後にタイプ名を指定します。リソース エディタに表示されるとおりの名前を使用してタイプを宣言します。
- フィールドまたはタグの最小出現頻度を定義するには、author:,Person1 または <place>:;Location5 のように、行の最後に数字を指定します。n は、定義する出現頻度を示し、フィールドまたはタグ内のキーワードは、抽出するドキュメントまたはレコードのセット全体で少なくとも n 回出現する必要があります。また、区切り文字を定義する必要もあります。
- コロンを含むタグがある場合、コロンの前にバックスラッシュを追加し、宣言が無視されないようにします。たとえば、<topic:source> というフィールドがある場合、<topic\:source> のように入力します。

シンタックスを説明するために、次のような、繰り返し出現する参考文献のフィールドがあると仮定しましょう。

```
author:Morel, Kawashima
abstract:この記事はフィールドの処理方法について説明しています。
publication:テキスト マイニング マニュアル
datepub:2010 年 3 月
```

この例では、抽出プロセスが作者および要約に焦点を当てるようにしたいが、残りの内容は無視したい場合、次のフィールドのみを宣言します。

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

この例の場合、author:,Person1 フィールドの宣言は、フィールドの内容についての言語処理が中断したことを示しています。代わりに、作者フィールドで、名前を区切るためにカンマを使用して複数の名前を含むことを指定します。そしてこれらの名前は「人名」タイプに割り当て、またドキュメントまたはレコードのセット全体で 1 回以上名前が出現した場合

は抽出する必要があることを宣言しています。フィールド abstract: が他に宣言せずに表示されているため、抽出時にフィールドがスキャンされ、標準的な言語処理およびタイプ指定が行われます。

XML テキストの書式設定

抽出プロセスを特定の XML タグ内のテキストのみに制限したい場合、[ドキュメント設定] ダイアログ ボックスの **[XML テキストの書式設定]** セクションで **[XML テキスト]** ドキュメント タイプのオプションを使用し、そのテキストを含むタグを宣言します。抽出されたキーワードは、これらのタグまたは下位タグに含まれるテキストからのみ派生します。

重要: 抽出プロセスをスキップしてキーワードの区切り文字に規則を指定する場合、タイプを抽出したテキストに割り当てるか抽出したキーワードに出現頻度を指定し、次で説明する **[構造のあるテキスト]** オプションを使用します。

XML テキストの書式設定のタグを宣言する場合、次の規則を使用します。

- 1 行ごとに宣言できるのは、1 つの XML タグだけです。
- タグの要素は、大文字小文字を区別します。
- タグに <title id="1234"> のような属性があり、すべての変異形、またはこの場合すべての ID を追加したい場合、属性および終わりの山カッコ（>）を除いたタグ（<title）を追加します。

シンタックスを説明するために、次のような XML ドキュメントがあると仮定しましょう。

```
<section>交通規則
     <title id="01234">交通信号</title>
     <p>道路標識は役に立ちます。</p>
</section>
<p>規則について学ぶことは重要です。</p>
```

この例の場合、次のタグを宣言します。

```
<section>
<title
```

この例では、タグ <section> を宣言しているため、このタグのテキストと入れ子になっているタグ、**交通信号**および**道路標識は役立ちます**が抽出プロセスでスキャンされます。ただし、タグ<p> が明示的に宣言されず、宣言されたタグの入れ子にもなっていないため、**規則について学ぶことは重要です。**は無視されます。

テキスト マイニング ノード:[モデル] タブ

[モデル] タブを使用して、ノード出力の作成方法と一般的なモデル設定を指定します。

図 3-3
[テキスト マイニング ノード] ダイアログ ボックス:[モデル] タブ

設定できるパラメータを次に示します。

モデル名： ターゲットまたは ID フィールド（その指定がない場合はモデルタイプ）に基づいてモデル名を生成、またはカスタム名を指定することができます。

区分されたデータを使用: データ区分フィールドが定義されている場合、このオプションでは学習用データ区分のデータのみを使用して、モデルを構築します。

ビルド モード: このテキスト マイニング ノードでストリームが実行される場合にモデル ナゲットが作成される方法を指定します。また、コンセプトを抽出、カテゴリを作成、および言語リソースを調整するだけでなく、テキスト リンク分析を実行してクラスタを検証できる**[インタラクティブに作成]**モードを使用して、より実践的な探索的アプローチを使用できます。

- **インタラクティブに作成:** ストリームを実行する場合、このオプションでインタラクティブ ワークベンチが起動します。インタラクティブ ワークベンチでは、コンセプトやパターンを抽出し、抽出結果を検証および調整、カテゴリを作成および調整、言語リソース（テンプレー

ト、類義語、タイプ、ライブラリなど）を調整、カテゴリ モデル ナゲットを作成することができます。 詳細は、 p. 43 インタラクティブに作成 を参照してください。

- **直接生成:** ストリームの実行時、モデルが自動的に作成され、[モデル] パレットに追加されることを指示します。インタラクティブ ワークベンチとは異なり、ノードで定義された設定のほか、実行時に必要な追加の操作はありません。このオプションを選択すると、作成するモデルのタイプを定義できるモデル特有のオプションが表示されます。 詳細は、 p. 46 直接生成 を参照してください。

リソースのコピー元: テキスト マイニング時、抽出は、[エキスパート] タブの設定だけではく、言語リソースに基づいて行われます。これらのリソースは、抽出時のテキストの処理方法の基本として機能し、コンセプト、タイプ、そしてときには TLA パターンを取得します。リソース テンプレートまたはテキスト分析パッケージのいずれかから、リソースをテキスト マイニングモデル作成ノードにコピーできます。いずれかを選択して **[読み込み]** をクリックし、リソースのコピー元となるパッケージまたはテンプレートを定義します。読み込んでいるときに、リソースのコピーがノードに保存されます。そのため、更新されたテンプレートまたは TAP を使用したい場合、ここで、インタラクティブ ワークベンチ セッションで再読み込みを行う必要があります。リソースがコピーされ、読み込まれた日時が、ノードに表示されます。 詳細は、 p. 46 テンプレートおよび TAP からのリソースのコピー を参照してください。

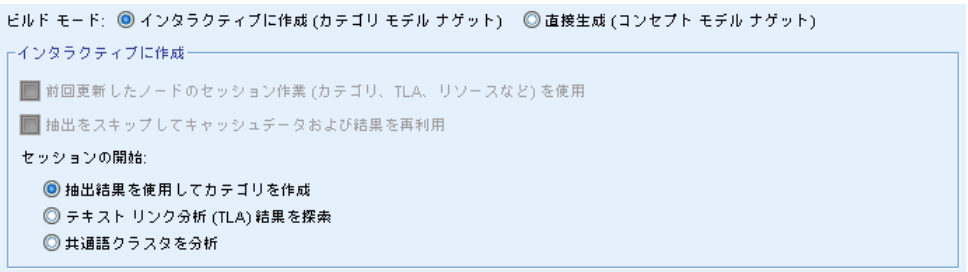
テキストの言語: マイニングされるテキストの言語を示します。ノードでコピーされたリソースが、表示される言語オプションを制御します。リソースを調整した言語を選択するか、**[すべて]** オプションを選択できます。テキスト データの正確な言語を指定することをお勧めしますが、不明な場合は、**[すべて]** オプションを選択できます。**[すべて]** は、日本語テキストには使用できません。自動言語認識を使用してすべてのドキュメントをスキャンして記録し、テキスト言語を最初に特定するため、この **[すべて]** オプションを選択すると、実行時間が長くなります。このオプションを使用して、サポートされライセンス許可された言語のすべてのレコードまたはドキュメントが、言語に適した内部辞書を使用して、抽出エンジンによって読み込まれます。 詳細は、 18 章 p. 366 言語の識別子 を参照してください。 現在使用できないサポート言語のライセンス購入については、営業担当者に連絡してください。

インタラクティブに作成

テキスト マイニング モデル作成ノードの［モデル］タブで、モデル ナゲットのビルド モードを選択できます。**[インタラクティブに作成]** を選択すると、ストリーム実行時にインタラクティブ インターフェイスが開きます。このインタラクティブ ワークベンチで、次の作業を行えます。

- コンセプトおよびタイプなどの抽出結果を抽出および検証し、テキスト データで目立つキーワードを探索します。
- さまざまな手法を使用してコンセプト、タイプ、TLA パターンおよび規則からカテゴリを作成および展開して、ドキュメントおよびレコードをこれらのカテゴリにスコアリングできるようにします。
- 言語リソース（リソース テンプレート、ライブラリ、辞書、類義語など）を調整し、コンセプトが抽出、検証、調整されるインタラクティブ プロセスによって結果が改善されるようにします。
- テキスト リンク分析（TLA）を実行し、検出された TLA パターンを使用して、よりよいカテゴリ モデル ナゲットを作成します。テキスト リンク分析ノードには、同じ探索オプションやモデル作成機能はありません。
- クラスタを生成して、[検証] パネルで新しい関係性を探索し、コンセプト、タイプ、パターン、およびカテゴリの間の関係性を検証します。
- IBM® SPSS® Modeler の [モデル] パレットに調整済みカテゴリ モデル ナゲットを生成し、それらを他のストリームで使用します。

図 3-4
インタラクティブに作成するための [モデル] タブのオプション



前回更新したノードのセッション作業（カテゴリ、TLA、リソースなど）を使用: インタラクティブ ワークベンチ セッションで作業している場合、セッション データ（抽出パラメータ、リソース、カテゴリ定義など）でノードを更新できます。[**セッション作業を使用**] オプションを選択すると、保存されたセッション データを使用してインタラクティブ ワークベンチを再起動できます。初めてこのノードを使用する場合、セッション データが保存されていないため、このオプションは無効になります。このオプションを使用できるセッション データでのノードの更新方法については、「モデル作成ノードの更新および保存」（ p. 148 ）を参照してください。.

このオプションを使用してセッションを起動すると、前回インタラクティブ ワークベンチ セッションからノードを更新した時の抽出設定、カテゴリ、リソースおよびその他の作業が、次回セッションを起動したときに使用できます。保存したセッションデータはこのオプションで使用されるため、下のテンプレートからコピーしたリソースなど、特定の内容やその他のタブが無効となり、無視されます。このオプションを使用せずにセッ

ションを起動すると、定義されたとおりのノードの内容のみが使用されます。つまり、ワークベンチで実行した前回の作業は使用できなくなります。

注:抽出結果を［**セッション作業を使用**］オプションを使用してキャッシュした後、ストリームの入力ノードを変更する際、抽出結果を更新する場合にインタラクティブ ワークベンチ セッションが起動したら新しい抽出を実行する必要があります。

抽出をスキップしてキャッシュデータおよび結果を再利用: インタラクティブ ワークベンチ セッションで、キャッシュされた抽出結果およびデータを再利用できます。このオプションは、セッションが起動したときに実行されるまったく新しい抽出を待機するのではなく、時間を節約して抽出結果を再利用したい場合に特に役立ちます。このオプションを使用するには、インタラクティブ ワークベンチ セッションからこのノードを事前に更新し、オプション［**セッション作業を保存し、再利用するために抽出結果とともにテキストデータをキャッシュに格納する**］を選択する必要があります。このオプションを使用できるセッション データでのノードの更新方法については、「モデル作成ノードの更新および保存」（ p. 148 ）を参照してください。.

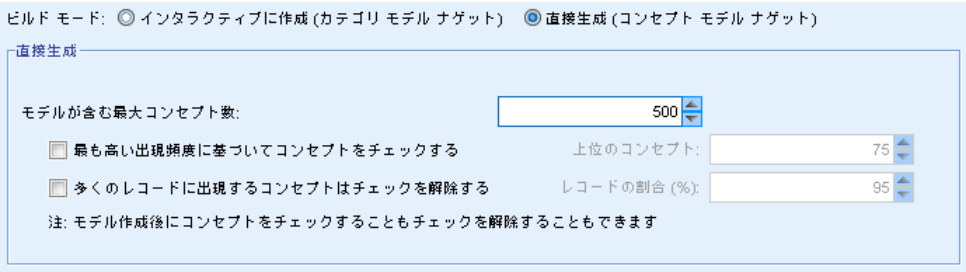
セッションの開始: インタラクティブ ワークベンチ セッションの起動時に最初に表示したいビューおよび実行したい操作を示します。開始時のビューに関係なく、セッション内の任意のビューに一度切り替えることができます。

- **抽出結果を使用してカテゴリを作成:** カテゴリとコンセプト ビューでインタラクティブ ワークベンチを起動し、必要に応じて抽出を実行します。このビューでは、カテゴリを作成し、カテゴリ モデルを生成できます。また、別のビューに切り替えることもできます。 詳細は、 8 章 p. 127 インタラクティブ ワークベンチ モード を参照してください。
- **テキスト リンク分析 (TLA) 結果を探索:** まず、意見またはテキスト リンク分析ビューの他のリンクなど、テキスト内のコンセプト間の関係性を抽出および特定します。このオプションを使用して結果を抽出するには、TLA パターン規則を含むテンプレートまたはテキスト分析パッケージを選択する必要があります。より大きいデータセットを扱う場合、TLA 抽出に時間がかかる場合があります。この場合、上流でサンプル ノードの使用を検討する必要があります。 詳細は、 12 章 p. 268 テキスト リンク分析の検証 を参照してください。
- **共通語クラスタを分析:** このオプションはクラスタ ビューで起動し、古い抽出結果を更新します。このビューで、共通語クラスタ分析を実行し、一連のクラスタを作成できます。共通語クラスタリングは、まず指定されたレコードまたはドキュメントの共起に基づいて 2 つのコンセプト間のリンク値の強度を評価し、最後に強くリンクしたコンセプトをクラスタにグループ化するプロセスです。 詳細は、 8 章 p. 127 インタラクティブ ワークベンチ モード を参照してください。

直接生成

テキスト マイニング モデル作成ノードの［モデル］タブで、モデル ナゲットのビルド モードを選択できます。**［直接生成］**を選択すると、ノードでオプションを設定し、ストリームを実行できます。出力はコンセプト モデル ナゲットで、［モデル］パレットに直接投入されます。インタラクティブ ワークベンチとは異なり、ノードのオプションで定義された出現頻度設定のほか、実行時に必要な追加の操作はありません。

図 3-5
［直接生成］の［モデル］タブのオプション



モデルが含む最大コンセプト数: 自動的にモデルを作成する場合（非インタラクティブ）にのみ適用され、コンセプト モデルを作成することを示します。また、このモデルには、指定した数以下ののコンセプトが含まれることも示します。

- **最も高い頻度に基づいてコンセプトをチェックする。上位のコンセプト:**チェックされるコンセプトの数です。最も頻度の高いコンセプトからチェックします。ここで、頻度は、ドキュメント/レコードのセット全体の中でコンセプト（およびすべての基本キーワード）が出現する回数を示します。レコード内にコンセプトが複数回出現する場合があるため、この数値がレコード数を上回る場合があります。
- **多くのレコードに出現するコンセプトはチェックを解除する。レコードの割合:**レコード数の割合が、指定した数を上回るコンセプトのチェックを解除します。このオプションは、テキストまたはすべてのレコードで頻繁に出現するが、分析においては重要でないコンセプトを除外する場合に役立ちます。

テンプレートおよび TAP からのリソースのコピー

テキスト マイニング時、抽出は、［エキスパート］タブの設定だけではく、言語リソースに基づいて行われます。これらのリソースは、抽出時のテキストの処理方法の基本として機能し、コンセプト、タイプ、そしてときには TLA パターンを取得します。リソースをリソース テンプレートからこのノードにコピーすることができます。また、テキスト マイニング ノードを使用している場合は、テキスト分析パッケージ（TAP）を選択することもできます。

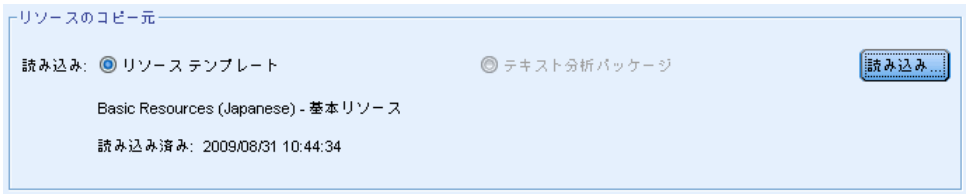
デフォルトでは、ノードを領域に追加すると、製品のライセンスされた言語の基本テンプレートからノードに、リソースがコピーされます。複数の言語のライセンスがある場合、最初に選択された言語を使用して、自動的に読み込むテンプレートを決定します。

読み込んでいるときに、選択したリソースのコピーがノードに保存されます。テンプレートまたは TAP の内容のみがコピーされますが、テンプレートまたは TAP 自体はノードにリンクしません。つまり、このテンプレートまたは TAP を後で更新すると、これらの更新が自動的にノードで使用できることはありません。ノードに読み込まれたリソースは、テンプレートまたは TAP のコピーが再読み込みされないかぎり、またはテキスト マイニング ノードを更新して［**セッション作業を使用**］オプションを選択しないかぎり、かならず使用されます。［**セッション作業を使用**］の詳細は、このトピックを参照してください。

テンプレートまたは TAP を選択する場合、テキスト データと言語が同じものを選択してください。ライセンスが付与された言語でのみテンプレートまたは TAP を使用できます。テキスト リンク分析を実行したい場合、TLA パターンを含むテンプレートを選択する必要があります。テンプレートに TLA パターンが含まれている場合、［リソース テンプレートを読み込む］ダイアログ ボックスの［TLA］列にアイコンが表示されます。

注:TAP をテキスト リンク分析ノードに読み込むことはできません。

図 3-6
テキスト マイニング ノード、［モデル］タブ:リソースをノードにコピーするオプション



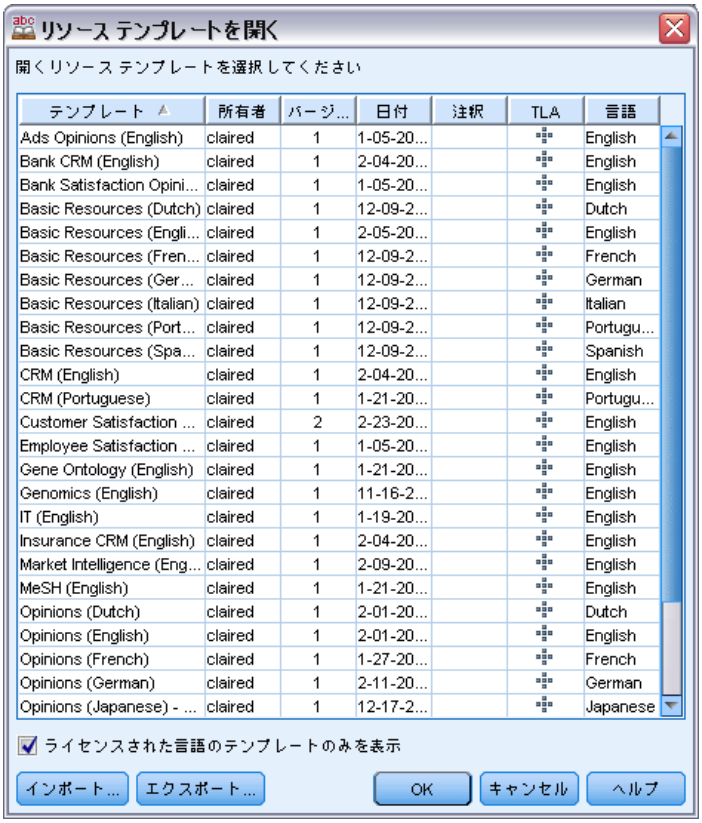
リソース テンプレート

リソース テンプレートは、特定のドメインまたは使用向けに調整された、事前定義済みライブラリおよび詳細な言語リソースおよび非言語リソースです。テキスト マイニング モデル作成ノードでは、ノードをストリームに追加するときには基本テンプレートのリソースのコピーが既にノードに読み込まれています。ただし、テンプレートを変更するか、［**リソース テンプレート**］または［**テキスト分析パッケージ**］を選択して［**読み込み**］をクリックし、テキスト分析パッケージを読み込むことができます。テンプレートの場合、［リソース テンプレートを読み込む］ダイアログ ボックスでテンプレートを選択できます。

注:リストに必要なテンプレートが表示されないにもかかわらず、コンピュータのコピーがエクスポートされている場合、今すぐインポートできます。このダイアログ ボックスからエクスポートして、他のユーザーと共

有することもできます。 詳細は、 15 章 p.309 テンプレートのインポートおよびエクスポート を参照してください。

図 3-7
[リソース テンプレートを読み込む] ダイアログ ボックス



テキスト分析パッケージ (TAP)

テキスト分析パッケージ (TAP) は、1 つまたは複数の事前定義されたカテゴリのセットとまとめられた、ライブラリと高度な言語リソースおよび非言語リソースの事前定義されたセットです。IBM® SPSS® Modeler Text Analytics では、英語テキスト向けおよび日本語テキスト向けの事前に作成された TAP がいくつか用意され、それぞれか特定のドメイン向けに調整されています。これらの TAP を編集できませんが、それらを使用してカテゴリ モデル作成を開始できます。インタラクティブ セッションで独自の TAP を作成することもできます。詳細は、 10 章 p.248 テキスト分析パッケージの読み込み を参照してください。注:日本語テキスト展開は IBM® SPSS® Modeler Premiumで利用可能です。

注:TAP をテキスト リンク分析ノードに読み込むことはできません。

[セッション作業を使用] オプションの使用 ([モデル] タブ)

[モデル] タブでリソースがノードにコピーされますが、後からインタラクティブ セッションでリソースを変更し、これら最近の変更でテキスト マイニング モデル作成ノードの更新が必要な場合があります。この場合、テキスト マイニング モデル作成ノードの [モデル] タブの **[セッション作業を使用]** オプションを選択します。

[セッション作業を使用] を選択すると、ノードの **[読み込み]** ボタンが無効となり、インタラクティブ ワークベンチのこれらのリソースが、以前ここで読み込まれたリソースの代わりに使用されることを示します。

[セッション作業を使用] オプションで選択したリソースに変更を行うには、リソース エディタ ビューを使用して、インタラクティブ ワークベンチ セッション内で直接リソースを編集または切り替えることができます。 詳細は、 15 章 p.306 読み込み後のノード リソースの更新 を参照してください。

テキスト マイニング ノード:[エキスパート] タブ

[エキスパート] タブには、テキストの抽出方法および処理方法に影響を与える高度なパラメータがあります。このダイアログ ボックスのパラメータは、抽出プロセスの基本的な操作、そしていくつかの高度な操作を制御します。ただし、使用できるオプションの部分のみを示します。また、抽出結果に影響を与える言語リソースやオプションも数多くあり、[モデル] タブで選択するリソース テンプレートによって制御します。 詳細は、p.41 テキスト マイニング ノード:[モデル] タブ を参照してください。

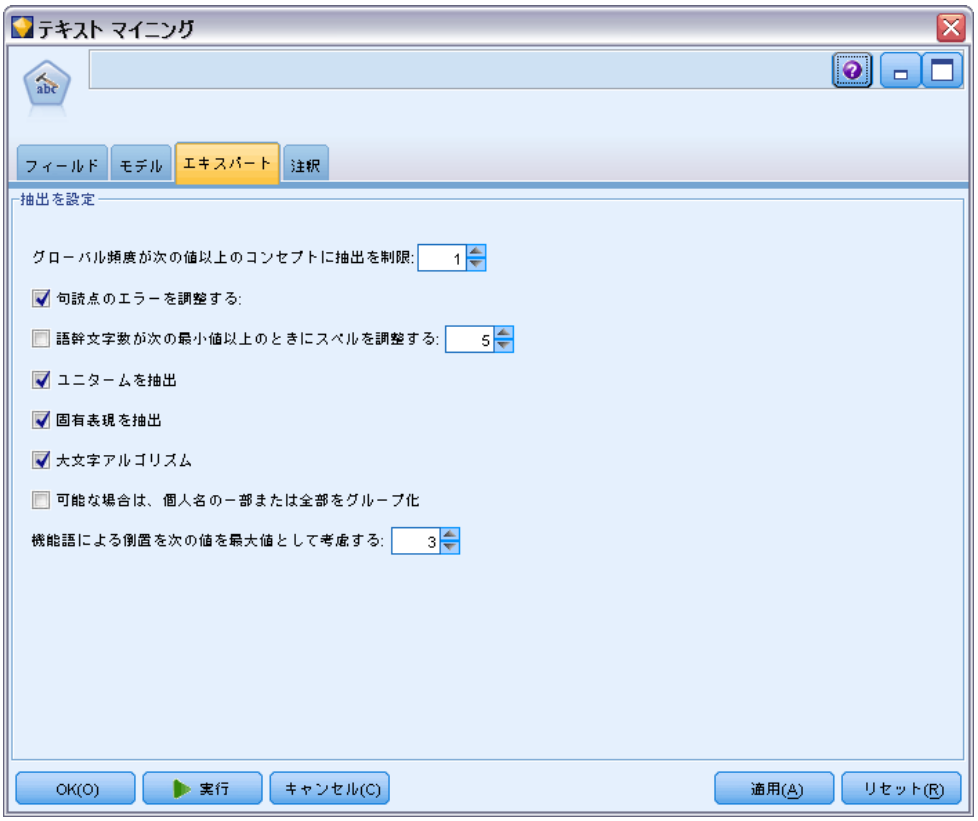
注:[モデル] タブで保存されたインタラクティブ ワークベンチ情報に基づいて **[インタラクティブに作成]** モードを選択した場合、このタブ全体が無効になります。この場合、抽出設定は、最近保存されたワークベンチ セッションから取得されます。

オランダ語、英語、フランス語、ドイツ語、イタリア語、ポルトガル語、スペイン語のテキストの場合

英語、スペイン語、フランス語、ドイツ語など、日本語以外の言語を抽出する場合、次のパラメータを設定できます。

注:日本語テキストのエキスパート 設定に関する詳細は、このトピックの後の項目を参照してください。日本語テキスト展開はIBM® SPSS® Modeler Premiumで利用可能です。

図 3-8
[テキスト マイニング ノード] ダイアログ ボックス:[エキスパート] タブ



グローバル頻度が次の値以上のコンセプトに抽出を制限: 抽出するために、単語または句が出現する必要がある最低限の回数を指定します。値に 5 を指定すると、抽出するこれらの単語または句が、レコードまたはドキュメントのセット全体で少なくとも 5 回出現するよう、制限します。

この制約を変更すると、抽出結果、つまり作成されるカテゴリに大きな違いが生じる場合があります。あるレストランのデータを処理し、このオプションの制約に1より大きい値を設定しないものとします。この場合、抽出結果がピザ (1)，薄いピザ (2)，ほうれん草のピザ (2)、および好きなピザ (2) となります。ただし、抽出のグローバル出現頻度を 5 以上に設定して抽出すると、これらのコンセプトのうち 3 つが取得されなくなります。代わりに、ピザが最も簡単な形で、この単語は考えられる候補として既に存在するため、ピザ (7) が取得されます。また、残りのテキストにピザという単語を含む他の句があるかどうかによって、7より大きい出現頻度がある場合があります。また、ほうれん草のピザがカテゴリの記述子である場合、すべてのレコードをキャプチャする代わりに、記述子としてピザの追加が必要な場合があります。このため、カテゴリが既に作成されている場合は、注意してこの制約を変更してください。

これは抽出のみの機能であるということを留意して下さい。テンプレートに用語が含まれているなら（たいていの場合含まれています）、テンプレートの用語がテキスト内に見つかるなら、その用語は頻度に関係なくインデックスされます。

コアライブラリー内の<Location>タイプの下にある、「ロサンゼルス」を含む基本リソースをテンプレートとして使用するとします。ドキュメント内に「ロサンゼルス」が一度しか含まれていない場合、「ロサンゼルス」はコンセプトリストの一部となります。これを防ぐには、最低でも**最低でも[n]回の頻度で起こるコンセプトに抽出を限定する**フィールドに入力したのと同じ値で起こるコンセプトを表示するフィルターを設定する必要があります。

句読点のエラーを調整する: 抽出時に句読点エラー（不適切な使用方法など）を含むテキストを一時的に正規化し、コンセプトの抽出可能性を向上させます。自由記述式アンケートの回答、電子メール、CRM データなど、テキストが短く品質が悪い場合、またはテキストに略語が多く含まれている場合に特に役立ちます。

語幹文字数が次の最小値以上のときにスペルを調整する: Fuzzy Grouping の手法を適用し、共通してスペルミスのある単語またはスペルの近い単語を 1 つのコンセプトにグループ化できるようにします。Fuzzy Grouping アルゴリズムでは、最初の母音を除くすべての母音を一時的に抜き取った後抽出した単語から 2 つ/3 つの子音を抜き取り、それらを比較して、それらが同じで modeling と modelling が同じグループに分けられるかどうかを確認します。ただし、各キーワードが <Unknown> タイプを除いて、別のタイプに割り当てられた場合、Fuzzy Grouping 手法は適用されません。

Fuzzy Grouping を使用する前に必要な、語幹文字数の制限を定義することもできます。キーワード内の語幹文字数は、すべての文字を合計し、活用語尾、複合語キーワードの場合は区切り文字および前置詞を形成する文字を差し引いて計算します。たとえば、キーワード exercises の語幹文字数は「exercise」という形式で 8 文字と数えられます。語末の s は活用語尾（複数形）であるためです。同様に、apple sauce の語幹文字は 10 文字（「apple sauce」）、そして manufacturing of cars の語幹文字は 16 文字（「manufacturing car」）となります。この算出方法は、Fuzzy Grouping を適用するべきかどうかを確認するためにのみ使用されますが、単語がどのように一致するかについては影響を与えません。

注:特定の単語が後で不適切にグループ化されていることが分かった場合、[アドバンス リソース] タブの Fuzzy Grouping: **例外** セクションで 明示的に宣言することによって、単語のペアをこの手法から除外できます。詳細は、 18 章 p.357 Fuzzy Grouping を参照してください。

ユニタームを抽出: 単語が複合語の一部でない限り、または名詞、またはスピーチ内の認識できない部分である場合、このオプションは単一の単語（ユニターム）を抽出します。

固有表現を抽出: 電話番号、セキュリティ番号、時間、日付、通貨、数字、パーセント、電子メールアドレス、HTTP アドレスなどの固有表現を抽出します。[アドバンス リソース] タブの **[固有表現: 設定]** セクションで、特定の種類の固有表現を追加したり除外したりできます。不要なエンティティを無効にすることにより、抽出エンジンは処理時間を節約できます。詳細は、 18 章 p. 362 構成 を参照してください。

大文字アルゴリズム: キーワードの最初の文字が大文字である場合、組み込み辞書にない単純キーワードおよび複合キーワードを抽出します。このオプションには、最も適切な名詞を抽出するのに優れた方法があります。

可能な場合は、個人名の一部または全部をグループ化: テキスト内で別々の形式で同時に出現する名前をグループ化します。名前はテキストの始めでは完全な形式で、後は短い形式でのみ参照されるため、この機能が役立ちます。このオプションでは、タイプが <Unknown> のユニタームが、タイプ <Person> の複合キーワードの最後の単語に一致するようにします。たとえば、doe があり、最初タイプが <Unknown> である場合、抽出エンジンは、<Person> タイプの複合キーワードに最後の単語として doe が含まれているかどうか（例: john doe）を確認します。ほとんどがユニタームとして抽出されることがないため、人の名前に適用されることはありません。

機能語による倒置を次の値を最大値として考慮する: 倒置手法を適用する場合に指定されている場合がある非機能的単語の最大数を指定します。この倒置手法では、活用語尾に関係なく、含まれる非機能的単語（of や the など）によってお互いに異なる類似した句をグループ化します。たとえば、この値を最大 2 単語に設定し、company officials および officials of the company が抽出されたとします。この場合、両方の抽出キーワードは、of the が無視されると同じであるとみなされるため、最終コンセプト リストに共にグループ化されます。

注:テキスト リンク分析結果の抽出を有効にするには、**[テキスト リンク分析結果を探索]** オプションでセッションを開始し、また TLA 定義を含むリソースも選択する必要があります。[抽出設定] ダイアログで、インタラクティブ ワークベンチ セッション中に後から TLA 結果を抽出できます。 詳細は、 9 章 p. 156 データの抽出 を参照してください。

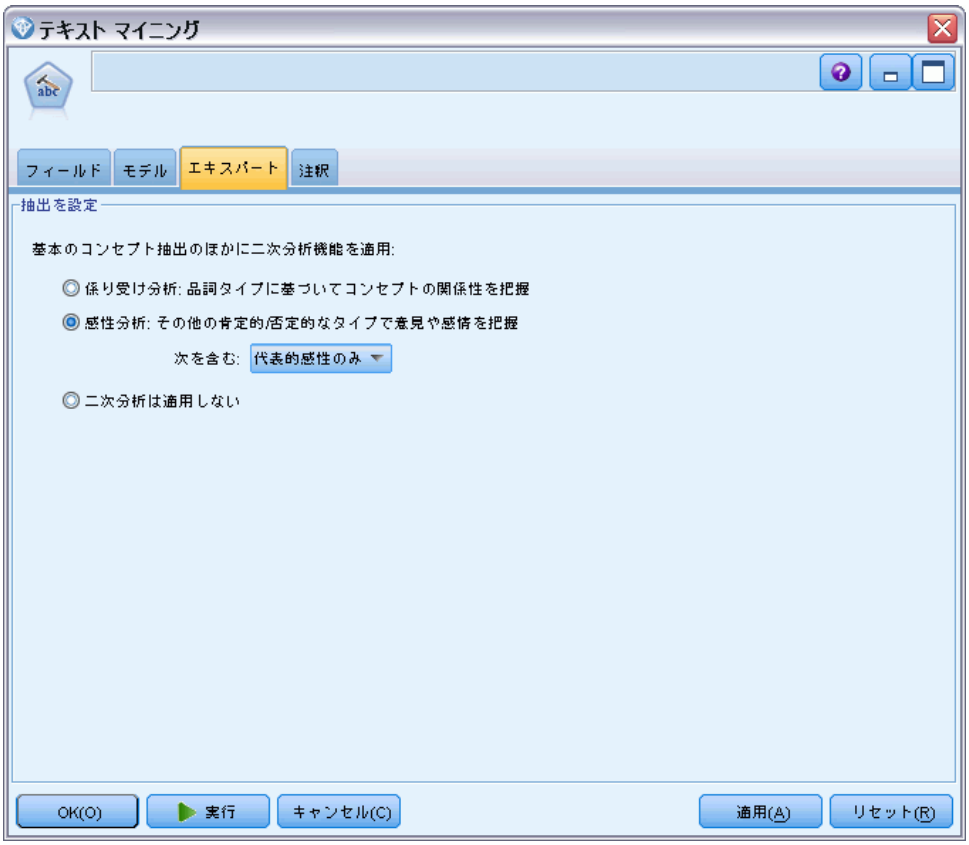
日本語テキストの場合

注:日本語テキスト展開はSPSS Modeler Premiumで利用可能です。

抽出プロセスにいくつか異なる点があるため、日本語テキストのダイアログには異なるオプションがあります。日本語テキストを処理するためには、このノードの［モデル］タブで日本語向けに調整されたテンプレートまたはテキスト分析パッケージを選択する必要があります。 詳細は、 p. 46 テンプレートおよび TAP からのリソースのコピー を参照してください。

設定できるパラメータを次に示します。

図 3-9
[テキスト マイニング ノード] ダイアログ ボックス:[エキスパート] タブ (日本語テキスト)



注:日本語テキスト展開はSPSS Modeler Premiumで利用可能です。

二次分析: 抽出が開始したとき、基本キーワード抽出が、タイプのデフォルト セットを使用して行われます。詳細は、 A 付録 p. 412 日本語テキストで使用できるタイプ を参照してください。 ただし、二次分析機能を選択すると、抽出機能にコンセプトの一部として助詞や助動詞が含まれているため、より多くの詳細なコンセプトを取得することができます。たとえば、「肩の荷が下りた」という文があるとします。この例では基本キーワード抽出は各コンセプトを個別に抽出します。例:**肩** (肩)、**荷** (重量)、**下りる** (上げ)、しかしこれらの単語の間の関係性は抽出されません。しかし、感性分析を適用すると、コンセプト =**肩の荷が下りた** というより高次の意味をもつコンセプトが抽出され、<**良い-安心**> という感性タイプとして割り当てられます。なお感性分析については、このほかに多くの感性タイプがあります。さらに、二次分析機能を選択すると、テキスト リンク分析結果も生成できます。

注:二次分析を呼び出すと、抽出プロセスにより時間がかかります。詳細は、 A 付録 p. 404 二次抽出の手順 を参照してください。

- **係り受け解析**:このオプションを選択すると、キーワード、およびキーワードに助詞等を加えた語を、基本の品詞タイプのコンセプトとして抽出します。また係り受けテキスト リンク分析 (TLA) に基づいたパターン結果を出力します。
- **感性分析**:この分析機能を選択すると、追加の抽出コンセプト、および可能な場合は、TLA パターン結果の抽出が行われます。基本タイプのほか、嬉しい、吉報、幸運、安心、幸福など 80 を超える感性タイプも利用できます。これらのタイプを使用して、感情、感性、意見の表現のテキストでコンセプトおよびパターンを検出します。感性分析に対するフォーカスを指示するオプションは３つあります：**全ての感性**、**代表的な感性のみ**、それと**結論のみ**。
- **二次分析は適用しない**:すべての二次分析をオフにします。TLA 結果の取得には二次分析が必要であるため、[モデル]タブで オプション [**テキスト リンク分析結果を探索**] が選択されている場合、このオプションは非表示になります。このオプションを後で選択し、[**テキスト リンク分析 (TLA) 結果を探索**] を選択すると、ストリーム実行時にエラーが発生します。

時間を削減する上流のサンプリング

大きいデータがある場合、特にインタラクティブ ワークベンチ セッションを使用している場合、処理時間は数分から数時間かかる場合があります。データのサイズが大きいほど、抽出およびカテゴリ化処理の時間がより長くなります。より効率的に作業するために、テキスト マイニング ノードの上流の IBM® SPSS® Modeler のサンプル ノードを追加します。このサンプル ノードを使用して、ドキュメントまたはレコードの小さい部分集合で無作為サンプルを取得し、最初のいくつかの通過を実行します。

小さいサンプルは、リソースの編集方法を決定し、すべてではなくても多くのカテゴリを作成するのに適しています。小さいデータセットで実行し、結果が適切であれば、カテゴリ作成の手法をデータのセット全体に適用できます。作成したカテゴリに適合しないドキュメントまたはレコードを検出し、必要に応じて調整を行うことができます。

注:サンプル ノードは、標準的な SPSS Modeler ノードです。

ストリーム内のテキスト マイニング ノードの使用

テキストマイニング モデル作成ノードを使用して、データにアクセスし、ストリーム内のコンセプトを抽出します。データベースノードなどのように、データにアクセスするにはどのソースノードも使用できます。ファイルノード、ウェブフィードノード、またはフィクスドファイルノード。外部ドキュメント内のテキストの場合、ファイル リスト ノードを使用できます。

例 1:コンセプト モデル ナゲットを直接作成するファイル リスト ノードおよびテキスト マイニング ノード

次の例では、テキスト マイニング モデル作成ノードと共にファイル リスト ノードを使用して、コンセプト モデル ナゲットを生成する方法が示されています。ファイル リスト ノード使用の詳細は、「2 章」を参照してください。

図 3-10
ストリームの例:テキスト マイニング ノードに接続したファイル リスト ノード

- **ファイル リスト ノード ([設定] タブ):** まず、このノードをストリームに追加して、テキスト ドキュメントが保存されている場所を指定しました。テキスト マイニングを実行するすべてのドキュメントを含むディレクトリを選択しました。

図 3-11
[ファイル リスト ノード] ダイアログ ボックス[設定] タブ

- **テキスト マイニング ノード ([フィールド] タブ):** 次に、テキスト マイニング ノードを ファイル リスト ノードに追加して接続しました。このノードで、入力形式、リソース テンプレート、および出力形式を定義しました。ファイル リストノードから作成されたフィールド名を選択し、テキスト フィー

ルドが [ドキュメントへのパス名] やその他の設定を表示するオプションを選択しました。詳細は、 p. 54 ストリーム内のテキスト マイニング ノードの使用 を参照してください。

図 3-12
[テキスト マイニング ノード] ダイアログ ボックス:[フィールド] タブ

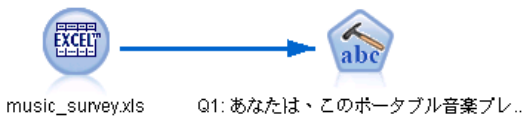
- **テキスト マイニング ノード ([モデル] タブ)**: 次に［モデル］タブで、ビルド モードとリソースを選択し、このノードから直接コンセプト モデル ナゲットを生成しました。異なるリソース テンプレートを選択できますが、この例では、基本リソースを保持しています。

図 3-13
[テキスト マイニング モデル作成ノード] ダイアログ ボックス:[モデル] タブ

例 2:インタラクティブにカテゴリモデルを作成するExcelファイルノードおよびテキストマイニングノード

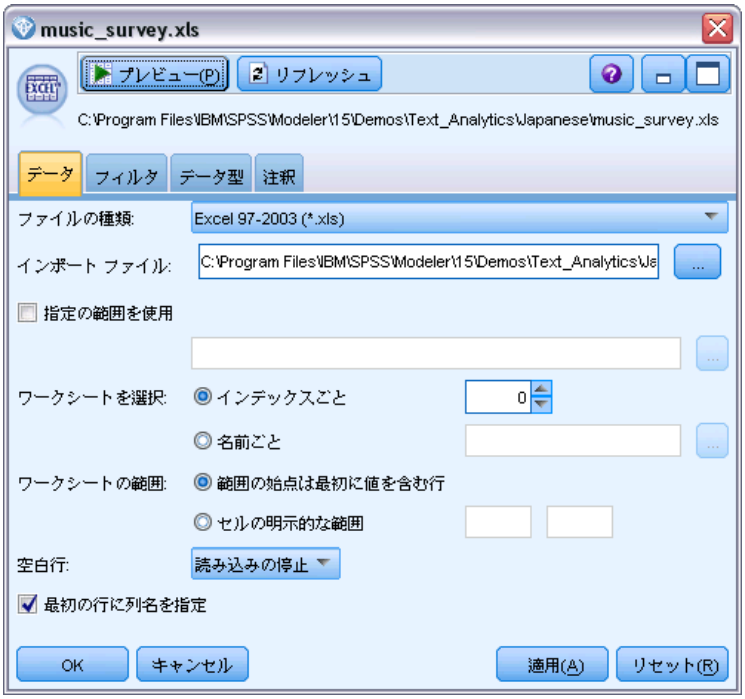
この例では、テキスト マイニング ノードがどのようにインタラクティブ ワークベンチ セッションを起動できるかを示しています。インタラクティブ ワークベンチの詳細は、8 章を参照してください。

図 3-14
ストリームの例:テキストマイニングノードとExcelファイルノード (インタラクティブに作成)



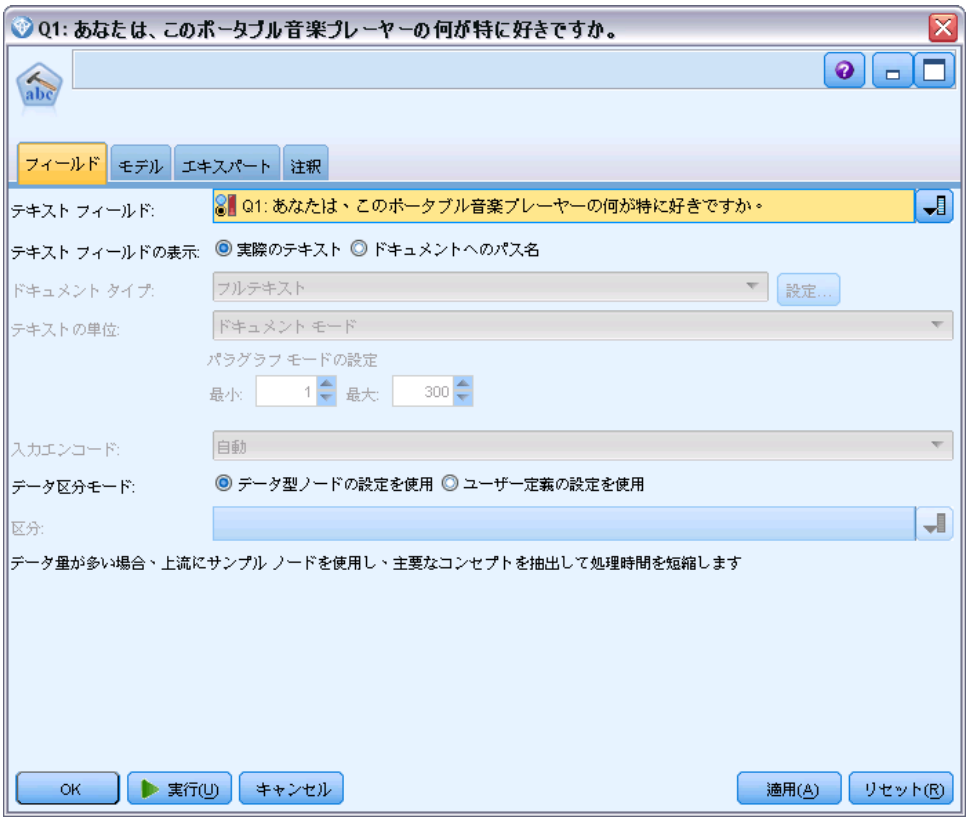
- **Excelソースノード ([データ] タブ):** まず、このノードをストリームに追加して、テキスが保存されている場所を指定しました。

図 3-15
Excelソースノードダイアログボックス : [データ] タブ



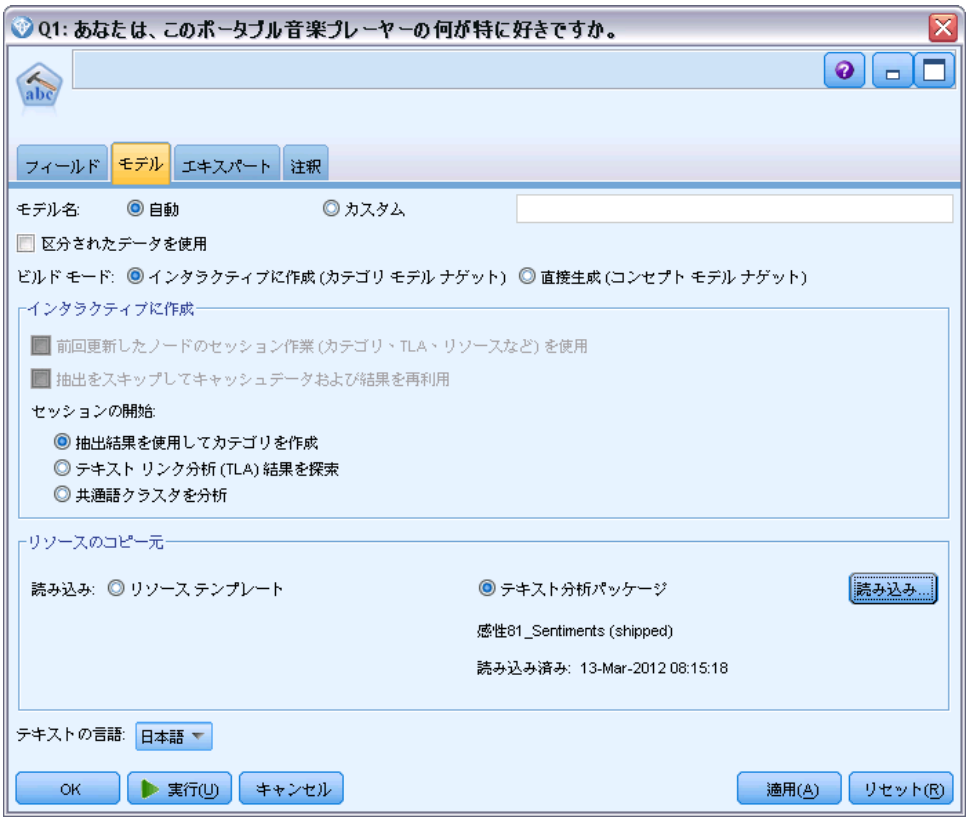
- **テキスト マイニング ノード ([フィールド] タブ):** 次に、テキスト マイニング ノードを追加して接続しました。最初のタブで入力形式を定義しました。入力ノードからフィールド名を選択し、Excelソースノードから直接データが送られるため、[テキスト] フィールドが**実際のテキスト**を示すオプションを選択しました。

図 3-16
[テキスト マイニング モデル作成ノード] ダイアログ ボックス:[フィールド] タブ



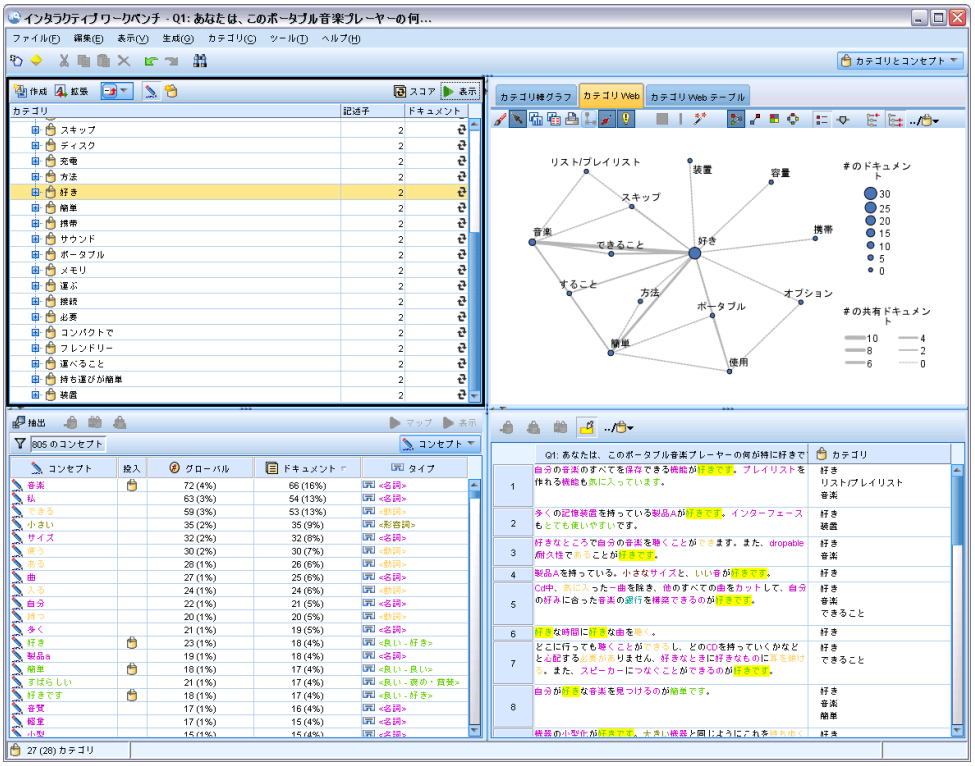
- **テキスト マイニング ノード ([モデル] タブ):** 次に、[モデル] タブで、インタラクティブにカテゴリ モデル ナゲットを作成し、抽出結果を使用して自動的にカテゴリを作成するよう選択しました。この例では、リソースのコピーおよび一連のカテゴリをテキスト分析パッケージから読み込みました。

図 3-17
[テキスト マイニング モデル作成ノード] ダイアログ ボックス:[モデル] タブ



► **インタラクティブ ワークベンチ セッション**:次に、ストリームを実行し、インタラクティブ ワークベンチ インターフェイスが開きました。抽出が実行された後、データの探索およびカテゴリの改善を開始しました。

図 3-18
インタラクティブ ワークベンチ セッション



テキスト マイニング モデル ナゲット:コンセプト モデル

テキスト マイニング コンセプト モデル ナゲットは、[モデル] タブで **[モデルを直接生成]** のオプションを選択してテキスト マイニング モデル モードが正常に実行されると作成されます。テキスト マイニング コンセプト モデル ナゲットは、コール センターのメモ帳データなど、その他のテキスト データの主要キーワードをリアルタイムで発見するために使用されます。

コンセプト モデル ナゲット自体は、タイプに割り当てられているコンセプトのリストを組み合わせます。そのモデルのいずれかまたはすべてのコンセプトを選択し、その他のデータに対してスコアリングできます。テキスト マイニング モデル ナゲットを含むストリームを実行すると、モデル作成の前に、テキスト マイニングモデル作成ノードの [モデル] タブで選択されたビルド モードに従って、新しいフィールドがデータに追加されます。 詳細は、 p. 62 コンセプト モデル:[モデル] タブ を参照してください。

モデル ナゲットが翻訳されたドキュメントを使用して生成された場合、翻訳された言語でスコアリングが実行されます。同様に、モデル ナゲットが英語で生成された場合、ドキュメントが英語に翻訳されるため、モデル ナゲットで翻訳言語を指定できます。

テキスト マイニング モデル ナゲットは生成時、モデル ナゲット パレット（IBM® SPSS® Modeler ウィンドウの右上の［モデル］タブ）内にあります。

結果の表示

モデル ナゲットに関する情報を表示するには、モデル ナゲット パレットでノードを右クリックし、コンテキスト メニューから［**ブラウズ**］を選択します（ストリーム中のノードの場合は［**編集**］）。

モデルのストリームへの追加

モデル ナゲットをストリームに追加するには、モデル ナゲット パレット内でアイコンをクリックし、ノードを配置するストリーム領域をクリックします。アイコンを右クリックし、コンテキスト メニューから［**ストリームに追加**］をクリックします。次に、ストリームをノードに接続すれば、データを渡して予測を生成する準備が整います。

図 3-19
テキスト モデル ナゲットを表示するモデル ナゲット パレット

コンセプト モデル:[モデル] タブ

コンセプト モデルの［モデル］タブには、抽出されたコンセプトのセットが表示されます。コンセプトは、各コンセプトに 1 行ずつのテーブル形式で表示されます。このタブでは、スコアリングに使用されるコンセプトを選択します。

注:代わりにカテゴリ モデル ナゲットを生成すると、このタブには異なる情報が表示されます。 詳細は、 p. 78 カテゴリ モデル ナゲット:[モデル] タブ を参照してください。

図 3-20
[コンセプト モデル ナゲット] ダイアログ ボックス:[モデル] タブ

一番左の列のチェック ボックスで示されているように、デフォルトではすべてのコンセプトがスコアリングに選択されています。チェック ボックスがオンの場合、コンセプトはスコアリングに使用されます。チェック ボックスがオフの場合、コンセプトはスコアリングから除外されます。複数の行を選択して、選択部分のいずれかのチェック ボックスをクリックすると、複数の行をオンにできます。

各コンセプトの詳細については、次の各列に表示された追加情報を参照してください。

コンセプト: 抽出された代表語句です。コンセプトが、このコンセプトに関連する基本キーワードのほか、コンセプト名を示す場合があります。コンセプトの一部である基本キーワードを確認するには、このタブ内の［基本キーワード］パネルを表示してコンセプトを選択し、ダイアログ ボックスの下部にある該当するキーワードを確認します。 詳細は、 p. 66 コンセプト モデルの基本キーワード を参照してください。

グローバル: ここで、グローバル（出現頻度）は、ドキュメント/レコードのセット全体の中でコンセプト（およびすべての基本キーワード）が出現する回数を示します。

- **棒グラフ**: 棒グラフで表示されたこのコンセプトがテキスト データに出現したグローバル出現頻度。棒の色は、タイプを視覚的に区別するためにコンセプトに割り当てられたタイプの色です。
- **%**. このコンセプトがテキスト データに出現したグローバル出現頻度（パーセント表示）。
- **N**: テキスト データにおけるこのコンセプトの出現数。

ドキュメント: ここで、ドキュメントはドキュメント数、つまりコンセプト（およびすべての基本キーワード）が出現するドキュメントまたはレコード数を示します。

- **棒グラフ**: 棒グラフとして表示されたこのコンセプトのドキュメント数。棒の色は、タイプを視覚的に区別するためにコンセプトに割り当てられたタイプの色です。
- **%**. パーセントで表示されたこのコンセプトのドキュメント数。
- **N**: このコンセプトを含むドキュメントまたはレコードの数。

タイプ: コンセプトが割り当てられるタイプ。各コンセプトについて、［グローバル］列および［ドキュメント］列は色付きで表示され、コンセプトに割り当てられたタイプを示します。**タイプ**は、コンセプトの意味上のグループです。 詳細は、 17 章 p. 332 キーワード辞書 を参照してください。

コンセプトの作業

テーブル内のセルを右クリックすると、次のようなコンテキスト メニューが表示されます。

- **すべて選択**: テーブル内のすべての行が選択されます。
- **コピー**: 選択したコンセプトがクリップ ボードにコピーされます。
- **フィールドもコピー** 選択したコンセプトが列の見出しと共にクリップ ボードにコピーされます。
- **選択項目をチェック**: スコアリングするこれらのコンセプトを含むテーブルで選択した行のすべてのチェック ボックスをオンにします。
- **選択項目をチェック解除**: テーブルで選択した行のすべてのチェック ボックスをオフにします。

- **すべてチェック:** テーブルのすべてのチェック ボックスをオンにします。これにより、すべてのコンセプトが最終的な出力に使用されます。
- **すべてチェック解除:** テーブルのすべてのチェック ボックスをオフにします。コンセプトのチェックを解除すると、最終出力では使用されません。
- **コンセプトを含む:** [コンセプトを含む] ダイアログ ボックスが表示されます。 詳細は、 p. 65 スコアリングのコンセプト追加のオプション を参照してください。

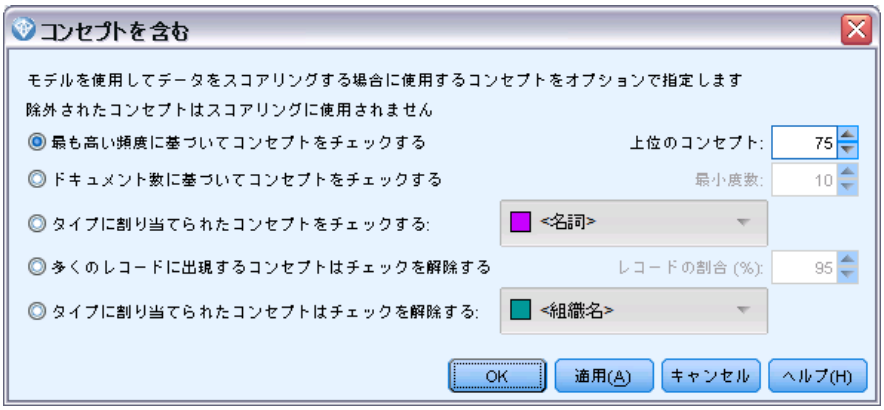
スコアリングのコンセプト追加のオプション

それらのコンセプトをすばやく選択または選択解除するには、**[コンセプトを含む]** のツールバー ボタンをクリックします。

図 3-21
[コンセプトを含む] ツールバー ボタン

このツールバー ボタンをクリックすると [コンセプトを含む] ダイアログ ボックスが開き、規則に基づいてコンセプトを選択できます。[モデル] タブでチェック マークの付いたすべてのコンセプトは、スコアリングに追加されます。このサブダイアログで規則を適用し、スコアリングに使用するコンセプトを変更します。

図 3-22
[コンセプトを含む] ダイアログ ボックス



次のオプションから選択できます。

最も高い頻度に基づいてコンセプトをチェックする。上位のコンセプト: チェックされるコンセプトの数です。最もグローバル頻度の高いコンセプトからチェックします。ここで、頻度は、ドキュメント/レコードのセット全体の中でコンセプト（およびすべての基本キーワード）が出現する回数を示します。

レコード内にコンセプトが複数回出現する場合があるため、この数値がレコード数を上回る場合があります。

ドキュメント数に基づいてコンセプトをチェックする。最小度数:チェックするコンセプトに必要な最低限のドキュメント数です。ここで、ドキュメント数は、コンセプト（およびすべての基本キーワード）が出現するドキュメントの数を示します。

タイプに割り当てられたコンセプトをチェックする: ドロップダウン リストからタイプを選択して、このタイプに割り当てられるすべてのコンセプトをチェックします。抽出時、コンセプトはタイプに自動的に割り当てられます。**タイプ**は、コンセプトの意味上のグループ化です。タイプには、上位レベルのコンセプト、肯定語および否定語および識別子、コンテキスト識別子、人名、地名、組織名などが含まれます。 詳細は、 17 章 p. 332 キーワード辞書 を参照してください。

多くのレコードに出現するコンセプトはチェックを解除する。レコードの割合:レコード数の割合が、指定した数を上回るコンセプトのチェックを解除します。このオプションは、テキストまたはすべてのレコードで頻繁に出現するが、分析においては重要でないコンセプトを除外する場合に役立ちます。

タイプに割り当てられたコンセプトはチェックを解除する: ドロップダウン リストから選択したタイプと合致するコンセプトのチェックを解除します。

コンセプト モデルの基本キーワード

テーブルで選択したコンセプトに定義されている基本キーワードが表示されます。ツールバーで基本キーワードの切り替えボタンをクリックすると、ダイアログの分割したパネルに基本キーワード表が表示されます。

これらの基本キーワードには、モデル ナゲットの生成に使用されるテキストにある抽出された複数形/単数形、置換キーワード、Fuzzy Grouping のキーワードなどのほか、テキスト内にあったかどうかに関係なく、言語リソースで定義された類義語が含まれます。

図 3-23
[基本キーワードを表示] ツールバー ボタン

注:基本キーワードのリストは編集できません。このリストは、すべて言語リソースで定義されている置換、類義語辞書、Fuzzy Grouping によって生成されます。コンセプトに基づいてキーワードをどのようにグループ化するか、またはそれらをどのように処理するかを変更するには、リソースで直接変更し（インタラクティブ ワークベンチの リソース エディタ または テンプレート エディタ で編集でき、ノードに再読み込み)、ストリームを再実行して、更新された結果を持つ新しいモデル ナゲットを取得します。

基本キーワードまたはコンセプトを含むテーブル内のセルを右クリックすると、次のようなコンテキスト メニューが表示されます。

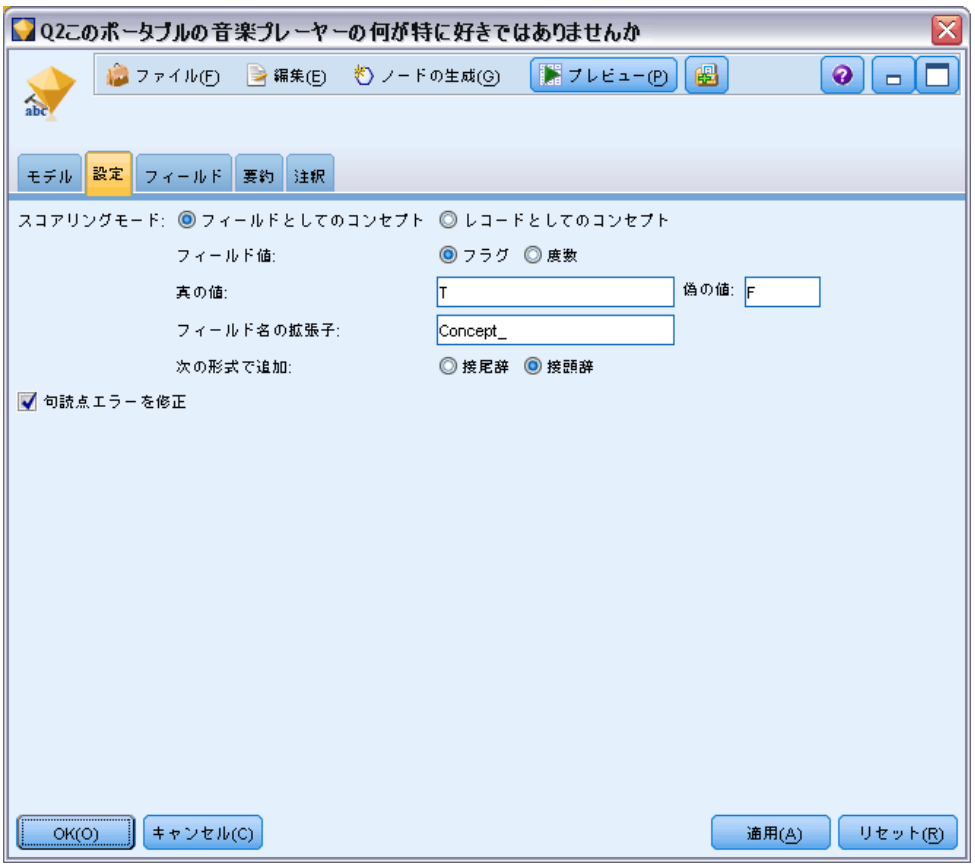
- **コピー:** 選択したセルをクリップボードにコピーします。
- **フィールドもコピー:** 選択したセルが列の見出しと共にクリップ ボードにコピーされます。
- **すべて選択:** テーブル内のすべてのセルが選択されます。

コンセプト モデル:[設定] タブ

必要に応じて、[設定] タブを使用して、新しい入力データのテキスト フィールド値を定義します。また、ここで出力のデータ モデル（スコアリング モード）も定義します。

注: このタブは、モデル ナゲットが領域内にある場合にのみ表示されます。[モデル] パレットでこのダイアログ ボックスを使用している場合、このタブは存在しません。

図 3-24
[テキスト マイニング コンセプト モデル ナゲット] ダイアログ ボックス:[設定] タブ



スコアリング モード:レコードとしてのコンセプト

このスコアリングモードで、新しいレコードが各concept/documentペアに作成されます。通常、入力に比べて出力に多くのレコードがあります。

入力フィールドに加えて、次の新しいフィールドがデータに追加されます。

テーブル 3-1
「レコードとしてのコンセプト」の出力フィールド

フィールド	説明
Concept	テキスト データ フィールドの抽出したコンセプト名が指定されます。
Type	地名または人名などの完全なタイプ名としてコンセプトのタイプを格納します。**タイプ**は、コンセプトの意味上のグループ化です。 詳細は、 17 章 p. 332 キーワード辞書 を参照してください。
Count	テキスト本文（レコード/ドキュメント）の該当するコンセプト（および基本キーワード）の出現数が表示されます。

このオプションを選択すると、[**句読点のエラーを調整する**] を除くすべてのオプションが無効になります。

スコアリング モード:フィールドとしてのコンセプト

各入力フィールドのコンセプト モデル で、新しいレコードが指定されたドキュメントで見つかった各コンセプトに作成されます。そのため、入力内と同じ数の出力レコードがあります。ただし、各レコード（行）では、[モデル] タブで選択された（チェック マークあり）コンセプトまたはカテゴリに 1 つずつ新しいフィールド（列）があります。各コンセプト フィールドの値は、このタブのフィールド値として [**フラグ**] または [**度数**] のどちらを選択するかによって異なります。

フィールド値: 各コンセプトの新しいフィールドに度数またはフラグ値を指定するかを選択します。

- **フラグ**:はい/いいえ、真/偽、T/F、または1 と 2 など、出力に 2 つの値を持つフラグを取得します。ストレージ タイプは、選択した値を反映するよう自動的に設定されます。たとえば、フラグに数値を入力すると、自動的に整数値として処理されます。フラグ型では、文字列、整数、実数、または日付/時間のストレージ タイプを利用することができます。[**真** (True)] および [**偽** (False)] のフラグ値を入力します。
- **度数:** 指定したレコードにコンセプトが出現した回数を取得します。

フィールド名拡張子: フィールド名の拡張子を指定します。コンセプト名に加えてこの拡張子を使用して、フィールド名が生成されます。

- **次の形式で追加:** 拡張子をフィールド名に追加する場所を指定します。[**接頭辞**] を選択すると、文字列の頭に拡張子が追加されます。[**接頭辞**] を選択すると、文字列の終わりに拡張子が追加されます。

句読点のエラーを調整する: 抽出時に句読点エラー（不適切な使用方法など）を含むテキストを一時的に正規化し、コンセプトの抽出可能性を向上させます。自由記述式アンケートの回答、電子メール、CRM データなど、テキストが短く品質が悪い場合、またはテキストに略語が多く含まれている場合に特に役立ちます。

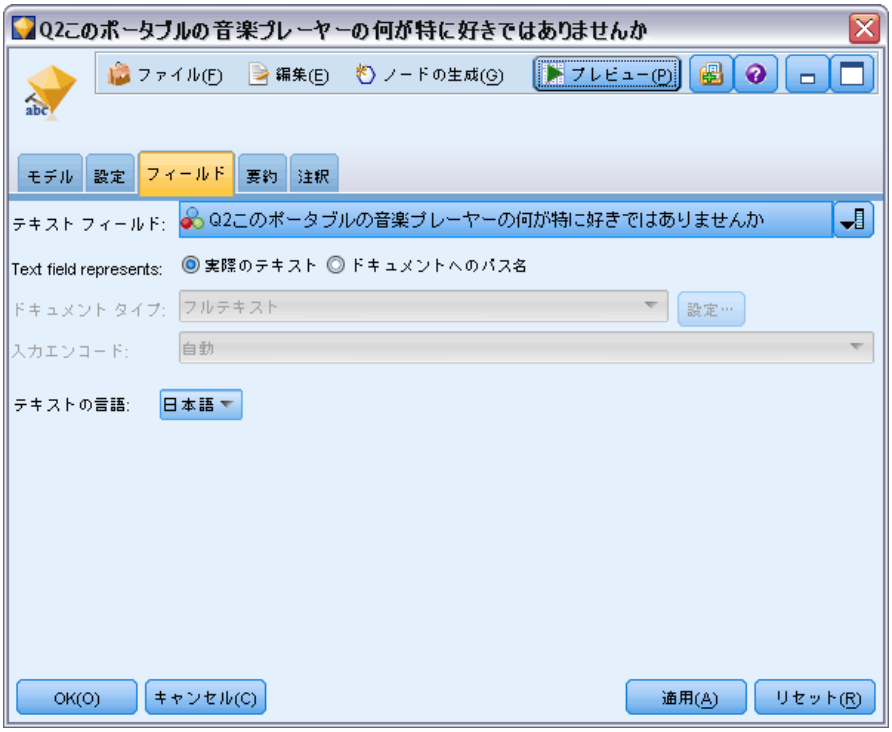
注:[句読点のエラーを調整する] オプションは、日本語テキストを処理している場合は適用されません。

コンセプト モデル:[フィールド] タブ

必要に応じて、[フィールド] タブを使用して、新しい入力データのテキスト フィールド値を定義します。

注: このタブは、モデル ナゲットがストリームにある場合にのみ表示されます。[モデル] パレットでこの出力を使用している場合、このタブは存在しません。

図 3-25
[テキスト マイニング コンセプト モデル ナゲット] ダイアログ ボックス:[フィールド] タブ



テキスト フィールド: マイニングするテキスト、ドキュメントのパス名、またはドキュメントへのディレクトリ パス名が入力されたフィールドを選択します。このフィールドはデータ ソースによって異なります。

テキスト フィールドの表示: これまでの結果で指定されたテキスト フィールドに何が入力されているかを示します。選択されるのは次のとおりです。

- **実際のテキスト:** コンセプトが抽出される正確なテキストをフィールドに入力する場合、このオプションを指定します。
- **ドキュメントへのパス名:** テキスト ドキュメントの場所へ1 つまたは複数のパス名をフィールドに入力する場合、このオプションを選択します。

ドキュメント タイプ: テキスト フィールドが [**ドキュメントへのパス名**] を示すよう指定した場合にのみ使用できます。ドキュメント タイプは、テキストの構造を指定します。次に示すタイプの 1 つを選択します。

- **フル テキスト:** 多くのドキュメントまたはテキスト ソースに使用します。テキストのセット全体をスキャンして抽出します。他のオプションとは異なり、このオプションに追加設定はありません。
- **構造のあるテキスト:** 参考文献形式、特許、特定および分析できる通常の構造を含むファイルに使用します。このドキュメントタイプを使用して、抽出プロセスのすべてまたは一部をスキップします。キーワードの区切り文字を定義、タイプを割り当て、出現頻度の最小値を指定できます。このオプションを選択する場合、[**設定**] ボタンをクリックして、[ドキュメント設定] ダイアログ ボックスの [**構造のあるテキストの書式設定**] 領域にテキストの区切り文字を入力します。詳細は、 p. 39［フィールド］タブのドキュメント設定 を参照してください。
- **XML テキスト:** 抽出するテキストを含む XML タグを指定します。他のタグはすべて無視されます。このオプションを選択する場合、[**設定**] ボタンをクリックして、[ドキュメント設定] ダイアログ ボックスの [**XMLテキストの書式設定**] 領域で、抽出プロセスで読み込まれるテキストを含む XML 要素を明示的に指定します。 詳細は、 p. 39［フィールド］タブのドキュメント設定 を参照してください。

入力エンコード: テキスト フィールドが [**ドキュメントへのパス名**] を示すよう指定した場合にのみ使用できます。デフォルトのテキスト エンコードを指定します。日本語以外のすべての言語について、指定された、または認識されたエンコードから ISO-8859-1 への変換が行われます。そのため、別のエンコードが指定されている場合であっても、抽出エンジンは処理前にテキストを ISO-8859-1 に変換します。ISO-8859-1 エンコード定義に一致しない文字は、スペースに変換されます。日本語テキストの場合、SHIFT_JIS、EUC_JP、UTF-8、または ISO-2022-JP のエンコード オプションから 1 つ選択できます。

テキストの言語: マイニング中のテキストの言語を認識；これが抽出中に検出される主な言語です。現在使用できないサポート言語のライセンス購入については、営業担当者に連絡してください。

コンセプト モデル:[要約] タブ

［要約］タブには、モデルそのもの（分析フォルダ）、モデルで使用するフィールド（フィールド フォルダ）、モデル作成時に使用する設定（ビルド設定 フォルダ）、およびモデルの学習（学習の要約 フォルダ）についての情報を表示します。

モデル作成ノードを初めて参照した場合、［要約］タブのフォルダは閉じられています。関心のある結果を表示するには、フォルダの左側にある展開コントロールを使用するか、または［**すべて展開**］ボタンをクリックしてすべての結果を表示してください。見終わった結果を隠すには、展開コントロールを使って特定のフォルダを閉じるか、または［**すべて閉じる**］ボタンをクリックしてすべてのフォルダを非表示にします。

図 3-26
[テキスト マイニング モデル ナゲット] ダイアログ ボックス:[要約] タブ

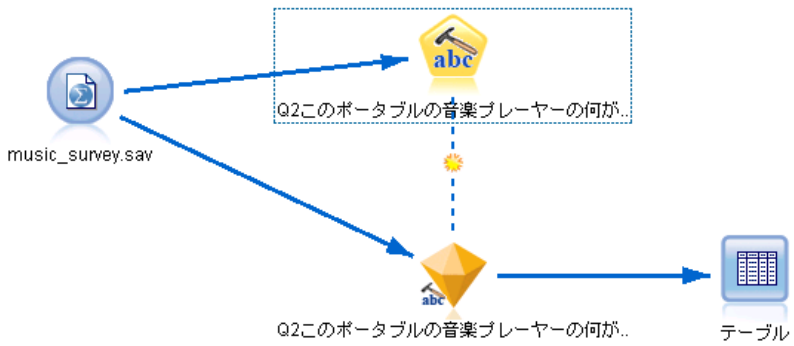
ストリームでのコンセプト モデル ナゲットの使用

テキスト マイニング モデル作成ノードを使用すると、コンセプト モデル ナゲットまたはカテゴリ モデル ナゲット（インタラクティブ ワークベンチ セッションを使用）のいずれかを生成できます。次の例では、単純なストリームでコンセプト モデルの使用方法について示しています。

例:コンセプト モデル ナゲットを含む統計ファイル ノード

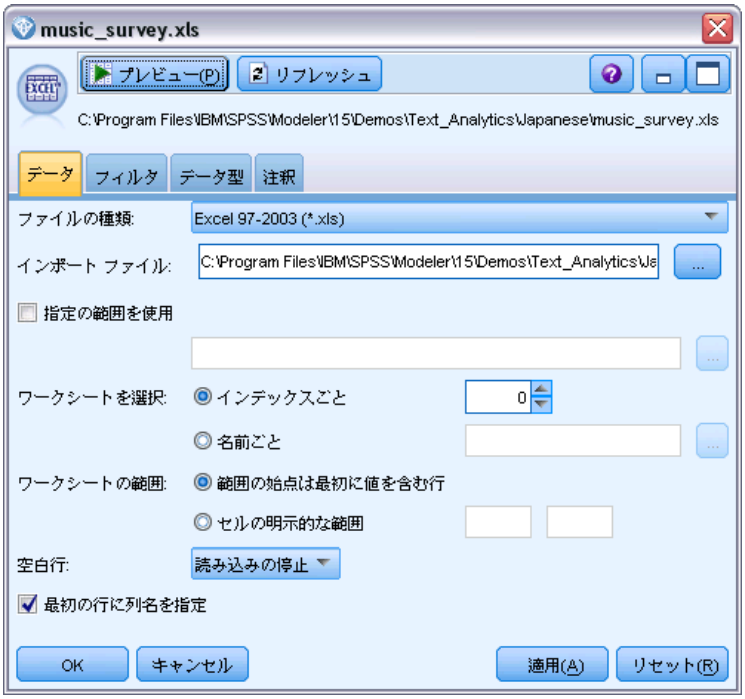
次の例は、テキストマイニングコンセプトモデルナゲットの使用方法を表示しています。

図 3-27
ストリームの例:テキスト マイニング コンセプト モデル ナゲットを含む統計ファイル ノード



- **統計ファイル ノード ([設定] タブ):** まず、このノードをストリームに追加して、テキスト ドキュメントが保存されている場所を指定しました。

図 3-28
Statistics ファイル ノード ダイアログ ボックス:[データ] タブ

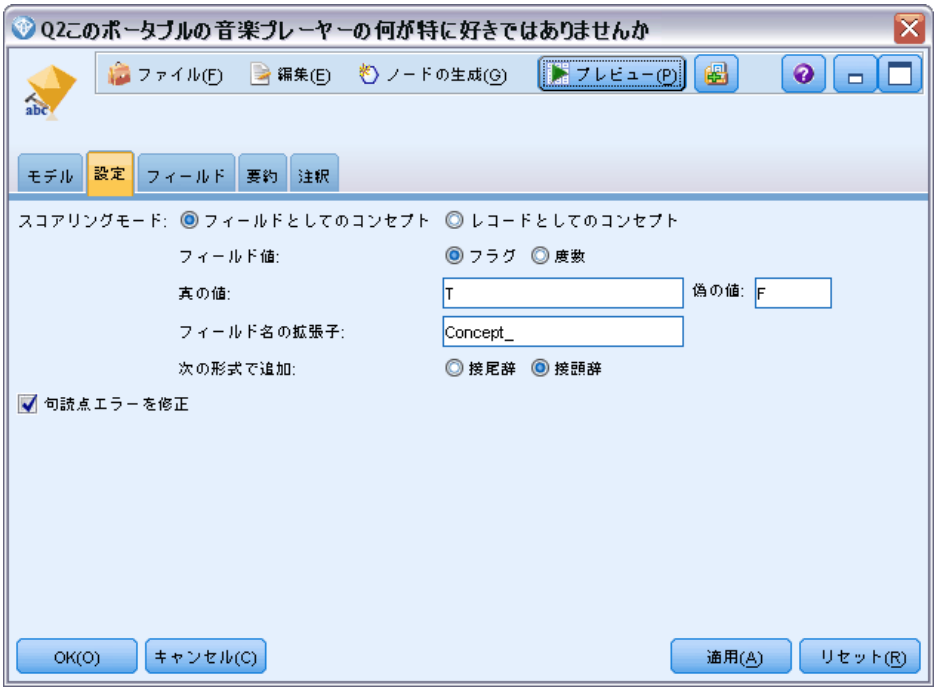


- **テキスト マイニング コンセプト モデル ナゲット ([モデル] タブ)**: 次に、コンセプト モデル ナゲットを統計ファイル ノードに追加して接続しました。データのスコアリングに使用したいコンセプトを選択しました。

図 3-29
[テキスト マイニング モデル ナゲット] ダイアログ ボックス:[モデル] タブ

- **テキスト マイニング コンセプト モデル ナゲット ([設定] タブ):** 次に出力形式を定義し、[フィールドとしてのコンセプト] を選択しました。1 つの新しいフィールドが、[モデル] タブで選択した各コンセプトの出力に作成されます。各フィールド名は、コンセプト名と、接頭辞「Concept_」で成り立っています。

図 3-30
[テキスト マイニング コンセプト モデル ナゲット] ダイアログ ボックス:[設定] タブ

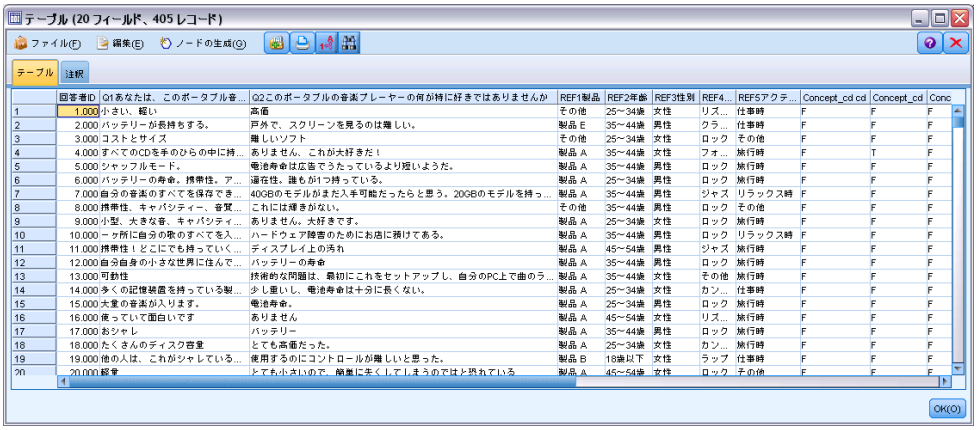


- **テキスト マイニング コンセプト モデル ナゲット ([フィールド] タブ):** 次に、テキスト フィールド [Q2_What_do_you_like_least_about_this_portable_music_player] を選択しました。lpmpフィールド名は統計ファイル ノードに由来しています。また、オプション[テキスト フィールドの表示: 実際のテキスト]を選択しました。実際のテキスト。

図 3-31
[テキスト マイニング コンセプト モデル ナゲット] ダイアログ ボックス:[フィールド] タブ

► **テーブル ノード:** 次に、 テーブル ノードを接続して結果を表示し、ストリームを実行しました。テーブル出力が画面上に表示されます。

図 3-32
コンセプト フラグを表示するためにスクロールしたテーブル出力



	回答者ID	Q1 あなたは、このポータブル音...	Q2 このポータブルの音楽プレーヤーの何が特に好きではありませんか	REF1製品	REF2年齢	REF3性別	REF4...	REF5アクテ...	Concept_cd	Concept_cd	Conc
1	1.000	小さい、軽い	高価	その他	25~34歳	女性	リズ...	仕事時	F	F	F
2	2.000	バッテリーが長持ちする。	戸外で、スクリーンを見るのは難しい。	製品 E	35~44歳	男性	クラ...	仕事時	F	F	F
3	3.000	コストとサイズ	難しいソフト	その他	25~34歳	女性	ロック	その他	F	F	F
4	4.000	すべてのCDを手のひらの中に持...	ありません、これが大好きだ！	製品 A	35~44歳	女性	フォ...	旅行時	F	T	F
5	5.000	シャッフルモード。	電池寿命は広告でうたっているより短いようだ。	製品 A	35~44歳	男性	ロック	旅行時	F	F	F
6	6.000	バッテリーの寿命。携帯性。ア...	遠在性。誰もが1つ持っている。	製品 A	25~34歳	男性	ロック	旅行時	F	F	F
7	7.000	自分の音楽のすべてを保存でき...	40GBのモデルがまた入手可能だったらと思う。20GBのモデルを持っ...	製品 A	35~44歳	男性	ジャズ	リラックス時	F	F	F
8	8.000	携帯性、キャパシティー、音質...	これには嫌きがない。	その他	35~44歳	男性	ロック	その他	F	F	F
9	9.000	小型、大きな音、キャパシティ...	ありません。大好きです。	製品 A	25~34歳	女性	ロック	旅行時	F	F	F
10	10.000	一ヶ所に自分の歌のすべてを入...	ハードウェア障害のためにお店に預けてある。	製品 A	35~44歳	男性	ロック	リラックス時	F	F	F
11	11.000	携帯性！どこにでも持っていく...	ディスプレイ上の汚れ	製品 A	45~54歳	男性	ジャズ	旅行時	F	F	F
12	12.000	自分自身の小さな世界に住んで...	バッテリーの寿命	製品 A	35~44歳	男性	ロック	旅行時	F	F	F
13	13.000	可動性	技術的な問題は、最初にこれをセットアップし、自分のPC上で曲のラ...	製品 A	35~44歳	女性	その他	旅行時	F	F	F
14	14.000	多くの記憶装置を持っている製...	少し重いし、電池寿命は十分に長くない。	製品 A	25~34歳	女性	カン...	仕事時	F	F	F
15	15.000	大量の音楽が入ります。	電池寿命。	製品 A	25~34歳	男性	ロック	旅行時	F	F	F
16	16.000	僕っていて面白いです	ありません	製品 A	45~54歳	女性	リズ...	旅行時	F	F	F
17	17.000	おシャレ	バッテリー	製品 A	35~44歳	男性	ロック	旅行時	F	F	F
18	18.000	たくさんのディスク容量	とても高価だった。	製品 A	25~34歳	女性	カン...	旅行時	F	F	F
19	19.000	他の人は、これがシャレている...	使用するのにコントロールが難しいと思った。	製品 B	18歳以下	女性	ラップ	仕事時	F	F	F
20	20.000	経営	とても小さいので、簡単に失くしてしまうのではと恐れている	製品 A	45~54歳	女性	ロック	その他	F	F	F

テキスト マイニング モデル ナゲット:カテゴリ モデル

テキスト マイニング カテゴリモデル ナゲットは、インタラクティブ ワークベンチからカテゴリ モデルが生成されると作成されます。このモデル作成ナゲットには、一連のカテゴリが含まれ、その定義はコンセプト、タイプ、TLA パターンおよびカテゴリ規則で構成されています。ナゲットを使用して、アンケートの回答、ブログ エントリ、その他の Web フィード、およびその他テキスト データをカテゴリ化します。

モデル作成ノードでインタラクティブ ワークベンチ セッションを起動すると、カテゴリ モデルを生成する前に抽出結果を探索し、リソースを調整、カテゴリを調整できます。テキスト マイニング モデル ナゲットを含むストリームを実行すると、モデル作成の前に、テキスト マイニングモデル作成ノードの［モデル］タブで選択されたビルド モードに従って、新しいフィールドがデータに追加されます。 詳細は、 p.78 カテゴリ モデル ナゲット:[モデル] タブ を参照してください。

モデル ナゲットが翻訳されたドキュメントを使用して生成された場合、翻訳された言語でスコアリングが実行されます。同様に、モデル ナゲットが英語で生成された場合、ドキュメントが英語に翻訳されるため、モデル ナゲットで翻訳言語を指定できます。

テキスト マイニング モデル ナゲットは生成時、モデル ナゲット パレット（IBM® SPSS® Modeler ウィンドウの右上の［モデル］タブ）内にあります。

結果の表示

モデル ナゲットに関する情報を表示するには、モデル ナゲット パレットでノードを右クリックし、コンテキスト メニューから **[ブラウズ]** を選択します（ストリーム中のノードの場合は **[編集]**）。

モデルのストリームへの追加

モデル ナゲットをストリームに追加するには、モデル ナゲット パレット内でアイコンをクリックし、ノードを配置するストリーム領域をクリックします。アイコンを右クリックし、コンテキスト メニューから [ストリームに追加] をクリックします。次に、ストリームをノードに接続すれば、データを渡して予測を生成する準備が整います。

図 3-33
テキスト モデル ナゲットを表示するモデル ナゲット パレット

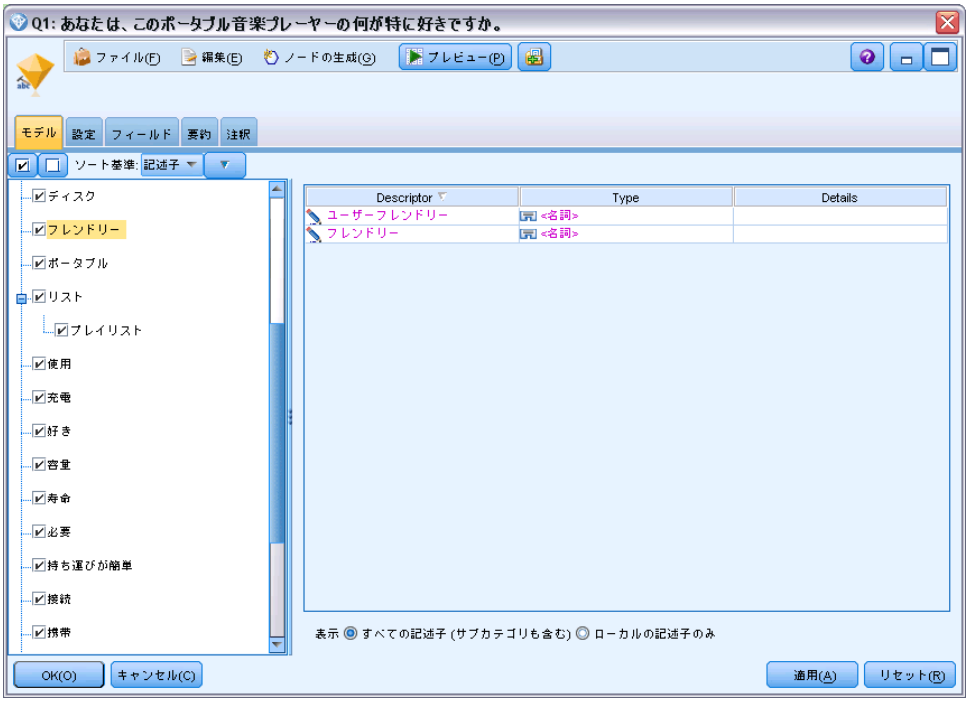
カテゴリ モデル ナゲット:[モデル] タブ

カテゴリ モデルの場合、[モデル] タブの左側にカテゴリ モデルのカテゴリのリスト、右側に選択したカテゴリの記述子のリストが表示されます。各カテゴリは、多くの記述子で構成されています。選択した各カテゴリについて、関連する記述子がテーブルに表示されます。記述子には、コンセプト、カテゴリ規則、タイプ、および TLA パターンが含まれます。記述子が示す内容の例のほか、各記述子のタイプも表示されます。

このタブでは、スコアリングに使用されるカテゴリを選択します。カテゴリ モデルについて、ドキュメントおよびレコードがカテゴリにスコアリングされます。ドキュメントまたはレコードにテキストまたは基本キーワードの 1 つまたは複数の記述子がある場合、そのドキュメントまたはレコードは記述子が含まれるカテゴリに割り当てられます。これらの基本キーワードには、モデル ナゲットの生成に使用されるテキストにある抽出された複数形/単数形のキーワード、置換キーワード、Fuzzy Grouping のキーワードなどのほか、テキスト内にあったかどうかに関係なく、言語リソースで定義された類義語が含まれます。

注:代わりにコンセプト モデル ナゲットを生成すると、このタブには異なる結果が表示されます。 詳細は、 p.62 コンセプト モデル:[モデル] タブ を参照してください。

図 3-34
[カテゴリ モデル ナゲット] ダイアログ ボックス:[モデル] タブ



カテゴリ ツリー

各カテゴリの詳細については、該当するカテゴリを選択し、そのカテゴリの記述子に表示される情報を参照してください。記述子については、次の情報が表示されます。

- **記述子**名:このフィールドには記述子名のほか記述子の種類を示すアイコンが指定されます。

  コンセプト　　TLA パターン

  タイプ　　カテゴリ規則

- **タイプ:** このフィールドには記述子のタイプ名が指定されます。タイプは、組織名、商品名、肯定的な意見など、類似したコンセプトの集合(意味的なグループ)です。条件規則はタイプに割り当てられません。
- **詳細:** その記述子に含まれる内容のリストが指定されます。合致数によっては、ダイアログ ボックスのサイズ制限のため、各記述子のリスト全体を表示できない場合があります。

カテゴリの選択およびコピー

左のパネルのチェック ボックスで示されているように、デフォルトではすべてのカテゴリがスコアリングに選択されています。チェック ボックスがオンの場合、カテゴリはスコアリングに使用されます。チェック ボックスがオフの場合、カテゴリはスコアリングから除外されます。複数の行を選択して、選択部分のいずれかのチェック ボックスをクリックすると、複数の行をオンにできます。また、カテゴリまたはサブカテゴリが選択されているが、サブカテゴリの 1 つが選択されていない場合、チェックボックスの背景が青で表示され、選択されたカテゴリの下位の選択が一部であることを示します。

テーブル内のカテゴリを右クリックすると、次のようなコンテキスト メニューが表示されます。

- **選択項目をチェック:** テーブルで選択した行のすべてのチェック ボックスをオンにします。
- **選択項目をチェック解除:** テーブルで選択した行のすべてのチェック ボックスをオフにします。
- **すべてチェック:** テーブルのすべてのチェック ボックスをオンにします。これにより、すべてのカテゴリが最終的な出力に使用されます。ツールバーの対応するチェックボックス アイコンを使用することもできます。
- **すべてチェック解除:** テーブルのすべてのチェック ボックスをオフにします。カテゴリのチェックを解除すると、カテゴリは最終的な出力で使用されなくなります。ツールバーの対応する空白のチェックボックス アイコンをス要することもできます。

記述子テーブル内のセルを右クリックすると、次のようなコンテキスト メニューが表示されます。

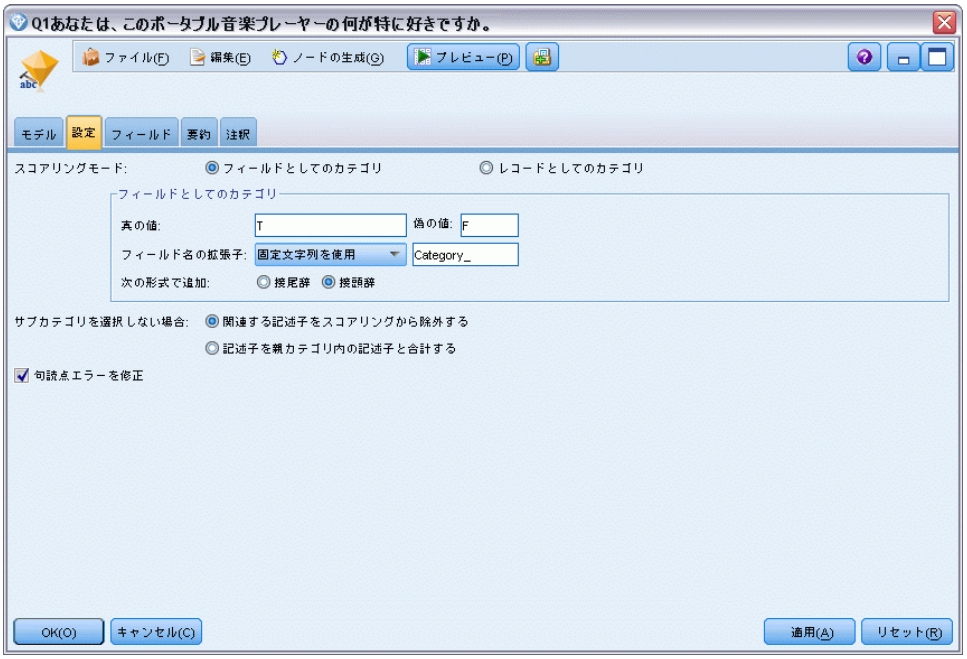
- **コピー:** 選択したコンセプトがクリップ ボードにコピーされます。
- **フィールドもコピー:** 選択した記述子が列の見出しと共にクリップ ボードにコピーされます。
- **すべて選択:** テーブル内のすべての行が選択されます。

カテゴリ モデル ナゲット:[設定] タブ

必要に応じて、[設定] タブを使用して、新しい入力データのテキスト フィールド値を定義します。また、ここで出力のデータ モデル（スコアリング モード）も定義します。

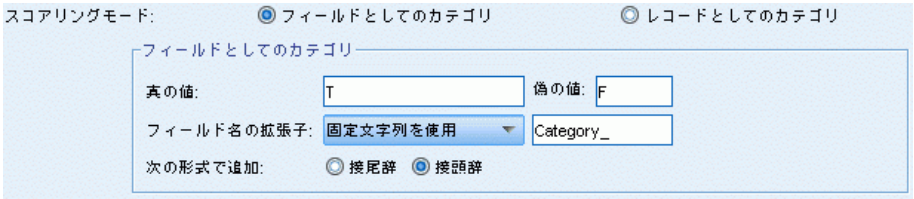
注: このタブは、モデル ナゲットがストリーム内の領域にある場合にのみ、ノードのダイアログ ボックスに表示されます。[モデル] パレットでこのナゲットを使用している場合、このタブは存在しません。

図 3-35
[テキスト マイニング カテゴリ モデル ナゲット] ダイアログ ボックス:[設定] タブ



スコアリング モード:フィールドとしてのカテゴリ

図 3-36
「フィールドとしてのカテゴリ」の [設定] タブ



このオプションの場合、入力内と同じ数の出力レコードがあります。しかし、これによって各レコードは、モデルタブで選択された全てのカテゴリ（チェックマークを使用）毎に新しいフィールドを含みます各フィールドに、はい/いいえ、True/False、T/F または 1 および 2 などの True および False のフラグ値を入力します。ストレージ タイプは、選択した値を反映するよう自動的に設定されます。たとえば、フラグに数値を入力すると、自動的に整数値として処理されます。フラグ型では、文字列、整数、実数、または日付/時間のストレージ タイプを利用することができます。

フィールド名拡張子: フィールド名の拡張接頭辞/接尾辞を指定したり、カテゴリ コードを使用したりできます。カテゴリ名に加えてこの拡張子を使用して、フィールド名が生成されます。

- **次の形式で追加:** 拡張子をフィールド名に追加する場所を指定します。[**接頭辞**] を選択すると、文字列の頭に拡張子が追加されます。[**接頭辞**] を選択すると、文字列の終わりに拡張子が追加されます。

サブカテゴリが選択されていない場合: スコアリングに選択されていないサブカテゴリに含まれる記述子の処理方法を指定できます。オプションは 2 つあります。

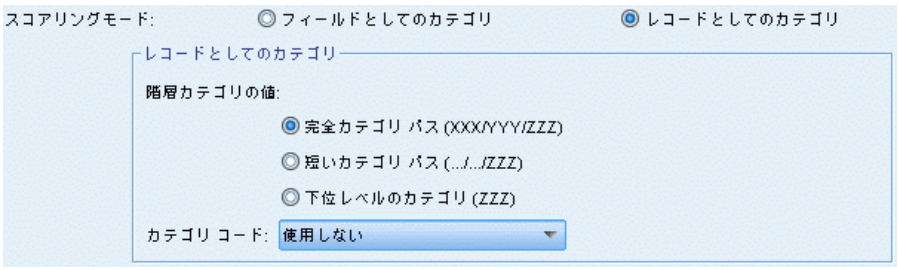
- オプション **[記述子をスコアリングから完全に除外する]** を選択すると、チェック記号のない（選択されていない）サブカテゴリは無視され、スコアリングに使用されません。
- オプション **[記述子を上位カテゴリ内の記述子と合計する]** を選択すると、チェック記号のない（選択されていない）サブカテゴリの記述子は上位カテゴリ（このサブカテゴリの上位にあるカテゴリ）の記述子として使用されます。複数レベルのサブカテゴリが選択されない場合、記述子は使用できる最初の上位カテゴリにロール アップされます。

句読点のエラーを調整する: 抽出時に句読点エラー（不適切な使用方法など）を含むテキストを一時的に正規化し、コンセプトの抽出可能性を向上させます。自由記述式アンケートの回答、電子メール、CRM データなど、テキストが短く品質が悪い場合、またはテキストに略語が多く含まれている場合に特に役立ちます。

注:[句読点のエラーを調整する] オプションは、日本語テキストを処理している場合は適用されません。

スコアリング モード:レコードとしてのカテゴリ

図 3-37
「レコードとしてのカテゴリ」の [設定] タブ



このオプションで、新しいレコードが各category、documentペアに作成されます。通常、入力に比べて出力に多くのレコードがあります。入力フィールドのほか、モデルの種類によって、新しいフィールドもデータに追加されます。

テーブル 3-2
「レコードとしてのカテゴリ」の出力フィールド

新しい出力フィールド	説明
Category	テキスト ドキュメントが割り当てられるカテゴリ名が指定されます。カテゴリが別のカテゴリのサブカテゴリである場合、ダイアログで選択した値によってカテゴリ名への完全パスが制御されます。

階層カテゴリの値: サブカテゴリの名前を出力内でどのように表示するかを制御します。

- **完全カテゴリパス:** カテゴリ名と、該当する場合、カテゴリ名とサブカテゴリ名をスラッシュを使用して区切り、上位カテゴリの完全パスを出力します。
- **短いカテゴリパス:** カテゴリ名のみを出力します。ただし、省略記号を使用して、該当するカテゴリの上位カテゴリ数を示します。
- **下位レベルのカテゴリ:** 完全パスまたは上位カテゴリを表示せず、カテゴリ名のみを出力します。

サブカテゴリが選択されていない場合: スコアリングに選択されていないサブカテゴリに含まれる記述子の処理方法を指定できます。オプションは 2 つあります。

- オプション **[記述子をスコアリングから完全に除外する]** を選択すると、チェック記号のない（選択されていない）サブカテゴリは無視され、スコアリングに使用されません。
- オプション **[記述子を上位カテゴリ内の記述子と合計する]** を選択すると、チェック記号のない（選択されていない）サブカテゴリの記述子は上位カテゴリ（このサブカテゴリの上位にあるカテゴリ）の記述子として使用されます。複数レベルのサブカテゴリが選択されない場合、記述子は使用できる最初の上位カテゴリにロール アップされます。

句読点のエラーを調整する: 抽出時に句読点エラー（不適切な使用方法など）を含むテキストを一時的に正規化し、コンセプトの抽出可能性を向上させます。自由記述式アンケートの回答、電子メール、CRM データなど、テキストが短く品質が悪い場合、またはテキストに略語が多く含まれている場合に特に役立ちます。

注:[句読点のエラーを調整する] オプションは、日本語テキストを処理している場合は適用されません。

カテゴリ モデル ナゲット:その他のタブ

カテゴリ モデル ナゲットの［フィールド］タブと［設定］タブは、コンセプト モデル ナゲットと同じです。

- ［フィールド］タブ。 詳細は、 p. 69 コンセプト モデル:[フィールド] タブ を参照してください。
- ［要約］タブ。 詳細は、 p. 71 コンセプト モデル:[要約] タブ を参照してください。

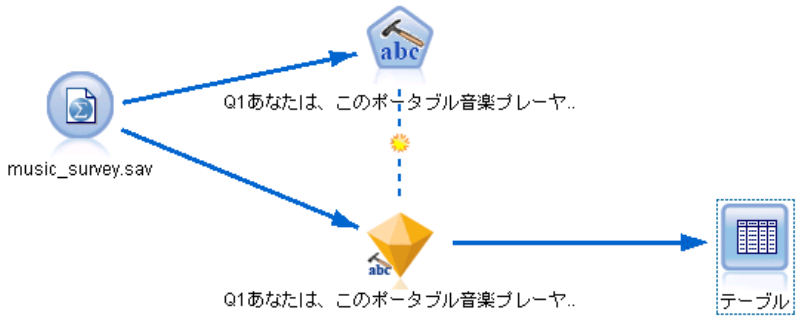
ストリームでのカテゴリ モデル ナゲットの使用

テキスト マイニング カテゴリ モデル ナゲットは、インタラクティブ ワークベンチ セッションから生成されます。このモデル ナゲットはストリームで使用できます。

例:カテゴリ モデル ナゲットを含む統計ファイル ノード

次の例では、テキスト マイニング モデル ナゲットの使用方法について示しています。

図 3-38
ストリームの例:テキスト マイニング カテゴリ モデル ナゲットを含む統計ファイル ノード

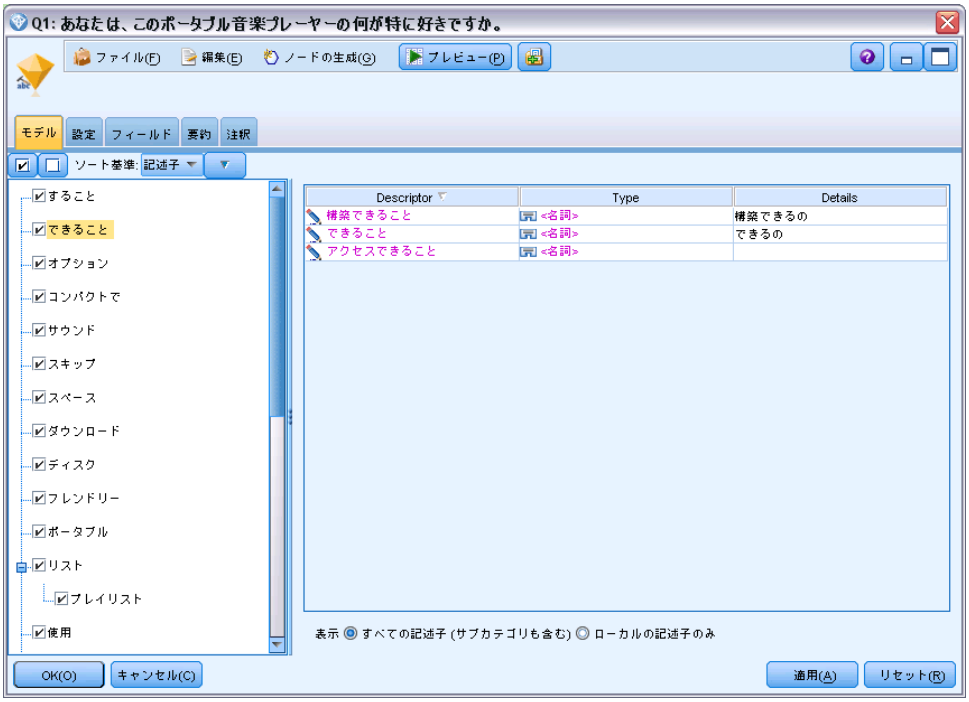


- **統計ファイル ノード ([設定] タブ):** まず、このノードをストリームに追加して、テキスト ドキュメントが保存されている場所を指定しました。

図 3-39
Statistics ファイル ノード ダイアログ ボックス:[データ] タブ

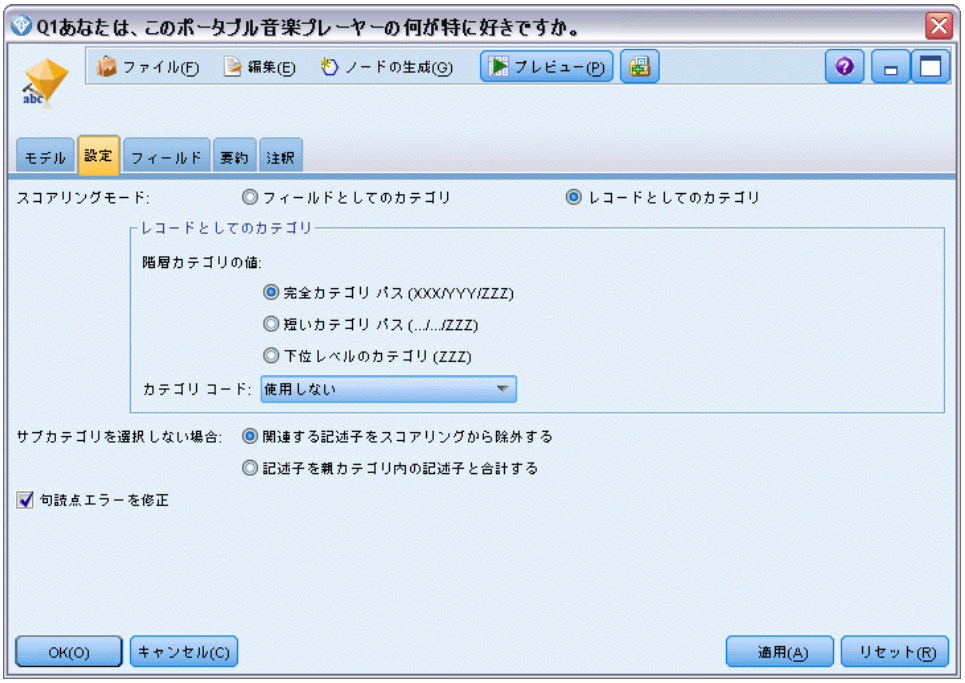
- **テキスト マイニング カテゴリ モデル ナゲット ([モデル] タブ):** 次に、カテゴリ モデル ナゲットを統計ファイル ノードに追加して接続しました。データのスコアリングに使用したいカテゴリを選択しました。

図 3-40
[テキスト マイニング モデル ナゲット] ダイアログ ボックス:[モデル] タブ



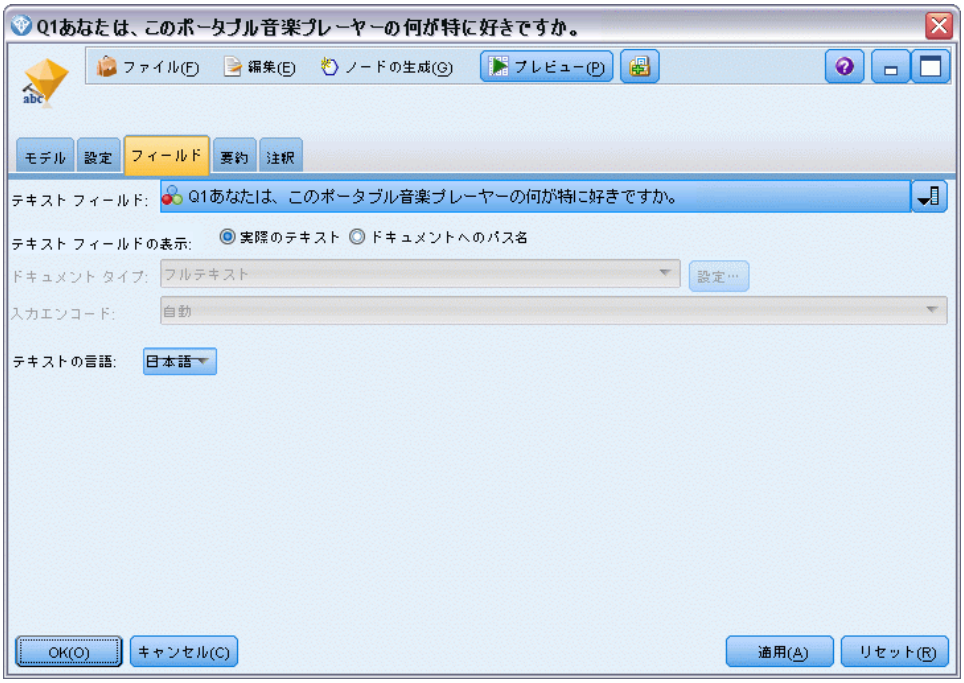
► **テキスト マイニング モデル ナゲット ([設定] タブ):** 次に出力形式 [フィールドとしてのカテゴリ] を定義しました。

図 3-41
[カテゴリ モデル ナゲット] ダイアログ ボックス:[設定] タブ



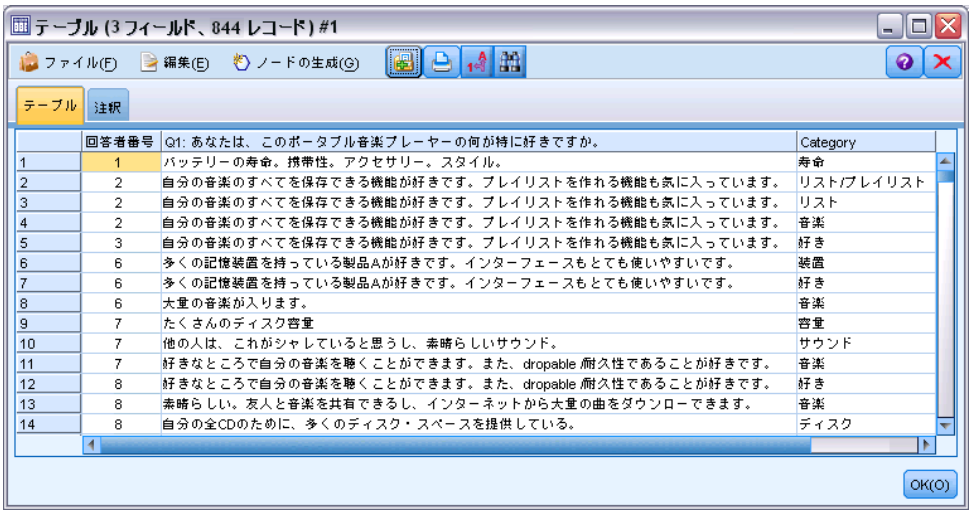
- **テキスト マイニング カテゴリ モデル ナゲット ([フィールド] タブ):** 次に統計ファイル ノードから作成されたフィールド名であるテキスト フィールド変数を選択し、テキスト フィールドが示す内容のオプション **[実際のテキスト]** やその他の設定を選択しました。

図 3-42
[テキスト マイニング モデル ナゲット] ダイアログ ボックス:[フィールド] タブ



► **テーブル ノード:** 次に、 テーブル ノードを接続して結果を表示し、ストリームを実行しました。

図 3-43
テーブル出力



	回答者番号	Q1: あなたは、このポータブル音楽プレーヤーの何が特に好きですか。	Category
1	1	バッテリーの寿命。携帯性。アクセサリー。スタイル。	寿命
2	2	自分の音楽のすべてを保存できる機能が好きです。プレイリストを作れる機能も気に入っています。	リスト/プレイリスト
3	2	自分の音楽のすべてを保存できる機能が好きです。プレイリストを作れる機能も気に入っています。	リスト
4	2	自分の音楽のすべてを保存できる機能が好きです。プレイリストを作れる機能も気に入っています。	音楽
5	3	自分の音楽のすべてを保存できる機能が好きです。プレイリストを作れる機能も気に入っています。	好き
6	6	多くの記憶装置を持っている製品Aが好きです。インターフェースもとても使いやすいです。	装置
7	6	多くの記憶装置を持っている製品Aが好きです。インターフェースもとても使いやすいです。	好き
8	6	大量の音楽が入ります。	音楽
9	7	たくさんのディスク容量	容量
10	7	他の人は、これがシャレていると思うし、素晴らしいサウンド。	サウンド
11	7	好きなところで自分の音楽を聴くことができます。また、dropable /耐久性であることが好きです。	音楽
12	8	好きなところで自分の音楽を聴くことができます。また、dropable /耐久性であることが好きです。	好き
13	8	素晴らしい。友人と音楽を共有できるし、インターネットから大量の曲をダウンローできます。	音楽
14	8	自分の全CDのために、多くのディスク・スペースを提供している。	ディスク

章
4

テキスト リンクのマイニング

テキスト リンク分析ノード

テキスト リンク分析（TLA）ノードは、パターンマッチ テクノロジをテキスト マイニングのコンセプト抽出に追加し、既知のパターンに基づいてテキスト データのコンセプト間の関連性を特定します。これらの関連性は、顧客が製品についてどのように感じているか、どの企業と組んでビジネスを行うか、または遺伝子または医薬品の間の関連性について説明が可能です。

たとえば、競合他社の製品名を抽出しても、重要でない場合があります。このノードを使用して、データ内に人々がこの製品に関してどのように感じているかという意見がある場合、それについて学習することができます。関連性および関係性は、既知のパターンをテキスト データに合致させることによって、特定および抽出します。

IBM® SPSS® Modeler Text Analytics に付属する特定のリソース テンプレート内の TLA パターンを使用または独自のパターンを作成/編集できます。パターン規則は、マクロ、単語リスト、およびブール型質問を形成する単語の空所、または入力テキストと比較される条件規則で構成されています。TLA パターン規則がテキストに一致する場合、このテキストを TLA 結果として抽出し、出力データとして再構築できます。 詳細は、 19 章 p.368 テキスト リンク規則について を参照してください。

テキスト リンク分析ノードを使用して、より直接的にテキストから TLA パターン結果を特定および抽出し、パターンの結果をストリーム内のデータセットに追加できます。ただし、テキスト リンク分析を実行できるのは、テキスト リンク分析ノードだけではありません。テキスト マイニングモデル作成ノードのインタラクティブ ワークベンチ セッションを使用することもできます。

インタラクティブ ワークベンチで、TLA パターン結果を検証してその結果をカテゴリ記述子として使用し、ドリルダウンおよびグラフを使用して結果についてより詳細に学習することができます。詳細は、 12 章 p.268 テキスト リンク分析の検証 を参照してください。 実際に、テキスト マイニング ノードを使用して TLA 結果を抽出すると、後で TLA ノードで直接使用するためにテンプレートを検証し、データ向けに調整することができます。

出力は、最大 6 つのスロット、または 6 つの部分で表示されます。日本語のパターンの出力は、1 つまたは 2 つのスロットだけです。 詳細は、 p.97 TLA ノード出力 を参照してください。

このノードは、IBM® SPSS® Modeler ウィンドウの下部にあるノード パレットの SPSS Modeler Text Analytics タブにあります。 詳細は、 1 章 p.12 IBM SPSS Modeler Text Analytics ノード を参照してください。

要件: テキスト リンク分析ノードは、標準的な入力ノード（データベース ノード、フラット ファイル ノードなど）のいずれかを使用してフィールドに読み込まれた、またはファイル リスト ノードまたは Web フィード ノードによって生成された外部ドキュメントへのフィールド リスト パスに読み込まれたテキスト データを受け入れます。

利点: テキスト リンク分析 ノードは、基本的なコンセプト抽出以上の機能により、コンセプト間の関連性、およびデータ内にあると考えられる関連する意見や識別子に関する情報を提供します。

テキスト リンク分析ノード:[フィールド] タブ

図 4-1
[テキスト リンク分析ノード] ダイアログ ボックス:[フィールド] タブ

[フィールド] タブを使用して、コンセプトを抽出するデータのフィールド設定を指定します。設定できるパラメータを次に示します。

ID フィールド: テキスト レコードの識別子を含むフィールドを選択します。識別子は整数でなければなりません。ID フィールドは、各テキスト レコードのインデックスとして機能します。テキスト フィールドがマイニングされるテキストを示す場合、ID フィールドを使用します。テキスト フィールドが [ドキュメントへのパス名] を示す場合、ID フィールドを使用します。

テキスト フィールド: マイニングするテキスト、ドキュメントのパス名、またはドキュメントへのディレクトリ パス名が入力されたフィールドを選択します。このフィールドはデータ ソースによって異なります。

テキスト フィールドの表示: これまでの結果で指定されたテキスト フィールドに何が入力されているかを示します。選択されるのは次のとおりです。

- **実際のテキスト:** コンセプトが抽出される正確なテキストをフィールドに入力する場合、このオプションを指定します。
- **ドキュメントへのパス名:** テキスト ドキュメントの場所へ1 つまたは複数のパス名をフィールドに入力する場合、このオプションを選択します。

ドキュメント タイプ: テキスト フィールドが [ドキュメントへのパス名] を示すよう指定した場合にのみ使用できます。ドキュメント タイプは、テキストの構造を指定します。次に示すタイプの 1 つを選択します。

- **フル テキスト:** 多くのドキュメントまたはテキスト ソースに使用します。テキストのセット全体をスキャンして抽出します。他のオプションとは異なり、このオプションに追加設定はありません。
- **構造のあるテキスト:** 参考文献形式、特許、特定および分析できる通常の構造を含むファイルに使用します。このドキュメントタイプを使用して、抽出プロセスのすべてまたは一部をスキップします。キーワードの区切り文字を定義、タイプを割り当て、出現頻度の最小値を指定できます。このオプションを選択する場合、[設定] ボタンをクリックして、[ドキュメント設定] ダイアログ ボックスの [構造のあるテキストの書式設定] 領域にテキストの区切り文字を入力します。詳細は、 3 章 p. 39 [フィールド] タブのドキュメント設定 を参照してください。
- **XML テキスト:** 抽出するテキストを含む XML タグを指定します。他のタグはすべて無視されます。このオプションを選択する場合、[設定] ボタンをクリックして、[ドキュメント設定] ダイアログ ボックスの [XMLテキストの書式設定] 領域で、抽出プロセスで読み込まれるテキストを含む XML 要素を明示的に指定します。 詳細は、 3 章 p. 39 [フィールド] タブのドキュメント設定 を参照してください。

テキストの単位: テキスト フィールドが [ドキュメントへのパス名] を示すよう指定し、ドキュメント タイプに [フル テキスト] を選択した場合にのみ使用できます。次の実行モードを選択します。

- **ドキュメント モード**: 通信社からの記事など、短く意味的に同質のドキュメントに使用します。
- **パラグラフ モード**: Web ページおよびタグのないドキュメントに使用します。抽出プロセスでは、内部タグやシンタックスなどの特徴を利用して、ドキュメントを意味的に分割します。このモードを選択すると、パラグラフごとにスコアリングが適用されます。そのため、**リンゴ**および**オレンジ**が同じパラグラフで見つかった場合にのみ、たとえば規則 **リンゴ & オレンジ** が当てはまります。

パラグラフ モードの設定: テキスト フィールドが [**ドキュメントへのパス名**] を示すよう指定し、テキストの単位オプションを [**パラグラフ モード**] に設定した場合にのみ使用できます。抽出で使用する文字の閾値を指定します。実際のサイズは、最も近いピリオドに丸められます。ドキュメント コレクションのテキストから作成される単語の関連性を典型とするには、抽出サイズが小さすぎないよう指定します。

- **最小**: 抽出で使用する文字の最小数を指定します。
- **最大**: 抽出で使用する文字の最大数を指定します。

入力エンコード: テキスト フィールドが [**ドキュメントへのパス名**] を示すよう指定した場合にのみ使用できます。デフォルトのテキスト エンコードを指定します。日本語以外のすべての言語について、指定された、または認識されたエンコードから ISO-8859-1 への変換が行われます。そのため、別のエンコードが指定されている場合であっても、抽出エンジンは処理前にテキストを ISO-8859-1 に変換します。ISO-8859-1 エンコード定義に一致しない文字は、スペースに変換されます。日本語テキストの場合、SHIFT_JIS、EUC_JP、UTF-8、または ISO-2022-JP のエンコード オプションから 1 つ選択できます。

リソースのコピー元: テキスト マイニング時、抽出は、[エキスパート] タブの設定だけではく、言語リソースに基づいて行われます。これらのリソースは、抽出時のテキストの処理方法の基本として機能し、コンセプト、タイプ、および TLA パターンを取得します。リソース テンプレートからリソースをテキスト マイニングモデル作成ノードにコピーできます。

リソース テンプレートは、特定のドメインまたは使用向けに調整された、事前定義済みライブラリおよび詳細な言語リソースおよび非言語リソースです。これらのリソースは、抽出時のデータの処理方法についての基本として機能します。[**読み込み**] をクリックし、リソースをコピーするテンプレートを選択します。

テンプレートを選択したときにテンプレートが読み込まれ、ストリームが実行されているときには読み込まれません。読み込んでいるときに、リソースのコピーがノードに保存されます。そのため、更新されたテンプレートを使用したい場合、ここで再読み込みを行う必要があります。 詳細は、 3 章 p.46 テンプレートおよび TAP からのリソースのコピー を参照してください。

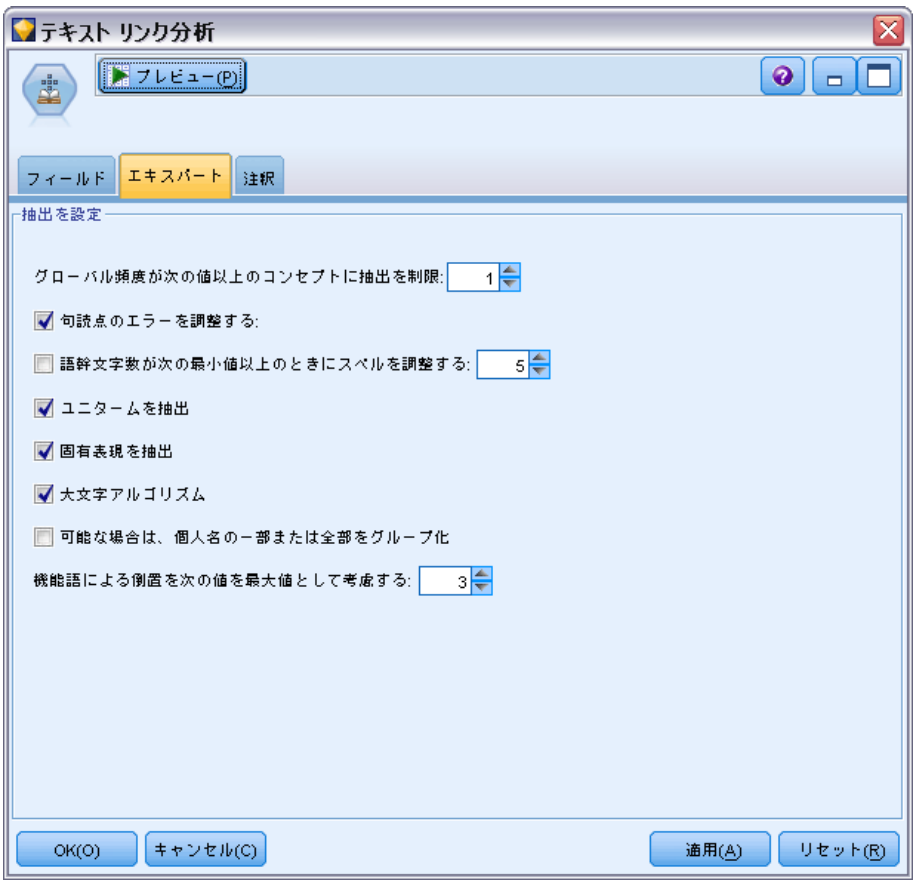
テキストの言語: マイニングされるテキストの言語を示します。ノードでコピーされたリソースが、表示される言語オプションを制御します。リソースを調整した言語を選択するか、[すべて] オプションを選択できます。テキスト データの正確な言語を指定することをお勧めしますが、不明な場合は、[すべて] オプションを選択できます。[すべて] は、日本語テキストには使用できません。自動言語認識を使用してすべてのドキュメントをスキャンして記録し、テキスト言語を最初に特定するため、この [すべて] オプションを選択すると、実行時間が長くなります。このオプションを使用して、サポートされライセンス許可された言語のすべてのレコードまたはドキュメントが、言語に適した内部辞書を使用して、抽出エンジンによって読み込まれます。詳細は、 18 章 p. 366 言語の識別子 を参照してください。 現在使用できないサポート言語のライセンス購入については、営業担当者に連絡してください。

テキスト リンク分析ノード:[エキスパート] タブ

このノードでは、テキスト リンク分析 (TLA) パターン結果の抽出が自動的に有効化されています。[エキスパート] タブには、テキストの抽出方法および処理方法に影響を与える追加パラメータがあります。このダイアログ ボックスのパラメータは、抽出プロセスの基本的な操作、そしていくつかの高度な操作を制御します。また、抽出結果にも影響を与える言語リソースやオプションも数多くあり、選択するリソース テンプレートによって制御します。

オランダ語、英語、フランス語、ドイツ語、イタリア語、ポルトガル語、スペイン語のテキストの場合

図 4-2
[テキスト リンク分析ノード] ダイアログ ボックス:[エキスパート] タブ



句読点のエラーを調整する: 抽出時に句読点エラー（不適切な使用方法など）を含むテキストを一時的に正規化し、コンセプトの抽出可能性を向上させます。自由記述式アンケートの回答、電子メール、CRM データなど、テキストが短く品質が悪い場合、またはテキストに略語が多く含まれている場合に特に役立ちます。

語幹文字数が次の最小値以上のときにスペルを調整する: Fuzzy Grouping の手法を適用し、共通してスペルミスのある単語またはスペルの近い単語を 1 つのコンセプトにグループ化できるようにします。Fuzzy Grouping アルゴリズムでは、最初の母音を除くすべての母音を一時的に抜き取った後抽出した単語から 2 つ/3 つの子音を抜き取り、それらを比較して、それらが同じで `modeling` と `modelling` が同じグループに分けられるかどうかを確認します。ただし、各キーワードが `<Unknown>` タイプを除いて、別のタイプに割り当てられた場合、Fuzzy Grouping 手法は適用されません。

Fuzzy Grouping を使用する前に必要な、語幹文字数の制限を定義することもできます。キーワード内の語幹文字数は、すべての文字を合計し、活用語尾、複合語キーワードの場合は区切り文字および前置詞を形成する文字を差し引いて計算します。たとえば、キーワード exercises の語幹文字数は「exercise」という形式で 8 文字と数えられます。語末の s は活用語尾（複数形）であるためです。同様に、apple sauce の語幹文字は 10 文字（「apple sauce」）、そして manufacturing of cars の語幹文字は 16 文字（「manufacturing car」）となります。この算出方法は、Fuzzy Grouping を適用するべきかどうかを確認するためにのみ使用されますが、単語がどのように一致するかについては影響を与えません。

注:特定の単語が後で不適切にグループ化されていることが分かった場合、[アドバンス リソース] タブの Fuzzy Grouping: **例外** セクションで 明示的に宣言することによって、単語のペアをこの手法から除外できます。詳細は、 18 章 p. 357 Fuzzy Grouping を参照してください。

ユニタームを抽出: 単語が複合語の一部でない限り、または名詞、またはスピーチ内の認識できない部分である場合、このオプションは単一の単語（ユニターム）を抽出します。

固有表現を抽出: 電話番号、セキュリティ番号、時間、日付、通貨、数字、パーセント、電子メールアドレス、HTTP アドレスなどの固有表現を抽出します。[アドバンス リソース] タブの **[固有表現: 設定]** セクションで、特定の種類の固有表現を追加したり除外したりできます。不要なエンティティを無効にすることにより、抽出エンジンは処理時間を節約できます。詳細は、 18 章 p. 362 構成 を参照してください。

大文字アルゴリズム: キーワードの最初の文字が大文字である場合、組み込み辞書にない単純キーワードおよび複合キーワードを抽出します。このオプションには、最も適切な名詞を抽出するのに優れた方法があります。

可能な場合は、個人名の一部または全部をグループ化: テキスト内で別々の形式で同時に出現する名前をグループ化します。名前はテキストの始めでは完全な形式で、後は短い形式でのみ参照されるため、この機能が役立ちます。このオプションでは、タイプが <Unknown> のユニタームが、タイプ <Person> の複合キーワードの最後の単語に一致するようにします。たとえば、doe があり、最初タイプが <Unknown> である場合、抽出エンジンは、<Person> タイプの複合キーワードに最後の単語として doe が含まれているかどうか（例: john doe）を確認します。ほとんどがユニタームとして抽出されることがないため、人の名前に適用されることはありません。

機能語による倒置を次の値を最大値として考慮する: 倒置手法を適用する場合に指定されている場合がある非機能的単語の最大数を指定します。この倒置手法では、活用語尾に関係なく、含まれる非機能的単語（of や the など）によってお互いに異なる類似した句をグループ化します。たとえば、この値を最大 2 単語に設定し、company officials および officials of the company が抽出されたとします。この場合、両方の抽出キーワード

は、of the が無視されると同じであるとみなされるため、最終コンセプト リストに共にグループ化されます。

日本語テキストの場合

図 4-3
[テキスト リンク分析ノード] ダイアログ ボックス:[エキスパート] タブ (日本語テキスト)

日本語テキストの場合、どの二次分析を適用するか選択することができます。

注:日本語テキスト展開はIBM® SPSS® Modeler Premiumで利用可能です。

二次分析: 抽出が開始したとき、基本キーワード抽出が、タイプのデフォルト セットを使用して行われます。詳細は、 A 付録 p. 412 日本語テキストで使用できるタイプ を参照してください。 ただし、二次分析機能を選択すると、抽出機能にコンセプトの一部として助詞や助動詞が含まれているため、より多くの詳細なコンセプトを取得することができます。たとえば、「肩の荷が下りた」という文があるとします。この例では基本

キーワード抽出は各コンセプトを個別に抽出します。例：**肩**（肩）、**荷**（重量）、**下りる**（上げ）、しかしこれらの単語の間の関係性は抽出されません。しかし、感性分析を適用すると、コンセプト **=肩の荷が下りた** というより高次の意味をもつコンセプトが抽出され、<**良い-安心**> という感性タイプとして割り当てられます。なお感性分析については、このほかに多くの感性タイプがあります。さらに、二次分析機能を選択すると、テキスト リンク分析結果も生成できます。

注:二次分析を呼び出すと、抽出プロセスにより時間がかかります。詳細は、 A 付録 p.404 二次抽出の手順 を参照してください。

- **係り受け解析**:このオプションを選択すると、キーワード、およびキーワードに助詞等を加えた語を、基本の品詞タイプのコンセプトとして抽出します。また係り受けテキスト リンク分析（TLA）に基づいたパターン結果を出力します。
- **感性分析**:この分析機能を選択すると、追加の抽出コンセプト、および可能な場合は、TLA パターン結果の抽出が行われます。基本タイプのほか、嬉しい、吉報、幸運、安心、幸福など 80 を超える感性タイプも利用できます。これらのタイプを使用して、感情、感性、意見の表現のテキストでコンセプトおよびパターンを検出します。感性分析に対するフォーカスを指示するオプションは３つあります：**全ての感性**、**代表的な感性のみ**、それと**結論のみ**。

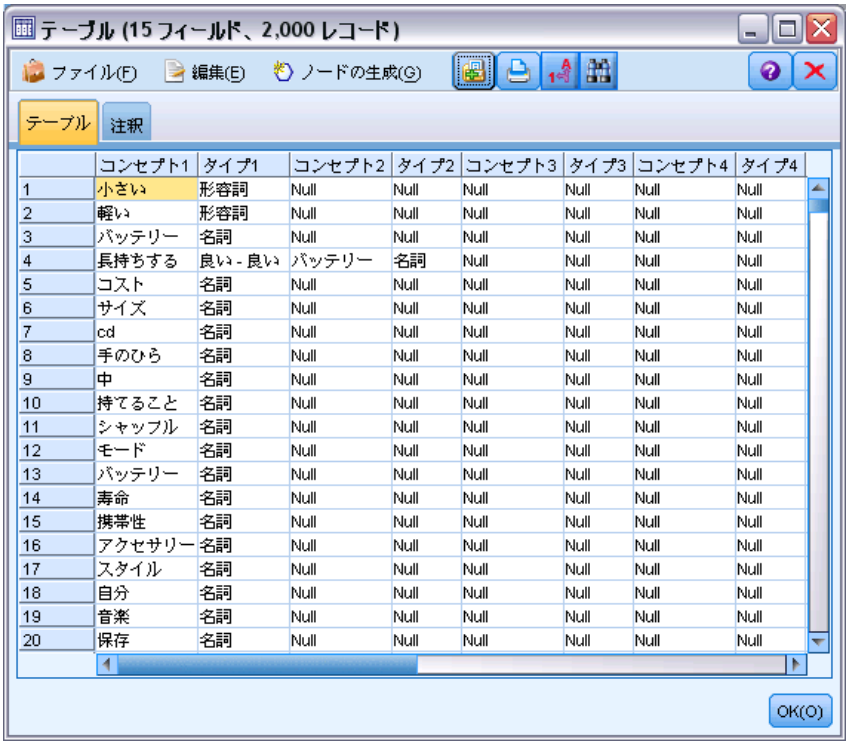
TLA ノード出力

テキスト リンク分析ノードを実行した後、データが再構築されます。テキスト マイニングでデータを再構築する方法を理解することは重要です。データ マイニングに異なる構造が必要な場合、[フィールド操作] パレットのノードを使用して、これを実行できます。たとえば、各行にテキスト レコードが表示されているデータを処理している場合、ソース テキスト データで見つかったパターンごとに 1 行ずつ作成されます。出力の各行について、次の 15 個のフィールドがあります。

- 6 つのフィールド（**コンセプト 1**、**コンセプト 2**、および**コンセプト 6** のような**コンセプト #**）は、パターン マッチで見つかったコンセプトを示します。
- 6 つのフィールド（**タイプ 1**、**タイプ 2**、および**タイプ 6** のような**タイプ #**）は、各コンセプトのタイプを示します。
- **条件規則名**は、テキストを合致させ、出力を生成するのに使用するテキスト リンク規則の名前を示します。
- ノードで指定した ID フィールドの名前を使用し、入力データと同じようにレコード ID またはドキュメント ID を示すフィールド。
- [**合致するテキスト**]は、元のレコードまたはドキュメント内にある、TLA パターンに合致したテキスト データの部分を示します。

注:日本語テキストのテキスト リンク分析パターン規則では、1 つまたは 2 つのスロットのパターン結果のみを生成します。

図 4-4
テーブル ノードに表示される出力



	コンセプト1	タイプ1	コンセプト2	タイプ2	コンセプト3	タイプ3	コンセプト4	タイプ4
1	小さい	形容詞	Null	Null	Null	Null	Null	Null
2	軽い	形容詞	Null	Null	Null	Null	Null	Null
3	バッテリー	名詞	Null	Null	Null	Null	Null	Null
4	長持ちする	良い - 良い	バッテリー	名詞	Null	Null	Null	Null
5	コスト	名詞	Null	Null	Null	Null	Null	Null
6	サイズ	名詞	Null	Null	Null	Null	Null	Null
7	cd	名詞	Null	Null	Null	Null	Null	Null
8	手のひら	名詞	Null	Null	Null	Null	Null	Null
9	中	名詞	Null	Null	Null	Null	Null	Null
10	持てること	名詞	Null	Null	Null	Null	Null	Null
11	シャッフル	名詞	Null	Null	Null	Null	Null	Null
12	モード	名詞	Null	Null	Null	Null	Null	Null
13	バッテリー	名詞	Null	Null	Null	Null	Null	Null
14	寿命	名詞	Null	Null	Null	Null	Null	Null
15	携帯性	名詞	Null	Null	Null	Null	Null	Null
16	アクセサリー	名詞	Null	Null	Null	Null	Null	Null
17	スタイル	名詞	Null	Null	Null	Null	Null	Null
18	自分	名詞	Null	Null	Null	Null	Null	Null
19	音楽	名詞	Null	Null	Null	Null	Null	Null
20	保存	名詞	Null	Null	Null	Null	Null	Null

注:リリース 5.0 以前のテキスト リンク分析ノードを含む既存のストリームは、ノードを更新しない限り、完全に実行できない可能性があります。IBM® SPSS® Modelerの最新バージョンでの特定の機能改善には、古いノードを新しいバージョンに置き換える必要があります。これにより、さらに展開可能で強力となります。

特定の言語の自動翻訳も実行できます。この機能を使用して、話せないまたは読めない言語のドキュメントをマイニングできます。翻訳機能を使用したい場合は、SDL Software as a Service (SaaS)へのアクセスが必要です。 詳細は、 5 章 p.106 翻訳設定 を参照してください。

TLA 結果のキャッシュ

キャッシュすると、テキスト リンク分析結果がストリーム内に置かれます。ストリームの実行ごとにテキスト リンク分析結果の抽出が繰り返されないようにするには、テキスト リンク分析ノードを選択して、メニューから **[編集] → [ノード] → [キャッシュ] → [有効化]** を選択します。次回ストリームが実行される場合、出力がノードにキャッシュされます。ノードのアイコン

には、小さい「ドキュメント」のグラフィックが表示され、キャッシュがいっぱいになると白から緑に変わります。キャッシュはセッションの間保存されます。ストリームを閉じて再び開いた後など、キャッシュを別の日にも保持するには、ノードを選択して、メニューから [**編集**] → [**ノード**] → [**キャッシュ**] → [**キャッシュの保存**] を選択します。次にストリームを開く場合、翻訳を再度行わずに保存されたキャッシュを再度読み込むことができます。

また、ノードを右クリックして、コンテキスト メニューから [**キャッシュ**] を選択して、ノードのキャッシュを保存したり有効にしたりできます。

ストリーム内のテキスト リンク分析ノードの使用

テキスト リンク分析ノードを使用して、データにアクセスし、ストリーム内のコンセプトを抽出します。入力ノードを使用して、データにアクセスできます。

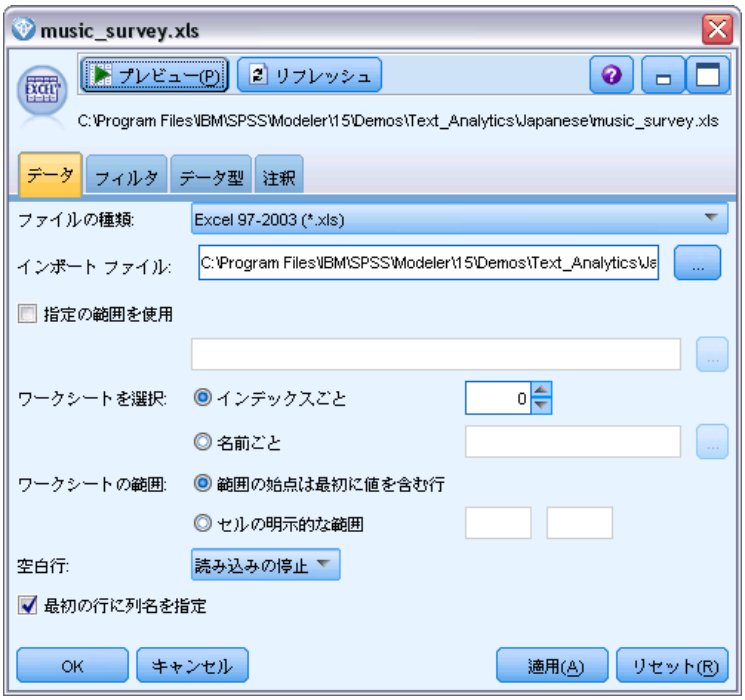
例:Statistics ファイル ノードとテキスト リンク分析ノード

次の例には、テキスト リンク分析ノードの使用方法が示されています。

図 4-5
例:Statistics ファイル ノードとテキスト リンク分析ノード

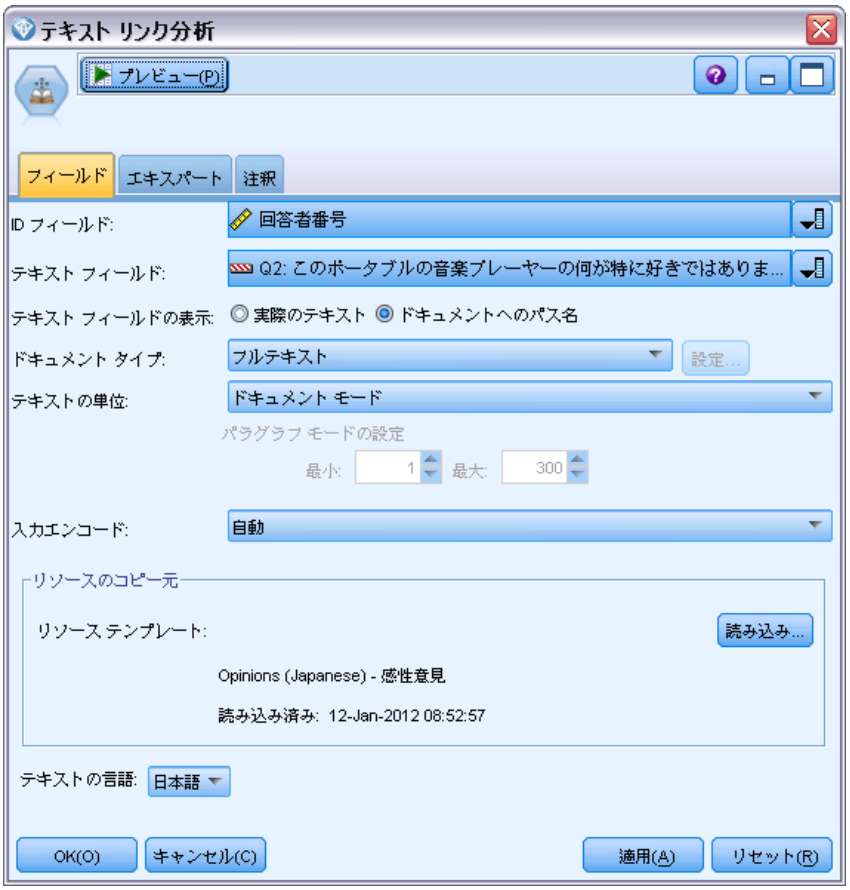
- **可変長ファイル ノード ([ファイル] タブ):** まず、このノードをストリームに追加して、テキスが保存されている場所を指定しました。

図 4-6
Statistics ファイル ノード ダイアログ ボックス:[データ] タブ



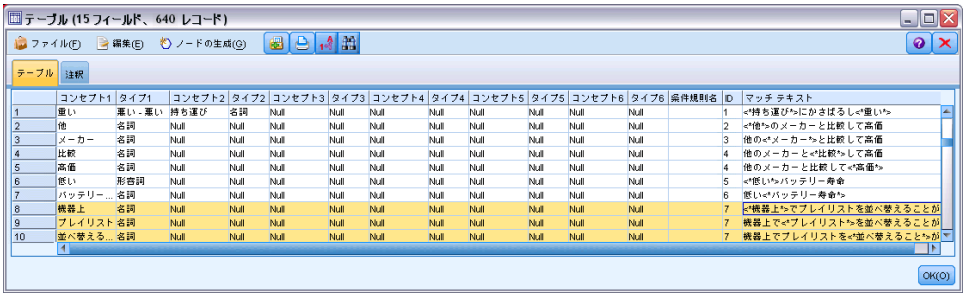
- **テキスト リンク分析ノード ([フィールド] タブ):** このノードをストリームに接続して、コンセプトを抽出し、下流でモデル作成または表示しました。ID フィールドおよびデータを含むテキスト フィールド名、そしてその他の設定を指定しました。

図 4-7
[テキスト リンク分析ノード] ダイアログ ボックス:[フィールド] タブ



- **テーブル ノード:** 最後にテーブル ノードを接続して、テキスト ドキュメントから抽出されたコンセプトを表示しました。表示されるテーブル出力で、このストリームがテキスト リンク分析ノードで実行された後、データ内の TLA パターン結果を確認できます。いくつかの結果で、合致したコンセプト/タイプが 1 つだけであることを示します。他の結果はより複雑で、いくつかのタイプおよびコンセプトが含まれています。また、テキスト リンク分析ノードを使用してデータを実行し、コンセプトを抽出した結果、データのいくつかの部分が変化しています。例の元のデータには、8 つのフィールドと 405 件のレコードが含まれていました。テキスト リンク分析ノードを実行した後、フィールド数は 15 で、レコード数は 640 件となります。TLA パターン結果ごとに 1 行ずつ割り当てらてます。たとえば、ID 7 は、3 つの TLA パターン結果が抽出されているため、3 行となります。この出力を元のデータに結合したい場合、結合ノードを使用できます。

図 4-8
テーブル出力ノード



	コンセプト1	タイプ1	コンセプト2	タイプ2	コンセプト3	タイプ3	コンセプト4	タイプ4	コンセプト5	タイプ5	コンセプト6	タイプ6	条件規則名	ID	マッチ テキスト
1	重い	重い - 重い	持ち運び	名詞	Null	Null	Null	Null	Null	Null	Null	Null		1	<$持ち運び>にかさばるし<$重い>
2	他	名詞	Null	Null	Null	Null	Null	Null	Null	Null	Null	Null		2	<$他>のメーカーと比較して高価
3	メーカー	名詞	Null	Null	Null	Null	Null	Null	Null	Null	Null	Null		3	他の<$メーカー>と比較して高価
4	比較	名詞	Null	Null	Null	Null	Null	Null	Null	Null	Null	Null		4	他のメーカーと<$比較>して高価
5	高価	名詞	Null	Null	Null	Null	Null	Null	Null	Null	Null	Null		4	他のメーカーと比較して<$高価>
6	低い	形容詞	Null	Null	Null	Null	Null	Null	Null	Null	Null	Null		5	<$低い>バッテリー寿命
7	バッテリー...	名詞	Null	Null	Null	Null	Null	Null	Null	Null	Null	Null		6	低い<$バッテリー寿命>
8	機器上	名詞	Null	Null	Null	Null	Null	Null	Null	Null	Null	Null		7	<$機器上>でプレイリストを並べ替えることが
9	プレイリスト	名詞	Null	Null	Null	Null	Null	Null	Null	Null	Null	Null		7	機器上で<$プレイリスト>を並べ替えることが
10	並べ替える...	名詞	Null	Null	Null	Null	Null	Null	Null	Null	Null	Null		7	機器上でプレイリストを<$並べ替えること>が

章
5

抽出するテキストの翻訳

翻訳ノード

翻訳ノードを使用して、IBM® SPSS® Modeler Text Analytics を使用して分析するアラビア語、中国語、ペルシア語などのサポート言語から、英語または他の言語にテキストを翻訳できます。このノードで、そのままではサポートされない 2 バイト言語のドキュメントのマイニングが可能になります。さらに、アナリストが該当する言語を理解できないとしても、外国語のドキュメントからコンセプトを抽出できるようになります。翻訳ノードを使えるようになるにはSDL’s Software as a Service (SaaS)にアクセスしなければいけないことにご注意下さい。

これらの言語のいずれかでテキスト マイニングを行う場合、ストリームのテキスト マイニングモデル作成ノードの前に翻訳ノードを追加してください。翻訳ノードでキャッシュし、ストリームを実行するごとに翻訳が繰り返されないようにすることもできます。

このノードは、IBM® SPSS® Modeler ウィンドウの下部にあるノード パレットの SPSS Modeler Text Analytics タブにあります。 詳細は、 1 章 p.12 IBM SPSS Modeler Text Analytics ノード を参照してください。

翻訳のキャッシュ: 翻訳をキャッシュする場合、翻訳されたテキストは、外部ファイルではなくストリームに保存されます。ストリームの実行ごとに翻訳が繰り返されないようにするには、翻訳ノードを選択して、メニューから **[編集] → [ノード] → [キャッシュ] → [使用する]** を選択します。次のストリームが実行される場合、翻訳の出力がノードにキャッシュされます。ノードのアイコンには、小さい「ドキュメント」のグラフィックが表示され、キャッシュがいっぱいになると白から緑に変わります。キャッシュはセッションの間保存されます。ストリームを閉じて再び開いた後など、キャッシュを別の日にも保持するには、ノードを選択して、メニューから **[編集] → [ノード] → [キャッシュ] → [キャッシュの保存]** を選択します。次にストリームを開く場合、翻訳を再度行わずに保存されたキャッシュを再度読み込むことができます。

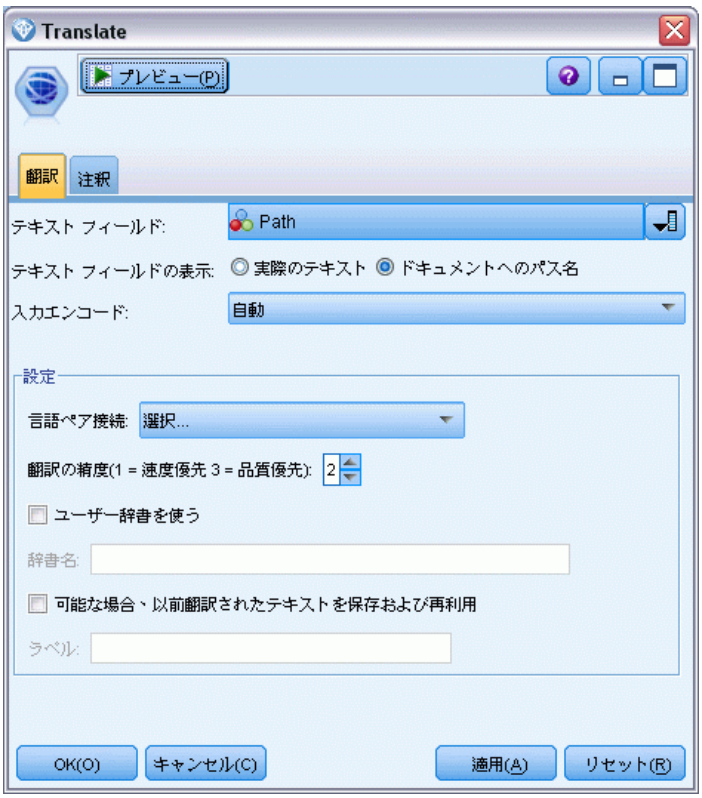
また、ノードを右クリックして、コンテキスト メニューから **[キャッシュ]** を選択して、ノードのキャッシュを保存したり有効にしたりできます。

重要! プロキシ サーバーを経由して Web の情報を取得しようとしている場合、SPSS Modeler Text Analytics Client および Server の net.properties ファイルでプロキシ サーバーを有効にする必要があります。このファイル内の説明に従ってください。これは、Web フィード ノードを使用して Web にアクセスする場合または Language Weaver ライセンスを取得する場合に適用されます。これらの場合、接続が を経由するためです。クライアントの場合、デフォルトでは、ファイルは C:¥Program

Files¥IBM¥SPSS¥Modeler¥15¥jre¥lib¥net.properties にあります。サーバーの場合、デフォルトでは、ファイルは C:¥Program Files¥IBM¥SPSS¥Modeler¥15ext¥bin¥spss.TMWBServer¥jre¥lib¥net.properties にあります。

翻訳ノード:[翻訳] タブ

図 5-1
[翻訳ノード] ダイアログ ボックス:[フィールド] タブ



テキスト フィールド: マイニングするテキスト、ドキュメントのパス名、またはドキュメントへのディレクトリ パス名が入力されたフィールドを選択します。このフィールドはデータ ソースによって異なります。Direction=None または Type=Typeless が指定されていても、任意の文字列フィールドを指定できます。

テキスト フィールドの表示: これまでの結果で指定されたテキスト フィールドに何が入力されているかを示します。選択されるのは次のとおりです。

- **実際のテキスト:** コンセプトが抽出される正確なテキストをフィールドに入力する場合、このオプションを指定します。
- **ドキュメントへのパス名:** 抽出するテキストを含む外部ドキュメントの場所へ1 つまたは複数のパス名をフィールドに入力する場合、このオプションを選択します。たとえば、ファイル リスト ノードを使用してドキュメントのリストを読み込む場合、このオプションを選択する必要があります。 詳細は、 2 章 p. 15 ファイル リスト ノード を参照してください。

入力エンコード: ソース テキストのエンコードを選択します。[**自動**] オプションを選択して開始できますが、いくつかのファイルが適切に処理されていないことがわかった場合、このリストから実際のエンコードを選択することをお勧めします。[自動] オプションを選択すると、短いデータベース レコードなど、短いテキストを処理する場合に、誤ってエンコードを特定する場合があります。このノードからのテキスト出力は、UTF-8 でエンコードされます。

設定: ストリームの翻訳設定を指定します。

- **言語ペア接続**:使用したい言語ペアを選んでください。利用可能な言語ペアはTranslation Settingsダイアログにあるサービスのリンクをセットした後に自動的にこのリストに表示されます。 詳細は、 p. 106 翻訳設定 を参照してください。
- **翻訳精度**:希望する精度に対する速度のレベルを示す 1 から 3 の値を選択し、希望する精度を指定します。値が小さいほど、翻訳結果が速く作成されますが、精度は劣ります。値が大きいと、精度の高い結果が作成されますが、時間がかかります。時間を最適化するには、低いレベルから始め、結果を見た後でより高い精度が必要がと感じた場合にのみ、レベルを高くすることをお勧めします。
- **カスタム辞書の使用**。以前に Language Weaver のカスタム辞書を作成している場合は、翻訳時に使用することができます。カスタム辞書を使用するには、**カスタム辞書を使用**チェックボックスを洗濯し、**辞書名**を入力します。複数の辞書を使用する場合は、辞書名をコンマで区切ってください。
- **可能な場合、以前の翻訳テキストを保存して再利用:** 翻訳結果を保存するよう指定し、また次回ストリームが実行されるときに同じ数のレコード/ドキュメントがある場合、コンテンツも同じであると想定し、翻訳結果を再利用して処理時間を短くします。実行時にこのオプションを選択し、レコード数が前回保存した場合と一致しない場合、テキストは完全に翻訳され、次回実行時のラベル名で保存されます。このオプションは、Language Weaver の翻訳言語を選択した場合にのみ使用できます。

注:テキストがストリーム内に保存されている場合、翻訳ノードでのキャッシュも可能です。この場合、キャッシュが有効なときは翻訳結果が再利用されるだけでなく、上級の結果は無視されます。

- **ラベル:[可能な場合、以前の翻訳テキストを保存して再利用]** を選択した場合、その結果のラベル名を指定する必要があります。このラベルを使用して、サーバーの以前に翻訳したテキストを特定します。ラベルが指定されていない場合、ストリームを実行すると警告がストリーム プロパティに追加され、再利用ができなくなります。

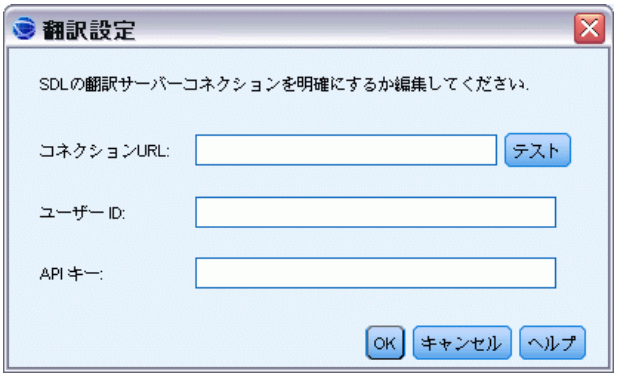
翻訳設定

このダイアログ ボックスで、翻訳するときにいつでも再利用できる Language Weaver 翻訳サーバーを定義および管理できます。接続をここで定義すると、すべての接続設定を再度入力することなく、翻訳時に言語ペア接続をすばやく選択できます。

言語ペア接続により、サーバーへの接続の詳細（LOC、WAN、または HTTP）のほか、ソース言語や翻訳言語を特定します。たとえば、[中国語 - 英語] は、ソース テキストが中国語で、翻訳テキストが英語であることを示します。SDLオンラインサービスを介してアクセスする各接続の定義を入力する必要があります。

重要! プロキシ サーバーを経由して Web の情報を取得しようとしている場合、IBM® SPSS® Modeler Text Analytics Client および Server の net.properties ファイルでプロキシ サーバーを有効にする必要があります。このファイル内の説明に従ってください。これは、Web フィード ノードを使用して Web にアクセスする場合または Language Weaver ライセンスを取得する場合に適用されます。これらの場合、接続が を経由するためです。クライアントの場合、デフォルトでは、ファイルは C:¥Program Files¥IBM¥SPSS¥Modeler¥15¥jre¥lib¥net.properties にあります。サーバーの場合、デフォルトでは、ファイルは C:¥Program Files¥IBM¥SPSS¥Modeler¥15ext¥bin¥spss.TMWBServer¥jre¥lib¥net.properties にあります。

図 5-2
[翻訳設定] ダイアログ



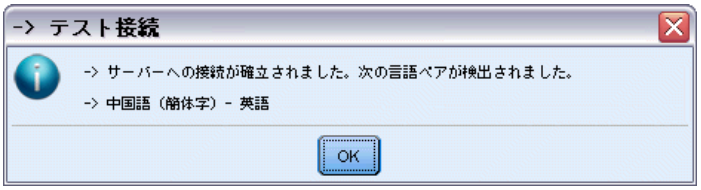
接続URL：SDL Software as a Serviceに接続するURLを入力してください。

ユーザー ID.:Language Weaverから入手した一意のIDを入力します。

API キー：Language Weaverから入手した一意のIDを入力します。

テスト接続を選択して [テスト] をクリックすると、接続が正しく設定されていることを検証し、その接続の言語ペアを表示します。

図 5-3
正常な接続のメッセージ



翻訳ノードの使用

アラビア語、中国語、ペルシア語など、サポートされた翻訳言語からコンセプトを抽出するには、ストリームのテキスト マイニング ノードの前に翻訳ノードを追加します。

例:外部ドキュメントのテキストの翻訳

翻訳するテキストが 1 つまたは複数の外部ファイルに含まれる場合、ファイル リスト ノードを使用して名前リスト内で読み込むことができます。この場合、翻訳ノードはファイル リスト ノードと後続のテキスト マイニング ノードの間に追加され、出力は翻訳テキストの場所にあります。

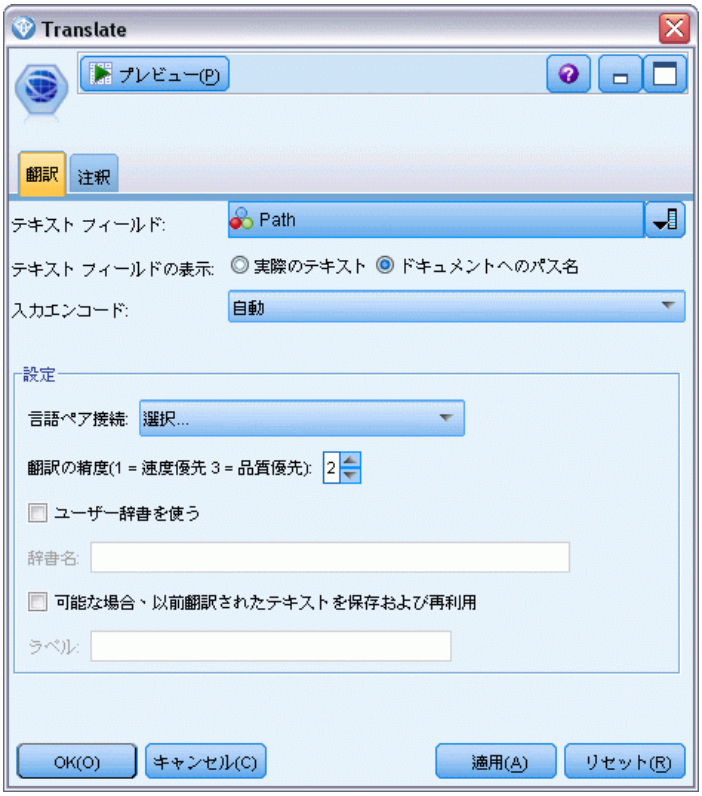
図 5-4
ストリームの例:翻訳ノードを使用した場合のファイル リスト ノード

- **ファイル リスト ノード ([設定] タブ):** ファイル リスト ノードでソース ファイルを選択しました。

図 5-5
[ファイル リスト ノード] ダイアログ ボックス[設定] タブ

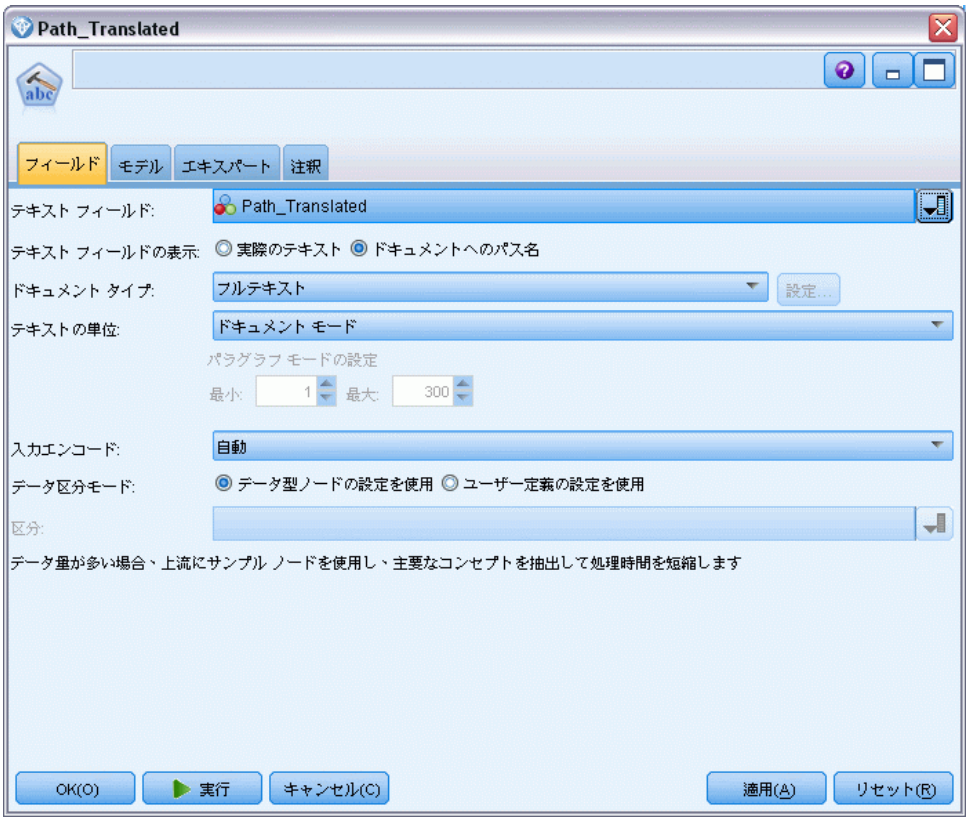
- **翻訳ノード ([翻訳] タブ):** 次に、翻訳ノードを追加して接続しました。ノードでは、ファイル リスト ノードで作成されたフィールドを選択しました。デフォルトでは、ファイルの元の場所を示すパスを指定します。同じタブで、事前定義された言語接続ペアを選択します。

図 5-6
[翻訳ノード] ダイアログ ボックス:[翻訳] タブ



- **テキスト マイニング ノード ([フィールド] タブ):** 後続のテキスト マイニング ノードで、翻訳ノードで作成された翻訳済みファイルの場所を指定する新しいフィールドを選択しました（_Translated の前にファイル リストノードのテキスト フィールドから命名）。

図 5-7
[テキスト マイニング モデル作成ノード] ダイアログ ボックス:[フィールド] タブ



- **テキスト マイニング ノード ([モデル] タブ):** ［モデル］ タブで、[英語] を言語に選択します。

図 5-8
[テキスト マイニング ノード] ダイアログ ボックス:[モデル] タブ

章
6

外部ソース テキストの参照

ファイル ビューア ノード

ドキュメントのコレクションをマイニングしている場合、ファイルの完全パス名をテキスト マイニング モデル作成ノードおよび翻訳ノードに直接指定できます。ただし、 テーブル ノードに出力する場合、テーブル ノード内のテキストではなく、ドキュメントの完全パス名のみ表示されます。ファイル ビューア ノードをテーブル ノードのアナログとして使用でき、すべてのドキュメントを 1 つのファイルに結合することなくドキュメント内の実際のテキストを使用できるようになります。

ストリームではアクセスできなため、ファイル ビューア ノードを使用して、コンセプトが抽出されたソース テキストまたは翻訳されていないテキストにアクセスを提供することによって、テキスト抽出の結果をより深く理解することができます。このノードは、ファイル リスト ノードの後のストリームに追加され、すべてのファイルへのリンクのリストを取得します。

このノードの結果として、コンセプトを抽出するために読み込み、使用されたすべてのドキュメント要素を示すウィンドウが表示されます。このウィンドウで、ツールバー アイコンをクリックして、ドキュメント名をハイパーリンクで表示する外部ブラウザでレポートを起動することができます。リンクをクリックして、コレクションの該当するドキュメントを開くことができます。 詳細は、 p. 113 ファイル ビューア ノードの使用 を参照してください。

このノードは、IBM® SPSS® Modeler ウィンドウの下部にあるノード パレットの IBM® SPSS® Modeler Text Analytics タブにあります。 詳細は、1 章 p. 12 IBM SPSS Modeler Text Analytics ノード を参照してください。

注:クライアントサーバー モードで作業し、ファイル ビューア ノードがストリームの一部である場合、ドキュメント コレクションはサーバーの Web サーバー ディレクトリに保存する必要があります。テキスト マイニング出力ノードは、Web サーバー ディレクトリ保存されているドキュメントのリストを作成するため、Web サーバーのセキュリティ設定により、これらのドキュメントに対する権限を管理します。

ファイル ビューア ノード設定

次のダイアログ ボックスを使用して、ファイル ビューア ノードの設定を指定します。

図 6-1
[ファイル ビューア ノード] ダイアログ ボックス[設定] タブ

ドキュメント フィールド: 表示するドキュメントの名前全体およびパスを含むデータからフィールドを選択します。

生成された HTML ページのタイトル: ドキュメントのリストを表示するページの最上部に表示されるタイトルを作成します。

ファイル ビューア ノードの使用

次の例には、ファイル ビューア ノードの使用方法が示されています。

例:ファイル リスト ノードおよびファイル ビューア ノード

図 6-2
ファイル ビューア ノードの使用を示すストリーム

- **ファイル リスト ノード ([設定] タブ):** まず、このノードを追加して、ドキュメントの場所を指定しました。

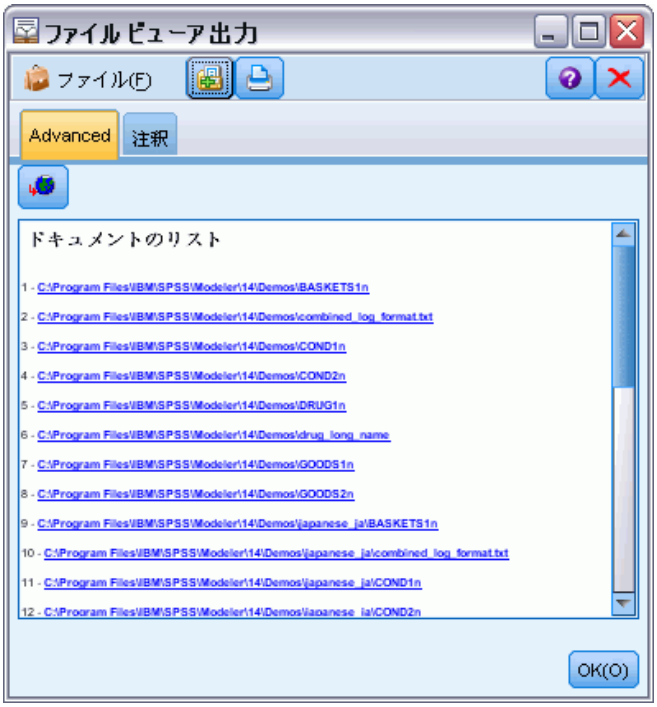
図 6-3
[ファイル リスト ノード] ダイアログ ボックス[設定] タブ

- **ファイル ビューア ノード ([設定] タブ):** 次に、 ファイル ビューア ノードを接続して、ドキュメントの HTML リストを作成しました。

図 6-4
[ファイル ビューア ノード] ダイアログ ボックス[設定] タブ

- **[ファイル ビューア出力] ダイアログ:** 次に、新しいウィンドウでドキュメントのリストを出力するストリームを実行しました。

図 6-5
ファイル ビューア出力



▶ ドキュメントを参照するために、赤い矢印の付いた地球儀の表示されたツールバーをクリックしました。ブラウザのドキュメントのハイパーリンクのリストが開きました。

章 7

スクリプト用のノードのプロパティ

IBM® SPSS® Modeler には、コマンド ラインからストリームを実行できるスクリプト言語があります。ここでは、IBM® SPSS® Modeler Text Analytics に付属する各ノードに固有のノードのプロパティについて説明します。SPSS Modeler に付属するノードの標準セットの詳細については、『Scripting and Automation Guide』を参照してください。

ファイル リスト ノード:filelistnode

スクリプトには、次の表のプロパティを使用できます。 このノードを filelistnode といいます。

テーブル 7-1
ファイル リスト ノードのスクリプトのプロパティ

スクリプトのプロパティ	データの型
path	文字列
recurse	フラグ
word_processing	フラグ
excel_file	フラグ
powerpoint_file	フラグ
text_file	フラグ
web_page	フラグ
xml_file	フラグ
pdf_file	フラグ
no_extension	フラグ

注：’Create list’ パラメータは使用できなくなり、そのオプションを含むスクリプトは自動的に ‘Files’ 出力に変換されます。

Web フィード ノード:webfeednode

スクリプトには、次の表のプロパティを使用できます。 このノードを webfeednode といいます。

テーブル 7-2
Web フィード ノードのスクリプトのプロパティ

スクリプトのプロパティ	データの型	プロパティの説明
urls	文字列1 文字列2 ... 文字列n	各 URL は、リスト構造で指定されます。 “¥n” で区切った URL リスト
recent_entries	フラグ	
limit_entries	整数	URL ごとに読み込む最新エントリ数:
use_previous	フラグ	Web フィード キャッシュを保存して再利用。
use_previous_label	文字列	保存した Web キャッシュの名前。
start_record	文字列	非 RSS の開始タグ.
urln.title	文字列	リスト内の各 URL について、ここでも定義する必要があります。 最初の URL は url1.title となり、番号は URL リストの位置に対応します。コンテンツのタイトルを含む開始タグです。
urln.short_description	文字列	urln.titleと同様です。
urln.description	文字列	urln.titleと同様です。
urln.authors	文字列	urln.titleと同様です。
urln.contributors	文字列	urln.titleと同様です。
urln.published_date	文字列	urln.titleと同様です。
urln.modified_date	文字列	urln.titleと同様です。
html_alg	None HTMLCleaner	コンテンツのフィルタリング方法。
discard_lines	フラグ	短い行を破棄。 min_wordsとともに使われる。
min_words	整数	最小単語数。
discard_words	フラグ	短い行を破棄。 min_avg_lenとともに使われる。
min_avg_len	整数	
discard_scw	フラグ	1 文字単語が多い行を破棄。 max_scwとともに使われる。
max_scw	整数	1 行につき最大 0 ～ 100 パーセントの割合の 1 文字単語
discard_tags	フラグ	特定のタグを含む行を破棄。
tags	文字列	特殊文字は、バックスラッシュ (¥) でエスケープする必要があります。
discard_spec_words	フラグ	特定の文字列を含む行を破棄。
words	文字列	特殊文字は、バックスラッシュ (¥) でエスケープする必要があります。

テキスト マイニング ノード:TextMiningWorkbench

次のパラメータを使用して、スクリプトを介してノードを定義または更新することができます。 このノードを TextMiningWorkbench といいます。

重要! スクリプトを介して異なるリソース テンプレートを指定できません。テンプレートが必要な場合は、ノードのダイアログ ボックスで選択する必要があります。

テーブル 7-3
テキスト マイニング モデル作成ノードのスクリプトのプロパティ

スクリプトのプロパティ	データの型	プロパティの説明
text	フィールド	
method	ReadText ReadPath	
docType	整数	正の数（0,1,2）を指定します。ここでは、0 = Full Text、1 = Structured Text、および 2 = XML となります。
encoding	Automatic "UTF-8" "UTF-16" "ISO-8859-1" "US-ASCII" "CP850" "EUC-JP" "SHIFT-JIS" "ISO2022-JP"	"UTF-8" のような特殊文字と値の組み合わせは、算術演算子と区別するために引用符（ “）で囲む必要があります。
unity	整数	正の数（0,1）を指定します。ここでは、0 = Paragraph、1 = Document となります。
para_min	整数	
para_max	整数	
mtag	文字列	すべての mtag 設定が含まれます(XML ファイルの［設定］ダイアログ ボックス)
mclef	文字列	すべての mclef 設定が含まれます(構造テキスト ファイルの［設定］ダイアログ ボックス)
partition	フィールド	
custom_field	フラグ	分割フィールドを指定するかどうかを示します。
use_model_name	フラグ	
model_name	文字列	
use_partitioned_data	フラグ	データ区分フィールドが定義されている場合、学習データだけがモデルの構築に使用されます。

スクリプトのプロパティ	データの型	プロパティの説明
model_output_type	Interactive Model	Interactiveはカテゴリモデルとなります。 Model はコンセプトモデルとなります。
use_interactive_info	フラグ	ワークベンチ セッションでインタラクティブに作成する場合のみ。
reuse_extraction_results	フラグ	ワークベンチ セッションでインタラクティブに作成する場合のみ。
interactive_view	Categories TLA Clusters	ワークベンチ セッションでインタラクティブに作成する場合のみ。
extract_top	整数	このパラメータは model_type = Conceptのときに使われます。
use_check_top	フラグ	
check_top	整数	
use_uncheck_top	フラグ	
uncheck_top	整数	
language	de en es fr it ja nl pt	
frequency_limit	整数	バージョン 14.0 では廃止。
concept_count_limit	整数	抽出を少なくともこの値以上のグローバル度数のコンセプトに制限 日本語テキストには使用不可
fix_punctuation	フラグ	日本語テキストには使用不可
fix_spelling	フラグ	日本語テキストには使用不可
spelling_limit	整数	日本語テキストには使用不可
extract_uniterm	フラグ	日本語テキストには使用不可
extract_nonlinguistic	フラグ	日本語テキストには使用不可
upper_case	フラグ	日本語テキストには使用不可
group_names	フラグ	日本語テキストには使用不可
permutation	整数	語順が異なる同義語で無視する非機能語の数（デフォルトは 3)。日本語テキストには使用不可。

スクリプトのプロパティ	データの型	プロパティの説明
jp_algorithmset conclusions only Representative only All Sentiments	0 1 2	日本語テキストの抽出の場合のみ。 注:IBM® SPSS® Modeler Premium内で利用可能。 0 = 感性二次抽出 1 = 依存関係抽出 2 = 二次分析は使用しない
jp_algorithm_sense_mode	0 1 2	日本語テキストの抽出の場合のみ。 注:SPSS Modeler Premium内で利用可能。 0 = 結論のみ 2 = 代表のみ 3 = すべての感性。

テキスト マイニング モデル ナゲット:TMWBModelApplier

スクリプトには、次の表のプロパティを使用できます。 このノードを TMWBModelApplier といいます。

テーブル 7-4
テキスト マイニング モデル ナゲットのプロパティ

スクリプトのプロパティ	データの型	プロパティの説明
scoring_mode	Fields Records	
field_values	Flags Counts	このオプションは、カテゴリのモデル ナゲットでは使用できません。 Flagsのためには、TRUEかFALSEをセットします。
true_value	文字列	Flags で、true の値を定義。
false_value	文字列	Flags で、true の値を定義。
extension_concept	文字列	フィールド名の拡張子を指定します。コンセプト名に加えてこの拡張子を使用して、フィールド名が生成されます。 add_as 値を使用して、この拡張子を使用する場所を指定します。
extension_category	文字列	フィールド名拡張子:フィールド名の拡張接頭辞/接尾辞を指定したり、カテゴリ コードを使用したりできます。カテゴリ名に加えてこの拡張子を使用して、フィールド名が生成されます。 add_as 値を使用して、この拡張子を使用する場所を指定します。
add_as	Suffix Prefix	
fix_punctuation	フラグ	

スクリプトのプロパティ	データの型	プロパティの説明
excluded_subcategories_descriptors	RollUpToParent Ignore	カテゴリ モデル のみ。サブカテゴリが選択されていない場合、スコアリングに選択されていないサブカテゴリに含まれる記述子の処理方法を指定できます。オプションは 2 つあります。 ■ オプション［記述子をスコアリングから完全に除外する］を選択すると、チェック記号のない（選択されていない）サブカテゴリは無視され、スコアリングに使用されません。 ■ オプション［記述子を上位カテゴリ内の記述子と合計する］を選択すると、チェック記号のない（選択されていない）サブカテゴリの記述子は上位カテゴリ（このサブカテゴリの上位にあるカテゴリ）の記述子として使用されます。複数レベルのサブカテゴリが選択されない場合、記述子は使用できる最初の上位カテゴリにロール アップされます。
check_model	フラグ	バージョン14 では廃止。
text	フィールド	
method	ReadText ReadPath	
docType	整数	ありえる値（0, 1, 2）0 = Full Text, 1 = Structured Text, 2 = XML
encoding	Automatic "UTF-8" "UTF-16" "ISO-8859-1" "US-ASCII" "CP850" "EUC-JP" "SHIFT-JIS" "ISO2022-JP"	"UTF-8" のような特殊文字と値の組み合わせは、算術演算子と区別するために引用符（ “）で囲む必要があります。
language	de en es fr it ja nl pt	

テキスト リンク分析ノード:textlinkanalysis

次の表のパラメータを使用して、スクリプトを介してノードを定義または更新することができます。 このノードを textlinkanalysis といいます。

重要! スクリプトを介してリソース テンプレートを指定できません。テンプレートを選択するには、ノードのダイアログ ボックスで選択する必要があります。

テーブル 7–5
テキスト リンク分析 (TLA) ノードのスクリプトのプロパティ

スクリプトのプロパティ	データの型	プロパティの説明
id_field	フィールド	
text	フィールド	
method	ReadText ReadPath	
docType	整数	正の数 (0, 1, 2) を指定します。ここでは、0 = Full Text、1 = Structured Text、および 2 = XML となります。
encoding	Automatic "UTF-8" "UTF-16" "ISO-8859-1" "US-ASCII" "CP850" "EUC-JP" "SHIFT-JIS" "ISO2022-JP"	"UTF-8" のような特殊文字と値の組み合わせは、算術演算子と区別するために引用符 (“) で囲む必要があります。
unity	整数	正の数 (0, 1) を指定します。ここでは、0 = Paragraph、1 = Document となります。
para_min	整数	
para_max	整数	
mtag	文字列	すべての mtag 設定が含まれます (XML ファイルの [設定] ダイアログ ボックス)
mclef	文字列	すべての mclef 設定が含まれます (構造テキスト ファイルの [設定] ダイアログ ボックス)
language	de en es fr it ja nl pt	
concept_count_limit	整数	抽出を少なくともこの値以上のグローバル度数のコンセプトに制限 日本語テキストには使用不可

スクリプトのプロパティ	データの型	プロパティの説明
fix_punctuation	フラグ	日本語テキストには使用不可
fix_spelling	フラグ	日本語テキストには使用不可
spelling_limit	整数	日本語テキストには使用不可
extract_uniterm	フラグ	日本語テキストには使用不可
extract_nonlinguistic	フラグ	日本語テキストには使用不可
upper_case	フラグ	日本語テキストには使用不可
group_names	フラグ	日本語テキストには使用不可
permutation	整数	語順が異なる同義語で無視する非機能語の数（デフォルトは 3）。日本語テキストには使用不可。
jp_algorithmset conclusions only Representative only All Sentiments	0 1 2	日本語テキストの抽出の場合のみ。 注:IBM® SPSS® Modeler Premium内で利用可能。 0 = 感性二次抽出 1 = 依存関係抽出 2 = 二次分析は使用しない
jp_algorithm_sense_mode	0 1 2	日本語テキストの抽出の場合のみ。 注:SPSS Modeler Premium内で利用可能。 0 = 結論のみ 2 = 代表のみ 3 = すべての感性。

翻訳ノード:translatenode

スクリプトには、次の表のプロパティを使用できます。 このノードを translatenode といいます。

テーブル 7-6
翻訳ノードのプロパティ

スクリプトのプロパティ	データの型	プロパティの説明
text	フィールド	
method	ReadText ReadPath	

スクリプトのプロパティ	データの型	プロパティの説明
encoding	Automatic "Big5", "Big5-HKSCS" , "UTF-8", "UTF-16", "US-ASCII", "Latin1", "CP850", "CP874", "CP1250", "CP1251", "CP1252", "CP1253", "CP1254", "CP1255", "CP1256", "CP1257", "CP1258", "GB18030", "GB2312", "GBK", "eucJP", "JIS7", "SHIFT_JIS", "eucKR", "TSCII", "ucs2", "KOI8-R", "KOI8-U", "ISO8859-1", "ISO8859-2", "ISO8859-3", "ISO8859-4", "ISO8859-5", "ISO8859-6", "ISO8859-7", "ISO8859-8", "ISO8859-8-i", "ISO8859-9", "ISO8859-10", "ISO8859-13", "ISO8859-14", "ISO8859-15", "IBM 850", "IBM 866", "Apple Roman", "TIS-620"	"UTF-8" のような特殊文字と値の組み合わせは、算術演算子と区別するために引用符（“）で囲む必要があります。
lw_server_type	LOC WAN HTTP	
lw_hostname	文字列	
lw_port	整数	
url	文字列	翻訳サーバーの URL
apiKey	文字列	
user_id	文字列	
lpid	整数	language_from または language_from_id が設定されている場合は、使用されません。
translate_from	Arabic, Chinese, Traditional Chinese, Czech, Danish, Dutch, English, French, German, Greek, Hindi, Hungarian, Italian, Japanese, Korean, Persian, Polish, Portuguese, Romanian, Russian, Spanish, Somali, Swedish	

スクリプトのプロパティ	データの型	プロパティの説明
translate_from_id	ara, chi, cht, cze, dan, dut, eng, fra, ger, gre, hin, hun, ita, jpn, kor, per, pol, por, rum, rus, som, spa, swe	
translate_to	English	
translate_to_id	eng	
translation_accuracy	整数	翻訳プロセスに必要な精度のレベルを、1 ～ 3 の値を選択して指定します。
use_previous_translation	フラグ	翻訳結果が以前の実行ですでに存在し、再使用できることを指定します。
translation_label	文字列	再利用するために翻訳結果を特定するラベルを入力します。

パート II:
インタラクティブワークベンチ

章

8

インタラクティブ ワークベンチ モード

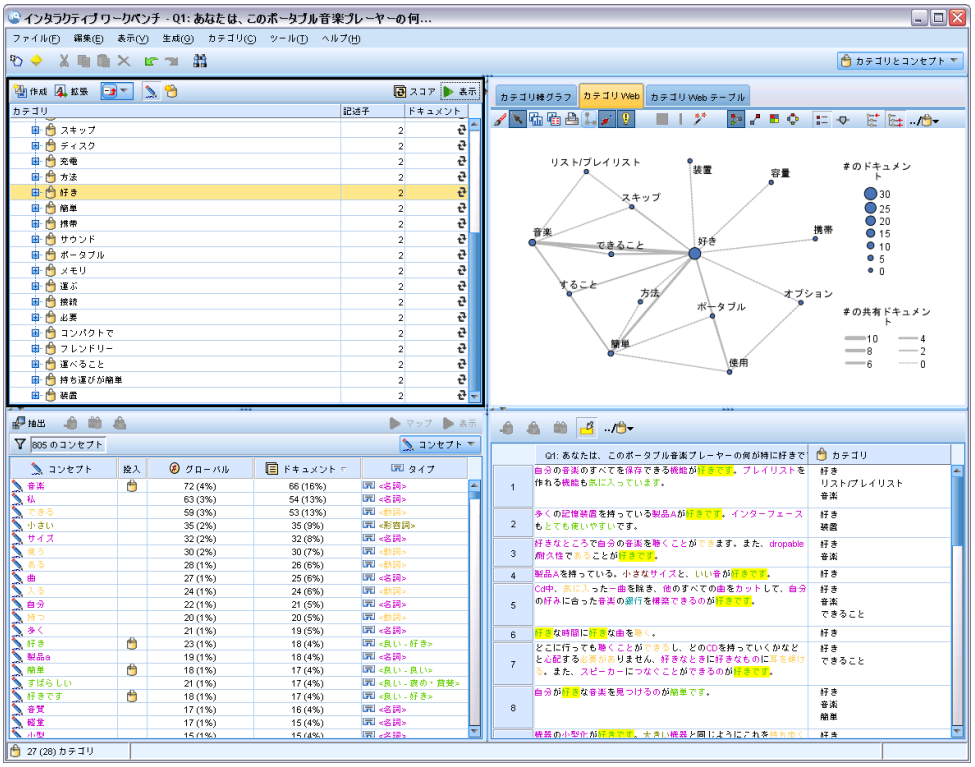
テキスト マイニングモデル作成ノードから、ストリーム実行時にインタラクティブ ワークベンチ セッションを起動することができます。このワークベンチでは、テキスト データから主要なコンセプトを抽出、カテゴリを作成、そしてテキスト リンク分析パターンおよびクラスタを検討、カテゴリ モデルを生成できます。この章では、次のような主要な要素とともに、ワークベンチ インターフェイスについて高レベルな視点から説明します。

- **抽出結果:** 抽出が実行された後、テキストデータから特定および抽出されるキー ワードおよび句で、「コンセプト」とも呼ばれます。これらのコンセプトは、「タイプ」にグループ化されます。これらのコンセプトおよびタイプを使用して、カテゴリを作成するほか、データを検討できます。これらは、**コンセプトおよびカテゴリ** ビューで管理されます。
- **カテゴリ型:** 抽出結果、パターン、ルールなどの記述子を定義として使用し、カテゴリ定義の一部を含むかどうかに基づいてドキュメントおよびレコードが割り当てられるカテゴリのセットを手動でまたは自動的に作成できます。これらは、**コンセプトおよびカテゴリ** ビューで管理されます。
- **クラスタ**クラスタは、間にリンクがあるコンセプトのグループ化で、コンセプト間の関係を示します。コンセプトは、その他の要素間で、2 つのコンセプトがそれぞれ現れる頻度と比較して 2 つのコンセプトが同時に現れる頻度を素養する複雑なアルゴリズムを使用してグループ化します。これらは、**クラスタ** ビューで管理されます。クラスタを構成するコンセプトをカテゴリに追加することもできます。
- **テキスト リンク分析パターン:** 言語リソースにテキスト リンク分析（TLA）パターンの規則がある場合、またはすでにいくつかの TLA 規則があるリソース テンプレートを使用している場合、テキスト データからパターンを抽出できます。これらのパターンを使用して、データのコンセプト間に興味深い関連を見つけることができます。また、これらのパターンをカテゴリの記述子として使用することもできます。これらは、**テキスト リンク分析** ビューで管理されます。日本語テキストの場合、二次分析を選択して、TLA 抽出をオンにする必要があります。注:日本語テキスト展開はIBM® SPSS® Modeler Premiumで利用可能です。
- **言語リソース:** 抽出プロセスは、テキストの抽出方法および処理方法を支配する一連のパラメータおよび言語定義によって異なります。これらは、**リソース エディタ** ビューのテンプレートおよびライブラリのフォームで管理されます。

カテゴリとコンセプト ビュー

アプリケーション インターフェイスは、いくつかのビューで構成されています。カテゴリとコンセプト ビューは、抽出結果を検討および調整するほか、カテゴリを作成および検討するウィンドウです。**カテゴリ**は、スコアリング プロセスでドキュメントおよびレコードが割り当てられる、密接に関連するキーワードおよびパターンのグループを参照します。**コンセプト**は、カテゴリの構築ブロック（記述子）として使用できる最も基本的なレベルの抽出結果を参照します。

図 8-1
分類とコンセプト ビュー



カテゴリとコンセプト ビューは 4 つのパネルで構成され、[表示] メニューから名前を選択して隠したり表示したりできます。 詳細は、 10 章 p. 177 テキスト データの分類 を参照してください。

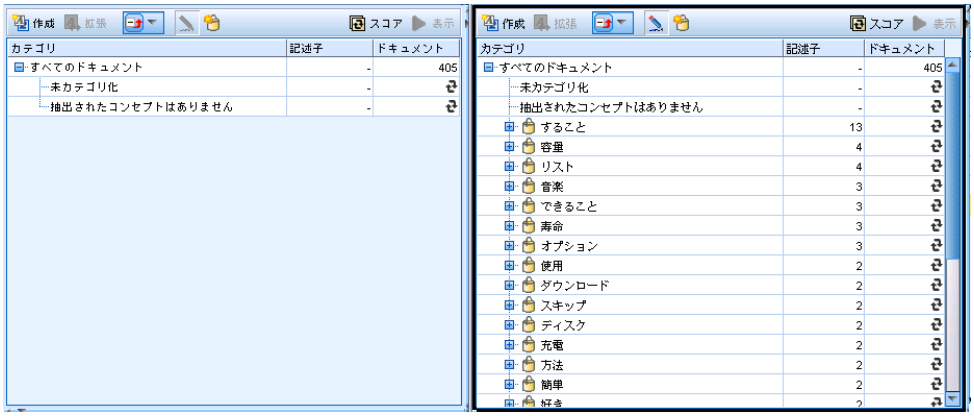
[カテゴリ] パネル

左上にあるこの領域は、構築したカテゴリを管理できる表が表示されます。テキスト データからコンセプトとタイプを抽出した後、セマンティック ネットワークや内包関係のコンセプトなどの方法を使用して、または手

動で作成してカテゴリを構築できます。カテゴリ名をダブルクリックすると、［カテゴリ定義］ダイアログボックスが開き、コンセプト、タイプ、規則など、定義を構成するすべての記述子が表示されます。詳細は、10 章 p. 177 テキスト データの分類 を参照してください。 すべての言語ですべての自動的手法が使用できるわけではありません。

パネルで 1 行選択すると、［データおよび視覚化］パネルに該当するドキュメント/レコードまたは記述子に関する情報が表示されます。

図 8-2
カテゴリとコンセプト ビュー:カテゴリがない場合とある場合の［カテゴリ］パネル



［抽出結果］パネル

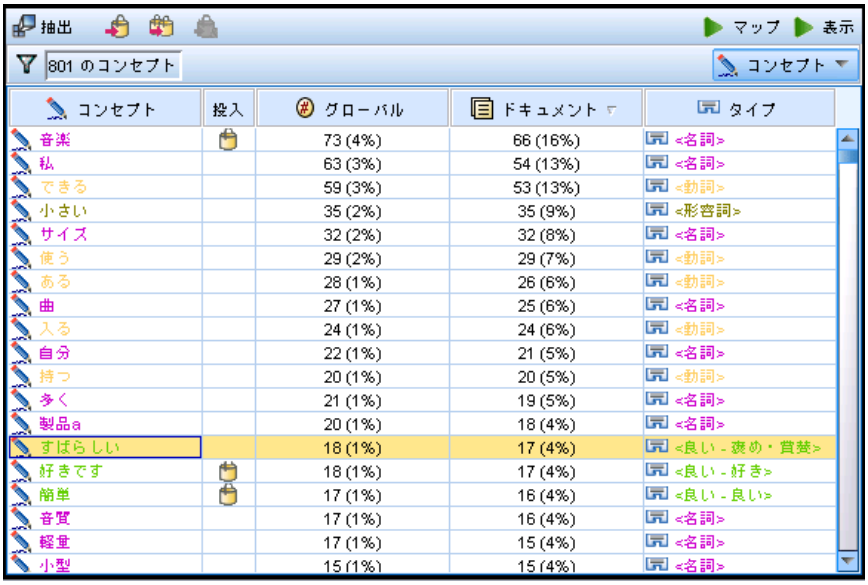
左下のこの領域には、抽出結果が表示されます。抽出を実行すると、抽出エンジンがテキスト データを読み込み、関連コンセプトを特定し、それぞれにタイプを割り当てます。**コンセプト**は、テキスト データから抽出した単語や句です。**タイプ**は、キーワード辞書の形式で保存されたコンセプトのセマンティック グループです。抽出が完了すると、コンセプトとタイプが［抽出結果］パネルにカラー コード化されて表示されます。 詳細は、9 章 p. 152 抽出結果:コンセプトとタイプ を参照してください。

コンセプト名の上にマウスポインタを置くと、コンセプトの基本キーワードのセットが表示されます。これにより、コンセプト名とそのコンセプトにグループ化された数行のキーワードを示すツールヒントが表示されます。これらの基本キーワードには、抽出された複数形/単数形のキーワード、置換キーワード、Fuzzy Grouping のキーワードなどのほか、テキスト内にあったかどうかに関係なく、言語リソースで定義された類義語が含まれます。コンセプト名を右クリックし、コンテキスト メニューのオプションを選択して、これらのキーワードをコピーしたり、基本キーワードの完全セットを表示したりできます。

テキスト マイニングは、抽出結果をテキスト データのコンテキストに従ってレビューし、新しい結果を作成するよう調整、そして再評価するインタラクティブ プロセスです。言語リソースを修正することによって、抽出結

果を修正できます。この調整は、[抽出結果] パネルまたは [データ] パネルから部分的に直接、またリソース エディタ ビューから直接実行できます。 詳細は、 p. 141 リソース エディタ ビュー を参照してください。

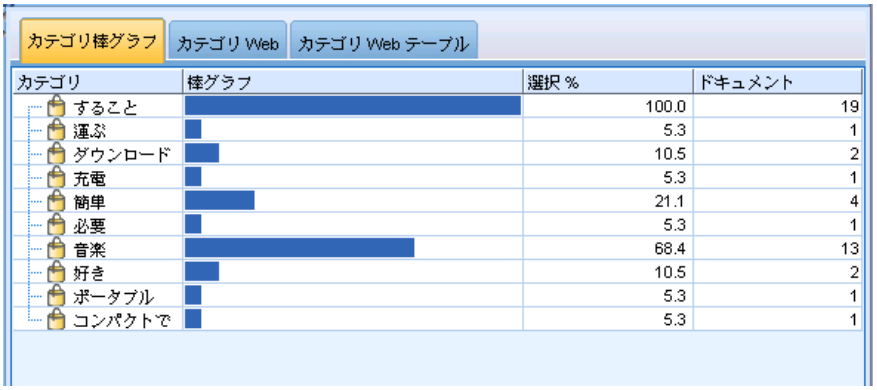
図 8-3
カテゴリとコンセプト ビュー:抽出後の [抽出結果] パネル



[視覚化] パネル

右上にあるこの領域には、ドキュメント/レコードのカテゴリ化の共通性について、さまざまな観点が表示されます。各グラフやチャートは類似の情報を提供しますが、異なる方法または異なる詳細レベルで表示します。これらの図表やグラフを使用して、カテゴリ化の結果を分析したり、カテゴリまたはレポートの調整を行うことができます。たとえば、グラフを使用して、あまりに類似している（75% を超えるレコードを共有しているなど）またはあまりに異なるカテゴリを見つけることができます。グラフまたは図表の内容は、その他のパネルでの選択内容に対応しています。 詳細は、13 章 p. 278 カテゴリ グラフおよび図表 を参照してください。

図 8-4
カテゴリとコンセプト ビュー:[視覚化] パネル

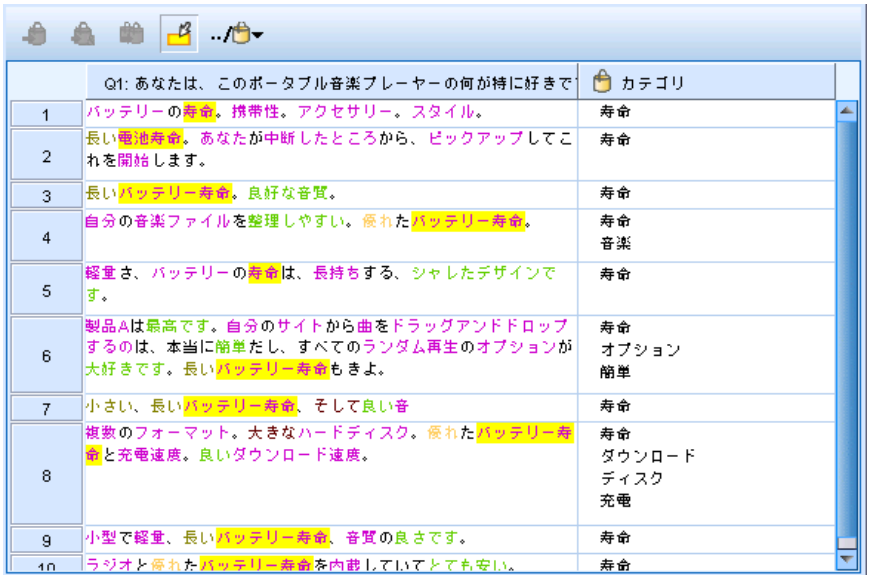


[データ] パネル

[データ] パネルは、右下に表示されます。このウィンドウには、ビューの別の領域での選択内容に対応するドキュメントまたはレコードを示すテーブルが表示されます。選択内容に応じて、対応するテキストのみが [データ] パネルに表示されます。選択した後、[**表示**] ボタンをクリックすると、[データ] パネルに対応するテキストが入力されます。

別のパネルで選択した場合、該当するドキュメントまたはレコードにはコンセプトが色付きで強調表示され、テキスト内のコンセプトを特定しやすくします。カラーコード化された項目上でマウス ポインタを停止させて、項目が抽出されたコンセプトの名前と、項目が割り当てられたタイプを示すヒントを表示することもできます。 詳細は、 10 章 p. 190 [データ] パネル を参照してください。

図 8-5
カテゴリとコンセプト ビュー:[データ] パネル



コンセプトとカテゴリ ビューでの検索

特定のセクションで、情報の迅速な検索が必要な場合があります。検索ツールバーを使用して、検索する文字列を入力し、大文字や小文字の区別および検索方向など、その他の検索基準を定義することができます。そして、検索するパネルを選択できます。

図 8-6
コンセプトとカテゴリ ビューの検索ツールバー

検索機能を使用するには

- カテゴリとコンセプト ビューのメニューで、[編集] → [検索] を選択します。[カテゴリ] パネルおよび [視覚化] パネルの上に検索ツールバーが表示されます。
- テキスト ボックスに検索したい文字列を入力します。ツールバー ボタンを使用して、大文字と小文字の区別、部分一致、検索の方向を制御します。
- ツールバーで、検索するパネル名をクリックします。一致が見つかった場合は、テキストがウィンドウで強調表示されます。
- 次の一致を検索するには、パネルの名前をもう一度クリックします。

クラスタ ビュー

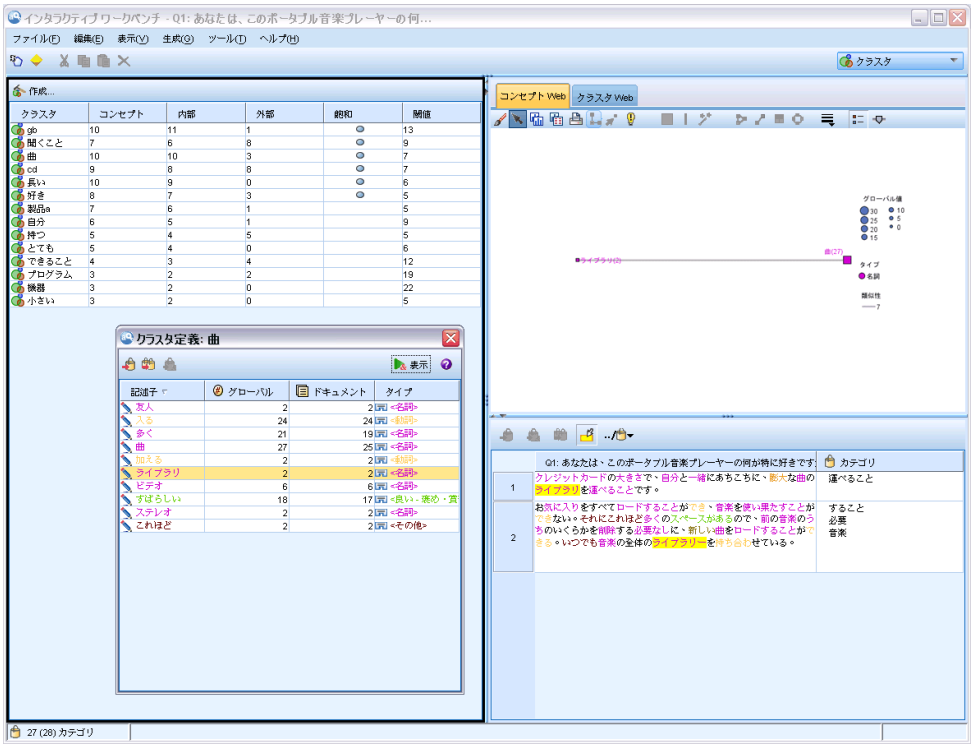
クラスタ ビューでは、テキスト データのクラスタ結果を構築および検討できます。**クラスタ**は、コンセプトが出現する頻度およびいっしょに出現する頻度に基づいてアルゴリズムをクラスタ化することによって生成されるコンセプトのグループです。カテゴリの目的は、含まれるテキストが各カテゴリの記述子（コンセプト、条件規則、パターン）にどのように合致するかに基づいてドキュメントまたはレコードをグループ化することですが、クラスタの目的は共起するコンセプトをグループ化することです。

クラスタ内のコンセプトが、その他のコンセプトと共に出現する低い頻度のコンセプトと共により頻繁に出現すればするほど、クラスタが興味深いコンセプトの関係性をより適切に特定します。2 つのコンセプトが同じドキュメントまたはレコードで現れると（または類義語かキーワードのいずれかが現れると）それらは共起します。 詳細は、 11 章 p. 259 クラスタの分析 を参照してください。

クラスタを構築し、見つけるのに時間がかかるコンセプト間の関係を発見することができる一連の図表およびグラフでクラスタを検討できます。クラスタ全体をカテゴリに追加できませんが、[クラスタ定義] ダイアログボックスを使用してクラスタのコンセプトをカテゴリに追加できます。 詳細は、 11 章 p. 266 クラスタ定義 を参照してください。

クラスタリングの設定に変更を行うと、結果に影響を与える場合があります。 詳細は、 11 章 p. 261 クラスタの作成 を参照してください。

図 8-7
クラスタ ビュー



クラスタ ビューは 3 つのパネルで構成され、[表示] メニューから名前を選択して隠したり表示したりできます。通常、表示されるのは [クラスタ] パネルと [視覚化] パネルだけです。

[クラスタ] パネル

左側にあるこの領域には、テキスト データで見つかったクラスタが表示されます。[作成] ボタンをクリックして、クラスタリングの結果を作成できます。クラスタは、クラスタリング アルゴリズムによって形成され、頻繁に共起するコンセプトを特定しようとします。

新しい抽出が行われると、クラスタ結果は消去され、クラスタを再構築して最新の結果を取得する必要があります。クラスタを構築している場合、作成すべき最大クラスタ数、含むことができる最大コンセプト数、使用できる外部コンセプトとの最大リンク数など、いくつかの設定を変更できます。 詳細は、 11 章 p. 265 クラスタの検証 を参照してください。

図 8-8
クラスタ ビュー:[クラスタ] パネル


作成...

クラスタ	コンセプト	内部	外部	飽和	閾値
gb	10	11	1	●	13
開くこと	7	6	8	●	9
曲	10	10	3	●	7
cd	9	8	8	●	7
長い	10	9	0	●	6
好き	8	7	3	●	5
製品a	7	6	1		5
自分	6	5	1		9
持つ	5	4	5		5
とても	5	4	0		6
できること	4	3	4		12
プログラム	3	2	2		19
機器	3	2	0		22
小さい	3	2	0		5

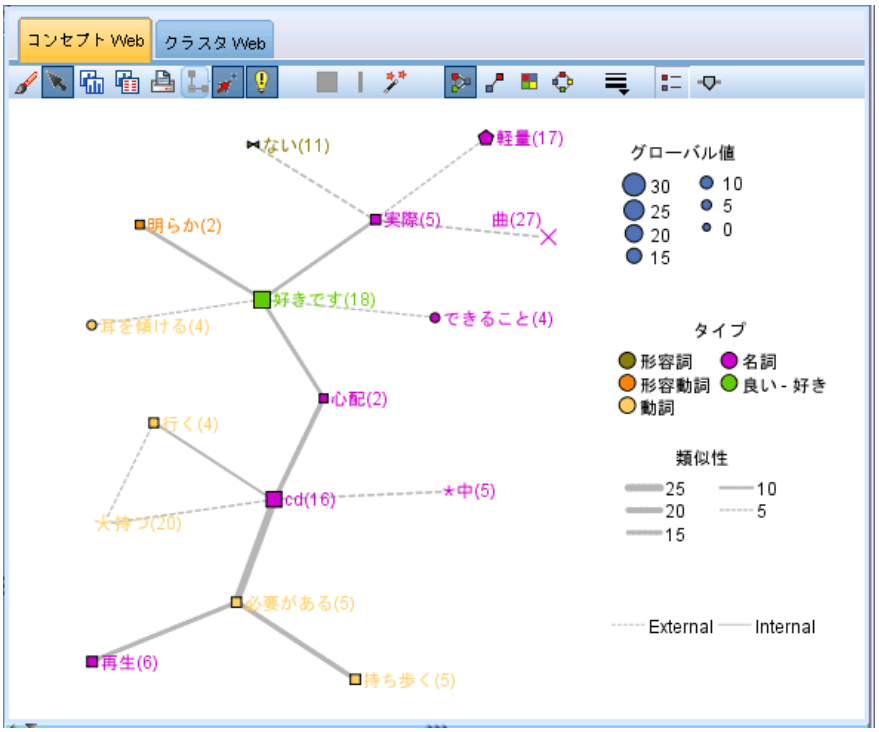
[視覚化] パネル

右上の角にあるこのパネルはクラスタ化の2つの観点を提供します：それは、コンセプト Web グラフと クラスタ Web グラフです。表示されない場合、[表示] メニュー（**[表示] → [視覚化]**）からこの領域にアクセスできます。クラスタのパネルでの選択内容によって、クラスタ間またはクラスタ内の該当する交互作用を表示できます。次のような形式で結果を表示します。

- **コンセプト Web:** クラスタ外のリンクしたコンセプトのほか、選択したクラスタ内のすべてのコンセプトを表示する Web グラフ。
- **クラスタ Web:** その他のクラスタ間のリンクのほか、選択したクラスタからのリンクを表示する Web グラフ。

注:クラスタ Web グラフを表示するには、外部リンクを持つクラスタを構築する必要があります。外部リンクは、別のクラスタのコンセプト ペア間のリンクです（あるクラスタのコンセプトと別のクラスタのコンセプトとの間）。 詳細は、 13 章 p. 282 クラスタ グラフ を参照してください。

図 8-9
クラスタ ビュー:[視覚化] パネル



[データ] パネル

[データ] パネルは、右下にあり、デフォルトでは表示されません。これらのクラスタの範囲は複数のドキュメント/レコードにわたり、データ結果は興味深いものではなくなるため、[クラスタ] 結果から [データ] パネルの結果を表示できません。ただし、[クラスタ定義] ダイアログ ボックス内の選択に対応するデータを表示できます。そのダイアログ ボックスの選択内容に応じて、[データ] パネルに対応するテキストのみが表示されます。選択を行うと、**[表示 &]** ボタンをクリックして、[データ] パネルに、すべてのコンセプトを同時に含むドキュメントまたはレコードを表示します。

該当するドキュメントまたはレコードには、コンセプトを色付きで強調表示し、テキスト内のコンセプトを特定しやすくします。カラーコード化された項目上でマウス ポインタを停止させて、項目が抽出されたコンセプトと、項目が割り当てられたタイプを表示することもできます。[データ] パネルには、複数の列が表示されますが、テキスト フィールド列は常に表示されます。抽出時に使用されたテキスト フィールドの名前、またはテキスト データがさまざまなファイルにある場合はドキュメント名が表示されます。その他の列も使用できます。 詳細は、 10 章 p. 190 [データ] パネル を参照してください。

テキスト リンク分析ビュー

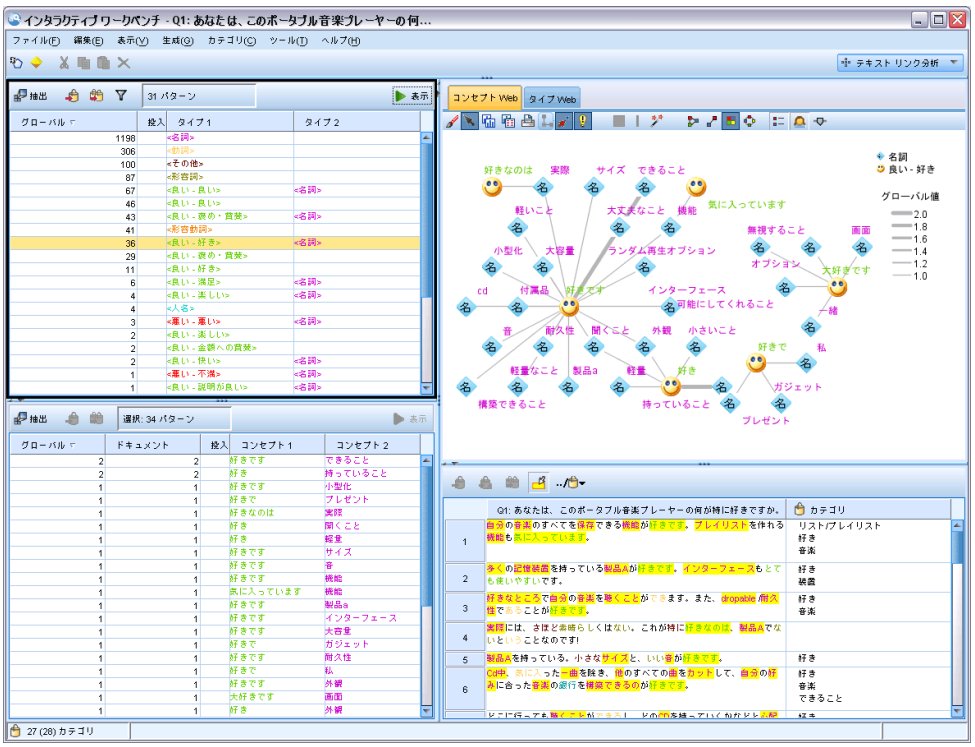
テキスト リンク分析ビューでは、テキスト データで見つかったテキスト リンク分析パターンを作成および検証できます。テキスト リンク分析 (TLA) はパターンマッチ手法で、TLA 規則を定義し、それらをテキスト内の実際の抽出されたコンセプトおよび関連性と比較することができます。

パターンは、コンセプト間の関連性または特定のサブジェクトに関する意見を探索する場合に最も役立ちます。たとえば、調査データで製品に関する意見、医療調査アンケートから遺伝子的関連性、または情報データから人名と地名との関連性を抽出したい場合などです。

TLA パターンを抽出すると、[データ] パネルまたは [視覚化] パネルで検証し、カテゴリとコンセプト ビューでそれらの結果をカテゴリに追加することができます。いくつかの TLA 規則が、TLA 結果を抽出するために使用するリソース テンプレートまたはライブラリで定義されています。 詳細は、 19 章 p. 368 テキスト リンク規則について を参照してください。

TLA パターン規則の抽出を選択している場合、結果はこのビューに表示されます。抽出を選択していない場合は、[**抽出**] ボタンを使用して、パターンの抽出を可能にするオプションを選択する必要があります。

図 8-10
テキスト リンク分析ビュー



テキスト リンク分析ビューは 4 つのパネルで構成され、[表示] メニューから名前を選択して隠したり表示したりできます。 詳細は、 12 章 p.268 テキスト リンク分析の検証 を参照してください。

[タイプ パターン] パネルおよび [コンセプト パターン] パネル

左側の [タイプおよびコンセプト パターン] パネルは、TLA パターン結果を検証および選択できる 2 つの相関したパネルです。パターンは、6 つのタイプまたは 6 つのコンセプトで構成されています。日本語テキストの場合、パターンのシリーズは、最大 1 ～ 2 のタイプまたはコンセプトです。言語リソースに定義されているため、TLA パターン規則は、パターン結果の複雑さを示します。詳細は、 19 章 p.368 テキスト リンク規則について を参照してください。注:日本語テキスト展開はIBM® SPSS® Modeler Premiumで利用可能です。

図 8-11
テキスト リンク分析ビュー:[タイプ パターン] パネルおよび [コンセプト パターン] パネル

パターン結果はまずタイプ レベルでグループ化され、コンセプト パターンに分割されます。このため、2つの異なる結果パネルがあります：それが、タイプ パターン（左上）とコンセプト パターン（左下）です。

- **タイプ パターン**：［タイプ パターン］は、TLA パターン規則を満たす 2 つ以上の関連タイプで構成されている抽出パターンが表示されます。タイプ パターンは、<**組織名**> + <**地名**> + <**肯定的**> と表され、特定の場所の組織について、肯定的なフィードバックを提供します。
- **コンセプト パターン**:［コンセプト パターン］パネルには、上の［タイプ パターン］で現在選択されているすべてのタイプ パターンのコンセプト レベルで抽出パターンが表示されます。コンセプト パターンは、**ホテル** + **パリ** + **すばらしい** などの構造に従います。

カテゴリとコンセプト ビューの抽出結果と同様、ここで結果を確認できます。これらのパターンを構成するタイプおよびコンセプトに調整を行う場合、カテゴリとコンセプト ビューの［抽出結果］パネルまたはリソース エディタで変更を行うか、パターンを再抽出します。

[視覚化] パネル

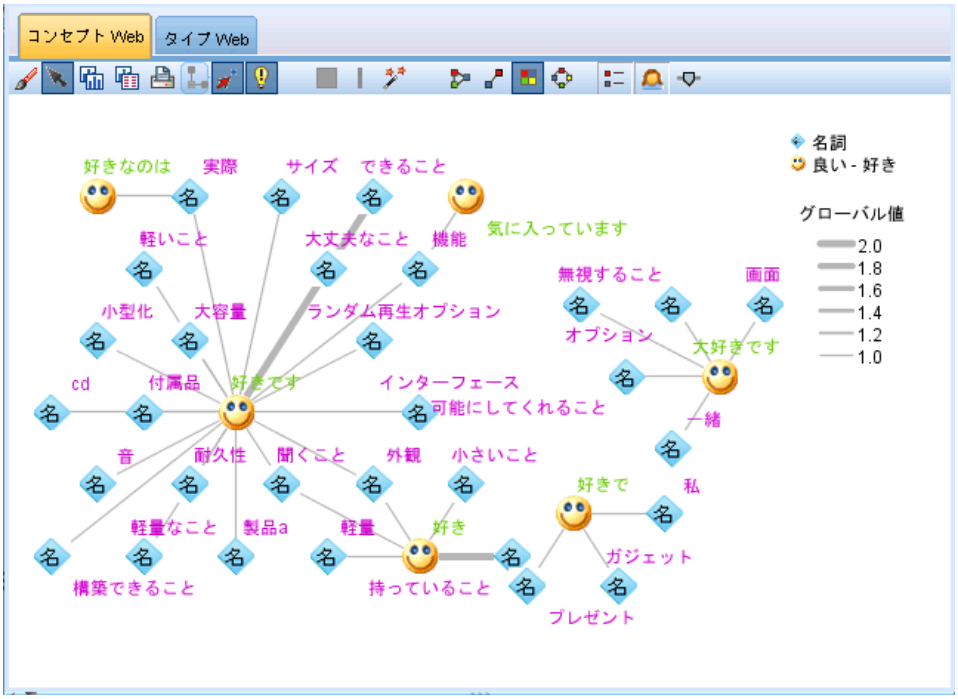
テキスト リンク分析ビューの右上のこのパネルには、選択したパターンの Web グラフがタイプ パターンまたはコンセプト パターンのいずれかとして表示されます。表示されない場合、［表示］メニュー（**[表示] → [視覚化]**）からこの領域にアクセスできます。その他のパネルでの選択内容によって、ドキュメント/レコードおよびパターンの間の該当する相互作用を表示できます。

次のような形式で結果を表示します。

- **コンセプト グラフ**: このグラフには、選択したパターンのすべてのコンセプトを示します。コンセプト グラフの線の幅およびノードのサイズ（タイプ アイコンが表示されていない場合）には、選択したテーブルのグローバル出現値を示します。
- **タイプ グラフ**: このグラフには、選択したパターンのすべてのタイプを示します。グラフの線の幅およびノードのサイズ（タイプ アイコンが表示されていない場合）には、選択したテーブルのグローバル出現値を示します。ノードは、タイプ カラーまたはアイコンによって示されます。

詳細は、 13 章 p. 285 テキスト リンク分析のグラフ を参照してください。

図 8-12
テキスト リンク分析[視覚化] パネル



[データ] パネル

[データ] パネルは、右下に表示されます。このウィンドウには、ビューの別の領域での選択内容に対応するドキュメントまたはレコードを示すテーブルが表示されます。選択内容に応じて、対応するテキストのみが [データ] パネルに表示されます。選択した後、[**表示**] ボタンをクリックすると、[データ] パネルに対応するテキストが入力されます。

別のパネルで選択した場合、該当するドキュメントまたはレコードにはコンセプトが色付きで強調表示され、テキスト内のコンセプトを特定しやすくします。カラーコード化された項目上でマウス ポインタを停止させて、項目が抽出されたコンセプトの名前と、項目が割り当てられたタイプを示すヒントを表示することもできます。 詳細は、 10 章 p. 190 [データ] パネル を参照してください。

リソース エディタ ビュー

IBM® SPSS® Modeler Text Analytics は、頑健な抽出エンジンを使用して、主要なコンセプトをテキストデータから迅速かつ正確にキャプチャします。このエンジンは、大容量の非構造化テキスト データをどのように分析および解釈するかを示す言語リソースによって大きく異なります。

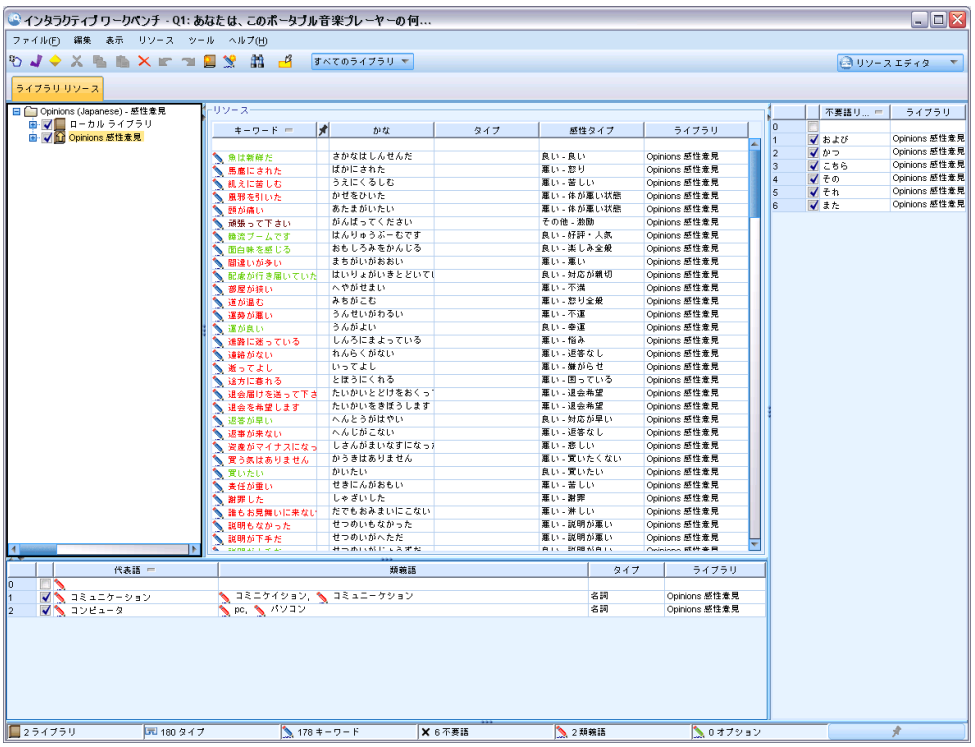
リソース エディタ ビューでは、コンセプトの抽出、タイプに基づいたグループ化、テキスト データでのパターンの検出などに使用する言語リソースを表示および調整できます。SPSS Modeler Text Analytics では、事前設定されたリソース テンプレートをいくつか用意しています。また、一部の言語では、テキスト分析パッケージのリソースを使用することもできます。詳細は、 10 章 p. 245 テキスト分析パッケージの使用 を参照してください。

これらのリソースは、常にデータのコンテキストに完全に対応しているとは限らないため、リソース エディタ で特定のコンテキストまたはドメインの独自のリソースを作成、編集および管理できます。詳細は、 16 章 p. 316 ライブラリの使用 を参照してください。

言語リソースの調整プロセスを簡略化するために、一般的な辞書タスクを、[抽出結果] パネルおよび [データ] パネルのコンテキスト メニューを使用して、カテゴリとコンセプトビューから直接実行できます。詳細は、 9 章 p. 168 抽出結果の調整 を参照してください。

注： 日本語テキスト向けのリソースのインターフェイスは、若干異なります。日本語テキスト展開はIBM® SPSS® Modeler Premiumで利用可能です。詳細は、 A 付録 p. 407 日本語テキストのリソースの編集 を参照してください。

図 8-13
リソース エディタ ビュー



リソース エディタ で実行する操作は、言語リソースの管理および調整を中心に展開しています。これらのリソースは、テンプレートおよびライブラリの形で保存されています。リソース エディタビューは４つに編成されています：ライブラリツリーペイン、タイプ辞書ペイン、代替辞書ペイン、それと除外辞書ペイン。

注: 詳細は、 15 章 p.299 エディタのインターフェイス を参照してください。

オプションの設定

[オプション] ダイアログ ボックスで IBM® SPSS® Modeler Text Analytics の一般的なオプションを設定できます。このダイアログボックスには、以下のようなタブがあります。

- **セッション**: このタブには、一般オプションおよび区切りがあります。
- **表示**: インターフェイスで使用される色についてのオプションがあります。
- **サウンド**: サウンド キューについてのオプションがあります。

オプションを編集するには:

▶ メニューの [**ツール**] > [**オプション**] を選択します。[オプション] ダイアログ ボックスが開きます。

▶ 変更する情報を含むタブを選択します。

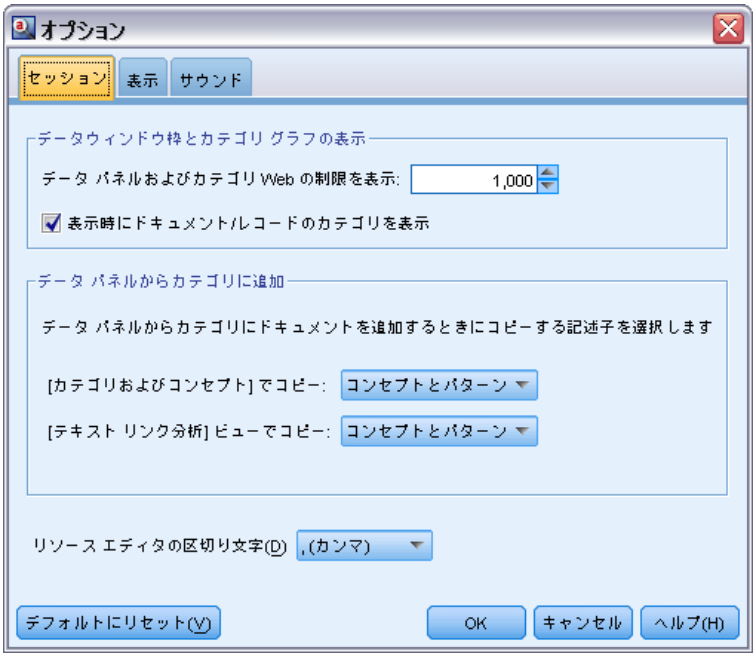
▶ オプションのいずれかを変更します。

▶ [OK] をクリックして、変更を保存します。

オプション: [セッション] タブ

このタブで、基本的な設定をいくつか定義できます。

図 8-14
[オプション] ダイアログ ボックス:[セッション] タブ



データウィンドウ枠とカテゴリ グラフの表示: カテゴリとコンセプト ビューの［データ］パネルおよび［視覚化］パネルにデータがどのように表示されるかを指定します。

- **データ パネルおよびカテゴリ Web の制限を表示:** カテゴリとコンセプト ビューの［データ］パネルまたはグラフおよび図表の入力に表示または使用するドキュメントの最大数を設定します。
- **表示時にドキュメント/レコードのカテゴリを表示:** ［表示］をクリックするとドキュメントまたはレコードがスコアリングされ、それらが属するカテゴリが［データ］パネルの［カテゴリ］列およびカテゴリ グラフに表示されます。特にデータセットが大きい場合は、データおよびグラフをより早く表示するよう、このオプションを無効にする必要があります。

データパネルからカテゴリに追加: ドキュメントおよびレコードが［データ］パネルから追加される場合、カテゴリに追加する内容を設定します。

- **［カテゴリおよびコンセプト］でコピー** このビューの［データ］パネルからドキュメントまたはレコードを追加すると、［**コンセプトのみ**］または［**コンセプトおよびパターン**］でコピーされます。
- **テキスト リンク分析ビューでコピー:** このビューの［データ］パネルからドキュメントまたはレコードを追加すると、［**パターンのみ**］または［**コンセプトおよびパターン**］でコピーされます。

リソース エディタの区切り文字: コンセプト、類義語、オプション要素などの要素をリソース エディタ ビューで入力する場合に区切り文字として使用する文字を選択します。

オプション:[表示] タブ

このタブでは、全体的な外観に関するオプションや、要素を区別するための色を編集できます。

注:製品の表示を以前のリリースのクラシック表示に切り替えるには、IBM® SPSS® Modeler メイン ウィンドウにある［ツール］メニューの［ユーザー オプション］ダイアログを開きます。

図 8-15
[オプション] ダイアログ ボックス:[表示] タブ

ユーザー定義の色: 画面上に表示される要素の色を編集します。表内のそれぞれの要素に関して、色を変更できます。ユーザー定義の色を指定するには、変更したい要素の右側にある色をクリックし、色のドロップダウンリストから色を変更します。

- **未抽出のテキスト:** ［データ］パネルで抽出されていないが表示されるテキスト データ。
- **強調背景:** パネルの要素または［データ］パネルのテキストを選択する場合のテキスト選択の背景色。

- **抽出が必要な背景:** [抽出結果]、[パターン]、[クラスタ] パネルの背景色はライブラリに変更が行われたことを示し、抽出が必要です。
- **カテゴリ フィード バック背景:** 操作後に出現するカテゴリ背景色。
- **デフォルト タイプ:** [データ] パネルおよび [抽出結果] パネルに出現するタイプおよびコンセプトのデフォルト色。この色は、リソース エディタで作成するカスタム タイプに適用されます。リソース エディタ でこれらのキーワード辞書のプロパティを編集し、カスタム キーワード辞書のこのデフォルト色を上書きします。 詳細は、 17 章 p.334 タイプの作成 を参照してください。
- **テーブルの縞 1:** 各セットの行を区別するための、[強制コンセプトを編集] ダイアログボックスのテーブルで、交互に使用される 2 つの色のうちの最初の色。
- **テーブルの縞 2:** 各セットの行を区別するための、[強制コンセプトを編集] ダイアログボックスのテーブルで、交互に使用される 2 つの色のうちの 2 番目の色。

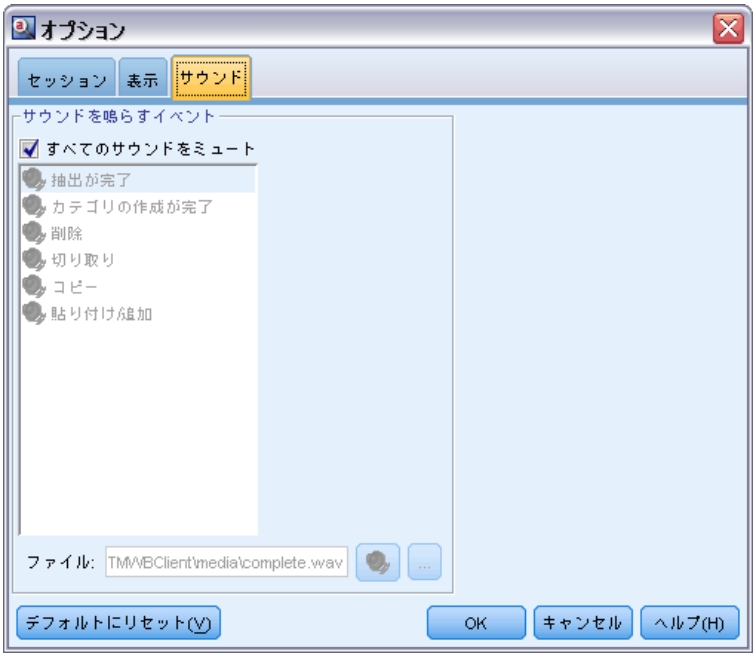
注:[**デフォルトにリセット**] ボタンをクリックすると、このダイアログ ボックスのすべてのオプションは、この製品を最初にインストールしたときに設定されていた値に戻されます。

オプション:[サウンド] タブ

このタブでは、サウンドに関するオプションを編集できます。サウンドを鳴らすイベントの部分で、あるイベントが起こった際にこれを知らせるサウンドを指定できます。使用できるサウンドはたくさんあります。サウンドを参照して選択する場合は、[…] ボタンを使用します。IBM® SPSS® Modeler Text Analytics のサウンドを作成するために使われる .wav ファイルは、インストール ディレクトリ中の media サブディレクトリにあります。サウンドを鳴らしたくない場合、[**すべてのサウンドをミュート**] を選択します。デフォルトでは、音が鳴らないようになっています。

注:[**デフォルトにリセット**] ボタンをクリックすると、このダイアログ ボックスのすべてのオプションは、この製品を最初にインストールしたときに設定されていた値に戻されます。

図 8-16
[オプション] ダイアログ ボックス:[サウンド] タブ



Microsoft Internet Explorer ヘルプの設定

Microsoft Internet Explorer の設定

このアプリケーションのほとんどのヘルプ機能では、Microsoft Internet Explorer に基づいたテクノロジが使用されています。Internet Explorer のバージョンによっては (Microsoft Windows XP、Service Pack 2 と共に提供されるバージョンも含む)、ローカル コンピュータ上の [Internet Explorer] ウィンドウ内で「アクティブなコンテンツ」と見なされる対象が、デフォルトにより封鎖されます。このデフォルトの設定により、ヘルプ機能内である種のコンテンツが表示されなくなります。すべてのヘルプ コンテンツを表示するために、Internet Explorer のデフォルトの動作を変更できます。

- Internet Explorer のメニューから次の項目を選択します。
  ツール > インターネット オプション
- **[詳細]** タブをクリックします。
- **[セキュリティ]** セクションまで下方へスクロールします。
- **[マイ コンピュータのファイルでのアクティブ コンテンツの実行を許可する]** を選択（チェック）します。

モデル ナゲットおよびモデル作成ノードの生成

インタラクティブ セッションの場合、実行した作業を使用して、次のいずれかを作成する必要があります。

- **テキスト マイニング モデル作成ノード:** インタラクティブ ワークベンチ セッションから生成したモデル作成ノードは、設定とオプションが、オープン インタラクティブ セッションに保存されている設定およびオプションを反映するテキスト マイニング ノードです。元のテキスト マイニング ノードがない場合、または新しいバージョンを作成したい場合、役立ちます。 詳細は、 3 章 p.33 コンセプトおよびカテゴリのマイニング を参照してください。
- **カテゴリ モデル ナゲット:** インタラクティブ ワークベンチ セッションから生成されたモデル ナゲットは、カテゴリ モデル ナゲットです。カテゴリ モデル ナゲットを生成するには、カテゴリとコンセプト ビューに 1 つ以上のカテゴリが必要です。 詳細は、 3 章 p.77 テキスト マイニング モデル ナゲット:カテゴリ モデル を参照してください。

テキスト マイニング モデル作成ノードを生成するには

▶ メニューの [生成] > [モデル作成ノードの生成] を選択します。テキスト マイニング モデル作成ノードは、ワークベンチ セッションで現在すべての設定を使用する作業キャンバスに追加されます。ノードには、テキスト フィールドの名前が付きます。

カテゴリ モデル ナゲットを生成するには

▶ メニューの [生成] > [モデルの生成] を選択します。モデル ナゲットは、モデル パレットに直接生成され、デフォルト名が付きます。

モデル作成ノードの更新および保存

インタラクティブ セッションで作業する場合、時々モデル作成ノードを変更して、変更を保存することをお勧めします。また、インタラクティブ ワークベンチ セッションで作業を終了し、作業を保存する場合も、モデル作成ノードを更新する必要があります。モデル作成ノードを更新する場合、ワークベンチ セッションの内容が、インタラクティブ ワークベンチ セッションに由来するテキスト マイニング ノードに保存します。更新しても、出力ウィンドウは閉じません。

重要! 更新するとストリームは保存されません。ストリームを保存するには、モデル作成ノードを更新した後、IBM® SPSS® Modeler のメイン ウィンドウで保存します。

モデル作成ノードを更新するには

▶ メニューの [ファイル] → [モデル作成ノードを更新] を選択します。オプションおよびカテゴリとともに、作成設定および抽出設定でモデル作成ノードを更新します。

セッションの終了

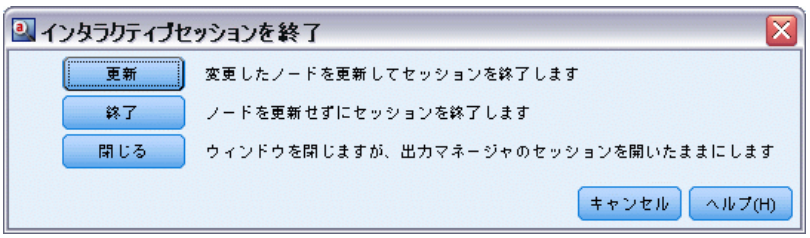
セッションの作業を終了する場合、次の 3 つの終了でセッションを離れることができます。

- **保存:** まず、別のセッションで再利用するためにライブラリを公開し、次のセッションのために元のモデル作成ノードに作業を保存します。詳細は、 16 章 p.326 ライブラリの共有 を参照してください。 保存した後、セッションが終了し、IBM® SPSS® Modeler ウィンドウの出力マネージャからセッションが削除されます。
- **終了:** 保存していない作業を破棄し、セッション ウィンドウを閉じて、SPSS Modeler ウィンドウの出力マネージャからセッションが削除されます。メモリを確保するために、重要な作業を保存して、セッションを終了することをお勧めします。
- **閉じる:** 作業は保存されず、また破棄もされません。セッション ウィンドウが閉じますが、セッションは稼動し続けます。SPSS Modeler ウィンドウの出力マネージャでこのセッションを選択すると、セッション ウィンドウを開くことができます。

ワークベンチのセッションを終了するには

▶ メニューから、[ファイル] → [閉じる] を選択します。

図 8-17
[インタラクティブ セッションを閉じる] ダイアログ ボックス



キーボードのアクセス機能

インタラクティブ ワークベンチのインターフェイルには、製品の機能によりアクセスしやすくするキーボード ショートカットが用意されています。基本的には、Alt キーと他の適切なキーを同時に押してメニュー項目を選択したり（例: Alt + F キーで［ファイル］メニューを選択）、

Tab キーを使ってダイアログ ボックス中のコントロール間を移動することができます。ここでは、もう 1 つのナビゲーションであるキーボードショートカットについて説明します。IBM® SPSS® Modeler インターフェイスには、その他のショートカットがあります。

テーブル 8-1
一般的なキーボード ショートカット

ショートカット キー	関数
Ctrl + 1	タブのあるパネルの最初のタブを表示します。
Ctrl + 2	タブのあるパネルの 2 番目のタブを表示します。
Ctrl + A	フォーカスのあるパネルのすべての要素を選択します。
Ctrl + C	選択したテキストをクリップボードにコピーします。
Ctrl + E	カテゴリとコンセプトビューおよびテキストリンク分析ビューの抽出を起動します。
Ctrl + F	リソース エディタ/テンプレート エディタ の［検索］ツールバーが表示されていない場合は表示し、フォーカスします。
Ctrl + I	カテゴリとコンセプト ビューの場合、選択したカテゴリの［カテゴリ定義］ダイアログ ボックスを起動します。クラスタ ビューの場合、選択したクラスタの［クラスタ定義］ダイアログ ボックスを起動します。
Ctrl + R	リソース エディタ/テンプレート エディタ の［キーワードを追加］を開きます。
Ctrl + T	リソース エディタ/テンプレート エディタ で［タイプのプロパティ］ダイアログ ボックスを開き、新しいタイプを作成します。
Ctrl + V	クリップボードの内容を貼り付けます。
Ctrl + X	リソース エディタ/テンプレート エディタ から選択した項目を切り取ります。
Ctrl+Y	ビューの最後のアクションをやり直します。
Ctrl + Z	ビューの最後のアクションを取り消します。
F1	ヘルプを表示するか、ダイアログ ボックスでは、その項目のコンテキスト ヘルプを表示します。
F2	テーブルのセルの編集モードを有効にしたり、無効にしたりします。
F6	アクティブなビューの主なパネル間でフォーカスを移動します。
F8	フォーカスをパネルの分割バーに移動し、サイズを変更します。
F10	メインの［ファイル］メニューを展開します。
上方向矢印、下方向矢印	分割バーが選択されているときに、ウィンドウを垂直方向にサイズ変更します。
左方向矢印、右方向矢印	分割バーが選択されているときに、ウィンドウを水平方向にサイズ変更します。
Home、End	分割バーが選択されているときに、ウィンドウを最大サイズまたは最小サイズに変更します。
Tab	ウィンドウ、パネル、ダイアログ ボックスの次の項目に移動します。

ショートカット キー	関数
Shift + F10	項目のコンテキスト メニューを表示します。
Shift + Tab	ウィンドウ、またはダイアログ ボックスの前の項目に移動します。
Shift + 矢印	編集モード (F2) のとき、編集フィールドの文字を選択します。
Ctrl + Tab	ウィンドウの次のメイン領域にフォーカスを移動します。
Shift + Ctrl + Tab	ウィンドウの前のメイン領域にフォーカスを移動します。

ダイアログ ボックスのショートカット

ダイアログ ボックスを使用している場合、いくつかのショートカットおよびスクリーン リーダー キーが役立ちます。ダイアログ ボックスに入力すると、Tab キーを押して、最初のコントロールにフォーカスし、スクリーン リーダーを起動する必要があります。特殊なキーボード ショートカットおよびスクリーン リーダーのショートカットの詳細について、次の表で説明しています。

テーブル 8-2
ダイアログ ボックスのショートカット

ショートカット キー	関数
Tab	ウィンドウ、またはダイアログ ボックスの次の項目に移動します。
Ctrl + Tab	テキスト ボックスから次の項目に移動します。
Shift + Tab	ウィンドウ、またはダイアログ ボックスの前の項目に移動します。
Shift + Ctrl + Tab	テキスト ボックスから前の項目に移動します。
スペース キー	フォーカスのあるコントロールまたはボタンを選択します。
Esc	変更をキャンセルして、ダイアログ ボックスを閉じます。
Enter	変更を確認して、ダイアログ ボックスを閉じます ([OK] ボタンと同じ)。テキスト ボックスで作業している場合、まず Ctrl + Tab を押して、テキスト ボックスから移動する必要があります。

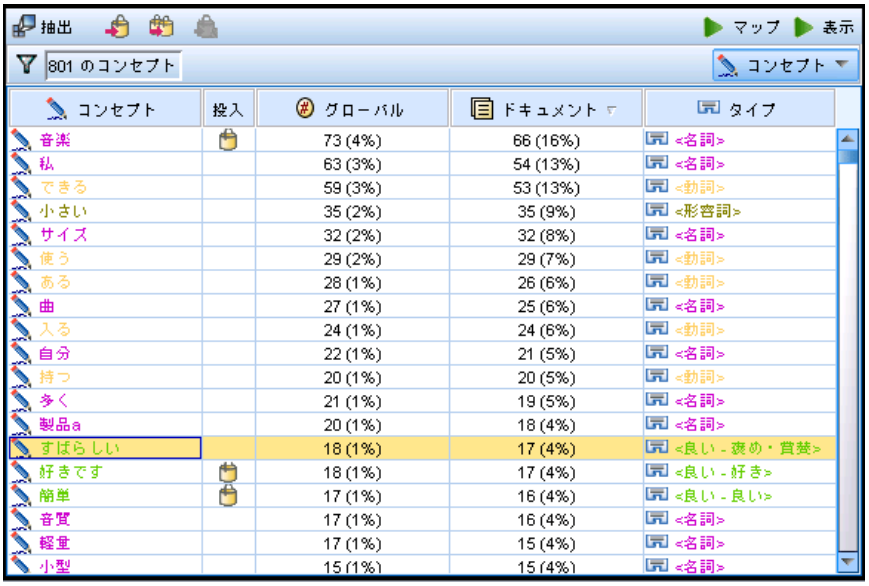
章

9

コンセプトとタイプの抽出

インタラクティブ ワークベンチを起動するストリームを実行する場合、抽出はストリームのテキスト データに実行されます。この抽出の最終結果は、一連のコンセプト、タイプ、そして TLA パターンが言語リソースにある場合はパターンとなります。[抽出結果] パネルでコンセプトおよびタイプを表示および処理できます。 詳細は、 1 章 p. 7 抽出の方法 を参照してください。

図 9-1
抽出後の [抽出結果] パネル



抽出結果を調整する場合、言語リソースを変更し、再抽出できます。詳細は、 p. 168 抽出結果の調整 を参照してください。 抽出プロセスは、結果の抽出および構成方法を指定する [抽出] ダイアログ ボックスのリソースおよびパラメータによって異なります。抽出結果を使用して、カテゴリ定義の（全部ではない場合）大部分を定義できます。

抽出結果:コンセプトとタイプ

抽出プロセスで、すべてのテキスト データがスキャンされ、関連するコンセプトが特定、抽出、そしてタイプに割り当てられます。抽出が完了すると、カテゴリとコンセプト ビューの左下隅にある [抽出結果] パネルに結果が表示されます。セッションを初めて起動した場合、ノードで選択

した言語リソース テンプレートを使用して、これらのコンセプトおよびタイプを抽出および構成します。

抽出されたコンセプト、タイプ、および TLA パターンは、まとめて**抽出結果**と呼ばれ、カテゴリの記述子、または構築ブロックとして機能します。また、カテゴリ規則でコンセプト、タイプ、およびパターンを使用することもできます。さらに、自動的手法では、コンセプトおよびタイプを使用してカテゴリを作成します。

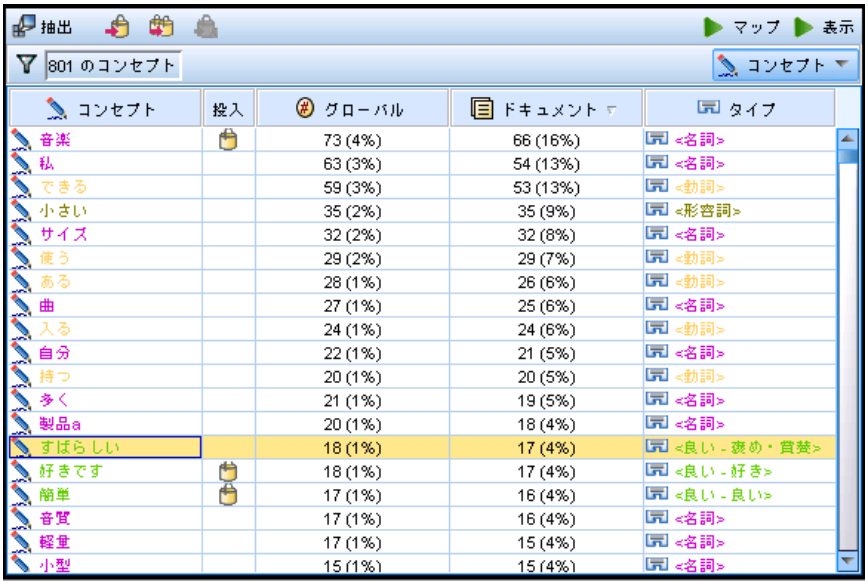
テキスト マイニングは、抽出結果をテキスト データのコンテキストに従ってレビューし、新しい結果を作成するよう調整、そして再評価するインタラクティブ プロセスです。抽出後、結果を表示し、必要に応じて言語リソースを修正することによって、結果を変更する必要があります。[抽出結果] パネル、[データ] パネル、[カテゴリ定義] ダイアログ ボックス、または [クラスタ定義] ダイアログ ボックスから直接、リソースをある程度調整できます。詳細は、 p. 168 抽出結果の調整 を参照してください。リソース エディタ ビューで直接調整することもできます。詳細は、 8 章 p. 141 リソース エディタ ビュー を参照してください。

調整した後、再抽出して新しい結果を表示できます。最初から抽出結果を調整することによって、再抽出するごとに、カテゴリ定義で同じ結果を取得することを確認し、データのコンテキストに完全に対応することができます。このようにして、ドキュメント/レコードをより正確で、繰り返し可能な方法で、カテゴリ定義に割り当てます。

コンセプト

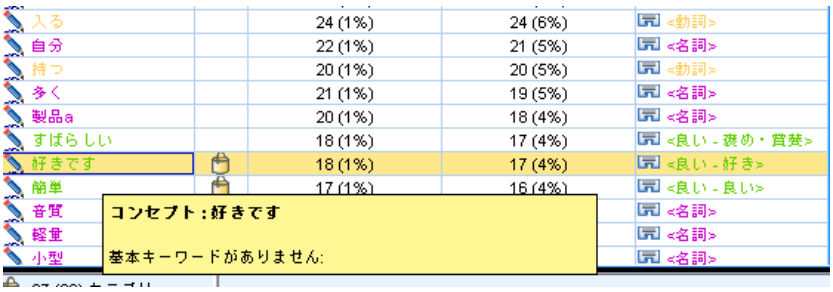
抽出プロセスで、テキスト データをスキャンして分析し、テキスト内の関心のあるまたは関連する 1 つの単語（`election` または `peace` など）や句（`presidential election`, `election of the president`、または `peace treaties` など）を特定します。これらの単語や句を、まとめて「キーワード」と呼びます。言語リソースを使用して、関連キーワードを抽出し、類似したキーワードを**コンセプト**と呼ばれる代表語でグループ化します。

図 9-2
抽出後の [抽出結果] パネル



コンセプト名の上にマウスポインタを置くと、コンセプトの基本キーワードのセットが表示されます。これにより、コンセプト名とそのコンセプトにグループ化された数行のキーワードを示すツールヒントが表示されます。これらの基本キーワードには、抽出された複数形/単数形のキーワード、置換キーワード、Fuzzy Grouping のキーワードなどのほか、テキスト内にあったかどうかに関係なく、言語リソースで定義された類義語が含まれます。コンセプト名を右クリックし、コンテキスト メニューのオプションを選択して、これらのキーワードをコピーしたり、基本キーワードの完全セットを表示したりできます。

図 9-3
抽出後の [抽出結果] パネル



デフォルトでは、ドキュメント数（[ドキュメント] 列）に従って、降順で並べられます。コンセプトが抽出されると、それらをタイプに割り当てて、同様のコンセプトをグループ化します。それらはこのタイプに従ってカラー

コード化されます。色は、リソース エディタ のタイプのプロパティで定義されます。詳細は、 17 章 p. 332 キーワード辞書 を参照してください。

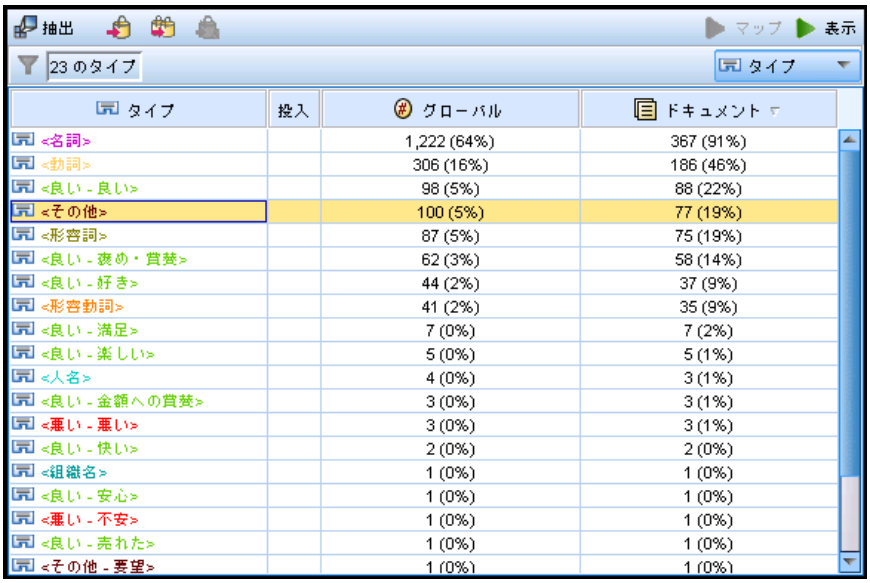
カテゴリ定義でコンセプト、タイプ、またはパターンが使用されている場合、アイコンが並べ替え可能な [投入] 列に表示されます。

タイプ

タイプは、コンセプトの意味上のグループ化です。コンセプトが抽出されると、それらをタイプに割り当てて、同様のコンセプトをグループ化します。<Location>、<Organization>、<Person>、<Positive>、<Negative>など、いくつかのビルトインのタイプが IBM® SPSS® Modeler Text Analytics に付属しています。たとえば、<Location> のタイプは、地理的なキーワードや地名をグループ化します。このタイプは、chicago、paris、および tokyo などのコンセプトに割り当てられます。 ほとんどの言語の場合、キーワード辞書になく、テキストから抽出されたコンセプトは、自動的に<Unknown>のタイプとなりますが、日本語テキストの場合は、自動的に<**名詞**>ノートのタイプとなります：日本語テキスト展開は IBM® SPSS® Modeler Premiumで利用可能です。詳細は、 17 章 p. 334 ビルトインのタイプ を参照してください。

タイプ ビューを選択すると、デフォルトでは抽出したタイプがグローバルな頻度の高い順に表示されます。タイプは区別できるよう、カラーコード化されます。色は、タイプのプロパティの一部です。詳細は、17 章 p. 334 タイプの作成 を参照してください。 独自のタイプを作成することもできます。

図 9-4
タイプ ビュー:[抽出結果] パネル



パターン

テキスト データからパターンを抽出することもできます。ただし、リソース エディタ にテキスト リンク分析（TLA）パターン規則を含むライブラリが必要です。また、[テキスト リンク分析のパターン抽出を有効にする] オプションを使用して、SPSS Modeler Text Analytics ノードの設定または［抽出］ダイアログ ボックスでこれらのパターンの抽出を選択する必要もあります。詳細は、 12 章 p.268 テキスト リンク分析の検証 を参照してください。

データの抽出

抽出が必要な場合、［抽出結果］パネルが黄色で表示され、メッセージ[抽出ボタンをクリックしてキーワードを抽出してください] というメッセージが、このウィドウ枠のツールバーの下に表示されます。

抽出結果がない場合、言語リソースに変更を行い抽出結果を更新する必要がない場合、または抽出結果を保存していないセッションを開く場合は、抽出が必要な場合があります（[ツール] > [オプション]）。

注:抽出結果を［セッション作業を使用］オプションを使用してキャッシュした後、ストリームの入力ノードを変更する際、抽出結果を更新する場合にインタラクティブ ワークベンチ セッションが起動したら新しい抽出を実行する必要があります。

抽出実行中には進行状況が表示されます。抽出している間、抽出エンジンはテキスト データをすべて読み込み、関連キーワードおよびパターンを特定し、それらを抽出して、タイプに割り当てます。そして、エンジンは、1 つの主要なキーワード、コンセプトに類義語のキーワードをグループ化します。プロセスが完了すると、生成されたコンセプト、タイプ、パターンが［抽出結果］パネルに表示されます。

抽出プロセスにより、一連のコンセプト、タイプ、そして有効な場合はテキスト リンク分析（TLA）パターンが作成されます。カテゴリとコンセプト ビューの［抽出結果］パネルでこれらのコンセプトおよびタイプを表示および処理できます。TLA パターンを抽出した場合、これらはテキスト リンク分析ビューにされます。

注:データセットのサイズと、抽出プロセスを完了するためにかかる時間の間には、関連性があります。上流にサンプル ノードを追加、またはコンピュータの構成を最適化することをいつでも検討することができます。

データを抽出するには

▶ メニューの［ツール］>［抽出］を選択します。または、［抽出］ツールバーボタンをクリックします。

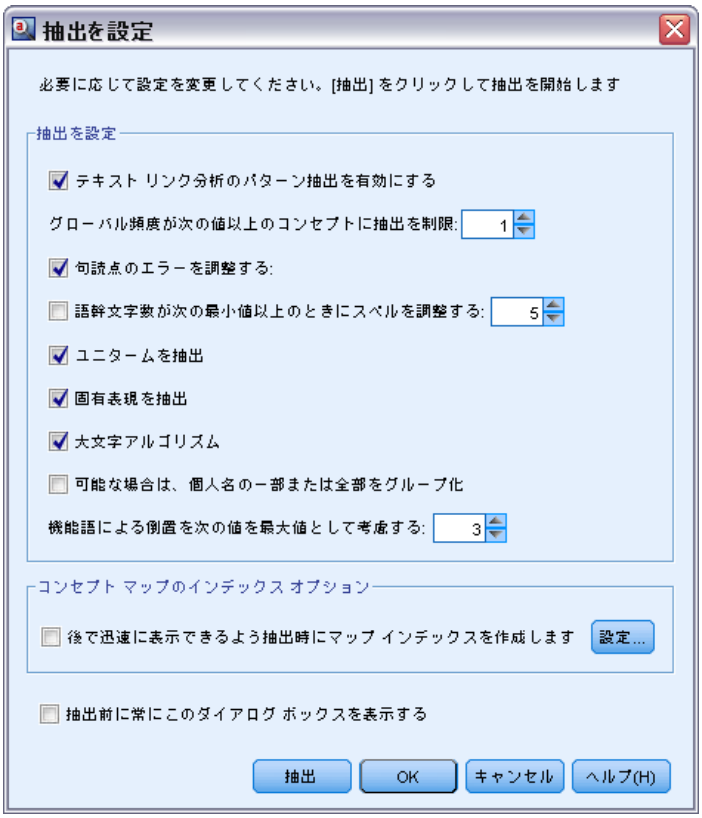
- ［抽出設定］ダイアログの表示を選択すると必ず、ダイアログが表示され、変更を行うことができます。各設定の記述子については、このトピックの後半を参照してください。
- **[抽出]** をクリックして、抽出プロセスを開始します。抽出が始まると、進捗状況のダイアログ ボックスが表示されます。抽出後、結果が［抽出結果］ウィンドウに表示されます。デフォルトでは、ドキュメント数（[ドキュメント］列）に従って、降順で並べられます。

  ツールバー オプションを使用して結果を確認し、結果を並べ替える、結果を絞り込む、または異なるビュー（コンセプト、またはタイプ）に切り替えることができます。言語リソースを処理して、抽出結果を調整することもできます。 詳細は、 p. 168 抽出結果の調整 を参照してください。

オランダ語、英語、フランス語、ドイツ語、イタリア語、ポルトガル語、スペイン語のテキストの場合

［抽出設定］ダイアログ ボックスには、基本的な抽出オプションがいくつか表示されます。

図 9-5
［抽出設定］ダイアログ ボックス



テキスト リンク分析のパターン抽出を有効にする: テキスト データから TLA パターンを抽出するよう指定します。また、リソース エディタのいずれかのライブラリに TLA パターン規則があることも想定します。このオプションを指定すると、抽出時間が大幅に長くなります。 詳細は、 12 章 p. 268 テキスト リンク分析の検証 を参照してください。

句読点のエラーを調整する: 抽出時に句読点エラー（不適切な使用方法など）を含むテキストを一時的に正規化し、コンセプトの抽出可能性を向上させます。自由記述式アンケートの回答、電子メール、CRM データなど、テキストが短く品質が悪い場合、またはテキストに略語が多く含まれている場合に特に役立ちます。

語幹文字数が次の最小値以上のときにスペルを調整する: Fuzzy Grouping の手法を適用し、共通してスペルミスのある単語またはスペルの近い単語を 1 つのコンセプトにグループ化できるようにします。Fuzzy Grouping アルゴリズムでは、最初の母音を除くすべての母音を一時的に抜き取った後抽出した単語から 2 つ/3 つの子音を抜き取り、それらを比較して、それらが同じで `modeling` と `modelling` が同じグループに分けられるかどうかを確認します。ただし、各キーワードが `<Unknown>` タイプを除いて、別のタイプに割り当てられた場合、Fuzzy Grouping 手法は適用されません。

Fuzzy Grouping を使用する前に必要な、語幹文字数の制限を定義することもできます。キーワード内の語幹文字数は、すべての文字を合計し、活用語尾、複合語キーワードの場合は区切り文字および前置詞を形成する文字を差し引いて計算します。たとえば、キーワード `exercises` の語幹文字数は「exercise」という形式で 8 文字と数えられます。語末の s は活用語尾（複数形）であるためです。同様に、`apple sauce` の語幹文字は 10 文字（「apple sauce」）、そして `manufacturing of cars` の語幹文字は 16 文字（「manufacturing car」）となります。この算出方法は、Fuzzy Grouping を適用するべきかどうかを確認するためにのみ使用されますが、単語がどのように一致するかについては影響を与えません。

注:特定の単語が後で不適切にグループ化されていることが分かった場合、[アドバンス リソース] タブの **Fuzzy Grouping: 例外** セクションで 明示的に宣言することによって、単語のペアをこの手法から除外できます。詳細は、 18 章 p. 357 Fuzzy Grouping を参照してください。

ユニタームを抽出: 単語が複合語の一部でない限り、または名詞、またはスピーチ内の認識できない部分である場合、このオプションは単一の単語（ユニターム）を抽出します。

固有表現を抽出: 電話番号、セキュリティ番号、時間、日付、通貨、数字、パーセント、電子メールアドレス、HTTP アドレスなどの固有表現を抽出します。[アドバンス リソース] タブの **[固有表現: 設定]** セクションで、特定の種類の固有表現を追加したり除外したりできます。不要なエンティティを無効にすることにより、抽出エンジンは処理時間を節約できます。詳細は、 18 章 p. 362 構成 を参照してください。

大文字アルゴリズム: キーワードの最初の文字が大文字である場合、組み込み辞書にない単純キーワードおよび複合キーワードを抽出します。このオプションには、最も適切な名詞を抽出するのに優れた方法があります。

可能な場合は、個人名の一部または全部をグループ化: テキスト内で別々の形式で同時に出現する名前をグループ化します。名前はテキストの始めでは完全な形式で、後は短い形式でのみ参照されるため、この機能が役立ちます。このオプションでは、タイプが <Unknown> のユニタームが、タイプ <Person> の複合キーワードの最後の単語に一致するようにします。たとえば、doe があり、最初タイプが <Unknown> である場合、抽出エンジンは、<Person> タイプの複合キーワードに最後の単語として doe が含まれているかどうか（例: john doe）を確認します。ほとんどがユニタームとして抽出されることがないため、人の名前に適用されることはありません。

機能語による倒置を次の値を最大値として考慮する: 倒置手法を適用する場合に指定されている場合がある非機能的単語の最大数を指定します。この倒置手法では、活用語尾に関係なく、含まれる非機能的単語（of や the など）によってお互いに異なる類似した句をグループ化します。たとえば、この値を最大 2 単語に設定し、company officials および officials of the company が抽出されたとします。この場合、両方の抽出キーワードは、of the が無視されると同じであるとみなされるため、最終コンセプト リストに共にグループ化されます。

コンセプト マップのインデックス オプション コンセプト マップを後ですぐに描画できるよう、抽出時間にマップの指標を作成することを指定します。インデックスの設定を編集するには、[**設定**] をクリックします。 詳細は、p. 167 コンセプト マップ インデックスの作成 を参照してください。

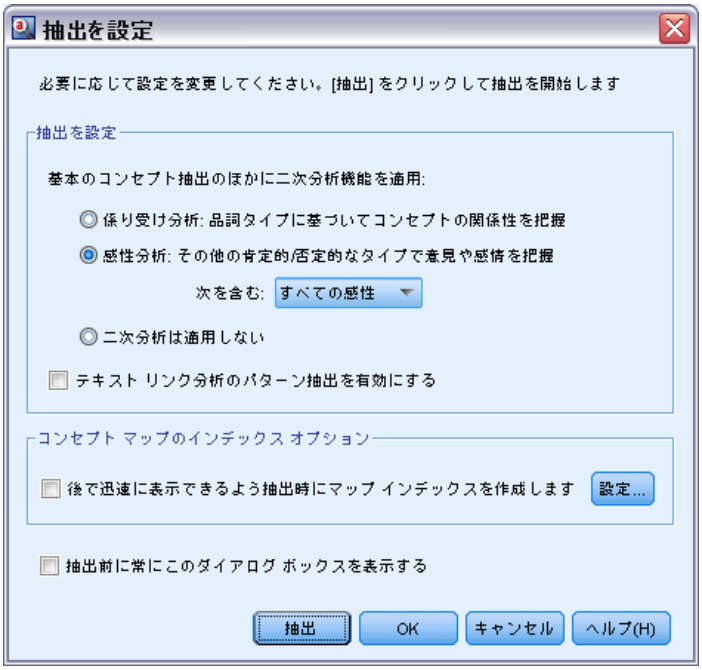
抽出前に常にこのダイアログ ボックスを表示する: ［ツール］ メニューを選択しない限り表示したくない場合、抽出ごとに［抽出設定］ダイアログを表示するかどうか、または抽出設定を編集する場合、抽出ごとに表示するかどうかを尋ねるかどうかを指定します。

日本語テキストの場合

注:日本語テキスト展開はIBM® SPSS® Modeler Premiumで利用可能です。

［抽出設定］ダイアログ ボックスには、日本語テキスト向けの基本的な抽出オプションがいくつか表示されます。デフォルトでは、ダイアログで選択された設定は、テキスト マイニング モデル作成ノードの［エキスパート］タブで選択された設定と同じです。日本語テキストを処理するためには、入力としてテキストを使用するほか、テキスト マイニング ノードの［モデル］タブで日本語テンプレートまたはテキスト分析パッケージを選択する必要があります。詳細は、 3 章 p. 46 テンプレートおよび TAP からのリソースのコピー を参照してください。

図 9-6
日本語テキストの [抽出設定] ダイアログ ボックス



注:日本語テキスト展開はSPSS Modeler Premiumで利用可能です。

二次分析: 抽出が開始したとき、基本キーワード抽出が、タイプのデフォルト セットを使用して行われます。詳細は、 A 付録 p. 412 日本語テキストで使用できるタイプ を参照してください。 ただし、二次分析機能を選択すると、抽出機能にコンセプトの一部として助詞や助動詞が含まれているため、より多くの詳細なコンセプトを取得することができます。たとえば、「肩の荷が下りた」という文があるとします。この例では基本キーワード抽出は各コンセプトを個別に抽出します。例：**肩** （肩）、**荷** （重量）、**下りる** （上げ）、しかしこれらの単語の間の関係性は抽出されません。しかし、感性分析を適用すると、コンセプト =**肩の荷が下りた** というより高次の意味をもつコンセプトが抽出され、<**良い-安心**> という感性タイプとして割り当てられます。なお感性分析については、このほかに多くの感性タイプがあります。さらに、二次分析機能を選択すると、テキスト リンク分析結果も生成できます。

注:二次分析を呼び出すと、抽出プロセスにより時間がかかります。詳細は、 A 付録 p. 404 二次抽出の手順 を参照してください。

- **係り受け解析**:このオプションを選択すると、キーワード、およびキーワードに助詞等を加えた語を、基本の品詞タイプのコンセプトとして抽出します。また係り受けテキスト リンク分析（TLA）に基づいたパターン結果を出力します。

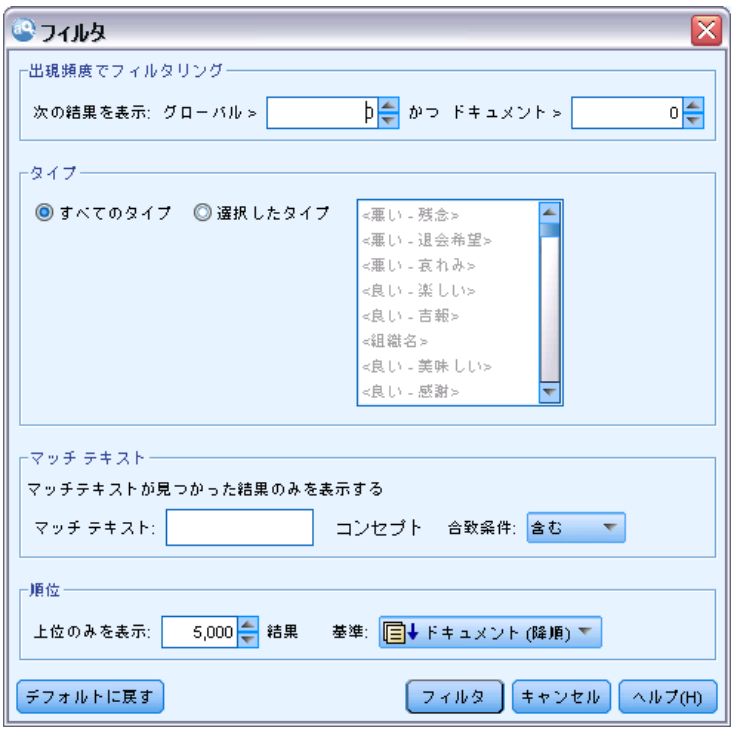
- **感性分析**：この分析機能を選択すると、追加の抽出コンセプト、および可能な場合は、TLA パターン結果の抽出が行われます。基本タイプのほか、嬉しい、吉報、幸運、安心、幸福など 80 を超える感性タイプも利用できます。これらのタイプを使用して、感情、感性、意見の表現のテキストでコンセプトおよびパターンを検出します。感性分析に対するフォーカスを指示するオプションは３つあります：**全ての感性**、**代表的な感性のみ**、それと**結論のみ**。
- **二次分析は適用しない**：すべての二次分析をオフにします。TLA の結果の取得には二次分析が必要であるため、オプション［**テキスト リンク分析のパターン抽出を有効にする**］が選択されている場合、このオプションは選択できません。

テキスト リンク分析のパターン抽出を有効にする: テキスト データから TLA パターンを抽出するよう指定します。また、リソース エディタのいずれかのライブラリに TLA パターン規則があることも想定します。このオプションを指定すると、抽出時間が大幅に長くなります。また、TLA のパターン結果を抽出するために、二次分析を選択する必要があります。 詳細は、12 章 p. 268 テキスト リンク分析の検証 を参照してください。

抽出結果のフィルタリング

非常に大きなデータセットを処理する場合、抽出プロセスでは、多数の結果が作成される場合があります。多くのユーザーによって、多数の結果が作成されると、結果を効率的に確認することが困難になります。そのため、最も関心のある結果に絞り込むために、［抽出結果］パネルで使用できる［フィルタ］ダイアログを使用してこれらの結果をフィルタリングできます。

図 9-7
[フィルタ]([抽出結果] パネル)



[フィルタ] ダイアログのすべての設定を同時に使用して、カテゴリに使用できる抽出結果をフィルタリングします。

出現頻度でフィルタリング:フィルタリングを実行して、特定のグローバル出現頻度値またはドキュメントの出現頻度の値を持つ結果のみを表示できます。

- **グローバル出現頻度**は、コンセプトがドキュメントまたはレコードの全体的なセットに出現する回数の合計で、[**グローバル**] 列に表示されます。
- **ドキュメント出現頻度**は、コンセプトが出現するドキュメントまたはレコードの合計数で、[**ドキュメント**] 列に表示されます。

たとえば、コンセプト nato が 500 件のレコードに 800 回出現した場合、このコンセプトのグローバル出現頻度は 800 で、ドキュメント出現頻度は 500 となります。

タイプ: 特定のタイプに属する結果のみを表示できます。すべてのタイプまたは特定のタイプのみを選択できます。

マッチ テキスト: ここで定義する規則に一致する結果のみを表示できます。[**マッチ テキスト**] フィールドに一致する文字のセットを入力し、マッチに適用する条件を選択します。

テーブル 9-1
マッチ テキストの条件

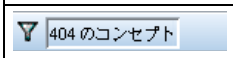
条件	説明
次を含む	文字列が任意の場所で出現する場合、テキストが一致します(デフォルトの選択)。
次で開始	コンセプトまたはタイプが特定のテキストで始まる場合にのみ、テキストが一致します。
次で終了	コンセプトまたはタイプが特定のテキストで終わる場合にのみ、テキストが一致します。
完全一致	文字列全体が、コンセプト名またはタイプ名に一致する必要があります。

順位: フィルタリングして、グローバル出現頻度（[グローバル]）またはドキュメント頻度（[ドキュメント]）に従って、上位のコンセプトのみを昇順または降順で表示することができます、

[抽出結果] パネルに表示される結果

フィルタリングに基づいて、結果が［抽出結果］ウィンドウにどのように表示されるかについて、いくつか例を示します。

テーブル 9-2
フィルタ フィードバックの例

フィルタ フィードバック	説明
	ツールバーには結果の数が表示されます。テキスト マッチ フィルタがなく、最大数に達していないため、追加のアイコンは表示されません。
	ツールバーには、結果がフィルタで指定された最大値に制限されていることを示します（この例では 300)。紫のアイコンが表示されている場合、コンセプトの最大数に達したことを示します。アイコンの上にポインタを置くと、詳細が表示されます。
	ツールバーには、マッチ テキスト フィルタを使用して、結果が制限されていることを示します。虫めがねのアイコンが表示されます。

結果を絞り込むには

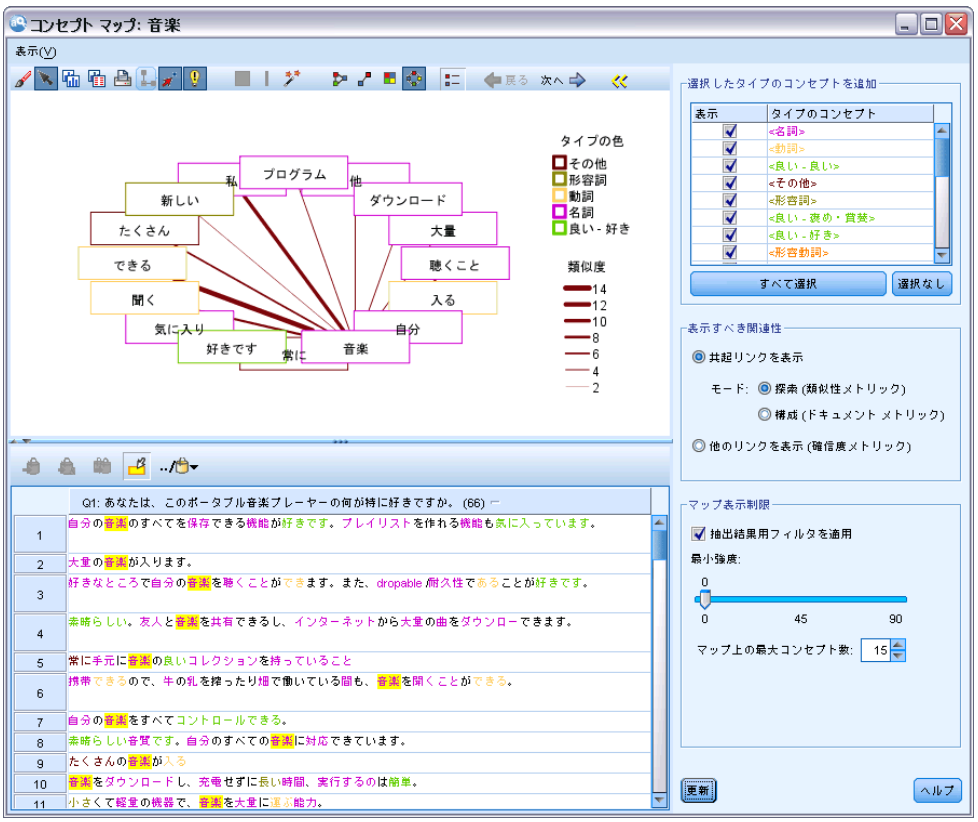
- メニューの [ツール] → [フィルタ] を選選択します。［フィルタ］ダイアログ ボックスが開きます。
- 使用するフィルタを選択および調整します。
- [OK] をクリックして、フィルタを適用すると、［抽出結果］ウィンドウに新しい結果が表示されます。

コンセプト マップの検証

コンセプト マップを作成して、コンセプトがどのように相関するかを検証できます。1 つのコンセプトを選択し、[**マップ**] をクリックすると、コンセプト マップのウィンドウが開き、選択したコンセプトに関連するコンセプトのセットを検証することができます。含めるタイプ、検索する関係性の種類など、設定を編集して表示するコンセプトを除外できます。

重要! マップを作成する前に、インデックスを生成する必要があります。これには数分かかることがあります。ただし、いったんインデックスを生成すると、再抽出するまで指標を再生成する必要がありません。抽出するごとにインデックスを自動的に生成したい場合は、抽出設定でそのオプションを選択します。 詳細は、 p. 156 データの抽出 を参照してください。

図 9-8
選択したコンセプトのコンセプト マップ



コンセプト マップを表示するには

▶ [抽出結果] ウィンドウで、1 つのコンセプトを選択します。

- このパネルのツールバーで、[**マップ**] ボタンをクリックします。マップ インデックスがすでに生成されている場合、コンセプト マップが個別のダイアログで開きます。マップ インデックスが生成されていない、または古い場合、インデックスを再度作成する必要があります。このプロセスには数分かかることがあります。

- 検証するマップの周辺をクリックします。リンクしたコンセプトをダブルクリックすると、マップが再描画再描画され、ダブルクリックしたコンセプトのリンクしたコンセプトが表示されます。

- 最上部のツールバーには、以前のマップに戻す、関係の強度に応じてリンクをフィルタリング、出現するコンセプトのタイプや表示する関係の種類を制御するフィルタ ダイアログを開くなど、基本的なマップ ツールがいくつかあります。2 番目のツールバーのラインには、グラフ編集ツールがあります。 詳細は、 13 章 p. 287 グラフのツールバーおよびパレットの使用 を参照してください。

- 検出されるリンクの種類が適切でない場合、マップに右側に表示されたこのマップの設定を確認してください。

マップの設定:選択したタイプのコンセプトを追加

テーブルの選択されたタイプに属するこれらのコンセプトのみがマップに表示されます。特定のタイプのコンセプトを隠すには、テーブルの該当するタイプの選択を解除します。

マップの設定:表示すべき関連性

共起リンクを表示:共起リンクを表示するには、モードを選択します。モードは、リンクの強度がどのように計算されたかに影響を与えます。

- 探索（類似性メトリック):類似性メトリックで、2 つのコンセプトが個別に出現する頻度と、同時に出現する頻度を考慮に入れた複雑な計算方法で、リンクの強度を算出します。高い強度の値は、コンセプトのペア

は、個別に出現するよりも、同時に出現する頻度が高いことを示します。この式により、浮動小数点値は整数に変換されます。

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

C_I は、コンセプト I が出現するドキュメントまたはレコードの数です。

C_J は、コンセプト J が出現するドキュメントまたはレコードの数です。

C_{IJ} は、コンセプトのペア I および J が出現するドキュメントまたはレコードの数です。

- 構成（ドキュメント メトリック）:ドキュメント メトリックを持つリンクの強度は、共起の行度数で決定します。一般的に、2 つのコンセプトがより頻繁に出現すると、同時に出現する確率が高くなります。強度の値が高いと、コンセプトのペアが頻繁に同時に出現します。

他のリンクを表示（確信度メトリック）:他のリンクの表示を選択することもできます。たとえば、セマンティック、派生（形態）、または、内包（シンタックス）であり、リンクされたコンセプトからいくつのステップのコンセプトが除去されたかに関連します。このことは、リソースの調整、特に同義語 や曖昧さの回避に役立ちます。このようなグループ化の手法の簡単な説明は、詳細言語設定 p. 198 を参照してください。

注:インデックスを作成するときにこれらのオプションが選択されなかった場合、または関連性が見つからなかった場合、何も表示されません。 詳細は、 p. 167 コンセプト マップ インデックスの作成 を参照してください。

マップの設定:マップ表示制限

抽出結果フィルタの適用:すべてのコンセプトを使用する場合、[抽出結果] ウィンドウでフィルタを使用して、表示内容を制限することができます。このオプションを選択すると、IBM® SPSS® Modeler Text Analytics が、フィルタリングされたセットを使用して、関連コンセプトを検索します。詳細は、 p. 161 抽出結果のフィルタリング を参照してください。

最小強度:ここで、最小強度を設定します。関連性の強度がこの制限値より低い関連コンセプトは、マップに表示されません。

マップ上の最大コンセプト数:マップに表示する関連性の最大数を指定します。

コンセプト マップ インデックスの作成

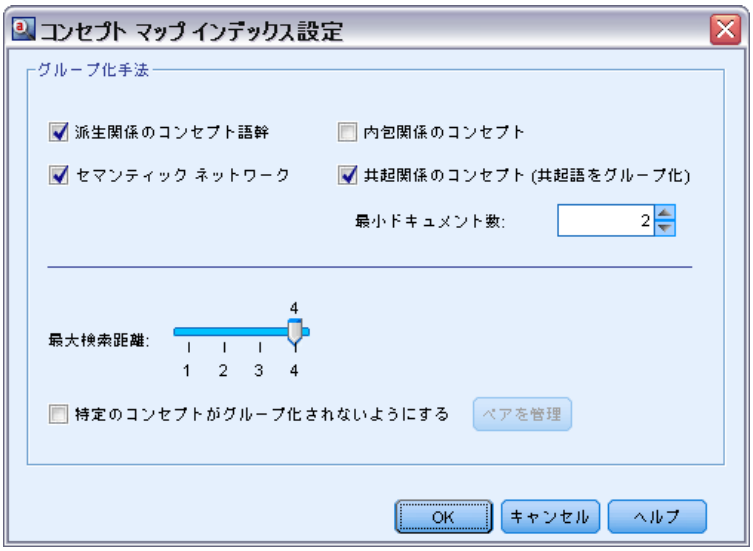
マップを作成する前に、コンセプトの関連性のインデックスを生成する必要があります。コンセプト マップを作成する場合、IBM® SPSS® Modeler Text Analytics は、このインデックスを参照します。このダイアログで手法を選択して、インデックスへの関連性を選択できます。

グループ化手法:1 つまたは複数の手法を選択します。これらの手法の簡単な説明は、「言語学的手法について」(p. 203) を参照してください。 すべてのテキスト言語ですべての手法が使用できるわけではありません。

特定のコンセプトがグループ化されないようにする: 出力の 2 つのコンセプトがグループ化またはペアとならないように処理を停止します。コンセプト ペアを作成または管理するには、[**ペアを管理**] をクリックします。 詳細は、10 章 p. 202 例外ペアのリンクの管理 を参照してください。

インデックスの作成には、数分かかる場合があります。ただし、いったんインデックスを生成すると、再抽出するまで、または設定を変更してより多くの関連性を追加しない限り、インデックスを再生成する必要がありません。抽出するごとにインデックスを生成したい場合は、抽出設定でそのオプションを選択します。 詳細は、 p. 156 データの抽出 を参照してください。

図 9-9
インデックスの設定



抽出結果の調整

抽出とは反復可能なプロセスで、結果を抽出、確認、結果を変更、および再抽出して結果を更新できます。テキスト マイニングおよびカテゴリ化には精度および継続性が不可欠であるため、最初から抽出結果を調整することによって、再抽出ごとに、カテゴリ定義でまったく同じ結果を取得します。このようにして、レコードおよびドキュメントをより正確で、繰り返し可能な方法で、カテゴリに割り当てます。

抽出結果は、カテゴリを作成するのための要素となります。これらの抽出結果を使用してカテゴリを作成すると、1 つまたは複数のカテゴリ記述子に一致するテキストが含まれる場合、レコードおよびドキュメントが自動的にカテゴリに割り当てられます。言語リソースを調整する前にカテゴリ化を開始できますが、開始前に少なくとも 1 回、抽出結果を確認しておくと役立ちます。

結果を確認すると、抽出エンジンが異なる方法で処理する必要のある要素が見つかる場合があります。以下のような例があります。

- **認識されない類義語:賢い、知的、頭脳明晰、博識**など、類義語と考えられるいくつかのコンセプトが見つかり、抽出結果に個別のコンセプトとしてすべて表示されたとします。**知的、頭脳明晰、博識**がすべて代表コンセプト**賢い**の名ですべてグループ化されるよう、類義語定義を作成できます。こうすることにより、これらのコンセプトをすべて**賢い**とグループ化し、グローバル出現頻度も高くなります。 詳細は、p. 169 類義語の追加 を参照してください。
- **ミスタイプ コンセプト**: 抽出結果のコンセプトがあるタイプに出現し、別のタイプに割り当てたい場合があります。また、抽出結果に 15 種類の野菜のコンセプトがあり、それらすべてを <Vegetable> という新しいタイプに追加したい場合もあります。 ほとんどの言語の場合、キーワード辞書になく、テキストから抽出されたコンセプトは、自動的に<Unknown>のタイプとなりますが、日本語テキストの場合は、自動的に<**名詞**>ノートのタイプとなります：日本語テキスト展開は IBM® SPSS® Modeler Premiumで利用可能です。 コンセプトをタイプに追加できます。 詳細は、 p. 172 コンセプトのタイプへの追加 を参照してください。
- **重要でないコンセプト**: 抽出されたコンセプトで非常に頻度の高い、つまり多くのレコードまたはドキュメントで見つかる場合があります。ただし、このコンセプトは分析には重要でないと見なされます。このコンセプトを抽出から除外できます。 詳細は、 p. 174 コンセプトの抽出からの除外 を参照してください。
- **不正な合致**: 特定のコンセプトを含むレコードまたはドキュメントを確認する場合、faculty（**能力**）と facility（**施設**）のように 2 つの単語が誤ってグループ化されているのを発見する場合があります。この合致は Fuzzy Grouping という、内部アルゴリズムによるものであり、2 つまたは 3 つの子音および母音を一時的に無視して、一般的

なスペルミスをグループ化します。これらの単語を無視する必要のある単語のペアのリストに追加できます。詳細は、 18 章 p.357 Fuzzy Grouping を参照してください。 Fuzzy grouping は、日本語テキストに対しては使用できません。

- **未抽出コンセプト**: 特定のコンセプトが抽出されるのを期待しているにもかかわらず、レコードまたはドキュメント テキストを確認しているときに一部の単語または句が抽出されていないことに気づく場合があります。これらの単語は重要でない動詞または形容詞でない場合が多くあります。ただし、抽出されなかった単語または句をカテゴリ定義として使用したい場合があります。コンセプトを抽出するために、キーワードをキーワード辞書に強制投入できます。詳細は、 p.175 単語を抽出に強制投入 を参照してください。

こうした変更の多くは、1 つまたは複数の要素を選択して右クリックし、コンテキスト メニューを使用することによって、[抽出結果] パネル、[データ] パネル、[カテゴリ定義] ダイアログ ボックス、または [クラスタ定義] ダイアログ ボックスから直接実行できます。

変更を行った後、パネルの背景色が変わり、変更を表示するには再抽出が必要であることを示します。詳細は、 p.156 データの抽出 を参照してください。 大きいデータセットを扱う場合、変更ごとではなく、複数の変更を行った後に再抽出を行うとより効率的です。

注:リソース エディタ ビューで抽出結果を作成するために使用する、編集可能な言語リソースのセット全体を表示できます (**[表示] → [リソース エディタ]**)。これらのリソースは、このビューにライブラリおよび辞書の形式で表示されます。ライブラリおよび辞書内で直接コンセプトおよびタイプをカスタマイズできます。詳細は、 16 章 p.316 ライブラリの使用 を参照してください。

類義語の追加

類義語は、同じ意味を持つ 2 つ以上の単語と関連があります。類義語はキーワードとその短縮形をまとめるのにもよく使用されます。またよくつづり間違いが起こる語を正しい書き方のもので置き換えるのにも使用されます。類義語を使用すると、代表コンセプトの頻度が高くなり、テキスト データ内のさまざまな方法で表示される類似した情報を見つけやすくなります。

製品に付属する言語リソース テンプレートおよびライブラリには、事前定義された多くの類義語が含まれています。ただし、認識されていない類義語が見つかった場合、次回抽出するときに認識されるよう、類義語を定義することができます。

まず、代表コンセプトまたは主要キーワードを決定します。**代表コンセプト**は、最終的な結果ですべての類義語のキーワードをグループ化したい単語または句です。抽出時、類義語は、この代表コンセプトの下でグルー

プ化されます。次に、このコンセプトのすべての類義語を特定します。代表コンセプトは、最終的な抽出で、すべての類義語と置き換えられます。類義語となるためにはキーワードは抽出されなければなりません。ただし、代表コンセプトを、置換えを行うために抽出する必要はありません。たとえば、**知的**という単語を**賢い**という単語に置き換えたい場合、**知的**が類義語となり、**賢い**が代表コンセプトとなります。

新しい類義語定義を作成する場合、新しい代表コンセプトが辞書に追加されます。その後、類義語をその代表コンセプトに追加する必要があります。類義語を作成または編集する場合、これらの変更が リソース エディタ の類義語辞書に記録されます。これらの類義語辞書の内容全体を表示したい場合、またはかなりの数の変更を行いたい場合、リソース エディタで直接作業することが必要な場合があります。 詳細は、 17 章 p.344 類義語辞書 を参照してください。

新しい類義語は、リソース エディタ ビューのライブラリ ツリーの表示された最初のライブラリに自動的に保存されます。デフォルトでは、これがローカル ライブラリとなります。

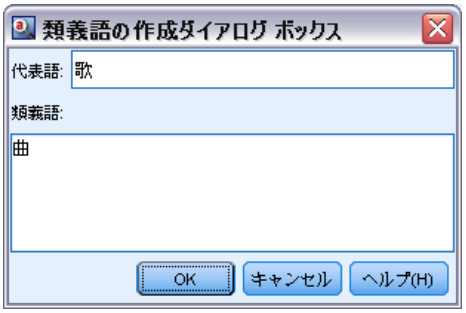
注:類義語定義を探していても、コンテキスト メニューまたは リソース エディタ で直接見つからない場合、内部の Fuzzy Grouping 手法によって、合致が発生している場合があります。 詳細は、 18 章 p.357 Fuzzy Grouping を参照してください。

新しい類義語を作成するには

▶ [抽出結果] パネル、[データ] パネル、[カテゴリ定義] ダイアログ ボックス、または [クラスタ定義] ダイアログ ボックスで、新しい類義語を作成したいコンセプトを選択します。

▶ メニューから **[編集] → [類義語への追加] → [新規]** を選択します。[類義語の作成] ダイアログ ボックスが開きます。

図 9-10
[類義語の作成] ダイアログボックス



▶ [代表語] テキスト ボックスに、代表語を入力します。これが、すべての類義語がグループ化されるコンセプトとなります。

- さらに類義語を追加したい場合、[類義語] リスト ボックスにそれらを入力します。辞書エディタの区切り文字（デフォルトでは「, 」（コンマ））を使って、各類義語を分けます。 詳細は、 8 章 p. 143 オプション：[セッション] タブ を参照してください。

- 日本語テキストを扱う場合、**[タイプの類義語]** フィールドでタイプ名を選択して、これらの類義語のタイプを指定します。ただし、代表語は、抽出時にタイプが割り当てられます。代表語がコンセプトとして抽出されない場合、抽出結果でこの列に表示されたタイプが代表語に割り当てられます。

  注: 日本語テキスト展開はIBM® SPSS® Modeler Premiumで利用可能です。

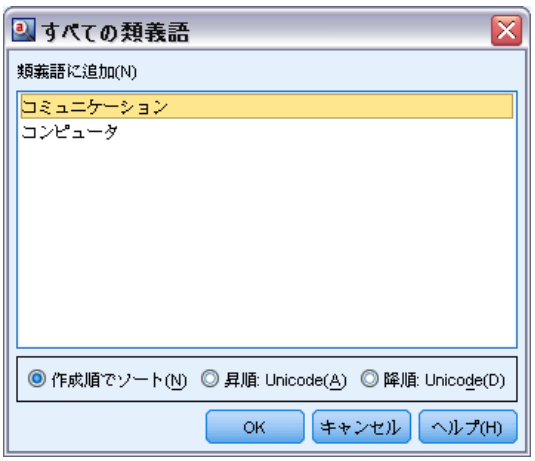
- [OK] をクリックし、変更を適用します。ダイアログ ボックスが閉じて、[抽出結果] パネルの背景色が変わり、変更を表示するには再抽出が必要であることを示します。変更が複数ある場合には、再抽出する前に変更を終わらせてください。

類義語に追加するには

- [抽出結果] パネル、[データ] パネル、[カテゴリ定義] ダイアログ ボックス、または [クラスタ定義] ダイアログ ボックスで、既存の類義語定義に追加したいコンセプトを選択します。

- メニューから **[編集] → [類義語への追加]** を選択します。メニューには、一番最後に作成された代表語がリストの一番上に表示されます。選択したコンセプトを追加したい類義語の名前を選択します。探している類義語が表示されたら、その名前を選択します。選択したコンセプトがその類義語の定義に追加されます。メニューに類義語が表示されない場合、**[もっと表示]** を選択すると [すべての類義語] ダイアログ ボックスが表示されます。

図 9-11
[すべての類義語] ダイアログ ボックス



▶ [すべての類義語] ダイアログ ボックスで、リストを自然な並び順（作成順）の昇順または降順で並べ替えることができます。選択したコンセプトを追加したい類義語の名前を選択し、[OK] をクリックします。ダイアログ ボックスが閉じ、コンセプトが類義語の定義に追加されます。

コンセプトのタイプへの追加

抽出を実行している場合、共通点を持つキーワードをグループ化するために、抽出されたコンセプトがタイプに割り当てられます。IBM® SPSS® Modeler Text Analytics は、多くのビルトインのタイプに付属しています。 詳細は、 17 章 p. 334 ビルトインのタイプ を参照してください。ほとんどの言語の場合、キーワード辞書になく、テキストから抽出されたコンセプトは、自動的に<Unknown>のタイプとなりますが、日本語テキストの場合は、自動的に<**名詞**>ノートのタイプとなります：日本語テキスト展開はIBM® SPSS® Modeler Premiumで利用可能です。

結果を確認しているとき、あるタイプに割り当てたいコンセプトがいくつか異なるタイプに割り当てられていたり、単語のグループが実際は新しいタイプに割り当てられていたりすることが分かる場合があります。こうした場合、コンセプトを別のタイプに割り当てなおすか、まとめて新しいタイプを作成する場合があります。 日本語テキストの新しいタイプを作成することはできません。

たとえば、自動車に関連する調査データを扱っており、車のさまざまな領域に焦点を当ててのカテゴリ化に関心があるとします。<Dashboard> というタイプを作成し、車のダッシュボードにある計測器およびノブに関連するすべてのコンセプトをグループ化することができます。そして、gas gauge、gas gauge、radio、および odometer などのコンセプトを、この新しいタイプに割り当てることができます。

またたとえば、大学に関連する調査データ、そして <Organization> ではなく <Person> というタイプとしての抽出 Johns Hopkins（大学）を扱っているとします。この場合、このコンセプトを <Organization> タイプに追加することができます。

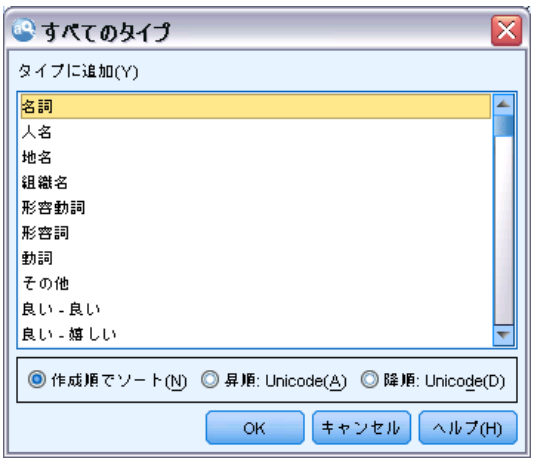
タイプを作成、またはコンセプトをタイプのキーワード リストに追加する場合、これらの変更は リソース エディタ の言語リソース ライブラリのキーワード辞書に記録されます。これらのライブラリの内容を表示したい場合、またはかなりの数の変更を行いたい場合、リソース エディタ で直接作業することが必要な場合があります。 詳細は、 17 章 p. 337 キーワードを追加 を参照してください。

コンセプトをタイプに追加するには

▶ [抽出結果] パネル、[データ] パネル、[カテゴリ定義] ダイアログ ボックス、または [クラスタ定義] ダイアログ ボックスで、既存のタイプに追加したいコンセプトを選択します。

- 右クリックしてコンテキスト メニューを開きます。
- メニューから [**編集**] → [**タイプへの追加**] を選択します。タイプ名が見つかったなら、これを選択します。選択したコンセプトを追加したいタイプの名前を選択します。探しているタイプの名前が表示されたら、その名前を選択します。選択したコンセプトがそのタイプに追加されます。メニューに類義語が表示されない場合、[**もっと表示**] を選択すると［すべてのタイプ］ダイアログ ボックスが表示されます。

図 9-12
[すべてのタイプ] ダイアログ ボックス



- ［すべてのタイプ］ダイアログ ボックスで、リストを自然な並び順（作成順）の昇順または降順で並べ替えることができます。選択したコンセプトを追加したいタイプの名前を選択し、[OK] をクリックします。ダイアログ ボックスが閉じ、コンセプトがタイプに追加されます。

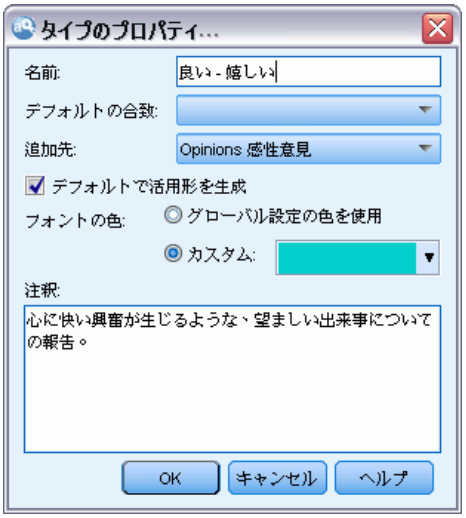
注: 日本語テキストでは、キーワードのタイワードを変更しても最終的な抽出リストで最終的に割り当てられるタイプは変更されないという例がいくつかあります。これは、いくつかの基本キーワードの抽出時に優先される内部辞書によるものです。

注: 日本語テキスト展開はSPSS Modeler Premiumで利用可能です。

新しいタイプを作成するには

- ［抽出結果］パネル、［データ］パネル、［カテゴリ定義］ダイアログ ボックス、または［クラスタ定義］ダイアログ ボックスで、新しいタイプを作成したいコンセプトを選択します。
- メニューから [**編集**] → [**タイプへの追加**] → [**新規**] を選択します。［タイプのプロパティ］ダイアログ ボックスが開きます。

図 9-13
[タイプのプロパティ] ダイアログ ボックス



- [名前] テキスト ボックスにこのタイプの新しい名前を入力し、他のフィールドを変更します。 詳細は、 17 章 p. 334 タイプの作成 を参照してください。

- [OK] をクリックし、変更を適用します。ダイアログ ボックスが閉じて、[抽出結果] パネルの背景色が変わり、変更を表示するには再抽出が必要であることを示します。変更が複数ある場合には、再抽出する前に変更を終わらせてください。

コンセプトの抽出からの除外

結果を確認しているときに、自動カテゴリ作成手法によって必要のないコンセプトが抽出または使用されていることが分かることがあります。これらのコンセプトの頻度が非常に高く、分析には全く重要でない場合もあります。この場合、コンセプトを不要なものとしてマークすることができます。通常、このリストに追加するコンセプトは、穴埋めのための単語または句で、一貫性を維持するためにテキストで使用され重要でなく、抽出結果を混乱させる場合があります、コンセプトを不要語辞書に追加することによって、それらが確実に抽出されないようにします。

不要語に追加することにより、次回抽出するときには除外コンセプトのすべての変化が抽出結果から除外されています。このコンセプトがカテゴリの記述子として出現している場合、再抽出後は 0 度数でカテゴリ内に残ります。

不要語に追加すると、これらの変更が リソース エディタ の不要語辞書に記録されます。不要語辞書のすべてを表示し、それらを編集したい場合、リソース エディタ で直接作業することが必要な場合があります。 詳細は、 17 章 p.350 不要語辞書 を参照してください。

注:日本語テキストでは、キーワードまたはタイプを除外しても、最終的には除外されない場合があります。これは、日本語リソースの基本キーワードの抽出時に優先される内部辞書によるものです。

注:日本語テキスト展開はIBM® SPSS® Modeler Premiumで利用可能です。

コンセプトを不要語に追加するには

- ［抽出結果］パネル、［データ］パネル、［カテゴリ定義］ダイアログ ボックス、または［クラスタ定義］ダイアログ ボックスで、不要語に追加したいコンセプトを選択します。
- 右クリックしてコンテキスト メニューを開きます。
- **［不要語に追加］**を選択します。コンセプトが リソース エディタ の不要語辞書に追加され、［抽出結果］パネルの背景色が変わり、変更を表示するには再抽出が必要であることを示します。変更が複数ある場合には、再抽出する前に変更を終わらせてください。

  注：不要語に追加した単語は、リソース エディタ のライブラリ ツリーに表示された最初のライブラリに自動的に保存されます。デフォルトでは、これがローカル ライブラリとなります。

単語を抽出に強制投入

抽出後、［データ］パネルでテキストデータを確認するとき、一部の単語または句が抽出されていないことが分かる場合があります。これらの単語は重要でない動詞または形容詞でない場合が多くあります。ただし、抽出されなかった単語または句をカテゴリ定義として使用したい場合があります。

これの単語および句を抽出したい場合、キーワードをタイプ ライブラリに強制投入できます。詳細は、 17 章 p.341 キーワードの強制 を参照してください。

重要! 辞書のキーワードを強制することが、絶対に確実というわけではありません。強制することによって、キーワードを辞書に明示的に追加している場合でも、再抽出後に［抽出結果］パネルに出現しない場合、あるいは出現しても宣言したとおりのものではない場合があります。抽出のプロセスにおいてある語や語句がより長い語句の一部として既に抽出されてしまっている場合や、語が品詞（名詞、動詞、形容詞、前置詞など々）に分割されている場合が考えられます。これを回避するために、**全体（複合語なし）**

マッチ オプションをキーワード辞書のこのキーワードに適用します。 詳細は、 17 章 p. 337 キーワードを追加 を参照してください。

章
10

テキスト データのカテゴリ化

テキスト データの分類

カテゴリとコンセプト ビューで、テキスト内の主要なキーワード、情報、属性をキャプチャする高いレベルのコンセプトまたはトピックを示す**カテゴリ**を作成できます。

IBM® SPSS® Modeler Text Analytics のリリース 14 の時点では、分類は階層的な構造を持つことができ、すなわち、サブカテゴリーを含むことができ、また、そのサブカテゴリーにもそれ自身のサブカテゴリーを更に下の階層に向かって持たせることができます。 製品内にこのような階層的な分類を構築することが可能であるだけでなく、階層的な分類を持ち、以前はコード フレームと呼ばれていた、定義済みの分類構造をインポートすることも可能です。

実際に、階層的カテゴリーにより、1個または複数のサブカテゴリーを持つツリー構造を構築して、たとえば異なるコンセプトやトピックの分野の項目をより正確にグループ化することができます。 レジャー活動に関して簡単な例を挙げることができます。「時間があればどんな活動がしたいですか？」という質問に対する答えとして、「スポーツ」、「日曜大工」、「釣り」などをトップ カテゴリに設定し、「スポーツ」の下の階層に「球技」、「水泳」などを設定できます。

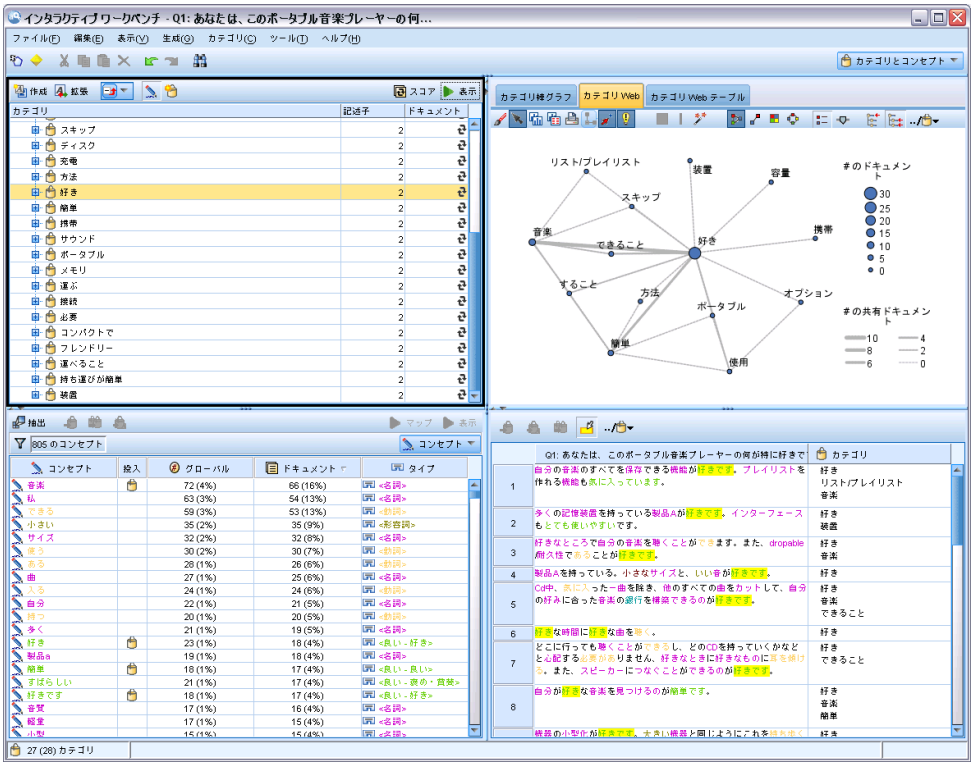
カテゴリは、コンセプト、タイプ、パターンおよびカテゴリ規則などの一連の記述子で構成されています。 また、これらの記述子を共に使用して、ドキュメントまたはレコードが指定されたカテゴリに属するかどうかを特定します。 ドキュメントまたはレコード内のテキストをスキャンして、テキストが記述子に一致するかどうかを確認することができます。 一致が見つかった場合は、ドキュメントまたはレコードはそのカテゴリに割り当てられます。 このプロセスを、**カテゴリ化**といいます。

カテゴリとコンセプト ビューの 4 つのパネルに表示されたデータを使用して、カテゴリを処理、作成および視覚的に検証することができます。また、[表示] メニューからその名前を選択して隠したり表示したりできます。

- **[カテゴリ] パネル:** このパネルでカテゴリを作成し、管理します。 詳細は、 p. 179［カテゴリ］パネル を参照してください。
- **[抽出結果] パネル:** このパネルで抽出したコンセプトおよびタイプを検証および処理します。 詳細は、 9 章 p. 152 抽出結果:コンセプトとタイプ を参照してください。

- **[視覚化] パネル**: このパネルでカテゴリについて、またカテゴリがどのように相互作用するかを視覚的に検証します。 詳細は、 13 章 p. 278 カテゴリ グラフおよび図表 を参照してください。
- **[データ] パネル**: このパネルでの選択に対応するドキュメントおよびレコード内に含まれるテキストを検証および確認できます。 詳細は、 p. 190 [データ] パネル を参照してください。

図 10-1
分類とコンセプト ビュー



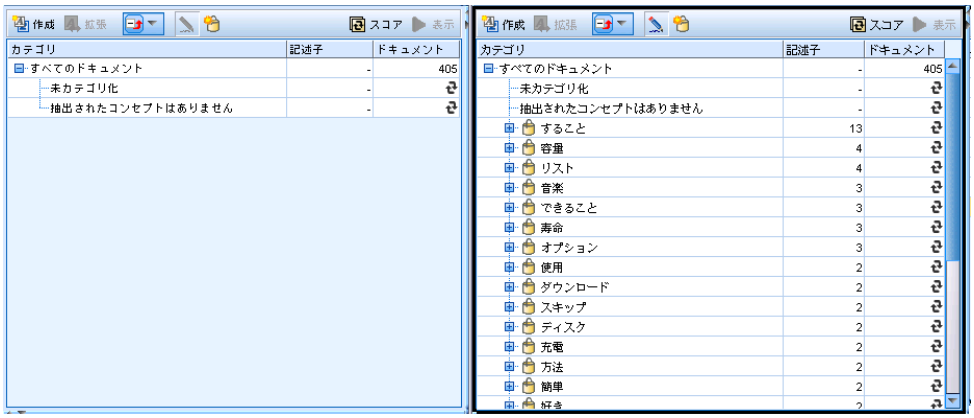
テキスト分析パッケージ（TAP）の一連のカテゴリからはじめることができますが、独自のカテゴリを作成する、または定義済みカテゴリ ファイルからインポートする必要もあります。 コンセプト、タイプ、パターンの抽出結果を使用する製品の自動化手法の頑健なセットを使用して、カテゴリおよびそれらの記述子を自動的に作成することができます。 データに関する追加の洞察を使用して、カテゴリを手動で作成することもできます。 ただし、インタラクティブ ワークベンチを使用してのみ、カテゴリを手動で作成し、調整できます。 詳細は、 3 章 p. 41 テキスト マイニング ノード:[モデル] タブ を参照してください。 抽出結果をカテゴリに手動でドラッグ アンド ドロップして、カテゴリ定義を作成できます。 カテゴリ規則をカテゴリに追加し、独自の事前定義済みカテゴリを使用して、これらのカテゴリまたは空のカテゴリの品質を向上させることができます。

それぞれの手法は、特定の種類のデータおよび状況に適していますが、同じ分析で手法を組み合わせて、全範囲のドキュメントまたはレコードをキャプチャすると役に立つ場合があります。 またカテゴリ化を行う際には、言語リソースに対して変更を加えたほうがよい場合もあります。

[カテゴリ] パネル

[カテゴリ] パネルでは、カテゴリを作成および管理できます。 このパネルは、カテゴリとコンセプト ビュー の左上に表示されます。 テキストデータからコンセプトとタイプを抽出した後、内包関係のコンセプト、共起など、自動的な方法を使用して、または手動で作成してカテゴリの作成を開始できます。 詳細は、 p. 193 カテゴリの作成 を参照してください。

図 10-2
カテゴリがない場合とある場合の [カテゴリ] パネル



カテゴリを作成または更新するごとに、ドキュメントまたはレコードが [スコア] ボタンをクリックするとスコアリングされ、いずれかのテキストが任意のカテゴリ内のデスクリプターと合致するか確認されます。 一致が見つかった場合は、ドキュメントまたはレコードはそのカテゴリに割り当てられます。 この最終結果は、ドキュメントまたはレコードのすべてではなくともその多くが、カテゴリの記述子に基づいて、カテゴリに割り当てられます、

カテゴリ ツリー一覧

このパネルのツリー一覧には、一連のカテゴリ、サブカテゴリ、記述子が表示されます。 ツリーには、各ツリー項目の情報を示す列があります。 表示できるのは次の列です。

- **コード**: 各カテゴリのコード値を表示します。 この列は、デフォルトでは非表示になっています。 この列は、次のメニューから表示できます。 **表示** > **[カテゴリ] パネル**を選択すると表示されます。

- **カテゴリ:** カテゴリ ツリーには、カテゴリおよびサブカテゴリの名前が表示されます。 また、記述子のツールバー アイコンをクリックすると、一連の記述子も表示されます。
- **記述子:** その定義を構成する記述子の数が表示されます。 この度数には、サブカテゴリの記述子数は含まれません。 [**カテゴリ**] 列に記述子名が表示されている場合、度数は表示されません。 ツリー内の記述子自体の表示または非表示をするには、次のメニューから行います。 **表示** > [**カテゴリ**] **パネル** > **すべての記述子**を選択すると表示されます。
- **ドキュメント:** スコアリング後、この列には、該当するカテゴリとすべてのサブカテゴリにカテゴリ化されているドキュメントまたはレコードの数が表示されます。 つまり、5 つのレコードが記述子に基づいて上位カテゴリに合致し、7 つの異なるレコードがその記述子に基づいてサブカテゴリに合致する場合、上位カテゴリのドキュメント数の合計は、この 2 つの数値の合計、この場合は 12 となります。ただし、同じレコードが上位カテゴリとそのサブカテゴリに合致する場合、度数は 11 となります。

カテゴリがない場合でも、テーブルには 2 つの行が表示されます。 上の行の [**すべてのドキュメント**]は、ドキュメント数またはレコード数の合計です。 2 番目の行 [**未カテゴリ化**] には、カテゴリがされていないドキュメント/レコード数が表示されます。

パネルの各カテゴリについて、小さい黄色のバケツのアイコンがカテゴリ名の前に表示されます。 カテゴリをクリックを選択、またはメニューで **表示** > **カテゴリ定義**を選択すると、カテゴリ定義ダイアログボックスが開き、**記述子**と呼ばれるすべての要素が表示されます。記述子はコンセプト、タイプ、パターン、カテゴリー ルールなどの定義を決定します。 詳細は、p. 188 カテゴリとは を参照してください。 デフォルトでは、カテゴリ ツリー一覧には、カテゴリの記述子は表示されません。［カテゴリ定義］ダイアログ ボックスではなくツリーで直接記述子を表示する場合、ツールバーの鉛筆のアイコンが表示された切り替えボタンをクリックします。 この切り替えボタンを選択すると、ツリーが展開され、記述子が表示されます。

カテゴリのスコアリング

カテゴリ ツリー一覧の [**ドキュメント**] 列には、特定のカテゴリにカテゴリ化されているドキュメント数またはレコード数が表示されます。 数値が過去のものまたは計算されていない場合、アイコンがその列に表示されます。パネル ツールバーの[**スコア**] ボタンをクリックして、ドキュメント数を再計算することができます。 大きいデータセットを使用する場合、スコアリング プロセスには時間がかかる場合があります。

ツリー内のカテゴリの選択

ツリー内で選択すると、横グループのカテゴリのみ選択できます。つまり、上位レベルのカテゴリを選択すると、サブカテゴリは選択できません。 または指定されたカテゴリの 2 つのサブカテゴリを選択すると、同時に別のカテゴリのサブカテゴリを選択できません。 不連続なカテゴリを選択すると、以前の選択内容が失われます。

[データ] パネルおよび [視覚化] パネルの表示

テーブルで 1 行選択すると、[**表示**] ボタンをクリックして、[視覚化] パネルおよび [データ] パネルを更新して、選択内容に対応する情報を表示します。 パネルが表示されない場合、[**表示**] をクリックしてパネルを開きます。

カテゴリの調整

カテゴリ化を行っても、最初から完全な結果が得られるとは限りません。削除したいカテゴリや、他のカテゴリとまとめたいカテゴリもあるでしょう。 また、抽出結果を確認して、役立つと思われるいくつかのカテゴリが作成されていないことが分かる場合があります。 その場合、結果を手動で変更し、特定の状況に対して結果を調整することができます。 詳細は、p. 253 カテゴリの編集および調整 を参照してください。

カテゴリ作成の方法と戦略

また抽出していない、または抽出結果が古い場合、カテゴリ作成方法または拡張方法のいずれかを使用すると、抽出についてのプロンプトが自動的に表示されます。 手法を適用した後、カテゴリにグループ化したコンセプトおよびタイプはその他の手法で構築したカテゴリに使用できます。 つまり、再利用しないことを選択しないかぎり、複数のカテゴリのコンセプトを表示することができます。

最適なカテゴリを作成するために、次のことを確認してください。

- **カテゴリ作成の方法**
- **カテゴリ作成の戦略**
- **カテゴリ作成のヒント**

カテゴリ作成の方法

すべてのデータセットが一意であるため、カテゴリ作成方法の数やそれらを適用する順序は、時間によって変わる場合があります。 また、テキストマイニングの目標が、データセットによって異なる場合があるため、それぞれの手法を検証して、指定したテキスト ータにとってどの手法が最良の結果を生み出すかを確認する必要があります。 自動的手法では、データ

を完全にカテゴリ化できません。そのため、データに合った 1 つまたは複数の自動的手法を見つけ、適用することをお勧めします。

事前作成されたカテゴリ セットを持つテキスト分析パッケージ（TAP、*.tap）を使用するほか、次の方法を組み合わせて回答をカテゴリ化することもできます。

- **自動作成手法:** いくつかの言語ベースおよび頻度ベースのカテゴリオプションを使用して、カテゴリを自動的に作成できます。 詳細は、p. 193 カテゴリの作成 を参照してください。
- **自動拡張手法:** いくつかの言語的手法を使用し、記述子を追加および拡張することによって既存のカテゴリを展開し、より多くのレコードをキャプチャすることができます。 詳細は、 p. 212 カテゴリの拡張 を参照してください。
- **手動による手法:** ドラッグアンドドロップなど、手動による手法がいくつかあります。 詳細は、 p. 217 手作業でのカテゴリの作成 を参照してください。

カテゴリ作成の方略

次のリストの方略は包括的ではありませんが、カテゴリの作成方法について、いくつかのキーワードが用意されています。

- テキスト マイニング ノードを定義する場合、テキスト分析パッケージからカテゴリを選択し、いくつかの作成済みカテゴリを使用して分析を開始します。 これらのカテゴリは、テキストを最初から十分にカテゴリ化することができます。 ただし、カテゴリを追加する場合、カテゴリ作成設定（**[カテゴリ] → [設定を行う]**）を編集することができます。 **詳細設定**: **言語**ダイアログを開き、カテゴリ入力オプションの**未使用の抽出結果**を選択し、追加カテゴリを作成します。
- ノードを定義するときに、インタラクティブ ワークベンチのカテゴリとコンセプト ビューのTAP からカテゴリを選択します。 次に、未使用のコンセプトまたはパターンを適切なカテゴリにドラッグ アンド ドロップします。 そして、編集した既存のカテゴリを展開（**[カテゴリ] → [カテゴリを展開]**）し、既存のカテゴリ記述子に関連するより多くの記述子を取得します。
- 詳細言語設定を使用して、自動的にカテゴリを作成します（**[カテゴリ] → [カテゴリを作成]**）。 生成されるカテゴリが適切なものとなるまで、記述子を削除、カテゴリを削除、または同様のカテゴリを結合してカテゴリを手動で調整します。 また、元来［**可能な場合ワイルドカードを使用して一般化**］オプションを使用しないでカテゴリを作成する場合、［**一般化**］オプションをオンにして［カテゴリを展開］を使用し、自動的にカテゴリを簡略化することもできます。

- 非常に説明的なカテゴリ名および注釈を持つ事前定義済みカテゴリ ファイルをインポートします。 また、元来そのオプションを選択しないでインポートし、カテゴリ名から記述子をインポートまたは生成する場合、後で［カテゴリを拡張］ダイアログを使用して、**[カテゴリ名から生成された記述子を使用して空白のカテゴリを拡張する]** オプションを選択することができます。 そして、これらのカテゴリの 2 回目の展開を行いますが、今回はグループ化手法を使用します。
- コンセプトまたはコンセプト パターンを頻度によって並べ替え、最も関心のあるコンセプトまたはコンセプト パターンを［カテゴリ］パネルにドラッグ アンド ドロップすることによって、カテゴリの最初のセットを手動で作成します。 カテゴリの最初のセットを作成したら、展開機能（**[カテゴリ] → [カテゴリを展開]**）を使用して、すべてのカテゴリを展開し、その他の関連記述子を含めてより多くのレコードに一致するように選択したカテゴリを調整します。

これらの手法を適用した後、作成されたカテゴリを確認、手動による手法を使用して小さな調整を行い、誤分類を削除、または欠損したと考えられるレコードまたは単語を追加することをお勧めします。 また、さまざまな手法を使用すると、重複したカテゴリを作成する場合もあるため、必要に応じてカテゴリを結合または削除することもできます。 詳細は、p. 253 カテゴリの編集および調整 を参照してください。

カテゴリ作成のヒント

より良いカテゴリを作成できるよう、方法を決定できるヒントをいくつか確認できます。

カテゴリ-to-ドキュメント比率のヒント

ドキュメントおよびレコードが割り当てられるカテゴリは、少なくとも 2 つの理由で、質的テキスト分析では相互に排他的である場合がよくあります。

- 1 つめの理由は、一般的に、テキスト ドキュメントまたはレコードが長いほど、表されるキーワードおよび意見がより明確なものとなることです。 そのため、ドキュメントまたはレコードに複数のカテゴリが割り当てられるという機会が大幅に増えます。
- 2 つめの理由は、論理的に分けられていないテキスト ドキュメントまたはレコードをグループ化および解釈するさまざまな方法があるということです。 回答者の政治的な信念に関する、自由記述式の質問を含んだ調査の場合、 「リベラル」および「保守的」、または「共和党」および「民主党」のようなカテゴリのほか、「社会的にリベラル」、「財政的に保守的」など、より特定的なカテゴリも作成できます。 こ

れらのカテゴリは相互に重複部分がなかったり、すべてをカバーしたりしている必要はありません。

作成するカテゴリ数のヒント

カテゴリ作成は、データから直接生じることはありません。データについて何か興味深いものがあった場合、カテゴリを作成してその情報を示すことができます。 一般的に、作成するカテゴリに数について、推奨される上限はありません。 ただし、カテゴリをあまりに多く作成すると、管理できない場合があります。 次の 2 つの原則が適用されます。

- **カテゴリ度数:** カテゴリが役立つものになるには、最低限のドキュメントまたはレコードが必要です。 1 つまたは 2 つのドキュメントには非常に興味深いものが含まれている場合がありますが、それが 1,000 件のドキュメントのうちの 1 つまたは 2 つである場合、含まれる情報は、実際に役立つほど母集団の中では頻繁ではありません。
- **複雑さ:** 作成したカテゴリが多いほど、分析が完了した後確認および要約する必要がある情報が多くなります。 ただし、カテゴリ数が多すぎる場合、複雑さが増しても、役立つ詳細は増えません。

カテゴリ数が多すぎることを判断する規則、またはカテゴリあたりの最小レコード数を決定する規則はありません。 個々の状況の必要性に応じて分析者が判断する必要があります。

しかしながら、基本的なアドバイスはあります。 まずカテゴリの個数は多すぎてはいけませんが、分析の早い段階においては、カテゴリが少なすぎるよりは多すぎるほうがいいでしょう。 比較的類似したカテゴリをまとめるほうが、ケースを分けて新しいカテゴリに細分化するよりも簡単なので、多くのカテゴリからより少ない個数のカテゴリになるように作業していくほうが、一般に容易だといえます。 テキスト マイニングの反復性およびこのソフトウェア プログラムを使用して達成できる容易さにより、より多くのカテゴリを作成することが、開始時点では推奨されます。

最適な記述子の選択

次に、カテゴリに最適な記述子（コンセプト、タイプ、TLA パターンおよびカテゴリ規則）の選択または作成におけるガイドラインをいくつか示します。 記述子とは、カテゴリの構築ブロックです。 ドキュメントまたはレコードの一部またはすべてのテキストが記述子に合致する場合、ドキュメントまたはレコードはカテゴリに合致します。

記述子が抽出されたコンセプトまたはパターンを含まないまたは対応しない場合、ドキュメントまたはレコードに合致しません。 そのため、次で説明しているように、コンセプト、タイプ、パターン、およびカテゴリ規則を使用します。

コンセプトが、それ自体だけでなく複数/単数形、類義語、およびスペルの変異形にいたる一連の基本キーワードも示すため、コンセプト自体は、記述子、または記述子の一部として使用する必要があります。 指定されたコンセプトの基本キーワードについての詳細を知るには、カテゴリとコンセプト ビューの［抽出結果］パネルのコンセプト名をクリックします。 コンセプト名でマウスポインタを停止すると、ツールヒントが表示され、そこに最後の抽出時にテキストで検出された基本キーワードが表示されます。すべてのコンセプトに基本キーワードがあるわけではありません。 たとえば、car と vehicle は類義語ですが car がコンセプトとして、vehicle が基本キーワードとして抽出された場合、vehicle を含むドキュメントまたはレコードに合致するため、記述子には car だけを使用します。

記述子としてのコンセプトとタイプ

コンセプト（または基本キーワードのいずれか）を含むすべてのドキュメントまたはレコードを検出する場合、そのコンセプトを記述子として使用します。 この場合、正確なコンセプト名で十分であるため、より複雑なカテゴリ規則を使用する必要はありません。 意見を抽出するリソースを使用する場合、文の真の意味を抽出する TLA パターン抽出時にコンセプトが変更される場合があるので注意してください（TLA に関する次の項の例を参照してください）。

たとえば、「Apple and pineapple are the best」のような、各個人の好きな果物を示す調査の回答によって、apple と pineapple が抽出されます。 コンセプト apple をカテゴリに記述子として追加すると、コンセプト apple（または基本キーワードのいずれか）を含むすべての回答がそのカテゴリに合致します。

ただし、とにかく apple について言及する回答を知りたい場合、カテゴリ規則を * apple * のように作成すると、apple、apple sauce、または french apple tart のようなコンセプトを含む回答もキャプチャできます。

また、<Fruit> のようにタイプを記述子として直接指定することによって、同じようにタイプ指定されたコンセプトを含むすべてのドキュメントまたはレコードをキャプチャすることもできます。 タイプには * は使用できませんので注意してください。

詳細は、 9 章 p.152 抽出結果:コンセプトとタイプ を参照してください。

記述子としてのテキスト リンク分析（TLA）パターン

より詳細で微妙なアイデアをキャプチャする場合、TLA パターン結果を記述子として使用します。 テキストが TLA 抽出中に分析されると、 テキスト全体（ドキュメントまたはレコード）を処理するのではなく、一度に 1 文、または 1 句ずつテキストが処理されます。 1 つの文の全部分を考慮することによって、TLA は意見、 2 つの要素間の関係性、または否定的な

表現を特定して、真の意味を理解できます。 コンセプト パターンまたはタイプ パターンを記述子として使用できます。 詳細は、 12 章 p.270 タイプ パターンおよびコンセプト パターン を参照してください。

たとえば、「the room was not that clean」というテキストがある場合、次のようなコンセプトが抽出されます： room and clean。 ただし、抽出設定で TLA 抽出が有効になっている場合、TLA で clean が否定文で使用されており、実際は not clean に対応し、コンセプト dirty の類義語であることを検出できます。 ここで、記述子であるコンセプト clean がこのテキストに合致しますが、清潔さについて示すその他のドキュメントまたはレコードのキャプチャできることが確認できます。 そのため、このテキストに合致し、より適切な記述子となるため、dirty を出力コンセプトに指定した TLA コンセプト パターンを使用することをお勧めします。

記述子としてのカテゴリ ビジネス規則

カテゴリ規則とは、抽出したコンセプト、タイプ、パターン、およびブール型演算子を使用した論理式に基づいて、ドキュメントまたはレコードを自動的に分類するステートメントです。 たとえば、「このカテゴリに、**アルゼンチン**ではなく、**大使館**という抽出したコンセプトを含む」という内容を意味する式を作成することができます。

カテゴリ規則をカテゴリの記述子として記述および使用し、&、|、および !() によってさまざまなアイデアを表現することができます。 ブール値。 これらの規則のシンタックスおよびそれらの作成、編集方法の詳細は、「」() を参照してください。 カテゴリ規則の使用 p.219

- & (AND) ブール演算子を含むカテゴリ規則を使用して、2 つ以上のコンセプトが出現するドキュメントまたはレコードを検出します。 & 演算子でつながった 2 つ以上のコンセプトは、同じ文またはフレーズで出現する必要はありませんが、同じドキュメントまたはレコードのどこかに出現するとカテゴリに合致すると見なされます。 たとえば、記述子にカテゴリ規則 food & cheap を作成すると、テキストに food と cheap が含まれているため、food が cheap という名詞でないにかかわらず、「the food was pretty expensive, but the rooms were cheap」というテキストを含むレコードに合致します。
- !() を含むカテゴリ規則を使用します (NOT) ブール演算子が、いくつかのコンセプトまたはタイプのいずれかを含むドキュメントまたはレコードを検出します。 コンテキストではなく、単語に基づいて関連すると思われる情報がグループ化されないようにします。 たとえば、カテゴリ規則 <Organization> & !(ibm) を記述子として作成すると、「SPSS Inc. was a company founded in 1967」というテキストには合致しますが、「the software company was acquired by IBM.」には合致しません。

- | (OR) ブール演算子を含むカテゴリ規則を使用して、いくつかのコンセプトまたはタイプのいずれかを含むドキュメントまたはレコードを検出します。 たとえば、カテゴリ規則 (personnel|staff|team|coworkers) & bad を記述子として作成すると、これらの名詞がコンセプト bad と共に検出されるドキュメントまたはレコードと合致します。
- カテゴリ規則にタイプを使用すると、その規則はより一般的になり、より展開しやすくなる場合があります。 たとえば、ホテルのデータを扱っている場合、ホテルのスタッフに対して顧客がどう思っているかについて、非常に関心があります。 関連するキーワードには、receptionist、waiter、waitress、reception desk、front desk（受付、ウェイター、ウェイトレス、受付デスク、フロント デスク）などがあります。 この場合、<HotelStaff> という新しいタイプを作成し、上記のキーワードをすべてこのタイプに追加します。 [* waitress * & nice]、[* desk * & friendly]、[* receptionist * & accommodating] のようなすべての種類のスタッフに 1 つのカテゴリ規則を作成することができますが、タイプ <HotelStaff> を使用して 1 つの、より一般的なカテゴリ規則を作成して[<HotelStaff> & <Positive>] の形式でホテル スタッフに対して好意的な意見の回答をすべてキャプチャできます。

注： カテゴリ規則に TLA パターンを含む場合、これらの規則に + と & を両方使用できます。 詳細は、 p. 222 カテゴリ規則内の TLA パターンの使用 を参照してください。

記述子であるコンセプト、TLA、またはカテゴリ規則がどのように合致するかについての例

次の例では、記述子としてコンセプト、カテゴリ規則、TLA パターンを使用するとドキュメントまたはレコードがどのようにカテゴリ化されるのかについて説明します。 次のような 5 つのレコードがあるとします。

- A: “awesome restaurant staff, excellent food and rooms comfortable and clean.” （素晴らしいレストラン スタッフ、食べ物もおいしく、部屋は快適で清潔）
- B: “restaurant personnel was awful, but rooms were clean.”（レストラン スタッフはひどいが、部屋はきれいだった）
- C: “Comfortable, clean rooms.”（快適で清潔な部屋）
- D: “My room was not that clean. “（私の部屋はきれいではなかった）
- E: “Clean.”（きれいだった）

レコードに「clean」という単語を含み、この情報をキャプチャしたいため、次の表に示す記述子のいずれかを作成します。 キャプチャしようとしている核心に基づき、ある種類の記述子を別の記述子に対して使用すると、どのように異なる結果を生成するかを確認できます。

テーブル 10-1
レコードが記述子に合致する例

記述子	A	B	C	D	E	説明
clean	合致	合致	合致	合致	合致	記述子は抽出されたコンセプトです。 TLA のないレコード D も含めてすべてのレコードがコンセプト clean を含み、TLA 規則によって「not clean」が「dirty」を意味することは自動的に認識されません。
clean + .	–	–	–	–	合致	記述子は clean を示す TLA パターンです。 clean が TLA 抽出時に関連するコンセプトなしで抽出されたレコードとのみ合致します。
[clean]	合致	合致	合致	–	合致	記述子は、それ自体またはその他の規則に clean を含む TLA 規則を探すカテゴリ規則です。 clean を含む TLA 出力が、clean が room のような別のコンセプトにリンクするかどうか、別のスロット位置にあるかどうかに関係なく、検出されたすべてのレコードに合致します。

カテゴリとは

カテゴリは、密接に関連したコンセプト、オピニオン、または属性のグループのことを言います。 短いことば（ラベル）を付けて、カテゴリの内容が簡単にわかるようにしておくと便利です。

たとえば、新しい洗濯用洗剤について消費者からのアンケートの回答を分析する場合、製品の香りを示すすべての回答を含む「香り」というラベルの付いたカテゴリを作成できます。 ただし、そのようなカテゴリは良い香りと感じた消費者と、香りが強いと感じた消費者とを差別化するものではありません。 IBM® SPSS® Modeler Text Analytics は適切なリソースを使用する場合意見の抽出ができるため、2 つのカテゴリを他に作成して、「香りが良い」と答えた回答者と「香りは好みではない」と答えた回答者を特定することができます。.

カテゴリとコンセプト ビューウィンドウの左上のパネルの［カテゴリ］パネルで、カテゴリを作成したり作業することができます。 各カテゴリは、1 つまたは複数の記述子で定義されます。 **記述子**は、カテゴリを定義するために使用されているコンセプト、タイプ、パターンおよびカテゴリ規則です。

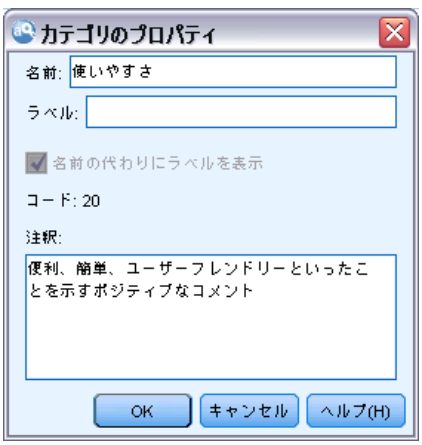
指定のカテゴリを構成する記述子を表示する場合、［カテゴリ］パネルのツールバーの鉛筆のアイコンをクリックし、ツリーを展開して記述子を表示します。 また、カテゴリを選択して［カテゴリ定義］ダイアログボックスを開きます（**[表示] → [カテゴリ定義]**）。

内包関係のコンセプトなどのカテゴリ作成手法を使用してカテゴリを自動的に作成する場合、その手法ではコンセプトおよびタイプを記述子として使用し、カテゴリを作成します。 TLA パターンを抽出すると、これらのパターンまたはパターンの一部をカテゴリ記述子として追加することもできます。詳細は、 12 章 p. 268 テキスト リンク分析の検証 を参照してください。 そしてクラスタを作成すると、クラスタのコンセプトを新しいまたは既存のカテゴリに追加することができます。 最後に、カテゴリ規則を手動で作成して、カテゴリの記述子として使用することができます。 詳細は、 p. 219 カテゴリ規則の使用 を参照してください。

カテゴリのプロパティ

記述子のほかカテゴリには、カテゴリの名前を変更、ラベルまたは注釈を追加するために編集できるプロパティもあります。

図 10-3
[カテゴリのプロパティ] ダイアログ



以下のようなプロパティがあります。

- **名前:** この名前は、デフォルトでツリーに表示されます。 自動化の手法でカテゴリを作成した場合、名前は自動的に付けられます。

- **ラベル:** ラベルの使用は、他の製品または他のテーブルまたはグラフで使用する、より重要なカテゴリ記述子を作成する場合に役立ちます。［名前の代わりにラベルを表示］を選択すると、インターフェイス内でカテゴリを特定する際にラベルが使用されます。
- **コード:** コード番号は、このカテゴリのコード値に対応します。 .
- **注釈:** このフィールドで各カテゴリの短い説明を追加できます。［カテゴリを作成］ダイアログでカテゴリを生成した場合、この注釈にメモが自動的に追加されます。 テキストを選択し、メニューから **[カテゴリ] → [注釈に追加]** を選択して、［データ］パネルからサンプル テキストを注釈に直接追加することもできます。

[データ] パネル

カテゴリを作成した場合、作業しているテキスト データを確認したい場合があります。たとえば、640 件のドキュメントがカテゴリ化されているカテゴリを作成する場合、実際にどのようなテキストが書かれているのかを確認するため、これらのドキュメントの一部またはすべてに目を通したい場合があります。右下の［データ］パネルでレコードまたはドキュメントを確認することができます。デフォルトで表示されない場合は、メニューから **[表示] → [パネル] → [データ]** を選択してください。

［データ］パネルには、特定の表示制限に応じて、［カテゴリ］パネル、［抽出結果］パネル、または［カテゴリ定義］ダイアログ ボックスの選択に該当するドキュメントまたはレコードごとに 1 行ずつ表示されます。デフォルトでは、［データ］パネルに表示されるドキュメントまたはレコード数が制限され、データをより迅速に表示できるようになります。ただし、これは［オプション］ダイアログ ボックスで調整できます。 詳細は、8 章 p. 143 オプション：［セッション］タブ を参照してください。

[データ] パネルの表示および更新

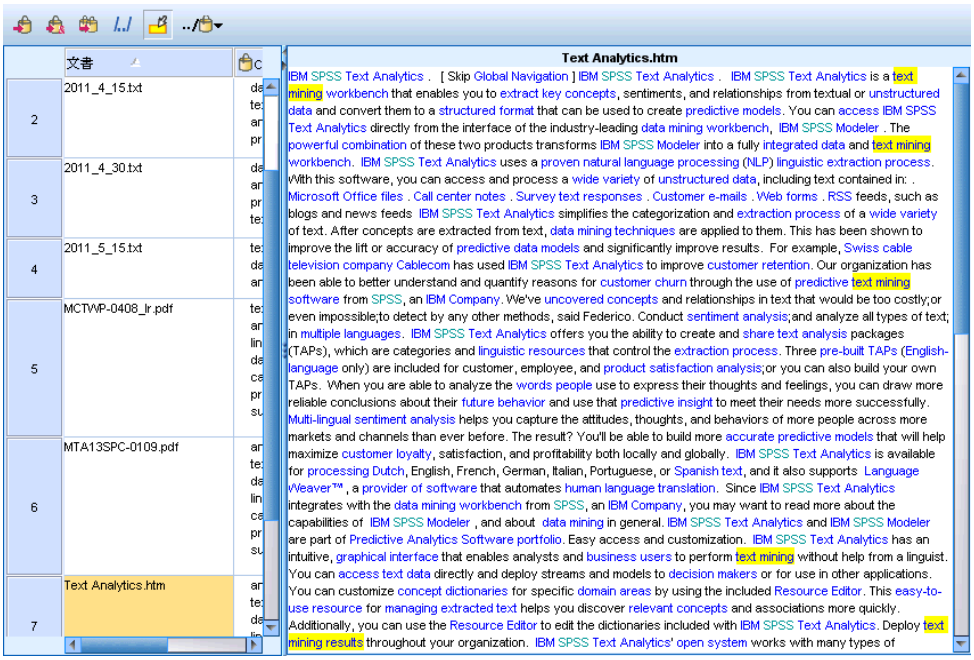
［データ］パネルでは、大きなデータセットの自動データ更新には時間がかかるため、自動的に表示の更新は行われません。そのため、このビューの別のパネルまたは［カテゴリ定義］ダイアログ ボックスで選択すると、**[表示]** をクリックして［データ］パネルの内容を更新します。

テキスト ドキュメントまたはレコード

テキスト データがレコードの形式で、テキストの長さが比較的短い場合、［データ］パネルのテキスト フィールドには、テキスト データの全体が表示されます。ただし、レコードおよび大きいデータセットを処理している場合、テキスト フィールドの列にはテキストの一部が表示され、右側の［テキスト プレビュー］パネルを開くと、テーブルで選択したレコードの大部分またはすべてが表示されます。テキスト データが個別ドキュメント

の形式の場合、[データ] パネルには、ドキュメントのファイル名が表示されます。ドキュメントを選択すると、[テキスト プレビュー] パネルには選択したドキュメントのテキストが表示されます。

図 10-4
[テキスト プレビュー] パネルが表示された [データ] パネル



色および強調表示

データを表示すると、該当するドキュメントまたはレコードのコンセプトおよび記述子は色付きで強調表示され、テキスト内のコンセプトおよび記述子を特定しやすくします。カラー コードは、コンセプトが属するタイプに対応します。カラーコード化された項目上でマウス ポインタを停止させて、項目が抽出されたコンセプトと、項目が割り当てられたタイプを表示することもできます。抽出されていないテキストは、黒で表示されます。通常、こうした抽出されていない単語は接続詞（「および」または「と」）、代名詞（「私」または「彼ら」）および 動詞（「いる」、「持つ」、または「取る」）の場合が多くあります。

[データ] パネルの列

テキスト フィールドの列は常に表示されていますが、その他の列も表示できます。その他の列を表示するには、メニューで **[表示] → [[データ] パネル]** を選択し、[データ] パネルに表示したい列を選択します。 表示できるのは次の列です。

- **「テキスト フィールド名」(#)/ドキュメント**: コンセプトおよびタイプが抽出されたテキスト データの列を追加します。データがドキュメントにある場合、列名は［ドキュメント］となり、ドキュメント ファイル名または完全パスのみが表示されます。これらのドキュメントのテキストを表示するには、［テキスト プレビュー］パネルを表示する必要があります。［データ］パネルの行数が、この列名の後のカッコ内に表示されます。読み込みの速度向上のために使用される［オプション］ダイアログの制約により、一部のドキュメントまたはレコードが表示されない場合があります。最大値に達すると、数値の後に［**最大**］と表示されます。詳細は、 8 章 p. 143 オプション：［セッション］タブ を参照してください。
- **カテゴリ**: レコードが属するカテゴリがそれぞれ表示されます。この列を表示する場合、最新の情報を示すため、［データ］パネルの更新に少し時間がかかる場合があります。
- **適合度順位**: 1 つのカテゴリの各レコードの順位が表示されます。この適合度順位は、カテゴリ内の他のレコードと比較して、レコードがカテゴリにどれだけ適合しているかを示します。［カテゴリ］パネル（左上のパネル）でカテゴリを選択すると、順位が表示されます。詳細は、 p. 192 カテゴリの関連性 を参照してください。
- **カテゴリの個数**: レコードが割り当てられているカテゴリ数が表示されます。

カテゴリの関連性

より良いカテゴリを作成するため、 各カテゴリのドキュメントまたはレコードの関連性のほか、ドキュメントまたはレコードが属するすべてのカテゴリの関連性を確認できます。

カテゴリのレコードに対する関連性

［データ］パネルにドキュメントまたはレコードが表示されると、それらすべてのカテゴリが［カテゴリ］列に表示されます。 ドキュメントまたはレコードが複数のカテゴリに含まれる場合、この列のカテゴリは、関連性が最も大きなものから小さなものの順に表示されます。 最初に表示されたカテゴリは、このドキュメントまたはレコードに最も対応すると考えられます。 詳細は、 p. 190［データ］パネル を参照してください。

レコードのカテゴリに対する関連性

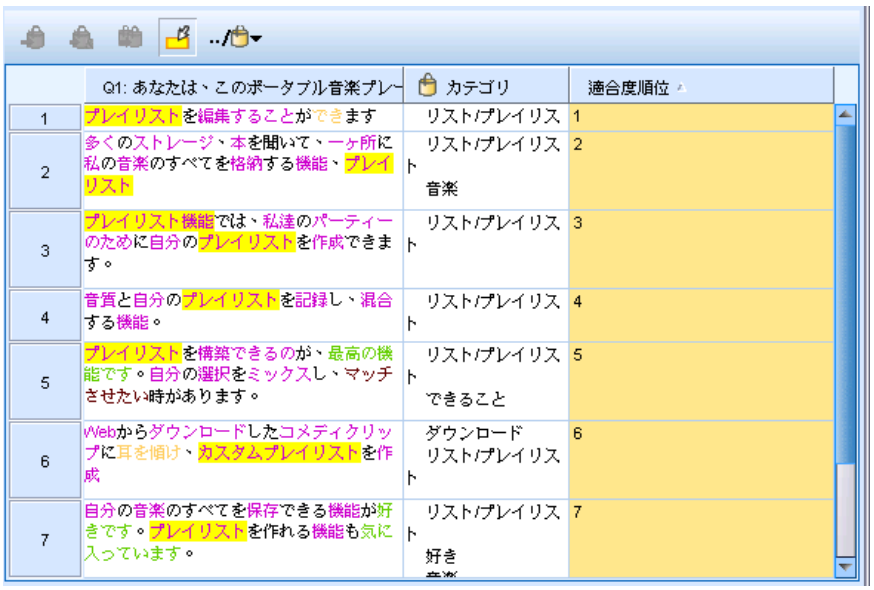
カテゴリを選択すると、［データ］パネルの［適合度順位］に各レコードの関連性が表示されます。 この適合度順位は、ドキュメントまたはレコードが選択したカテゴリにどれだけ適合しているかを、そのカテゴリのほかのレコードと比較して示します。 単一のカテゴリのレコードの順位を確認す

るには、左上の［カテゴリ］パネルでそのカテゴリを選択します。ドキュメントまたはレコードの順位が列に表示されます。 デフォルトでは、この列は表示されませんが、列が表示されるよう選択することができます。 詳細は、 p. 190［データ］パネル を参照してください。

レコードの順位が低いほど、このレコードの、選択したカテゴリに対する適合度または関連性が大きくなり、1 が最も適合度が高くなります。 複数のレコードの関連性が同じ場合、それぞれが同じ順位で表示され、その後に関連性が同じであることを示す等号 (=) が追加されます。 たとえば、1=、1=、3、4 などのようになります。このカテゴリに最もマッチするレコードが 2 つあることを意味します。

ヒント: 最も関連性の高いレコードのテキストをカテゴリの注釈に追加して、カテゴリの説明をより分かりやすくすることができます。 テキストを選択し、メニューから **[カテゴリ] > [注釈に追加]** を選択して、［データ］パネルからテキストを直接追加します。

図 10-5
カテゴリおよび適合度順位を示す [データ] パネル

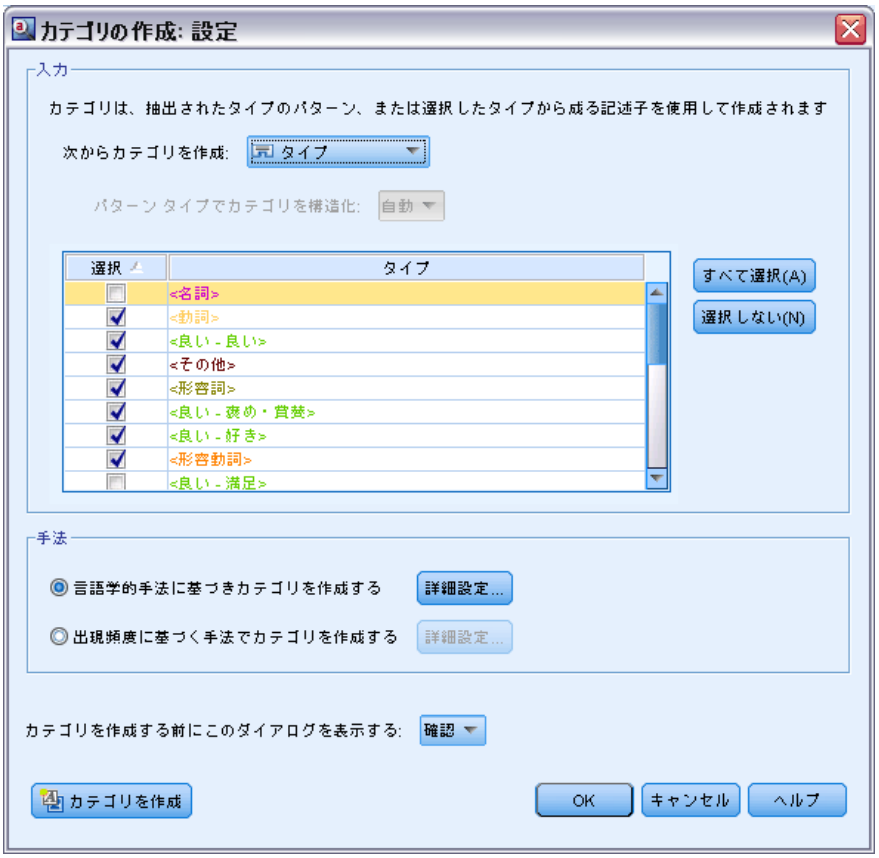


	Q1: あなたは、このポータブル音楽プレ	カテゴリ	適合度順位
1	プレイリストを編集することができます	リスト/プレイリスト	1
2	多くのストレージ、本を聞いて、一ヶ所に私の音楽のすべてを格納する機能、プレイリスト	リスト/プレイリスト 音楽	2
3	プレイリスト機能では、私達のパーティーのために自分のプレイリストを作成できます。	リスト/プレイリスト	3
4	音質と自分のプレイリストを記録し、混合する機能。	リスト/プレイリスト	4
5	プレイリストを構築できるのが、最高の機能です。自分の選択をミックスし、マッチさせたい時があります。	リスト/プレイリスト できること	5
6	Webからダウンロードしたコメディクリップに耳を傾け、カスタムプレイリストを作成	ダウンロード リスト/プレイリスト	6
7	自分の音楽のすべてを保存できる機能が好きです。プレイリストを作れる機能も気に入っています。	リスト/プレイリスト 好き	7

カテゴリの作成

テキスト分析パッケージのカテゴリがある場合でも、さまざまな言語学的手法および出現頻度に基づく手法を使用して、カテゴリを自動的に作成することもできます。**[カテゴリ作成設定]** ダイアログを使用して、自動的な言語的手法および出現頻度に基づく手法を適用し、コンセプトまたはコンセプト パターンのいずれかよりカテゴリを作成することができます。

図 10-6
[カテゴリを作成] ダイアログ ボックス



一般的に、カテゴリはさまざまな記述子（タイプ、コンセプト、TLA パターン、カテゴリ規則）で構成されます。自動化されたカテゴリ作成方法を使用してカテゴリを作成する場合、作成されたカテゴリは、選択した入力に応じてコンセプトまたはコンセプトのパターンから命名され、それぞれ一連の記述子を使用します。これらの記述子は、カテゴリ規則またはコンセプトの形式で、手法によって発見されたすべての関連コンセプトを含む場合があります。

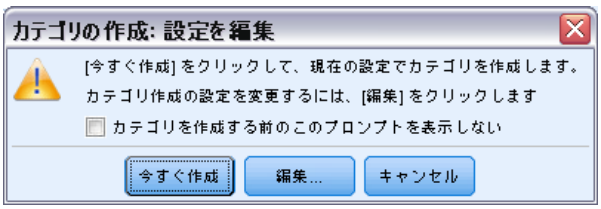
カテゴリを作成した後、[カテゴリ] パネルで確認またはグラフや図表を使用して検討することにより、カテゴリについて多くのことを学ぶことができます。手作業で若干の調整を行い、間違った分類を除去したり、把握されなかったレコードや語を追加することもできます。手法を適用した後、カテゴリにグループ化したコンセプト、種類、パターンは他の手法にも使用できます。また、さまざまな手法を使用すると、重複したカテゴリまたは不適切なカテゴリを作成する場合もあるため、カテゴリを結合または削除することもできます。 詳細は、 p. 253 カテゴリの編集および調整 を参照してください。

重要: 以前のリリースでは、共起規則および類義語規則は、大カッコで囲まれていました。このリリースの場合、大カッコはテキスト リンク分析パターン結果を示します。 その代わり、共起規則や類義語規則は、(**スピーカー システム | スピーカー**) のように、カッコで囲まれます。

カテゴリを作成するには

▶ メニューから、[**カテゴリ**] → [**カテゴリを作成**] を選択します。プロンプトが表示されないよう設定している場合、メッセージ ボックスが表示されます。

図 10-7
作成する前のプロンプト



▶ 今すぐ作成するか、左記に設定を編集するかを選択します。

- [**今すぐ作成**] をクリックすると、現在の設定でカテゴリの作成が開始されます。デフォルトで選択されている設定で、十分カテゴリ化プロセスを開始できます。カテゴリ作成のプロセスが開始し、進捗状況のダイアログが表示されます。
- [**編集**] をクリックして、ビルド設定を確認し、変更します。

注:最大で表示可能なカテゴリの数は10,000です。この数字に到達するか超えると警告が表示されます。これが発生した場合、ビルドまたはカテゴリ展開オプションを変更し、作成されたカテゴリの数を減らしてください。

入力

カテゴリは、タイプ パターンまたはタイプのいずれかより派生した記述子から作成されます。 表内で、カテゴリ作成プロセスに使用する各タイプまたはパターンを選択できます。

タイプ パターン :タイプ パターンを選択すると、タイプおよびコンセプトではなくパターンからカテゴリが独自に作成されます。このように、選択したタイプ パターンに属するコンセプト パターンを含むレコードまたはドキュメントがカテゴリされます。そのため、表で `<Budget>` タイプ パターンおよび `<Positive>` タイプ パターンを選択した場合、`cost & <Positive>` または `&Pos:rates & excellent` などのカテゴリを作成することができます。

図 10-8
使用できるタイプ パターンを示す [カテゴリを作成] ダイアログ

自動カテゴリ作成でタイプ パターンを入力として使用すると、その手法によってカテゴリ構造を形成する複数の方法が特定される場合があります。技術的には、カテゴリを作成する適切な方法はありませんが、分析により適した構造がある場合があります。この場合の出力をカスタマイズするために、タイプを優先的に指定できます。作成されたすべての上位レベルのカテゴリは、ここで選択したタイプのコンセプトのみに由来します。すべてのサブカテゴリには、このタイプのテキスト リンク パターンが含まれています。このタイプをパターン タイプ：フィールドにより **構造カテゴリで選択します**すると、テーブルが更新され、選択されたタイプを含む適用パターンのみを表示します。多くの場合、<Unknown> が事前に選択されています。この結果、タイプ <Unknown> を含むすべてのパターン（日本語以外のテキスト）が選択されます。日本語テキストの場合、<**名詞**> がプログラムによって事前に選択されます。注:日本語テキスト展開はIBM® SPSS® Modeler Premiumで利用可能です。表には、タイプ、レコード数またはドキュメント数（Doc. の数）の最も多いものから降順で表示されます。

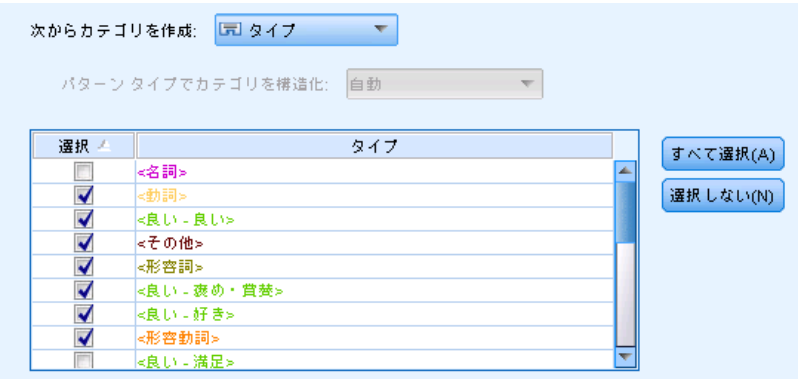
タイプ ：タイプを選択すると、カテゴリは選択したタイプに属するコンセプトから作成されます。そのため、表で <Budget> タイプを選択した場合、cost および price は <Budget> に割り当てられたこんせぷとであるため、cost または price のようなカテゴリを作成できます。

デフォルトでは、最も多いレコードまたはドキュメントをキャプチャするタイプのみが選択されます。このように事前選択すると、最も関心の高いタイプにすばやく焦点をあて、関心の低いカテゴリが作成されないようにすることができます。表には、タイプ、レコード数またはドキュメント数（Doc. の数）の最も多いものから降順で表示されます。**意見**ライブラリのタイプは、デフォルトではタイプ テーブルで選択されていません。

入力の選択は、取得するカテゴリに影響します。入力としてタイプを選んだ場合は、明確に関連付けられたコンセプトをより簡単に見ることができます。たとえば、タイプを入力として使用してカテゴリを作成する場合、**果物** というカテゴリを、**リンゴ**、**梨**、**柑橘類**、**オレンジ**などのコンセプトと

ともに取得できます。他方、タイプ パターンを入力として選び、パターンとしてたとえば <**不明**> + <**肯定的**> を選択した場合には、**果物** + <**肯定的**> というカテゴリーに、**果物** + **おいしい** や **リンゴ** + **良い**などの 1、2 種類の果物を伴ったものを取得することになるでしょう。この第 2 の結果は、2 つのコンセプト パターンを示すのみです。それは、他の果物の出現が必ずしも肯定的に評価されるものとは限らないからです。また、現在手元にあるテキスト データについてはこれで十分であったとしても、異なるドキュメント セットを使用する経時的な研究においては、**柑橘類** + **肯定的**のような別の記述子を手動で追加したり、タイプを使いたいと考えることもあるでしょう。タイプだけを入力として使用することは、すべての可能な果物を見つけ出すのに役に立ちます。

図 10-9
使用できるタイプを示す [カテゴリを作成] ダイアログ



手法

すべてのデータセットが一意であるため、手法の数やそれらを適用する順序は、時間によって変わる場合があります。テキスト マイニングの目標が、データセットによって異なる場合があるため、それぞれの手法を検証して、指定したテキスト データにとってどの手法が最良の結果を生み出すかを確認する必要があります。

これらの設定について詳しく知らなくても、使用することができます。デフォルトでは、最も一般的で平均的な設定がすでに選択されています。そのため、高度な設定のダイアログを省略して、カテゴリをすぐに作成することができます。同様に、ここで変更を行うと、最新の設定が常に保持されるため、設定ごとに設定ダイアログに戻る必要はありません。

言語的手法または出現頻度に基づく手法のいずれかを選択し、[詳細設定] ボタンをクリックして、選択した手法の設定を表示します。 自動的手法では、データを完全にカテゴリ化できません。そのため、データに合った 1 つまたは複数の自動的手法を見つけ、適用することをお勧めします。 言語的手法および出現頻度に基づく手法を同時に使用して作成することはできません。

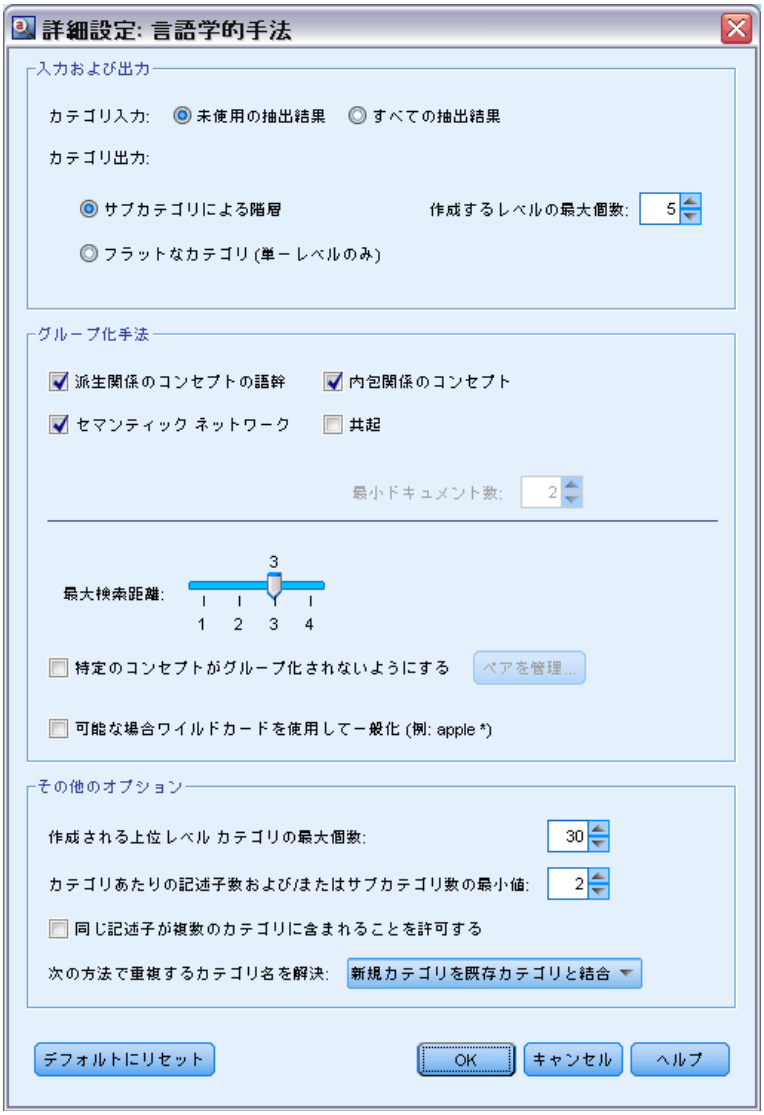
- **詳細言語法**:詳細は、 p. 198 を参照してください。
- **高度な出現頻度に基づく手法**:詳細は、 p. 210 を参照してください。

詳細言語設定

カテゴリを作成する場合、「派生関係のコンセプトの語幹」（日本語には使用不可)、「内包関係のコンセプト」、 「セマンティック ネットワーク」（英語テキストのみ)、および「共起規則」など、さまざまな詳細言語カテゴリ作成手法から選択することができます。これらの手法は個別に、またはそれぞれを組み合わせてカテゴリを作成することができます。

すべてのデータセットが一意であるため、手法の数やそれらを適用する順序は、時間によって変わる場合がありますので、注意してください。テキスト マイニングの目標が、データセットによって異なる場合があるため、それぞれの手法を検証して、指定したテキスト データにとってどの手法が最良の結果を生み出すかを確認する必要があります。 自動的手法では、データを完全にカテゴリ化できません。そのため、データに合った 1 つまたは複数の自動的手法を見つけ、適用することをお勧めします。

図 10-10
詳細設定:カテゴリ作成のための言語ダイアログボックス



入力および出力

カテゴリ入力: カテゴリが作成される内容を選択します。

- **未使用の抽出結果**:既存のカテゴリで使用されていない抽出結果からカテゴリを作成できます。レコードが、複数のカテゴリと合致する傾向が最も小さくなり、作成されるカテゴリの数が制限されます。
- **すべての抽出結果**:抽出結果のいずれを使用してもカテゴリを作成できます。カテゴリがないまたは少ない場合に最も役立ちます。

カテゴリ出力: 作成されるカテゴリの一般的な構造を選択します。

- **サブカテゴリによる階層**:サブカテゴリおよびサブサブカテゴリの作成を有効にします。作成できる最大数のレベル（[**作成するレベルの最大個数**] フィールド）を選択して、カテゴリの深度を設定できます。3 を選択すると、カテゴリ内にサブカテゴリを作成でき、またこれらのサブカテゴリ内にもサブカテゴリを作成できます。
- **フラットなカテゴリ（単一レベルのみ）**:1 レベルのみのカテゴリを作成できます。つまり、サブカテゴリは生成できません。

グループ化手法

使用できるそれぞれの手法は、特定の種類のデータおよび状況に適していますが、同じ分析で手法を組み合わせて、全範囲のドキュメントまたはレコードをキャプチャすると役に立つ場合があります。 複数のカテゴリのコンセプトを表示したり、重複するカテゴリを見つけることができます。

派生関係のコンセプトの語幹: コンセプト コンポーネントが形態的に関連するか、または語幹を共有するかどうかを分析するとき、コンセプトを取得し、そのコンセプトに関連するその他のコンセプトを検索することによって、カテゴリを作成します。 この手法は、生成された各カテゴリのコンセプトが類義語または意味の上で密接に関連しているため、類義語の複合語コンセプトを特定するのに非常に役立ちます。 長さの異なるデータを処理し、コンパクトなカテゴリをより少なく生成します。 たとえば、コンセプト `opportunities to advance` は、コンセプト `opportunity for advancement` および `advancement opportunity` とグループ化されます。詳細は、 p. 204 派生関係のコンセプトの語幹 を参照してください。このオプションは、日本語テキストに対しては使用できません。

セマンティック ネットワーク: 各コンセプトの考えられえる意味を、単語の関係の拡張インデックスから特定することによって開始し、関連するコンセプトをグループ化することによってカテゴリを作成します。 この手法は、コンセプトがセマンティック ネットワークに認識され、あまり曖昧でない場合に最も適しています。 テキストに、ネットワークが認識していない特殊な用語または専門用語が含まれている場合はあまり役に立ちません。 たとえば、コンセプト `granny smith apple` は、`granny smith` と横の関係があるため、`gala apple` および `winesap apple` とグループ化されます。 またあるいは、コンセプト `animal`（**動物**）は、その下位語である `cat`（**ネコ**）および`kangaroo`（**カンガルー**）とグループ化されます。 このリリースでは、英語テキストにのみ使用できます。 詳細は、 p. 207 セマンティック ネットワーク を参照してください。

内包関係のコンセプト: この手法では、一方の共通の文字列である単語を含むかどうかに基づき、マルチタームのコンセプト（複合語）をグループ化することによってカテゴリを作成します。 たとえば、コンセプト `seat`（**シート**）は、コンセプト `safety seat`（**セーフティ シート**）、`seat`

belt（**シート ベルト**）、および seat belt buckle（**シート ベルトのバックル**）とグループ化されます。 詳細は、 p. 206 内包関係のコンセプト を参照してください。

共起: この手法では、テキスト内の共起関係のコンセプトからカテゴリを作成します。 ドキュメントおよびレコードでコンセプトまたはコンセプトパターンがいっしょに出現することが多いとき、共起関係のコンセプトはおそらくカテゴリ定義の値のものである基底となる関連を反映します。 単語が頻繁に共起する場合、共起規則が作成され、新しいサブカテゴリのカテゴリの記述子として使用できます。 たとえば、多くのレコードに単語 price（**価格**）および availability（**有効性**）が含まれている場合（ただし、一方を含み、もう一方を含まないレコードはほとんどない）、これらのコンセプトを共起規則（price & available）にグループ化し、たとえばカテゴリ price のサブカテゴリに割り当てることができます。詳細は、 p. 209 共起規則 を参照してください。

- **最小ドキュメント数:**共起関係のコンセプトの重要性を判断できるようにするため、カテゴリの記述子として使用されるよう、指定の共起関係のコンセプトを含む必要のあるレコードまたはドキュメントの最小数を定義します。

最大検索距離: カテゴリ作成前の手法による検索の距離を選択します。値が小さいほど、取得する結果は少なくなります。ただし、これらの結果はノイズが少なく、またリンクや関連性が大きくなります。値が大きいほど、取得する結果は多くなります。ただし、これらの結果の信頼性または関連性が弱くなります。このオプションはすべての手法にグローバルに適用されますが、共起およびセマンティック ネットワークに対する効果は最も大きくなります。

特定のコンセプトがグループ化されないようにする: 出力の 2 つのコンセプトがグループ化またはペアとならないように処理を停止します。コンセプト ペアを作成または管理するには、[**ペアを管理**] をクリックします。 詳細は、 p. 202 例外ペアのリンクの管理 を参照してください。

可能な場合ワイルドカードを使用して一般化: アスタリスク（*）のワイルドカードを使用して、製品が一般的な規則をカテゴリに生成することができるようになります。たとえば、[**アップル タルト** + .] や [**アップル ソース** + .] などの複数の記述子を作成する代わりに、[**アップル** * + .] のようにワイルドカードを使用します。ワイルドカードを使用して一般化すると、以前と同じように、ちょうど同じ数のレコードまたはドキュメントを取得する場合が多くなります。ただし、このオプションには、数の縮小やカテゴリの記述子の簡略化という利点があります。また、このオプションを使用すると、新しいテキスト データ（例: 長期的/周期的研究）にこれらのカテゴリを使用してより多くのレコードまたはドキュメントをカテゴリ化する機能を拡大します。

カテゴリを作成するその他のオプション

適用するグループ化手法を選択するほか、次のように、その他の作成オプションを編集することができます。

最大数の上位レベル**カテゴリがさくせいされました。**このオプションを使用て、[カテゴリを作成] をクリックすると作成できるカテゴリ数を制限します。この値を高く設定し、関心の低いカテゴリを削除すると、よりよい結果が生成される場合があります。

カテゴリあたりの記述子数またはサブカテゴリ数の最小値: カテゴリが作成するために含む必要のある記述子数およびサブカテゴリ数の最小値を定義します。多くのレコードまたはドキュメントをキャプチャしないカテゴリの作成が制限されます。

同じ記述子が複数のカテゴリに含まれることを許可する: このオプションを選択すると、記述子を次の作成される複数のカテゴリに使用できるようにします。項目が一般的にまたは「自然に」 2 つ以上のカテゴリになり、より良い品質のカテゴリを作成するため、このオプションが一般的に選択されます。このオプションを選択しない場合、複数のカテゴリのレコードの重複が少なくなり、データのタイプによっては、これが望ましい設定となります。ただし、多くのデータ タイプでは、記述子を 1 つのカテゴリに制限すると、品質またはカテゴリの範囲が損なわれます。たとえば、`car seat manufacturer` というコンセプトがあったとします。このオプションを指定すると、このコンセプトは、テキスト `car seat` に基づいてあるカテゴリに、また `manufacturer` というテキストに基づいて別のカテゴリに使用されます。ただし、このオプションが選択されていない場合、2 つのカテゴリを取得できますが、コンセプト `car seat manufacturer` は、`car seat` および `manufacturer` がそれぞれ出現するレコード数など、いくつかの要素に基づいて、最も一致するカテゴリにのみ、記述子として使用されます。

次の方法で重複するカテゴリ名を解決: 名前が既存のカテゴリと同じ新規カテゴリまたはサブカテゴリの処理方法を選択します。新規カテゴリ（およびその記述子）と同じ名前を持つ既存カテゴリとを結合できます。あるいは、既存のカテゴリに重複する名前があった場合、カテゴリの作成をスキップすることもできます。

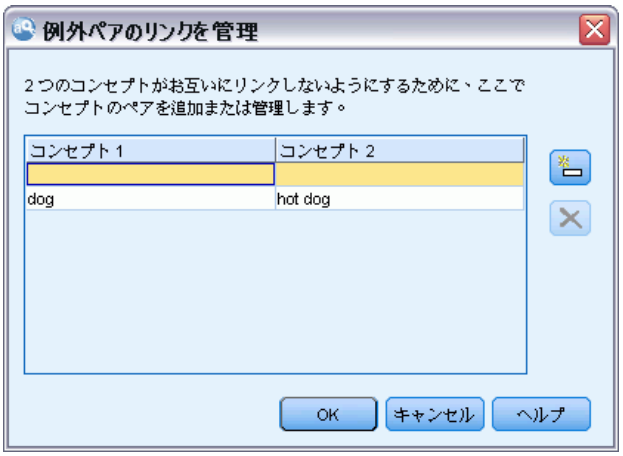
例外ペアのリンクの管理

カテゴリ作成、クラスタリングおよびコンセプト マッピングの間、既知の関連性によって内部アルゴリズムが語をグループ化します。2 つのコンセプトが対応しないように、またはお互いにリンクしないようにするためには、[**カテゴリ作成詳細設定**] ダイアログ、[**クラスタを作成**] ダイアログ、および [**コンセプト マップ インデックス設定**] ダイアログでこの機能をオンにして、[**ペアを管理**] ボタンをクリックします。

表示される [**リンクの例外を管理**] ダイアログでコンセプト ペアを追加、編集または削除できます。1 行につき 1 つのペアを入力します。ここでペアを入力すると、カテゴリ作成または拡張時のグループ化、クラスタリング、コンセプト マッピングが行われなくなります。必要に応じて、単語を正確に入力します。たとえば、単語のアクセント付きバージョンがアクセントのないバージョンとは同じではありません。

たとえば、hot dog および dog がグループ化されていないことを確認したい場合、ペアをテーブル内の各行に追加できます。

図 10-11
[例外ペアのリンクの管理] ダイアログ



言語学的手法について

カテゴリを作成または拡張する場合、「派生関係のコンセプトの語幹」（日本語には使用不可）、「内包関係のコンセプト」、「セマンティック ネットワーク」（英語テキストのみ）、および「共起規則」など、さまざまな詳細言語カテゴリ作成手法から選択することができます。これらの手法は個別に、またはそれぞれを組み合わせてカテゴリを作成することができます。

これらの設定について詳しく知らなくても、使用することができます。デフォルトでは、最も一般的で平均的な設定がすでに選択されています。必要に応じて、高度な設定のダイアログを省略して、カテゴリをすぐに作成または拡張することができます。同様に、ここで変更を行うと、最後に使用した設定が記憶されているため、設定ごとに設定ダイアログに戻る必要はありません。

ただし、すべてのデータセットが一意であるため、手法の数やそれらを適用する順序は、時間によって変わる場合がありますので、注意してください。テキスト マイニングの目標が、データセットによって異なる場合があるため、それぞれの手法を検証して、指定したテキスト データにとってどの手法が最良の結果を生み出すかを確認する必要があります。 自動的手

法では、データを完全にカテゴリ化できません。そのため、データに合った 1 つまたは複数の自動的手法を見つけ、適用することをお勧めします。

カテゴリを作成する、主な自動化言語法は、次のとおりです。

- **派生関係のコンセプトの語幹:** コンセプト コンポーネントが形態的に関連するかどうかを分析するとき、コンセプトを取得し、そのコンセプトに関連するその他のコンセプトを検索することによって、カテゴリを作成します。詳細は、 p. 204 派生関係のコンセプトの語幹 を参照してください。このオプションは、日本語テキストに対しては使用できません。
- **ないほう関係のコンセプト:** コンセプトを取得し、そのコンセプトを含むその他のコンセプトを見つけることによって、カテゴリを作成します。詳細は、 p. 206 内包関係のコンセプト を参照してください。
- **セマンティック ネットワーク:** 各コンセプトの考えられえる意味を、単語の関係の拡張インデックスから特定することによって開始し、関連するコンセプトをグループ化することによってカテゴリを作成します。詳細は、 p. 207 セマンティック ネットワーク を参照してください。このオプションは、英語テキストにのみ使用できます。
- **共起:** 新しいカテゴリを作成、カテゴリを拡張するために、または別のカテゴリ手法の入力として使用できる共起規則を作成します。 詳細は、 p. 209 共起規則 を参照してください。

派生関係のコンセプトの語幹

注：この手法は、日本語テキストに対しては使用できません。

派生関係のコンセプトの語幹による手法は、コンセプト コンポーネントが形態的に関連するかどうかを分析するとき、コンセプトを取得し、そのコンセプトに関連するその他のコンセプトを検索することによって、カテゴリを作成します。コンポーネントは単語です。コンセプトの各コンポーネントの末尾（接尾辞）を確認し、それらから派生するその他のコンセプトを見つけることによって、コンセプトのグループ化を試みます。単語がお互いに派生している場合、共有するか、意味の上で近い傾向にあります。末尾を特定するために、内部の言語固有の規則が使用されます。たとえば、コンセプト `opportunities to advance` は、コンセプト `opportunity for advancement` および `advancement opportunity` とグループ化されます。

いかなる種類のテキストにも派生関係のコンセプトの語幹を使用できます。それ自体によって作成されるカテゴリはごくわずかであり、各カテゴリに含まれるコンセプトも少数です。各カテゴリのコンセプトは、類義語または状況的に関連しています。手動でカテゴリを作成する場合でも、このアルゴリズムを作成すると役立ちます。見つかった類義語は、特に関心のあるコンセプトの類義語である場合があります。

注:それらを明示的に指定して、コンセプトがグループ化されないようにすることができます。 詳細は、 p. 202 例外ペアのリンクの管理 を参照してください。

キーワードのコンポーネント化および活用の無効化

派生関係のコンセプトの語幹による手法または内包関係のコンセプトの手法が適用されると、キーワードはまずコンポーネント（単語）に分割され、コンポーネントの活用が無効化されます。手法が適用されると、コンセプトおよびそれらに関連したキーワードが読み込まれ、スパース、ハイフン、アポストロフィなどの区切り文字に基づいて、コンポーネントに分割されます。たとえば、system administrator というキーワードは、{administrator, system} のように、コンポーネントに分割されます。

ただし、元のキーワードの一部は使用できず、ストップ ワードとして参照されます。英語の場合、こうした無視できるコンポーネントには、a、and、as、by、for、from、in、of、on、or、the、to、および with などがあります。

たとえば、キーワード examination of the data のコンポーネントは {data, examination} のようになり、of および the は無視できるとみなされます。また、コンポーネット セットでは、コンポーネントの順序は意味を持ちません。これにより、次の 3 つのキーワードは同等とすることができます。cough relief for child、child relief from a cough、および relief of child cough。これらはすべて、{child, cough, relief} という同じコンポーネント セットです。キーワードのペアが同等のものとして特定されるごとに、対応するコンセプトを結合して、すべてのキーワードを参照する新しいコンセプトを形成します。

また、キーワードのコンポーネントは活用している場合があるため、言語固有の規則が内部的に適用され、複数形など、活用した変異形にかかわらず、同等のキーワードを特定します。このようにして、活用のない単数形が level であるため、キーワード level of support および support levels を同等のものとして特定することができます。

派生関係のコンセプトの語幹の機能

キーワードがコンポーネント化され、活用がなくなった場合（前セクションを参照）、派生関係のコンセプトの語幹アルゴリズムがコンポーネントの末尾、または接尾辞を分析し、コンポーネントの語幹を検索して、そのコンセプトを、同じ、または類似したルートを持つ他のコンセプトとグループ化します。末尾は、テキスト言語に特有の言語派生規則を使用して特定されます。たとえば、接尾辞 ical で終わるコンセプト コンポーネントは、同じ語源を持ち、接尾辞 ic で終わるコンセプトから派生するという、英語のテキストの派生規則があるとします。この規則（および活用の無効化）を使用すると、アルゴリズムはコンセプト epidemiologic study および epidemiological studies をグループ化できます。

キーワードはすでにコンポーネント化され、(in や of などの) 無視できるコンポーネントが特定されているため、派生関係のコンセプトの語幹アルゴリズムは、コンセプト studies in epidemiology を epidemiological studies とグループ化することもできます。

一連のコンポーネント派生関係の規則は、このアルゴリズムでグループ化されるコンセプトの大半が同義語となるように選ばれました。たとえば、epidemiologic studies、epidemiological studies、 studies in epidemiologyという 3 つのコンセプトはすべて同義のキーワードです。完全性を高くするために、一部の派生規則を使用すると、アルゴリズムによって、状況的に関連するコンセプトをグループ化できます。たとえば、アルゴリズムは empire builder や empire building などのコンセプトをグループ化できます。

内包関係のコンセプト

内包関係のコンセプトの手法は、コンセプトを取得し、語彙系列のアルゴリズムを使用してカテゴリを作成し、その他のコンセプトに含まれるコンセプトを特定します。コンセプト内の単語が別のコンセプトの部分集合である場合、規定となるセマンティックの関係を反映します。内包関係のコンセプトは、どんな種類のテキストにも使用できる強力な手法です。

この方法は、セマンティック ネットワークと組み合わせるとうまく動作しますが、個別に使用することもできます。ドキュメントまたはレコードにドメイン固有の用語または専門用語が多く含まれている場合、内包関係のコンセプトを使用するとより良い結果が出ます。これは事前に類義語辞書を使用して、特別なキーワードが適切に抽出・グループ化されるように調整してある場合に、特にいい結果が得られます。

内包関係のコンセプトの機能

内包関係のコンセプト アルゴリズムを適用する前に、キーワードはコンポーネント化され、活用がなくなります。詳細は、 p. 204 派生関係のコンセプトの語幹 を参照してください。 次に、内包関係のコンセプト アルゴリズムはコンポーネント セットを分析します。各コンポーネント セットについて、アルゴリズムは最初のコンポーネント セットの部分集合である別のコンポーネント セットを検索します。

たとえば、コンポーネント セットが {breakfast, continental} のコンセプト continental breakfast があり、コンポーネント セットが {breakfast} のコンセプト breakfast がある場合、アルゴリズムは、continental breakfast は breakfast の一種であると結論付け、これらをともにグループ化します。

より大きな例では、[抽出結果] パネルにコンセプト seat が表示され、このアルゴリズムを適用する場合、safety seat、leather seat、seat belt、seat belt buckle、infant seat carrier、および car seat laws のようなコンセプトは該当するカテゴリにグループ化されます。

キーワードはすでにコンポーネント化され、(in や of などの) 無視できるコンポーネントが特定されているため、内包関係のコンセプト アルゴリズムは、コンセプト advanced spanish course にコンセプト course in spanish が含まれると認識します。

注:それらを明示的に指定して、コンセプトがグループ化されないようにすることができます。 詳細は、 p. 202 例外ペアのリンクの管理 を参照してください。

セマンティック ネットワーク

このリリースでは、セマンティック ネットワーク手法は、英語テキストにのみ使用できます。

この手法では、単語の関係の組み込みネットワークを使用してカテゴリを作成します。このため、キーワードが具体的で、あまりあいまいでなければ、この手法を使用すると、非常に良い結果を生成することができます。ただし、この手法が非常に技術的/専門的なコンセプト間に多くのつながりを見つけることを期待することができません。こうしたコンセプトを処理する場合、内包関係のコンセプトおよび派生関係のコンセプトの語幹による手法がより有用な場合があります。

セマンティック ネットワークの機能

セマンティック ネットワーク手法は、既知の単語の関係を利用して、類義語または下位語のカテゴリを作成します。**下位語**は、1 つのコンセプトがある種の 2 番目のコンセプトである場合、階層の関係性があり、ISA リレーションシップとも呼ばれます。たとえば、animal がコンセプトである場合、動物の種類である cat、kangaroo は animal の下位語となります。

類義語および下位語の関係性のほか、セマンティック ネットワークの手法では、<Location> タイプからコンセプト間の部分的なリンクおよび全体のリンクを検証します。たとえば、この手法ではコンセプト normandy、provence、および france を、ノルマンディおよびプロバンスは、フランスの一部であるため、1 つのカテゴリにグループ化します。

セマンティック ネットワークは、セマンティック ネットワークの各コンセプトの考えられる意味を特定することから始めます。コンセプトが類義語または下位語として特定されると、1 つのカテゴリにグループ化されます。たとえば、この手法をつかうと、次の 3 つのコンセプトからなる 1 つのカテゴリを作成されます。**生食用リンゴ**、**デザートのリンゴ**、および**グラニー スミス**。なぜならば、セマンティック ネットワークには次のような情報が含まれるからです。1) **デザートのリンゴ**は**生食用リンゴ**

の類義語であり、 2）**グラニー スミス**は**生食用リンゴ** の一種である（**生食用リンゴの下位語という意味で**）。

個別にみると、多くのコンセプト、特にユニタームがあいまいです。たとえば、コンセプト buffet は食事の種類、あるいは家具を表す場合があります。一連のコンセプトに meal、furniture、および buffet がある場合、アルゴリズムは meal または furniture のいずれかによる buffet のグループ化を選択するよう強制します。アルゴリズムによる選択は、レコードまたはドキュメントのコンテキストにおいては適切でない場合があります。

セマンティック ネットワーク手法は、特定の種類のデータによる内包関係のコンセプトにおいて優れています。セマンティック ネットワークと内包関係のコンセプトでは、apple pie が pie の一種であることを認識しますが、tart も pie の一種であることを認識できるのはセマンティック ネットワークだけです。

セマンティック ネットワークは、他の手法を組み合わせて機能します。たとえば、セマンティック ネットワーク手法と内包関係のコンセプトの手法を選択し、セマンティック ネットワークによりコンセプト teacher をコンセプト tutor とグループ化した（tutor は teacher の一種であるため）とします。内包アルゴリズムはコンセプトを graduate tutor と tutor にグループ分けし、結果として、2つのアルゴリズムが共同してアウトプット カテゴリを作成します。アウトプットカテゴリには、tutor, graduate tutor, and teacherが含まれます。

セマンティック ネットワークのオプション

この手法では、さまざまな追加設定が重要である場合があります。

- **［最大検索距離］**を変更します。カテゴリ作成前に手法による検索の距離を選択します。値が小さいほど、作成される結果は少なくなります。ただし、これらの結果はノイズが少なく、またリンクや関連性が大きくなります。値が大きいほど、取得する結果は多くなります。ただし、これらの結果の信頼性または関連性が弱くなります。

  たとえば、距離に応じて、Danish pastry から coffee roll（上位）まで、そして bun（祖父母）および bread まで上方に検索します。

  作成されるカテゴリが大きすぎる、あるいはあまりに多くのものがグループ化されていると感じられる場合は、検索距離を短縮するれば、より小さなカテゴリを作成でき、作業がしやすくなります。

重要! 誤ったグループ化と行うと結果に大きな悪影響をおよぼす場合があるため、この手法を手法する場合は、オプション **語幹文字数が次の最小値以上のときにスペルを調整する**（［抽出］ダイアログ ボックスまたはノードの［エキスパート］タブで定義）を適用せず、Fuzzy Grouping を行うことをお勧めします。

共起規則

共起規則を使用すると、ドキュメントまたはレコードのセット内で強い関連を持つコンセプトを見つけ、グループ化することができます。ドキュメントおよびレコードでコンセプトが共に頻繁に見つかる場合、共起はおそらくカテゴリ定義の値のものである基底となる関連を反映します。この手法はカテゴリを作成、カテゴリを拡張するために、または別のカテゴリ手法の入力として使用できる共起規則を作成します。レコードのあるセット内で頻繁に同時に現れ、他のレコードのいずれにも個別にあまり現れない場合、2 つのコンセプトは強力に共起します。この手法を使用して、少なくとも数百のドキュメントまたはレコードを持つ大きなデータセットを使用して良い結果を作成することができます。

たとえば、多くのレコードに単語 price および availability が含まれている場合、これらのコンセプトを共起規則 (price & available) にグループ化できます。また、コンセプト peanut butter、jelly、sandwich が個別で現れるより頻繁に同時に現れる場合、これらはコンセプト共起規則 (peanut butter&jelly & sandwich) にグループ化されます。

重要: 以前のリリースでは、共起規則および類義語規則は、大カッコで囲まれていました。このリリースの場合、大カッコはテキスト リンク分析パターン結果を示します。 その代わり、共起規則や類義語規則は、(**スピーカー システム | スピーカー**) のように、カッコで囲まれます。

共起規則の機能

この手法では、ドキュメントまたはレコードをスキャンし、同時に現れる傾向のある 2 つ以上のコンセプトを探します。ドキュメントまたはレコードのあるセット内で頻繁に同時に現れ、他のドキュメントまたはレコードのいずれにも個別にあまり現れない場合、2 つ以上のコンセプトは強力に共起します。

共起するコンセプトが見つかった場合、カテゴリ規則が形成されます。これらの規則は、& ブール型演算子を使用してつながっている 2 つ以上のコンセプトで構成されています。これらの規則は、規則内のコンセプトのセットがドキュメントまたはレコードですべて共起する場合、自動的にドキュメントまたはレコードをカテゴリに自動的に分類するという、論理ステートメントです。

共起規則のオプション

共起規則の手法を使用している場合、作成される規則に影響を与えるいくつかの設定を調整できます。

- **[最大検索距離]** を変更します。手法による共起の検索の範囲を選択します。検索範囲を拡大すると、それぞれの共起の最低相似値が下がります。結果として、複数の共起規則が作成される場合がありますが、相似

値の低いものは多くの場合さほど重要ではありません。検索範囲を減少すると、最低相似値が上がります。結果として、より少ない共起規則が作成されますが、より重要（強力）である傾向になります。

- **最小ドキュメント数**:共起として見做される為に、一定の組み合わせのコンセプトを含むべきレコードまたは ドキュメントの最小数。このオプションを低く設定するほど共起を検索し易くなります。値を増やすと、少数の重要な共起を検索します。例として、「りんご」と「なし」というコンセプトが一緒に２つのレコードから見つかったとします（そしてどちらのコンセプトも他のレコードでは起こらないとします）。**最少数のドキュメント。**が２（デフォルト）に設定されていると共起規則手法はカテゴリ規則を作成します（りんごとなし）。値が３に増加すると、その規則は作成されません。

注:少ないデータセット（<１０００レスポンス）では、デフォルト設定では共起を見つけられない可能性があります。その場合は、検索範囲値を上げてみてください。

注:それらを明示的に指定して、コンセプトがグループ化されないようにすることができます。 詳細は、 p. 202 例外ペアのリンクの管理 を参照してください。

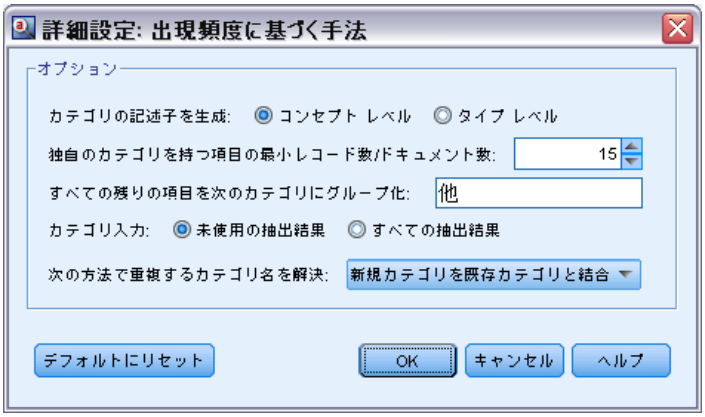
頻度の詳細設定

単純で機械的な出現頻度に基づく手法に基づいて、カテゴリを作成できます。この手法を使用して、指定されたレコードまたはドキュメントの数を超えて見つかった項目（タイプ、コンセプト、またはパターン）ごとに 1 つずつカテゴリを作成できます。また、あまり頻繁に出現しない項目すべてに 1 つカテゴリを作成できます。度数ごとに、テキスト全体の出現数の合計に対し、対象の抽出されたコンセプト（およびその類義語）、タイプ、またはパターンを含むレコードまたはドキュメントの数を参照します。

一般的または重要な回答を示すため、頻繁に出現する項目をグループ化すると、関心のある結果を生成できます。他の手法が適用されると、この手法が未使用の抽出結果に非常に役立ちます。他にカテゴリが存在しない場合、別のアプリケーションが抽出直後にこの手法を実行し、結果を編集して、関心のないカテゴリを削除し、これらのカテゴリを拡張して、より多くのレコードまたはドキュメントに一致するようにします。 詳細は、 p. 212 カテゴリの拡張 を参照してください。

この手法を使用する代わりに、[抽出結果] パネルのレコードまたはドキュメント数の多い順にコンセプトまたはコンセプト パターンを並べ替え。上位のコンセプトまたはコンセプト パターンを [カテゴリ] パネルにドラッグ アンド ドロップして、対応するカテゴリを作成することができます。

図 10-12
詳細設定:[度数分布表] ダイアログ ボックス



カテゴリの記述子を生成: 記述子の入力の種類を選択します。 詳細は、p. 193 カテゴリの作成 を参照してください。

- **コンセプト レベル**:このオプションを選択すると、コンセプトまたはコンセプト パターンの頻度が使用されます。タイプがカテゴリ作成の入力として選択されている場合はコンセプトが使用され、タイプ パターンが選択された場合はコンセプト パターンが使用されます。一般的には、この手法をコンセプト レベルに適用すると、コンセプトおよびコンセプト パターンがより低いレベルの尺度を示すため、より特定の結果を作成します。
- **タイプ レベル** :このオプションを選択すると、タイプまたはタイプ パターンの頻度が使用されます。タイプがカテゴリ作成の入力として選択されている場合はタイプが使用され、タイプ パターンが選択された場合はタイプ パターンが使用されます。この手法をタイプ レベルに適用すると、指定された情報の種類に関してすばやく表示することができます。

独自のカテゴリを持つ項目の最小ドキュメント数:このオプションを使用すると、頻繁に出現する項目からカテゴリを作成できます。出力が、少なくとも X 個のレコードまたはドキュメントで出現する記述子を含むカテゴリのみに制限されます。X は、このオプションに入力する値を示します。

すべての残りの項目を次のカテゴリにグループ化: このオプションを使用すると、あまり頻繁に出現しないすべてのコンセプトまたはタイプを、選択した名前の付いた「キャッチオール」カテゴリにグループ化します。デフォルトでは、カテゴリ名はその他です。

カテゴリ入力: 手法を適用するグループを選択します。

- **未使用の抽出結果**:既存のカテゴリで使用されていない抽出結果からカテゴリを作成できます。レコードが、複数のカテゴリと合致する傾向が最も小さくなり、作成されるカテゴリの数が制限されます。
- **すべての抽出結果**:抽出結果のいずれを使用してもカテゴリを作成できます。カテゴリがないまたは少ない場合に最も役立ちます。

次の方法で重複するカテゴリ名を解決: 名前が既存のカテゴリと同じ新規カテゴリまたはサブカテゴリの処理方法を選択します。新規カテゴリ（およびその記述子）と同じ名前を持つ既存カテゴリとを結合できます。あるいは、既存のカテゴリに重複する名前があった場合、カテゴリの作成をスキップすることもできます。

カテゴリの拡張

拡張は、記述子を自動的に追加または拡張して、既存のカテゴリを「拡大」するプロセスです。その目的は、本来カテゴリに割り当てられていなかった関連レコードまたはドキュメントをキャプチャするより良いカテゴリを作成することです。

選択した自動グループ化手法では、既存のカテゴリ記述子に関連するコンセプト、TLA パターン、およびカテゴリ規則を特定しようとします。これらの新しいコンセプト、パターン、カテゴリ規則が新しい記述子として追加されるか、既存の記述子に追加されます。拡張するためのグループ化手法には、 「派生関係のコンセプトの語幹」(日本語には使用不可)、, 「内包関係のコンセプト」、「セマンティック ネットワーク」（英語テキストのみ)、および「共起規則」が含まれます。**[カテゴリ名から生成された記述子を使用して空白のカテゴリを拡張する]** の手法を使用すると、カテゴリ名の単語を使用して記述子を生成します。そのため、カテゴリ名が記述的であるほど、結果が良いものとなります。

注:カテゴリを拡張する場合、出現頻度に基づく手法は使用できません。

拡張は、カテゴリをインタラクティブに改善する重要な方法です。次に、カテゴリを拡張する場合の例をいくつか示します。

- ［カテゴリ］パネルでコンセプト パターンをドラッグ/ドロップしてカテゴリを作成した後
- 手動でカテゴリを作成し、簡単なカテゴリ規則および記述子を追加した後
- 非常に記述的な名前を持つ 事前定義済みカテゴリ ファイルをインポートした後
- 選択した TAP に由来するカテゴリを修正した後

カテゴリを複数回使用できます。たとえば、非常に記述的な名前を持つ事前定義済みカテゴリ ファイルをインポートした場合、**[カテゴリ名から生成された記述子を使用して空白のカテゴリを拡張する]** オプションを使用して拡張

子、記述子の最初のセットを取得して、これらのカテゴリを再度拡張します。ただし、複数回拡張すると、記述子が拡張されて幅広くなると、あまりに一般的なカテゴリが生成される場合があります。作成グループ化手法および拡張グループ化手法では類似した基底のアルゴリズムを使用するため、カテゴリの作成後に直接拡張すると、より関心の高い結果の作成は期待できません。

ヒント

- 拡張を試みるが結果の使用は望まない場合、拡張を行った直後に操作をいつでも取り消す（[**編集**] > [**取り消し**]）ことができます。
- プロセス中、規則は個別に作成されるため、ドキュメントの同じセットに正確に一致するカテゴリのカテゴリ規則を 2 つ以上作成します。必要に応じて、カテゴリを確認し、カテゴリ記述子を手動で編集して重複を削除できます。 詳細は、 p. 255 カテゴリ記述子の編集 を参照してください。

カテゴリを展開するには

▶ ［カテゴリ］パネルで、展開するカテゴリを選択します。

▶ メニューから、[**カテゴリ**] → [**カテゴリを展開**] を選択します。プロンプトが表示されないようオプションを選択している場合、メッセージ ボックスが表示されます。

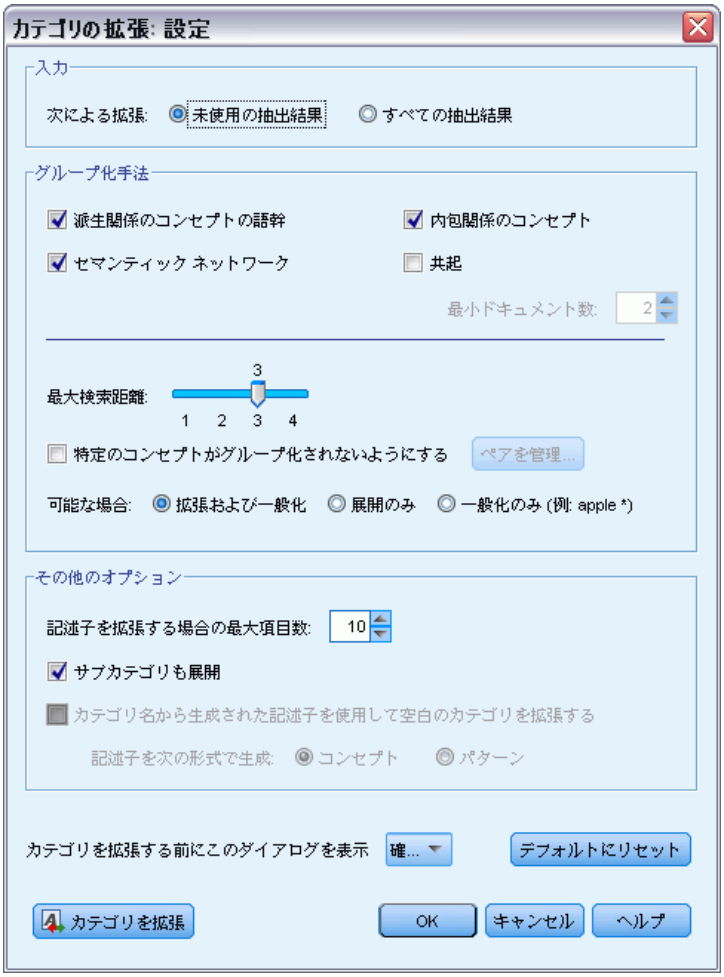
▶ 今すぐ作成するか、左記に設定を編集するかを選択します。

- [**今すぐ拡張**] をクリックすると、現在の設定でカテゴリの拡張が開始されます。プロセスが開始し、進捗状況のダイアログが表示されます。
- [**編集**] をクリックして、設定を確認し、変更します。

拡張しようとした後、新しい記述子が見つかったカテゴリには、［カテゴリ］パネルで [**展開**] という単語のフラグが立てられ、すばやくカテゴリを特定できます。［展開］というテキストは、再度展開するか、別の方法で編集、またはコンテキスト メニューを使用してこれらを解除するまで残ったままです。

注:最大で表示可能なカテゴリの数は10,000です。この数字に到達するか超えると警告が表示されます。これが発生した場合、ビルドまたはカテゴリ展開オプションを変更し、作成されたカテゴリの数を減らしてください。

図 10-13
[カテゴリを展開] ダイアログ ボックス



カテゴリの作成時または拡張時に使用できるそれぞれの手法は、特定の種類のデータおよび状況に適していますが、同じ分析で手法を組み合わせて、全範囲のドキュメントまたはレコードをキャプチャすると役に立つ場合があります。インタラクティブ ワークベンチで、カテゴリにグループ化されたコンセプトおよびタイプは、次にカテゴリを作成する場合も使用できます。つまり、複数のカテゴリのコンセプトを表示したり、重複するカテゴリを見つけることができます。

拡張。カテゴリの展開に使用する入力を選択します。

- **未使用の抽出結果**:既存のカテゴリで使用されていない抽出結果からカテゴリを作成できます。レコードが、複数のカテゴリと合致する傾向が最も小さくなり、作成されるカテゴリの数が制限されます。
- **すべての抽出結果**:抽出結果のいずれを使用してもカテゴリを作成できます。カテゴリがないまたは少ない場合に最も役立ちます。

グループ化手法

これらの手法の簡単な説明は、「詳細言語設定」(p. 198) を参照してください。これらの手法には、次のものが含まれています。

- **派生関係のコンセプトの語幹**(日本語テキストには使用不可)
- **セマンティックネットワーク**(英語テキストのみ、一般化のみオプションが選択されている場合は使用されません。)
- **内包関係のコンセプト**
- **共起** および **最小ドキュメント数**のサブオプション

これらのタイプは関連する結果を作成しないため、多くのタイプがセマンティック ネットワークから永続的に除外します。それらのタイプには、<Positive>、<Negative>, <IP>、その他の非言語的タイプなどがあります。

最大検索距離: カテゴリ作成前の手法による検索の距離を選択します。値が小さいほど、取得する結果は少なくなります。ただし、これらの結果はノイズが少なく、またリンクや関連性が大きくなります。値が大きいほど、取得する結果は多くなります。ただし、これらの結果の信頼性または関連性が弱くなります。このオプションはすべての手法にグローバルに適用されますが、共起およびセマンティック ネットワークに対する効果は最も大きくなります。

特定のコンセプトがグループ化されないようにする: 出力の 2 つのコンセプトがグループ化またはペアとならないように処理を停止します。コンセプト ペアを作成または管理するには、[**ペアを管理**] をクリックします。 詳細は、p. 202 例外ペアのリンクの管理 を参照してください。

可能な場合、 単純に拡張するか、ワイルドカードを使用して記述子を一般化するか、またはその両方を選択します。

- **拡張および一般化**: このオプションは、選択したカテゴリを拡張し、 記述子を一般化します。一般化を選択した場合、アスタリスク (*) のワイルドカードを使用して、製品が一般的なカテゴリ規則をカテゴリに作成することができるようになります。たとえば、[**アップル タルト** + .] や [**アップル ソース** + .] などの複数の記述子を作成する代わりに、[**アップル** * + .] のようにワイルドカードを使用します。ワイルドカードを使用して一般化すると、以前と同じように、ちょうど同じ数のレコードまたはドキュメントを取得する場合が多くなります。ただし、このオプションには、数の縮小やカテゴリの記述子の簡略化という

利点があります。また、このオプションを使用すると、新しいテキストデータ（例: 長期的/周期的研究）にこれらのカテゴリを使用してより多くのレコードまたはドキュメントをカテゴリ化する機能を拡大します。

- **展開のみ**:一般化せずにカテゴリを展開します。手動で作成したカテゴリには［**展開のみ**］オプションを選択し、［**拡張および一般化**］オプションを使用して同じカテゴリをもう一度展開すると便利です。
- **一般化のみ**: 別の方法でカテゴリを展開せずに、記述子を一般化します。

  注：このオプションを選択することで**セマンティックネットワーク**オプションを無効にします。これは**セマンティックネットワーク**オプションは記述を展開する時のみ使用可能である為です。

カテゴリを拡張するその他のオプション

適用するグループ化手法を選択するほか、次のように、その他のオプションを編集することができます。

記述子を拡張する場合の最大項目数: 項目（コンセプト、タイプおよびその他の式）で記述子を拡張する場合、単一の記述子に追加できる項目の最大数を定義します。この制限値を 10 に設定すると、最大 10 件の追加項目を既存の記述子に追加できます。10 件を超える項目を追加しようとする場合、10 番目の項目が追加されると、新しい項目の追加を停止します。そうすることにより、記述子のリストが短くなりますが、最も関心の高い項目が最初に使用されたことを保障するものではありません。［**可能な場合ワイルドカードを使用して一般化**］オプションを使用して、品質を落とすことなく拡張のサイズを縮小することが必要な場合があります。このオプションは、ブール値 & (AND) または! (NOT) を含む記述子にのみ適用されます。

サブカテゴリも展開: 選択したカテゴリ下のサブカテゴリも展開します。

カテゴリ名から生成された記述子を使用して空白のカテゴリを拡張する: 記述子が 0 件の、空のカテゴリにのみ適用されます。カテゴリにすでに記述子が含まれている場合、この方法では拡張されません。このオプションを選択すると、カテゴリ名を構成する単語に基づいて、各カテゴリの記述子を自動的に作成しようとします。カテゴリ名をスキャンして、名前の単語が抽出されたコンセプトに一致するかどうかを確認します。コンセプトが認識されると、そのコンセプトを使用して、合致するコンセプト パターンを検索し、コンセプトとパターンを使用してカテゴリの記述子を形成します。このオプションを選択すると、カテゴリ名が長く記述的である場合に、最良の結果を作成します。迅速にカテゴリの記述子を生成し、またカテゴリはこれらの記述子を含むレコードをキャプチャすることができます。別の場所からカテゴリをインポートしたり、長く記述的な名前を使用して手動でカテゴリを作成する場合に最も役立つオプションです。

記述子を次の形式で生成: このオプションは、先行のオプションがオンの場合のみ適用されます。

- **コンセプト:** 入力テキストから抽出されているかどうかに関係なく、記述子をコンセプトの形式で作成します。
- **パターン:** パターンが抽出されているかどうかに関係なく、記述子をパターンの形式で作成します。

手作業でのカテゴリの作成

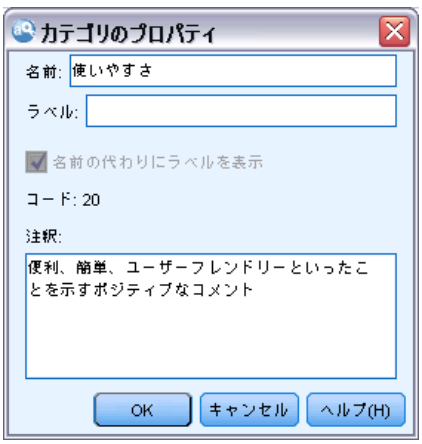
自動カテゴリ作成手法および条件規則エディタを使用してカテゴリを作成するほか、カテゴリを手動で作成することもできます。 手動で作成する方法は、次のとおりです。

- 要素を 1 つずつ追加する空のカテゴリを作成する。 詳細は、 p. 217 カテゴリの新規作成または名前の変更 を参照してください。
- キーワード、タイプ、パターンを［カテゴリ］ウィンドウにドラッグする。 詳細は、 p. 218 ドラッグ＆ドロップによるカテゴリの作成 を参照してください。

カテゴリの新規作成または名前の変更

空のカテゴリを作成して、コンセプトおよびタイプをカテゴリに追加できます。 カテゴリの名前を変更することもできます。

図 10-14
[カテゴリのプロパティ] ダイアログ



新規の空白カテゴリを作成するには

▶ ［カテゴリ］パネルに移動します。

▶ メニューから、**[カテゴリ] > [空白カテゴリを作成]** を選択します。 ダイアログ ボックスが開きます。

▶ ［名前］フィールドでそのカテゴリの名前を入力します。

▶ [OK] をクリックすると名前が適用され、ダイアログ ボックスが閉じます。 ダイアログ ボックスが閉じ、に新しいカテゴリ名がパネルに表示されます。

これでこのカテゴリに追加していくことができます。 詳細は、 p. 254 記述子のカテゴリへの追加 を参照してください。

カテゴリの名前を変更するには

▶ カテゴリを選択して、[カテゴリ] > [カテゴリ名を変更]. ダイアログ ボックスが開きます。

▶ [名前] フィールドでそのカテゴリの新しい名前を入力します。

▶ [OK] をクリックすると名前が適用され、ダイアログ ボックスが閉じます。 ダイアログ ボックスが閉じ、に新しいカテゴリ名がパネルに表示されます。

ドラッグ &ドロップによるカテゴリの作成

ドラッグ & ドロップは手作業の手法であって、アルゴリズムに基づいたものではありません。 次のように、[カテゴリ] ウィンドウでカテゴリを作成できます。

- コンセプト、タイプ、またはパターンを [抽出結果] パネルから [カテゴリ] パネルにドラッグ。
- 抽出されたコンセプトを [データ] パネルから [カテゴリ] パネルにドラッグ。
- 行全体を [データ] パネルから [カテゴリ] パネルにドラッグ。 その行に含まれる抽出されたすべてのコンセプトおよびパターンで構成されたカテゴリを作成します。

注: [抽出結果] パネルでは、複数選択をサポートし、複数要素のドラッグ アンド ドロップを可能にします。

重要! テキストから抽出されていないコンセプトを [データ] パネルからドラッグ アンド ドロップできません。 データで検出したコンセプトの抽出を強制する場合、そのコンセプトをタイプに追加する必要があります。 その後、抽出を再度実行します。 新しい抽出結果には、追加したばかりのコンセプトが含まれます。 その結果をカテゴリに使用できます。 詳細は、9 章 p. 172 コンセプトのタイプへの追加 を参照してください。

ドラッグ &ドロップでカテゴリを作成するには:

▶ [抽出結果] パネルまたは [データ] パネルから、1 つまたは複数のコンセプト、パターン、タイプ、レコード、または一部のレコードを選択します。

- マウス ボタンを押したまま、要素を既存のカテゴリまたはパネルの領域にドラッグして、新しいカテゴリを作成します。

- 要素をドロップする領域の上で、マウス ボタンを離します。 要素が［カテゴリ］パネルに追加されます。 変更が加えられたカテゴリは背景色が変わります。 この色は、[**カテゴリ フィードバック背景**]と呼ばれます。 詳細は、8 章 p. 143 オプションの設定 を参照してください。

  注: 結果として得られたカテゴリには、自動的に名前が付けられます。 詳細は、7 章p. 114 “カテゴリのプロパティの編集” を参照してください。

  どのレコードがカテゴリに割り当てられているかを確認する場合は、［カテゴリ］パネルで該当するカテゴリを選択します。［データ］パネルは自動的に更新され、そんカテゴリのすべてのレコードが表示されます。

カテゴリ規則の使用

カテゴリを作成するには、さまざまな方法があります。それらの方法の 1 つには、キーワードを表すカテゴリ規則を定義することがあります。カテゴリ規則とは、抽出したコンセプト、タイプ、パターン、およびブール型演算子を使用した論理式に基づいて、ドキュメントまたはレコードを自動的に分類するステートメントです。たとえば、「このカテゴリに、**アルゼンチン**ではなく、**大使館**という抽出したコンセプトを含む」という内容を意味する式を作成することができます。

カテゴリ規則の中には、共起 および 派生関係のコンセプトの語幹 などのグループ化手法を用いたカテゴリ作成（**カテゴリ > カテゴリ作成設定 > 詳細設定**: **言語学的手法**）時に、自動的に作成されるものもありますが、他方、データやコンテキストの独自のカテゴリ理解に従って、カテゴリ エディタを使って手動でカテゴリ規則を作成することもできます。各規則は単一のカテゴリに関連付けられるため、規則を満たすドキュメントまたはレコードがそのカテゴリにスコアリングされます。

カテゴリ規則により、より大きな特異性を持つ回答をカテゴリ化できることで、テキスト マイニングの結果およびより高度な数量分析の品質および生産性を向上させます。経験およびビジネス情報により、データおよびコンテキストに対する特定の理解を与える場合があります。このように理解することによって、情報をカテゴリに変換し、ブール型ロジックと抽出した要素を結合して、ドキュメントまたはレコードをより効率的かつ正確にカテゴリ化できます。

これらの規則を作成する機能により、ビジネス情報を製品の抽出テクノロジに重ねて、コード化の精度、効率性および生産性を拡張することができます。

注:条件規則がどのようにテキストに合致するかについての例は、「カテゴリ規則の例」（ p. 227 ）を参照してください。

カテゴリ規則シンタックス

カテゴリ規則の中には、共起 および 派生関係のコンセプトの語幹 などのグループ化手法を用いたカテゴリ作成（**カテゴリ > カテゴリ作成設定 > 詳細設定: 言語学的手法**）時に、自動的に作成されるものもありますが、他方、カテゴリエディタで手動でカテゴリ規則を作成することもできます。各規則は単一のカテゴリの説明であるため、規則を満たすドキュメントまたはレコードが自動的にそのカテゴリにスコアリングされます。

注:条件規則がどのようにテキストに合致するかについての例は、「カテゴリ規則の例」（ p. 227 ）を参照してください。

規則を作成または編集する場合、規則を条件規則エディタで開く必要があります。コンセプト、タイプ、またはパターンを追加し、またワイルドカードを使用して一致を拡張することができます。抽出されたコンセプト、タイプ、パターンを使用すると、すべての関連コンセプトを検出することができ便利です。

重要: 一般的なエラーを回避するために、コンセプトを［抽出結果］パネル、［テキスト リンク分析］パネルまたは［データ］パネルから直接条件規則エディタにドラッグアンドドロップしたり、使用できる場合はコンテキスト メニューを使用して追加することを推奨します。

コンセプト、タイプ、パターンが認識されると、テキストの隣にアイコンが表示されます。

 抽出コンセプト　  抽出タイプ　  抽出パターン

条件規則シンタックスおよび演算子

次の表には、条件規則シンタックスを定義する文字を示しています。これらの文字をコンセプト、タイプ、パターンと共に使用して、規則を作成します。

テーブル 10-2
サポートされるシンタックス

文字	説明
&	「and」ブール型演算子。たとえば、次のような a & b には、aおよびb が含まれます。 - **侵略** & **アメリカ合衆国** - 2016 & **オリンピック** - **良い** & **リンゴ**
|	「or」ブール型演算子は包含的演算子で、要素の一部またはすべてが見つかった場合に一致することを意味します。たとえば、次のような a | b には、aまたは b が含まれます。 - **攻撃** | **フランス** - **コンドミニアム** | **アパート**

文字	説明
!()	「not」ブール型演算子。たとえば、次のような !(a) には、 a が含まれません。 !(**良い** & **ホテル**)、**暗殺** & !(**オーストリア**)、または !(**金**) & !(**銅**)
*	使用方法に応じて、1 文字から単語全体にいたるまでの文字を示すワイルドカード 詳細は、 p. 225 カテゴリ規則におけるワイルドカードの使用 を参照してください。
()	式の区切り文字。カッコ内の式を最初に評価します。
+	順序特有のパターンを形成するために使用するパターン コネクタ。この演算子がある場合は、大カッコを使用する必要があります 詳細は、 p. 222 カテゴリ規則内の TLA パターンの使用 を参照してください。
[]	カテゴリ規則内部の抽出した TLA パターンに基づいて合致を検出する場合、パターン区切り文字が必要です。ブラケット内の内容は TLA パターンを参照し、単純な共起に基づいてコンセプトまたはタイプに合致しません。この TLA パターンを抽出していない場合、合致は発生しません。詳細は、 p. 222 カテゴリ規則内の TLA パターンの使用 を参照してください。 パターンではなくコンセプトとタイプの合致に着目している場合は、角カッコを使用しないでください。 注:古いバージョンの場合、カテゴリ作成手法で生成された共起規則や類義語規則は、大カッコに囲まれていました。すべての新しいバージョンの場合、大カッコは TLA パターンの存在を示します。その代わり、共起規則による手法や類義語規則による手法で作成された規則は、(**スピーカー システム**|**スピーカー**) のように、カッコで囲まれます。

& 演算子および | 演算子は、a & b = b & a および a | b = b | a のように相互的です。

バックスラッシュによる文字のエスケープ

シンタックス文字でもある文字がコンセプトに含まれている場合、その文字の前にバックスラッシュを追加して、規則が正しく解釈されるようにする必要があります。バックスラッシュ (¥) 文字を使用して、特別な意味を持つ文字をエスケープします。エディタにドラッグ アンド ドロップすると、自動的にバックスラッシュが追加されます。

次の条件規則シンタックス文字を条件規則シンタックスとしてでなく、そのまま扱う場合は、その文字の前にバックスラッシュを追加する必要があります。

&!|+<>()[]*

たとえば、コンセプト r&d に「and」演算子 (&) が使用されているため、条件規則エディタで入力する場合は、r\&d となるようにバックスラッシュが必要です。

カテゴリ規則内の TLA パターンの使用

テキスト リンク分析パターンを、カテゴリ規則で明示的に定義し、より具体的で文脈上の結果を取得することができます。カテゴリ規則でパターンを定義すると、より単純な抽出結果を省略し、抽出されたテキスト リンク分析パターン結果に基づいたドキュメントおよびレコードのみを合致させます。

重要: カテゴリ規則で TLA パターンを使用してドキュメントを合致させる場合、テキスト リンク分析を有効にして、抽出を実行する必要があります。カテゴリ規則では、そのプロセス時に検出される合致を検索します。テキスト マイニング ノードの［モデル］タブで TLA 結果の探索を選択していない場合、インタラクティブ セッションの抽出設定で TLA 抽出を有効にして、再抽出することができます。 詳細は、 9 章 p. 156 データの抽出 を参照してください。

大カッコで区切る: カテゴリ規則ではなく TLA パターンを使用している場合、TLA パターンを大カッコ［ ］で囲む必要があります。抽出した TLA パターンに基づいて合致を検出する場合、パターン区切り文字が必要です。カテゴリ規則には、タイプ、コンセプト、またはパターンが含まれるため、カッコはカッコ内の内容が抽出された TLA パターンを参照するということを明確にします。この TLA パターンを抽出していない場合、合致は発生しません。［カテゴリ］パネルに**リンゴ + 良い** のようにカッコのないパターンがあった場合、パターンがカテゴリ規則エディタ外部でカテゴリに直接追加されたことを示します。たとえば、コンセプト パターンをテキスト リンク分析ビューからカテゴリに直接追加する場合、大カッコは表示されません。ただし、規則内でパターンを使用する場合、［**バナナ + !(良い)**］のようにカテゴリ規則内の大カッコで囲む必要があります。

パターンで + 記号を使用:IBM® SPSS® Modeler Text Analytics では、最大 6 つの部分（スロット）のパターンを作成できます。順序が重要であることを示す場合、［**会社**1 + **買収** + **会社**2］のように、+ 記号を使用して、要素を接続します。ここでは、どの企業が買収するかの意味が変わってしまうため、順序が重要となります。文の構造ではなく、TLA パターン出力の構造がどのようになっているのかによって順序が決まります。たとえば、「I love Paris」というテキストがあり、このキーワードを抽出する場合、デフォルトの意見リソースは通常、意見を 2 つの部分で構成されるパターンの 2 番目の位置に配置するため、TLA パターンは [<Positive> + <Location>] ではなく、[paris + like] または [<Location> + <Positive>] となります。そのため、問題を回避するためにパターンを直接記述子として使用すると役に立つ場合があります。ただし、より複雑な表現の一部としてパターンを使用する必要がある場合、合致が検出されるかどうかにおいて、順序が大きな役割を果たすため、テキスト リンク分析ビューのパターン内の要素の順序には特に注意をしてください。

たとえば、”I like pineapple” というテキストと、 “I hate pineapple. However, I like strawberries” という 2 つのサンプル テキストがあったとします。like & pineapple という式はこの両方のテキストに一致します。それはコンセプト式であり、テキスト リンク規則ではないからです（角カッコでくくられていない）。式 pineapple + like は “I like pineapple” とのみ一致します。なぜならば、2 番目のテキストでは、like という語は今度は strawberries 関連づけられるからです。

パターンによりグループ化: 独自のパターンを使用して規則を簡略化することができます。3 つの式、cayenne peppers + like、chili peppers + like、および peppers + like の式をキャプチャするとします。これらを、[* peppers & like] のようにして、単一のカテゴリ規則にグループ化できます。hot peppers + good というもう 1 つの式がある場合、これら 4 つの式を [* peppers + <Positive>] のような規則でグループ化することができます。

パターン内の順序: 出力をよりよく構成ために、製品とともにインストールされたテンプレートに提供されているテキスト リンク分析規則が、文の語順に関係なく、同じ順序で基本パターンを出力しようとします。たとえば、テキスト「Good presentations.」を含むレコードおよび「the presentations were good」を含む別のレコードがある場合、いずれのテキストも同じ規則で合致し、出力の順序は、presentation + good や good + presentation ではなく、コンセプトパターン規則のpresentation + good と同じ順序になります。また、例のパターンのような 2 スロット パターンで、意見ライブラリのタイプに割り当てられたコンセプトは、デフォルトでは apple + bad のように出力の最後に表示されます。

テーブル 10-3
パターン シンタックスおよびブール型演算子の使用

式	ドキュメントまたはレコードが一致する条件
[]	任意の TLA パターンを含む。抽出した TLA パターンに基づいて合致を検出する場合、カテゴリ規則内でパターン区切り文字が必要です。ブラケット内の内容は、単なるコンセプトおよびタイプではなく TLA パターンを参照します。この TLA パターンを抽出していない場合、合致は発生しません。 そのため、パターンを含んでいない規則を作成する場合、!([]) を使用することができます。
[a]	パターン内の位置に関係なく、少なくとも 1 つの要素が a であるパターンを含む。たとえば、[deal] は、[deal + good] または [deal + .] に一致します。
[a + b]	コンセプト パターンを含む。たとえば、[deal + good] となります。 注: 他の要素を追加せずにこのパターンをキャプチャする場合、パターンを使用して規則を作成するのではなく、パターンを直接カテゴリに追加することをお勧めします。

式	ドキュメントまたはレコードが一致する条件
[a + b + c]	コンセプト パターンを含む。+ 記号は、一致する要素の順序が重要であることを示します。たとえば、[**会社**1 + **買収** + **会社** 2] となります。
[<A> + <B>]	最初のスロットがタイプ <A> 、2 番目のスロットがタイプ <B> のパターンを含み、ちょうど 2 つのパターンがある。+ 記号は、一致する要素の順序が重要であることを示します。たとえば、[<Budget> + <Negative>] となります。 注: 他の要素を追加せずにこのパターンをキャプチャする場合、パターンを使用して規則を作成するのではなく、パターンを直接カテゴリに追加することをお勧めします。
[<A> & <B>]	タイプ <A> および <B> のタイプ パターンを含む。たとえば、[<Budget> & <Negative>] となります。この TLA パターンは抽出されませんが、そのように記述された場合、[<Budget> + <Negative>]|[<Negative> + <Budget>] のようになります。+ 記号は、一致する要素の順序が重要であることを示します。また、他の要素がパターン内にある場合がありますが、少なくとも <Budget> and <Negative> があります。
[a + .]	a が唯一のコンセプトであるパターンを含み、そのパターンの他のスロットには何もない。たとえば、[deal + .] は、唯一の出力がコンセプト deal であるコンセプト パターンに一致します。コンセプト deal をカテゴリ記述子として追加した場合、deal を含むすべてのレコードを、deal に関して肯定的な記述を含むコンセプトとして取得します。ただし、[deal + .] を使用すると、deal を示すこれらのレコード パターン結果のみに合致し、他の関係性または意見は deal + fantastic と合致しません。 注: 他の要素を追加せずにこのパターンをキャプチャする場合、パターンを使用して規則を作成するのではなく、パターンを直接カテゴリに追加することをお勧めします。
[<A> + <>]	<A> が唯一のタイプであるパターンが含まれている。たとえば、[<Budget> + <>] は、唯一の出力がタイプ <Budget> のコンセプトであるパターンに一致します。 注:[price + <>] ではなく、[<Budget> + <>] のように、タイプ パターンでパターンの + 記号の後に使用する場合にのみ、<> を使用して空のタイプを示すことができます。 注: 他の要素を追加せずにこのパターンをキャプチャする場合、パターンを使用して規則を作成するのではなく、パターンを直接カテゴリに追加することをお勧めします。
[a + !(b)]	コンセプト a を含み、コンセプト b を含まないパターンを少なくとも 1 つ含む。少なくとも 1 つのパターンを含む必要があります。 たとえば、[price + !(high)] またはタイプの場合は [!(<Fruit>|<Vegetable>) + <Positive>] となります。
!([<A> & <B>])	特定のパターンを含まない。たとえば、!([<Budget> & <Negative>]) となります。

注:条件規則がどのようにテキストに合致するかについての例は、「カテゴリ規則の例」（ p. 227 ）を参照してください。

カテゴリ規則におけるワイルドカードの使用

ワイルドカードを規則のコンセプトに追加して、マッチング機能を拡張することができます。アスタリスク * ワイルカードを単語の前および/または後に追加して、コンセプトがどのように一致するかを指定できます。ワイルドカードの使用方法には次の 2 種類があります。

- **ワイルドカードの接頭（尾）辞**：ワイルドカードが接頭辞または接尾辞として使用されます。文字列とアスタリスクの間にスペースはありません。たとえば、operat* は、operat、operate、operates、operations、operational などと一致します。
- **単語ワイルドカード**：ワイルドカードがコンセプトの前または後ろに使用されます。文字列とアスタリスクの間にスペースがあります。たとえば、* operation は、operation、surgical operation、post operation などと一致します。たとえば、* operat* * のように、ワイルドカードが接頭（尾）辞とともに単語ワイルドカードを使用できます。この場合、operation、surgical operation、telephone operator、operatic aria などと一致します。最後の例のように、範囲があまりに幅広くなったり、不要なマッチをキャプチャしないよう、ワイルドカードを注意して使用することをお勧めします。

例外:

- ワイルドカードは、単独で使用できません。たとえば、(apple | *) は無効です。
- ワイルドカードをタイプ名の一致に使用することはできません。<Negative*> は、タイプ名に合致しません。
- 特定のタイプをワイルドカードで検索されたコンセプトに対する合致から除外することはできません。コンセプトが割り当てられるタイプは自動的に使用されます。
- ワイルドカードは、語の終わりであっても初めであっても（open* account）、または独立した要素であっても（open * account）語の連鎖の途中に使用することはできません。タイプ名にワイルドカードを使用することはできません。たとえば、apple* recipe など、word* word は「applesauce recipe」や他の言葉にも合致しません。ただし、apple* * は、applesauce recipe、apple pie、apple などの言葉に合致します。また、apple * toast など、word* word は 2 つの語の間にアスタリスクを使用しているため、apple cinnamon toast という語には合致しません。ただし、apple* * は、apple cinnamon toast、apple、apple pie などに合致します。

テーブル 10-4
ワイルドカードの使用方法

式	ドキュメントまたはレコードが一致する条件
*apple	文字で終了し、接頭辞として任意の数の文字を使用しているコンセプトを含む。例:*apple は、apple で終了し、次のように接頭辞を使用します。 - apple - pineapple - crabapple
apple*	文字で開始し、接尾辞として任意の数の文字を使用しているコンセプトを含む。例:*apple は、apple で開始し、次のように接尾辞を使用、または使用しません。 - apple - applesauce - applejack たとえば、apple* & !(pear* | quince) には、文字 apple で始まるコンセプトが含まれますが、文字 pear またはコンセプト quince で始まるコンセプトは含まれません。そのため、次のコンセプトとは一致しません。apple & quince 次のコンセプトとは一致します。 - applesauce - apple & orange
product	product という文字を含み、接頭辞または接尾辞、または両方として任意の数の文字を使用しているコンセプトを含む。 例:*product* は、次のコンセプトと一致します。 - product - byproduct - unproductive
* loan	単語 loan を含み、単語の前に別の単語と組み合わせる場合があるコンセプトを含む。たとえば、* loan は次のコンセプトと一致します。 - loan - car loan - home equity loan たとえば、[* delivery + <Negative>] は、前半が単語 delivery で終わり、後半にタイプ <Negative> を含むコンセプトを含み、次のコンセプト パターンと一致します。 - package delivery + slow - overnight delivery + late
event *	単語 event を含み、単語の後に別の単語が続く場合があるコンセプトを含む。たとえば、event * は次のコンセプトと一致します。 - event - event location - event planning committee
* apple *	任意の単語で始まり、次に apple で始まる単語が続き、別の単語が続く場合があるコンセプトを含む。* は 0 または n を意味するため、apple にも合致します。たとえば、* apple* は次のコンセプトと一致します。 - gala applesauce - granny smith apple crumble - famous apple pie - apple

式	ドキュメントまたはレコードが一致する条件
	たとえば、[* reservation* * + <Positive>] は、前半が単語 reservation のあるコンセプトを含み（コンセプト内の場所は関係ない）、後半にタイプ <Positive> を含み、次のコンセプト パターンと一致します。 - reservation system + good - online reservation + good

注:条件規則がどのようにテキストに合致するかについての例は、「カテゴリ規則の例」（ p. 227 ）を参照してください。

カテゴリ規則の例

それらを表現するために使用するシンタックスに基づき、規則がレコードにどのように合致するかを示すために、次の例ついて考えてみます。

例のレコード

次の 2 つのレコードがあるとします。

- レコード A: “when I checked my wallet, I saw I was missing 5 dollars.”
- レコード B: “$5 was found at the picnic area, but the blanket was missing.”

次の 2 つの表には、コンセプト パターンおよびタイプ パターンのほかにコンセプトおよびタイプに抽出される内容を示しています。

例から抽出されるコンセプトとタイプ

テーブル 10-5
コンセプトとタイプの抽出例

抽出コンセプト	コンセプトのタイプ
wallet	<不明>
missing	<否定的>
USD5	<通貨>
blanket	<不明>
picnic area	<不明>

例から抽出される TLA パターン

テーブル 10-6
TLA パターン出力の抽出例

抽出されるコンセプト パターン	抽出されるタイプパターン	レコード
picnic area + .	<不明> + <>	レコード B
wallet + .	<不明> + <>	レコード A

抽出されるコンセプトパターン	抽出されるタイプパターン	レコード
blanket + missing	<**不明**> + <**否定的**>	レコード B
USD5 + .	<**通貨**> + <>	レコード B
USD5 + missing	<**通貨**> + <**否定的**>	レコード A

カテゴリ規則の合致

次の表には、カテゴリ規則エディタに入力できるシンタックスをいくつか示しています。すべての規則が機能するわけではなく、またすべてが同じレコードに合致するわけではありません。異なるシンタックスが合致したレコードにどのように影響するかを確認してください。

テーブル 10-7
条件規則のサンプル

条件規則シンタックス	結果
USD5 & missing	抽出コンセプト missing および抽出コンセプト USD5 の両方が含まれているため、レコード A および B のいずれにも合致します。これは以下と同じになります。 (USD5 & missing)
missing & USD5	抽出コンセプト missing および抽出コンセプト USD5 の両方が含まれているため、レコード A および B のいずれにも合致します。これは以下と同じになります。 (missing & USD5)
missing & <**通貨**>	抽出コンセプト missing およびタイプ <**通貨**> に合致するコンセプト両方が含まれているため、レコード A および B のいずれにも合致します。これは以下と同じになります。 (missing & <**通貨**>)
<**通貨**> & missing	抽出コンセプト missing およびタイプ <**通貨**> に合致するコンセプト両方が含まれているため、レコード A および B のいずれにも合致します。これは以下と同じになります。 (<**通貨**> & missing)
[USD5 + missing]	レコード B は USD5 + missing を含む TLA パターン出力を作成していないため、A とは合致しますが B 都は合致しません（前の表を参照）。これは次の TLA パターン出力と同じになります。 USD5 + missing
[missing + USD5]	抽出された TLA パターン（前の表参照）が、最初の位置に missing を使用して表現した順序に合致しないため、レコード A にも B にも合致しません。これは次の TLA パターン出力と同じになります。 USD5 + missing

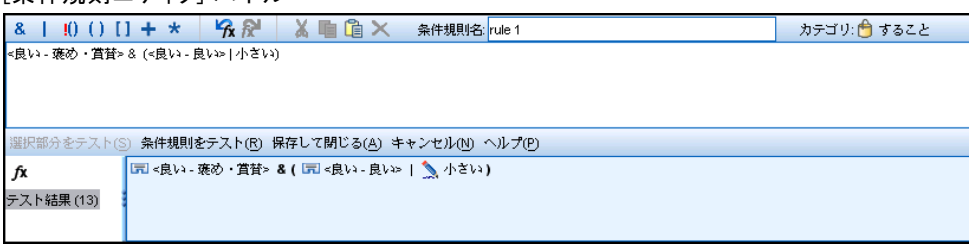
条件規則シンタックス	結果
[missing & USD5]	こうした TLA パターンがレコード B から抽出されているため、A とは合致しますが B とは合致しません。& 文字を使用すると、合致時の順序が重要でないことを示すため、この規則では [missing + USD5] または [USD5 + missing] のいずれかを検索します。レコード A の [USD5 + missing] のみに合致があります。
[missing + <**通貨**>]	抽出された TLA パターンがこの順序に合致していないため、レコード A にも B にも合致しません。TLA 出力はキーワード (USD5 + missing) またはタイプ (<**通貨**> + <Negative>) にのみ基づくため、同等のものはありませんが、コンセプトおよびタイプを組み合わせません。
[<**通貨**> + <**否定的**>]	TLA パターンがレコード B から抽出されているため、レコード A に合致しますが B には合致しません。以下の TLA 出力と同じになります。 <**通貨**> + <**否定的**>
[<**否定的**> + <**通貨**>]	抽出された TLA パターンがこの順序に合致していないため、レコード A にも B にも合致しません。**意見**テンプレートの場合、デフォルトでは、トピックが意見とともに検出されると、トピック (<**通貨**>) は最初のスロットに位置し、意見 (<**否定的**>) は 2 番目のスロットに位置します。

カテゴリ規則の作成

規則を作成または編集する場合、規則を条件規則エディタで開く必要があります。コンセプト、タイプ、またはパターンを追加し、またワイルドカードを使用して一致を拡張することができます。認識されたコンセプト、タイプ、パターンを使用すると、すべての関連コンセプトを検出するため、便利です。たとえば、コンセプトを使用すると、そのすべての関連キーワード、複数形、および類義語も規則を満たします。同様に、タイプを使用すると、そのすべてのコンセプトも規則にキャプチャされます。

既存の規則を編集するか、カテゴリ名を右クリックして **[条件規則を作成]** を選択して、条件規則エディタを開くことができます。

図 10-15
[条件規則エディタ] パネル



コンテキスト メニューを使用、ドラッグアンドドロップ、または手動でコンセプト、タイプおよびパターンをエディタに入力します。これらのブール型演算子（&，!()，|）やブラケットを使用して、条件規則式を形成します。一般的なエラーを回避するために、コンセプトを［抽出結果］パネルまたは［データ］パネルから直接条件規則エディタにドラッグアンドドロップすることを推奨します。エラーを回避するため、規則のシンタックスには十分注意してください。 詳細は、 p. 220 カテゴリ規則シンタックス を参照してください。

注:条件規則がどのようにテキストに合致するかについての例は、「カテゴリ規則の例」（ p. 227 ）を参照してください。

規則を作成するには

- データがまだ抽出されていない、または抽出が過去のものである場合は、抽出してください。 詳細は、 9 章 p. 156 データの抽出 を参照してください。
- ［カテゴリ］パネルで、規則を追加するカテゴリを選択します。
- メニューから、**[カテゴリ] → [条件規則の作成]** を選択します。ウィンドウにエディタのカテゴリ条件規則エディタのパネルが開きます。
- ［条件規則名］フィールドに、規則の名前を入力します。名前を入力しない場合、自動的に式を名前として使用します。規則の名前は、後で変更できます。
- より大きい式のテキスト フィールドで、次の作業を実行できます。
  - テキストをフィールドに直接入力するか、別のパネルからドラッグアンドドロップします。抽出されたコンセプト、タイプ、パターンのみを使用します。たとえば、cats という単語を入力しても単数形の cat のみが［抽出結果］パネルに表示される場合、エディタは cats を認識できなくなります。単数形には自動的に複数形が含まれる場合があり、そうでない場合はワイルドカードを使用することができます。 詳細は、p. 220 カテゴリ規則シンタックス を参照してください。
  - 規則に追加するコンセプト、タイプ、またはパターンを選択してメニューを使用します。
  - ブール型演算子を規則のリンク要素に追加します。ツールバー ボタンを使用して、「and」のブール型演算子 &、「or」のブール型演算子 |、「not」のブール型演算子 !()、カッコ ()、パターンのブラケット [ ] を規則に追加します。
- **[条件規則をテスト]** ボタンをクリックして、規則が適格であることを確認します。詳細は、 p. 220 カテゴリ規則シンタックス を参照してください。 見つかったドキュメントまたはレコードの数は、テキスト **[テスト結果]** の隣のカッコの中に表示されます。このテキストの右側に、認識された規則の要

素またはエラー メッセージが表示されます。タイプ、パターン、またはコンセプトの隣のグラフィックに赤い疑問符が表示されている場合、要素が既知の抽出に一致しないことを示します。一致しない場合、規則によってレコードは検出されません。

- 規則の一部をテストするには、該当する部分を選択して、[**選択部分をテスト**] をクリックします。
- 問題が見つかった場合は、必要な変更を行い、規則を再度テストします。
- 終了したら、[**保存して閉じる**] をクリックして、規則をもう一度保存し、エディタを閉じます。新しい規則名がカテゴリに表示されます。

規則の編集および削除

規則を作成および保存した後、その規則をいつでも編集することができます。 詳細は、 p. 220 カテゴリ規則シンタックス を参照してください。

規則が必要ない場合は、削除することができます。

規則を編集するには

- [カテゴリ定義] ダイアログ ボックスの [記述子] テーブルで、規則を選択します。
- メニューから、[**カテゴリ**] → [**条件規則の編集**] を選択するか、規則名をダブルクリックします。エディタが開き、選択された規則が表示されます。
- 既存の結果およびツールバー ボタンを使用して、変更を行います。
- 規則を再テストして、期待される結果が返されることを確認します。
- [**保存して閉じる**] をクリックして、規則をもう一度保存し、エディタを閉じます。

規則を削除するには

- [カテゴリ定義] ダイアログ ボックスの [記述子] テーブルで、規則を選択します。
- メニューから [**編集**] → [**削除**] を選択します。規則がカテゴリから削除されます。

定義済みのインポートおよびエクスポート

独自のカテゴリを Microsoft Excel (*.xls, *.xlsx) ファイルに保存している場合、それらを IBM® SPSS® Modeler Text Analytics にインポートできます。

また、使用中のインタラクティブ ワークベンチ セッション内のカテゴリを Microsoft Excel (*.xls, *.xlsx) ファイルにエクスポートすることもできます。 カテゴリをエクスポートすると、記述子やスコアなど、いくつかの追加情報を含めたり除外したりできます。 詳細は、 p.241 カテゴリのエクスポート を参照してください。

定義済みカテゴリにコードがない場合、または新しいコードが必要な場合、メニューから **[カテゴリ] > [カテゴリを管理] > [コードを自動生成]** を選択して、[カテゴリ] パネルでカテゴリ セットの新しいコードのセットを自動的に生成できます。 これにより、既存のコードがすべて削除され、自動的に番号が指定しなおされます。

定義済みカテゴリのインポート

定義済みカテゴリを IBM® SPSS® Modeler Text Analytics にインポートできます。 インポートする前に、定義済みカテゴリ ファイルが Microsoft Excel (*.xls, *.xlsx) ファイルであり、サポート可能な形式のいずれかの構造であることを確認してください。 形式を自動的に検知するよう選択することもできます。 次の形式がサポートされています。

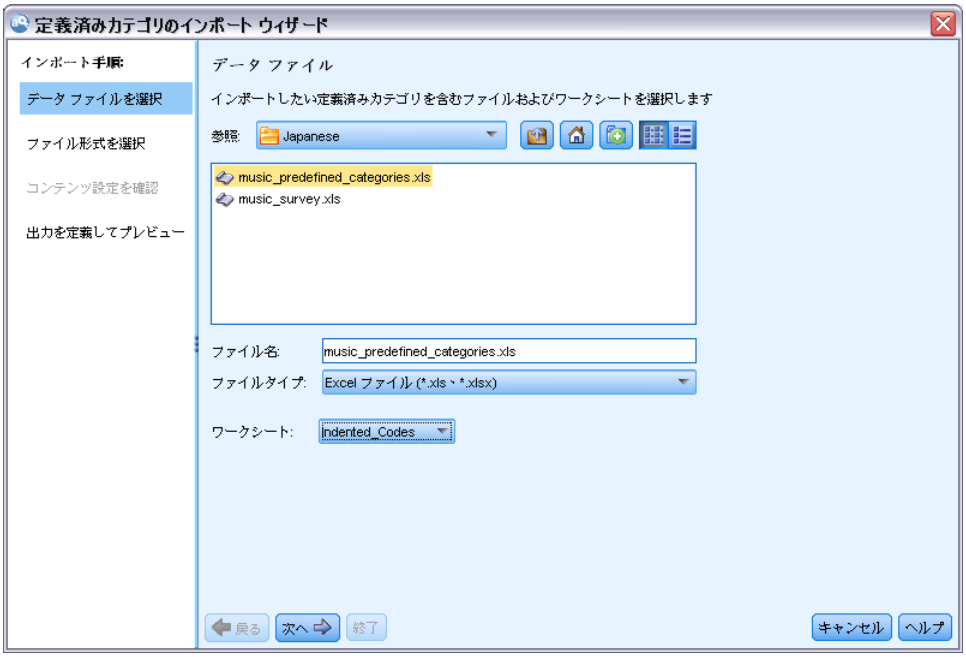
- **フラット リスト形式**: 詳細は、 p.237 フラット リスト形式 を参照してください。
- **コンパクト形式**: 詳細は、 p.238 コンパクト形式 を参照してください。
- **インデント形式**: 詳細は、 p.240 インデント形式 を参照してください。

注： 多くの言語の場合、事前定義されたカテゴリをどのようにインポートするかについて説明するデモ ストリームまたはデータ ファイルがあります。 ご使用言語の `<modeler_installation_directory>\Demos\Text_Analytics\` サブディレクトリを参照してください。

定義済みカテゴリをインポートするには

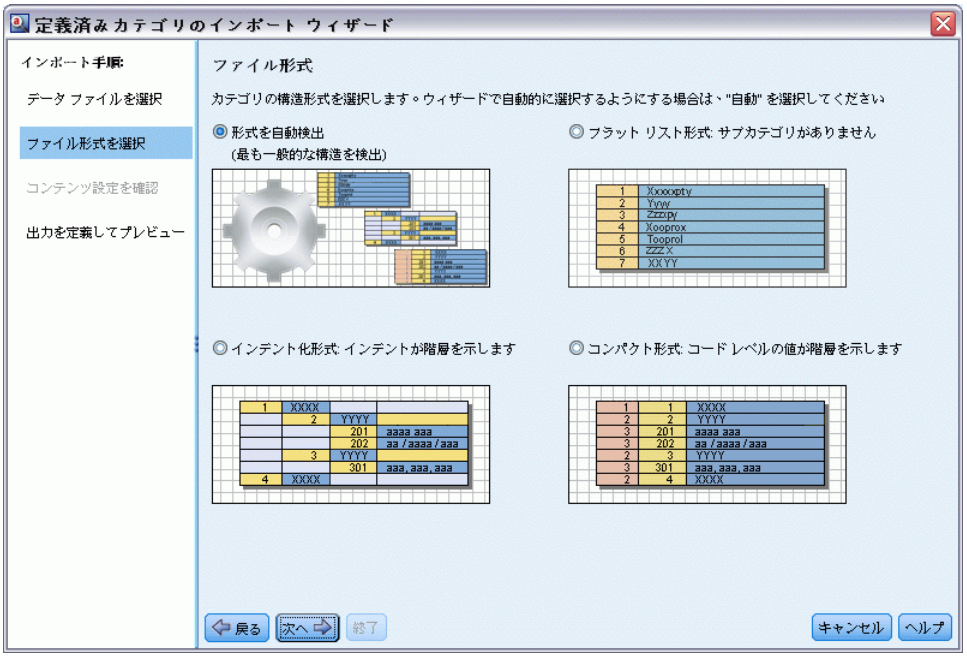
▶ メニューから、**カテゴリ > カテゴリを管理 > 定義済みカテゴリのインポート** を選択します。 定義済みカテゴリのインポート ウィザードが表示されます。

図 10-16
定義済みカテゴリのインポート ウィザード



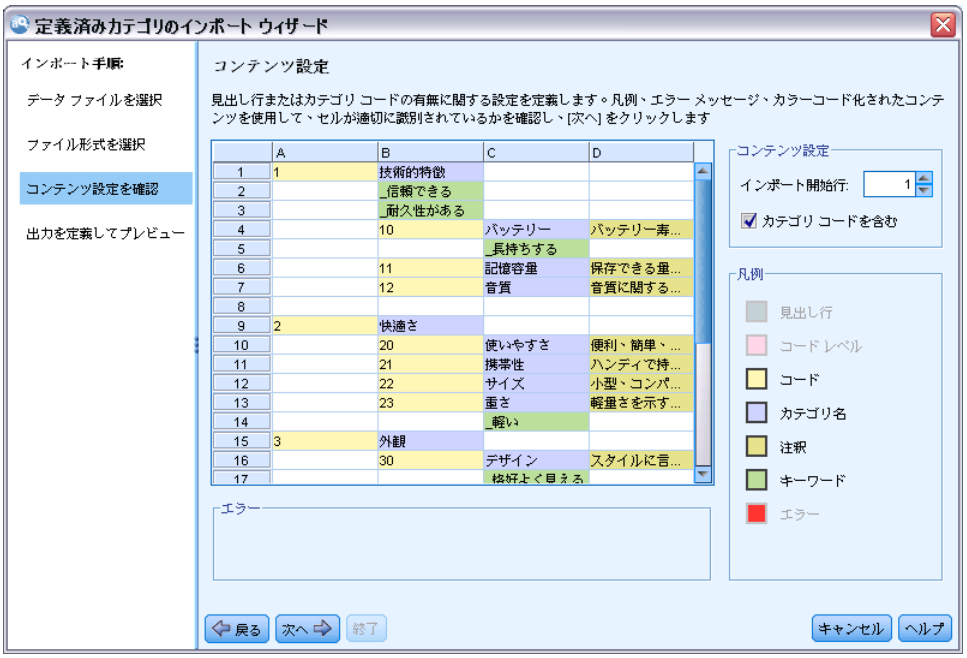
- ［参照］ドロップダウン リストから、ファイルを投入するドライブとフォルダを選択します。
- リストからファイルを選択します。［ファイル名］テキスト ボックスにファイルの名前が表示されます。
- リストから、定義済みカテゴリを含むワークシートを選択します。 ワークシート名が［ワークシート］フィールドに表示されます。
- **[次へ]** をクリックして、データ形式の選択を開始します。

図 10-17
[定義済みカテゴリのインポート] ダイアログボックス、データ形式のステップ



- ファイルの形式を選択するか、自動的に形式を検知しようとするオプションを選択します。 自動検知は、最も一般的な形式に最も適しています。
  - **フラット リスト形式**： 詳細は、 p. 237 フラット リスト形式 を参照してください。
  - **コンパクト形式**： 詳細は、 p. 238 コンパクト形式 を参照してください。
  - **インデント形式**： 詳細は、 p. 240 インデント形式 を参照してください。
- [次へ] をクリックして、追加のインポート オプションを定義します。 形式の自動検知を選択した場合、最後の手順に進みます。

図 10-18
定義済みカテゴリのインポート、インポート オプションの手順



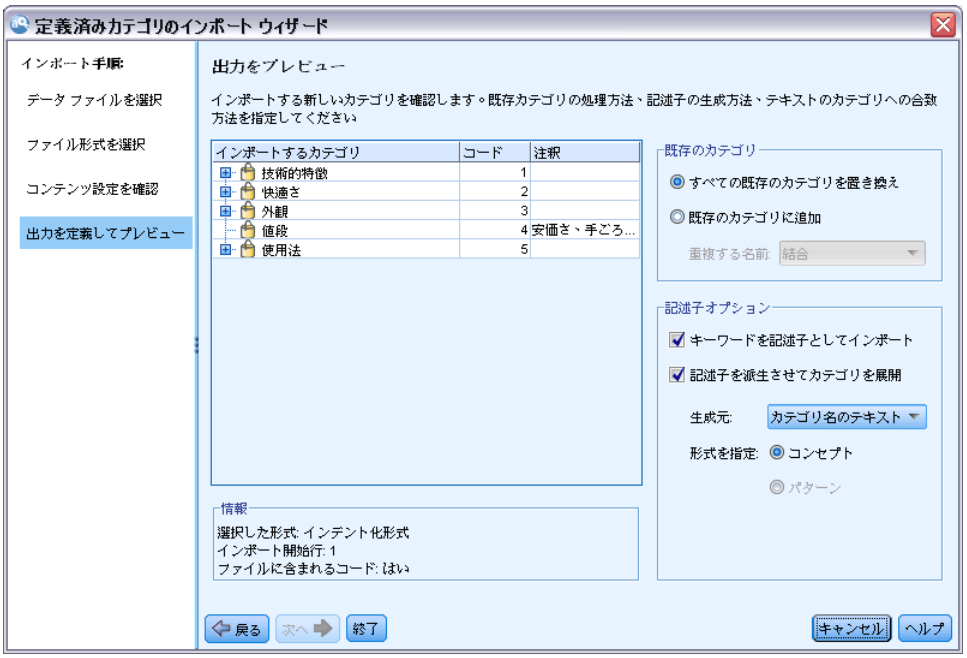
- このワークシートの 1 行または複数行が列見出しまたはその他の外部情報である場合、[**インポート開始行**] オプションでインポート開始場所である行番号を選択します。 たとえば、カテゴリ名が行 7 で始まる場合、ファイルを正しくインポートするためには、このオプションで行番号 7 を入力する必要があります。

- ファイルにカテゴリ コードが含まれている場合、オプション [**カテゴリ コードを含む**] を選択します。 これにより、ウィザードがデータを正しく認識します。

- カラーコード化されたセルおよび凡例を確認し、データが正しく特定されるようにします。 ファイルで検出されたエラーは赤で表示され、形式プレビュー テーブルの下に表示されます。 不正な形式が選択された場合、戻って別の形式を選択します。 ファイルを修正する必要がある場合、変更を行い、ファイルをもう一度選択してウィザードを再起動します。 ウィザードを終了する前にすべてのエラーを修正してください。

- [**次へ**] をクリックして、インポートされる一連のカテゴリおよびサブカテゴリを確認し、これらのカテゴリの記述子の作成方法を定義します。

図 10-19
[定義済みカテゴリのインポート] ダイアログボックス、プレビューのステップ



- テーブルでインポートされる一連のカテゴリを確認します。 記述子として表示されるはずのキーワードが表示されない場合、インポート時に認識されなかったことが考えられます。 正しく接頭辞が使用され、正しいセルに表示されていることを確認してください。

- セッションの既存のカテゴリの処理方法を選択します。
  - **すべての既存のカテゴリを置き換え**: 既存のカテゴリすべてを削除し、新しくインポートされたカテゴリが代わりに使用されます。
  - **既存のカテゴリに追加** カテゴリをインポートし、既存カテゴリと共通カテゴリを結合します。 既存のカテゴリに追加する場合、重複カテゴリの処理方法を決定する必要があります。 オプション [**結合**] を選択すると、カテゴリ名を共有する場合、インポートされたカテゴリは既存のカテゴリと結合されます。 オプション [**インポートから除外**] は、同じ名前が存在する場合、カテゴリのインポートを禁止します。

- [**キーワードを記述子としてインポート**] は、関連するカテゴリの記述子としてデータで特定されるキーワードをインポートするオプションです。

- [**記述子から派生してカテゴリを拡張**] は、カテゴリ、またはサブカテゴリの名前を示す単語、および注釈を構成する単語空記述子を生成するオプションです。 単語が抽出結果に合致する場合、記述子としてカテゴリに追加されます。 このオプションを選択すると、カテゴリ名または注釈が長く記述的である場合に、最良の結果を作成します。 迅速にカテゴリの記述

子を生成し、またカテゴリはこれらの記述子を含むレコードをキャプチャすることができます。

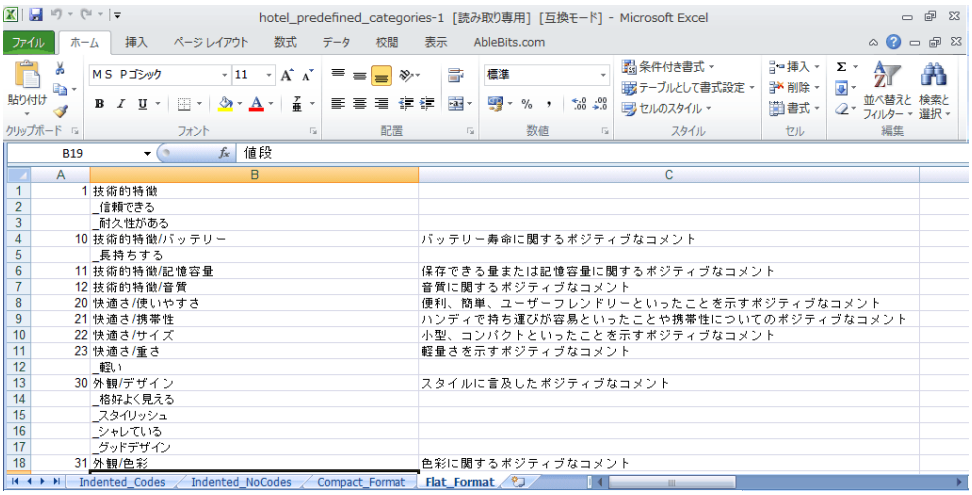
- [**派生元**] フィールドを使用して、記述子が派生するテキスト、名前またはカテゴリおよびサブカテゴリ、注釈内の単語から選択できます。
- [**形式を指定**] フィールドを使用して、これらの記述子をコンセプトまたは TLA パターンの形式で作成することを選択できます。 TLA 抽出が行われない場合、ウィザードの [**パターン**] オプションが無効となります。

▶ [**完了**] をクリックすると、定義済みカテゴリが [カテゴリ] パネルにインポートされます。

フラット リスト形式

このフラット リスト形式では、階層のない、上位レベルのカテゴリのみがあります。つまり、サブカテゴリやサブネットはありません。 カテゴリ名は 1 つの列に表示されます。

図 10-20
フラット リスト形式の例



	A	B	C
1	1	技術的特徴	
2		_信頼できる	
3		_耐久性がある	
4	10	技術的特徴/バッテリー	バッテリー寿命に関するポジティブなコメント
5		_長持ちする	
6	11	技術的特徴/記憶容量	保存できる量または記憶容量に関するポジティブなコメント
7	12	技術的特徴/音質	音質に関するポジティブなコメント
8	20	快適さ/使いやすさ	便利、簡単、ユーザーフレンドリーといったことを示すポジティブなコメント
9	21	快適さ/携帯性	ハンディで持ち運びが容易といったことや携帯性についてのポジティブなコメント
10	22	快適さ/サイズ	小型、コンパクトといったことを示すポジティブなコメント
11	23	快適さ/重さ	軽量さを示すポジティブなコメント
12		_軽い	
13	30	外観/デザイン	スタイルに言及したポジティブなコメント
14		_格好よく見える	
15		_スタイリッシュ	
16		_シャレている	
17		_グッドデザイン	
18	31	外観/色彩	色彩に関するポジティブなコメント

この形式のファイルには、次の情報が含まれます。

- オプションの**コード**列には、各カテゴリを一意に特定する数値が入力されます。 データ ファイルにコードを含まないよう指定（[**コンテンツ設定**] で [**カテゴリ コードを含む**] オプションを選択）した場合、カテゴリ名の左隣のセルに各カテゴリの一意のコードを含む列がなければなりません。 データにコードが含まれず、後でコードをいくつか作成したい場合、 後でいつでもコードを生成できます（**カテゴリ > カテゴリを管理 > コードを自動生成**）。
- 必須の**[カテゴリ名]** 列には、カテゴリのすべての名前が入力されています。 この列は、この形式を使用してインポートする場合に必要です。

- カテゴリ名のすぐ右にあるオプションの **[注釈]** セル。 この注釈は、カテゴリ/サブカテゴリを説明するテキストで構成されています。
- オプションの **[キーワード]** は、カテゴリの記述子としてインポートできます。 認識できるようにするために、これらのキーワードは関連するカテゴリ/サブカテゴリ名のすぐ下にあるセルになければならず、「_firearms, weapons / guns」のようにキーワードの前にアンダースコア（_）を追加する必要があります。 キーワード セルには、各カテゴリの説明に使用する 1 つまたは複数の単語を入力できます。 これらの単語は、ウィザードの最後の手順の指定内容に応じて、記述子としてインポートしたり、無視されます。 後で、記述子はテキストから抽出された結果と比較されます。 合致が見つかった場合、レコードまたはドキュメントがこの記述子を含むカテゴリにスコアリングされます。

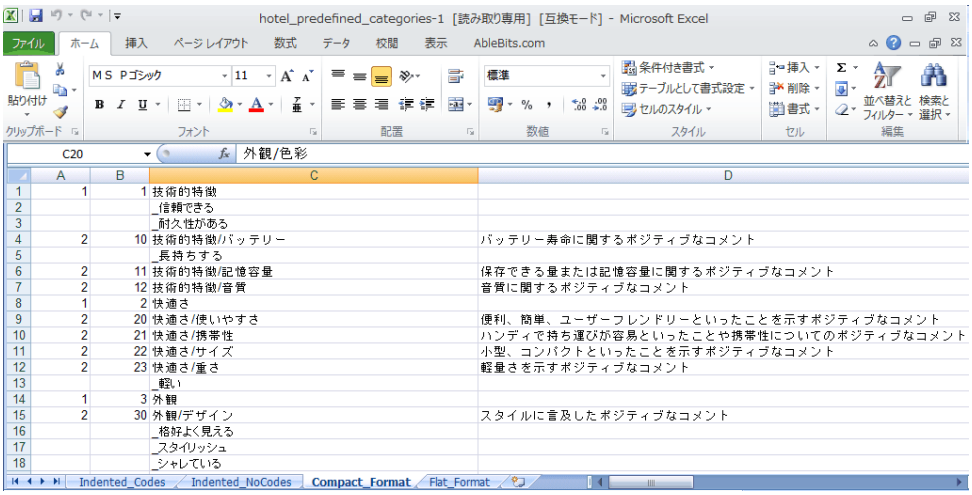
テーブル 10-8
コード、キーワード、および注釈を含むフラット リスト形式

列 A	列 B	列 C
カテゴリ コード（オプション）	カテゴリ名	注釈
	_記述子/キーワード リスト（オプション）	

コンパクト形式

コンパクト形式は、階層カテゴリで使用される点を除いて、フラット リスト形式と同じ構造です。 そのため、各カテゴリおよびサブカテゴリの階層レベルを定義するには、コード レベル列が必要です。

図 10-21
Microsoft Excel のコンパクト形式の定義済みカテゴリ ファイルの例



	A	B	C	D
1	1	1	技術的特徴	
2			_信頼できる	
3			_耐久性がある	
4	2	10	技術的特徴/バッテリー	バッテリー寿命に関するポジティブなコメント
5			_長持ちする	
6	2	11	技術的特徴/記憶容量	保存できる量または記憶容量に関するポジティブなコメント
7	2	12	技術的特徴/音質	音質に関するポジティブなコメント
8	1	2	快適さ	
9	2	20	快適さ/使いやすさ	便利、簡単、ユーザーフレンドリーといったことを示すポジティブなコメント
10	2	21	快適さ/携帯性	ハンディで持ち運びが容易といったことや携帯性についてのポジティブなコメント
11	2	22	快適さ/サイズ	小型、コンパクトといったことを示すポジティブなコメント
12	2	23	快適さ/重さ	軽量さを示すポジティブなコメント
13			_軽い	
14	1	3	外観	
15	2	30	外観/デザイン	スタイルに言及したポジティブなコメント
16			_格好よく見える	
17			_スタイリッシュ	
18			_シャレている	

この形式のファイルには、次の情報が含まれます。

- 必須の **[コード レベル]**列には、その行の後続の情報の階層の位置を示す番号が入力されます。 たとえば、値 1、2、3、が指定され、カテゴリおよびサブカテゴリの両方がある場合、1 はカテゴリ、2 はサブカテゴリ、3 はサブ-サブカテゴリを示します。 カテゴリおよびサブカテゴリのみがある場合、1 はカテゴリを、2 はサブカテゴリを示します。 カテゴリの深さの限り続きます。
- オプションの**コード**列には、各カテゴリを一意に特定する値が入力されます。 データ ファイルにコードを含まないよう指定（**[コンテンツ設定]**で **[カテゴリ コードを含む]** オプションを選択）した場合、カテゴリ名の左隣のセルに各カテゴリの一意のコードを含む列がなければなりません。 データにコードが含まれず、後でコードをいくつか作成したい場合、 後でいつでもコードを生成できます（**カテゴリ > カテゴリを管理 > コードを自動生成**）。
- 必須の**[カテゴリ名]** 列には、カテゴリおよびサブカテゴリのすべての名前が入力されています。 この列は、この形式を使用してインポートする場合に必要です。
- カテゴリ名のすぐ右にあるオプションの **[注釈]** セル。 この注釈は、カテゴリ/サブカテゴリを説明するテキストで構成されています。
- オプションの **[キーワード]** は、カテゴリの記述子としてインポートできます。 認識できるようにするために、これらのキーワードは関連するカテゴリ/サブカテゴリ名のすぐ下にあるセルになければならず、「_firearms, weapons / guns」のようにキーワードの前にアンダースコア（_）を追加する必要があります。 キーワード セルには、各カテゴリの説明に使用する 1 つまたは複数の単語を入力できます。 これらの単語は、ウィザードの最後の手順の指定内容に応じて、記述子としてインポートしたり、無視されます。 後で、記述子はテキストから抽出された結果と比較されます。 合致が見つかった場合、レコードまたはドキュメントがこの記述子を含むカテゴリにスコアリングされます。

テーブル 10-9
コードを含むコンパクト形式の例

列 A	列 B	列 C
階層コード レベル	カテゴリ コード（オプション）	カテゴリ名
階層コード レベル	サブカテゴリ コード（オプション）	サブカテゴリ名

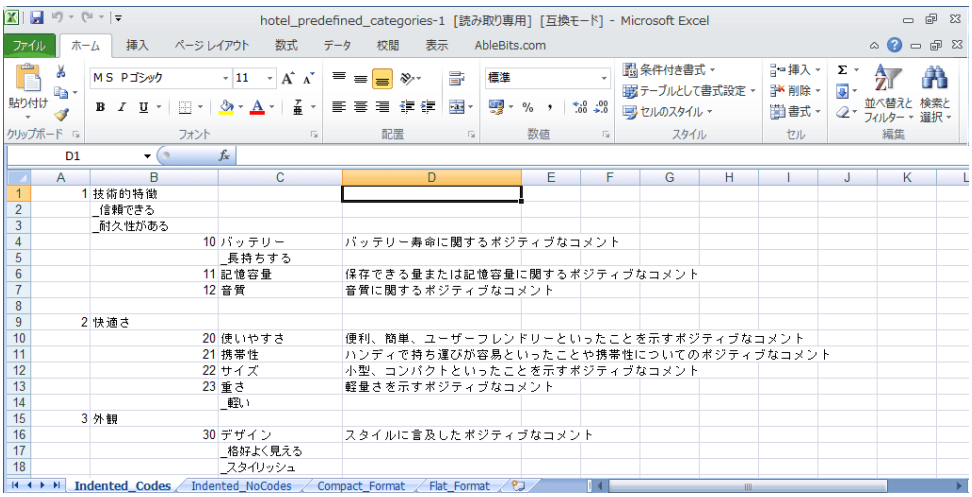
テーブル 10-10
コードを含まないコンパクト形式の例

列 A	列 B
階層コード レベル	カテゴリ名
階層コード レベル	サブカテゴリ名

インデント形式

インデント ファイル形式の場合、コンテンツには階層があります。つまり、カテゴリと、1 レベルまたは複数レベルのサブカテゴリがあります。さらに、構造がこの階層を示すよう、インデント化されています。 ファイルの各行にはカテゴリまたはサブカテゴリが含まれますが、サブカテゴリはカテゴリからインデント化され、サブ-サブカテゴリはサブカテゴリからインデント化されています。 この構造を Microsoft Excel で手動で作成したり、別の製品からエクスポートし Microsoft Excelにインポートした構造を使用することもできます。

図 10-22
Microsoft Excel のインデント構造のカテゴリの例



- **最上位レベルのカテゴリ コードおよびカテゴリ名**は、それぞれ列 A および列 B に表示されます。 またはコードがない場合、カテゴリ名が列 A に表示されます。
- **サブカテゴリ コードおよびサブカテゴリ名**は、それぞれ列 B および列 C に表示されます。 またはコードがない場合、サブカテゴリ名が列 B に表示されます。サブカテゴリはカテゴリのメンバーです。 上位レベルのカテゴリがない場合、サブカテゴリはありません。

テーブル 10-11
コードを含むインデント構造

列 A	列 B	列 C	列 D
カテゴリ コード（オプション）	カテゴリ名		
	サブカテゴリ コード（オプション）	サブカテゴリ名	
		サブ-サブカテゴリ コード（オプション）	サブ-サブカテゴリ名

テーブル 10-12
コードのないインデント構造

列 A	列 B	列 C
カテゴリ名		
	サブカテゴリ名	
		サブ-サブカテゴリ名

この形式のファイルには、次の情報が含まれます。

- オプションの**コード**は、各カテゴリまたはサブカテゴリを一意に特定する数値でなければなりません。 データ ファイルにコードを含まないよう指定（[**コンテンツ設定**] で [**カテゴリ コードを含む**] オプションを選択）した場合、カテゴリ名の左隣のセルに各カテゴリまたはサブカテゴリの一意のコードがなければなりません。 データにコードが含まれず、後でコードをいくつか作成したい場合、 後でいつでもコードを生成できます（**カテゴリ > カテゴリを管理 > コードを自動生成**）。
- 各カテゴリおよびサブカテゴリの必須の**[名前]**。 サブカテゴリは、カテゴリから、各行の右側に 1 セルインデントされていなければなりません。
- カテゴリ名のすぐ右にあるオプションの **[注釈]** セル。 この注釈は、カテゴリ/サブカテゴリを説明するテキストで構成されています。
- オプションの **[キーワード]** は、カテゴリの記述子としてインポートできます。 認識できるようにするために、これらのキーワードは関連するカテゴリ/サブカテゴリ名のすぐ下にあるセルになければならず、「_firearms, weapons / guns」のようにキーワードの前にアンダースコア（_）を追加する必要があります。 キーワード セルには、各カテゴリの説明に使用する 1 つまたは複数の単語を入力できます。 これらの単語は、ウィザードの最後の手順の指定内容に応じて、記述子としてインポートしたり、無視されます。 後で、記述子はテキストから抽出された結果と比較されます。 合致が見つかった場合、レコードまたはドキュメントがこの記述子を含むカテゴリにスコアリングされます。

重要! 1 つのレベルでコードを使用する場合、各カテゴリおよびサブカテゴリにコードを含む必要があります。 そうでない場合、インポート プロセスが失敗します。

カテゴリのエクスポート

また、使用中のインタラクティブ ワークベンチ セッション内のカテゴリを Microsoft Excel（*.xls, *.xlsx）ファイル形式にエクスポートすることもできます。 エクスポートされるデータは、大部分がカテゴリ プロパティの［カテゴリ］パネルの現在のコンテンツのデータです。 その

ため、**ドキュメント** スコア値もエクスポートする場合は、もう一度スコアリングを行うことをお勧めします。

常にエクスポート...

- ある場合はカテゴリ コード
- カテゴリ（およびサブカテゴリ）名
- ある場合はコード レベル（フラット/コンパクト形式）
- 列見出し（フラット/コンパクト形式）

オプションでエクスポート...

- **ドキュメント** スコア
- カテゴリの注釈
- 記述子名
- 記述子数

重要！ 記述子をエクスポートする場合、それらはテキスト文字列とアンダースコアの接頭辞に変換されます。 この製品に再度インポートする場合、パターン、条件規則、単純なコンセプトである記述子と区別する機能が失われます。 これらのカテゴリの本製品で再利用する場合、テキスト分析パッケージ（TAP）を使用して、現在定義されているすべての記述子（使用されているすべてのカテゴリ、コード、言語リソース）を保持することをお勧めします。 TAP ファイルは IBM® SPSS® Modeler Text Analytics と IBM® SPSS® Text Analytics for Surveys の両方で使用できます。 詳細は、 p.245 テキスト分析パッケージの使用 を参照してください。

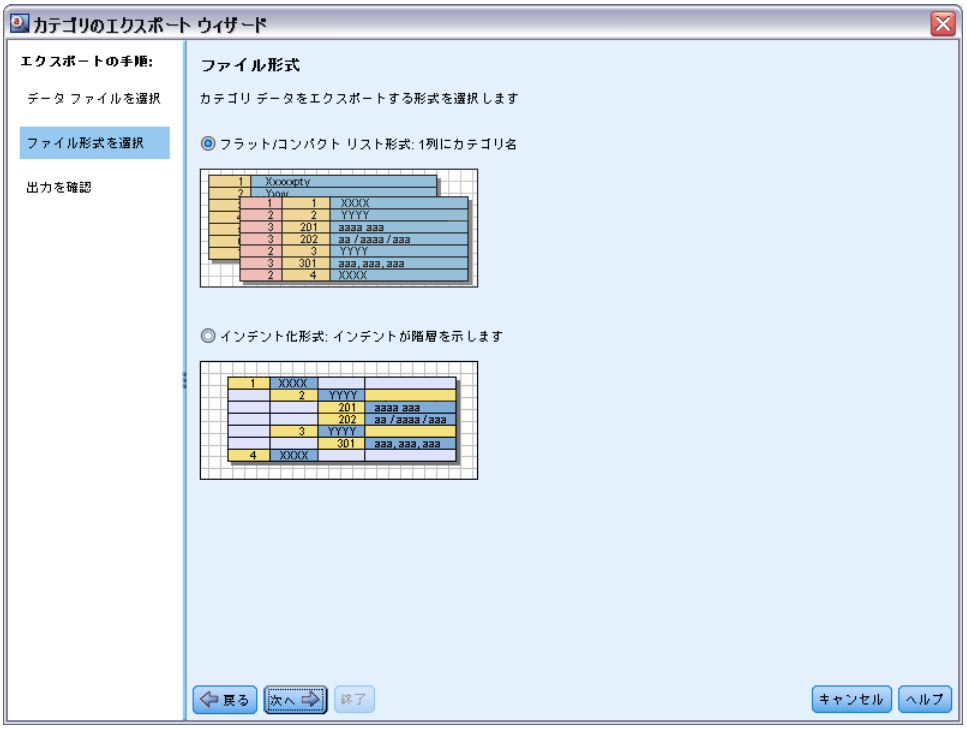
定義済みカテゴリをエクスポートするには

▶ メニューから、**カテゴリ** > **カテゴリを管理** > **カテゴリのエクスポート** を選択します。カテゴリのエクスポート ウィザードが表示されます。

図 10-23
カテゴリのエクスポート ウィザード、手順 1

- 場所を選択して、エクスポートするファイルの名前を入力します。
- ［ファイル名］テキスト ボックスに出力ファイルの名前を入力します。
- **[次へ]** をクリックして、カテゴリ データをエクスポートする形式を選択します。

図 10-24
カテゴリのエクスポート ウィザード、手順 2



▶ 次のいずれかの形式を選択します。

- **フラット リスト形式またはコンパクト リスト形式**： 詳細は、 p. 237 フラット リスト形式 を参照してください。フラット リスト形式には、サブカテゴリはありません。 詳細は、 p. 238 コンパクト形式 を参照してください。 コンパクト リスト形式には階層カテゴリが含まれています。
- **インデント形式**： 詳細は、 p. 240 インデント形式 を参照してください。

▶ [次へ] をクリックして、エクスポートするコンテンツを選択し、提案されたデータを確認します。

図 10-25
カテゴリのエクスポート ウィザード、手順 3

	A	B	C	D	E
1	コード レベル	コード	カテゴリ名	注釈	記述子数
2	1	1	技術的特徴		3
3			_耐久性がある		
4			_信頼できる		
5	2	10	バッテリー	バッテリー寿...	1
6			_長持ちする		
7	2	11	記憶容量	保存できる量...	0
8	2	12	音質	音質に関する...	0
9	1	2	快適さ		1
10	2	20	使いやすさ	便利、簡単、...	0
11	2	21	携帯性	ハンディで持...	0
12	2	22	サイズ	小型、コンパ...	0
13	2	23	重さ	軽量さを示す...	1
14			_軽い		
15	1	3	外観		4
16	2	30	デザイン	スタイルに言...	4
17			_シャレている		
18			_グッドデザ...		
19			_スタイリッ...		
20			_格好よく見...		
21	2	31	色彩	色彩に関する...	0
22	1	4	値段	安価さ、手ご...	0
23	1	5	使用法		1
24	2	50	音楽	音を聴くこと...	0
25	2	51	ビデオ		1

- エクスポート ファイルの内容を確認します。
- **注釈**または**記述子名**など、エクスポートする追加内容を選択または選択解除します。
- [**完了**] をクリックすると、カテゴリがエクスポートされます。

テキスト分析パッケージの使用

TAP とも呼ばれるテキスト分析パッケージは、テキスト回答のカテゴリ化を行うためのテンプレートとして機能します。TAP には多くのレコードを迅速かつ自動的にコード化するために必要な事前に作成されたカテゴリ セットおよび言語リソースが含まれているため、TAP を使用すると、最小限の介入でテキスト データをカテゴリ化できます。言語リソースを使用して、テキスト データを分析およびマイニングし、主要キーワードを抽出します。テキストの主要キーワードおよびパターンに基づき、レコードを TAP で選択したカテゴリ セットにカテゴリ化できます。独自の TAP を作成またはTAP を更新できます。

TAP は、次の要素で構成されています。

- **カテゴリ セット**:カテゴリ セットは、定義済みカテゴリ、カテゴリ コード、各カテゴリの記述子、カテゴリ セット全体の名前で構成されています。記述子とは、キーワード安いやパターン高価などの言語学的要素です。記述子を使用してカテゴリを定義し、テキストがカテゴリ記述子に一致すると、ドキュメントまたはレコードがカテゴリに投入されます。
- **言語リソース**:言語リソースは、一連のライブラリ、および主要キーワードおよびパターンを抽出するために調整された高度なリソースです。これらの抽出コンセプトおよびパターンは、レコードをカテゴリ セットのカテゴリに投入できる記述子として使用されます。

独自の TAP を作成、TAP を更新、または TAP を読み込むことができます。

TAP を選択してカテゴリ セットを選択した後、IBM® SPSS® Modeler Text Analytics でレコードを抽出およびカテゴリ化できます。

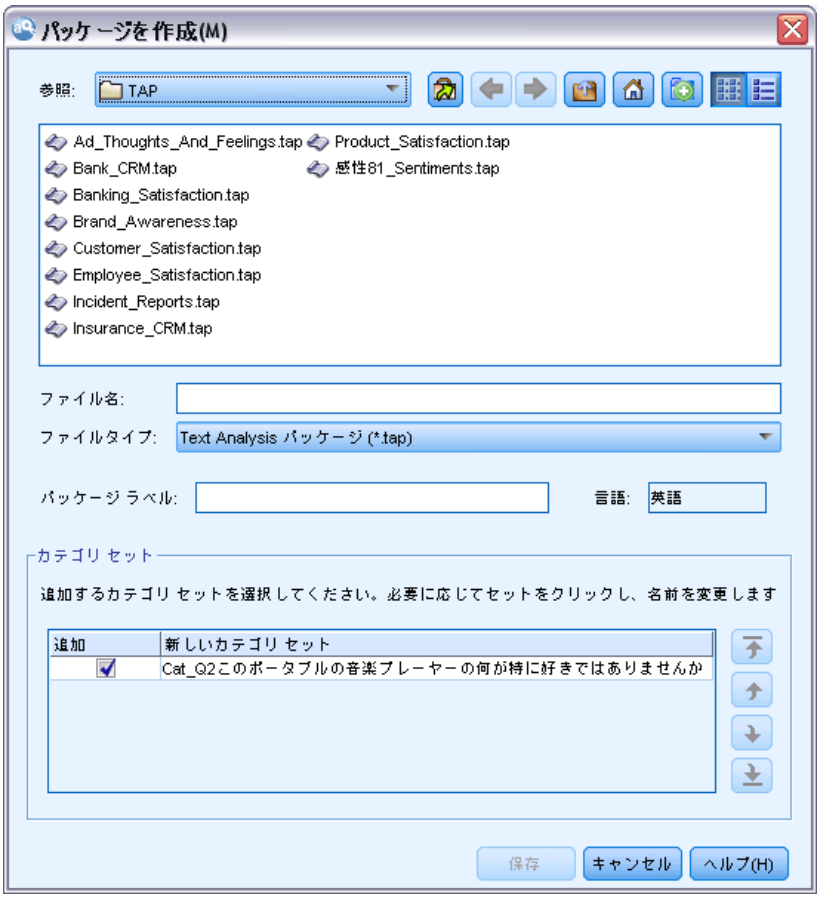
注:TAP を作成し、IBM® SPSS® Text Analytics for Surveys および SPSS Modeler Text Analytics の製品間で交互に使用できます。

テキスト分析パッケージの作成

少なくとも 1 つのカテゴリといくつかのリソースを含むセッションがある場合、オープン インタラクティブ ワークベンチ セッションのコンテンツからテキスト分析パッケージ (TAP) を作成できます。カテゴリおよび記述子 (コンセプト、タイプ、条件規則または TLA パターン出力) のセットをリソース エディタで開かれたすべての言語リソースと共に使用して TAP を作成できます。

リソースが作成された言語を表示できます。言語は、テンプレート エディタ または リソース エディタ ビューの [高度なリソース] タブで設定します。

図 10-26
[テキスト分析パッケージの作成] ダイアログ



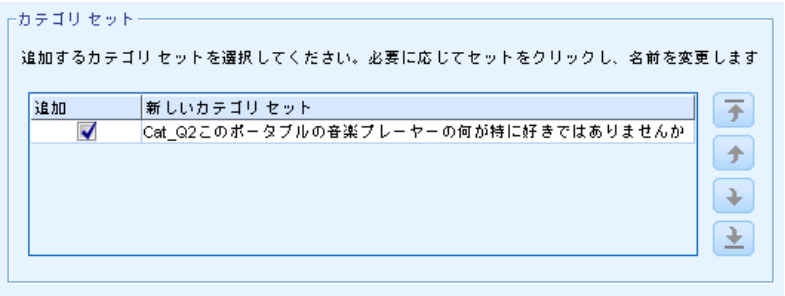
テキスト分析パッケージを作成するには

- メニューで、[ファイル] → [テキスト分析パッケージ] → [パッケージを作成] を選択します。[パッケージを作成] ダイアログが表示されます。
- TAP を保存するディレクトリを参照します。デフォルトでは、TAP は製品のインストール ディレクトリの \TAP サブディレクトリに保存されます。
- [ファイル名] フィールドに TAP の名前を入力します。
- [パッケージ ラベル] フィールドにラベルを入力します。ファイル名を入力すると、この名前がラベルとして自動的に表示されますが、このラベルは変更できます。
- TAP からカテゴリ セットを除外するには、[追加] チェックボックスをオフにします。カテゴリ セットを除外すると、カテゴリセットがパッケージに追加されなくなります。デフォルトでは、質問ごとに 1 つのカテゴ

リ セットが TAP に追加されます。TAP には、1 つ以上のカテゴリ セットが必要です。

- カテゴリ セットの名前を変更します。[新しいカテゴリ セット] 列にはデフォルトで一般名が入力されています。一般名はテキスト変数名に Cat_ 接頭辞を追加して生成されます。セルを 1 回クリックすると、名前を編集できます。入力して他の場所をクリックすると、名前の変更が適用されます。カテゴリ セットの名前を変更すると、TAP のみの名前が変更され、オープン セッションの編集名は変わりません。

図 10-27
カテゴリ セットの名前変更



- 必要に応じて、カテゴリ セット テーブルの右側にある矢印キーを使用して、カテゴリ セットを並べ替えます。

- [保存] をクリックして、テキスト分析パッケージを作成します。ダイアログ ボックスが閉じます。

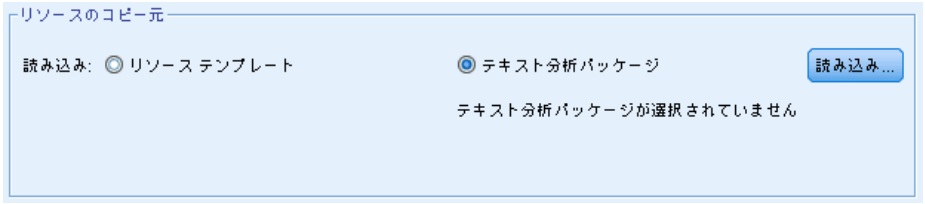
テキスト分析パッケージの読み込み

テキスト マイニング モデル作成ノードを設定している場合、抽出時に使用するリソースを指定する必要があります。リソース テンプレートを選択する代わりに、テキスト分析パッケージ (TAP) を選択して、そのリソースだけでなく、カテゴリ セットをノードにコピーすることができます。

カテゴリ セットをカテゴリ化の開始ポイントとして使用できるため、カテゴリ モデルをインタラクティブに作成する場合、TAP が最も重要となります。ストリームを実行すると、インタラクティブ ワークベンチ セッションが起動し、このセットのカテゴリが [カテゴリ] パネルに表示されます。このように、これらのカテゴリを使用してすぐにドキュメントおよびレコードをスコアリングを行い、これらのカテゴリが適切なものとなるまで調整、作成、拡張を続行します。 詳細は、 p. 181 カテゴリ作成の方法と戦略 を参照してください。

バージョン 14 以降、[読み込み] をクリックして TAP を選択すると、この TAP のリソースが定義された言語を表示することもできます。

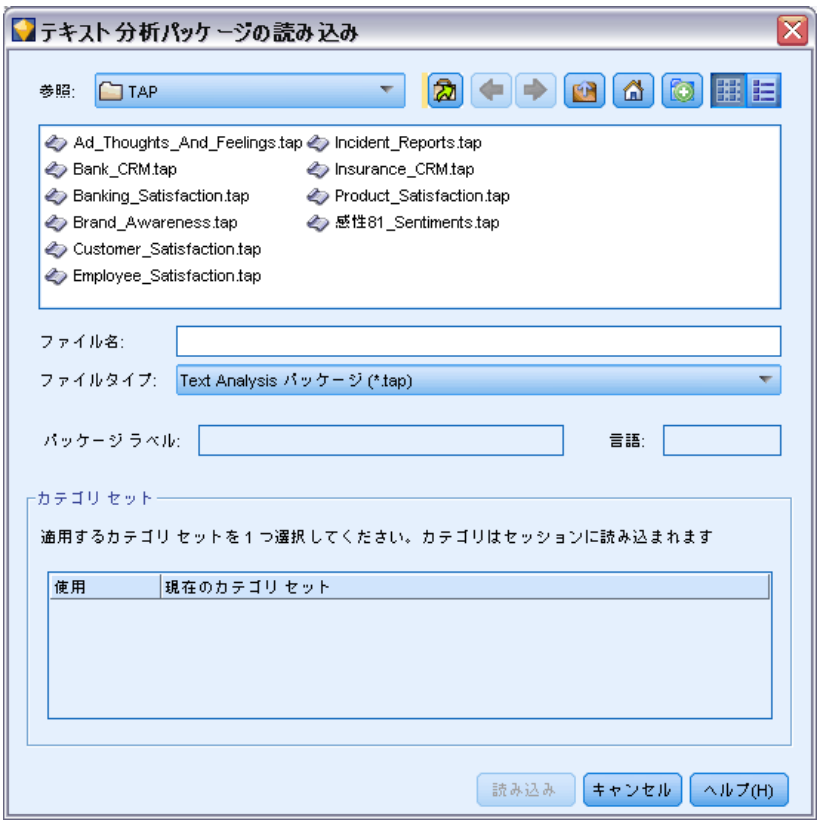
図 10-28
リソースをノードにコピーするための [モデル] タブのオプション



テキスト分析パッケージを読み込むには

- テキスト マイニング モデル作成ノードを編集します。
- ［モデル］タブの **[リソースのコピー元]** で、［テキスト分析パッケージ］を選択します。
- **[読み込み]** をクリックします。［テキスト分析パッケージの読み込み］ダイアログが表示されます。

図 10-29
[テキスト分析パッケージの読み込み] ダイアログ

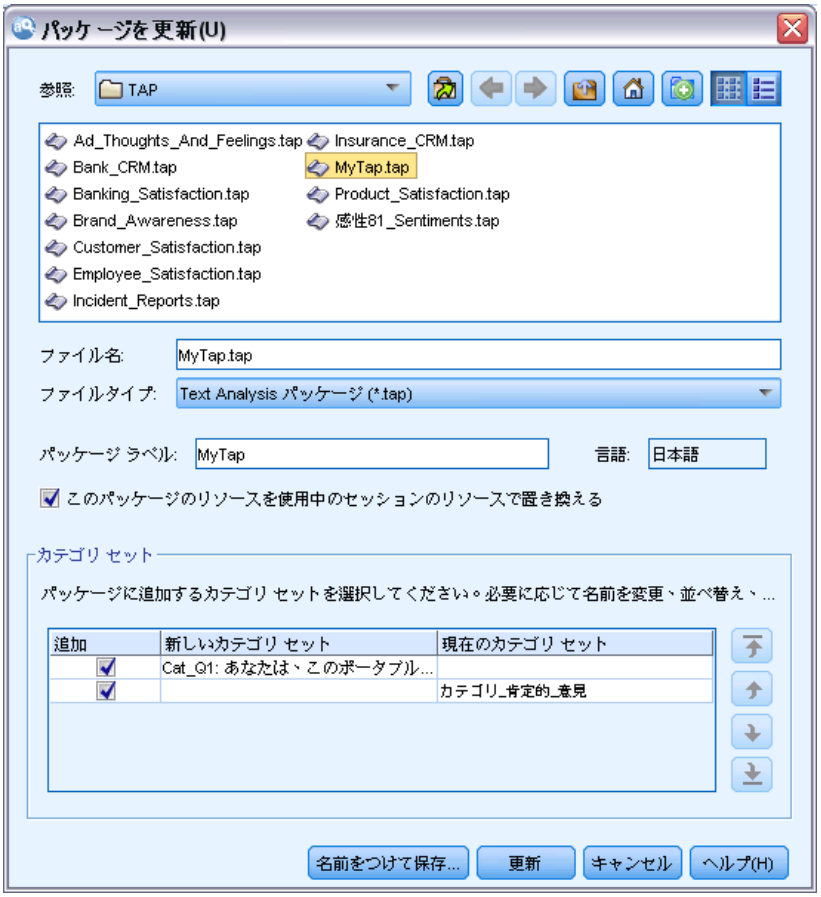


- ノードにコピーしたいリソースおよびカテゴリ セットを含む TAP の場所を参照します。デフォルトでは、TAP は製品のインストール ディレクトリの \TAP サブディレクトリに保存されます。
- [ファイル名] フィールドに TAP の名前を入力します。ラベルが自動的に表示されます。
- 使用したいカテゴリ セットを選択します。インタラクティブ ワークベンチ セッションで表示されるカテゴリのセットです。手動で、またはカテゴリの作成または展開のオプションを使用して、これらのカテゴリを調整および改善できます。
- [読み込み] をクリックして、テキスト分析パッケージの内容をノードにコピーします。ダイアログ ボックスが閉じます。TAP が読み込まれると、TAP のコピーがノードにコピーされます。そのため、リソースおよびカテゴリに行った変更は、明示的に更新および再読み込みしないかぎり TAP に反映されます。

テキスト分析パッケージの更新

カテゴリセット、言語リソースを改善するか、新しいカテゴリ セットを作成する場合、テキスト分析パッケージ (TAP) を更新して、これらの改善点を後で再利用しやすくすることができます。TAP を更新するには、TAP に追加したい情報を含むオープン セッションで作業している必要があります。更新する場合、カテゴリ セットの追加、リソースの置き換え、パッケージ ラベルの変更、またはカテゴリ セットの名前変更/並べ替えを選択できます。

図 10-30
[テキスト分析パッケージの更新] ダイアログ



テキスト分析パッケージを更新するには

- メニューで、[ファイル] → [テキスト分析パッケージ] → [パッケージを更新] を選択します。[テキスト分析パッケージの更新] ダイアログが表示されます。
- 更新したいテキスト分析パッケージを含むディレクトリを参照します。
- [ファイル名] フィールドに TAP の名前を入力します。
- TAP 内の言語リソースと現在のセッションの言語リソースと置き換えるには、[このパッケージのリソースを使用中のセッションのリソースで置き換える] オプションを選択します。通常、言語リソースはカテゴリ定義の作成に使用される主要キーワードおよびパターンを抽出するために使用されるため、言語リソースの更新は重要です。最新の言語リソースがあれば、レコードのカテゴリ化において最善の結果を得ることができます。このオプションを選択しない場合、パッケージ内にすでにある言語リソースは変更されません。

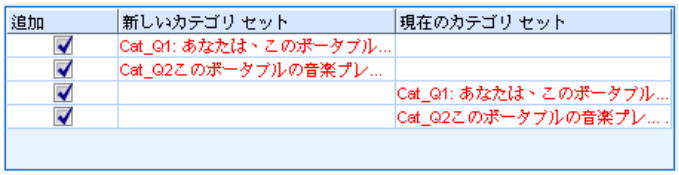
- 言語リソースのみを更新するには、かならず **[このパッケージのリソースを使用中のセッションのリソースで置き換える]** オプションを選択し、TAP 内にすでにある現在のカテゴリ セットのみを選択してください。

- 使用中のセッションの新しいカテゴリ セットを TAP に追加するには、追加する各カテゴリのチェックボックスをオンにしてください。1 つ、複数のカテゴリ セットを追加できますし、追加しなくてもかまいません。

- TAP からカテゴリ セットを削除するには、対応する **[追加]** チェックボックスをオンにします。改善されたカテゴリ セットを追加するため、すでに TAP にあるカテゴリ セットを削除する場合もあります。カテゴリ セットを削除するには、[現在のカテゴリ セット] 列の該当するカテゴリ セットの **[追加]** チェックボックスをオフにします。TAP には、1 つ以上のカテゴリ セットが必要です。

- 必要に応じて、カテゴリ セットの名前を変更します。セルを 1 回クリックすると、名前を編集できます。入力して他の場所をクリックすると、名前の変更が適用されます。カテゴリ セットの名前を変更すると、TAP のみの名前が変更され、オープン セッションの編集名は変わりません。2 つのカテゴリ セットに同じ名前が付いている場合、重複を修正するまで名前が赤で表示されます。

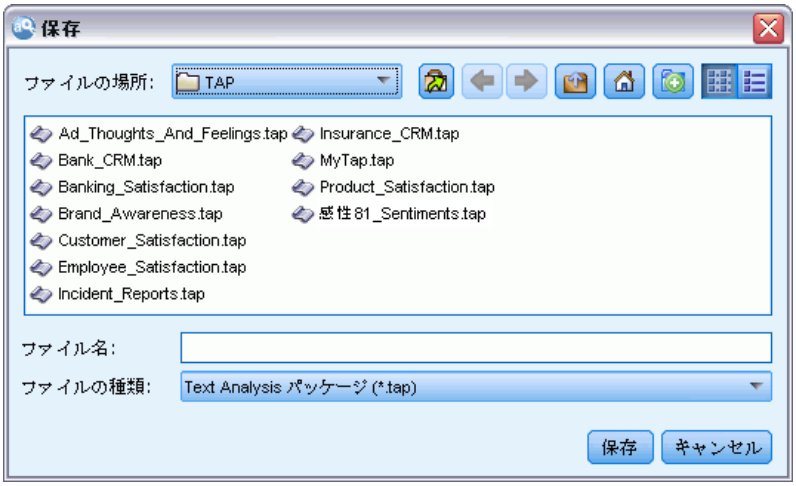
図 10-31
重複する名前



- 選択した TAP の内容と結合されたセッションの内容で新しいパッケージを作成するには、**[名前をつけて保存]** をクリックします。[テキスト分析パッケージの名前を付けて保存] ダイアログが表示されます。次の説明を参照してください。

- **[更新]** をクリックすると、選択した TAP に行われた変更が保存されます。

図 10-32
[テキスト分析パッケージの名前を付けて保存] ダイアログ



テキスト分析パッケージを保存するには

- TAP ファイルを保存するディレクトリを参照します。デフォルトでは、TAP ファイルはインストール ディレクトリの TAP サブディレクトリに保存されます。
- [ファイル名] フィールドに TAP ファイルの名前を入力します。
- [パッケージ ラベル] フィールドにラベルを入力します。ファイル名を入力すると、この名前がラベルとして自動的に表示されます。ただし、このラベルの名前は変更できます。ラベルはかならず指定する必要があります。
- [保存] をクリックして新しいパッケージを作成します。

カテゴリの編集および調整

カテゴリを作成すると、それらを検証して、何らかの調整を行う必要が常にあります。 言語リソースを調整するほか、定義を結合またはクリーンアップしたり、カテゴリ化されたドキュメントまたはレコードの一部をチェックするための方法を探すことによってカテゴリを確認する必要があります。 また、ニュアンスや特徴が分かるようにカテゴリが定義されるよう、カテゴリのドキュメントまたはレコードの確認を行うこともできます。

ビルトインで、自動のカテゴリ作成手法を使用して、カテゴリを作成できますが、これらのカテゴリに何らかの調整が必要な場合がよくあります。 1 つまたは複数の手法を使用した後、多くの新規カテゴリがウィンドウに表示されます。 カテゴリ定義が適切なものとなるまで、カテゴリのデータを確認して調整を行うことができます。 詳細は、 p. 188 カテゴリとは を参照してください。

カテゴリ調整のオプションがいくつかあります。その多くについては次のページで説明します。

- 記述子のカテゴリへの追加
- カテゴリの編集
- カテゴリの移動
- 階層カテゴリのフラット化
- カテゴリの結合
- カテゴリの削除
- 言語リソースの変更および再抽出
- カテゴリをどのように組み合わせるかを視覚化して調整 詳細は、13 章 p. 278 カテゴリ グラフおよび図表 を参照してください。

記述子のカテゴリへの追加

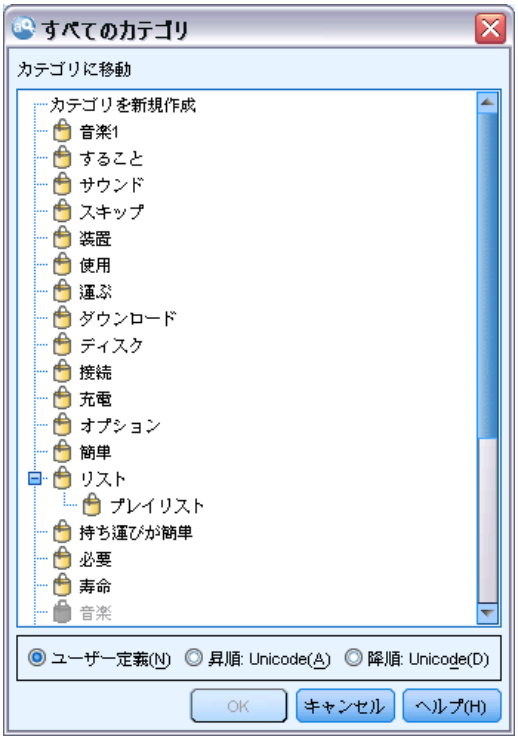
自動的手法を使用した後に、カテゴリ定義のいずれにも使用されなかった抽出結果が作成される場合があります。 結果のリストを［抽出結果］ウィンドウで確認する必要があります。 カテゴリに含ませたい要素があった場合、これを既存のあるいは新規のカテゴリに追加します。

コンセプトまたはタイプをカテゴリに追加するには

▶ ［抽出結果］パネルおよび［データ］パネルから、新規または既存のカテゴリに追加する要素を選択します。

▶ メニューから、**[カテゴリ] → [カテゴリに追加]** を選択します。 ［すべてのカテゴリ］ダイアログボックスにカテゴリのセットが表示されます。 選択した要素を追加したいカテゴリを選択します。 要素を新規カテゴリに追加する場合、**[新規カテゴリ]** を選択します。 最初に選択した要素を使用した新しいカテゴリが、［カテゴリ］パネルに表示されます。

図 10-33
[すべてのカテゴリ] ダイアログ ボックス



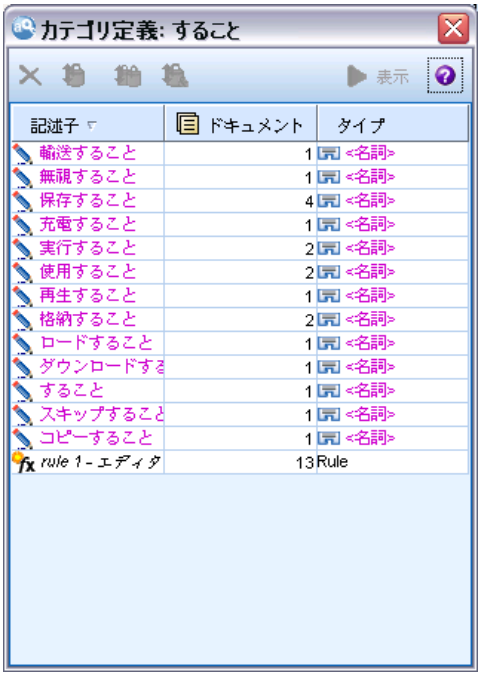
カテゴリ記述子の編集

いくつかのカテゴリを作成すると、各カテゴリを開いてその定義を構成するすべての記述子を表示できます。［カテゴリ定義］ダイアログ ボックスで、カテゴリ記述子にあらゆる編集を行うことができます。 また、カテゴリがカテゴリ ツリーに表示されている場合、ここで処理することもできます。

カテゴリを編集するには

- ［カテゴリ］パネルで編集したいカテゴリを選択します。
- メニューから［**表示**］→［**カテゴリ定義**］を選択します。［カテゴリ定義］ダイアログ ボックスが開きます。

図 10-34
[カテゴリ定義] ダイアログ ボックス



▶ 編集したい記述子を選択し、該当するツールバー ボタンをクリックします。

次の表で、カテゴリ定義を編集できるツールバーボタンについて説明します。

テーブル 10-13
ツールバー ボタンおよび説明

アイコン	説明
	選択した記述子をカテゴリから削除します。
	選択した記述子を新規または既存のカテゴリに移動します。
	選択した記述子を & カテゴリ規則の形式でカテゴリに移動します。詳細は、 p. 219 カテゴリ規則の使用 を参照してください。
	選択した各記述子を、独自の新規カテゴリとして移動します。
 表示	選択した記述子に従って、[データ] パネルおよび [視覚化] パネルの表示内容を更新します。

カテゴリの移動

カテゴリを別の既存カテゴリに投入または記述子を別のカテゴリに移動したい場合、カテゴリを移動できます。

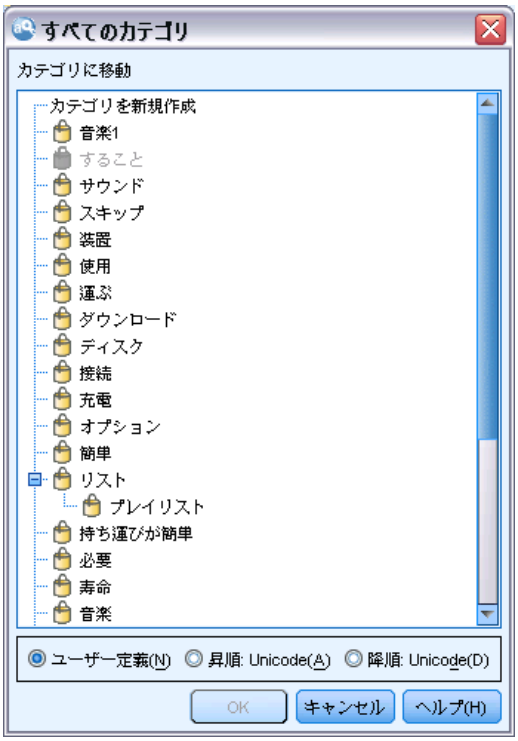
カテゴリを移動するには

▶ ［カテゴリ］パネルで、別のカテゴリに移動したいカテゴリを選択します。

▶ メニューから、**[カテゴリ] → [カテゴリに移動]** を選択します。 メニューに一連のカテゴリが表示され、リストの上部には最近作成されたカテゴリが表示されます。 選択したコンセプトを移動したいカテゴリ名を選択します。

- 探している名前が表示されたら、その名前を選択します。選択した要素がそのカテゴリに追加されます。
- 名前が表示されない場合、**[もっと表示]** を選択すると［すべてのカテゴリ］ダイアログ ボックスが表示され、リストからカテゴリを選択します。

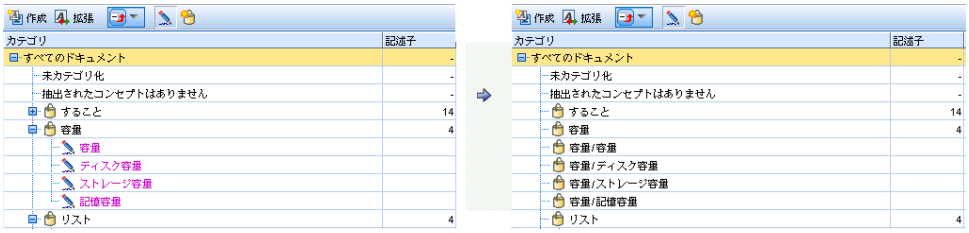
図 10-35
[すべてのカテゴリ] ダイアログ ボックス



カテゴリのフラット化

カテゴリおよびサブカテゴリを持つ階層カテゴリ構造がある場合、構造をフラットにすることができます。 カテゴリをフラット化する場合、そのカテゴリのサブカテゴリのすべての記述子が選択されたカテゴリに移動し、空のサブカテゴリが削除されます。 このように、サブカテゴリへの合致に使用されるすべてのドキュメントは、選択したカテゴリにカテゴリ化されます。

図 10-36
フラット化カテゴリ



カテゴリをフラット化するには

- ［カテゴリ］パネルで、フラットにするカテゴリ（上位レベルまたはサブカテゴリ）を選択します。
- メニューから、**[カテゴリ] → [カテゴリをフラット化]** を選択します。 サブカテゴリが削除され、記述子が選択したカテゴリに結合されます。

カテゴリの結合・組み合わせ

2 つ以上の既存カテゴリを 1 つの新規カテゴリに結合したい場合、それらを結合できます。 カテゴリを結合する場合、一般名の付いた新規カテゴリが作成されます。 カテゴリ記述子に使用されているすべてのコンセプト、タイプ、およびパターンがこの新規カテゴリに移動します。 このカテゴリ名は、カテゴリのプロパティを編集することで後から変更できます。

カテゴリまたはカテゴリの一部を結合するには

- ［カテゴリ］パネルで、結合したい要素を選択します。
- メニューから、**[カテゴリ] → [カテゴリの結合]** を選択します。 カテゴリ プロパティ ダイアログボックスが表示されるので、新しく作成したカテゴリの名前を入力します。 選択したカテゴリがサブカテゴリとして新しいカテゴリに結合されます。

カテゴリの削除

カテゴリを保持しない場合、削除することができます。

カテゴリを削除するには

- ［カテゴリ］パネルで、削除したいカテゴリを選択します。
- メニューから **[編集] → [削除]** を選択します。

章

11

クラスタの分析

クラスタ ビューで、コンセプトのクラスタを作成および検討できます（[表示] → [クラスタ]）。**クラスタ**は、ドキュメント/レコード セットでこれらのコンセプトが出現する頻度、および**共起**とも呼ばれる、同じドキュメントで同時に出現する頻度に基づいてアルゴリズムをクラスタリングすることによって生成される関連コンセプトのグループです。クラスタ内の各コンセプトは、クラスタ内の 1 つ以上の他のコンセプトと共に出現します。カテゴリの目的は、含まれるテキストが各カテゴリの記述子（コンセプト、条件規則、パターン）にどのように合致するかに基づいてドキュメントまたはレコードをグループ化することですが、クラスタの目的は共起するコンセプトをグループ化することです。

良いクラスタとは、リンクが強く頻繁に共起するコンセプトを含み、他のクラスタのコンセプトへのリンクが少ないクラスタです。大きなデータセットを扱う場合、この手法の処理時間が大幅に長くなる場合があります。

注:[クラスタを作成] ダイアログ ボックスの **[クラスタの計算に使用する最大ドキュメント数]** オプションを使用して、すべてのドキュメントまたはレコードの部分集合のみで作成します。

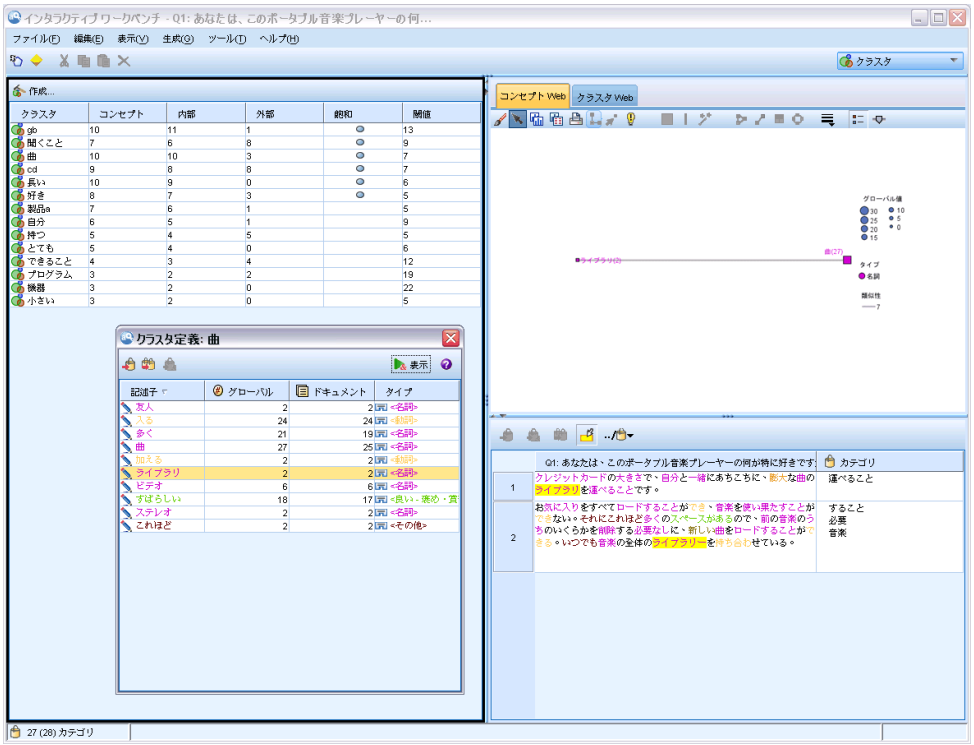
クラスタリングは、コンセプトのセットを分析し、ドキュメントで頻繁に共起するコンセプトを探すことから始まります。ドキュメント内で共起する 2 つのコンセプトは、コンセプト ペアと見なされます。次に、クラスタリング プロセスで、ペアが同時に出現するドキュメント数を各コンセプトが出現するドキュメント数と比較して、各コンセプト ペアの**類似度値**を評価します。 詳細は、 p. 264 類似度リンク値の計算 を参照してください。

最後に、リンク値と［クラスタを作成］ダイアログ ボックスで定義された設定を集約および考慮し、類似したコンセプトをグループ化します。集約とは、コンセプトを追加、またはクラスタが飽和するまで小さいクラスタを大きいクラスタに結合することです。コンセプトまたは小さいクラスタのさらなる結合によってクラスタが［クラスタを作成］ダイアログ ボックスの設定（コンセプト、内部リンク、外部リンクの数）を超えると、クラスタが**飽和**します。クラスタは、クラスタ内の他のコンセプトへのリンク数全体が最も大きいクラスタ内のコンセプトの名前を使用します。

別のクラスタにより強いリンクがある場合、そして飽和によってコンセプトが出現するクラスタの結合が行われない場合があるため、同じクラスタのすべてのコンセプト ペアが同時に出現するとはかぎりません。このため、内部リンクと外部リンクの両方が存在します。

- **内部リンク**は、クラスタ内のコンセプト ペア間のリンクです。すべてのコンセプトがクラスタ内のお互いのコンセプトにリンクしているわけではありません。ただし、各コンセプトは、クラスタ内の 1 つ以上の他のコンセプトにリンクしています。
- **外部リンク**は、別のクラスタのコンセプト ペア間のリンクです（あるクラスタのコンセプトと別のクラスタのコンセプトとの間）。

図 11-1
クラスタ ビュー



クラスタ ビューは次の 3 つのパネルで構成され、[表示] メニューから名前を選択して隠したり表示したりできます。

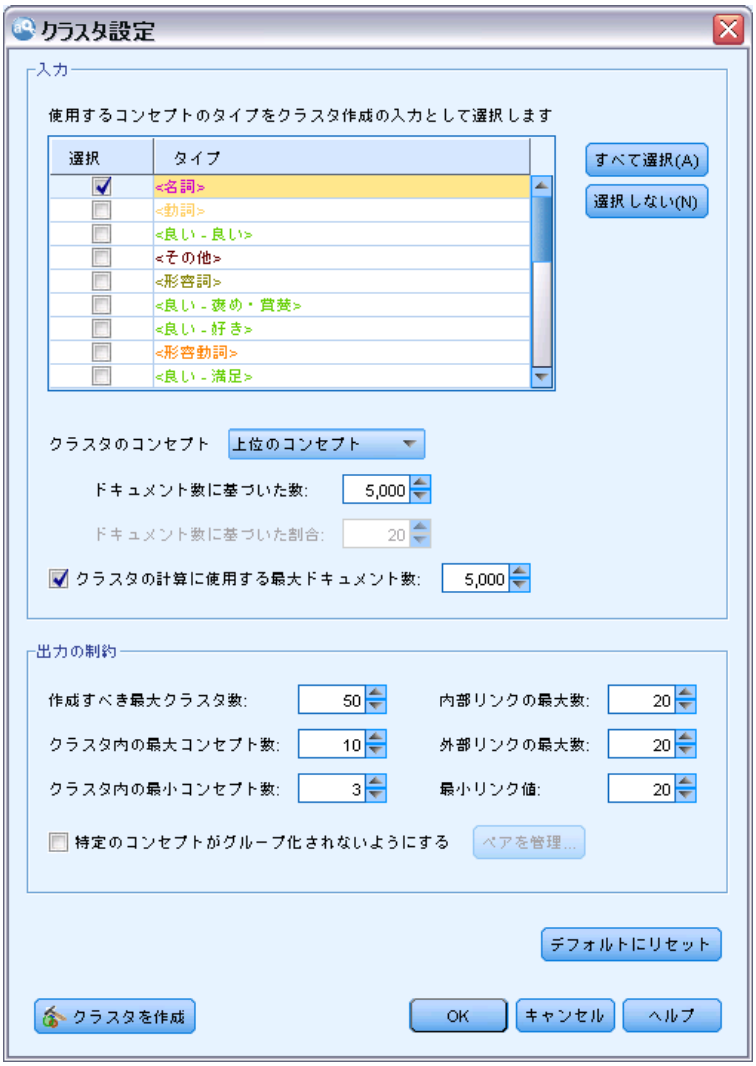
- **[クラスタ] パネル**: このパネルでカテゴリを作成し、管理できます。 詳細は、 p. 265 クラスタの検証 を参照してください。
- **[視覚化] パネル**: このパネルでカテゴリについて、またカテゴリがどのように相互作用するかを視覚的に検証できます。 詳細は、 13 章 p. 282 クラスタ グラフ を参照してください。
- **[データ] パネル**: このパネルでの選択に対応するドキュメントおよびレコード内に含まれるテキストを検証および確認できます。 詳細は、 p. 266 クラスタ定義 を参照してください。

クラスタの作成

クラスタ ビューを初めて開くと、クラスタは表示されません。メニュー（[**ツール**] → [**クラスタを作成**]）を使用するか、ツールバーの [**作成...**] をクリックして、クラスタを作成できます。 [**クラスタを作成**] ダイアログ ボックスが開き、クラスタ作成の設定および制約を定義できます。

注:抽出結果がリソースに一致しない場合、このパネルが［抽出結果］パネルと同じように黄色で表示されます。再抽出を実行して、最新の抽出結果を取得すると、黄色の表示は解除されます。ただし、抽出が実行されるごとに［クラスタ］パネルは解除され、クラスタを再作成する必要があります。同様に、あるセッションのクラスタは別のセッションに保存されません。

図 11-2
[クラスタを作成] ダイアログ ボックス



入力

[入力] テーブル:クラスタは、特定のタイプから派生した記述子から作成されます。テーブル内で、作成プロセスに使用するタイプを選択できます。デフォルトでは、最も多くのレコードまたはドキュメントをキャプチャするタイプが事前に選択されています。

クラスタのコンセプト: クラスタリングに使用するコンセプトの選択方法を選択します。コンセプトの数を減らすと、クラスタリング プロセスの速度を向上させることができます。さまざまな上位コンセプト、上位コン

セプトの割合を使用して、またはすべてのコンセプトを使用してクラスタリングを行うことができます。

- **ドキュメント数に基づいた数:**[上位のコンセプト] を選択した場合、クラスタリングに考慮するコンセプトの数を入力します。コンセプトは、ドキュメント数の値が最も大きいコンセプトに基づいて選択します。ドキュメント数は、コンセプトが出現するドキュメントまたはレコードの数です。
- **ドキュメント数に基づいた割合:**[上位コンセプトの割合] を選択した場合、クラスタリングに考慮するコンセプトの割合を入力します。コンセプトは、ドキュメント数の値が最も大きいコンセプトの割合に基づいて選択します。

クラスタの計算に使用する最大ドキュメント数: デフォルトでは、リンク値はドキュメントまたはレコードのセット全体を使用して計算します。ただし、リンクの計算に使用するドキュメントまたはレコードの数を制限して、クラスタリング プロセスの速度の向上が必要な場合があります。ドキュメント数を制限すると、クラスタの品質が低下する場合があります。このオプションを使用するには、左側のチェック ボックスをオンにして、使用するドキュメントまたはレコードの最大数を入力します。

出力の制約

作成すべき最大クラスタ数: この値は、生成し、[クラスタ] パネルに表示される最大クラスタ数です。クラスタリング プロセスで、飽和したクラスタは不飽和クラスタの前に表示されます。つまり、生成される多くのクラスタが飽和します。より多くの不飽和クラスタを表示するために、この設定を飽和クラスタの数より大きい値に変更できます。

クラスタ内の最大コンセプト数: この値は、クラスタが含むことができる最大コンセプト数です。

クラスタ内の最小コンセプト数: この値は、クラスタを作成するためにリンクする必要のある最小コンセプト数です。

内部リンクの最大数: この値は、クラスタが含むことができる内部リンクの最大数です。内部リンクは、クラスタ内のコンセプト ペア間のリンクです。

外部リンクの最大数: この値は、クラスタ外部のコンセプトへのリンクの最大数です。外部リンクは、別のクラスタのコンセプト ペア間のリンクです。

最小リンク値: この値は、クラスタリングに考慮されるコンセプト ペアに受け入れられる最小リンク値です。リンク値は、類似度評価式を使用して計算します。 詳細は、 p. 264 類似度リンク値の計算 を参照してください。

特定のコンセプトがグループ化されないようにする: 出力の 2 つのコンセプトがグループ化またはペアとならないように処理を停止します。コンセプト ペアを作成または管理するには、[ペアを管理] をクリックします。 詳細は、 10 章 p. 202 例外ペアのリンクの管理 を参照してください。

類似度リンク値の計算

コンセプト ペアが出現するドキュメント数がわかっているだけでは、2 つのコンセプトがどの程度類似しているかはわかりません。この場合、類似度値が役に立ちます。類似度リンク値は、関連性における各コンセプトの共起ドキュメント数を個別ドキュメント数に比較して測定します。類似度を計算する場合、測定の単位はコンセプトまたはコンセプト数が見つかったドキュメント数です。コンセプトまたはコンセプト ペアが「少なくとも」 1 回ドキュメント内に出現した場合、コンセプトまたはコンセプト ペアがドキュメント内で「見つかった」といえます。コンセプト グラフのラインの太さを、グラフの類似度リンク値を示すよう選択できます。

アルゴリズムを使用して、最も強いこれらの関連性を明らかにします。つまり、コンセプトがテキスト データで同時に出現する傾向は、個別に出現する傾向より高くなります。内部的に、アルゴリズムは 0 ～ 1 の類似度係数を生成します。1 の値は 2 つのコンセプトが常に同時に出現し、個別には出現しないことを意味します。類似度係数の結果に 100 をかけ、最も近い整数に丸められます。類似度係数は、次の図に示された式を使用して計算されます。

図 11-3
類似度係数の式

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

この場合、次のようになります。

- C_I は、コンセプト I が出現するドキュメントまたはレコードの数です。
- C_J は、コンセプト J が出現するドキュメントまたはレコードの数です。
- C_{IJ} は、コンセプトのペア I および J が同時に出現するドキュメントまたはレコードの数です。

たとえば、5,000 件のドキュメントがあるとします。I および J が抽出されたコンセプト、IJ が I および J のコンセプト ペアの共起であるとします。次の表では、係数とリンク値を計算する方法を示す 2 つのシナリオを示しています。

テーブル 11-1
コンセプトの頻度の例

コンセプト/ペア	シナリオ A	シナリオ B
コンセプト:I	20 件のドキュメントに出現	30 件のドキュメントに出現
コンセプト:J	20 件のドキュメントに出現	60 件のドキュメントに出現
コンセプト ペア:IJ	20 件のドキュメントに共起	20 件のドキュメントに共起

コンセプト/ペア	シナリオ A	シナリオ B
類似度係数	1	0.22222
類似度リンク値	100	22

シナリオ A の場合、コンセプト I と J、および IJ は 20 件のドキュメントに出現します。この場合類似度係数は 1 となり、コンセプトは常に同時に出現します。このペアの類似度リンク値は 100 となります。

シナリオ B の場合、コンセプト I は 30 件のドキュメントに出現し、コンセプト J は 60 件のドキュメント、ペア IJ は 20 件のドキュメントにのみ出現します。その結果、類似度係数は 0.22222 となります。このペアの類似度リンク値は、丸められて 22 となります。

クラスタの検証

クラスタを作成した後、[クラスタ] パネルで一連の結果を確認できます。各クラスタについて、次の情報がテーブルに表示されます。

- **クラスタ**: クラスタの名前です。クラスタは、内部リンク数が最も多いコンセプトから名前を付けられます。
- **コンセプト**: クラスタ内のコンセプト数です。 詳細は、 p.266 クラスタ定義 を参照してください。
- **内部**: クラスタ内の内部リンクの数です。内部リンクは、クラスタ内のコンセプト ペア間のリンクです。
- **外部**: クラスタ内の外部リンクの数です。外部リンクは、あるクラスタのコンセプトと、別のクラスタのコンセプトとのコンセプト ペア間のリンクです。
- **飽和**: 記号が表示されている場合、このクラスタが大きく、1 つまたは複数の制約を超えていることを示します。そのため、そのクラスタのクラスタリング プロセスは終了し、「飽和」していると見なされます。クラスタリング プロセスの終了時、飽和したクラスタは不飽和クラスタの前に表示されます。つまり、生成される多くのクラスタが飽和します。より多くの不飽和クラスタを表示するためには、[**作成すべき最大クラスタ数**] の設定を飽和クラスタの数より大きい値に変更するか、[**最小リンク値**] の値を小さくできます。 詳細は、 p.261 クラスタの作成 を参照してください。
- **閾値**: クラスタ内のすべての共起コンセプト ペアについて、クラスタで最も低い類似度リンク値です。詳細は、 p.264 類似度リンク値の計算 を参照してください。 閾値の最も大きいクラスタは、そのクラスタのコンセプトの全体の類似度が高く、閾値が小さいクラスタのコンセプトより密接に関連していることを示します。

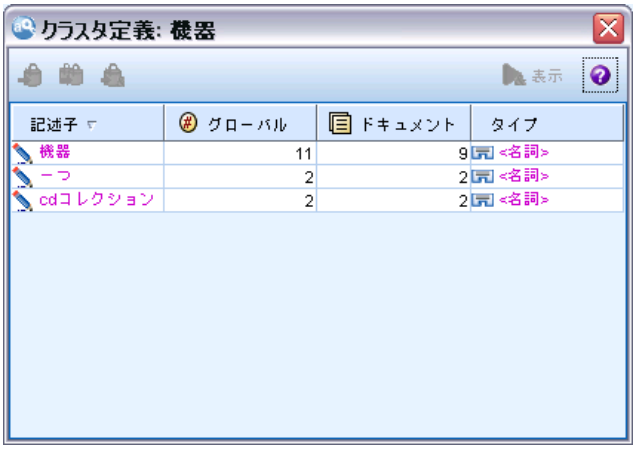
指定されたクラスタについての詳細を知るには、クラスタを選択すると、右側の視覚化パネルにクラスタを検証するための 2 つのグラフが表示されます。詳細は、 13 章 p. 282 クラスタ グラフ を参照してください。 テーブルの内容を切り取り、別のアプリケーションに貼り付けることができます。

抽出結果がリソースに一致しない場合、このパネルが［抽出結果］パネルと同じように黄色で表示されます。再抽出を実行して、最新の抽出結果を取得すると、黄色の表示は解除されます。ただし、抽出が実行されるごとに［クラスタ］パネルは解除され、クラスタを再作成する必要があります。同様に、あるセッションのクラスタは別のセッションに保存されません。

クラスタ定義

［クラスタ］パネルでクラスタをせんたくし、［クラスタ定義］ダイアログ ボックスを開くと、クラスタ内のすべてのコンセプトが表示されます（**［表示］→［クラスタ定義］**）。

図 11-4
［クラスタ定義］ダイアログ ボックス



選択したクラスタのすべてのコンセプトが［クラスタ定義］ダイアログ ボックスに表示されます。［クラスタ定義］ダイアログ ボックスの 1 つまたは複数のコンセプトを選択し、**［表示］**をクリックすると、［データ］ウィンドウに「選択したすべてのコンセプトがいっしょに出現する」すべてのレコードまたはドキュメントが表示されます。ただし、［クラスタ］パネルでクラスタを選択した場合、［データ］パネルにはテキスト レコードまたはドキュメントは表示されません。［データ］パネルに関する一般情報については、10 章の「［データ］パネル」を参照してください。

このダイアログ ボックスでコンセプトを選択すると、コンセプト Web グラフも変わります。詳細は、 13 章 p. 282 クラスタ グラフ を参照してください。 同様に、［クラスタ定義］ダイアログ ボックスで 1 つまたは

複数のコンセプトを選択すると、[視覚化] パネルにこれらのコンセプトの外部リンクおよび内部リンクがすべて表示されます。

列の説明

各記述子を容易に特定できるよう、アイコンが表示されます。

テーブル 11-2
列および記述子アイコン

列	説明
記述子	コンセプトの名前
グローバル	データセット全体にこの記述子が出現する回数を示します。グローバル出現頻度とも呼ばれます。
ドキュメント	この記述子が出現するドキュメントまたはレコードの数を示します。ドキュメント数とも呼ばれます。
Type	記述子が属するタイプを示します。記述子がカテゴリ規則である場合、この列にタイプ名は表示されません。

ツールバーの操作

このダイアログ ボックスから、カテゴリで使用する 1 つまたは複数のコンセプトを選択することもできます。コンセプトを選択する方法はいくつかありますが、クラスタ内で共起するコンセプトを選び、カテゴリ規則としてそれらを追加することが最も興味深い方法です。詳細は、 10 章 p. 209 共起規則 を参照してください。 ツールバー ボタンを使用して、コンセプトをカテゴリに追加できます。

テーブル 11-3
コンセプトをカテゴリに追加するツールバー ボタン

アイコン	説明
	選択したコンセプトを新規または既存のカテゴリに追加します。
	選択したコンセプトを新規または既存のカテゴリに & カテゴリ規則の形式で追加します。 詳細は、 10 章 p. 219 カテゴリ規則の使用 を参照してください。
	選択した各コンセプトを、独自の新規カテゴリとして追加します。
 表示 &	選択した記述子に従って、[データ] パネルおよび [視覚化] パネルの表示内容を更新します。

注:コンテキスト メニューを使用して、コンセプトをタイプに、類義語、または不要語項目として追加することもできます。

章
12
テキスト リンク分析の検証

テキスト リンク分析（TLA）ビューでは、テキスト リンク分析パターンを検証できます。テキスト リンク分析はパターンマッチ手法で、パターン規則を定義し、それらをテキスト内の実際の抽出されたコンセプトおよび関連性と比較することができます。

たとえば、組織に関するキーワードを抽出しても、重要でない場合があります。TLA を使用して、この組織と他の組織、または組織内の人々の間のリンクについて学習することができます。TLA を使用して、製品に関する意見、または遺伝子間の関連性についていくつかの言語で抽出することもできます。

TLA パターン結果を抽出すると、テキスト リンク分析ビューの［タイプ パターン］パネルまたは［コンセプト パターン］パネルでそれらを確認できます。詳細は、 p.270 タイプ パターンおよびコンセプト パターン を参照してください。 このビューの［データ］パネルまたは［視覚化］パネルで TLA パターン結果をさらに検証できます。そしておそらく最も重要なことですが、TLA パターン結果をカテゴリに追加できます。

また TLA パターンの抽出を選択していない場合、**［抽出］**をクリックして、［抽出設定］ダイアログ ボックスで **［テキスト リンク分析のパターン抽出を有効にする］**を選択できます。 詳細は、 p.269 TLA パターン結果の抽出 を参照してください。

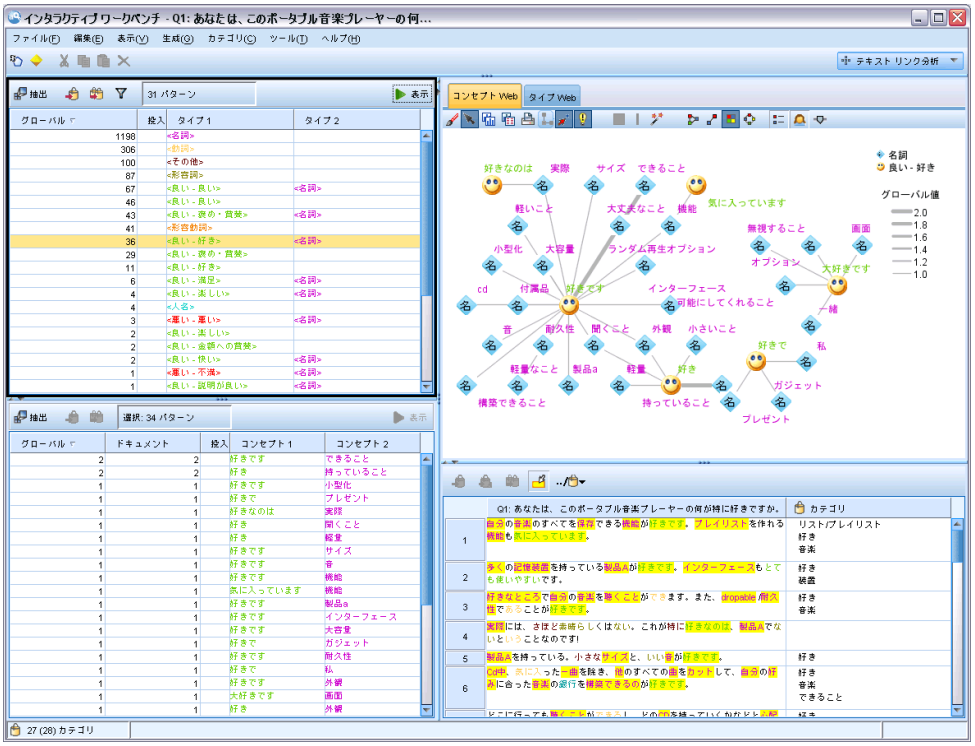
いくつかの TLA パターン規則が、TLA パターン結果を抽出するために使用するリソース テンプレートまたはライブラリで定義されています。IBM® SPSS® Modeler Text Analytics に付属する特定のリソース テンプレートで TLA パターンを使用できます。抽出できる関連性およびパターンの種類は、全体的にリソースで定義された TLA 規則によって異なります。日本語以外のすべてのテキスト言語の独自の TLA 規則を定義できます。パターンは、マクロ、単語リスト、およびブール型質問を形成する単語の空所、または入力テキストと比較される条件規則で構成されています。 詳細は、19 章 p.368 テキスト リンク規則について を参照してください。

TLA パターン規則がテキストに一致する場合、このテキストをパターンとして抽出し、出力データとして再構築できます。そして結果は、テキスト リンク分析ビューのパネルで表示されます。［表示］メニューでパネルの名前を選択して、各パネルを隠したり表示することができます。

- **［タイプ パターン］パネルおよび［コンセプト パターン］パネル:** これら 2 つのパネルでパターンを作成し、検証できます。 詳細は、 p.270 タイプ パターンおよびコンセプト パターン を参照してください。

- **[視覚化] パネル:** このパネルで、パターンのコンセプトおよびタイプがどのように相互作用するかを視覚的に検証できます。 詳細は、 13 章 p. 285 テキスト リンク分析のグラフ を参照してください。
- **[データ] パネル:** 別のパネルでの選択に対応するドキュメントおよびレコード内に含まれるテキストを検証および確認できます。 詳細は、 p. 275 [データ] パネル を参照してください。

図 12-1
テキスト リンク分析ビュー



TLA パターン結果の抽出

抽出プロセスにより、一連のコンセプト、タイプ、そして有効な場合はテキスト リンク分析 (TLA) パターンが作成されます。TLA パターンを抽出した場合、これらはテキスト リンク分析ビューにされます。抽出結果がリソースと同期していない場合、[パターン] パネルが黄色で表示され、再抽出をすると異なる結果が生成されることを示します。

[テキスト リンク分析のパターン抽出を有効にする] オプションを使用して、ノードの設定または [抽出] ダイアログ ボックスでこれらのパターンの抽出を選択する必要があります。 詳細は、 9 章 p. 156 データの抽出 を参照してください。

注:データセットのサイズと、抽出プロセスを完了するためにかかる時間の間には、関連性があります。パフォーマンス統計および推奨事項については、インストール手順を参照してください。上流にサンプル ノードを追加、またはコンピュータの構成を最適化することをいつでも検討することができます。

データを抽出するには

- メニューの [**ツール**] > [**抽出**] を選択します。または、[**抽出**] ツールバー ボタンをクリックします。
- 使用するオプションを変更します。このタブで TLA パターン結果を抽出するには、オプション [**テキスト リンク分析のパターン抽出を有効にする**] を選択し、テンプレートに TLA 規則を使用する必要があります。 詳細は、 9 章 p. 156 データの抽出 を参照してください。
- [**抽出**] をクリックして、抽出プロセスを開始します。

  抽出が始まると、進捗状況のダイアログ ボックスが表示されます。抽出を中断する場合は、[**キャンセル**] をクリックします。抽出が完了すると、ダイアログ ボックスが閉じられ、結果がパネルに表示されます。 詳細は、 p. 270 タイプ パターンおよびコンセプト パターン を参照してください。

タイプ パターンおよびコンセプト パターン

パターンは、2 つの部分、コンセプトとタイプを組み合わせて構成されています。パターンは、特定のサブジェクトに関する意見またはコンセプト間の関連性を探索する場合に最も役立ちます。競合他社の製品名を抽出しても、重要でない場合があります。この場合、抽出したパターンを参照し、ドキュメントまたはレコードに、製品が良い、悪い、または高いことを示すテキストが含まれている例があるかどうかを確認することができます。

図 12-2
テキスト リンク分析ビュー:[タイプ パターン] パネルおよび [コンセプト パターン] パネル

パターンは最大 6 つのタイプまたは 6 つのコンセプトから構成されます。そのため、両方のパターンのワイン同枠の行には、最大 6 つのスロットまたは場所があります。各スロットは、言語リソースで定義されているように、TLA パターンの要素固有の場所に対応しています。インタラクティブ ワークベンチでは、スロットに値がない場合、スロットはテーブルに表示されません。たとえば、最も長いパターン結果に 4 つのスロット

がある場合、後半 2 つのスロットは表示されません。 詳細は、 19 章 p. 368 テキスト リンク規則について を参照してください。

パターン結果を抽出する場合、まずタイプ レベルでグループ化され、コンセプト パターンに分割されます。そのため、**[タイプ パターン]**（左上）および **[コンセプト パターン]**（左下）の 2 つの結果パネルが表示されます。返されたすべてのコンセプト パターンを表示するには、タイプ パターンをすべて選択します。一番下のコンセプト パターンのパネルには、順位の最大値（[フィルタ] ダイアログ ボックスで定義）までのコンセプト パターンがすべて表示されます。

タイプ パターン： TLA パターン規則を満たす 1 つまたは複数の関連タイプで構成されているパターン規則が表示されます。タイプ パターンは、<**組織名**> + <**地名**> + <**肯定的**> と表され、特定の場所の組織について、肯定的なフィードバックを提供します。シンタックスは、次のようになります。

<**タイプ** 1> + <**タイプ** 2> + <**タイプ** 3> + <**タイプ** 4> + <**タイプ** 5> + <**タイプ** 6>

コンセプト パターン: 上の［タイプ パターン］で現在選択されているすべてのタイプ パターンのコンセプト レベルでパターンの結果が表示されます。コンセプト パターンは、**ホテル** + **パリ** + **すばらしい** などの構造に従います。シンタックスは、次のようになります。

コンセプト 1 + **コンセプト** 2 + **コンセプト** 3 + **コンセプト** 4 + **コンセプト** 5 + **コンセプト** 6

パターン結果が 6 つ未満の最大スロットを使用する場合、必要な数だけのスロット（または列）が表示されます。<**タイプ** 1>+<>+<**タイプ** 2>+<>+<>+<> のように、2 つの入力されたスロットの間の空白のスロットは破棄され、<Type1>+<Type3> のようになります。コンセプト パターンの場合、**コンセプト** 1+.+**コンセプト** 2 となります（. は空値を示します）。

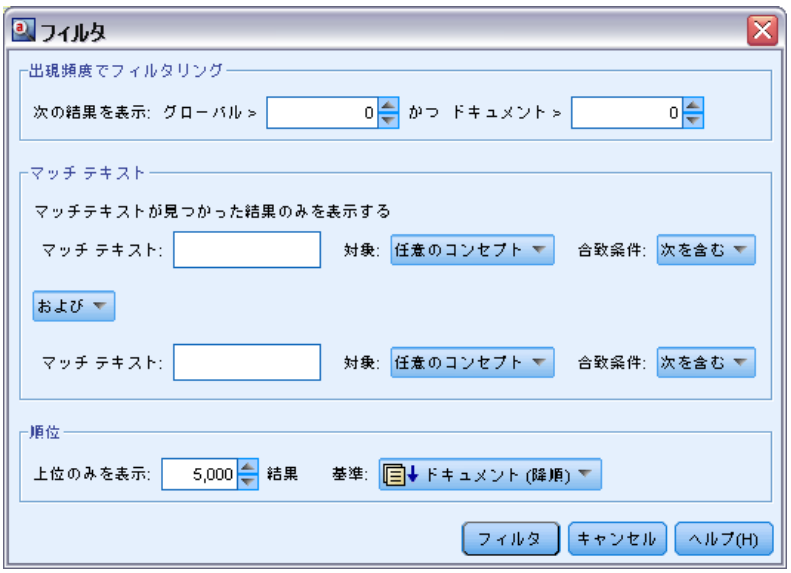
カテゴリとコンセプト ビューの抽出結果と同様、ここで結果を確認できます。これらのパターンを構成するタイプおよびコンセプトに調整を行う場合、カテゴリとコンセプト ビューの［抽出結果］パネルまたはリソース エディタで変更を行うか、パターンを再抽出します。カテゴリ定義でコンセプト、タイプ、またはパターンが使用されている場合、カテゴリまたは条件規則のアイコンが［パターン］テーブルまたは［抽出結果］テーブルの **[投入]** 列に表示されます。

TLA 結果のフィルタリング

非常に大きなデータセットを処理する場合、抽出プロセスでは、多数の結果が作成される場合があります。多くのユーザーによって、多数の結果が作成されると、結果を効率的に確認することが困難になります。ただし、これらの結果をフィルタリングして、最も関心の高い結果に焦点を当てることができます。［フィルタ］ダイアログ ボックスの設定を変

更して、表示されるパターンを制限できます。これらの設定はすべていっしょに使用されます。

図 12-3
[フィルタ] ダイアログ ボックス (TLA ビュー)



出現頻度でフィルタリング:フィルタリングを実行して、特定のグローバル出現頻度値またはドキュメントの出現頻度の値を持つ結果のみを表示できます。

- **グローバル出現頻度**は、パターンがドキュメントまたはレコードの全体的なセットに出現する回数の合計で、[**グローバル**] 列に表示されます。
- **ドキュメント出現頻度**は、パターンが出現するドキュメントまたはレコードの合計数で、[**ドキュメント**] 列に表示されます。

たとえば、あるパターンが 500 件のレコードに 300 回出現した場合、このパターンのグローバル出現頻度は 300 で、ドキュメント出現頻度は 500 となります。

マッチ テキスト: ここで定義する規則に一致する結果のみを表示できます。[**マッチ テキスト**] フィールドに合致する文字のセットを入力し、スロット番号またはそれらのすべての特定して、コンセプト名またはタイプ名のどちらでこのテキストを検索するかを選択します。合致を適用する条件を選択します（タイプ名の開始と終了を示す各カッコを使用する必要はありません)。ドロップダウン リストから [**かつ**] または [**または**] を選択して条件規則が両方の文またはいずれかに一致するようにし、最初の文と同じ方法で、2 番目のテキスト マッチ文を定義します。

テーブル 12-1
マッチ テキストの条件

条件	説明
次を含む	文字列が任意の場所で出現する場合、テキストが一致します (デフォルトの選択)。
次で開始	コンセプトまたはタイプが特定のテキストで始まる場合にのみ、テキストが一致します。
次で終了	コンセプトまたはタイプが特定のテキストで終わる場合にのみ、テキストが一致します。
完全一致	文字列全体が、コンセプト名またはタイプ名に一致する必要があります。

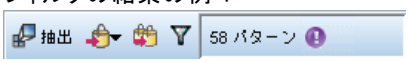
順位: フィルタリングして、グローバル出現頻度（[**グローバル**]）またはドキュメント頻度（[**ドキュメント**]）に従って、上位のパターンのみを昇順または降順で表示することができます、この順位の最大値は、表示に返されるパターンの合計数を制限します。

フィルタが適用されると、コンセプト パターンの最大合計数（最大順位）を超えるまで、タイプ パターンを追加します。まず最初に、順位が上位のタイプ パターンを参照し、対応するコンセプト パターンの合計を取得します。この合計が順位の最大値を超えていない場合、パターンがビューに表示されます。そして、次のタイプ パターンのコンセプト パターン数が合計されます。この数値に前のタイプ パターンのコンセプト パターンの合計数を加えた数値が順位の最大値より少ない場合、これらのパターンもビューに表示されます。これを、順位の最大値を超えることなく、できるかぎり多くのパターンが表示されるまで続行します。

[パターン] パネルに表示される結果

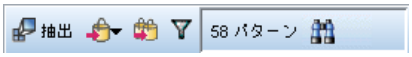
フィルタリングに基づいて、結果が［パターン］ウィンドウにどのように表示されるかについて、いくつか例を示します。

図 12-4
フィルタの結果の例 1



この例では、フィルタで指定された順位の最大値により、返されるパターンの数が制限されていることがツールバーに示されています。紫色のアイコンが表示されている場合、パターンの最大数に達していることを示します。アイコンの上にポインタを置くと、詳細が表示されます。[**順位**] フィルタに関する前述の説明を参照してください。

図 12-5
フィルタの結果の例 2



この例では、マッチ テキスト フィルタを使用して、結果が制限されていることがツールバーに示されています（虫めがねのアイコンを参照）。アイコンにポイントを置くと、マッチ テキストの内容が表示されます。

結果を絞り込むには

- メニューの [**ツール**] → [**フィルタ**] を選択します。［フィルタ］ダイアログ ボックスが開きます。
- 使用するフィルタを選択および調整します。
- [OK] をクリックするとフィルタが適用され、新しい結果が表示されます。

[データ] パネル

テキスト リンク分析パターンを抽出および検証すると、作業しているデータをいくつか確認したい場合があります。たとえば、パターンのグループが発見された実際のレコードを確認したい場合があります。右下の［データ］パネルでレコードまたはドキュメントを確認することができます。 デフォルトで表示されない場合は、メニューから [**表示**] → [**パネル**] → [**データ**] を選択してください。

［データ］パネルには、特定の表示制限に応じて、ビュー内の選択に該当するドキュメントまたはレコードごとに 1 行ずつ表示されます。デフォルトでは、［データ］パネルに表示されるドキュメントまたはレコード数が制限され、データをより迅速に表示できるようになります。ただし、これは［オプション］ダイアログ ボックスで調整できます。 詳細は、 8 章 p. 143 オプション：［セッション］タブ を参照してください。

[データ] パネルの表示および更新

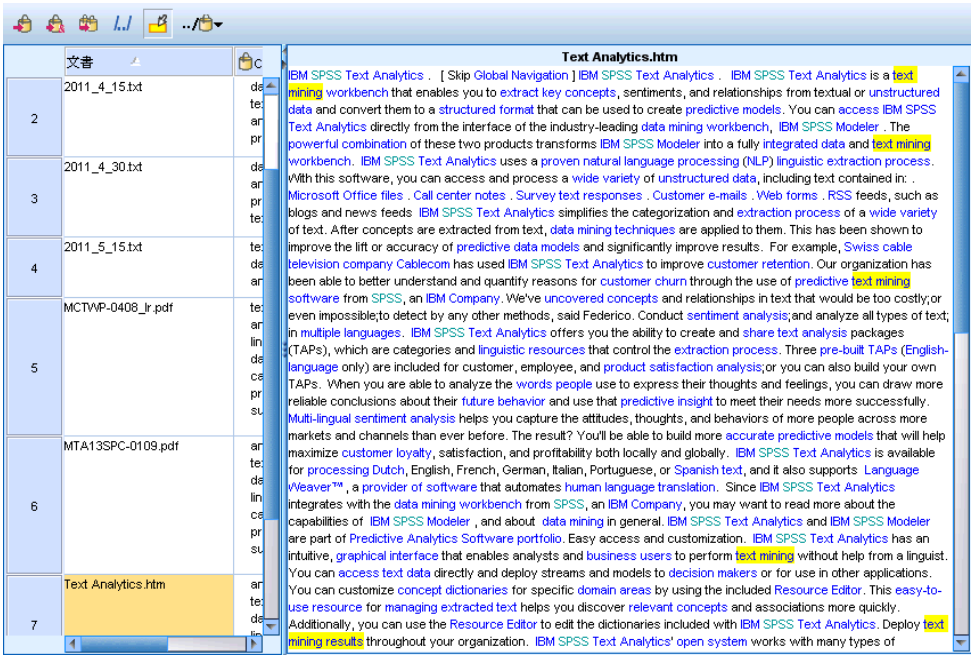
［データ］パネルでは、大きなデータセットの自動データ更新には時間がかかるため、自動的に表示の更新は行われません。そのため、このビューでタイプ パターンまたはコンセプト パターンを選択すると、[**表示**] をクリックして［データ］パネルの内容を更新できます。

テキスト ドキュメントまたはレコード

テキスト データがレコードの形式で、テキストの長さが比較的短い場合、［データ］パネルのテキスト フィールドには、テキスト データの全体が表示されます。ただし、レコードおよび大きいデータセットを処理している場合、テキスト フィールドの列にはテキストの一部が表示され、右側の［テキスト プレビュー］パネルを開くと、テーブルで選択したレコードの大部分またはすべてが表示されます。テキスト データが個別ドキュメントの形式の場合、［データ］パネルには、ドキュメントのファイル名が表示さ

れます。ドキュメントを選択すると、[テキスト プレビュー] パネルには選択したドキュメントのテキストが表示されます。

図 12-6
[テキスト プレビュー] パネルが表示された [データ] パネル



色および強調表示

データを表示すると、該当するドキュメントまたはレコードのコンセプトおよび記述子は色付きで強調表示され、テキスト内のコンセプトおよび記述子を特定しやすくします。カラー コードは、コンセプトが属するタイプに対応します。カラーコード化された項目上でマウス ポインタを停止させて、項目が抽出されたコンセプトと、項目が割り当てられたタイプを表示することもできます。抽出されていないテキストは、黒で表示されます。通常、こうした抽出されていない単語は接続詞（「および」または「と」）、代名詞（「私」または「彼ら」）および 動詞（「いる」、「持つ」、または「取る」）の場合が多くあります。

[データ] パネルの列

テキスト フィールドの列は常に表示されていますが、その他の列も表示できます。その他の列を表示するには、メニューで **[表示] → [[データ] パネル]** を選択し、[データ] パネルに表示したい列を選択します。 表示できるのは次の列です。

- **「テキスト フィールド名」(#)/ドキュメント**: コンセプトおよびタイプが抽出されたテキスト データの列を追加します。データがドキュメントにある場合、列名は［ドキュメント］となり、ドキュメント ファイル名または完全パスのみが表示されます。これらのドキュメントのテキストを表示するには、［テキスト プレビュー］パネルを表示する必要があります。［データ］パネルの行数が、この列名の後のカッコ内に表示されます。読み込みの速度向上のために使用される［オプション］ダイアログの制約により、一部のドキュメントまたはレコードが表示されない場合があります。最大値に達すると、数値の後に [- **最大**] と表示されます。詳細は、 8 章 p. 143 オプション：［セッション］タブ を参照してください。
- **カテゴリ**: レコードが属するカテゴリがそれぞれ表示されます。この列を表示する場合、最新の情報を示すため、［データ］パネルの更新に少し時間がかかる場合があります。
- **適合度順位**: 1 つのカテゴリの各レコードの順位が表示されます。この適合度順位は、カテゴリ内の他のレコードと比較して、レコードがカテゴリにどれだけ適合しているかを示します。［カテゴリ］パネル（左上のパネル）でカテゴリを選択すると、順位が表示されます。詳細は、10 章 p. 192 カテゴリの関連性 を参照してください。
- **カテゴリの個数**: レコードが割り当てられているカテゴリ数が表示されます。

章
13

グラフの視覚化

カテゴリとコンセプト ビュー、クラスタ ビュー、およびテキスト リンク分析ビューにはすべて、ウィンドウの右上隅に［視覚化］パネルがあります。このパネルを使用して、データを視覚的に検証することができます。次のグラフおよび図表を使用できます。

- **カテゴリとコンセプト ビュー:** このビューには、カテゴリ バー、 カテゴリ Web、および カテゴリ Web テーブルの 3 つのグラフ・図表があります。このビューでは、[**表示**] をクリックす場合にのみグラフが更新されます。 詳細は、 p.278 カテゴリ グラフおよび図表 を参照してください。
- **クラスタ ビュー:**このビューには、コンセプト Web グラフ および クラスタ Web グラフ の 2 つの Web グラフがあります。 詳細は、 p.282 クラスタ グラフ を参照してください。
- **テキスト リンク分析ビュー:**このビューには、コンセプト Web グラフ および タイプ Web グラフ の 2 つの Web グラフがあります。 詳細は、 p.285 テキスト リンク分析のグラフ を参照してください。

グラフの編集に使用する一般的なツールバーおよびパレットの詳細は、オンライン ヘルプまたはファイル SourceProcessOutputNodes.pdf の「グラフの編集」を参照してください。このファイルは、IBM® SPSS® ModelerDVD の ¥Documentation¥en フォルダにあります。

カテゴリ グラフおよび図表

カテゴリを作成する場合、時間をかけてカテゴリ定義、含まれるドキュメントまたはレコード、およびカテゴリの重複を確認することが重要になります。［視覚化］パネルには、カテゴリに関するいくつかの視点が表示されています。［視覚化］パネルは、カテゴリとコンセプト ビューの右上隅に表示されます。表示されない場合、［表示］メニュー（[**表示**] → [**パネル**] → [**視覚化**]）からこのパネルにアクセスできます。

このビューの［視覚化］パネルには、ドキュメントまたはレコードのカテゴリ化における共通性について 3 つの視点が表示されています。このパネルの図表やグラフを使用して、カテゴリ化の結果を分析したり、カテゴリまたはレポートの調整を行うことができます。カテゴリを調整する場合、このパネルを使用して、カテゴリ定義を確認し、あまりに類似している（ドキュメントまたはレコードの 75% 以上を共有しているなど）またはあまりに異なるカテゴリを明らかにできます。2 つのカテゴリがあまりに似ている

場合、2 つのカテゴリの結合することができます。また、一方のカテゴリから特定の記述子を削除して、カテゴリ定義の調整することもできます。

[抽出結果] パネル、[カテゴリ] パネルまたは [カテゴリ定義] ダイアログ ボックスで選択した内容に応じて、このパネルの各タブで、ドキュメント/レコードとカテゴリの間の該当する交互作用を表示できます。それぞれは、同様の情報を、異なる方法で、またはさまざまなレベルの詳細情報とともに表示されます。ただし、現在選択している部分のグラフを更新するには、選択を行ったパネルまたはダイアログ ボックスのツールバーの **[表示]** をクリックします。

カテゴリとコンセプト ビューの [視覚化] パネルには、次のようなグラフおよび図表が表示されます。

- **カテゴリ棒グラフ:** テーブルおよび棒グラフを使用して、選択に該当するドキュメント/レコードと関連するカテゴリとの間の重なりを示します。また、棒グラフは、カテゴリ内のドキュメント/レコード数の、ドキュメント/レコード数の合計に対する比率も示します。詳細は、p. 279 カテゴリ棒グラフ を参照してください。
- **カテゴリ Web グラフ:** このグラフは、その他のパネルの選択部分に従って、ドキュメント/レコードが属するカテゴリのドキュメント/レコードの重なりを示します。詳細は、 p. 280 カテゴリ Web グラフ を参照してください。
- **カテゴリ Web テーブル:** このテーブルは、[カテゴリ Web グラフ] タブと同じ情報をテーブル形式で表示します。このテーブルには 3 つの列があり、列の見出しをクリックするとソートできます。 詳細は、p. 281 カテゴリ Web テーブル を参照してください。

詳細は、 10 章 p. 177 テキスト データの分類 を参照してください。

カテゴリ棒グラフ

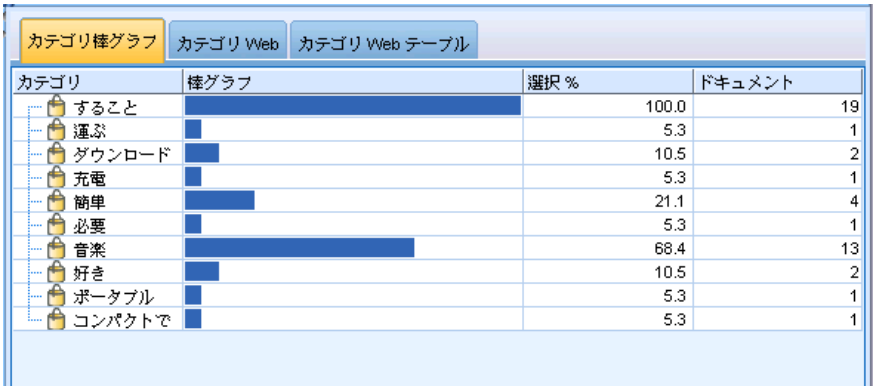
このタブには、選択に該当するドキュメント/レコードと関連するカテゴリとの間の重なりを示すテーブルおよび棒グラフが表示されます。また、棒グラフは、カテゴリ内のドキュメント/レコード数の、ドキュメントまたはレコード数の合計に対する比率も示します。このグラフのレイアウトは編集できません。ただし、列の見出しをクリックして、列をソートすることはできます。

テーブルには、次の列が表示されます。

- **カテゴリ:** 選択したカテゴリの名前が表示されます。デフォルトでは、選択した中で最も一般的なカテゴリが最初に表示されます。
- **棒グラフ:** 指定されたカテゴリのドキュメントまたはレコード数の、ドキュメントまたはレコード数の合計に対する比率を視覚的に表示します。

- **選択 %:** カテゴリのドキュメントまたはレコード数の、選択部分に表示されたドキュメントまたはレコード数の合計に対する比率に基づいたパーセンテージを示します。
- **ドキュメント:** 指定したカテゴリの選択部分のドキュメントまたはレコードの数を示します。

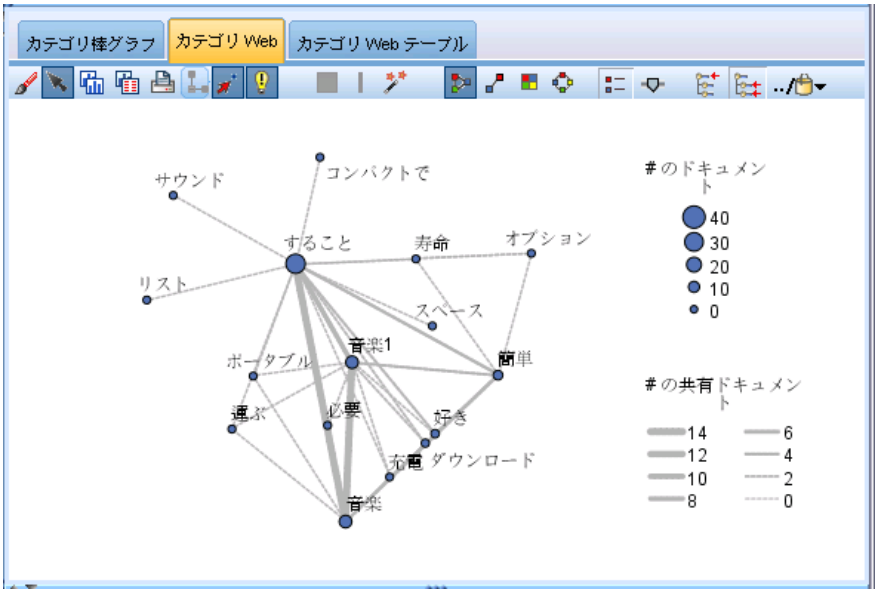
図 13-1
カテゴリ棒グラフ



カテゴリ Web グラフ

このタブには、カテゴリ Web グラフが表示されます。Web グラフは、その他のパネルの選択部分に従って、ドキュメントまたはレコードが属するカテゴリのドキュメントまたはレコードの重なりを示します。カテゴリ ラベルがある場合は、グラフにこれらのラベルが表示されます。このパネルのツールバー ボタンを使用して、グラフのレイアウト（ネットワーク、サークル、有向、またはグリッド）を選択できます。

図 13-2
カテゴリ Web グラフのグリッド レイアウト



Web グラフで、各ノードはカテゴリを示します。マウスを使用し、パネル内のノードを選択して移動できます。ノードのサイズは、選択部分のカテゴリのドキュメントまたはレコードの数に基づいた相対的なサイズを示します。カテゴリ間の線の太さと色は、含まれている共通のドキュメントまたはレコードの数を示します。探索的分析モードでノードの上にマウス ポインタを停止させると、ヒントにカテゴリの名前（またはラベル）およびカテゴリ内のドキュメントまたはレコードの全体数が表示されます。

注:デフォルトでは、ノードを移動できるグラフの探索的分析モードが有効化されています。ただし、編集モードに切り替えて、色、フォント、凡例など、グラフのレイアウトを編集できます。 詳細は、 p. 287 グラフのツールバーおよびパレットの使用 を参照してください。

カテゴリ Web テーブル

このタブには、[カテゴリ Web グラフ] タブと同じ情報をテーブル形式で表示されます。このテーブルには次の 3 つの列があり、列の見出しをクリックするとソートできます。

- **度数:** 2 つのカテゴリで共有している、または共通のドキュメントまたはレコードの数を表示します。
- **カテゴリ 1:** 最初のカテゴリの名前、そして含まれるドキュメントまたはレコード数の合計がカッコ内に表示されます。
- **カテゴリ 2:** 2 番目のカテゴリの名前、そして含まれるドキュメントまたはレコード数の合計がカッコ内に表示されます。

図 13-3
カテゴリ Web テーブル

度数	カテゴリ 1	カテゴリ 2
2	音楽(14)	ダウンロード(2)
1	音楽(14)	充電(1)
4	音楽(14)	簡単(6)
14	音楽(14)	音楽1(14)
14	音楽(14)	すること(32)
1	ダウンロード(2)	充電(1)
1	ダウンロード(2)	簡単(6)
2	ダウンロード(2)	音楽1(14)
2	ダウンロード(2)	すること(32)
1	充電(1)	簡単(6)
1	充電(1)	音楽1(14)
1	充電(1)	すること(32)
1	簡単(6)	音楽1(14)
2	簡単(6)	すること(32)
8	音楽1(14)	すること(32)
1	リスト(1)	すること(32)
6	すること(32)	音楽1(14)
1	スペース(1)	すること(32)

クラスタ グラフ

クラスタを作成した後、[視覚化] パネルの Web グラフで視覚的にクラスタを検証できます。 [視覚化] パネルは、[コンセプト Web グラフ] および [クラスタ Web グラフ] の 2 つのクラスタ化のパースペクティブを提供します。このパネルで Web グラフを使用して、クラスタリングを分析し、カテゴリに追加するコンセプトおよび規則を見つけることができます。[視覚化] パネルは、クラスタ ビューの右上隅にあります。表示されない場合、[表示] メニュー（**[表示] → [パネル] → [視覚化]**）からこのパネルにアクセスできます。[クラスタ] パネルでクラスタを選択すると、[視覚化] パネルに該当するグラフを自動的に表示できます。

注:デフォルトでは、グラフはインタラクティブ/選択モードで、ノードを移動できます。ただし、編集モードで、色、フォント、凡例など、グラフのレイアウトを編集できます。 詳細は、 p. 287 グラフのツールバーおよびパレットの使用 を参照してください。

クラスタ ビューには 2 つの Web グラフがあります。

- **コンセプト Web グラフ**:このグラフには、選択したクラスタ内のすべてのコンセプトおよびクラスタ外のリンクしたコンセプトが表示されます。このグラフを使用して、クラスタ内のコンセプトがどのようにリンクし

ているか、そして外部リンクを確認することができます。 詳細は、 p. 283 コンセプト Web グラフ を参照してください。

- **クラスタ Web グラフ**:このグラフには、選択したクラスタと、表示される選択したクラスタ間のすべての外部リンクを点線で表示します。 詳細は、 p. 284 クラスタ Web グラフ: を参照してください。

詳細は、 11 章 p. 259 クラスタの分析 を参照してください。

コンセプト Web グラフ

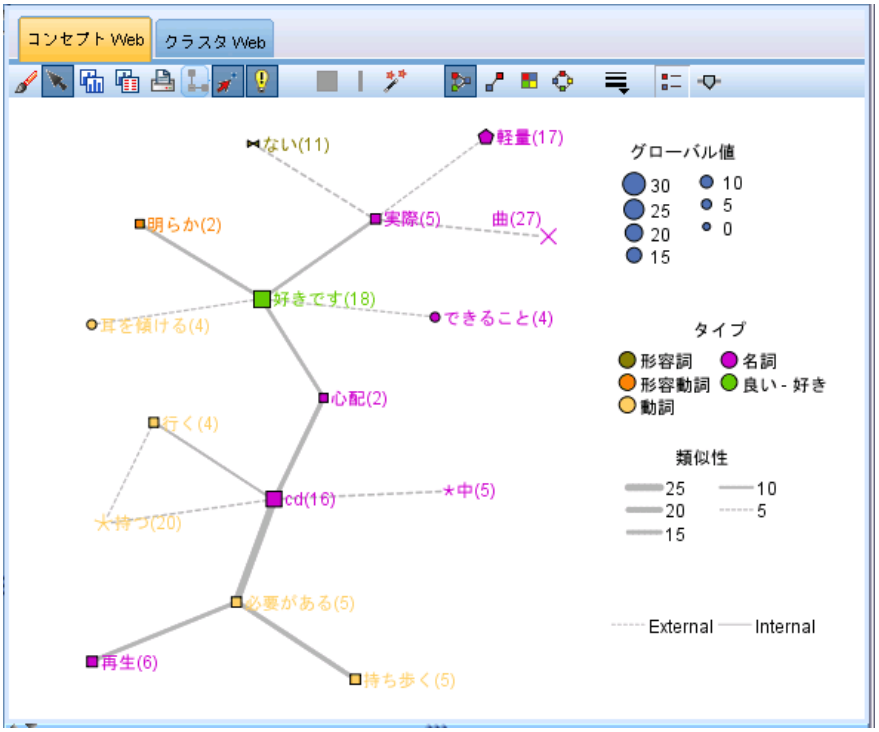
このタブには、クラスタ外のリンクしたコンセプトのほか、選択したクラスタ内のすべてのコンセプトを表示する Web グラフが表示されます。このグラフを使用して、クラスタ内のコンセプトがどのようにリンクしているか、そして外部リンクを確認することができます。クラスタ内の各コンセプトはノードとして表示され、タイプの色によって色分けされます。 詳細は、 17 章 p. 334 タイプの作成 を参照してください。

クラスタ内のコンセプト間の内部リンクが描画され、各リンクの線の太さは、グラフ ツールバーの選択に応じて、各コンセプトペアの共起のドキュメント数または類似度リンク値に直接関連します。クラスタのコンセプトおよびクラスタ外のこれらのコンセプト間の外部リンクも表示されます。

［クラスタ定義］ダイアログ ボックスでコンセプトを選択した場合、コンセプト Web グラフには、これらのコンセプトと、それらに関連する内部リンクおよび外部リンクが表示されます。選択したコンセプトのいずれかを含まないその他のコンセプト間のリンクは、グラフには表示されません。

注:デフォルトでは、グラフはインタラクティブ/選択モードで、ノードを移動できます。ただし、編集モードで、色、フォント、凡例など、グラフのレイアウトを編集できます。 詳細は、 p. 287 グラフのツールバーおよびパレットの使用 を参照してください。

図 13-4
コンセプト Web グラフ



クラスタ Web グラフ:

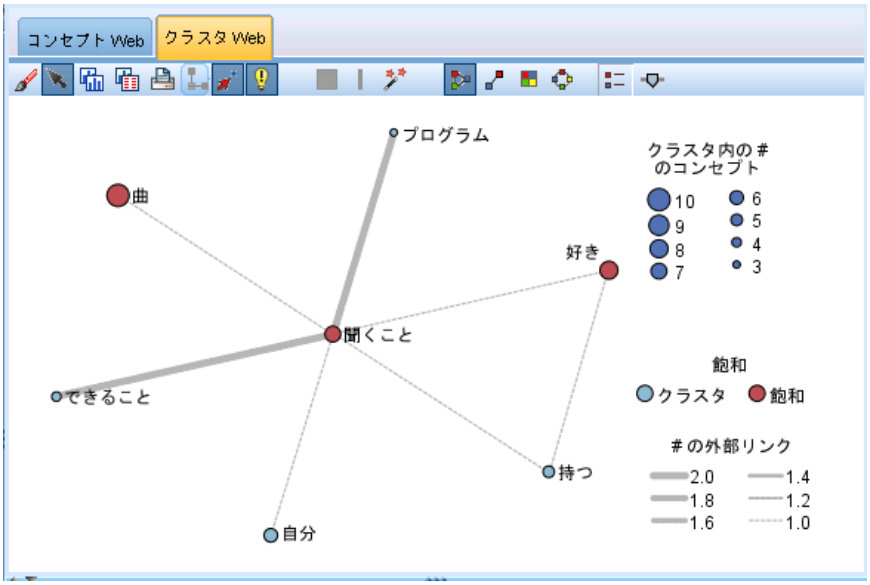
このタブには、選択したクラスタを示す Web グラフが表示されます。他のクラスタ間のリンクのほか、選択したクラスタ間の外部リンクがすべて点線で表示されます。クラスタ Web グラフでは、各ノードはクラスタ全体を表し、線の太さは、2 つのクラスタ間の外部リンク数を示します。

重要: クラスタ Web グラフを表示するには、外部リンクを持つクラスタを構築する必要があります。外部リンクは、別のクラスタのコンセプトペア間のリンクです（あるクラスタのコンセプトと別のクラスタのコンセプトとの間）。

たとえば、2 つのクラスタがあるとします。 **クラスタ** A には 3 つのコンセプト、A1、A2、およびA3 があります。 **クラスタ** B には 2 つのコンセプト、B1 および B2 があります。 コンセプト間のリンクは、A1-A2、A1-A3、A2-B1（外部）、A2-B2（外部）、A1-B2（外部）、および B1-B2 です。クラスタ Web グラフでは、線の太さが 3 つの外部リンクを示すことを意味します。

注:デフォルトでは、グラフはインタラクティブ/選択モードで、ノードを移動できます。ただし、編集モードで、色、フォント、凡例など、グラフのレイアウトを編集できます。 詳細は、 p. 287 グラフのツールバーおよびパレットの使用 を参照してください。

図 13-5
クラスタ Web グラフ



テキスト リンク分析のグラフ

テキスト リンク分析（TLA）パターンを抽出した後、[視覚化] パネルの Web グラフで視覚的にクラスタを検証できます。 視覚化パネルは、コンセプト（パターン）Web グラフ、およびタイプ（パターン）Web グラフ の 2 つのTLA パターンのパースペクティブを提供します。このパネルの Web グラフを使用して、パターンを視覚的に表すことができます。[視覚化] パネルは、テキスト リンク分析の右上隅にあります。表示されない場合、[表示] メニュー（**[表示] → [パネル] → [視覚化]**）からこのパネルにアクセスできます。選択項目がない場合、グラフ領域が空になります。

注:デフォルトでは、グラフはインタラクティブ/選択モードで、ノードを移動できます。ただし、編集モードで、色、フォント、凡例など、グラフのレイアウトを編集できます。 詳細は、 p. 287 グラフのツールバーおよびパレットの使用 を参照してください。

テキスト リンク分析ビューには 2 つの Web グラフがあります。

- **コンセプト Web グラフ**:このグラフには、選択したパターンのすべてのコンセプトを示します。コンセプト グラフの線の幅およびノードのサイズ（タイプ アイコンが表示されていない場合）には、選択したテー

ブルのグローバル出現値を示します。 詳細は、 p. 286 コンセプト Web グラフ を参照してください。

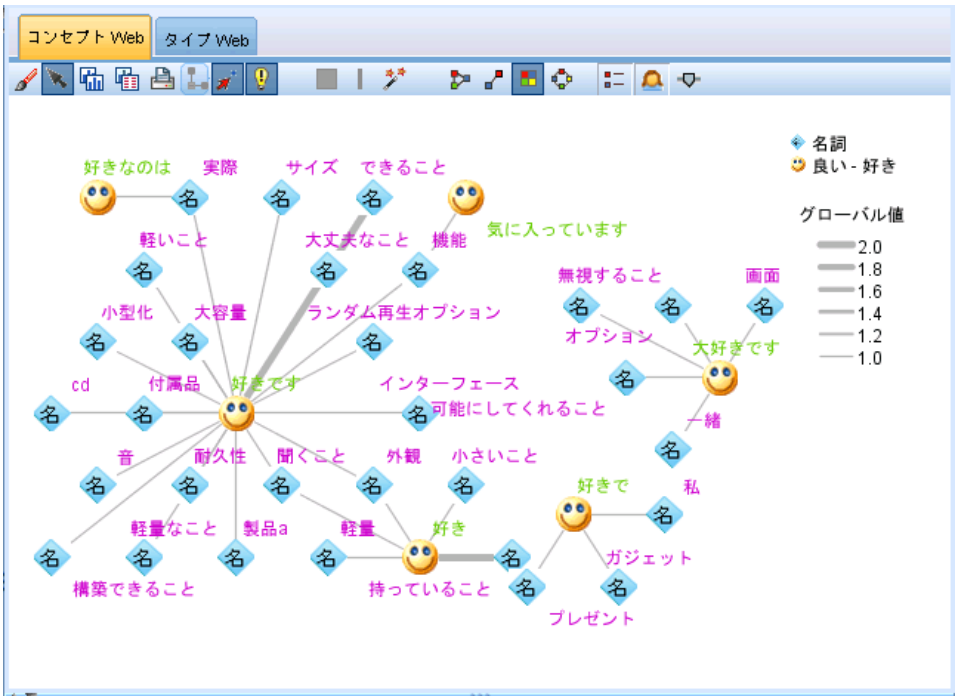
- **タイプ Web グラフ**: このグラフには、選択したパターンのすべてのタイプを示します。グラフの線の幅およびノードのサイズ（タイプ アイコンが表示されていない場合）には、選択したテーブルのグローバル出現値を示します。ノードは、タイプ カラーまたはアイコンによって示されます。 詳細は、 p. 287 タイプ Web グラフ を参照してください。

詳細は、 12 章 p. 268 テキスト リンク分析の検証 を参照してください。

コンセプト Web グラフ

Web グラフは、現在の選択で示されているすべてのコンセプトを表示します。たとえば、3 つの一致コンセプト パターンがあるタイプ パターンを選択した場合、このグラフにはリンクしたコンセプトが 3 セット表示されます。コンセプト グラフの線の幅およびノード サイズは、グローバル頻度値を示します。グラフには、パターンのパネルで選択されたものと同じ情報が表示されます。各コンセプトのタイプは、グラフ ツールバーの選択内容に応じて、色またはアイコンによって表示されます。 詳細は、 p. 287 グラフのツールバーおよびパレットの使用 を参照してください。

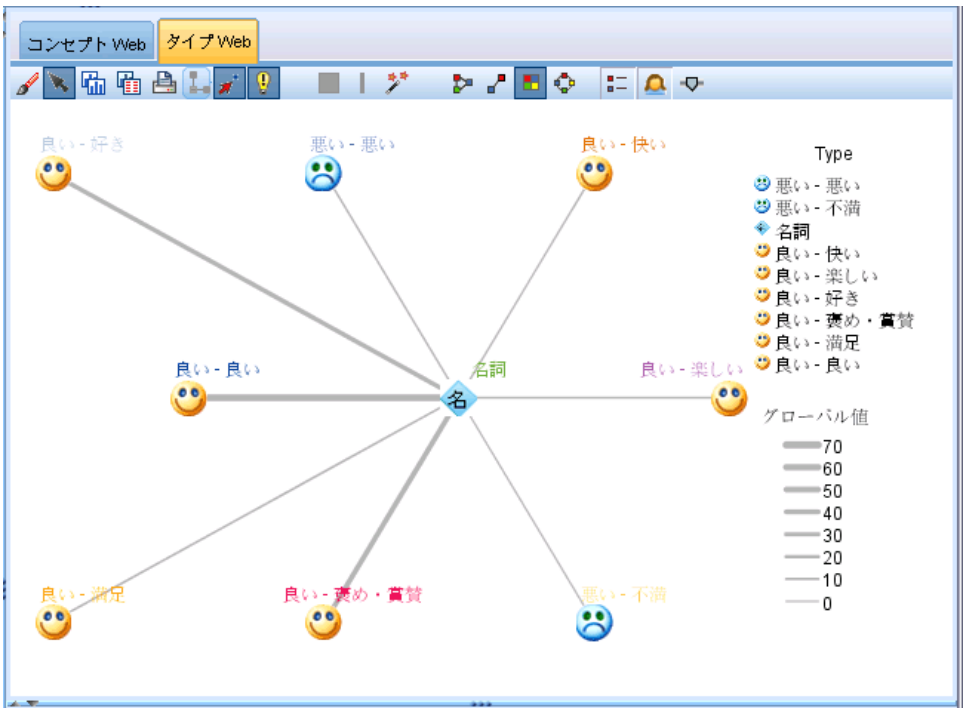
図 13-6
コンセプト Web グラフ



タイプ Web グラフ

この Web グラフは、現在の選択の各タイプ パターンを示します。たとえば、2 つのコンセプト パターンを選択した場合、このグラフには選択したパターンのタイプごとに 1 つのノードおよび同じパターンで見つかったタイプ間のリンクを示します。線の幅およびノード サイズは、セットのグローバル値を示します。グラフには、パターンのパネルで選択されたものと同じ情報が表示されます。グラフに表示されるタイプ名のほか、グラフ ツールバーで選択した内容によって色またはタイプ アイコンによって識別されます。 詳細は、 p. 287 グラフのツールバーおよびパレットの使用 を参照してください。

図 13-7
タイプ Web グラフ



グラフのツールバーおよびパレットの使用

各グラフにツールバーがあり、グラフに行うさまざまな操作を実行できるいくつかの共通パレットをすぐに使用できます。各ビュー（カテゴリおよびコンセプト、クラスタ、テキスト リンク分析）には、若干異なるツールバーがあります。探索的分析ビュー モードまたは編集ビュー モードから選択できます。

探索モードでは、視覚化によって表現されたデータや値を分析的に検討することができます。一方、編集モードでは、視覚化のレイアウトや外観を変更することができます。たとえば、フォントや色を自分の組織のスタイル ガイドに合わせて変更することが可能です。このモードを選択するには、メニューから [**表示**] → [**視覚化パネル**] → [**編集モード**] を選ぶか、ツールバーにあるアイコンをクリックします。

編集モードには、視覚化のレイアウトのさまざまな要素に影響を与えるいくつかのツールバーがあります。使用しないものがある場合は、そのツールバーを非表示にしてダイアログ ボックスにおけるグラフの表示領域を増やすことができます。ツールバーの選択または選択解除を行うには、[表示] メニューで目的のツールバーまたはパレットの名前をクリックします。

グラフの編集に使用する一般的なツールバーおよびパレットの詳細は、オンライン ヘルプまたはファイル SourceProcessOutputNodes.pdf の「グラフの編集」を参照してください。このファイルは、IBM® SPSS® ModelerDVD の ¥Documentation¥en フォルダにあります。

テーブル 13-1
Text Analytics のツールバー ボタン

ボタン/リスト	説明
	編集モードを有効化します。編集モードに切り替えて、フォントの拡大、会社のスタイル ガイドに合った色への変更、またはラベルや凡例の削除など、グラフの外観を変更できます。
	探索的分析モードを有効化します。デフォルトでは、探索的分析モードがオンになっており、グラフの周囲でノードを移動およびドラッグ、そしてグラフ オブジェクトの上でマウス ポインタを定義させて、ヒントの詳細情報を表示できます。
	カテゴリとコンセプト ビューおよびテキスト リンク分析ビューでグラフの Web 表示の種類を選択します。 ■ **サークル レイアウト**: どのようなグラフにも適用できる一般的なレイアウト。リンクに方向がないことを想定してグラフをレイアウトし、すべてのノードを同様に扱います。ノードは円の周囲にのみ配置されます。 ■ **ネットワーク レイアウト**: どのようなグラフにも適用できる一般的なレイアウト。リンクに方向がないことを想定してグラフをレイアウトし、すべてのノードを同様に扱います。ノードは、レイアウト内に自由に配置されます。 ■ **有向レイアウト**: 方向のあるグラフにのみ使用できるレイアウト。このレイアウトは、ルート ノードから葉ノードへのツリー上の構造を作成し、色別に構成します。このレイアウトを使用すると、階層データが適切に表示されます。 ■ **グリッド レイアウト**: どのようなグラフにも適用できる一般的なレイアウト。リンクに方向がないことを想定してグラフをレイアウトし、すべてのノードを同様に扱います。ノードは領域内のグリッド ポイントにのみ配置されます。

ボタン/リスト	説明
	リンクの太さの表示。グラフで線の太さが示す内容を選択します。これは、クラスタ ビューにのみ適用されます。クラスタ Web グラフは、クラスタ間の外部リンク数のみを表示します。以下から選択できます。 ■ **類似度**: 2 つのクラスタ間の外部リンク数を示します。 ■ **共起**: 記述子の共起が出現するドキュメント数を示します。
	凡例を表示する切り替えボタン。ボタンを押さない場合、判例は表示されません。
	タイプの色ではなくグラフ内のタイプのアイコンを表示する切り替えボタン。これは、テキスト リンク分析ビューにのみ適用されます。
	グラフの下にリンク スライダーを表示する切り替えボタン。矢印をスライドして、結果を絞り込むことができます。
	サブカテゴリではなく、選択されたカテゴリの最上位レベルのグラフが表示されます。
	選択されたカテゴリの最下位レベルのグラフが表示されます。
	サブカテゴリの名前を出力内でどのように表示するかを制御します。 ■ **完全カテゴリパス**: カテゴリ名と、該当する場合、カテゴリ名とサブカテゴリ名をスラッシュを使用して区切り、上位カテゴリの完全パスを出力します。 ■ **短いカテゴリパス**: カテゴリ名のみを出力します。ただし、省略記号を使用して、該当するカテゴリの上位カテゴリ数を示します。 ■ **下位レベルのカテゴリ**: 完全パスまたは上位カテゴリを表示せず、カテゴリ名のみを出力します。

章
14

セッション リソース エディタ

IBM® SPSS® Modeler Text Analytics は、主要キーワードをテキスト データから迅速にかつ正確にキャプチャします。この抽出プロセスは、テキストデータからの情報抽出を管理する言語リソースに大きく依存しています。デフォルトでは、これらのリソースはリソース テンプレートによって決まります。

SPSS Modeler Text Analytics は、言語リソースおよび非言語リソースを含む、ライブラリおよび高度なリソースの形式で専門的な**リソース テンプレート**のセットに付属しており、データの処理方法および抽出方法を定義できます。 詳細は、 15 章 p.297 テンプレートとリソース を参照してください。

ノードのダイアログ ボックスで、テンプレートのリソースをノードに読み込むことができます。一度インタラクティブ ワークベンチ セッションに入ると、必要に応じてこのノードのデータ用にこれらのリソースをカスタマイズできます。インタラクティブ ワークベンチ セッションの間、リソース エディタ ビューでリソースを作業できます。インタラクティブ セッションが起動すると、ノードにデータおよび抽出結果をキャッシュしていない場合は、ノードのダイアログ ボックスで読み込まれたリソースを使用して抽出を実行します。

リソース エディタを使用したリソースの編集

リソース エディタ では、インタラクティブ ワークベンチ セッションの抽出結果（コンセプト、タイプ、およびパターン）を作成に使用するリソースのセットへのアクセスが用意されています。このエディタは、テンプレート エディタ と非常に類似していますが、リソース エディタ では、インタラクティブ ワークベンチ セッションでリソースを編集するという点で異なります。リソースの作業および実行した他の作業を終了している場合、モデル作成ノードを更新してこの作業を保存し、後続のインタラクティブ ワークベンチ セッションで復元できます。 詳細は、 8 章 p.148 モデル作成ノードの更新および保存 を参照してください。

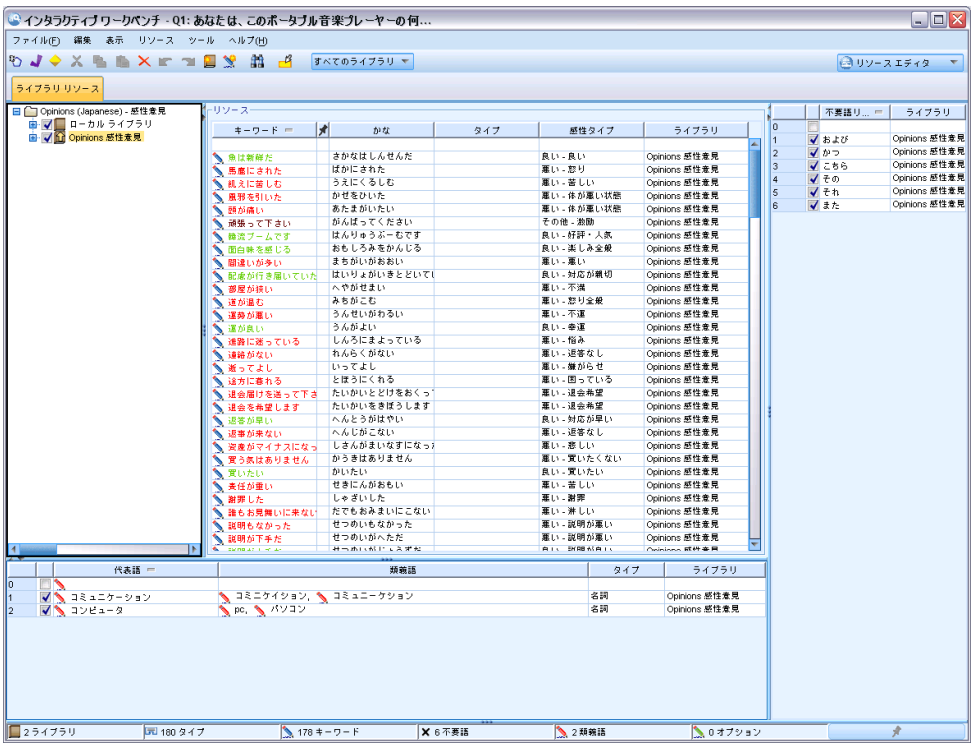
ノードにリソースを読み込むために使用するテンプレートで直接作業する場合は、テンプレート エディタ を使用することをお勧めします。次のような リソース エディタ 内で実行できる多くのタスクは、テンプレート エディタ と同じように実行されます。

- **ライブラリの使用**。 詳細は、 16 章 p.316 ライブラリの使用 を参照してください。

- **キーワード辞書の作成**。 詳細は、 17 章 p. 334 タイプの作成 を参照してください。
- **キーワードを辞書に追加**。 詳細は、 17 章 p. 337 キーワードを追加 を参照してください。
- **類義語の作成**。 詳細は、 17 章 p. 346 類義語の定義 を参照してください。
- **テンプレートのインポートおよびエクスポート**。 詳細は、 15 章 p. 309 テンプレートのインポートおよびエクスポート を参照してください。
- **ライブラリの公開**。 詳細は、 16 章 p. 328 ライブラリの公開 を参照してください。

オランダ語、英語、フランス語、ドイツ語、イタリア語、ポルトガル語、スペイン語のテキストの場合

図 14-1
日本語以外のリソース エディタ ビュー

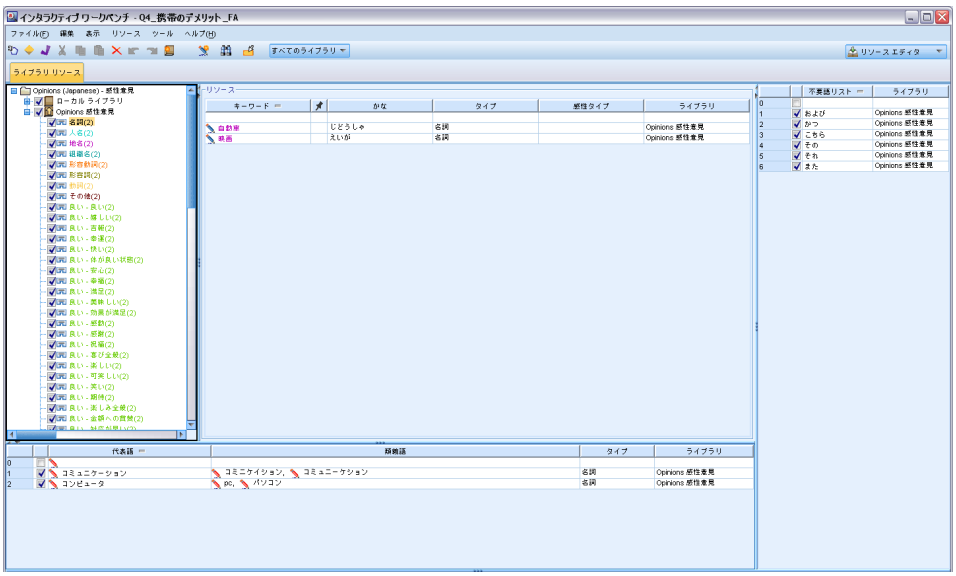


日本語テキストの場合

日本語テキストのエディタのインターフェイスは、他のテキスト言語と異なります。 詳細は、 A 付録 p. 407 日本語テキストのリソースの編集 を参照してください。

注:日本語テキスト展開はIBM® SPSS® Modeler Premiumで利用可能です。

図 14-2
日本語テキストのリソース エディタ ビュー

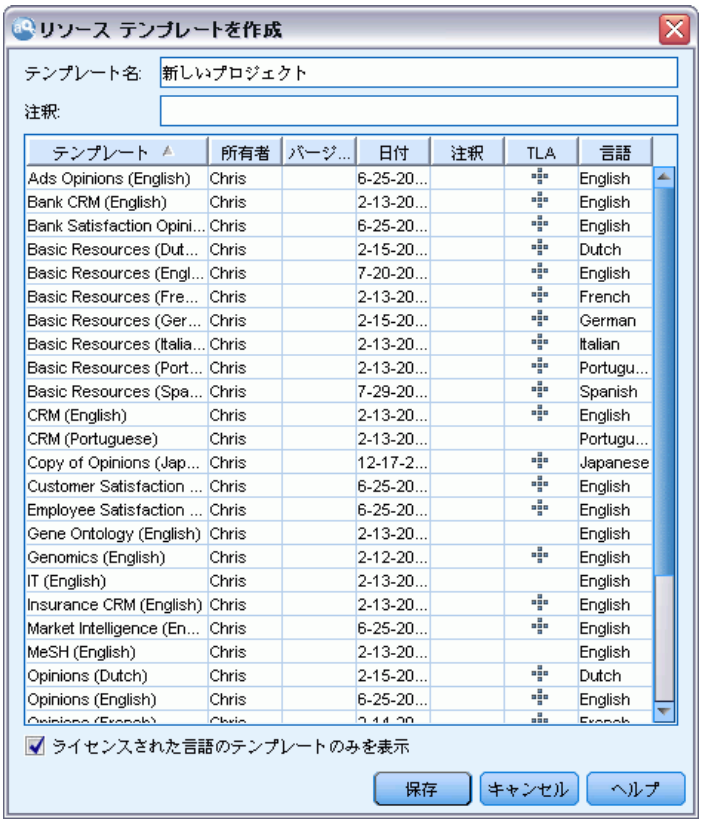


テンプレートの作成および更新

リソースに変更を行い、今後それらを再利用したい場合、リソースをテンプレートとして保存することができます。保存する場合は、既存のテンプレート名を使用して保存するのか、新しい名前を付けるのかを選択できます。その後、テンプレートを読み込むと、同じリソースを取得することができます。 詳細は、 3 章 p. 46 テンプレートおよび TAP からのリソースのコピー を参照してください。

注:ライブラリを公開して共有することもできます。 詳細は、 16 章 p. 326 ライブラリの共有 を参照してください。

図 14-3
[リソース テンプレートを作成] ダイアログ ボックス



テンプレートを作成 (または更新) するには

- リソース エディタ ビューのメニューで、[リソース] > [リソース テンプレートを作成] を選択します。[リソース テンプレートを作成] ダイアログ ボックスが開きます。

- 新しいテンプレートを作成する場合は、[テンプレート名] フィールドに新しい名前を入力します。既存のテンプレートを現在読み込まれたリソースで上書きする場合は、テーブルでテンプレートを選択します。

- [保存] をクリックして、テンプレートを作成します。

重要! テンプレートはノードで選択されたときに読み込まれ、ストリームが実行されているときは読み込まれないため、最新の変更を取得したい場合にリソース テンプレートを使用するその他のノードのリソース テンプレートを再読み込みしてください。 詳細は、 15 章 p. 306 読み込み後のノード リソースの更新 を参照してください。

リソース テンプレートの切り替え

現在セッションで読み込まれたリソースを別のテンプレートからのコピーに置き換えたい場合、これらのリソースに切り替えることができます。これによって、現在セッション内に読み込まれているリソースを上書きすることができますリソースを切り替えて定義済みのテキスト リンク分析（TLA）パターン ルールを使用する場合、必ずTLA列内でマークされたテンプレートを選択してください。

重要!日本語テンプレートから日本語以外のテンプレートへの切り替え、またその逆の切り替えを行うことはできません。注:日本語テキスト展開はIBM® SPSS® Modeler Premiumで利用可能です。

セッション作業（カテゴリ、パターン、リソース）を復元したいが他のセッション作業を失わずにテンプレートからリソースのコピーを読み込んで更新したい場合特に、リソースの切り替えが役立ちます。リソース エディタに内容をコピーしたいテンプレートを選択し、[OK] をクリックします。これにより、このセッションのリソースが置き換えられます。次回インタラクティブ ワークベンチ セッションを起動するときにこれらの変更を保持したい場合、セッションの終わりにモデル作成ノードを更新してください。

注:インタラクティブ セッションで別のテンプレートの内容に切り替える場合、ノードに表示されるテンプレートの名前は、最後に読み込まれ、コピーされたテンプレートの名前となります。これらのリソースまたは他のセッション作業を利用するには、セッションを終了する前にモデル作成ノードを更新し、ノードで [**セッション作業を使用**] オプションを選択します。 詳細は、 8 章 p. 148 モデル作成ノードの更新および保存 を参照してください。

図 14-4
[リソースを切り替え] ダイアログ ボックス

リソースを切り替えるには

- リソース エディタ ビューのメニューで、[リソース] > [リソース テンプレートの切り替え] を選択します。[テンプレートの切り替え] ダイアログ ボックスが開きます。
- テーブルに表示されたテンプレートから、使用したいテンプレートを選択します。
- [OK] をクリックして、現在読み込まれているこれらのリソースを中止し、代わりに選択したテンプレートのリソースのコピーを読み込みます。リソースに変更を行い、今後使用するためにライブラリを保存したい場合、切り替える前にそれらを公開、更新、共有することができます。 詳細は、16 章 p. 326 ライブラリの共有 を参照してください。

パート III:
テンプレートとリソース

章
15

テンプレートとリソース

IBM® SPSS® Modeler Text Analytics は、主要キーワードをテキスト データから迅速にかつ正確にキャプチャします。この抽出プロセスは、テキストデータからの情報抽出を管理する言語リソースに大きく依存しています。詳細は、 1 章 p.7 抽出の方法 を参照してください。リソース エディタ ウィンドウで、これらのリソースを調整できます。

ソフトウェアをインストールすると、専門的なリソースも取得します。これらの付属リソースは、特定の言語と特定の応用分野で、数年にわたる調査と調整の結果得られたもので、ユーザーはその恩恵を受けることができます。ただし、これらの特別なリソースは、使用するデータの文脈には、完全に一致してはいまないので、ユーザー側の組織のデータ向けに、これらのリソーステンプレートを編集したり、独自に調整したカスタムライブラリを作成して使用することができるようになっています。これらのリソースの形式は多岐にわたり、それぞれセッションで使用できます。リソースは、次の中にあります。

- **リソース テンプレート:**テンプレートは、製品への意見といったように、ある特定の領域や文脈に特化したリソースをまとめた形で、ライブラリのセット、タイプ、およびアドバンズ リソースで構成されています。
- **テキスト分析パッケージ (TAP):**テンプレートに保存されているリソースに加え、リソースをもとに作成した専門的カテゴリセットをまとめたテキスト分析パッケージ(TAP)は、カテゴリとリソースをいっしょに保存して再利用することを可能にします。 詳細は、 10 章 p.245 テキスト分析パッケージの使用 を参照してください。
- **ライブラリ:**ライブラリは、TAP およびテンプレートの構成要素として使用されます。それらは、セッションのリソースに個別に追加されます。各ライブラリはいくつかの辞書で構成され、タイプのリスト、類義語リスト、不要語リストを定義、管理するために使用されます。ライブラリは個別に提供されていますが、テンプレートおよびTAP でパッケージ化されています。 詳細は、 16 章 p.316 ライブラリの使用 を参照してください。

注:抽出時、いくつかのコンパイル済み内部辞書も使用されます。これらのコンパイル済み辞書には、コア ライブラリのタイプを補完する多くの定義が含まれています。これらのコンパイル済み辞書は編集できません。

リソース エディタ を用いることで、抽出結果(キーワード、タイプ、およびパターン) を出力する際に適用されるリソースセットへのアクセスが可能となります。リソース エディタ で実行するタスクには、次のような数多くのものがあります。

- **ライブラリの使用**。 詳細は、 16 章 p. 316 ライブラリの使用 を参照してください。
- **キーワード辞書の作成**。 詳細は、 17 章 p. 334 タイプの作成 を参照してください。
- **キーワードを辞書に追加**。 詳細は、 17 章 p. 337 キーワードを追加 を参照してください。
- **類義語の作成**。 詳細は、 17 章 p. 346 類義語の定義 を参照してください。
- **TAP のリソースの更新**。 詳細は、 10 章 p. 250 テキスト分析パッケージの更新 を参照してください。
- **テンプレートの作成**。 詳細は、 14 章 p. 292 テンプレートの作成および更新 を参照してください。
- **テンプレートのインポートおよびエクスポート**。 詳細は、 p. 309 テンプレートのインポートおよびエクスポート を参照してください。
- **ライブラリの公開**。 詳細は、 16 章 p. 328 ライブラリの公開 を参照してください。

テンプレート エディタとリソース エディタの比較

テンプレート、ライブラリ、リソースを使用および編集する方法は主な方法は 2 つあります。テンプレート エディタ または リソース エディタ で言語リソースの作業ができます。

テンプレート エディタ

テンプレート エディタ を使用すると、インタラクティブ ワークベンチ セッションがなく特定のノードまたはストリームから独立しているリソース テンプレートを作成および編集できます。このエディタを使用して、テキスト リンク分析ノードおよびテキスト マイニング モデル作成ノードに読み込む前にリソース テンプレートを作成または編集できます。

テンプレート エディタ は、IBM® SPSS® Modeler のメイン ツールバーを使用、または [ツール] → テンプレート エディタ] メニューを選択して使用できます。

リソース エディタ

インタラクティブ ワークベンチ セッション内で使用できる リソース エディタ によって、特定のノードおよびデータセットのコンテキストでリソースを使用できます。テキスト マイニング モデル作成ノードをストリームに追加すると、リソース テンプレートの内容のコピー、またはテキスト分析パッケージ（カテゴリ セットおよび言語リソース）のコピーを読み込んで、テキスト マイニングに使用するテキストの抽出方法を制御できます。インタラクティブ ワークベンチ セッションを起動すると、カテゴリの作成、テキスト リンク分析パターンの抽出、カテゴリ モデルの作成のほか、統合された リソース エディタ ビューでそのセッションのデータのリソースを調整することもできます。 詳細は、 14 章 p. 290 リソース エディタを使用したリソースの編集 を参照してください。

インタラクティブ ワークベンチ セッションでリソースの作業を行うと、それらの変更はそのセッションにのみ適用されます。後続のセッションで継続できるよう、作業（リソース、カテゴリ、パターンなど）を保存したい場合、モデル作成ノードを更新する必要があります。 詳細は、 8 章 p. 148 モデル作成ノードの更新および保存 を参照してください。

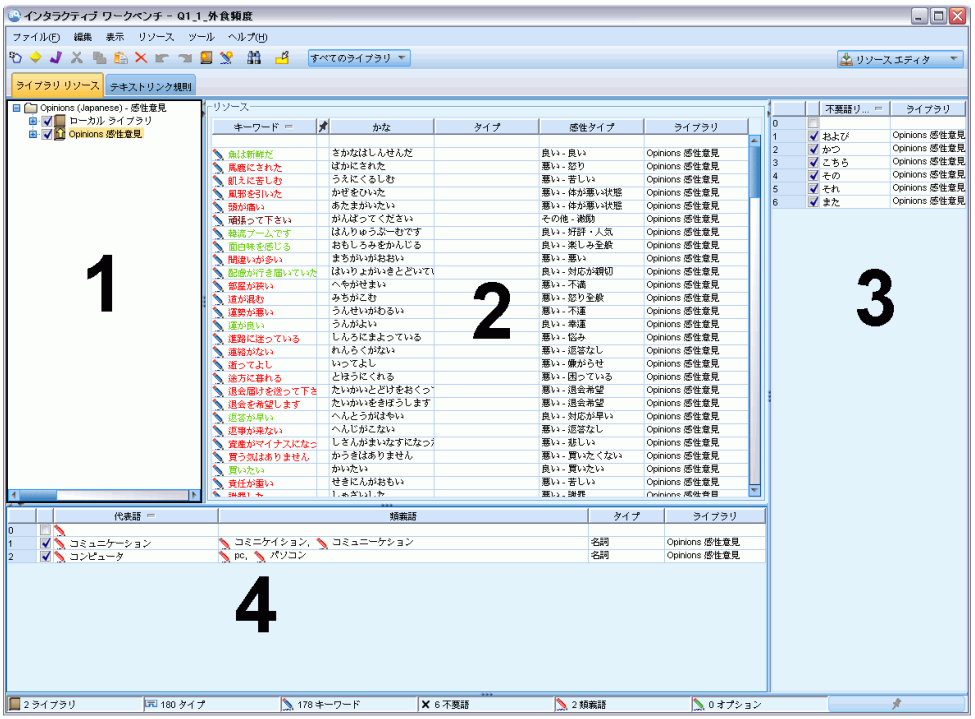
変更を保存して元のテンプレートの内容がモデル作成ノードにコピーされている元のテンプレートに戻る場合、更新されたテンプレートを他のノードに読み込むことができ、リソースからテンプレートを作成できます。 詳細は、 14 章 p. 292 テンプレートの作成および更新 を参照してください。

エディタのインターフェイス

テンプレート エディタまたはリソース エディタ で実行する操作は、言語リソースの管理および調整を中心に展開しています。これらのリソースは、テンプレートおよびライブラリの形で保存されています。 詳細は、 17 章 p. 332 キーワード辞書 を参照してください。

[ライブラリ リソース] タブ

図 15-1
テキスト マイニング テンプレート エディタ



インターフェイスは、次のような 4 つの部分で構成されています。

1. **[ライブラリ ツリー] パネル:** 左上のこのパネルにはライブラリのツリーが表示されます。このツリーでライブラリを有効化および無効化し、ツリーのライブラリを選択して、その他のパネルのビューをフィルタリングできます。コンテキスト メニューを使用して、このツリーで多くの操作を実行できます。ツリーのライブラリを展開して、含まれるタイプのセットを表示できます。特定のライブラリのみに焦点を当てたい場合、[**表示**] メニューからこのリストをフィルタリングすることもできます。

2. **[キーワード辞書のキーワード リスト] パネル:** ライブラリ ツリーの右側にあるこのパネルには、ツリーで選択されたライブラリのキーワード辞書のキーワード リストが表示されます。**キーワード辞書**は、1 つのラベル、またはタイプ、名前に基づいてグループ化されたキーワードの集合です。抽出エンジンがテキスト データを読み取る場合、テキストの単語を、キーワード辞書のキーワードと比較します。抽出したコンセプトがキーワード辞書でキーワードとして表示されている場合、そのタイプ名が割り当てられます。キーワード辞書を、共通点のあるキーワードの独立した辞書として見なすことができます。たとえば、コア ライブラリの `<Location>` タイプには、`new orleans`、`great britain`、および `new york` などのコン

セプトが含まれます。これらのキーワードはすべて、地名を示します。ライブラリには、1 つまたは複数のキーワード辞書が含まれます。詳細は、17 章 p. 332 キーワード辞書 を参照してください。

3. [不要語辞書] パネル: 右側にあるこのパネルには、最終的な抽出結果から除外されるキーワードの集合が表示されます。不要語辞書のキーワードは、[抽出結果] パネルには表示されません。不要語キーワードは選択するライブラリに保存できます。ただし、[不要語辞書] パネルには、ライブラリ ツリーに表示されるすべてのライブラリの不要語登録されたすべてのキーワードが表示されます。詳細は、 17 章 p. 350 不要語辞書 を参照してください。

4. [類義語辞書] パネル: 左下にあるこのパネルには、類義語およびオプションの要素がそれぞれのタブに表示されます。類義語およびオプションの要素を使用すると、最終的な抽出結果の代表語に基づいて類似したキーワードをグループ化できます。この辞書には既知の類義語やユーザー定義の類義語および要素、そして一般的なスペルミスと正しいスペルのペアが含まれています。類義語の定義およびオプションの要素は、選択するライブラリに保存できます。ただし、[類義語辞書] パネルには、ライブラリ ツリーに表示されるすべてのライブラリのすべての内容が表示されます。このパネルにはすべてのライブラリのすべての類義語またはオプションの要素が表示されますが、ツリーのすべてのライブラリの類義語は、このパネルでいっしょに表示されます。ライブラリには、含まれる類義語辞書は 1 つだけです。詳細は、 17 章 p. 344 類義語辞書 を参照してください。［オプションの要素］タブは、日本語テキストの言語リソースには適用されませんので、注意してください。

注:

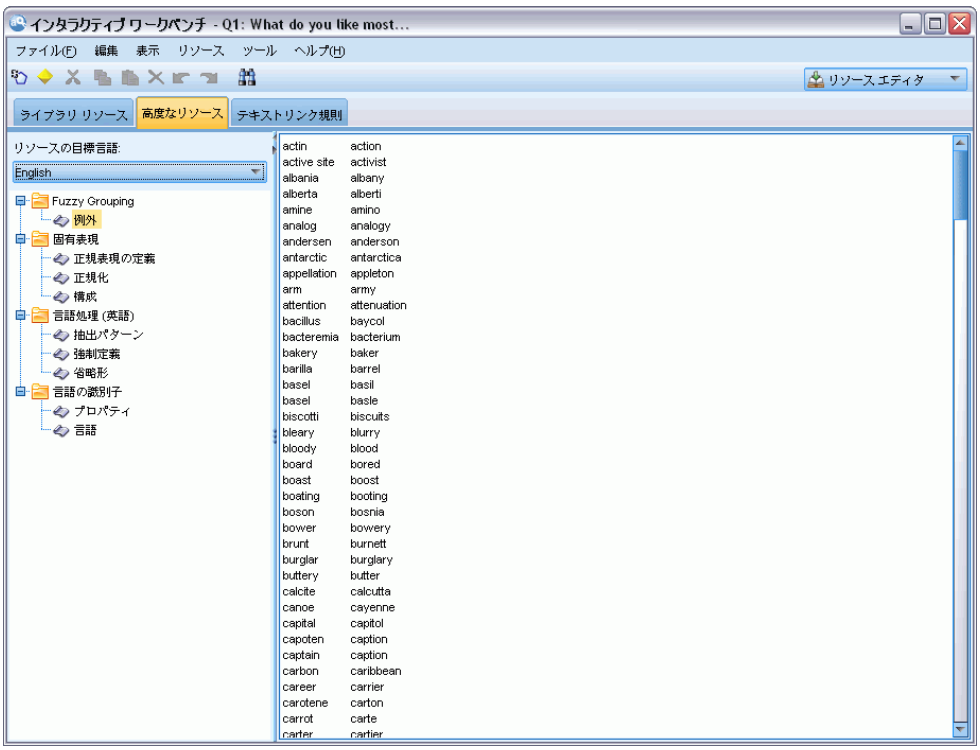
- 1 つのライブラリに関する情報のみ表示されるようフィルタリングしたい場合、ツールバーのドロップダウン リストを使用して、ライブラリ ビューを変更できます。[**すべてのライブラリ**] という上位レベルのエントリおよび各ライブラリの追加エントリが含まれます。 詳細は、 16 章 p. 321 ライブラリの表示 を参照してください。
- 日本語テキストのエディタのインターフェイスは、他のテキスト言語と異なります。詳細は、 A 付録 p. 407 日本語テキストのリソースの編集を参照してください。注:日本語テキスト展開はIBM® SPSS® Modeler Premiumで利用可能です。

[高度なリソース] タブ

エディタ ビューの 2 番目のタブで高度なリソースを使用できるようになりました。このタブで高度なリソースを確認および編集することができます。詳細は、 18 章 p. 353 アドバンス リソースについて を参照してください。

重要! このタブは、日本語テキストに対して調整されたリソースには使用できません。

図 15-2
テキスト マイニング テンプレート エディタ - [高度なリソース] タブ

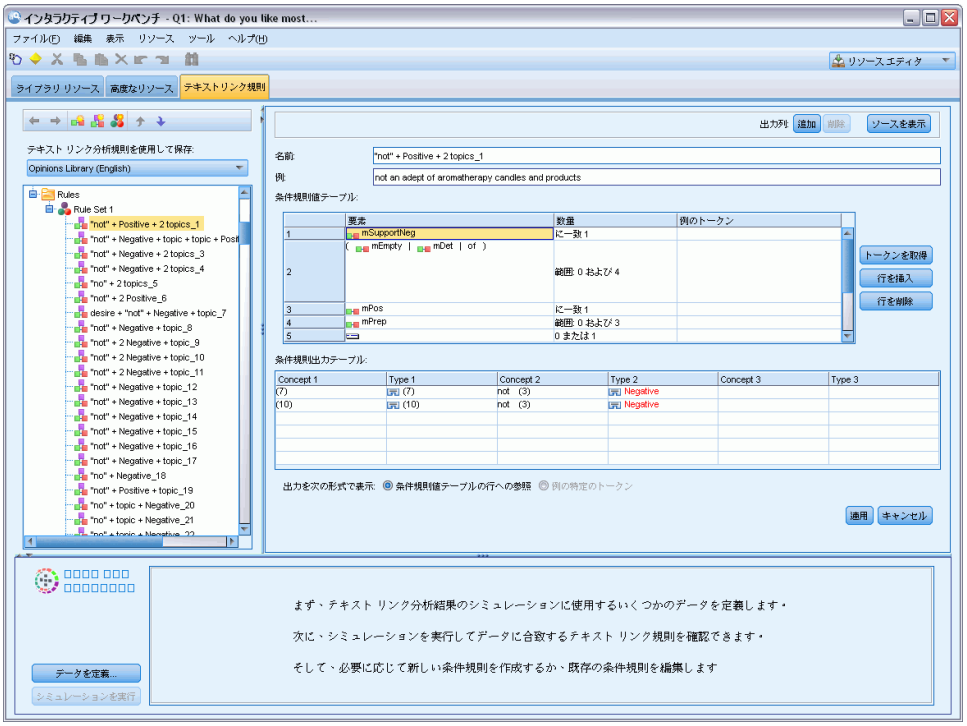


[テキスト リンク規則] タブ

バージョン 14 以降、テキスト リンク分析規則はエディタ ビューの独自のタブで編集できます。条件規則エディタで作業し、独自の規則を作成、シミュレーションを実行して、規則が TLA 結果にどのような影響を与えるかを確認できます。詳細は、 19 章 p. 368 テキスト リンク規則について を参照してください。

重要! このタブは、日本語テキストに対して調整されたリソースには使用できません。

図 15-3
テキスト マイニング テンプレート エディタ – [テキスト リンク規則] タブ

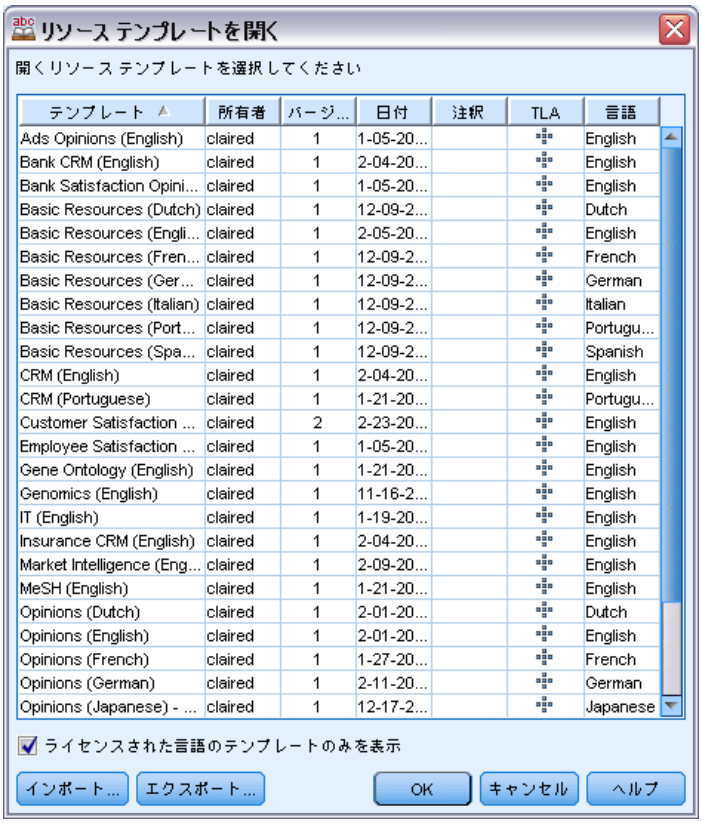


テンプレートを開く

テンプレート エディタ を起動すると、テンプレートを開くことを確認するメッセージが表示されます。同様に、[ファイル] メニューからテンプレートを開くことができます。いくつかのテキスト リンク分析 (TLA) 規則を含むテンプレートが必要な場合、[TLA] 列にアイコンのあるテンプレートを選択してください。テンプレートが作成された言語は、[言語] 列に表示されます。

テーブルに表示されたいテンプレートをインポートしたい場合、テンプレートをエクスポートしたい場合、[テンプレートを開く] ダイアログ ボックスのボタンを使用します。 詳細は、 p. 309 テンプレートのインポートおよびエクスポート を参照してください。

図 15-4
[リソース テンプレートを開く] ダイアログ ボックス



テンプレートを開くには

- テンプレート エディタ のメニューから [ファイル] → [リソース テンプレートを開く] を選択します。[リソース テンプレートを開く] ダイアログ ボックスが開きます。
- テーブルに表示されたテンプレートから、使用したいテンプレートを選択します。
- [OK] をクリックすると、このテンプレートが開きます。現在エディタで別のテンプレートが開いている場合、[OK] をクリックすると開いていたテンプレートが中断し、ここで選択したテンプレートが表示されます。リソースに変更を行い、今後使用するためにライブラリを保存したい場合、別のテンプレートを開く前にそれらを公開、更新、共有することができます。詳細は、 16 章 p. 326 ライブラリの共有 を参照してください。

テンプレートの保存

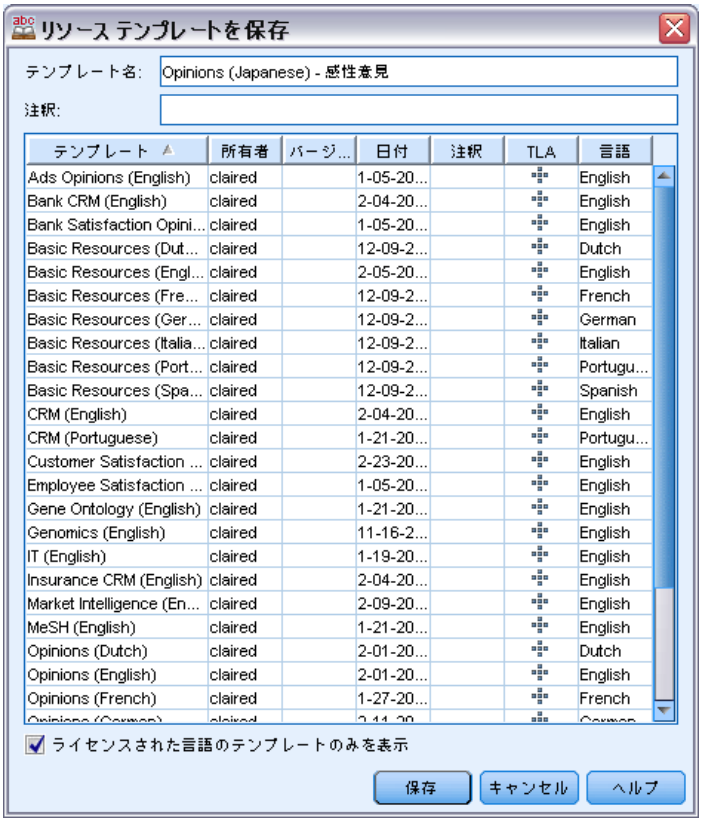
テンプレート エディタ で、テンプレートに行った変更を保存できます。既存のテンプレート名を使用して保存するのか、新しい名前を付けるのかを選択できます。

以前すでにノードに読み込んだテンプレートに変更を行った場合、最新の変更を取得するには、テンプレートの内容をノードに再読み込みする必要があります。 詳細は、 3 章 p. 46 テンプレートおよび TAP からのリソースのコピー を参照してください。

または、テキスト マイニングノードの［モデル］タブのオプション［**保存されたインタラクティブ作業を使用**］を使用している場合、つまり以前のインタラクティブ ワークベンチ セッションのリソースを使用している場合、インタラクティブ ワークベンチ セッションからこのテンプレートのリソースに切り替える必要があります。 詳細は、 14 章 p. 294 リソース テンプレートの切り替え を参照してください。

注:ライブラリを公開して共有することもできます。 詳細は、 16 章 p. 326 ライブラリの共有 を参照してください。

図 15-5
［リソース テンプレートを保存］ダイアログ ボックス



テンプレートを保存するには

- テンプレート エディタ のメニューから [**ファイル**] → [**リソース テンプレートを保存**] を選択します。[リソース テンプレートを保存] ダイアログ ボックスが開きます。
- このテンプレートを新しいテンプレートとして保存する場合は、[テンプレート名] フィールドに新しい名前を入力します。既存のテンプレートを現在読み込まれたリソースで上書きする場合は、テーブルでテンプレートを選択します。
- 必要に応じて、テーブルにコメントまたは注釈として表示する説明を入力します。
- [**保存**] をクリックして、テンプレートを保存します。

  重要! テンプレートまたは TAP のリソースがノードに読み込まれ/コピーされるため、テンプレートに変更を行い、既存のストリームでこれらの変更を活用する場合、それらを再読み込みすることによってリソースを更新する必要があります。 詳細は、 p. 306 読み込み後のノード リソースの更新 を参照してください。

読み込み後のノード リソースの更新

デフォルトでは、ノードをストリームに追加する場合、デフォルト テンプレートの一連のリソースが読み込まれ、ノードに組み込まれます。テンプレートを変更または TAP を使用する場合、それらを読み込むとこれらのリソースのコピーがリソースを上書きします。テンプレートおよび TAP がノードに直接リンクしていないため、テンプレートまたは TAP に行った変更は、以前のきそんのーどでは自動的に活用できません。これらの変更を活用するには、そのノードでリソースを更新する必要があります。リソースは、次の 2 通りの方法のいずれかで更新できます。

方法 1:[モデル] タブでリソースを再読み込みする

新しいまたは更新されたテンプレートまたは TAP を使用してノードのリソースを行進する場合、ノードの [モデル] タブでテンプレートを再読み込みできます。再読み込みをして、ノードのリソースのコピーを最新のコピーと置き換えます。元のテンプレート名のほか、更新日時が [モデル] タブに表示されます。 詳細は、 3 章 p. 46 テンプレートおよび TAP からのリソースのコピー を参照してください。

ただし、テキスト マイニング モデル作成ノードでインタラクティブ セッション データの作業を行っており、[モデル] タブの [**セッション作業を使用** オプションを選択している場合、保存されたセッション作業およびリソースが使用され、[**読み込み**] ボタンが無効になります。インタラクティブ ワークベンチ セッション時に一度、[**モデル作成ノードを更新**] オプションを選

択して、カテゴリ、リソースおよびその他のセッション作業を保持しているため無効になります。これらのリソースを変更または更新したい場合、次の方法で リソース エディタ のリソースを切り替える必要があります。

方法 2:リソース エディタ のリソースの切り替え

インタラクティブ セッションで異なるリソースを使用したい場合はいつでも、[リソースを切り替え] ダイアログ ボックスを使用してこれらのリソースを交換することができます。これは、既存のカテゴリ作業を再利用したいがリソースを置き換える場合に特に役立ちます。この場合、テキスト マイニング モデル作成ノードの [モデル] タブの [セッション作業を使用] オプションを選択できます。オプションを選択すると、ノードのダイアログ ボックスを使用してテンプレートを再読み込みする機能が無効になり、代わりにセッション時に行った設定および変更が保持されます。ストリームを実行してインタラクティブ ワークベンチ セッションを起動し、リソース エディタ のリソースを切り替えることができます。 詳細は、14 章 p. 294 リソース テンプレートの切り替え を参照してください。

リソースなど後続のセッションにセッション作業を保持するために、リソースをノードに保存するようインタラクティブ ワークベンチ セッション内からモデル作成ノードを更新する必要があります。 詳細は、 8 章 p. 148 モデル作成ノードの更新および保存 を参照してください。

注:インタラクティブ セッションで別のテンプレートの内容に切り替える場合、ノードに表示されるテンプレートの名前は、最後に読み込まれ、コピーされたテンプレートの名前となります。これらのリソースまたは他のセッション作業を利用するには、セッションを終了する前にモデル作成ノードを更新します。

テンプレートの管理

テンプレート名の変更、テンプレートのインポートおよびエクスポート、または古いテンプレートの削除など、テンプレートを扱う上で基本的な管理方法がいくつかあります。これらの管理は [テンプレートを管理] ダイアログ ボックスで実行されます。テンプレートをインポートおよびエクスポートすると、テンプレートを他のユーザーと共有できます。 詳細は、 p. 309 テンプレートのインポートおよびエクスポート を参照してください。

注:この製品とともにインストールされた（付属の）テンプレートの名前を変更したり、削除することはできません。その代わりとして、名前を変更したい場合には、インストールしたテンプレートを開き、新しい名前でテンプレートを作成します。ユーザー定義のテンプレートは削除することができますが、付属のテンプレートを削除しようとすると、このライブラリがインストールされたときの一番最初のバージョンで置き換えられます。

図 15-6
[テンプレートを管理] ダイアログ ボックス

テンプレートの名前を変更するには

- メニューの [リソース] → [リソース テンプレートの管理] を選択します。[テンプレートを管理] ダイアログ ボックスが開きます。
- 名前を変更したいテンプレートを選択し、[名前を変更] をクリックします。表の名前のセルが編集可能なフィールドとなります。
- 新しい名前を入力し、Enter キーを押します。確認のダイアログ ボックスが開きます。
- 名前を変更する場合は、[はい] をクリックします。そうでない場合は、[いいえ] をクリックします。

テンプレートを削除するには

- メニューの [リソース] → [リソース テンプレートの管理] を選択します。[テンプレートを管理] ダイアログ ボックスが開きます。
- [テンプレートを管理] ダイアログ ボックスで、削除したいテンプレートを選択します。
- [削除] をクリックします。確認のダイアログ ボックスが開きます。
- [はい] をクリックすると削除され、[いいえ] をクリックすると操作はキャンセルされます。[はい] をクリックすると、テンプレートが削除されます。

テンプレートのインポートおよびエクスポート

テンプレートをインポートおよびエクスポートすることによって、テンプレートを他のユーザーまたはマシンと共有できます。テンプレートは内部データベースに格納されますが、ハード ドライブに *.lrt ファイルとしてエクスポートできます。

テンプレートをインポートまたはエクスポートする環境があるため、これらの機能を提供するダイアログ ボックスがいくつかあります。

- テンプレート エディタ の［テンプレートを開く］ダイアログ ボックス
- テキスト マイニング モデル作成ノードおよびテキスト リンク分析ノードの［リソースを読み込む］ダイアログ ボックス
- テンプレート エディタ および リソース エディタ の［テンプレートを管理］ダイアログ ボックス

テンプレートをインポートするには

▶ ダイアログ ボックスの **[インポート]** をクリックします。［テンプレートをインポート］ダイアログ ボックスが開きます。

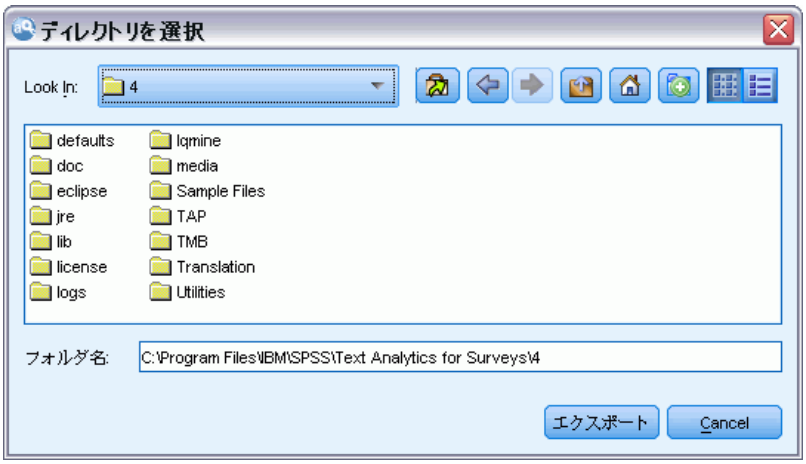
図 15-7
[テンプレートをインポート] ダイアログ ボックス

▶ インポートするリソース テンプレート ファイル（*.lrt）を選択して、**[インポート]** をクリックします。インポートしているテンプレートを別の名前で保存するか、既存のテンプレートを上書きできます。ダイアログ ボックスが閉じ、表にテンプレートが表示されます。

テンプレートをエクスポートするには

▶ ダイアログ ボックスで、エクスポートしたいテンプレートを選択し、[**エクスポート**] をクリックします。[ディレクトリを選択] ダイアログ ボックスが開きます。

図 15-8
[ディレクトリを選択] ダイアログ ボックス



▶ エクスポート先のディレクトリを選択し、[**エクスポート**] をクリックします。ダイアログ ボックスが閉じ、テンプレートがエクスポートされ、ファイルの拡張子 (*.lrt) がつきます。

テンプレート エディタ の終了

テンプレート エディタ の作業を終了したら、作業を保存してエディタを終了できます。

テンプレート エディタ を終了するには

▶ メニューから、[**ファイル**] → [**閉じる**] を選択します。[保存して閉じる] ダイアログ ボックスが開きます。

図 15-9
[保存して閉じる] ダイアログ ボックス

- エディタを閉じる前に開いているテンプレートを保存するには、[**テンプレートへの変更を保存**] を選択します。
- エディタを閉じる前に開いているテンプレートのライブラリを公開するには、[**ライブラリを公開**] を選択します。このオプションを選択すると、公開するライブラリを選択するよう要求するメッセージが表示されます。 詳細は、 16 章 p. 328 ライブラリの公開 を参照してください。

リソースのバックアップ

セキュリティ上の観点から、リソースのバックアップが必要な場合があります。

重要! 復元を実行すると、データベースの内容はすべて完全に消去され、バックアップファイルの内容しか使用できなくなります。これには処理中の作業も含まれます。

リソースをバックアップするには

- メニューの [**リソース**] → [**バックアップ ツール**] → [**リソースをバックアップ**] を選択します。[バックアップ] ダイアログ ボックスが開きます。

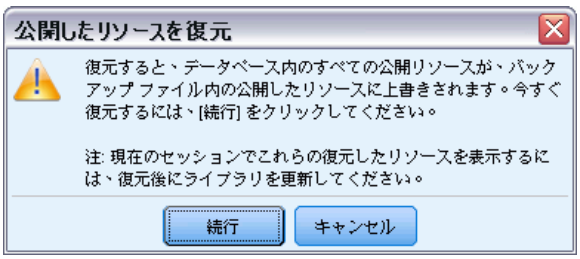
図 15-10
[リソースをバックアップ] ダイアログ ボックス

- バックアップ ファイルの名前を入力して、[**保存**] をクリックします。ダイアログ ボックスが閉じ、バックアップ ファイルが作成されます。

リソースを復元するには

▶ メニューの [リソース] → [バックアップ ツール] → [リソースを復元] を選択します。復元するとデータベースの現在の内容が上書きされることの警告が表示されます。

図 15-11
上書きの警告メッセージ



▶ [はい] をクリックして、先に進みます。 ダイアログ ボックスが開きます。

図 15-12
[リソースを復元] ダイアログ ボックス

▶ 復元したいバックアップ ファイルを選択し、[開く] をクリックします。ダイアログ ボックスが閉じ、リソースが復元されます。

重要! 復元を実行すると、データベースの内容はすべて完全に消去され、バックアップファイルの内容しか使用できなくなります。これには処理中の作業も含まれます。

リソース ファイルのインポート

この製品以外のところでリソースファイルに直接変更を加えた場合、そのライブラリを選択してインポートの手順を踏むことにより、それらのファイルを選択したライブラリにインポートできます。ディレクトリごとインポートする場合、対象となるファイルすべてを、特定の使用中のライブラリにインポートすることもできます。インポートできるのは、*.txt ファイルのみです。

重要! 日本語ファイルの場合、インポートする.txt ファイルはUTF8でエンコードする必要があります。また、日本語の不要語リストはインポートできません。注:日本語テキスト展開はIBM® SPSS® Modeler Premiumで利用可能です。

インポートしようとするファイルには、各行に項目1つを記載し、その内容は、以下の構造となります。

- 語または語句のリスト（各行に 1 つずつ）。ファイルはキーワード辞書のキーワード リストとしてインポートされ、キーワード 辞書の名前はファイル名から拡張子を除いたものとなります。日本語テキストの場合、ファイル名はインポートする既知の日本語タイプと合致する必要があります。フェイル名が基本の日本語タイプではなく感性タイプの名前に合致する場合、ファイルのキーワードが感性タイプに割り当てられ、デフォルトの基本タイプ 名詞 にも割り当てられます。
- term1<TAB>term2 のようなエントリのリストとして構成されている場合、類義語のリストとしてインポートされます。term1 は基本キーワードで term2 は代表語です。日本語テキストの場合、デフォルト値 名詞 は代表語に割り当てられます。

1 つのリソース ファイルをインポートするには

▶ メニューの [リソース] → [ファイルをインポート] → [単一ファイルをインポート] を選択します。[ファイルトをインポート] ダイアログ ボックスが開きます。

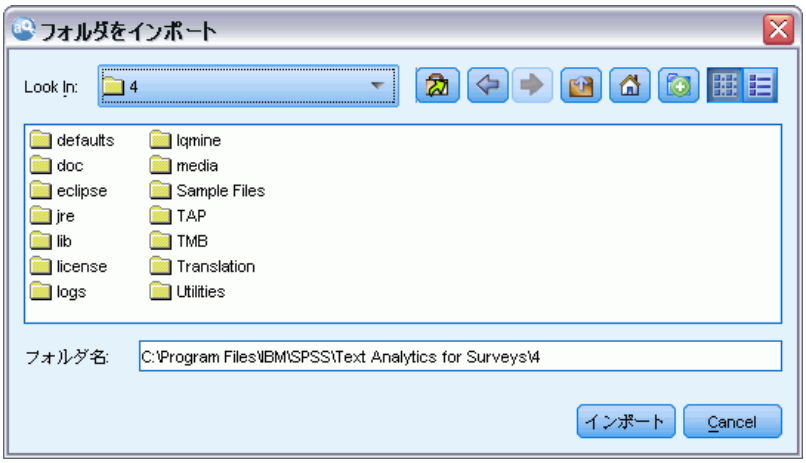
図 15-13
[ファイルをインポート] ダイアログ ボックス

▶ インポートしたいファイルを選択し、[**インポート**] をクリックします。ファイルの内容は内部形式に変換され、ライブラリに追加されます。

ディレクトリ内のすべてのファイルをインポートするには

▶ メニューの [**リソース**] → [**ファイルをインポート**] → [**フォルダ全体をインポート**] を選択します。[フォルダをインポート] ダイアログ ボックスが開きます。

図 15-14
[フォルダをインポート] ダイアログ ボックス



▶ [**インポート**] リストから、インポートしたいすべてのリソースについて、ライブラリを選択します。[**デフォルト**] オプションを選択すると、ディレクトリの名前で、新しいライブラリが作成されます。

- ファイルをインポートするディレクトリを選択します。サブディレクトリは読み込まれません。
- [**インポート**] をクリックします。ダイアログ ボックスが閉じ、インポートされたリソース ファイルの内容が辞書およびアドバンス リソース ファイルの形式でエディタに表示されます。

章
16

ライブラリの使用

テキスト データからキーワードを抽出してグループ化するために抽出エンジンで使用するリソースには、常に 1 つ以上のライブラリが含まれています。テンプレート エディタ および リソース エディタ の左上部分にあるライブラリ ツリーに一連のライブラリが表示されます。ライブラリは、３種類の辞書で構成されています：キーワード、類義語、および不要語の3種類の辞書で構成されています。 詳細は、 17 章 p. 332 ライブラリ辞書について を参照してください。

リソース テンプレートまたは選択したTAP のリソースには、複数のライブラリが含まれており、テキスト データからすぐにキーワード抽出を開始できるようになっています。しかしユーザーは、独自のライブラリを作成して、それらを再利用できるよう公開することもできます。 詳細は、 p. 328 ライブラリの公開 を参照してください。

たとえば、自動車産業に関連するテキスト データを頻繁に扱っているとします。データを分析した後、カスタム化された言語リソースを作成し、業界特有の用語や隠語を扱えるようにします。テンプレート エディタ を使用して、新しいテンプレートを作成し、テンプレート内にライブラリを作成して自動車に関するキーワードを抽出し、グループ化します。このライブラリの情報が再び必要になるため、ライブラリを中央リポジトリに公開し、**[ライブラリを管理]** ダイアログ ボックスで使用できるようにします。また異なるストリーム セッションで独立して再利用できるようになります。

また、その業界のさらに各分野(電子機器、エンジン、冷却装置、あるいは場合によっては特定の製造業者や市場) などに特有のキーワードをグループ化する必要性が出てくる場合もあります。グループごとにライブラリを作成して公開することで、テキスト データでこれを使用できます。こうすれば、自分のテキスト データのコンテキストの状況に最も適切したライブラリを追加できます。

注:追加リソースは、[アドバンス リソース] タブで設定および管理できます。一部のリソースはすべてのライブラリに適用され、固有表現、Fuzzy Grouping の例外などを管理します。また、[テキスト リンク規則] タブで、ライブラリ固有のテキスト リンク分析のパターン規則を編集できます。 詳細は、 18 章 p. 353 アドバンス リソースについて を参照してください。

付属ライブラリ

デフォルトでは、複数のライブラリが IBM® SPSS® Modeler Text Analytics と共にインストールされます。これらの事前に形式設定されたライブラリを使用して、さまざまなタイプのほか、多くの事前定義されたキーワード

や類義語を使用できます。これらの付属ライブラリは、複数の異なるドメイン向けに調整されます。また、複数の言語で使用できます。

注:日本語テキストリソースおよび例外の詳細は日本語テキストの例外 p.401 ノートを参照してください：日本語テキスト展開はIBM® SPSS® Modeler Premiumで利用可能です。

多くのライブラリがありますが、最も一般的に使用されているのは次のとおりです。

- **ローカル ライブラリ**: ユーザー定義の辞書の格納に使用します。デフォルトではすべてのリソースに追加される空のライブラリです。空白のキーワード辞書も含まれます。カテゴリとコンセプト ビュー、クラスタ ビュー、およびテキスト リンク分析ビューのリソースに直接変更または調整を行う（単語をタイプに追加するなど）場合に役立ちます。この場合、これらの変更および調整は、リソース エディタ のライブラリ ツリーに表示された最初のライブラリに自動的に保存されます。デフォルトでは、これがローカル ライブラリとなります。このライブラリはセッション データ固有であるため、このライブラリは公開できません。この内容を公開したい場合は、まずライブラリの名前を変更する必要があります。
- **コア ライブラリ**: 人名、地名、組織名、商品名、そして不明を示す 5 つの基本的なビルトインのタイプで構成されているため、多くのケースで使用されます。キーワード辞書の 1 つに記載されているのは少数のキーワードですが、コア ライブラリに記載されているタイプは、テキストマイニング製品に付属する内部のコンパイル済み辞書の頑健なタイプを補います。これらの内部コンパイル済み辞書には、各タイプの多くのキーワードが含まれています。このため、キーワード辞書のキーワードリストにキーワードは表示されませんが、コア タイプで抽出およびタイプ指定できます。つまり、Johnのみがコア ライブラリの <Person> キーワード辞書で出現する場合に、George をタイプ <Person> として抽出できます。同様に、コア ライブラリがない場合でも、それらのタイプを含むコンパイル済み辞書が抽出エンジンで使用されるため、抽出結果にそれらのタイプが表示される場合があります。
- **意見ライブラリ**: テキスト データの意見パターンを抽出する場合、最も一般的に使用されます。このライブラリには、嗜好、識別子、優先順位を示す単語が数多く含まれています。これらは-他のキーワードと連携して使用された場合に、-主題についての意見を示すものです。このライブラリには、多くのビルトインのタイプ、類義語および不要語が含まれています。また、テキスト リンク分析に使用されるパターン規則の大きなセットも含まれています。このライブラリのテキスト リンク分析規則および生成されるパターン結果を利用するには、このライブラリを[テキスト リンク規則] タブで指定する必要があります。 詳細は、19 章 p.368 テキスト リンク規則について を参照してください。

- **予算ライブラリ:** コストを参照するキーワードを抽出するために使用します。このライブラリには、価格または品質に関する形容詞、識別子、意見を示す多くの単語および句が含まれています。
- **バリエーション ライブラリ:** 特定の言語バリエーションが適切にグループ化するために類義語定義が必要なケースを追加するために使用します。ライブラリには、類義語定義のみが含まれます。

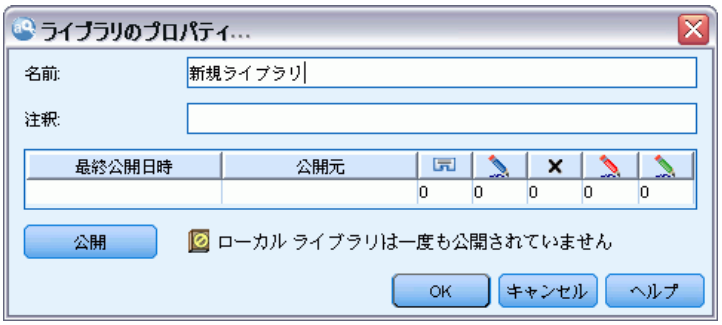
テンプレート外の付属ライブラリの一部はいくつかのテンプレートの内容に似ていますが、テンプレートは特定のアプリケーション向けに調整され、追加のアドバンス リソースを含んでいます。一般的なテンプレートに個別のライブラリを追加するのではなく、処理しているテキスト データの種類向けに作成されたテンプレートを使用し、これらのリソースに変更を行うことをお勧めします。

コンパイル済み辞書も、SPSS Modeler Text Analytics に付属しています。コンパイル済み辞書は抽出プロセスで常に使用され、デフォルト ライブラリのビルトインのキーワード辞書に対する多くの補足的定義が含まれています。これらのリソースはコンパイルされていないため、表示も編集もできません。ただし、これらのコンパイル済み辞書にタイプが割り当てられたキーワードを他の辞書に強制投入することができます。 詳細は、17 章 p. 341 キーワードの強制 を参照してください。

ライブラリの作成

ライブラリはいくつでもできます。新しいライブラリを作成した後、このライブラリ内で辞書を作成し、キーワード、類義語、不要語を入力できます。

図 16-1
[ライブラリのプロパティ] ダイアログ ボックス



ライブラリを作成するには

▶ メニューの [リソース] → [新規ライブラリ] を選択します。[ライブラリのプロパティ] ダイアログが開きます。

▶ [名前] テキスト ボックスにライブラリの名前を入力します。

- 必要に応じて、[注釈] テキスト ボックスにコメントを入力します。
- ライブラリに入力する前にこのライブラリを公開したい場合は、[**公開**] をクリックします。詳細は、 p. 326 ライブラリの共有 を参照してください。 後でいつでも公開することができます。
- [OK] をクリックしてライブラリを作成します。ダイアログ ボックスが閉じ、ツリー ビュー内にライブラリが表示されます。ツリー内のこのライブラリを展開すると、空白のキーワード辞書が自動的に含まれていることがわかります。すぐにキーワードを追加することもできます。 詳細は、17 章 p. 337 キーワードを追加 を参照してください。

パブリック ライブラリを追加

別のセッション データからライブラリを再利用したい場合、パブリック ライブラリであれば、ライブラリを現在のリソースに追加することができます。**パブリック ライブラリ**とは、公開されているライブラリです。 詳細は、 p. 328 ライブラリの公開 を参照してください。

重要! 日本語ライブラリを日本語以外のライブラリに追加、またその逆を行うことはできません。注:日本語テキスト展開はIBM® SPSS® Modeler Premiumで利用可能です。

パブリック ライブラリを追加すると、**ローカル** コピーがセッション データに埋め込まれます。このライブラリに変更することはできますが、これらの変更を共有したい場合には、このライブラリのパブリック バージョンを再度公開する必要があります。

パブリック ライブラリを追加すると、このライブラリと他のライブラリ間でキーワードのタイプが同じでない場合、[競合を解決してください] ダイアログ ボックスが表示されます。これらの競合を自分で解決するか、あるいはそこに提示された解決方法を承認して、この操作を完了する必要があります。 詳細は、 p. 330 競合の解決 を参照してください。

注:インタラクティブ ワークベンチ セッションを起動する、またはセッションを閉じて公開するときにライブラリを更新する場合、ライブラリが同期しない可能性が低くなります。 詳細は、 p. 326 ライブラリの共有 を参照してください。

図 16-2
[ライブラリを追加] ダイアログ ボックス

ライブラリを追加するには

- ▶ メニューの [**リソース**] → [**ライブラリを追加**] を選択します。[ライブラリを追加] ダイアログ ボックスが開きます。
- ▶ リストのライブラリを選択します。
- ▶ [**追加**] をクリックします。新しく追加されたライブラリと既存のライブラリとの間に競合箇所がある場合には、この競合を解決するか、ライブラリを変更しないと次に進めません。 詳細は、 p. 330 競合の解決 を参照してください。

キーワードおよびタイプの検索

エディタ内の [検索] 機能を使用して、様々な場所を検索できます。エディタのメニューから [**編集**] → [**検索**] を選択すると、検索ツールバーが表示されます。このツールバーを使用して、一回に 1 つの出現ずつ検索できます。もう一度 [**検索**] をクリックすると、次に出現しているキーワードが検索できます。

検索するとき、エディタは検索ツールバーのドロップダウン リストに表示されている1つあるいは複数のライブラリのみを検索します。[**すべてのライブラリ**] が選択されている場合、エディタ内のすべてが検索されます。

検索を開始すると、対象となっている部分から検索を始めます。検索はセッションごとに行われ、開始位置であるアクティブなセルに戻るまでループします。矢印を使うことで、検索の順番を逆にできます。検索で大文字と小文字を区別するかどうかを選択することもできます。

ウィンドウ内の文字列を検索するには

- ▶ メニューから [**編集**] → [**検索**] を選択します。検索ツールバーが表示されます。

- 検索したい文字列を入力します。
- [検索] ボタンをクリックして検索を開始します。該当するキーワードまたはタイプの出現が強調表示されます。
- ボタンをもう一度クリックして次の出現しているものに移動します。

ライブラリの表示

ある特定のライブラリまたはすべてのライブラリの内容を表示できます。これは、ライブラリがたくさんあったり、あるいは、ある特定のライブラリの内容を公開前に確認する際に便利です。ビューを変更しても、この［ライブラリ リソース］タブの表示内容が変わるだけで、これによって抽出のプロセスでライブラリが使用されなくなる、というものではありません。詳細は、 p. 322 ローカル ライブラリを使用不可に を参照してください。

デフォルトのビューは [すべてのライブラリ] で、これはツリー内にすべてのライブラリを、また他のウィンドウにその内容が表示するものです。ツールバーのドロップダウン リストまたはメニューの選択（[表示] → [ライブラリ]）によってこの選択範囲を変更できます。1 つのライブラリが表示されている場合、他のライブラリのすべての項目がビューから表示されなくなりますが、抽出時に読み取ることができます。

ライブラリ ビューを変更するには

- ［ライブラリ リソース］タブのメニューで [表示] → [ライブラリ] を選択します。すべてのローカル ライブラリを含んだメニューが開きます。
- 1つのライブラリを選択するか、あるいは [すべてのライブラリ] を選択して、その内容を表示させます。ウィンドウの内容はこの選択によって変わってきます。

ローカル ライブラリの管理

パブリック ライブラリに対し、ローカル ライブラリはインタラクティブ ワークベンチ セッション内またはテンプレート内のライブラリです。詳細は、 p. 323 パブリック ライブラリの管理 を参照してください。 ローカル ライブラリの基本的な管理方法としては次のようなものがあります。ローカル ライブラリの名前の変更、無効化、削除。

ローカル ライブラリの名前の変更

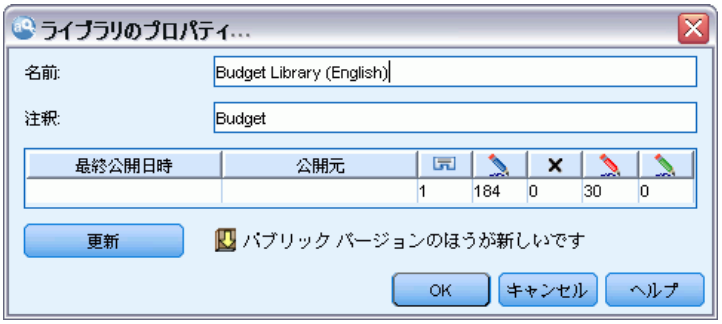
ローカル ライブラリの名前は変更できます。ローカルライブラリの名前を変更した場合、これと同じパブリックバージョンがあった場合、それらの関係性はなくなってしまいます。つまり、それ以降の変更はこのパブリッ

クライブラリと共有されることはないということです。ローカルライブラリを新しい名前のものとして再度公開することはできます。この場合にも、このローカルライブラリにおける変更は、元の名前のパブリックライブラリに対して反映されません。

注:パブリック ライブラリの名前は変更できません。

- メニューから [編集] → [ライブラリのプロパティ] を選択します。[ライブラリのプロパティ] ダイアログ ボックスが開きます。

図 16-3
[ライブラリのプロパティ] ダイアログ ボックス



ローカル ライブラリの名前を変更するには

- ツリー ビュー内で、名前を変更したいライブラリを選択します。
- [名前] テキスト ボックスにライブラリの新しい名前を入力します。
- [OK] をクリックし、ライブラリの新しい名前を確定します。ダイアログ ボックスが閉じ、ツリービュー内にあるライブラリ名が更新されます。

ローカル ライブラリを使用不可に

抽出プロセスからライブラリを一時的に除外したい場合には、ツリー ビュー内で、このライブラリの名前の左側にあるチェックボックスをオフにします。これによって、このライブラリはプロジェクト内に保持されますが、その内容は、競合のチェックならびに抽出のプロセスでは無視されるようになります。

ライブラリを無効化するには

- ライブラリ ツリー パネルで、使用しないライブラリを選択します。
- スペースバーをクリックします。名前の左側にあるチェック ボックスがオフになります。

ローカル ライブラリの削除

パブリック バージョンのライブラリを削除せずにライブラリを削除することができます。その逆も可能です。ローカル ライブラリを削除すると、セッションのみのライブラリおよびすべての内容が削除されます。ローカル バージョンのライブラリを削除しても、他のセッションまたはパブリック バージョンのライブラリは削除されません。 詳細は、 p. 323 パブリック ライブラリの管理 を参照してください。

ローカル ライブラリを削除するには

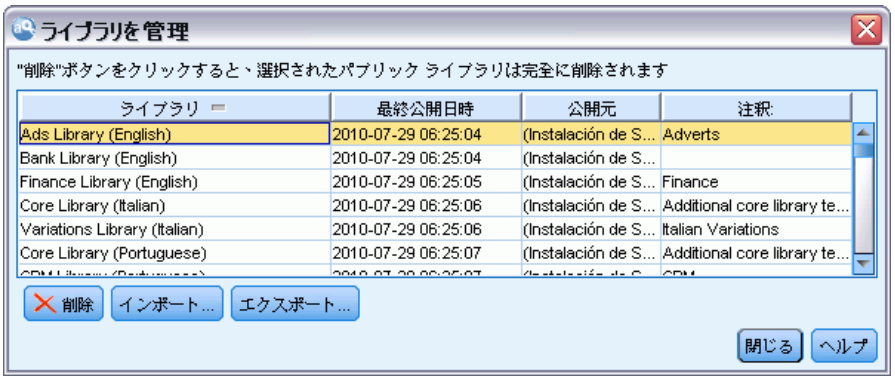
- ツリー ビューで、削除したいライブラリを選択します。
- ライブラリを削除するには、メニューから、[**編集**] → [**削除**] を選択します。ライブラリは削除されます。
- このライブラリを公開したことがない場合には、このライブラリを削除するか保存するかを尋ねるメッセージが表示されます。[**削除**] をクリックして次に進むか、[**保持**] をクリックし、このライブラリを保持します。

  注:1 つのライブラリは必ず保持する必要があります。

パブリック ライブラリの管理

ローカル ライブラリを再利用するために、ローカル ライブラリを公開して処理し、[ライブラリを管理] ダイアログ ボックスに表示することができます ([**リソース**] → [**ライブラリを管理**])。詳細は、 p. 326 ライブラリの共有 を参照してください。 パブリックライブラリの基本的な管理方法としては、パブリック ライブラリのインポート、エクスポート、または削除があります。パブリック ライブラリの名前は変更できません。

図 16-4
[ライブラリを管理] ダイアログ ボックス



パブリック ライブラリのインポート

▶ ［ライブラリを管理］ダイアログ ボックスの［**インポート...**］をクリックします。［ライブラリをインポート］ダイアログ ボックスが開きます。

図 16-5
［ライブラリをインポート］ダイアログ ボックス

▶ インポートしたいライブラリ（*.lib）を選択し、このライブラリをローカルに追加したい場合、［**現在のプロジェクトにライブラリを追加**］のチェックボックスをオンにします。

▶ ［**インポート**］をクリックします。ダイアログ ボックスが閉じます。同じ名前のパブリック ライブラリがすでに存在する場合、インポートしようとしているライブラリの名前を変更するか、あるいは現在のパブリックライブラリを上書きするのかをたずねるメッセージが表示されます。

パブリック ライブラリのエクスポート

パブリック ライブラリを .lib 形式でエクスポートすると、ライブラリを共有できるようになります。

▶ ［ライブラリを管理］ダイアログ ボックスで、リストからエクスポートしたいライブラリを選択します。

▶ ［**エクスポート**］をクリックします。［ディレクトリを選択］ダイアログ ボックスが開きます。

図 16-6
[ディレクトリを選択] ダイアログ ボックス

▶ エクスポート先のディレクトリを選択し、[**エクスポート**] をクリックします。ダイアログ ボックスが閉じ、ライブラリ ファイル (*.lib) がエクスポートされます。

パブリック ライブラリの削除

パブリック バージョンのライブラリを削除せずにローカル ライブラリを削除することができます。その逆も可能です。ただし、ライブラリがこのダイアログ ボックスから削除されると、ローカル バージョンがもう一度公開されるまでセッション リソースに追加できなくなります。

製品とともにインストールされたライブラリを削除すると、最初にインストールされていたバージョンが復元されます。

▶ [ライブラリを管理] ダイアログ ボックスで、削除したいライブラリを選択します。該当する見出しをクリックして、リストをソートすることができます。

▶ [**削除**] をクリックしてライブラリを削除します。IBM® SPSS® Modeler Text Analytics が、ローカル バージョンのライブラリがパブリック ライブラリと同じかどうかを検証します。。同じであった場合には、警告なくこのライブラリが削除されます。しかしライブラリのバージョンが異なっている場合、パブリックバージョンを保持するのか、あるいは削除するのかをたずねる警告が表示されます。

ライブラリの共有

ライブラリを使用して、複数のインタラクティブ ワークベンチ セッション間で共有しやすい方法でリソースを扱うことができます。ライブラリには２つの状態、すなわち２つのバージョン（版）があります。エディタで編集可能でインタラクティブ ワークベンチ セッションの一部であるライブラリは**ローカル ライブラリ**と呼ばれます。インタラクティブ ワークベンチ セッションで作業している間、たとえば 野菜ライブラリに多くの変更を加えることができます。変更が他のデータでも役立つ場合、この野菜というライブラリの**パブリック ライブラリ** 版を作成することで、これらのリソースを他でも使用できるようになります。パブリック ライブラリは、他のインタラクティブ ワークベンチ セッションのリソースに使用可能です。

［ライブラリを管理］ダイアログ ボックスにパブリック ライブラリが表示されます。このようなパブリックバージョンのライブラリは、他の文脈のリソースに追加できます。こうすることで、ユーザーが作成したカスタムの言語リソースを他でも活用できます。

付属ライブラリ（インストール時に含まれるライブラリ）は、最初はパブリックライブラリです。これらのライブラリ内のリソースを編集してから、これを新しいパブリックバージョンとすることも可能です。これらの新しいバージョンを、他のインタラクティブ ワークベンチ セッションで使用できるようになります。

自分のライブラリを使って作業をし、これに変更を加えていった場合、このライブラリと他のバージョンのライブラリが同期しなくなってきます。場合によっては、ローカルバージョンのほうがパブリックバージョンより最新であったり、また反対に、パブリックバージョンのほうがローカルバージョンよりも最新であったりします。他のインタラクティブ ワークベンチ セッション内からパブリック バージョンが更新された場合、一方のライブラリに含まれていない変更がパブリック バージョンおよびローカル バージョンに含まれている場合もあります。ライブラリの個々のバージョンの同期がなくなった場合、これらを再度同期させることができます。ライブラリバージョンの同期は、ローカルライブラリの再公開や更新によって行います。

インタラクティブ・ワークベンチ・セッションをローンチ、あるいは閉じる、する際にはいつでも、アップデートや再発行が必要なライブラリの同期化を試みる必要があります。またローカルライブラリの同期状況は、ツリービュー内のライブラリ名の隣にあるアイコンや、［ライブラリのプロパティ］ダイアログボックスを表示することによって簡単にわかります。また、同期はメニューからいつでも行うことができます。次の表は、あり得る5つの状態とそれに対応するアイコンです。

テーブル 16-1
ローカル ライブラリの同期の状態

アイコン	ローカル ライブラリの状態の説明
	未公開–ローカル ライブラリは公開されていません。
	同期–ローカル ライブラリのバージョンおよびパブリック ライブラリのバージョンが同じです。ローカル ライブラリにも適用されます。ローカル ライブラリはセッション固有のリソースのみを含むとされているため、公開できません。
	古い–パブリック ライブラリ バージョンの方がローカル バージョンに比べて新しいものです。ローカルバージョンを更新して、これらの変更を反映させます。
	新しい–ローカル ライブラリ バージョンの方がパブリック バージョンに比べて新しいものです。このローカルバージョンをパブリックバージョンとして再公開できます。
	非同期–ローカル ライブラリおよびパブリック ライブラリに、一方には含まれていない変更があります。この場合、ローカルライブラリを更新するのか、あるいはこれを公開するのかを決めなくてはなりません。更新を選んだ場合、最後に更新あるいは公開した以降の変更はすべて失われます。一方、公開を選んだ場合、パブリックバージョンに加えられた変更は上書きされて失われます。

注:インタラクティブ ワークベンチ セッションを起動する、またはセッションを閉じて公開するときにライブラリを更新する場合、ライブラリが同期しない可能性が低くなります。

ライブラリに変更するとこのライブラリを含む他のストリームに役立つと考えた場合はいつでも、ライブラリを再公開できます。変更すると他のストリームに役立つ場合は、それらのストリームのローカル バージョンを更新できます。このように、新しいライブラリを作成または多くのパブリック ライブラリをリソースに追加することによってデータに適用される各コンテキストまたはドメインのストリームを作成できます。

あるパブリックライブラリが共有されている場合、ローカルバージョンとパブリックバージョンの間で差異が出てくる可能性は高くなります。インタラクティブ ワークベンチ セッションから起動または閉じて公開する、または テンプレート エディタ からテンプレートを開くまたは閉じる場合、[ライブラリを管理] ダイアログ ボックスでライブラリと同期していないバージョンのライブラリを公開または更新できるメッセージが表示されます。パブリックライブラリのバージョンがこのローカルバージョンよりも新しい場合、更新するかどうかをたずねるダイアログボックスが開きます。パブリックバージョンで更新する代わりに、現在のローカルバージョンを保持するのか、あるいは更新を現在のローカルライブラリに反映させるのかを選択できます。

ライブラリの公開

ライブラリをこれまで公開していなかった場合、これを公開するとデータベース内に、ローカルライブラリのパブリックコピーが作成されます。ライブラリを再度公開した場合、ローカルライブラリの内容が、既存のパブリックバージョンの内容を置き換えます。再公開した後、他のストリームセッションのこのライブラリを更新し、ローカル バージョンがパブリックバージョンと同期するようにできます。ライブラリを公開できる場合でも、ローカル バージョンは常にセッションに格納されます。

重要! ローカルライブラリが変更されており、またこれに対応するパブリックライブラリも変更されている場合、これは同期していない（非同期）と見なされます。このような場合、まず変更されたパブリックバージョンでローカルバージョンを更新し、次にローカルバージョンを再度公開することで、両方のバージョンを全く同じにすることを推奨します。変更が加えられたローカルバージョンを先に公開してしまうと、パブリックバージョンの変更が上書きされてしまいます。

図 16-7
[ライブラリを公開] ダイアログ ボックス

ローカル ライブラリをデータベースに公開するには

- メニューの [リソース] → [ライブラリを公開] を選択します。［ライブラリを公開］ダイアログ ボックスが開き、デフォルトでは、公開の必要があるライブラリがすべて選択されています。
- 公開または再公開したい各ライブラリの左側にあるチェック ボックスをオンにします。
- [公開] をクリックし、このライブラリを［ライブラリを管理］データベースに公開します。

ライブラリの更新

インタラクティブ ワークベンチ セッションを起動または閉じる場合、パブリック バージョンと同期しないライブラリを更新または公開できます。パブリックライブラリのバージョンがローカルバージョンよりも新しい場合、このライブラリを更新するかどうかをたずねるダイアログボックスが開きます。パブリックバージョンで更新せずに現在のローカルバージョンを保持するのか、あるいは現在のプロジェクト内のローカルバージョンをパブリックのもので置き換えるのかを選択できます。パブリックバージョンがローカルバージョンよりも新しい場合、ローカルバージョンを更新して、パブリックバージョンの内容と同期させることができます。更新とは、パブリックバージョン内の変更を、ローカルバージョンに適用することです。

注:インタラクティブ ワークベンチ セッションを起動する、またはセッションを閉じて公開するときにライブラリを更新する場合、ライブラリが同期しない可能性が低くなります。 詳細は、 p. 326 ライブラリの共有 を参照してください。

図 16-8
[ライブラリを更新] ダイアログ ボックス

ローカル ライブラリを更新するには

- メニューの [リソース] → [ライブラリを更新] を選択します。。[ライブラリを更新] ダイアログボックスが開き、デフォルトでは、更新の必要のあるライブラリがすべて選択されています。
- 公開または再公開したい各ライブラリの左側にあるチェック ボックスをオンにします。
- [更新] をクリックし、ローカル ライブラリを更新します。

競合の解決

ローカル ライブラリとパブリック ライブラリの競合

ストリーム セッションを起動すると、IBM® SPSS® Modeler Text Analytics が、ローカル ライブラリと［ライブラリを管理］ダイアログ ボックスに表示されたライブラリとの比較を行います。セッションのローカル ライブラリがパブリック バージョンと同期していない場合、［ライブラリの同期］ダイアログ ボックスが開きます。ここで使用したいライブラリのバージョンを選択する際には、次のようないくつかの方法があります。

- **すべてをローカル ライブラリで:** このオプションでは、ローカルライブラリをすべてそのまま保持します。いつでもこれらを再公開あるいは更新できます。
- **すべてをこのマシン上の公開済みライブラリで:** このオプションでは、表示されたローカルライブラリをデータベース内のバージョンで置き換えます。
- **すべてを最新のライブラリで:** このオプションでは、表示されたローカルライブラリをデータベース内のバージョンで置き換えます。
- **その他:** このオプションでは、使用したいバージョンをユーザーが表から選びます。

強制キーワードの競合

パブリックライブラリを追加したり、あるいはローカルライブラリを更新した場合、リソース内で、当該ライブラリと他のライブラリとの間で、キーワードやタイプの競合や重複が発見されることがあります。この場合、提示された競合の解決方法から選択します。あるいは競合や重複を変更するための、［強制キーワードを編集］ダイアログボックスが表示されます。 詳細は、 17 章 p.341 キーワードの強制 を参照してください。

図 16-9
［強制キーワードを編集］ダイアログ ボックス

［強制キーワードを編集］ダイアログ ボックスには、競合するキーワードまたはタイプのペアが表示されます。各ペアを見やすくするために、異なる背景色が使われています。これらの色は、［オプション］ダイアログ ボックスで変更できます。詳細は、 8 章 p. 145 オプション：［表示］タブを参照してください。［強制キーワードを編集］ダイアログ ボックスには、次の 2 つのタブがあります。

- **重複:** このタブには、このライブラリ内に含まれる重複するキーワードが表示されます。キーワード部分にある押しピンのアイコンは、キーワードのこの出現が強制されていることを表わします。黒いXのアイコンがある場合、他のところで強制されているため、キーワードのこの出現は抽出時に無視されることを示します。
- **ユーザー定義:** このタブには、競合ではなく、キーワード辞書のキーワードパネルにおいて手作業で強制されたキーワードのリストが含まれています。

注:ライブラリを追加・更新すると、［強制キーワードを編集］ダイアログ ボックスが開きます。このダイアログ ボックスをキャンセルしても、ライブラリの更新または追加はキャンセルされません。

競合を解決するには

▶ ［強制キーワードを編集］ダイアログ ボックスで、強制したいキーワードの［使用］列で選択します。

▶ 完了したら、[OK] をクリックして強制キーワードを適用し、ダイアログ ボックスを閉じます。[**キャンセル**] をクリックすると、このダイアログ ボックスで行った変更がキャンセルされます。

章
17

ライブラリ辞書について

テキスト データの抽出に使用されるこれらのリソースは、テンプレートおよびライブラリの形式で保存されています。ライブラリは、3 つの辞書で構成されています。

- The**キーワード辞書**には、1 つのラベル、またはタイプ名に基づいてグループ化されたキーワードの集合が含まれています。抽出エンジンがテキスト データを読み取る場合、テキストの単語を、キーワード辞書で定義したキーワードと比較します。抽出時、タイプのキーワードおよび類義語の活用形が、コンセプトという代表語にグループ化されます。抽出されたコンセプトは、キーワードとして出現するキーワード辞書に割り当てられます。エディタの左上のウィンドウと中央のパネル（ライブラリ ツリーとキーワードのパネル）でキーワード辞書を管理できます。詳細は、 p. 332 キーワード辞書 を参照してください。
- **類義語辞書**には、最終抽出結果で、類義語、またはコンセプトという 1 つの代表語の下で類似したキーワードをグループ化するために使用する類義語またはオプションの要素として定義される単語の集合が含まれています。[類義語] タブおよび [オプション] タブを使用してエディタの左下のパネルで、類義語辞書を管理できます。詳細は、 p. 344 類義語辞書 を参照してください。
- **不要語辞書**には、最終抽出結果から削除されるキーワードおよびタイプの集合が含まれています。エディタの一番右側のパネルで不要語辞書を管理できます。 詳細は、 p. 350 不要語辞書 を参照してください。

詳細は、 16 章 p. 316 ライブラリの使用 を参照してください。

キーワード辞書

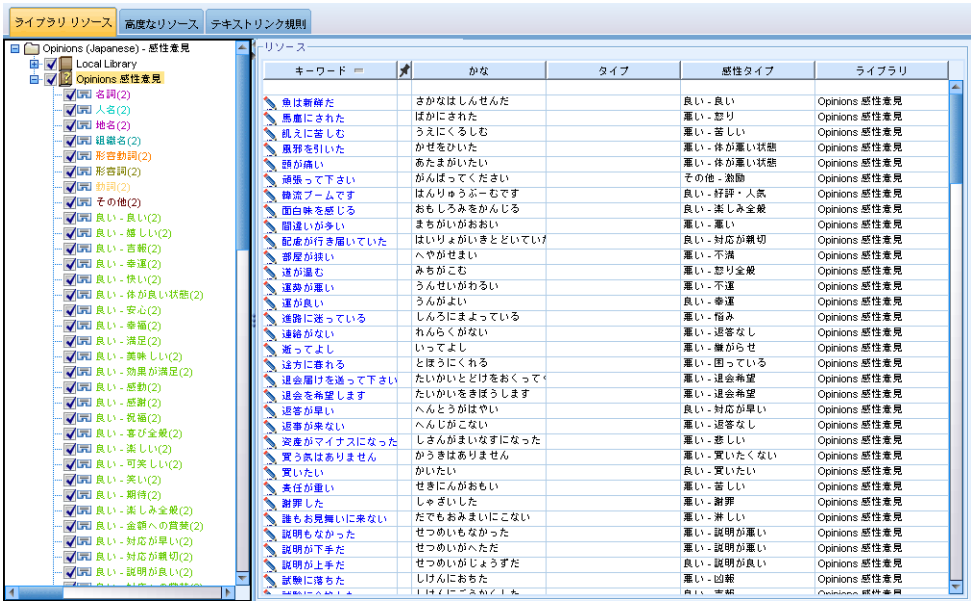
キーワード辞書は、1 つのタイプ名、またはラベル、およびキーワードのリストで構成されています。キーワード辞書は、エディタの [ライブラリ リソース] タブの左上のパネルおよび中央のパネルで管理されます。 インタラクティブワークベンチセッションで作業中の場合、メニュー内の**ビュー >リソースエディタ**からこのビューにアクセスできます。そうでない場合は、テンプレート エディタ内で該当のテンプレート用の辞書を編集できます。

抽出エンジンがテキスト データを読み取る場合、テキストの単語を、キーワード辞書で定義したキーワードと比較します。キーワードは言語リソースのキーワード辞書にある語および句です。

単語がキーワードに一致した場合、そのキーワードにタイプ名が割り当てられます。抽出時にリソースが読み込まれた場合、テキスト内で見つかったキーワードはいくつかの処理手順を経て、[抽出結果] パネルでコン

セプトとなります。同じキーワード辞書に含まれる複数のキーワードが抽出エンジンによって類義語と判断される場合、最も頻繁に出現するキーワードに基づいてグループ化され、[抽出結果] パネルで**コンセプト**として表示されます。たとえば、キーワード question および query は、最終的にコンセプト名 question で表示されます。

図 17-1
ライブラリ ツリーおよびキーワードパネル



キーワード辞書のリストが、左側のライブラリ ツリー パネルに表示されます。各キーワード辞書の内容は、中央のパネルに表示されます。キーワード辞書には、キーワードのリスト以上のものが含まれています。テキスト データの語および語句がキーワード辞書で定義されたキーワードに一致する方法は、定義されたマッチ オプションによって決まります。**マッチ オプション**は、キーワードがテキストデータの候補語またあは候補句に関してどのように固定されているかを指定します。 詳細は、 p. 337 キーワードを追加 を参照してください。

注:マッチ オプションや活用形など、すべてのオプションが日本語テキストに適用されるわけではありません。詳細は、 A 付録 p. 416 日本語のタイプのプロパティの編集 を参照してください。注:日本語テキスト展開はIBM® SPSS® Modeler Premiumで利用可能です。

また、キーワードの活用形を自動的に生成して辞書に追加するかどうかを指定して、キーワード辞書のキーワードを拡張することができます。活用形を生成して、単数形の複数形、複数形の単数形、および形容詞をキー

ワードに自動的に追加することができます。 詳細は、 p. 337 キーワードを追加 を参照してください。

注: ほとんどの言語の場合、キーワード辞書になく、テキストから抽出されたコンセプトは、自動的に<Unknown>のタイプとなりますが、日本語テキストの場合は、自動的に<**名詞**>ノートのタイプとなります：日本語テキスト展開はSPSS Modeler Premiumで利用可能です。

ビルトインのタイプ

IBM® SPSS® Modeler Text Analytics は付属ライブラリおよびコンパイル済み辞書の形式で一連の言語リソースに付属しています。付属ライブラリには、<**地名**>、<**組織名**>、<**人名**>、および <**商品名**> を含むビルトインのキーワード辞書が含まれています。

注:デフォルトのセット、ビルトインのタイプは日本語テキストでは異なります。日本語リソースに付属するタイプの詳細は、「日本語テキストで使用できるタイプ」を参照してください。注:日本語テキスト展開はIBM® SPSS® Modeler Premiumで利用可能です。

これらのキーワード辞書は、抽出エンジンによって使用され、タイプ <Location> をコンセプト paris に割り当てるように、タイプを抽出したコンセプトに割り当てます。多くのキーワードがビルトインのキーワード辞書で定義されていますが、すべての可能性をカバーしているわけではありません。そのため、辞書に追加するか独自に作成することができます。特定の付属キーワード辞書の内容の詳細は、[タイプのプロパティ] ダイアログ ボックスの注釈を参照してください。ツリーでタイプを選択し、コンテキスト メニューから **[編集] → [プロパティ]** をクリックします。

注:付属ライブラリのほか、(抽出エンジンにも使用される) コンパイル済み辞書には、ビルトインのキーワード辞書に対する定義の多くの補足が含まれていますが、その内容は製品では表示されません。ただし、これらのコンパイル済み辞書にタイプが割り当てられたキーワードを他の辞書に強制投入することができます。 詳細は、 p. 341 キーワードの強制 を参照してください。

タイプの作成

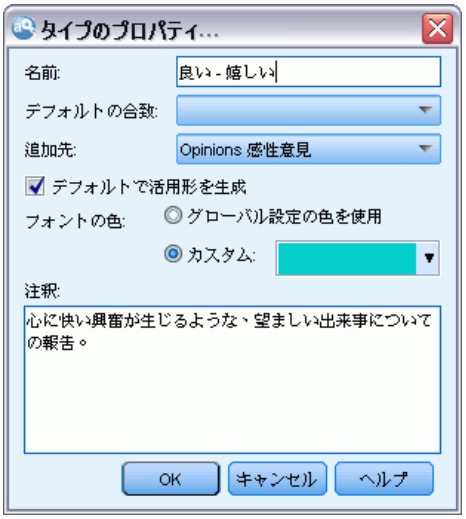
キーワード辞書を作成して、類似したキーワードをグループ化できます。この辞書に出現するキーワードが抽出プロセスで見つかった場合、キーワードはそのタイプ名に割り当てられ、コンセプト名で抽出されます。ライブラリを作成すると、すぐにキーワードを入力できるような空白のキーワード辞書が含まれます。

重要:日本語リソースの新しいタイプを作成することはできません。日本語リソースのキーワード辞書の詳細は、「日本語のライブラリ ツリー、タイプ、キーワードのパネル」を参照してください。注:日本語テキスト展開はIBM® SPSS® Modeler Premiumで利用可能です。

食べ物に関するテキストを分析しており、野菜に関連するキーワードをグループ化する場合、独自の <Vegetables> キーワード辞書を作成できます。テキストに出現する重要なキーワードと感じた場合は、carrot、broccoli、および spinach などのキーワードを追加できます。抽出時、これらのキーワードのいずれかが見つかった場合、コンセプトとして抽出され、<Vegetables> タイプに割り当てられます。

キーワードの活用形を生成できるため、単語または表現のすべての形式を定義する必要はありません。このオプションを選択すると、抽出エンジンは、他の形式からキーワードの単数形または複数形を、このタイプに割り当てられているものとして自動的に認識します。このオプションは、動詞または形容詞の活用形が必要な場合が少ないため、タイプには主に名詞が含まれている場合に役立ちます。

図 17-2
[タイプのプロパティ] ダイアログ ボックス



名前: 作成しているキーワード辞書に指定する名前。複数のタイプ名が同じ単語で始まる場合は特に、タイプ名にスペースを使用しないことをお勧めします。

注:タイプ名や記号の使用にはいくつかの制限があります。例えば、タイプ名に「@」や「!」などの記号は使用しないでください。

デフォルトの合致: デフォルトの合致属性は、抽出エンジンにこのキーワードがテキスト データにどのように合致するかを指示します。キーワードをキーワード辞書に追加すると、これが自動的キーワードに割り当てられる

合致属性となります。キーワード リストで合致の選択を手動でいつでも選択することができます。オプションの内容：オプションには、[**キーワード全体**]、[**語頭**]、[**語末**]、[**どこでも**]、[**語頭あるいは語末**]、[**完全かつ語頭**]、[**完全かつ語末**]、[**完全かつ（語頭あるいは語末）**]、および [**全体（複合語なし）**] があります。詳細は、 p. 337 キーワードを追加 を参照してください。 このオプションは、日本語リソースには適用されません。

追加先: 新しいキーワード辞書を作成するライブラリを指定します。

デフォルトで活用形を生成: 抽出エンジンに、文法的形態論を使用して、キーワードの単数形または複数形など、この辞書に追加するキーワードの類似した形式をキャプチャおよびグループ化するよう指示します。このオプションは、タイプにほとんど名詞が含まれている場合に特に役立ちます。このオプションを選択すると、このタイプに追加されたすべての新しいキーワードには自動的にこのオプションが与えられますが、リストで手動で変更できます。このオプションは、日本語リソースには適用されません。

フォントの色: このフィールドを指定すると、インターフェイスでこのタイプの結果と他のタイプの結果とを区別できるようになります。[**グローバル設定の色を使用**] を選択すると、タイプのデフォルト色がこのキーワードに使用されます。このデフォルト色は、[オプション] ダイアログ ボックスで設定されます。詳細は、 8 章 p. 145 オプション：[表示] タブ を参照してください。[**カスタム**] を選択すると、ドロップダウン リストから色を選択できます。

注釈: このフィールドはオプションで、任意のコメントまたは説明に使用できます。

キーワード辞書を作成するには

- 新しいキーワード辞書を作成するライブラリを選択します。
- メニューの [**ツール**] → [**新規タイプ**] を選択します。[タイプのプロパティ] ダイアログ ボックスが開きます。
- [**名前**] テキスト ボックスにキーワード辞書の名前を入力し、必要なオプションを選択します。
- [OK] をクリックしてキーワード辞書を作成します。新しいタイプがライブラリ ツリー パネルに表示され、中央パネルに表示されます。すぐにキーワードを追加できます。詳細は、キーワードを追加を参照してください。

  注：ここでは、リソース エディタ ビューまたは テンプレート エディタでどのように変更を行うかについて説明します。[抽出結果] パネル、[データ] パネル、[カテゴリ] パネル、または他のビューの [クラスタ定義] ダイアログ ボックスで直接、この種類の調整を行うこともできます。 詳細は、 9 章 p. 168 抽出結果の調整 を参照してください。

キーワードを追加

ライブラリ ツリー パネルにはライブラリが表示され、ツリーを展開すると、ツリーに含まれているキーワード辞書が表示されます。ツリーの選択によって、中央のパネルのキーワードリストに選択したライブラリまたはキーワード辞書のキーワードが表示されます。

重要! 日本語リソースのキーワードは異なった方法で定義されます。詳細は、 A 付録 p. 409 日本語のライブラリ ツリー、タイプ、キーワードのパネル を参照してください。注:日本語テキスト展開はIBM® SPSS® Modeler Premiumで利用可能です。

図 17-3
[キーワード] パネル

Resources

Term	Match	Inflect	Type	Library
non-optimal	Entire Term		NegativeFunctioning	Opinions Library (English)
non-operative	Entire Term		NegativeFunctioning	Opinions Library (English)
non-metallic	Entire Term		Contextual	Opinions Library (English)
non-invasive	Entire Term		PositiveFeeling	Opinions Library (English)
non-intuitive	Entire Term		Negative	Opinions Library (English)
non-intrusive	Entire Term		PositiveFeeling	Opinions Library (English)
non-hostile	Entire Term		PositiveAttitude	Opinions Library (English)
non-functioning	Entire Term		NegativeFunctioning	Opinions Library (English)
non-friendly	Entire Term		NegativeAttitude	Opinions Library (English)
non-fat	Entire Term		Contextual	Opinions Library (English)
non-failing	Entire Term		PositiveFunctioning	Opinions Library (English)
non-existent	Entire Term		Negative	Opinions Library (English)
non-existant	Entire Term		Negative	Opinions Library (English)
non-essential	Entire Term		Negative	Opinions Library (English)
non-equal	Entire Term		Negative	Opinions Library (English)
non-enthusiastic	Entire (no compounds)		NegativeAttitude	Opinions Library (English)
non-enhanced	Entire Term		Negative	Opinions Library (English)
non-enforceable	Entire Term		NegativeFunctioning	Opinions Library (English)
non-effective	Entire Term		Negative	Opinions Library (English)
non-corrected	Entire Term		NegativeCompetence	Opinions Library (English)
non-cooperative	Entire Term		NegativeAttitude	Opinions Library (English)
non-conventional	Entire Term		Positive	Opinions Library (English)
non-constrained	Entire Term		Positive	Opinions Library (English)
non-complexed	Entire Term		Positive	Opinions Library (English)

リソース エディタ では、[キーワード] パネルで直接または [新しいキーワードを追加] ダイアログ ボックスを使用して、キーワードをキーワード辞書に追加できます。追加するキーワードは、単語でも複合語でもかまいません。リストの一番上に空白の行があり、そこに新しいキーワードを追加できます。

注:ここでは、リソース エディタ ビューまたは テンプレート エディタでどのように変更を行うかについて説明します。[抽出結果] パネル、[データ] パネル、[カテゴリ] パネル、または他のビューの [クラスタ定義] ダイアログ ボックスで直接、この種類の調整を行うこともできます。 詳細は、 9 章 p. 168 抽出結果の調整 を参照してください。

[キーワード] 列

この列のセルに、単語または複合語を入力します。キーワードが表示される色は、キーワードが保存または強制投入されるタイプの色によって異なります。[タイプのプロパティ] ダイアログ ボックスでタイプの色を変更できます。 詳細は、 p. 334 タイプの作成 を参照してください。

[強制] 列

このセルで押しピンのアイコンを押すと、抽出エンジンは、他のライブラリのこの同じキーワードの他の出現を無視します。 詳細は、 p. 341 キーワードの強制 を参照してください。

[合致] 列

この列ではマッチ オプションを選択して、抽出エンジンにこのキーワードがテキスト データにどのように合致するかを指示します。例については表を参照してください。タイプのプロパティを編集して、デフォルト値を変更できます。詳細は、 p. 334 タイプの作成 を参照してください。 メニューから [**編集**] → [**合致方法を変更**] を選択します。次に示すのは基本的なマッチ オプションで、これらを組み合わせて使用することもできます。

- **語頭**:辞書のキーワードがテキストから抽出したコンセプトの最初の文字に合致する場合、このタイプが割り当てられます。たとえば、apple と入力すると、apple tart が合致します。
- **語末**:辞書のキーワードがテキストから抽出したコンセプトの最後の文字に合致する場合、このタイプが割り当てられます。たとえば、apple と入力すると、cider apple が合致します。
- **どこでも**:辞書のキーワードがテキストから抽出したコンセプトの一部に合致する場合、このタイプが割り当てられます。たとえば、apple と入力すると、[**どこでも**] のタイプが、apple tart、cider apple、および cider apple tart に割り当てられます。
- **キーワード全体**:テキストから抽出したコンセプト全体が辞書キーワードにそのまま合致する場合、このタイプが割り当てられます。キーワードを [**キーワード全体**] として追加すると、[**完全かつ語頭**]、[**完全かつ語末**]、[**完全かつどこか**]、または [**全体 (複合語なし)**] がキーワードの抽出を強制します。

  さらに、<Person> タイプは、edith piaf や mohandas gandhi のように名前の 2 つの部分を抽出するため、姓について記載されていないときに名を抽出する場合、名をこのキーワード辞書に明示的に追加したい場合があります。たとえば、edith のすべてのインスタンスを名前として取得したい場合、[**キーワード全体**] または [**完全かつ語頭**] を使用して、edith を <Person> タイプに追加する必要があります。
- [**全体 (複合語なし)**]:テキストから抽出したコンセプト全体が辞書キーワードにそのまま合致する場合、このタイプが割り当てられ、キーワードが長い複合語に合致しないよう、抽出が停止します。たとえ

ば、apple と入力すると、[全体（複合語なし）] オプションは、apple にタイプを割り当て、別に強制されていない限り複合語 apple sauce は抽出されません。

次の表では、キーワード apple がキーワード辞書にあると想定します。マッチ オプションによって、この表にはテキスト内で見つかった場合に抽出およびタイプが割り当てられるコンセプトが表示されます。

テーブル 17-1
合致の例

マッチ オプション - キーワード: apple	抽出コンセプト			
	apple	apple tart	ripe apple	homemade apple tart
キーワード全体	✔			
語頭		✔		
語末			✔	
語頭あるいは語末		✔	✔	
完全かつ語頭	✔	✔		
完全かつ語末	✔		✔	
完全かつ（語頭あるいは語末）	✔	✔	✔	
どこでも		✔	✔	✔
完全かつどこか	✔	✔	✔	✔
全体（複合語なし）	✔	抽出なし	抽出なし	抽出なし

[活用] 列

この列では、抽出エンジンが抽出時にこのキーワードの活用形を生成し、すべてをグループ化するかどうかを選択します。この列のデフォルト値は［タイプのプロパティ］で定義されますが、場合によっては、列で直接このオプションを変更できます。メニューから［編集］→［活用形を変更］を選択します。

[タイプ] 列

この列で、ドロップダウン リストからキーワード辞書を選択します。タイプのリストは、ライブラリ ツリー パネルでの選択に応じてフィルタリングされます。リストの最初のタイプは、ライブラリ ツリー パネルで選択されたデフォルト タイプです。メニューから［編集］→［タイプを変更］を選択します。

[ライブラリ] 列

キーワードが格納されているライブラリが表示されます。ライブラリ ツリー パネルでキーワードを別のタイプにドラッグ アンド ドロップして、そのライブラリを変更できます。

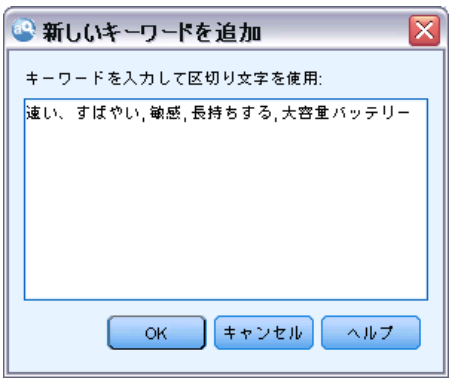
キーワード辞書に 1 つのキーワードを追加するには

- ライブラリ ツリー パネルで、キーワードを追加したいキーワード辞書を選択します。
- 中央のパネルのキーワード リストで、使用できる最初の空白セルにキーワードを入力し、このキーワードに必要なオプションを設定します。

キーワード辞書に複数のキーワードを追加するには

- ライブラリ ツリー パネルで、キーワードを追加したいキーワード辞書を選択します。
- メニューの [ツール] → [新規キーワード] を選択します。[新しいキーワードを追加] ダイアログ ボックスが開きます。

図 17-4
[新しいキーワードを追加] ダイアログ ボックス



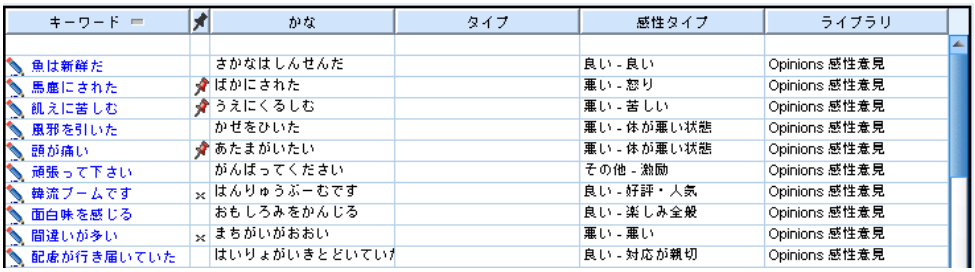
- キーワードを入力するか、キーワードのセットをコピーして貼り付けて、選択したキーワード辞書に追加したいキーワードを入力します。複数のキーワードを入力する場合、[オプション] ダイアログで定義された区切り文字を使用してキーワードを区切るか、各キーワードを新しい行に追加するつようがあります。 詳細は、 8 章 p. 143 オプションの設定 を参照してください。
- [OK] をクリックすると、キーワードが辞書に追加されます。マッチ オプションは、このキーワード ライブラリのデフォルトのオプションに自動的に設定されます。ダイアログ ボックスが閉じ、辞書に新しいキーワードが表示されます。

キーワードの強制

キーワードを特定のタイプに割り当てる場合、対応するキーワード辞書に追加することができます。ただし、同じ名前を持つ複数のキーワードがある場合、抽出エンジンはどのタイプを使用するかを認識する必要があります。そのため、使用するタイプを選択するよう要求するメッセージが表示されます。この操作をタイプへのキーワードの**強制**といいます。このオプションは、コンパイル済み（内部、編集不可能）辞書からのタイプの割り当てを上書きする場合に役立ちます。通常、キーワードが重複しないようにすることをお勧めします。

強制しても、このキーワードの他の出現を「削除」するわけではありません。抽出エンジンによって無視されます。キーワードを強制または強制を解除することによって、使用する出現を後で変更することができます。また、パブリック ライブラリを追加またはパブリック ライブラリを更新する場合、キーワードをキーワード辞書に投入することが必要な場合もあります。

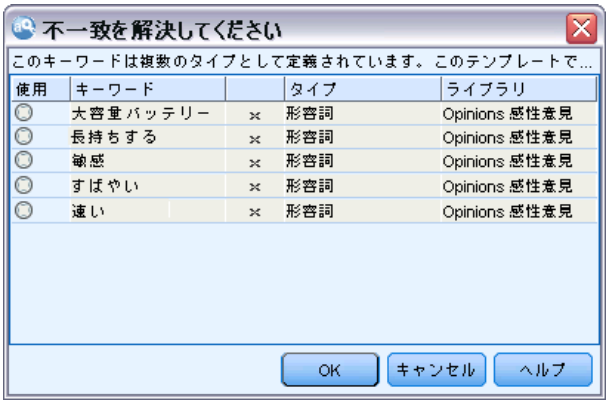
図 17-5
強制の状態のアイコン



キーワード パネルの 2 番目の列、[強制] 列で、度のキーワードが強制されているか、または無視されているかを確認できます。押しピンのアイコンが表示されている場合、キーワードのこの出現が強制されていることを示します。黒い X のアイコンがキーワードの後に表示されている場合、他の場所で強制されているため、キーワードのこの出現は抽出時に無視されることを示します。また、キーワードを強制すると、強制されたタイプの色で表示されます。**タイプ** 1 および **タイプ** 2 のキーワードが **タイプ** 1 に強制投入されると、**タイプ** 1 に定義されたフォントの色で表示されます。

アイコンをダブルクリックして、状態を変更できます。キーワードが他の場所で表示されている場合、[不一致を解決してください] ダイアログボックスが開き、使用する出現を選択できます。

図 17-6
[不一致を解決してくださ] ダイアログ ボックス



タイプの名前の変更

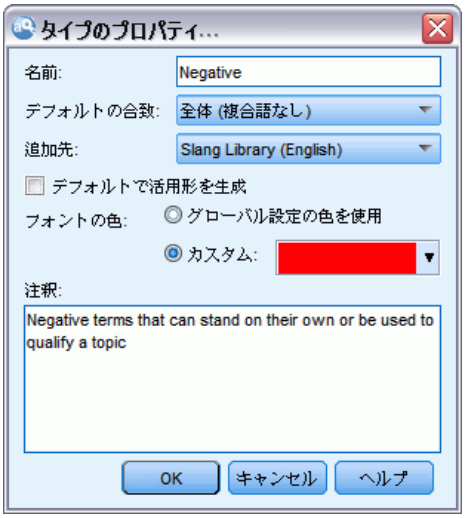
［タイプのプロパティ］を編集して、キーワード辞書の名前を変更したり、その他の辞書の設定を変更することができます。

重要! 複数のタイプ名が同じ単語で始まる場合は特に、タイプ名にスペースを使用しないことをお勧めします。コア ライブラリまたは意見ライブラリでタイプの名前を変更したり、デフォルトの合致属性を変更しないことをお勧めします。

タイプの名前を変更するには

- ライブラリ ツリー パネルで、名前を変更したいキーワード辞書を選択します。
- マウスを右クリックし、コンテキスト メニューから **[タイプのプロパティ]** をクリックします。［タイプのプロパティ］ダイアログ ボックスが開きます。

図 17-7
[タイプのプロパティ] ダイアログ ボックス



- ［名前］テキスト ボックスにキーワード辞書の新しい名前を入力します。
- [OK] をクリックして、新しい名前を承認します。新しいタイプ名がライブラリ ツリー パネルに表示されます。

タイプの移動

キーワード辞書をライブラリ内の別の場所、またはツリー内の別のライブラリにドラッグすることができます。

ライブラリ内のタイプの順序を変更するには

- ライブラリ ツリー パネルで、移動したいキーワード辞書を選択します。
- メニューで [**編集**] → [**1つ上に移動**] を選択すると、ライブラリ ツリー パネルでキーワード辞書が 1 つ上の位置に移動します。[**編集**] → [**1つ下に移動**] を選択すると 1 つ下の位置に移動します。

タイプを別のライブラリに移動するには

- ライブラリ ツリー パネルで、移動したいキーワード辞書を選択します。
- マウスを右クリックし、コンテキスト メニューから [**タイプのプロパティ**] をクリックします。［タイプのプロパティ］ダイアログ ボックスが開きます。(タイプを別のライブラリにドラッグ アンド ドロップすることができます)。

- ［追加先］リスト ボックスで、キーワード辞書を移動したいライブラリを選択します。
- [OK] をクリックします。ダイアログ ボックスが閉じ、タイプが選択したライブラリに移動します。

タイプの無効化および削除

キーワード辞書を一時的に削除したい場合、ライブラリ ツリー パネルの辞書名の左側にあるチェック ボックスをオフにして無効化することができます。これは、ライブラリ内に辞書を保存したいが、競合を検証する場合および抽出プロセス時に内容を無視することを示します。

ライブラリからキーワード辞書を永続的に削除することもできます。

キーワード辞書を無効化するには

- ライブラリ ツリー パネルで、無効化したいキーワード辞書を選択します。
- スペースバーをクリックします。タイプ名前の左側にあるチェック ボックスがオフになります。

キーワード辞書を削除するには

- ライブラリ ツリー パネルで、削除したいキーワード辞書を選択します。
- メニューから、**［編集］→［削除］**を選択すると、キーワード辞書が削除されます。

類義語辞書

類義語辞書は、1 つの代表語に基づいて、類似したキーワードをグループ化できるキーワードの集合です。類義語辞書は、［ライブラリ リソース］タブの一番下のパネルで管理されます。 インタラクティブワークベンチセッションで作業中の場合、メニュー内の**ビュー >リソースエディタ**からこのビューにアクセスできます。そうでない場合は、テンプレート エディタ内で該当のテンプレート用の辞書を編集できます。

当辞書では類義語の定義の方法が 2 つあります。**類義語**および**オプションの要素**。このパネルでタブをクリックして切り替えることができます。

テキスト データの抽出を実行した後、他のコンセプトの類義語または活用形であるコンセプトをいくつか見つけることができます。オプションの要素および類義語を特定して、抽出エンジンがこれらを 1 つの代表語にマップされるよう強制できます。

類義語やオプションの要素を使用すると、より頻度の高いドキュメント数のより重要で代表的なコンセプトに結合することによって、[抽出結果]パネルのコンセプト数を削減することができます。

注： 日本語リソースの場合、オプションの要素は適用されず、使用できません。また、日本語テキストの場合、類義語の処理が若干異なります。詳細は、 A 付録 p. 417 日本語テキストの類義語辞書の使用 を参照してください。注:日本語テキスト展開はIBM® SPSS® Modeler Premiumで利用可能です。

類義語

類義語とは、同じ意味を持つ複数の語を関連付けたものです。また、類義語はキーワードを略語とグループ化したり、一般的にスペルミスのある単語と正しいスペルの単語とをグループ化したりするために使用できます。[類義語] タブでこれらの類義語を定義できます。

類義語定義はは、2 つの部分で構成されています。1 つ目は**代表語**で、抽出エンジンがすべての類義語キーワードをグループ化する基準となるキーワードです。この代表語が別の代表語の類義語として使用されていない限り、または不要語として登録されていない限り、[抽出結果] パネルに表示されるコンセプトとなります。2 つめは、代表語の下にグループ化される類義語のリストです。

たとえば、 `automobile` という単語を `vehicle` という単語に置き換えたい場合、`automobile` が類義語となり、`vehicle` が代表コンセプトとなります。

[類義語] 列にはどんな言葉も入力できますが、その語が抽出時に見つからず、キーワードのマッチ オプションが **[キーワード全体]** である場合、類義語は出現しません。ただし、類義語をこのキーワードの下にグループ化するために、代表語を抽出する必要がありません。

図 17-8
類義語辞書、[類義語] タブ

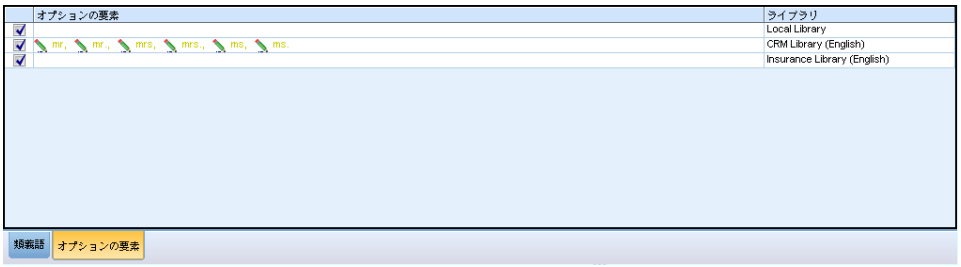
オプションの要素

オプションの要素は、テキスト内で若干異なったように出現する場合でも類義語を保持するために抽出時に無視できる結合キーワード内のにんいの語を示します。オプションの要素は単語で、結合語から削除された場合、別のキーワードと合致を作成できます。これらの単語は、結合語内のどこでも（語頭、語中、語末）出現します。[オプション] タブでオプションの要素を定義できます。

たとえば、ibm および ibm corp のキーワードをグループ化する場合、corp をオプションの要素として扱うよう宣言する必要があります。また別の例を示すと、キーワード access をオプションの要素として指定し、抽出時に internet access speed および internet speed が見つかった場合、最も頻繁に出現するキーワードに基づいてグループ化されます。

注： 日本語リソースの場合、オプションの要素が適用されないため、[オプションの要素] タブはありません。日本語テキスト展開はSPSS Modeler Premiumで利用可能です。

図 17-9
類義語辞書、[オプション] タブ



類義語の定義

[類義語] タブで、 テーブルの一番上の空白行に類義語定義を入力できます。まず代表語とその類義語を定義します。この定義を格納するライブラリも選択できます。抽出時、類義語のすべての出現を、最終的な抽出で代表語に基づいてグループ化します。 詳細は、 p. 337 キーワードを追加 を参照してください。

たとえば、テキストデータにて、多くの電気通信情報が含められている場合には、下記の単語が出てくる可能性があります：**セルラーフォン**、**ワイアレスフォン**および **モバイルフォン**. この場合、cellular および mobile を wireless の類義語として定義する必要があります。これらの類義語を定義すると、cellular phone および mobile phone の抽出されたすべての出現は、wireless phone と同じキーワードとして扱われ、キーワードリストにいっしょに表示されます。

キーワード辞書を作成している場合、キーワードを入力し、そのキーワードに対して 3 つまたは 4 つの類義語を考えることができます。この場合、類義語辞書にすべてのキーワードと代表語を入力し、類義語をドラッグすることができます。

注:日本語テキストの場合、類義語の処理が若干異なります。詳細は、A 付録 p. 417 日本語テキストの類義語辞書の使用 を参照してください。
注:日本語テキスト展開はIBM® SPSS® Modeler Premiumで利用可能です。

類義語は、類義語の活用形（複数形など）にも適用されます。コンテキストに応じて、キーワードをどのように代用するかについて制限が必要な場合があります。特定の文字を使用して、類義語の処理の程度に制限を加えることができます。

- **感嘆符 (!):**!synonym のように類義語の前に直接感嘆符を追加すると、る意義のの活用形は代表語によって代用されないことを示します。ただし、!target-term のように代表語の前に感嘆符を追加すると、複合代表語の一部または変異形がさらる類義語を受け入れないようにすることを示します。
- **アスタリスク (*):**synonym* のようにアスタリスクを類義語の直後に置くと、この語を代表語に置き換えることを示します。たとえば、manage* を類義語に、management を代表語に定義すると、associate managers が代表語 associate management に置き換えられます。また、internet * のようにごとアスタリスクの間にスペースを追加することもできます（synonym *）。代表語に internet、類義語に internet * * および web * を定義すると、internet access card および web portal は internet に置き換えられます。この辞書では、語または文字列の頭をアスタリスクにすることはできません。
- **カレット (^):**^ synonym のように類義語の前にカレットとスペースを追加すると、キーワードが類義語から始まる場合にのみ、類義語のグループ化が適用されることを示します。たとえば、 ^ wage を類義語に、income を代表語に定義して両方が抽出される場合、キーワード income に基づいてグループ化されます。ただし、minimum wage と income が抽出されると、minimum wage が wage で始まっていないため、それらはいっしょにグループ化されません。この記号と類義語の間にスペースを追加する必要があります。
- **ドル記号 ($):**synonym $ のように類義語の後にスペースドル記号を追加すると、キーワードが類義語で終わる場合にのみ、類義語のグループ化が適用されることを示します。たとえば、 cash $ を類義語に、money を代表語に定義して両方が抽出される場合、キーワード money に基づいてグループ化されます。ただし、cash cow と money が抽出されると、cash cow が cash で終わっていないため、それらはいっ

しょにグループ化されません。この記号と類義語の間にスペースを追加する必要があります。

- **カレット (^) およびドル記号 ($):** カレットとドル記号を ^ synonym $ のように同時に使用すると、完全一致の場合にのみキーワードが類義語と合致します。つまり、類義語のグループ化が行われるよう、抽出されたキーワードの類義語の前後に語があってはいけないことを意味します。たとえば、^ van $ を類義語に、truck を代表語に定義すると、van は truck とグループ化されますが、marie van guerin がそのままになります。また、カレットとドル記号を使用して類義語を定義して、この語がソース テキスト内のどこにでも出現する場合、類義語は自動的に抽出されます。

図 17-10
類義語辞書、[類義語] タブの例

注: これらの特殊文字およびワイルドカードは、日本語テキストではサポートされていません。詳細は、 A 付録 p. 417 日本語テキストの類義語辞書の使用 を参照してください。 注: 日本語テキスト展開はSPSS Modeler Premiumで利用可能です。

類義語エントリを追加するには

- ▶ 類義語パネルを表示し、左下の [類義語] タブをクリックします。
- ▶ テーブルの一番上の空白行で、[代表語] 列に代表語を入力します。入力した代表語が色つきで表示されます。この色は、キーワードが表示されるまたは強制されるタイプを示します。キーワードが黒で表示されている場合、キーワード辞書にはこのキーワードがないことを意味します。
- ▶ 代表語の右側の 2 番目のセルをクリックして、類義語のセットを入力します。[オプション] ダイアログ ボックスで定義したグローバル区切り文字を使用して、各エントリを区切ります。詳細は、 8 章 p. 143 オプションの設定 を参照してください。 入力したキーワードが色つきで表示されます。この色は、キーワードが出現するタイプを示します。キーワードが黒で表示されている場合、キーワード辞書にはこのキーワードがないことを意味します。
- ▶ 最後のセルをクリックして、この類義語定義を格納するライブラリを選択します。

注:ここでは、リソース エディタ ビューまたは テンプレート エディタでどのように変更を行うかについて説明します。[抽出結果] パネル、[データ] パネル、[カテゴリ] パネル、または他のビューの [クラスタ定義] ダイアログ ボックスで直接、この種類の調整を行うこともできます。 詳細は、 9 章 p. 168 抽出結果の調整 を参照してください。

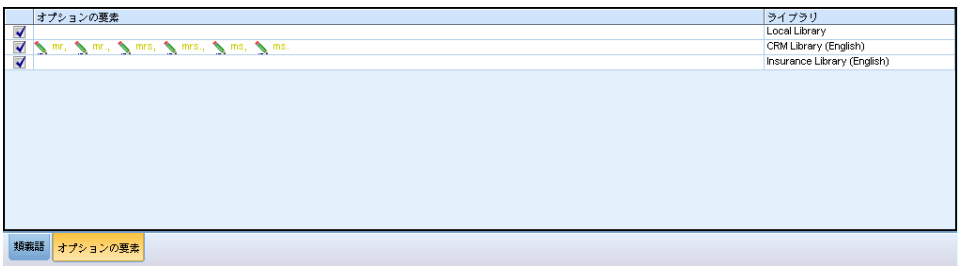
オプションの要素の定義

[オプション] タブで、必要なライブラリのオプションの要素を定義します。これらのエントリは、各ライブラリでグループ化されます。ライブラリがライブラリ ツリー パネルに追加されると、空白のオプションの要素行が [オプション] タブに追加されます。

すべてのエントリが、自動的に小文字の語に変換されます。抽出エンジンは、エントリをテキスト内の小文字および大文字の語に合致させます。

注： 日本語リソースの場合、オプションの要素は適用されず、使用できません。

図 17-11
類義語辞書、[オプション] タブ



注:[オプション] ダイアログ ボックスで定義した区切り文字を使用して、キーワードを区切ります。詳細は、 8 章 p. 143 オプションの設定 を参照してください。 入力しているオプションの要素にキーワードの一部と同じ区切り文字が含まれている場合、その前にバックスラッシュを追加する必要があります。

エントリを追加するには

- 類義語パネルを表示し、エディタの左下にある [オプション] タブをクリックします。
- このエントリを追加するライブラリの [オプションの要素] 列のセルをクリックします。
- オプションの要素を入力します。[オプション] ダイアログ ボックスで定義したグローバル区切り文字を使用して、各エントリを区切ります。 詳細は、 8 章 p. 143 オプションの設定 を参照してください。

類義語の無効化および削除

辞書でエントリを無効化して、一時的に削除することができます。エントリを無効化すると、そのエントリは抽出時に無視されます。

類義語辞書の古いエントリを削除することもできます。

エントリを無効化するには

- 辞書で、無効化したいエントリを選択します。
- スペースバーをクリックします。エントリの左側にあるチェック ボックスがオフになります。

  注:エントリの左側にあるチェック ボックスをオフにして、無効化することもできます。

類義語エントリを削除するには

- 辞書で、削除したいエントリを選択します。
- メニューから、[**編集**] → [**削除**] を選択または Del キーを押すと、キーワード辞書が削除されます。エントリは辞書内から除外されます。

オプションの要素エントリを削除するには

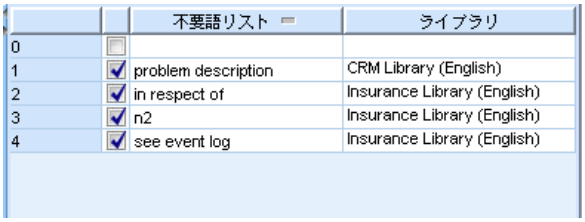
- 辞書で、削除したいエントリをダブルクリックします。
- キーワードを手動で削除します。
- Enter を押して変更を適用します。

不要語辞書

不要語辞書は、語、句、一部の文字列のリストです。不要語辞書のエントリに合致またはエントリを含むキーワードは無視されるか、抽出から除外されます。不要語辞書は、エディタの右側のパネルで管理されます。通常、このリストに追加するキーワードは、穴埋めのための単語または句で、一貫性を維持するためにテキストで使用されテキストにとって重要でなく、抽出結果を混乱させる場合があります、こういったキーワードを不要語辞書に追加することで、全く抽出されないようにできます。

不要語辞書は、エディタの［ライブラリ リソース］タブの右上のパネルで管理されます。 インタラクティブワークベンチセッションで作業中の場合、メニュー内の**ビュー >リソースエディタ**からこのビューにアクセスできます。そうでない場合は、テンプレート エディタ内で該当のテンプレート用の辞書を編集できます。

図 17-12
[不要語辞書] パネル



不要語辞書で、 テーブルの一番上の空白行に語、句、または一部の文字列を入力できます。文字列を不要語辞書に 1 つまたは複数の語として、またはアスタリスクをワイルドカードとして使用し、単語の一部として追加することができます。不要語辞書で宣言されたエントリを使用して、コンセプトを抽出から除外します。エントリが、キーワード辞書などインターフェイスの別の場所でも宣言されている場合、他の辞書では、現在除外されていることを示す取り消し線が表示されます。この文字列は、テキストデータに出現する必要はなく、また適用されるキーワード辞書の一部として宣言する必要はありません。

注:類義語エントリで代表語としても機能するコンセプトを不要語辞書に追加すると、代表語およびそのすべての類義語も不要語として登録されます。 詳細は、 p. 346 類義語の定義 を参照してください。

ワイルドカード (*) の使用

日本語のほか、すべてのテキスト言語で、不要語のエントリを一部の文字列として扱うことを示すためにアスタリスクのワイルドカードを使用します。抽出エンジンで見つかった、不要語辞書で入力した文字列で始まるまたは終わる語を含むキーワードは、最終的な抽出からは除外されます。ただし、ワイルドカードの使用が認められない場合が 2 つあります。

- *- のようにアスタリスクの前にダッシュ (-) が追加されている場合
- *’s のようにアスタリスクの前にアポストロフィー (‘) が追加されている場合

テーブル 17-2
不要語エントリの例

エントリ	例	結果
語	next	next という語が含まれている場合、コンセプト (またはそのキーワード) は抽出されません。
句	for example	for example という句が含まれている場合、コンセプト (またはそのキーワード) は抽出されません。

エントリ	例	結果
一部	copyright*	copyrighted、copyrighting、copyrights、または copyright 2010 のように、copyright の変異形に合致または変異形を含むコンセプト（またはそのキーワード）は不要語として登録されます。
一部	*ware	freeware、shareware、software、hardware、beware、または silverware のように、ware の変異形に合致または変異形を含むコンセプト（またはそのキーワード）は不要語として登録されます。

エントリを追加するには

- テーブルの一番上の空白行で、キーワードを入力します。入力したキーワードが色つきで表示されます。この色は、キーワードが出現するタイプを示します。キーワードが黒で表示されている場合、キーワード辞書にはこのキーワードがないことを意味します。

エントリを無効化するには

不要語辞書でエントリを無効化して、エントリを一時的に削除することができます。エントリを無効化すると、そのエントリは抽出時に無視されます。

- 不要語辞書で、無効化したいエントリを選択します。
- スペースバーをクリックします。エントリの左側にあるチェック ボックスがオフになります。

  注:エントリの左側にあるチェック ボックスをオフにして、無効化することもできます。

エントリを削除するには

不要語辞書の不要なエントリを削除することもできます。

- 不要語辞書で、削除したいエントリを選択します。
- メニューから［**編集**］→［**削除**］を選択します。エントリは辞書内から除外されます。

章
18

アドバンス リソースについて

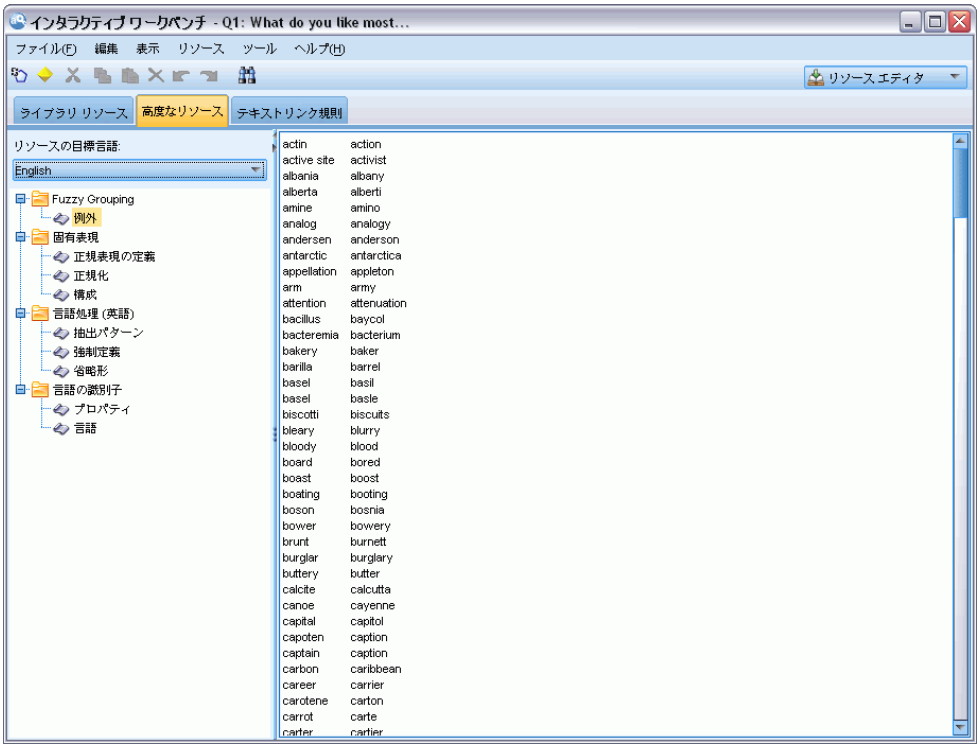
キーワード辞書、不要語辞書および類義語辞書のほか、Fuzzy Grouping 設定や固有表現キーワード辞書など、さまざまなアドバンス リソースの設定を使用することもできます。これらのリソースは、テンプレート エディタ または リソース エディタ ビューの［高度なリソース］タブで処理することができます。

重要! このタブは、日本語テキストに対して調整されたリソースには使用できません。

［アドバンス リソース］タブに移動すると、次の情報を編集できます。

- **リソースの対象言語。**リソースが作成され調整される言語の選択に使用されます。 詳細は、 p. 356 リソースの目標言語 を参照してください。
- **Fuzzy Grouping (例外):** Fuzzy Grouping（スペル修正）アルゴリズムから単語のペアを除外するために使用します。. 詳細は、 p. 357 Fuzzy Grouping を参照してください。
- **固有表現:** 抽出時に適用される正規表現や正規化規則のほか、どの固有表現を抽出できるかを有効化および無効化するために使用します。 詳細は、 p. 358 固有表現 を参照してください。
- **言語処理:** 文を構築する（抽出パターンおよび強制定義）そして選択した言語の略語を使用する特別な方法を宣言するために使用します。 詳細は、 p. 364 言語処理 を参照してください。
- **言語の識別子:** 言語が［すべて］に設定された場合に呼び出される自動的言語識別子を設定する場合に使用します。詳細は、 p. 366 言語の識別子 を参照してください。

図 18-1
テキスト マイニング テンプレート エディタ - [高度なリソース] タブ



注:検索/置換ツールバーを使用して、情報を迅速に検索したり、一様の変更をセクションに行ったりすることができます。 詳細は、 p. 355 置換 を参照してください。

アドバンスリソースを編集するには

- 編集するリソース セクションを検索および選択します。内容が右側のパネルに表示されます。
- メニューまたはツールバーのボタンを使用して、必要に応じて内容を切り取り、コピー、または貼り付けることができます。
- このセクションで、書式設定規則を使用して、変更したいファイルを編集します。編集を行うとすぐに、変更が保存されます。ツールバーの取り消しまたはやり直しの矢印を使用して、以前の変更に戻ります。

検索

特定のセクションで、情報の迅速な検索が必要な場合があります。たとえば、テキスト リンク分析を実行する場合、多くのマクロおよびパターン定義がある場合があります。検索機能を使用して、特定の規則をすぐ

に見つけることができます。セクション内の情報を検索するために、検索ツールバーを使用できます。

図 18-2
検索ツールバー

検索機能を使用するには

- 検索するリソース セクションを検索および選択します。内容がエディタの右側のパネルに表示されます。
- メニューから [**編集**] → [**検索**] を選択します。[アドバンス リソースを編集] ダイアログ ボックスの右上に、検索ツールバーが表示されます。
- テキスト ボックスに検索したい文字列を入力します。ツールバー ボタンを使用して、大文字と小文字の区別、部分一致、検索の方向を制御します。
- [**検索**] を選択して、検索を開始します。一致が見つかった場合は、テキストがウィンドウで強調表示されます。
- [**検索**] もう一度クリックすると、次の一致を検索します。

注:[テキスト リンク規則]タブ使用時は、[検索]オプションは、ソースコードの表示中のみ使用できます。

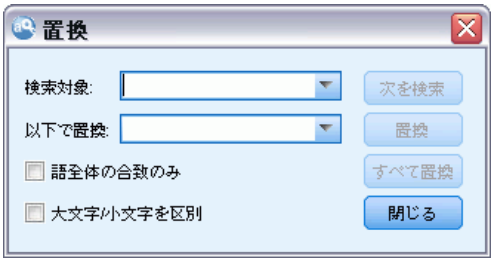
置換

アドバンス リソースへより広い更新が必要な場合があります。置換機能を使用すると、内容に一様の更新を行うことができます。

置換機能を使用するには

- 検索および置換するリソース セクションを検索および選択します。内容がエディタの右側のパネルに表示されます。
- メニューから [**編集**] → [**置換**] を選択します。[置換] ダイアログ ボックスが開きます。

図 18-3
[置換] ダイアログ ボックス



- [**検索対象**] テキスト ボックスで、検索する文字列を入力します。
- [**以下で置換**] テキスト ボックスで、見つかったテキストの代わりに使用する文字列を入力します。
- 完全な語のみを検索または置換する場合、[**語全体の合致のみ**] を選択します。
- 大文字と小文字が完全に一致する語のみを検索または置換する場合、[**大文字/小文字を区別**] を選択します。
- [**次を検索**] を選択して、一致を検索します。一致が見つかった場合は、テキストがウィンドウで強調表示されます。この一致を置換したくない場合は、置換したい一致が見つかるまで [**次を検索**] をクリックします。
- [**置換**] を選択して、選択した一致を置換します。
- [**置換**] を選択して、セクション内のすべての一致を置換します。置換が行われた数を示したメッセージが表示されます。
- 置換が終了したら、[**閉じる**] をクリックします。ダイアログ ボックスが閉じます。

  注:誤って置換を行った場合、ダイアログ ボックスを閉じ、メニューの [**編集**] → [**取り消し**] を選択して、置換を取り消すことができます。取り消したい変更ごとに 1 回ずつ実行する必要があります。

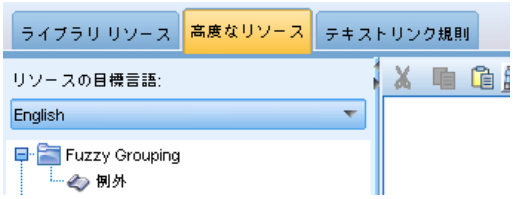
リソースの目標言語

リソースは、特定のテキスト言語で作成されます。これらのリソースが調整される言語は、[アドバンス リソース] タブで定義されます。必要に応じて、[**リソースの目標言語**] コンボボックスでその言語を選択し、別の言語に切り替えることができます。また、ここに示された言語は、これらのリソースで作成するテキスト分析パッケージの言語として表示されます。

重要! リソースの言語を変更する必要は、めったにありません。言語を変更すると、リソースが抽出言語と合致しない場合に問題が発生する場合があります。また、あまり採用されませんが、複数の言語のテキストが期待され

るため、抽出時に**すべて**の言語オプションを使用する場合に、言語を変更する場合があります。言語を変更することにより、たとえば、関心のある二次言語に関して、抽出パターン、省略形、強制定義などの言語処理リソースにアクセスすることができます。ただし、リソースに対して行った変更を公開または保存したり、その他の抽出を実行する前に、その言語を、抽出対象として関心のある一次言語に設定し戻すように留意してください。

図 18-4
目標言語の設定



Fuzzy Grouping

テキスト マイニング ノードおよび抽出設定で、**[語幹文字数が次の最小値以上のときにスペルを調整する]** を選択している場合、Fuzzy Grouping アルゴリズムが有効化されます。

Fuzzy Grouping を使用すると、抽出語から最初の母音を除くすべての母音および二重および三重の子音を一時的に取り除いて比較し、残りが同じがどうかを確認して、一般的にスペルミスのある語やスペルの近い語をグループ化することができます。抽出プロセス時、Fuzzy Grouping 機能は、抽出キーワードに適用され、結果を比較して、一致があるかどうかを判断します。一致があった場合、元のキーワードは最終的な抽出リストにいっしょにグループ化されます。それらのキーワードは、データ内で最も頻繁に出現するキーワードに基づいてグループ化されます。

注:比較されている 2 つのキーワードが <Unknown> タイプを除く異なるタイプに割り当てられている場合、このペアに Fuzzy Grouping 手法は適用されません。つまり、この手法を適用するには、キーワードが同じタイプまたは <Unknown> タイプに割り当てられている必要があります。

この機能を有効にして、類似したスペルの 2 つのキーワードが不正にグループ化されていることがわかった場合、それらを Fuzzy Grouping から除外する必要があります。不正に一致したペアを［アドバンス リソース］タブの［例外］セクションに入力することによって、Fuzzy Grouping から除外することができます。 詳細は、 p. 353 アドバンス リソースについて を参照してください。

次の例は、Fuzzy Grouping がどのように実行されるかについて説明しています。Fuzzy Grouping が有効な場合、これらの語が同じであると示され、以下のように一致します。

```
color   -> colr                  mountain -> montn
colour  -> colr                  montana  -> montn

modeling  -> modlng              furniture -> furntr
modelling -> modlng              furnature -> furntr
```

前者の例では、mountain および montana をグループ化から除外することが考えられます。次のように［例外］セクションに入力します。

```
mountain      montana
```

重要! Fuzzy Grouping の例外では、特定の類義語規則が適用されているため、2 つの単語がグループ化される場合があります。この場合、感嘆符のワイルドカード（!） を使用して類義語を入力し、単語が出力で類義語とならないようにする必要があります。 詳細は、 17 章 p. 346 類義語の定義 を参照してください。

Fuzzy Grouping の例外の書式規則

- 1 行につき 1 つの例外ペアのみを定義する。
- 単語または複合語を使用する。
- 語には小文字の実を使用する。大文字は無視されます。
- ペア内の各語を区切るには TAB 文字を使用する。

固有表現

特定の種類のデータを扱っている場合、日付、社会保障番号、またがその他の固有表現を抽出したいと考える場合があります。これらのエンティティは、エンティティを有効化または無効化できる構成ファイルで明示的に宣言されています。詳細は、 p. 362 構成 を参照してください。 抽出エンジンから出力を最適化するために、非言語処理からの入力を正規化し、事前に定義された形式に従って同様のエンティティをグループ化します。詳細は、 p. 362 正規化 を参照してください。

注:抽出設定で、固有表現の抽出を有効にしたり無効にしたりできます。

使用できる固有表現

次の表の固有表現を抽出できます。カッコ内はタイプ名です。

住所（<Address>）	組織名（<Organization>）
アミノ酸（<Aminoacid>）	割合（<Percent>）
通貨（<Currency>）	商品名（<Product>）
日付（<Date>）	プロテイン（<Gene>）
遅延（<Delay>）	電話番号（<PhoneNumber>）

数値（<Digit>）	時間（<Time>）
電子メール アドレス（<email>）	米国社会保障番号（<SocialSecurityNumber>）
HTTP/URL アドレス（<url>）	計量（<Weights-Measures>）
IP アドレス（<IP>）	

処理するテキストの削除

固有表現の抽出を行う前に、入力テキストが削除されます。この手順で、次の一時的な変更が行われ、固有表現が以下のようなものとして特定および抽出されます。

- 複数のスペースの行列は、1 つのスペースに置き換えられます。
- 表形式はスペースに置き換えられます。
- 単一の行末文字または配列文字はスペースの置き換えられ、複数の行末文字の配列はパラグラフの最後としてマークされます。行末は、復帰改行（CR）および改行（LF）またはそれらの両方で示されます。
- HTML タグおよび XML タグは一時的に除外および無視されます。

正規表現の定義

固有表現抽出時、正規表現の特定に使用される正規表現定義の編集またはそれへの追加が必要な場合があります。これは［アドバンス リソース］タブの［**正規表現の定義**］セクションで行われます。 詳細は、 p. 353 アドバンス リソースについて を参照してください。

ファイルはいくつかのセクションに分割されます。最初のセクションは [macros] です。このセクションのほか、固有表現ごとにセクションが存在する場合があります。このファイルにセクションを追加できます。各セクション内で、規則には番号が付けられています（regexp1、regexp2、など）。これらの規則は 1 ～ n の順番に番号を付けなければなりません。番号が欠けていると、このファイルの処理が一度に中断します。

エンティティが言語に依存する場合があります。構成ファイルの言語パラメータの値が 0 以外の場合、エンティティは言語に依存すると見なされます。詳細は、 p. 362 構成 を参照してください。 エンティティが言語に依存する場合、[english/PhoneNumber] のようにセクション名の前に言語を示す必要があります。PhoneNumber エンティティの言語の値に 2 が指定されている場合、このセクションには英語の電話番号にのみ適用される規則が含まれます。

重要! エディタでこのファイルや他のファイルに変更を行い、抽出エンジンが必要に応じて機能しない場合、ツールバーの［**元のものにリセット**］オプションを使用して、ファイルを付属の下の内容に戻します。このファイルは、正規表現に対する特定のレベルの親密度が必要です。この領域においてさらに支援が必要な場合、IBM Corp. にご連絡ください。

特殊文字: [] {} () ¥ * + ? | ^ $

以下の特殊文字を除くすべての文字はそれ自身に一致します。これらの特殊文字は、表現内の特別な目的に使用されます。.[{()*+?|^$これらの文字をその文字として使用するには、定義で文字の前にバックスラッシュ (¥) を追加する必要があります。

たとえば、Web アドレスを抽出しようとしている場合、終止符はエンティティにとって非常に重要な文字であるため、次のようにバックスラッシュを追加する必要があります。

```
www\.[a-z]+\.[a-z]+
```

繰り返し演算子および識別子: ? + * {}

定義をより柔軟性のあるものにするために、正規表現に標準的なワイルドカードをいくつか使用できます。使用できるのは * ? +

- アスタリスク * は、0 以上の先行文字列があることを示します。たとえば、ab*c は、「ac」、「abc」、「abbbc」などに一致します。
- プラス記号 + は、1 つ以上の先行文字列があることを示します。たとえば、ab+c は、「abc」、「abbc」、「abbbc」などに一致しますが、「ac」には一致しません。
- 疑問符 + は、0 または 1 つの先行文字列があることを示します。たとえば、modell?ing は、「modeling」、および「modeling」のどちらにも一致します。
- 繰り返し制限を示す中カッコ {} は、繰り返しの境界を示します。たとえば、
  - [0-9]{n} は、ちょうど n 回繰り返された数値に一致します。たとえば、[0-9]{4} は “1998” に一致し、“33” および “19983” には一致しません。
  - [0-9]{n} は、 n 回以上繰り返された数値に一致します。たとえば、[0-9]{3,} は “199” または “1998” に一致しますが “19” には一致しません。
  - [0-9]{n,m} は、n ～ m 回繰り返された数値に一致します。たとえば、[0-9]{3,5} は “199”、“1998” または “19983” に一致しますが、“19” および “199835” には一致しません。

任意のスペースおよびハイフン

定義内に任意のスペースを追加したい場合があります。たとえば、「uruguayan pesos」、「uruguayan peso」、「uruguay pesos」、「uruguay peso」、「pesos」または「peso」などの通貨を抽出したい場合、スペースで区切られた 2 つの単語があるという事実に対処する必要があります。この場合、この定義は (uruguayan |uruguay)?pesos? として記述されます。pesos/peso と共に使用する場合、uruguayan または uruguay の後

にはスペースが続くため、任意のスペースは、任意の行列 (uruguayan |uruguay) 内で定義される必要があります。(uruguayan|uruguay)? pesos? のように任意の行列内にスペースがない場合、スペースが必要であるため「pesos」または「peso」とは一致しません。

リスト内にハイフンを含む一連の文字を探している場合、ハイフンを最後に定義する必要があります。たとえば、カンマ (,) またはハイフン (-) を検索する場合、[,-] を使用し、[-,] は決して使用しません。

リストおよびマクロの文字列の順序

短い行列の前に最も長い行列を定義する必要があります。短い行列に一致があるため、最も長い行列が読み取られなくなるためです。たとえば、「billion」または「bill」という文字列を検索している場合、「billion」を「bill」の前に定義する必要があります。つまり、(billion|bill) ではなく、(bill|billion) となります。マクロは一連の文字列であるため、これはマクロにも適用されます。

定義セクションの規則の順序

1 行ごとに 1 つの規則を定義します。各セクション内で、規則には番号が付けられています (regexp1、regexp2、など)。これらの規則は 1 ～ n の順番に番号を付けなければなりません。番号が欠けていると、このファイルの処理が一度に中断します。エントリを無効にするには、正規表現の定義に使用する各行の初めに # 記号を追加します。エントリを有効にするには、行の前の # 文字を削除します。

各セクションで、確実に処理するために、最も具体的な規則を最も一般的な規則の前に定義する必要があります。たとえば、「month year」の形式および「month」の形式で日付を検索したい場合、「month year」の規則を「month」の規則の前に定義する必要があります。次に、その定義を示します。

```
#@# January 1932
regexp1=$(MONTH),?[0-9]{4}

#@# January
regexp2=$(MONTH)
```

そして、次のようには定義してはいけません。

```
#@# January
regexp1=$(MONTH)

#@# January 1932
regexp2=$(MONTH),? [0-9]{4}
```

規則でのマクロの使用

いくつかの規則で特定の行列を使用している場合、マクロを使用できます。この行列の定義を変更する必要がある場合、該当する箇所を一度だけ変更する必要があり、それを参照するすべての規則を変更する必要はありません。たとえば、次のようなマクロがあるとします。

```
MONTH=((january|february|march|april|june|july|august|september|october|
november|december)|(jan|feb|mar|apr|may|jun|jul|aug|sep|oct|nov|dec)(\.)?)
```

マクロの名前を参照する場合、必ず `$()` で囲みます。例:`regexp1=$(MONTH)`

すべてのマクロは `[macros]` で定義する必要があります。

正規化

固有表現を抽出する場合、出現したエンティティは正規化され、事前定義された形式に従って同様のエンティティとグループ化されます。たとえば、語内の通貨記号および同格の文字は同じように扱われます。正規化エントリは、[アドバンス リソース] タブの **[正規化]** セクションで行われます。詳細は、 p. 353 アドバンス リソースについて を参照してください。ファイルはいくつかのセクションに分割されます。

重要! このファイルはアドバンス ユーザーのみが使用できます。このファイルの変更が必要な可能性はほとんどありません。この領域においてさらに支援が必要な場合、IBM Corp. にご連絡ください。

正規化の書式規則

- 1 行につき正規化エントリを 1 つだけ追加する。
- このファイルのセクションを厳密に重視する。新しいセクションを追加することはできません。
- エントリを無効にするには、その行の初めに # 記号を追加する。エントリを有効にするには、行の前の # 文字を削除します。

構成

固有表現構成ファイルに抽出したい固有表現タイプを有効化および無効化することができます。必要のないエンティティを無効にすることによって、必要な処理時間を短縮することができます。これは [アドバンス リソース] タブの **[構成]** セクションで行われます。詳細は、 p. 353 アドバンス リソースについて を参照してください。 非言語学的抽出が有効になると、抽出エンジン抽出時にこの構成ファイルを読み取り、固有表現タイプを抽出するかどうかを判断します。

このファイルのシンタックスは次のようになります。

#name<TAB>言語<TAB>Code

テーブル 18-1
構成ファイルのシンタックス

列ラベル	説明
#name	固有表現の語は、固有表現抽出の他の 2 つの必須ファイルで参照されます。ここで使用される名前は大文字と小文字が区別されます。
言語	ドキュメントの言語。特定の言語を選択することが最適ですが、[任意] オプションもあります。指定できるオプションは次のとおりです。0 = IP/URL/電子メール アドレスなど、正規表現が言語特有でなく、さまざまな言語のいくつかのテンプレートで使用できる任意のオプション、1 = フランス語、2 = 英語、4 = ドイツ語、5 = スペイン語、6 = オランダ語、8 = ポルトガル語、10 = イタリア語です。
Code	品詞コード。多くのエンティティは、一部の場合を除いて “s” の値をとります。指定できる値は次のとおりです。s = ストップワード、a = 形容詞、n = 名詞です。有効にすると、固有表現が最初に抽出され、抽出パターンを適用し、より大きいコンテキストでその役割を特定します・たとえば、割合は “a” の値が与えられます。固有表現として 30% が抽出されるとします。それは形容詞として特定されます。テキストに「30% salary increase」という文字列が含まれていた場合、“30%” の固有表現は “ann”（形容詞、名詞、名詞）の品詞パターンに適合します。

エンティティ定義の順序

このファイルでエンティティが宣言される順序は重要であり、それらがどのように抽出されるかに影響を与えます。表示された順序に適用されます。順序を変更すると、結果も変わります。最も具体的な固有表現は、最も一般的な固有表現の前に定義する必要があります。

たとえば、固有表現 “Aminoacid” は次のように定義されます。

```
regexp1=($(AA)-?$(NUM))
```

$(AA) は、特定のアミノ酸に対応する固有の 3 文字の行列である、“(ala|arg|asn|asp|cys|gln|glu|gly|his|ile|leu|lys|met|phe|pro|ser)” に対応します。

一方、固有表現 Gene はより一般的であり、次のように定義されます。

```
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?
```

[構成] セクションで「Gene」を「Aminoacid」の前に定義すると、「Gene」の「regexp3」が最初に一致するため、「Aminoacid」は一致しなくなります。

こうせいの書式規則

- 列内の各語を区切るには TAB 文字を使用する。
- 行を削除しない。
- 前述の表示示されたシンタックスを重視する。
- エントリを無効にするには、その行の初めに # 記号を追加する。エンティティを有効にするには、行の前の # 文字を削除します。

言語処理

現在使用される各言語には、キーワードを表現、文を構築、省略形を使用する特別な方法があります。[言語処理] セクションでは、抽出パターンを編集、これらのパターンの定義を強制、[言語] ドロップダウン リストで選択した言語の省略形を宣言することができます。

- 抽出パターン
- 強制定義
- 省略形

抽出パターン

ドキュメントから情報を抽出する場合、抽出エンジンは、品詞抽出パターンのセットをテキストの単語の「積み重ね」に適用して、抽出の候補のキーワード（単語および句）を特定します。抽出パターンを追加または変更できます。

品詞には、名詞、形容詞、過去分詞、決定詞、前置詞、人の名、イニシャル、助詞など、文法的な要素が含まれます。これら一連の要素が、品詞の抽出パターンを構成しています。IBM Corp. のテキスト マイニング製品では、各品詞が 1 つの文字で表され、パターンが定義しますくなります。たとえば、形容詞は小文字の「a」で表されます。デフォルトでは、サポートされているコードのセットがデフォルトの抽出パターンの上位に、パターンのセットと、各パターンの例と共に表示され、使用される各コードについて理解しやすくなります。

抽出パターンの書式設定規則

- 1 行ごとに 1 つのパターン。
- 行頭に # を使用してパターンを無効化する。

単語の指定された順序は抽出エンジンによって一度だけ読み込まれ、エンジンが合致を検出した最初の抽出パターンに割り当てるため、抽出パターンの表示順が非常に重要になります。

強制定義

ドキュメントから情報を抽出するとき、抽出エンジンはテキストをスキャンし、出現するすべての単語の品詞を特定します。状況によっては、単語がいくつかの異なる役割に適合する場合があります。単語が特定の品詞の役割を取得するよう強制する場合、または処理からその単語を完全に除外する場合、[アドバンス リソース] タブの [強制定義] セクションで指定できます。 詳細は、 p. 353 アドバンス リソースについて を参照してください。

指定した単語の品詞を強制するには、次のシンタックスを使用して、このセクションに 1 行追加します。

```
term:code
```

テーブル 18-2
シンタックスの説明

エントリ	説明
term	キーワード名。
code	品詞の役割を示す 1 文字のコード。ユニタームあたり最大 6 つの品詞コードを表示できます。また、additional:s のように小文字のコード s を使用して、単語が複合語/句に抽出されないようにできます。

強制定義の書式設定規則

- 単語ごとに 1 行。
- キーワードにコロンの使用不可。
- 単語を抽出しないようにするには小文字の s を品詞コードに使用する。
- 1 行あたり最大 6 つの品詞コードを使用する。サポートされている品詞コードは、[抽出パターン] セクションに表示されます。 詳細は、 p. 364 抽出パターン を参照してください。
- 部分一致の場合、文字列の最後にアスタリスク (*) をワイルドカードとして使用する。たとえば、add*:s と入力すると、add、additional、additionally、addendum、および additive のような単語はキーワードとして、あるいは複合語キーワードの一部として抽出されなくなります。ただし、単語の合致がコンパイル済み辞書または強制定義でキーワードとして明示的に宣言されている場合は、抽出されます。たとえば、add*:s および addendum:n を入力すると、addendum がテキスト内で見つかった場合に抽出されます。

省略形

抽出エンジンがテキストを処理しているとき、検出されたピリオドは、文の終了を示す指標として認識されます。これは通常、正常な処理ですが、省略形がテキスト内にある場合、ピリオド文字のこうした処理は適用されません。
テキストからキーワードを抽出し、特定の省略形の処理が不適切であったことがわかった場合、このセクションで省略形を明示的に宣言する必要があります。

注:省略形が類義語定義に出現、またはキーワード辞書でキーワードとして定義されている場合、ここで省略形の投入を追加する必要はありません。

省略形の書式設定規則

- 1 行ごとに 1 つの省略形を定義する。

言語の識別子

分析しているテキスト データに特定の言語を選択することが最適ですが、テキストがさまざまな言語または不明の言語である場合、[すべて] オプションを指定することもできます。[すべて] 言語オプションでは、言語の識別子と呼ばれる言語の自動認識エンジンを使用します。言語の識別子はドキュメントをスキャンし、提案された言語のテキストを識別し、抽出時に自動的に各ファイルに最適な内部辞書を適用します。[すべて] オプションは［プロパティ］オプションのパラメータによって支配されます。

プロパティ

言語の識別子は、このセクションのパラメータを使用して構成されます。次の表では、［アドバンス リソース］タブの [言語の識別子 - プロパティ] セクションで設定できるパラメータについて示しています。 詳細は、 p. 353 アドバンス リソースについて を参照してください。

テーブル 18-3
パラメータの説明

パラメータ	説明
NUM_CHARS	テキストの言語を判断するために、抽出エンジンが読み取る文字数を指定します。数が小さいほど、言語が速く識別されます。数が大きいほど、言語が正確に識別されます。値を 0 に設定すると、ドキュメントのテキスト全体が読み取られます。
USE_FIRST_SUP-PORTED _LANGUAGE	抽出エンジンが言語の識別子で検出された最初の提案言語を使用するかどうかを指定します。値を 1 に設定すると、最初に提案された言語を使用します。値を 0 に設定すると、代替言語を使用します。
FALLBACK_LANGUAGE	識別子が返した言語がサポートされていない場合に使用する言語を指定します。指定できる値は、english、french、german、spanish、dutch、italian、および ignore です。値を ignore に設定すると、言語がサポートされていないドキュメントは無視されます。

言語

言語の識別子は、さまざまな言語をサポートしています。［アドバンス リソース］タブの［**言語の識別子 - 言語**］セクションで、言語のリストを編集できます。

多くの言語があるほど、正でないものが誤って正と認識されるケースが多くなり、パフォーマンスが低下するため、使用の可能性が低い言語をこのリストから削除することを検討することができます。ただし、このファイルに新しい言語を追加できません。使用頻度が高い言語をリストの上位に移動して、言語の識別子がドキュメントの一致をより速く見つけられるようにしてください。

章
19
テキスト リンク規則について

テキスト リンク分析（TLA）はパターンマッチ手法で、一連の条件規則を使用して、テキスト内の関係性を抽出するのに使用されます。テキスト リンク分析が抽出に対して有効である場合、テキスト データがこれらの規則に対して比較されます。合致が検出されると、テキスト リンク分析パターンが抽出され、表示されます。これらの条件規則は、[テキスト リンク規則] タブで定義されます。

たとえば、組織に関する単純なアイデアを示すコンセプト抽出してもそれが重要でない場合がありますが、TLA を使用して、さまざまな組織または組織に関連する人々の間のつながりについて学習することができます。TLA を使用して、指定された製品または経験についてどう思うかなどのトピックについての意見を抽出することもできます。

TLA を利用するには、テキスト リンク分析（TLA）規則を含むリソースが必要です。テンプレートを選択すると TLA 列にアイコンがあるかどうかによって、どのテンプレートに TLA 規則があるかを確認できます。

図 19-1
TLA 列が表示されているテンプレートのダイアログ

テンプレート	所有者	バージ...	日付	注釈	TLA	言語
Bank CRM (English)	claired	1	9-30-20...			English
Insurance CRM (English)	claired	1	9-30-20...			English
Ads Opinions (English)	claired	1	9-17-20...			English
Bank Satisfaction Opinio...	claired	1	9-17-20...			English
Security Intelligence (En...	claired	1	10-05-2...			English

テキスト データのテキスト リンク分析パターンは、抽出プロセスのパターン マッチの段階で検出されます。この段階で、条件規則がテキスト データと比較され、合致が検出されると情報がパターンとして抽出されます。テキスト リンク分析からより多くの情報を抽出したり、どのように合致するかを変更したりすることが必要な場合があります。こうした場合、条件規則が特定のニーズに適応するよう、規則を処理することができます。これは［テキスト リンク規則］タブで実行します。

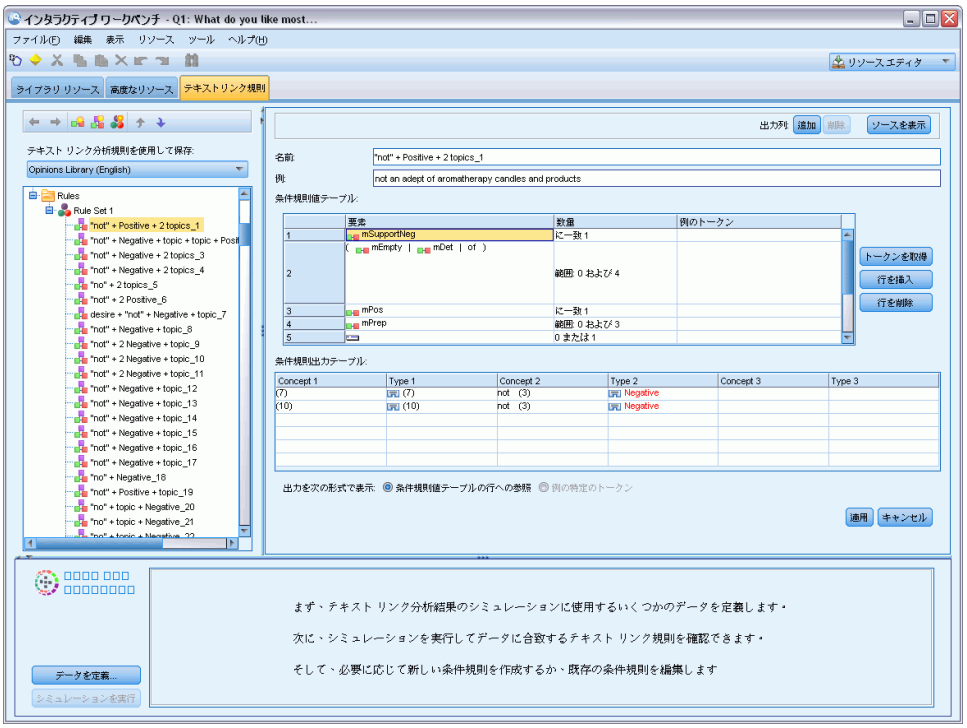
注:変数のサポートは、バージョン 13 では継続されていません。代わりにマクロを使用してください。 詳細は、 p. 378 マクロの作業 を参照してください。

テキスト リンク規則を処理するには

テンプレート エディタ または リソース エディタ ビューの［テキスト リンク規則］タブで直接、条件規則を編集および作成できます。条件規則がテキストとどのように合致するかを確認するために、このタブでシミュレーションを実行できます。シミュレーション時、抽出はサンプルのシミュレーション データでのみ実行され、テキスト リンク規則を適用してパターン マッチがあるかどうかを確認します。テキストに合致する規則は、シミュレーション パネルに表示されます。合致に基づいて、条件規則およびマクロを編集し、テキストがどのように合致するかを変更することができます。

他の高度なリソースとは異なり、TLA 規則はライブラリ固有の定義です。そのため、一度に 1 つのライブラリの TLA 規則を使用できます。テンプレート エディタ または リソース エディタ で、**［テキスト リンク規則］**タブに移動します。使用したい、または編集したい TLA 規則を含むテンプレートのライブラリを選択します。このため、特別な理由がない限り、すべての規則を 1 つのライブラリに保存することを強くお勧めします。

図 19-2
［テキスト リンク規則］タブ



重要! このタブは、日本語リソースには使用できません。

作業の開始

[テキスト リンク規則タブ] エディタで作業を開始するにはさまざまな方法があります。

- いくつかのサンプル テキストで結果をシミュレーションし、現在のセットの規則がシミュレーションからパターンをどのように抽出するかに基づいて合致規則を編集または作成する。
- スクラッチから新しい規則を作成するか、既存の規則を編集する。
- ソース ビューで直接作業する。

規則の編集または作成が必要な場合

各テンプレートに付属するテキスト リンク分析規則は、多くの単純または複雑な関係性をテキストから抽出することに適している場合が多いですが、これらの規則に変更を加えるか、独自の規則をいくつか作成することが必要な場合があります。例:

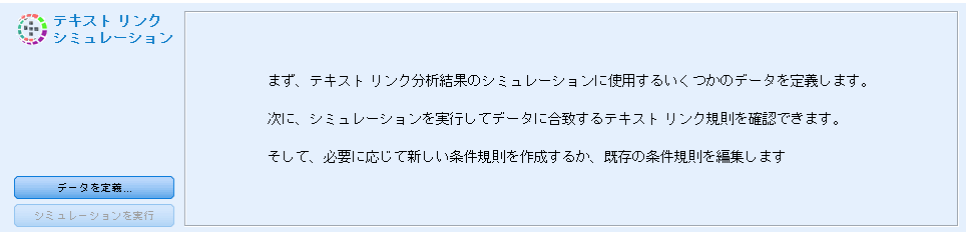
- 新しい規則またはマクロを作成して既存の規則で抽出されていないアイデアまたは関係性をキャプチャする。
- リソースに追加したタイプのデフォルトの動作を変更する。通常、mTopic または mNonLingEntities などのマクロを編集することが必要です。 詳細は、 p.382 特殊マクロ：mTopic、mNonLingEntities、SEP を参照してください。
- 既存のテキスト リンク分析規則およびマクロに新しいタイプを追加する。たとえば、タイプ<**組織名**>が広すぎる場合、<**医薬品**>、<**自動車メーカー**>、<**金融**>など、いくつかの企業部門に分かれた組織の新しいタイプを作成することができます。この場合、テキスト リンク分析規則を編集またはマクロを作成して、これらの新しいタイプを考慮に入れ、適宜処理する必要があります。
- 既存のテキスト リンク分析規則およびタイプを追加する。たとえば、「john doe called jane doe」というテキストをキャプチャする規則があり、通話をキャプチャするこの規則を使用して、電子メールのやり取りもキャプチャします。電子メールの固有表現タイプを規則に追加し、次のようなテキストもキャプチャするようにできます：johndoe@ibm.com emailed janedoe@ibm.com.
- 規則を新規作成するのではなく、既存の規則を若干変更する。たとえば、「xyz is very good」というテキストに合致する規則があり、この規則を使用して、「xyz is very, very good」もキャプチャします。

テキスト リンク分析結果のシミュレーション

新しいテキスト リンク規則を定義できるようにするために、またはテキスト リンク分析時に特定の文がどのように合致するかを理解できるようにするために、テキストのサンプル部分を抽出してシミュレーションを実行すると役立ちます。シミュレーション時、抽出はサンプルのシミュレーション データでのみ実行され、テキスト リンク規則を適用してパターン マッチがあるかどうかを確認します。目標は、シミュレーションの結果を取得し、これらの結果を使用して規則を改善、新しい規則を作成するか、どのように合致が出現するかをより良く理解することです。テキストの各部分（状況によって文、語、句）に対し、シミュレーションの出力には、トークンの集合およびそのテキストのパターンを明らかにしなかった TLA 規則を示します。**トークン**は、抽出プロセス時に特定される単語または語句として定義されます。

他の高度なリソースとは異なり、TLA 規則はライブラリ固有の定義です。そのため、一度に 1 つのライブラリの TLA 規則を使用できます。テンプレート エディタ または リソース エディタ で、[**テキスト リンク規則**] タブに移動します。使用したい、または編集したい TLA 規則を含むテンプレートのライブラリを選択します。このため、特別な理由がない限り、すべての規則を 1 つのライブラリに保存することを強くお勧めします。

図 19-3
データ定義前の [テキスト リンク シミュレーション] パネル



重要! データ ファイルを使用する場合、含まれるテキストを処理時間短縮のために短くなっていることを確認してください。シミュレーションの目的は、テキストの一部がどのように読み取られるかを確認し、このテキストに条件規則がどのように合致するかを理解することです。この情報を使用して、条件規則を作成および編集できます。テキスト リンク分析ノードを使用、または TLA 抽出を有効にしてインタラクティブ セッションのあるストリームを実行し、より完全なデータ セットの結果を取得します。このシミュレーションは、テストおよび条件規則作成のためだけに行われます。

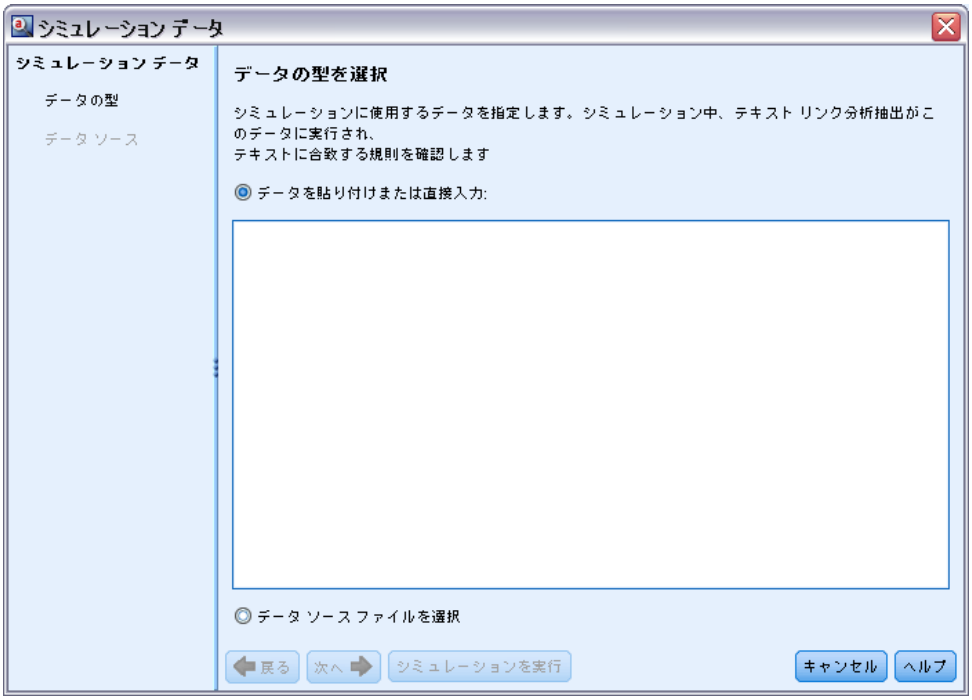
シミュレーションのデータ定義

条件規則がテキストとどのように合致するかを確認するために、サンプルデータを使用してシミュレーションを実行できます。まず、データを定義します。

データの定義

▶ [テキスト リンク規則] タブの一番下にある［シミュレーション］パネルで [データを定義] をクリックします。または、データがまだ定義されていない場合、メニューから [ツール] → [シミュレーションの実行] を選択します。シミュレーション データ ウィザードが開きます。

図 19-4
シミュレーション ウィザード



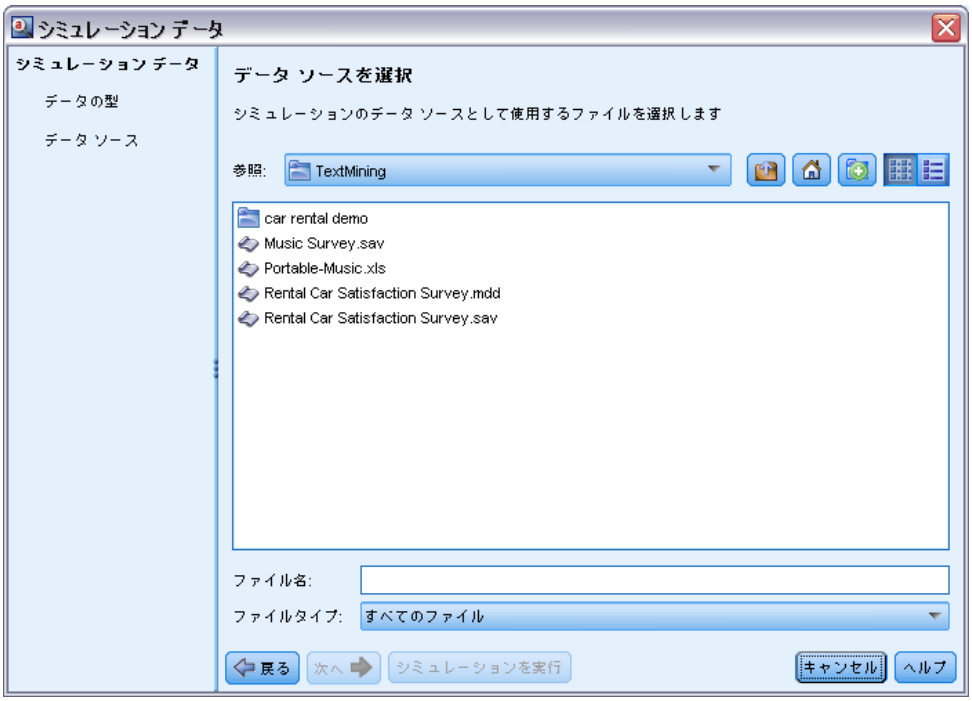
▶ 次のいずれかを選択して、データの種類を指定します。

- **データを貼り付けまたは直接入力**:テキストボックスで、クリップボードからテキストを貼り付けるか、処理するテキストを手動で入力します。1 行ごとに 1 文ずつ入力するか、ピリオドやカンマなど、句読点を使用して文を区切ります。テキストを入力すると、[シミュレーションを実行] をクリックして、シミュレーションを開始できます。
- **ファイル データ ソースを指定**:テキストを含むファイルを処理します。[次へ] をクリックすると、処理するファイルを定義できるウィザードの手順に進みます。ファイルを選択すると、[シ

ミュレーションを実行] をクリックして、シミュレーションを開始できます。以下のファイル タイプがサポートされます：.rtf、.doc、.docx、.docm、.xls、.xlsx、.xlsm、.htm、.html、.txt および拡張子のないファイルです。選択したデータ ファイルは、シミュレーション時にそのまま読み込まれます。たとえば、Microsoft Excel ファイルを選択すると、特定のワークシートまたは列を選択できません。代わりに、ワークブック全体が、IBM® SPSS® Modeler の Microsoft Excel 入力ノードを使用した場合と同じように読み取られます。ファイル全体は、ファイル リスト ノードをテキスト マイニング ノードに接続した場合と同じように扱われます。

重要! データ ファイルを使用する場合、含まれるテキストを処理時間短縮のために短くなっていることを確認してください。シミュレーションの目的は、テキストの一部がどのように読み取られるかを確認し、このテキストに条件規則がどのように合致するかを理解することです。この情報を使用して、条件規則を作成および編集できます。テキスト リンク分析ノードを使用、または TLA 抽出を有効にしてインタラクティブ セッションのあるストリームを実行し、より完全なデータ セットの結果を取得します。このシミュレーションは、テストおよび条件規則作成のためだけに行われます。

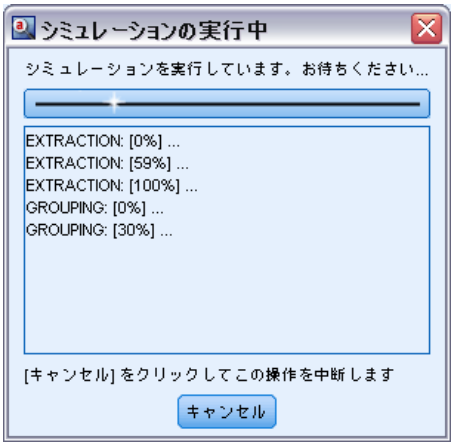
図 19-5
シミュレーション ウィザード - データ ソースを選択



▶ [シミュレーションを実行] をクリックして、シミュレーション プロセスを開始します。プロセス ダイアログが表示されます。インタラクティブ セッショ

ンで作業している場合、シミュレーション時に使用する抽出設定は、インタラクティブ セッションで現在選択している抽出設定となります（コンセプトとカテゴリ ビューの [**ツール**] → [**抽出設定**] を参照）。テンプレート エディタ で作業している場合、シミュレーション時に使用する抽出設定は、デフォルトの抽出設定で、テキスト リンク分析ノードの［エキスパート］タブに表示されている設定と同じです。 詳細は、 p. 374 シミュレーション結果の理解 を参照してください。

図 19-6
[シミュレーション プロセス] ダイアログ



シミュレーション結果の理解

条件規則がテキストとどのように合致するかを確認するために、サンプルデータを使用してシミュレーションを実行し、結果を確認できます。ここで、データにより正確に適用するよう一連の規則を変更できます。抽出プロセスおよびシミュレーション プロセスが完了すると、シミュレーションの結果が表示されます。

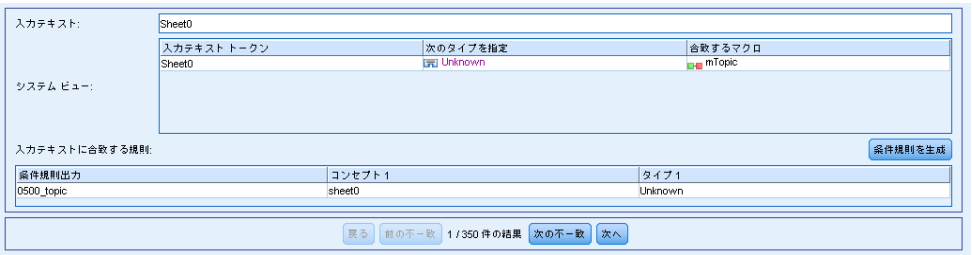
抽出時に特定された各「文」に対し、正確な「文」を含む情報のいくつかの部分が、この入力テキストの文にあるトークンの分析結果、そしてその文のテキストに合致した規則を含むいくつかの情報が表示されます。「**文**」は、抽出機能でテキストを読み取り可能な単位にどのように分割するかによって、語、文、句を意味します。

トークンは、抽出プロセス時に特定される単語または語句として定義されます。たとえば、文「My uncle lives in New York」 の場合、my、uncle、lives、in、および new york というトークンがあります。また、uncle はコンセプトとして抽出され、タイプは <**不明**>、そして new york はコンセプトとして抽出され、タイプは <**地名**> となります。すべてのコンセプトはトークンとなりますが、すべてのトークンがコンセプトとなるわけではありません。トークンは、マクロ、リテラル文字列、そして

単語の空所の場合もあります。タイプを指定された単語または語句のみがコンセプトとなります。

インタラクティブ セッションまたはリソース エディタで作業している場合、コンセプト レベルで対応します。TLA 規則はより細かく、抽出されない、またタイプ指定されない場合であっても、文内の各トークンを規則の定義に使用できます。コンセプトでないトークンを使用できると、テキストの複雑な関係性をキャプチャする場合により柔軟な規則となります。

図 19-7
規則に対する合致を示すサンプル シミュレーションの結果



シミュレーション データに複数の文がある場合、[**次へ**] および [**戻る**] をクリックして、結果を前後に移動することができます。

選択したライブラリ（このタブのツリー上のライブラリ名を参照）の TLA 規則に文が合致しない場合、結果は不一致と見なされ、ボタン [**次の不一致**] および [**前の不一致**] が有効になり、合致を検出した規則のないテキストがあることを示し、これらのインスタンスにすばやく移動できます。

新しい規則を作成、規則を編集、またはリソースや抽出設定を変更したと、シミュレーションの再実行が必要な場合があります。シミュレーションを再実行するには、[シミュレーション] パネルで [**シミュレーションを実行**] をクリックします。同じ入力データが再度使用されます。

次のフィールドおよびテーブルがシミュレーション結果に表示されます。

入力テキスト: ウィザードで定義した、シミュレーション データからの抽出プロセスで特定される実際の「文」。「文」は、抽出機能でテキストを読み取り可能な単位にどのように分割するかによって、語、文、句を意味します。

システム ビュー: 抽出プロセスが特定したトークンの集合。

- **入力テキスト トークン**: 入力テキストで検出される各トークン。トークンについては、このトピックの前の項で定義されています。
- **次のタイプを指定**: トークンがコンセプトとして特定およびタイプ指定されている場合、(<**不明**>、<**人名**>、<**地名**> など) 関連するタイプ名がこの列に表示されます。
- **合致するマクロ**: トークンが既存のマクロに合致する場合、関連するマクロ名がこの列に表示されます。

入力テキストに合致する規則: このテーブルには、入力テキストに対して合致した TLA 規則が表示されます。合致する各規則に対して、[**条件規則出力**] 列に条件規則の名前と、その規則に関連する出力値（コンセプト + タイプのペア）が表示されます。合致する条件規則の名前をダブルクリックすると、シミュレーション パネル上のエディタ パネルに規則が表示されます。

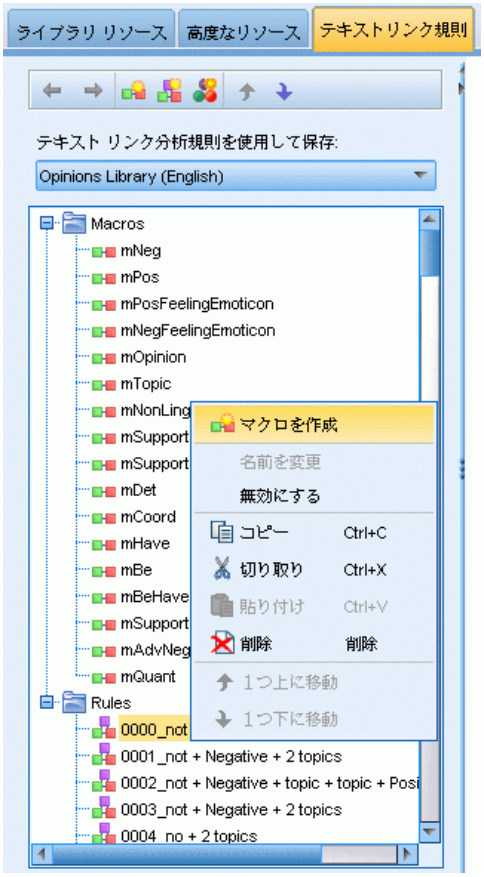
[条件規則を生成] ボタン:シミュレーション パネルでこのボタンをクリックすると、シミュレーション パネルの上の条件規則エディタ パネルに新しい規則が表示されます。入力テキストを例として採用します。同様に、シミュレーション時にタイプ指定またはマクロに合致したトークンは、[**条件規則値テーブル**] の［要素］列に自動的に挿入されます。トークンがタイプ指定され、かつマクロに合致した場合、マクロ値は条件規則を単純化するために条件規則内で使用されるマクロ値となります。たとえば、Basic English リソースを使用している場合、「I like pizza」という文は、シミュレーション時に <**不明**> とタイプ指定され、mTopic のマクロに合致します。この場合、生成された規則で mTopic が要素として使用されます。詳細は、 p. 383 テキスト リンク規則の使用 を参照してください。

ツリー内の規則およびマクロのナビゲート

抽出時にテキスト リンク分析が実行されると、[**テキスト リンク規則**] タブで選択されたライブラリに保存されているテキスト リンク規則が使用されます。

他の高度なリソースとは異なり、TLA 規則はライブラリ固有の定義です。そのため、一度に 1 つのライブラリの TLA 規則を使用できます。テンプレート エディタ または リソース エディタ で、[**テキスト リンク規則**] タブに移動します。使用したい、または編集したい TLA 規則を含むテンプレートのライブラリを選択します。このため、特別な理由がない限り、すべての規則を 1 つのライブラリに保存することを強くお勧めします。

図 19-8
[テキスト リンク規則] タブ:規則およびマクロ ツリー



[テキスト リンク分析規則を使用して保存:] 次のリストで該当するライブラリを選択して、[テキスト リンク規則] タブで処理するライブラリを指定できます：このタブのドロップ ダウン リスト。抽出時にテキスト リンク分析が実行されると、**[テキスト リンク規則]** タブで選択されたライブラリに保存されているテキスト リンク規則が使用されます。そのため、複数のライブラリにテキスト リンク規則（TLA 規則）を定義した場合、TLA 規則がある最初のライブラリのみがテキスト リンク分析に使用されます。このため、特別な理由がない限り、すべての規則を 1 つのライブラリに保存することを強くお勧めします。

ツリーでマクロまたは帰っ足を選択する場合、右側のエディタ パネルに内容が表示されます。ツリー内の項目を右クリックすると、次のようなタスクを示すコンテキスト メニューが表示されます。

- ツリー内に新しいマクロを作成し、右側のエディタで開く。
- ツリー内に新しい規則を作成し、右側のエディタで開く。
- ツリー内に新しいルール セットを作成する。

- 項目を切り取り、コピーおよび貼り付けて、編集を簡単にする。
- マクロ、規則、ルール セットを削除して、リソースから除外する。
- マクロ、規則、ルール セットを無効にして、処理時に無視されるようにする。
- 規則を上下に移動させて処理の順序を変更する。

ツリー上の警告

警告はツリー上で黄色の三角を伴って表示されます。これはどこかに問題があることを知らせるために現れます。マウス ポインターをエラーのあるマクロやルールに当てると、説明がポップアップ表示されます。大抵の場合は、以下のように表示されます：**警告:例がありません。例を入力してください** この場合は例を入力します。

例を忘れた場合、または、例がルールにマッチしない場合には、トークン取得機能を使用することができませんので、ルールに沿った例を1つ入力することをお奨めします。

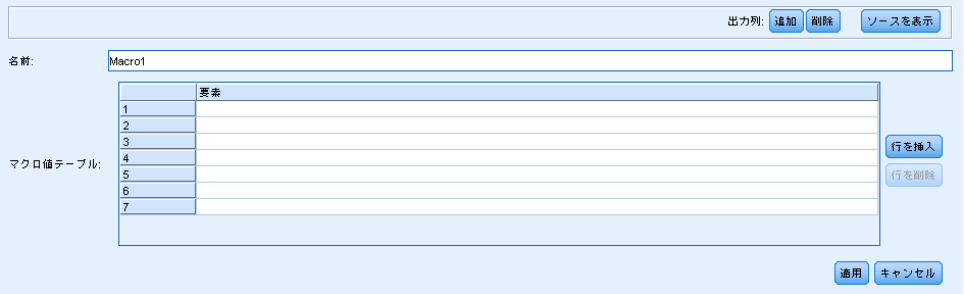
例が黄色でハイライト表示された場合は、TLA エディタに不明なタイプまたはマクロであることを示します。以下のメッセージも同様です：**警告:不明なタイプまたはマクロです**。これは、ソース ビューの $something で定義された項目（たとえば、$myType）が使用するライブラリのレガシー タイプではなく、マクロでもないことを知らせるものです。

シンタックス チェッカーを更新するには、他のルールまたはマクロに切り替える必要があります。コンパイルは何も必要ありません。ですから、たとえばルール A が例がないための警告を表す場合には、例を追加する必要があります。上位または下位のルールをクリックしてから、ルール A に戻り今は適正となったことを確認します。

マクロの作業

マクロは、タイプ、その他のマクロおよびリテラル（単語）文字列を OR 演算子（|）とグループ化することによって、テキスト リンク分析規則の出現を簡略化することができます。マクロを使用すると、複数のテキスト リンク分析規則でマクロを再利用して簡略化するだけでなく、テキスト リンク分析規則全体で更新する必要はなく、1 つのマクロを更新することができます。付属の TLA 規則の多くには、事前定義されたマクロが含まれています。マクロは、[テキスト リンク規則] タブの左端のパネルの上位ツリーに表示されます。

図 19-9
[テキスト リンク規則] タブ:マクロ エディタ



次のフィールドおよびテーブルがシミュレーション結果に表示されます。

名前:このマクロを示す一意の名前。マクロ名の接頭辞に小文字の m を指定して、規則内でマクロであることをすぐに特定できるようにすることをお勧めします。手動で規則内のマクロを参照する場合(インライン編集またはソース ビュー)、抽出プロセスでこの特殊な名前を認識するよう、 $ 文字の接頭辞を使用する必要があります。ただし、マクロ名をドラッグ アンド ドロップまたはコンテキスト メニューを使用して追加する場合、その名前はマクロとして自動的に認識され、$ は追加されません。

マクロ値テーブル:

- このマクロが示すことができるすべての値を示す多くの行。これらの値では大文字と小文字が区別されます。
- 値は、タイプ、リテラル文字列、単語の空所またはマクロの組み合わせがあります。 詳細は、 p. 392 条件規則およびマクロにサポートされている要素 を参照してください。
- マクロの要素に値を入力するには、使用する行をダブルクリックします。タイプの参照、マクロの参照、リテラル文字列、または単語の空所を入力できるテキスト ボックスが表示されます。または、セルを右クリックすると、共通のマクロ、タイプ名、固有表現キーワード名のリストを提供するコンテキスト メニューが表示されます。タイプまたはマクロを参照するには、マクロ名またはタイプ名の先頭に ‘$’ 文字を追加する必要があります(例: マクロ mTopic の場合 $mTopic)。引数を結合する場合、カッコ ( ) を使用して引数およびブール型演算子 OR を示す文字 | をグループ化する必要があります。
- 右側のボタンを使用して、マクロ値テーブルの行を追加または削除できます。
- それぞれの行に要素を入力します。たとえば、am OR was OR is のような 3 つのリテラル文字列のいずれかを示すマクロを作成する場合、ビューの各行にそれぞれのリテラル文字列を入力すると、マクロ テーブルには3 つの行が作成されます。

マクロの作成および編集

新しいマクロを作成したり、既存のマクロを編集できます。マクロ エディタのガイドラインおよび説明に従います。 詳細は、 p. 378 マクロの作業 を参照してください。

マクロの新規作成

- メニューの [**ツール**] → [**新規マクロ**] を選択します。または、ツリー ツールバーの［新規マクロ］アイコンをクリックすると、エディタに新しいマクロが表示されます。
- 一意の名前を入力して、マクロ値の要素を定義します。
- エラーの有無を確認したら、[**適用**] をクリックします。

マクロの編集

- ツリー内のマクロ名をクリックします。右側のエディタ パネルにマクロが表示されます。
- 変更を加えます。
- エラーの有無を確認したら、[**適用**] をクリックします。

マクロの無効化および削除

マクロの無効化

処理中にマクロが無視されるようにするには、マクロを無効にすることができます。マクロを無効にすると、この無効なマクロを参照する規則で警告またはエラーが発生する場合があります。マクロを削除および無効にする場合は注意してください。

- ツリー内のマクロ名をクリックします。右側のエディタ パネルにマクロが表示されます。
- 名前を右クリックします。
- コンテキスト メニューから、[**無効にする**] を選択します。マクロのアイコンがグレーになり、マクロ自体を編集できなくなります。

マクロの削除

マクロを除外する場合、削除することができます。マクロを削除すると、このマクロを参照する規則でエラーが発生する場合があります。マクロを削除および無効にする場合は注意してください。

- ツリー内のマクロ名をクリックします。右側のエディタ パネルにマクロが表示されます。
- 名前を右クリックします。
- コンテキスト メニューから、[削除] を選択します。リストからマクロが消去されます。

エラーのチェック、保存およびキャンセル

マクロの変更の適用

マクロ エディタの外をクリックまたは [適用] をクリックすると、マクロが自動的にスキャンされ、エラーの有無が確認されます。エラーが見つかると、アプリケーションの別の部分に移る前に修正する必要があります。

ただし、あまり深刻でないエラーが検出された場合は、警告のみが表示されます。たとえば、マクロにタイプまたはその他のマクロに対する不完全または参照されない定義が含まれている場合、警告メッセージが表示されます。[適用] をクリックすると、未修正の警告により、左側パネルの規則とマクロ ツリーのマクロ名の左側に警告アイコンが表示されます。

マクロを適用しても、マクロが永続的に保存されるわけではありません。適用すると、検証プロセスで、エラーおよび警告の有無をチェックします。

インタラクティブ ワークベンチ セッション内のリソースの保存

- インタラクティブ ワークベンチ セッションでリソースに行った変更を保存し、次回ストリーム実行時にそれらの変更を取得できるようにするには、次の手順を実行する必要があります。

  モデル作成ノードを更新して、次回ストリーム実行時にこれらの同じリソースを取得できるようにします。詳細は、 8 章 p.148 モデル作成ノードの更新および保存 を参照してください。 その後ストリームを保存します。ストリームを保存するには、モデル作成ノードを更新した後、IBM® SPSS® Modeler のメイン ウィンドウで保存します。

- インタラクティブ ワークベンチ セッションでリソースに行った変更を保存し、他のストリームでそれらを使用できるようにするは、次の手順を実行します。

- 使用していたテンプレートを更新するか、新しいテンプレートを作成します。詳細は、 14 章 p. 292 テンプレートの作成および更新 を参照してください。 現在のノードの変更を保存するわけではありません（前の手順を参照）。
- または、使用していた TAP を更新します。 詳細は、 10 章 p. 250 テキスト分析パッケージの更新 を参照してください。

テンプレート エディタ 内のリソースの保存

▶ まず、ライブラリを公開します。 詳細は、 16 章 p. 328 ライブラリの公開 を参照してください。

▶ そして、メニューで [**ファイル**] → [**リソース テンプレートを保存**] を使用してテンプレートを保存します。

マクロの変更のキャンセル

▶ 変更を破棄する場合は、[**キャンセル**] をクリックします。

特殊マクロ：mTopic、mNonLingEntities、SEP

基本リソース テンプレートおよび意見テンプレートは、mTopic および mNonLingEntities という 2 つの特別なマクロに付属しています。

mTopic

デフォルトでは、mTopic マクロは、何らかの意見に結びつけられる可能性のあるテンプレート付属タイプをすべてグループ化します。このような Core ライブラリ タイプには次のようなものがあります。<Person>、<Organization>、<Location>などです。これらが該当するのは、タイプが意見タイプではない（たとえば、<Negative> または <Positive>）か、または［アドバンス リソース］で固有表現として定義されているタイプです。

意見（または同様の）テンプレートで新しいタイプを作成する場合、は、このタイプが別のマクロまたは［アドバンス リソース］タブの固有表現セクションで指定されていない場合、マクロ mTopic で定義された他のタイプと同じ方法で処理されると想定します。

たとえば、意見テンプレートからリソースに新しいタイプを作成したとします：<**野菜**> と <**果物**>としましょう。変更を行うことなく、新しいタイプは mTopic として扱われ、新しいタイプに関する肯定的、否定的、中立的、文脈上の意見を自動的に明らかにすることができます。抽出時、たとえば文「I enjoy broccoli, but I hate grapefruit」が次の出力パターンを作成するとします。

```
broccoli <Vegetables> + like <Positive>
grapefruit <Fruit> + dislike <Negative>
```

ただし、mTopic の他のタイプと異なる方法でこれらのタイプを処理する場合、タイプ名を mPos のような既存マクロに追加してすべての肯定的意見タイプをグループ化するか、1 つまたは複数の規則で後で参照できる新しいマクロを作成することができます。

重要!<Vegetables> のような新しいタイプを作成した場合、この新しいタイプは mTopic にタイプとして含まれますが、このタイプ名はマクロ定義に明示的に表示されるわけではありません。

mNonLingEntities

同様に、[アドバンス リソース] タブの **[固有表現]** セクションに新しい固有表現を追加すると、それらは特に指定のない限り mNonLingEntities として自動的に処理されます。 詳細は、 18 章 p. 358 固有表現 を参照してください。

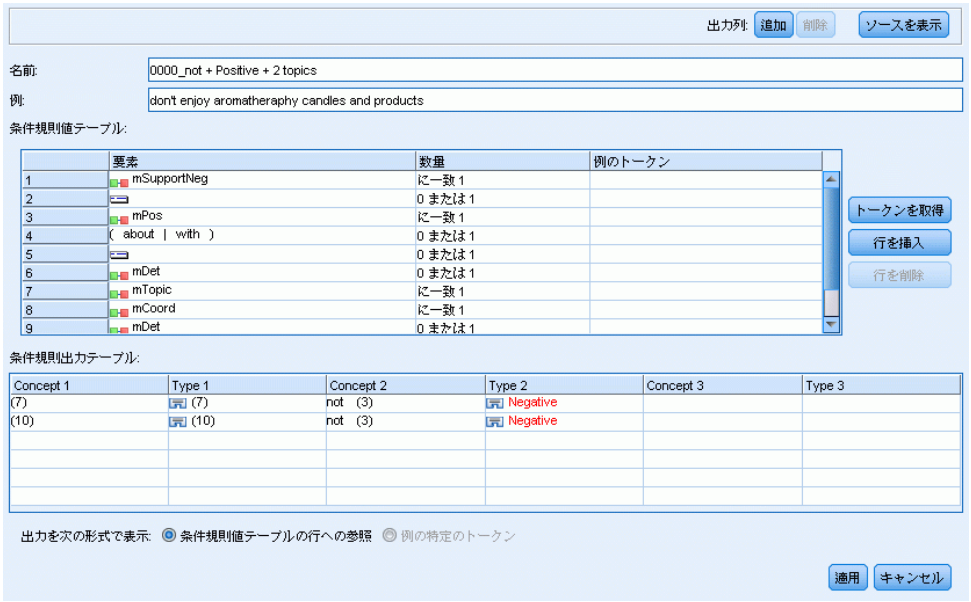
SEP

定義済みのマクロ SEP も使用できます。これは、ローカル コンピュータで定義されたグローバル区切り文字、通常カンマ（,）に対応します。

テキスト リンク規則の使用

テキスト リンク分析規則は、文にマッチを実行するために使用するブール型質問です。テキスト リンク分析には、1つまたは複数の引数が含まれます：タイプ、マクロ、理れらる文字列、単語の空所などです。TLA 結果を抽出するには、テキスト リンク分析規則が少なくとも 1 つ必要です。

図 19-10
[テキスト リンク規則] タブ:条件規則エディタ



[名前] フィールド:テキスト リンク分析規則の一意の名前。

[例] フィールド:オプションで、この規則でキャプチャされる例の文または語の連鎖を含むことができます。例の使用をお勧めします。このエディタ内でこのテキスト例からトークンを生成し、規則にどのように合致し、どのように出力されるかを確認できます。 **トークン**は、抽出プロセス時に特定される単語または語句として定義されます。たとえば、文「My uncle lives in New York」 の場合、my、uncle、lives、in、および new york というトークンがあります。また、uncle はコンセプトとして抽出され、タイプは <**不明**>、そして new york はコンセプトとして抽出され、タイプは <**地名**> となります。すべてのコンセプトはトークンとなりますが、すべてのトークンがコンセプトとなるわけではありません。トークンは、マクロ、リテラル文字列、そして単語の空所の場合もあります。タイプを指定された単語または語句のみがコンセプトとなります。

条件規則値テーブル:このテーブルには、文に対し規則を合致させるために使用する規則の要素が含まれています。右側のボタンを使用して、テーブルの行を追加または削除できます。テーブルは、次の 3 つの列で構成されています。

- **[要素]** 列:1 つのタイプ、リテラル文字列、単語の空所 (<**任意のトークン**>)、またはマクロとして値を入力します。詳細は、 p. 392 条件規則およびマクロにサポートされている要素 を参照してください。 要素セルをダブルクリックして、情報を直接入力します。または、セルを右クリックすると、共通のマクロ、タイプ名、固有表現キーワード名のリス

トを提供するコンテキスト メニューが表示されます。セルに情報を入力する場合、マクロまたはタイプ名の先頭に ‘$’ 文字を追加する必要があります（例: マクロ mTopic の場合 $mTopic)。要素の行を作成する順序は、規則がテキストにどのように合致するかにおいて重要です。引数を結合する場合、カッコ （ ） を使用して引数およびブール型演算子 OR を示す文字 | をグループ化する必要があります。値では大文字と小文字が区別されます。

- **［数量］**列:合致が発生するために要素が検出される必要のある回数の、最大値および最小値を示します。たとえば、0 ～ 3 つの単語のいずれかの箇所の 2 つの要素間に空所、または一連の単語を定義する場合、リストから **[範囲: 0 and 3]** を選択するか、ダイアログ ボックスに直接数値を入力することができます。デフォルトは「**1 に一致**」です。要素をオプションにすることが必要な場合があります。その場合、最小数量が 0 および最大数量が 0 より大きくなります（例: 0 または 1、0 ～ 2)。規則の最初の要素をオプションにはできません。つまり最初の要素の数量を 0 にすることはできません。
- **［例のトークン］**列:**[トークンを取得]** をクリックすると、**[例]** のテキストをトークンに分割し、これらのトークンを使用して、定義した要素と合致するトークンを列に入力します。出力テーブルにこれらのトークンを表示することもできます。

条件規則出力テーブル:このテーブルの各行は、TLA パターン出力が結果にどのように表示されるかを定義します。条件規則出力では、それぞれ「スロット」を示す最大 6 つのコンセプト/タイプ列のペアのパターンを生成できます。たとえば、タイプ パターン <**地名**> + <**肯定的**> は、 2 つのコンセプト/タイプ列のペアで構成される 2 スロットのパターンです。

言語によって、同じ基本的な考えをさまざまな方法で自由に表現できますが、同じ基本的な考えをキャプチャするために多くの条件規則を定義する場合があります。たとえば、「Paris is a place I love」と「I really, really like Paris and Florence」というテキストは同じ基本の考え「Paris is liked」を示しますが、異なる方法で表現され、両方をキャプチャするには 2 つの異なる条件規則が必要です。ただし、類似した考えがグループ化されている場合、パターン結果を処理するとより簡単です。このため、2 つの異なる条件規則でこれら 2 つの語句をキャプチャしますが、タイプ パターン <**地名**> + <**肯定的**> のようにいずれのテキストも示すよう、2 つの条件規則に同じ出力を定義することができます。そして、出力が元のテキストの構造または語順と必ずしも同じでないことがわかります。さらに、このようなタイプのパターンは他の句と一致する可能性があり、以下のようなコンセプトを生み出す可能性があります：たとえば、paris + like および tokyo + likeです。

少ないエラーで出力をすばやく定義するには、コンテキスト メニューを使用して、出力に表示する要素を選択できます。また、条件規則値テーブルから要素を出力にドラッグ アンド ドロップすることもできます。たとえば、条件規則値テーブルの行 2 に mTopic マクロへの参照を含む規則が

あり、値を出力に投入する場合、mTopic の要素を条件規則出力テーブルの最初の列のペアにドラッグ/ドロップできます。このようにすると、選択したペアのコンセプトとタイプが自動的に投入されます。または、条件規則値テーブルの 3 番目の要素（行 3）で定義されたタイプから出力する場合、そのタイプを条件規則値テーブルから出力テーブルの [**タイプ 1**] セルにドラッグします。テーブルが更新され、(3) に行の参照が表示されます。

また、出力する [**コンセプト**] 列をダブルクリックし、$ の後に行番号を入力（例: $2、条件規則値テーブルの行 2 で定義された要素を参照）して、これらの参照を手動でテーブルに入力することもできます。情報を手動で入力する場合、[**タイプ**] 列を定義し、# の後に行番号を入力（例: #2、条件規則値テーブルの行 2 で定義された要素を参照）する必要もあります。

さらに、これらの方法を組み合わせることもできます。条件規則値テーブルの行 4 にタイプ <**肯定的**> があるとします。そのタイプを [**タイプ 2**] 列にドラッグし、[**コンセプト 2**] 列のセルをクリックしてそのタイプの前に「not」という単語を手動で入力できます。テーブルの出力列は not (4)、あるいは編集モードまたは入力モードの場合は not $4 となります。[タイプ 1] 列を右クリックして、たとえば mTopic というマクロを選択します。するとこの出力で、以下のようなコンセプト パターンが生成されます：car + bad。

多くの条件規則には出力が 1 つしかありませんが、複数の出力を生成できるまたは生成する必要がある場合があります。この場合、条件規則出力テーブルの行ごとに 1 つの出力を定義します。

重要! TLA パターンの間、他の言語処理が実行されます。そのため、出力が t$3\t#3 を読み込む場合、パターンは 3 番目の要素の最終コンセプトと 3 番目の最終タイプを、言語処理が適用された後（類義語およびその他のグループ）に表示します。

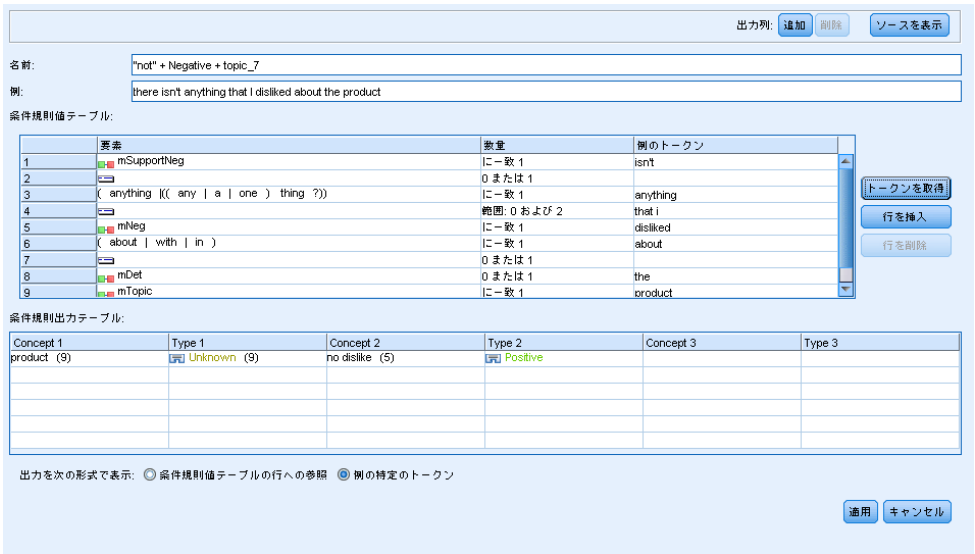
- **出力を次の形式で表示:** デフォルトでは、[**条件規則値テーブルの行への参照**] オプションが選択され、[条件規則値] タブで定義されているような行への数値参照を使用して出力が表示されます。前に [トークンを取得] をクリックして、条件規則値テーブルの [例のトークン] 列にトークンがある場合、このオプションを選択して、これら特定のトークンの出力を表示できます。

注:出力テーブルに表示できるコンセプト/タイプの出力ペアが十分にない場合、エディタ ツールバーの [追加] ボタンをクリックして別のペアをクリックできます。現在 3 つのペアが表示され、[追加] をクリックすると、2 つの列（コンセプト 4 およびタイプ 4）がテーブルに追加されます。つまり、すべての条件規則の出力テーブルに 4 つのペアが表示されます。このライブラリのルール セットのルールがそのペアを使用しないかぎり、未使用のペアを削除することもできます。

条件規則の例

リソースに次のテキスト リンク分析規則が含まれ、TLA 結果の抽出を有効にしたとします。

図 19-11
[テキスト リンク規則] タブ:条件規則エディタ



抽出すると、抽出エンジンが各文を読み込み、次の順序に合致させようとします。

要素 (行)	引数の説明
1	マクロ mPos または mNeg、タイプ <**不確定**> で示されるいずれかのタイプのコンセプト。
2	マクロ mTopic で示されるいずれかのタイプに指定されたコンセプト。
3	マクロ mBe で示されるいずれかの語。
4	オプションの要素、0 個または 1 つの語で、単語の空所または <**任意のトークン**> として参照。
5	マクロ mTopic で示されるいずれかのタイプに指定されたコンセプト。

出力テーブルには、この規則の必要なすべてが条件規則値テーブルの行 5 で定義された mTopic マクロに対応するコンセプトまたはタイプ + 条件規則値テーブルの行 1 で定義された mPos、mNeg、または <**不確定**> に対応するコンセプトまたはタイプのパターンであることを示します。sausage + like または <**不明**> + <**肯定的**> となります。

条件規則の作成および編集

新しい条件規則を作成したり、既存のマクロを編集できます。条件規則エディタのガイドラインおよび説明に従います。 詳細は、 p. 383 テキスト リンク規則の使用 を参照してください。

条件規則の新規作成

- メニューの［**ツール**］→［**新しい条件規則**］を選択します。また、ツリー ツールバーの［新規条件規則］アイコンをクリックすると、エディタに新しい条件規則が表示されます。
- 一意の名前を入力して、条件規則値の要素を定義します。
- エラーの有無を確認したら、［**適用**］をクリックします。

条件規則の編集

- ツリー内の条件規則名をクリックします。右側のエディタ パネルに条件規則が表示されます。
- 変更を加えます。
- エラーの有無を確認したら、［**適用**］をクリックします。

条件規則の無効化および削除

条件規則の無効化

処理中に条件規則が無視されるようにするには、マクロを無効にすることができます。条件規則を削除および無効にする場合は注意してください。

- ツリー内の条件規則名をクリックします。右側のエディタ パネルに条件規則が表示されます。
- 名前を右クリックします。
- コンテキスト メニューから、［**無効にする**］を選択します。条件規則のアイコンがグレーになり、条件規則自体を編集できなくなります。

条件規則の削除

条件規則を除外する場合、削除することができます。条件規則を削除および無効にする場合は注意してください。

- ツリー内の条件規則名をクリックします。右側のエディタ パネルに条件規則が表示されます。

- 名前を右クリックします。
- コンテキスト メニューから、[削除] を選択します。リストから条件規則が消去されます。

エラーのチェック、保存およびキャンセル

条件規則の変更の適用

条件規則エディタの外をクリックまたは [適用] をクリックすると、条件規則が自動的にスキャンされ、エラーの有無が確認されます。エラーが見つかると、アプリケーションの別の部分に移る前に修正する必要があります。

ただし、あまり深刻でないエラーが検出された場合は、警告のみが表示されます。たとえば、条件規則にタイプまたはマクロに対する不完全または参照されない定義が含まれている場合、警告メッセージが表示されます。[適用] をクリックすると、未修正の警告により、左側パネルのツリーの条件規則名の左側に警告アイコンが表示されます。

条件規則を適用しても、条件規則が永続的に保存されるわけではありません。適用すると、検証プロセスで、エラーおよび警告の有無をチェックします。

インタラクティブ ワークベンチ セッション内のリソースの保存

- インタラクティブ ワークベンチ セッションでリソースに行った変更を保存し、次回ストリーム実行時にそれらの変更を取得できるようにするには、次の手順を実行する必要があります。

  モデル作成ノードを更新して、次回ストリーム実行時にこれらの同じリソースを取得できるようにします。詳細は、 8 章 p.148 モデル作成ノードの更新および保存 を参照してください。 その後ストリームを保存します。ストリームを保存するには、モデル作成ノードを更新した後、IBM® SPSS® Modeler のメイン ウィンドウで保存します。

- インタラクティブ ワークベンチ セッションでリソースに行った変更を保存し、他のストリームでそれらを使用できるようにするは、次の手順を実行します。
  - 使用していたテンプレートを更新するか、新しいテンプレートを作成します。詳細は、 14 章 p.292 テンプレートの作成および更新 を参照してください。 現在のノードの変更を保存するわけではありません（前の手順を参照）。
  - または、使用していた TAP を更新します。 詳細は、 10 章 p.250 テキスト分析パッケージの更新 を参照してください。

テンプレート エディタ 内のリソースの保存

- まず、ライブラリを公開します。 詳細は、 16 章 p. 328 ライブラリの公開 を参照してください。
- そして、メニューで [ファイル] → [リソース テンプレートを保存] を使用してテンプレートを保存します。

条件規則変更のキャンセル

- 変更を破棄する場合は、エディタ パネルで [キャンセル] をクリックします。

条件規則の処理順序

テキスト リンク分析が抽出時に実行されると、合致が見つかるまで、またはすべての条件規則が終了するまで、「文」（句、語、語句）が各条件規則に対して順番にマッチされます。ツリー内の位置は、条件規則が使用される順序を示します。条件規則の順序は具体的なものから一般的なものへの順序がベスト プラクティスです。最も具体的な条件規則がツリーの最上位になります。特定の条件規則またはルール セットを変更するには、条件規則とマクロ ツリーのコンテキスト メニューで [1つ上に移動] または [1つ下に移動] を選択するか、ツールバーの上方向矢印または下方向矢印を選択します。

ソース ビュー内で作業している場合、エディタ内を移動することによって条件規則の順序を変更することはできません。ソース ビュー内で上位に表示されている条件規則から処理されます。コピー/貼り付けの問題を回避するために、条件規則の順序の変更はツリー内でのみ行うことを強くお勧めします。

重要! 以前のバージョンの IBM® SPSS® Modeler Text Analytics では、一意で数値の条件規則 ID が必要でした。バージョン 15 以降では、条件規則をツリー内の上下に移動して、あるいはソース ビュー内の位置によってのみ処理順を示します。

次の 2 つの文を含むテキストがあるとします。

```
I love anchovies
I love anchovies and green peppers
```

また、値が次のようになる 2 つのテキスト リンク分析規則があるとします。

図 19-12
2 つの条件規則の例

A

	要素	数量	例のトークン
1	Positive	に一致 1	
2	mDet	0 または 1	
3	mTopic	に一致 1	
4			
5			
6			
7			

B

	要素	数量	例のトークン
1	Positive	に一致 1	
2	mDet	0 または 1	
3	mTopic	に一致 1	
4	(SEP | and | or)	0 および 2	
5	mDet	0 または 1	
6	mTopic	に一致 1	
7			

ソース ビューで、条件規則値が次のようになります。

```
  A:value = $Positive $mDet?$mTopic
  B:value = $Positive $mDet?$mTopic ($SEP|and|or){1,2}
$mDet?$mTopic
```

ルール A が B よりツリー上で高い位置（頂点近く）にある場合は、A が先に処理され、I love anchovies and green peppersの分は最初に`$Positive $mDet?$mTopic` とマッチングされ、完全なパターン出力（anchovies + like）を生成します。これは、2つの `$mTopic` 一致を検索しないというルールによるものです。

そのため、テキストの核心をキャプチャするには、最も具体的な規則（この場合 B）がより一般的な規則（この場合 A）よりツリー内の上位になければいけません。

ルール セットの使用（多段階処理）

ルール セットとは、多段階処理を実行するために条件規則とマクロ ツリーの関連する一連の条件規則をグループ化する有用な方法です。ルール セットには、名前以外にそれ自体の定義はありませんが、ルール セットを使用して、条件規則を意味のあるグループに編成します。一度の通過で処理するにはテキストが多すぎ、また多様である場合があります。たとえば、セキュリティに関する情報データを処理する場合、連絡方法（x が y に電話）、家族関係（y の義理の兄弟 x）、金銭のやり取り（x wired $100 to y）などで明らかになる個人間の繋がりがテキストに含まれている場合があります。この場合、連絡先が明らかになる定義、家族構成が明らかに

なる定義など、特定の種類の関係性に焦点を当てたテキスト リンク分析規則の特殊なセットを作成すると役立ちます。

ルール セットを作成するには、条件規則とマクロ ツリーのコンテキスト メニューまたはツールバーで［ルール セットを作成］を選択します。ツリーのルール セット ノードで直接新しい条件規則を作成するか、既存の規則をルール セットに移動できます。

条件規則がルール セットにグループ化されるリソースを使用して抽出を実行すると、抽出エンジンは、テキスト全体で複数の通過を行い、通過ごとにさまざまな種類のパターンを合致させます。このように、「文」を各ルール セットの条件規則に合致させることができます。ルール セットがない場合は単一の条件規則にのみ合致させることができます。

注:ルール セットごとに条件規則を、512 個まで定義することができます。

ルール セットの新規作成

- メニューの［ツール］→［新規ルール セット］を選択します。または、ツリーのツールバーで［新規ルール セット］アイコンをクリックします。条件規則ツリーにルール セットが表示されます。
- このルール セットの新しい条件規則を追加するか、既存の条件規則をセット内に移動します。

ルール セットの無効化

- ツリー内のルール セット名をクリックします。
- コンテキスト メニューから、［無効にする］を選択します。ルール セットのアイコンがグレーになり、そのルール セットに含まれるすべての条件規則も無効となって、処理時に無視されるようになります。

ルール セットの削除

- ツリー内のルール セット名をクリックします。
- コンテキスト メニューから、［削除］を選択します。ルール セットと、それに含まれるすべての条件規則がリソースから削除されます。

条件規則およびマクロにサポートされている要素

次の引数は、テキスト リンク分析規則およびマクロの値パラメータに受け入れられます。

マクロ

テキスト リンク分析規則または別のマクロで直接マクロを使用できます。コンテキスト メニューからマクロ名を選択するのではなく、手動でまたはソース ビューからマクロ名を入力する場合、$mTopic のように、名前の前にドル記号（$）を必ず追加してください。マクロ名では大文字と小文字が区別されます。コンテキスト メニューでマクロを選択する場合、現在の［テキスト リンク規則］タブで定義されたマクロから選択できます。

タイプ

テキスト リンク分析規則またはマクロのタイプを直接使用できます。コンテキスト メニューからタイプ名を選択するのではなく、手動でまたはソース ビューからタイプ名を入力する場合、$Person のように、名前の前にドル記号（$）を必ず追加してください。タイプ名では大文字と小文字が区別されます。コンテキスト メニューを使用する場合、使用されているリソースの現在のセットからタイプを選択できます。

不明のタイプを参照する場合、警告メッセージが表示され、修正されるまで条件規則とマクロ ツリーの条件に警告メッセージが表示されます。

リテラル文字列

抽出されなかった情報を追加するために、抽出エンジンが検索するリテラル文字列を定義できます。抽出されたすべての単語または句はタイプに割り当てられるため、それらをリテラル文字列で使用することはできません。抽出された単語を使用すると、そのタイプが <**不明**> であっても、その単語は無視されます。

リテラル文字列は、1 単語または複数の単語の場合があります。次の規則は、リテラル文字列のリストを定義する場合に適用されます。

- (his) のように、文字列のリストをカッコで囲む。リテラル文字列を選択する場合、各文字列を (a|an|the) または (his|hers|its) のように OR 演算子で区切る必要があります。
- 単語または複合語を使用する。
- リスト内の単語を、ブール型演算子 OR と同様に機能する | 文字で区切る。
- 単数形および複数形に一致させたい場合は両方を入力する。活用形は自動的に生成されません。
- 小文字の実を使用する。

- リテラル文字列を再利用するには、マクロとしてそれらを定義してから他のマクロおよびテキスト リンク分析規則でそのマクロを使用する。
- 文字列にピリオド（終止符）またはハイフンを使用している場合は、それらを含める必要がある。たとえば、テキストの a.k.a を一致させるには、文字 a.k.a と共にピリオドをリテラル文字列として入力してください。

排他演算子

排他演算子として ! を使用し、否定の式が特定のスロットに含まれないようにします。インライン セル編集（条件規則値テーブルまたはマクロ値テーブルのセルをダブルクリック）またはソース ビューから手動でのみ排他演算子を追加できます。たとえば、$mTopic @{0,2} !($Positive) $Budget をテキスト リンク分析に追加すると、(1) mTopic マクロのいずれかのタイプに割り当てられたキーワードを含み、(2) 0 ～ 2 語の単語の空所を含み、(3) <Positive> タイプに割り当てられたキーワードのインスタンスを含まず、(4) <Budget> タイプに割り当てられたキーワードを含むテキストを検索します。「cars have an inflated price tag」はキャプチャされますが、「store offers amazing discounts」は無視されます。

この演算子を使用するには、セルをダブルクリックして、要素のセルに感嘆符のポイントおよびカッコを入力する必要があります。

単語の空所（<任意のトークン>）

単語の空所（または <**任意のトークン**>）は、2 つの要素間に存在するトークンの数値範囲を定義します。単語の空所は、追加の決定詞、前置詞句、形容詞またはその他の単語の有無によってわずかに異なる非常に類似した句と合致させる場合に役立ちます。

テーブル 19-1
単語の空所のない条件規則値テーブルの要素の例

#	要素
1	**不明**
2	mBeHave
3	**肯定的**

注: ソースビューではこの値は以下のように定義されます：$Unknown $mBeHave $Positive

この値は「the hotel staff was nice」のような文と合致します。ここで、「hotel staff」のタイプは <**不明**> となり、「was」はマクロ mBeHave、「nice」は <**肯定的**> となります。ただし、「the hotel staff was very nice」とは一致しません。

テーブル 19-2
<任意のトークン> のある条件規則値テーブルの要素の例

#	要素
1	不明
2	mBeHave
3	
4	肯定的

注: ソースビューではこの値は以下のように定義されます：$Unknown $mBeHave @{0,1} $Positive

単語の空所を条件規則値に追加すると、「the hotel staff was nice」および「the hotel staff was very nice」のいずれの文にも合致します。

ソース ビューまたはインライン編集の場合、単語の空所のシンタックスは @{#,#} です。この場合、@ は単語の空所を示し、{#,#} は、先行する要素と後続の要素の間に受け入れられる単語数の最小値および最大値を定義します。たとえば、@{1,3} は、定義された 2 つの要素の間に少なくとも 1 ～ 3 つの単語がある場合、この 2 つの要素の間に合致がある可能性があることを示します。@{0,3} は 0、1、2 または 3 この単語がある場合に定義された 2 つの要素の間の合致があると考えられることを示します。

入力モードでの表示および作業

各条件規則およびマクロについて、TLA エディタは TLA 出力を合致および作成する抽出機能で使用する基底のソース コードを生成します。コード自体を処理する場合、エディタ上部の “View Source” ボタンをクリックして、このソースコードを表示したり直接編集したりすることができます。ソース ビューが、現在選択している条件規則またはマクロにジャンプして、強調表示します。ただし、エディタ パネルを使用してエラーの発生を少なくすることをお勧めします。

ソースの表示または編集を終了するには、[**ソースを終了**] をクリックします。条件規則に無効なシンタックスを生成した場合、ソース ビューを終了する前にそのシンタックスを修正する必要があります。

重要! ソース ビューで編集する場合、一度に 1 つずつ条件規則およびマクロを編集することを強くお勧めします。マクロを編集した後、抽出して結果を検証してください。結果が適切である場合、他の変更を行う前にテンプレートを保存することをお勧めします。結果が適切でない場合、またはエラーが発生した場合、保存したリソースに戻ってください。

ソース ビューのマクロ

[macro]
name = macro_name
value = ([type_name|macro_name|literal_string|word_gap])

[macro]	各マクロは、[macro] とマークされた行から開始し、マクロの開始を示す必要があります。
name	マクロ定義の名前。それぞれの名前は一意である必要があります。
value	1 つまたは複数のタイプ、リテラル文字列、単語の空所またはマクロの組み合わせ。詳細は、 p. 392 条件規則およびマクロにサポートされている要素 を参照してください。引数を結合する場合、カッコ ( ) を使用して引数およびブール型演算子 OR を示す文字	をグループ化する必要があります。

マクロのセクションで説明されているガイドラインおよびシンタックスのほか、ソース ビューにはエディタ ビューでの作業に不要のガイドラインがいくつかあります。入力モードで作業する場合、マクロは次のことを重点に置く必要があります。

- 各マクロは、[macro] とマークされた行から開始し、マクロの開始を示す必要があります。
- 要素を無効にするには、行の前にコメントを示すインジケータ (#) を追加します。

例: この例では、mTopic というマクロを定義します。mTopic の値は、次のうちの 1つに一致するキーワードの存在を示します。<Product>、<Person>、<Location>、<Organization>、<Budget>、または <Unknown>。

```
[macro]
name=mTopic
value=($Unknown|$Product|$Person|$Location|$Organization|$Budget|$Currency)
```

ソース ビューの条件規則

[pattern(ID)]
name = pattern_name
value = [$type_name|macro_name|word_gaps|literal_strings]
output = $digit[\t]#digit[\t]$digit[\t]#digit[\t]$digit[\t]#digit[\t]

[pattern (<ID>)]	テキスト リンク分析規則の開始を示し、処理の順序を決定する一意の数値 ID を与えます。
name	テキスト リンク分析規則の一意の名前を示します。

value　テキストに一致するシンタックスおよび引数を提供します。 詳細は、 p. 392 条件規則およびマクロにサポートされている要素 を参照してください。

output　テキストで見つかった合致パターンの出力形式。出力は、ソース テキストの要素の正確な元の位置と必ずしも似ているわけではありません。また、各行に出力を配置することで、指定されたテキスト リンク分析規則に複数の出力行を指定することができます。

出力のシンタックス:

- `$1\t#1\t$3\t#3` のように、出力をタブ コード `\t` で区切る。
- 該当する位置の値パラメータで定義された引数に一致するキーワードの `$` および数値呼び出し。つまり、`$1` は値に定義された最初の引数に一致するキーワードを意味します。
- 該当する位置の要素のタイプ名の `#` および数値呼び出し。項目がリテラル文字列のリストである場合、タイプ <**不明**> が割り当てられます。
- `Null\tNull` の値は、出力を作成しません。

条件規則のセクションで説明されているガイドラインおよびシンタックスのほか、ソース ビューにはエディタ ビューでの作業に不要のガイドラインがいくつかあります。入力モードで作業する場合、条件規則は次のことを重点に置く必要があります。

- 複数の要素が定義されている場合、オプションであるかどうかに関係なく、カッコで囲む必要があります（例: `($Negative|$Positive)` または `($mCoord|$SEP)?`）. `$SEP` はカンマを示します。
- テキスト リンク分析規則の最初の要素を、オプションの要素にすることはできません。たとえば、`value = $mTopic?` または `value = @{0,1}` から始めることはできません。
- 数量（またはインスタンス数）をトークンに関連付けることができます。これは、ケースごとに個別の規則を作成するのではなく、すべてのケースを網羅する規則を 1 つだけ作成する場合に役立ちます。たとえば、`,`（カンマ）または `and` のいずれかに合致させる場合、リテラル文字列 `($SEP|and)` を使用できます。数量を追加してリテラル文字列が `($SEP|and){1,2}` となるよう拡張すると、以下のいずれかに一致するようになります。”`,`” “**および**” “`, and`”.
- テキスト リンク分析規則 `value` のマクロ名および `$` および `?` との間にスペースは使用できません。
- テキスト リンク分析規則 `output` にスペースは使用できません。
- 要素を無効にするには、行の前にコメントを示すインジケータ (#) を追加します。

例:リソースに次の TLA テキスト リンク分析規則が含まれ、TLA 結果の抽出を有効にしたとします。

```
## Jean Doe was the former HR director of IBM in France
[pattern(201)]
name= 1_201
value = $Person ($SEP|$mDet|$mSupport|as|then){1,2} @{0,1} $Function
 (of|with|for|in|to|at) @{0,1} $Organization @{0,2} $Location
output = $1\t#1\t$4\t#4\t$7\t#7\t$9\t#9
```

抽出すると、抽出エンジンが各文を読み込み、次の順序に合致させようとします。

位置	引数の説明
1	個人の名前（$Person）
2	以下のうち1つまたは2つとします：カンマ（$SEP）、区切り文字（$mDet）、助動詞（$mSupport）、文字列 “then” または “as”
3	0 または 1 つの単語（@{0,1}）
4	役職（$Function）
5	いかの文字列のうち一つです：文字列 “of”、“with”、“for”、“in”、“to”、または “at”
6	0 または 1 つの単語（@{0,1}）
7	組織の名前（$Organization）
8	0, 1 または 2 つの単語（@{0,2}）
9	場所の名前（$Location）

このテキスト リンク分析規則のサンプルは、次のような文または句に合致します。

Jean Doe, the HR director of IBM in France
Jean Doe was the former HR director of IBM in France
IBM appointed Jean Doe as the HR director of IBM in France

このテキスト リンク分析規則のサンプルは、次のような出力を作成します。

```
jean doe <Person> hr director <Function> ibm <Organization> france <Location>
```

この場合、次のようになります。

- jean doe は、$1（テキスト リンク分析規則の最初の要素）に対応するキーワードで、<Person> は jean doe（#1）のタイプです。
- director は、$4（テキスト リンク分析規則の 4 番目の要素）に対応するキーワードで、<Function> は hr director（#4）のタイプです。
- ibm は、$7（テキスト リンク分析規則の 7 番目の要素）に対応するキーワードで、<Organization> は ibmの以下となります。（#7),
- france は、$9（テキスト リンク分析規則の 9 番目の要素）に対応するキーワードで、<Location> は france（#9）のタイプです。

ソース ビューのルール セット

[set(<ID>)]

[set (<ID>)]	ルール セットの開始を示し、そのセットの処理の順序を決定する一意の数値 ID を与えます。

例:次の文には、人名、会社内の役職、その会社の合併/買収の活動が含まれています。

```
IBM has entered into a definitive merger agreement with SPSS,
said Jack Noonan, CEO of SPSS.
```

いくつかの出力で条件規則を作成し、次のような出力をすべて処理できます。

```
## IBM entered into a definitive merger agreement with SPSS, said
Jack Noonan, CEO of SPSS.

[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
```

この場合、次の 2 つの出力パターンが出力されます。

- ibm <Organization> + merges with <ActiveVerb> + spss <Organization>
- jack noonan <Person> + ceo <Function> + spss <Organization>

重要! TLA パターンの間、他の言語処理が実行されます。この場合、merger は、抽出プロセスの類義語グループ段階で merges with の下にグループ化されます。merges with は <ActiveVerb> タイプに含まれるため、このタイプ名は最終 TLA パターン出力に表示されます。そのため、出力が t$3\t#3 を読み込む場合、パターンは 3 番目の要素の最終コンセプトと 3 番目の最終タイプを、言語処理が適用された後（類義語およびその他のグループ）に表示します。

前述のような複雑な条件規則を作成する代わりに、2 つの条件規則を使用して容易に管理できます。1 つ目は、会社間で合併者/買収者を検出します。

```
[set(1)]
## IBM has entered into a definitive merger agreement with SPSS
[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
```

ibm <Organization> + merges with <ActiveVerb> + spss <Organization> を生成します。

2 つ目は人名/役職/会社に特化されています。

```
[set(2)]
## said Jack Noonan, CEO of SPSS
[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
```

```
jack noonan <Person> + ceo <Function> + spss <Organization>
```

を生成します。

付録

A

日本語テキストの例外

注:本章の機能はIBM® SPSS® Modeler Premiumでのみ利用可能です。

日本語テキストは IBM® SPSS® Modeler Text Analytics の他のサポート言語と同じような方法で処理およびマイニングされますが、多くの相違点があります。小さな相違点については、本マニュアルのほかのすべての言語の指示とともに記載されますが、大きな相違点については、この付録の章で説明します。

注:Text Mining for Clementine 日本語バージョン 2.2 以降では、多くの新しい改善点があります。これらの改善点の詳細については、営業担当者に連絡してください。

日本語テキストの抽出およびカテゴリ化

日本語テキストをマイニングする場合、サポートされている他の言語と同じような処理となります。詳細は、 1 章 p.2 テキスト マイニングについて を参照してください。 ただし、日本語の処理では、次のような異なる部分があります。

抽出の方法

回答の主要キーワードの抽出時、IBM® SPSS® Modeler Text Analytics は言語学に基づくテキスト分析に依存します。このアプローチを用いると統計に基づくシステムがもたらすようなスピードと費用対効果が得られます。また人の手を介することがほとんどないので、極めて高い精度が得られます。言語学に基づくテキスト分析は、自然言語処理とも、計量言語学とも呼ばれる研究の分野に基づきます。

注:日本語テキスト展開はIBM® SPSS® Modeler Premiumで利用可能です。

日本語テキストにおける抽出プロセス時の統計的アプローチと言語学的アプローチとの差を、例として**沈む**という単語を使用して説明することができます。この単語を使用した、**日が沈む**や、**気分が沈む**という表現があります。統計的アプローチのみを使用すると、**日**、**気分**、および**沈む**という単語がそれぞれ個別に抽出されます。ただし、言語学的手法に基づく感性分析を使用すると、 **日**、**気分**、および **沈む**が抽出されるだけでなく、**気分が沈む**も抽出され、タイプ <**悪い － 悲しみ全般**> に割り当てられます。感性分析で言語学に基づいた手法を使用すると、より意味のある表現を抽出できます。感性の分析およびキャプチャを行い、定義と

いうより信頼できるアプローチによって、言語学的テキスト マイニングを実行してテキストの曖昧さに切り込みます。

抽出プロセスがどうなっているのかを理解しておくと、言語リソース（ライブラリ、タイプ、類義語など）を微調整する際に役に立ちます。抽出プロセスのステップには以下のものがあります。

- ソース データの標準フォーマットへの変換
- 候補のキーワードの特定
- 類義語の等価クラスおよび統合の特定
- タイプの割り当て
- 二次分析によるインデックスの付与、および必要に応じてパターン マッチ

手順 1: ソース データの標準フォーマットへの変換

最初のステップでは、後続の分析に利用できるように、インポートしたデータを決まった形式に変換します。この変換は内部的に実行され、元のデータは変更されません。

手順 2: 候補のキーワードの特定

言語学的抽出において、候補となるキーワードを特定する際の言語リソースの役割を理解しておくのは大切なことです。言語リソースは、抽出が実行されるごとに使用されます。言語リソースは、テンプレート、ライブラリ、およびコンパイル済み辞書の形式で保存されています。ライブラリには、語のリスト、関係性、また抽出の実行や調整に使用される情報が含まれています。基幹辞書は表示・編集ができません。ただし、残りのリソースをテンプレート エディタ で、またはインタラクティブ ワークベンチ セッションの場合はリソース エディタ で編集できます。

コンパイル済み辞書は、SPSS Modeler Text Analytics の抽出エンジンの主要な、内部コンポーネントです。これらのリソースには、品詞コード（名詞、動詞、形容詞など）を含む基本形のリストを収めた一般辞書が含まれています。また、リソースには、<**地名**>、<**組織**>、または<**人名**>に多くの抽出されたキーワードを割り当てるために使用する、予約済みのビルトインのタイプも含まれています。詳細は、 p. 412 日本語テキストで使用できるタイプ を参照してください。注:日本語テキスト展開はSPSS Modeler Premiumで利用可能です。

これらコンパイル済み辞書のほか、製品にはいくつかのライブラリが付属し、それらを使用して、コンパイル済み辞書のタイプ定義およびコンセプト定義を補い、また類義語を提供することができます。これらのライブラリ、および作成したユーザー指定のライブラリは、いくつかの辞書で構成されています。これらには、キーワード辞書、類義語辞書、および不要語辞書が含まれています。詳細は、 p. 407 日本語テキストのリソースの編集 を参照してください。注:日本語テキスト展開はSPSS Modeler Premiumで利用可能です。

データがインポートおよび変換されると、抽出エンジンは抽出の候補のキーワードの特定を開始します。候補となるキーワードとは、テキスト内の概念を特定するのに使用される語や、語の集まりのことです。テキストを処理しているとき、単語（**ユニターム**）および複合語（**マルチターム**）は、品詞パターン抽出を使用して特定されます。たとえば、品詞パターンが「<**地名**> + <**名詞**>」のマルチターム**青森りんご**は、2 つの部分に分けられます。そして、候補の感性キーワードは、感性テキスト リンク分析を使用して特定されます。

たとえば、次のような日本語のテキストがあるとします。**写真が新鮮で良かった**。この場合、抽出エンジンは、感性テキスト リンク規則のいずれかを使用して（**品物**）+ **が** + **良い** を合致させた後、感性タイプ **良い - 褒め・賞賛** を割り当てます。

注:前述のコンパイル済み一般辞書にあるキーワードは、ユニタームとして重要でないまたは言語学的にあいまいであるすべての単語を示します。これらの単語は、ユニタームを特定するときに不要語に追加されます、ただし、それらは、品詞を決定またはより長い候補の複合語（マルチターム）を参照している場合に再評価されます。

手順 3: 類義語の等価クラスおよび統合の特定

候補のユニタームおよびマルチタームが特定された後、正規化辞書を使用して、等価クラスを特定します。等価クラスは、ある語句の基本形、すなわち同じ語句の２つの表現を１つの形で表わしたものです。句を等価クラスに割り当てる目的は、たとえば、side effect と **副作用** を別のコンセプトとして扱わないようにすることです。等価クラスのどのコンセプトを使用するか、つまり、side effect または**副作用**のどちらを主要キーワードとして使用するかを判断するために、抽出エンジンは、次の規則を順に適用します。

- ライブラリのユーザー指定の形式。
- コンパイル済みリソースで定義されている最も頻度の高い形式。

手順 4: タイプの割り当て

次に、抽出されたコンセプトにタイプを割り当てます。タイプは、コンセプトの意味上のグループ化です。基幹辞書ならびにライブラリの両方がこのステップで使用されます。タイプには、上位レベルのコンセプト、肯定的な単語および否定的な単語、人名、地名、組織名などが含まれます。詳細は、 17 章 p. 332 キーワード辞書 を参照してください。

日本語リソースには、タイプの異なるセットがあります。詳細は、p. 412 日本語テキストで使用できるタイプ を参照してください。注:日本語テキスト展開はSPSS Modeler Premiumで利用可能です。

手順 5: イベント抽出によるインデックスの付与およびパターン マッチ

レコードまたはドキュメントのセット全体に、テキストの位置と各等価クラスの代表キーワードの間にポインタを確定してインデックスを付けます。候補のコンセプトの活用形インスタンスはすべて、候補の基本型としてインデックスが付けられます。基本形ごとに全体の出現頻度が計算されます。

SPSS Modeler Text Analytics は、タイプやコンセプトだけでなく、それらの関係性も見つけることができます。この製品ではいくつかのアルゴリズムおよびライブラリを使用でき、またタイプおよびコンセプトの間のテキスト リンク分析の関係性パターンを抽出する機能が用意されてます。特定の意見（製品の反応など）を検出しようとする場合、特に役立ちます。

二次抽出の手順

日本語テキストで抽出を実行する場合、基本キーワードおよび人名、地名、組織名、名詞、形容詞、動詞、形容動詞、およびその他の 8 つの基本タイプからコンセプトを自動的に取得します。ただし、日本語テキスト向けに用意されたデフォルトのリソースを活用するために、以下の二次分析機能のいずれかを選択必要があります：**[感性分析]**または**[係り受け分析]**。

二次分析を選択すると、テキスト リンク分析パターンを抽出して、テキスト内のキーワード間の関係性を明らかにすることができます。インタラクティブ ワークベンチ セッションでノードを定義または抽出オプションを選択すると、抽出プロセスに二次分析機能を追加することができます。

注:日本語テキスト展開はIBM® SPSS® Modeler Premiumで利用可能です。

二次分析: 抽出が開始したとき、基本キーワード抽出が、タイプのデフォルト セットを使用して行われます。詳細は、 p.412 日本語テキストで使用できるタイプ を参照してください。 ただし、二次分析機能を選択すると、抽出機能にコンセプトの一部として助詞や助動詞が含まれているため、より多くの詳細なコンセプトを取得することができます。たとえば、「肩の荷が下りた」という文があるとします。この例では基本キーワード抽出は各コンセプトを個別に抽出します。例：**肩**（肩）、**荷**（重量）、**下りる**（上げ）、しかしこれらの単語の間の関係性は抽出されません。しかし、感性分析を適用すると、コンセプト **=肩の荷が下りた** というより高次の意味をもつコンセプトが抽出され、<**良い-安心**> という感性タイプとして割り当てられます。なお感性分析については、このほかに多くの感性タイプがあります。さらに、二次分析機能を選択すると、テキストリンク分析結果も生成できます。

注:二次分析を呼び出すと、抽出プロセスにより時間がかかります。詳細は、 p.404 二次抽出の手順 を参照してください。

- **係り受け解析**:このオプションを選択すると、キーワード、およびキーワードに助詞等を加えた語を、基本の品詞タイプのコンセプトとして抽出します。また係り受けテキスト リンク分析（TLA）に基づいたパターン結果を出力します。
- **感性分析**:この分析機能を選択すると、追加の抽出コンセプト、および可能な場合は、TLA パターン結果の抽出が行われます。基本タイプのほか、嬉しい、吉報、幸運、安心、幸福など 80 を超える感性タイプも利用できます。これらのタイプを使用して、感情、感性、意見の表現のテキストでコンセプトおよびパターンを検出します。感性分析に対するフォーカスを指示するオプションは３つあります：**全ての感性**、**代表的な感性のみ**、それと**結論のみ**。

感性分析のオプション

日本語テキストを処理するとき、感性分析機能を使用して、追加のコンセプトおよびタイプを抽出できます。この分析機能には、意見、間隔、感性をテキスト データから抽出できる、80 を超える追加タイプが含まれます。また、二次分析として **感性分析** を選択した場合、抽出エンジンに抽出する感性を指示する次のオプションからいずれかを選択する必要があります。

- **すべての感性**
- **代表的感性のみ**
- **結論のみ**

抽出時、感性分析機能はレコードまたはドキュメントを、それぞれに述語が含まれる節に分割します。たとえば、「4月になったがまだ寒い。」というテキストは、文末記号「。」を 1 つだけ使用しているにもかかわらず、分析機能によって 2 つの節があると解釈されます。各節が抽出エンジンによって検証され、選択したサブオプションに適しているかどうかを確認します。

次のサンプル テキストを使用して、3 つのオプションを検証してみましょう。”案内してくれた仲居さんは無愛想だったが、部屋は広くて申し分なかった。夕食も満足。「このテキストは次のように翻訳されます」：A serving lady was not friendly, but the room was large and quite satisfactory.I satisfied with the dinner, too.抽出時、元のテキストは次の節に分割されます。

- **案内してくれた仲居さんは無愛想だったが、**
- **部屋は広くて申し分なかった。**
- **夕食も満足。**

すべての感性

このオプションを指定すると、リソースおよび感性テキスト リンク規則に一致するすべての感性、意見および感情が抽出されます。この例では、サンプル テキストから次のコンセプトが抽出されます。

テーブル A-1
[すべての感性] オプションを使用した場合の出力の例

Concept	タイプ
仲居さんは無愛想だった	〈悪い－対応が不親切〉
部屋は広くて	〈良い－満足〉
申し分なかった	〈良い－満足〉
満足	〈良い－満足〉

注:前述の表の 2 番目および 3 番目の行で、どのようにして同じ節から 2 つのコンセプトを取得するかを示しています。

代表的感性のみ

このオプションを指定すると、各節で表現されているより代表的な意見または感性のみを抽出します。テキストに複数の意見または感性がある場合、アルゴリズムが適用されます。このアルゴリズムは、節で見つかった感性および単語の位置の重要度を判断しようとします。重要度が同じ感性キーワードが 2 つ見つかった場合、節の最初の感性キーワードではなく後の方の場所にある感性キーワードが抽出されます。

内部アルゴリズムおよび単語の位置が適用された場合、この節の 2 番目の申し分なかったより重要であると認識されるため、部屋は広くては適用されません。

テーブル A-2
[代表的感性のみ] オプションを使用した場合のテキストの出力の例

Concept	タイプ
仲居さんは無愛想だった	〈悪い－対応が不親切〉
申し分なかった	〈満足〉
満足	〈満足〉

結論のみ

このオプションを指定すると、感性キーワードをレコードまたはドキュメント全体の結論を示すものとして特定および抽出されるよう強制します。すべてのテキストに結論があるわけではなく、このオプションを指定してもテキストの指定した部分から何も抽出されない場合もあります。また、レコードまたはドキュメントが長いほど、分析機能で主な結

論を特定することが困難になります。めったにありませんが、複数の結論が抽出される場合があります。

満足 の場合、これが表現された感性の重要な結論であると見なされます。

テーブル A-3
[結論のみ] オプションを使用した場合のテキストの出力の例

Concept	タイプ
満足	〈満足〉

カテゴリ化の方法

IBM® SPSS® Modeler Text Analytics でカテゴリモデルを作成する場合、いくつかの手法から選択して、カテゴリを作成できます。すべてのデータセットが一意であるため、手法の数やそれらを適用する順序は、変わる場合があります。結果の解釈が、他の人とは異なる場合があるため、テキスト データにとってどの手法が最良の結果を生み出すか、それぞれの手法を検証する必要があります。SPSS Modeler Text Analytics では、カテゴリをさらに検証し、調整できるインタラクティブ ワークベンチ セッションでカテゴリ モデルを作成できます。

このガイドの場合、**カテゴリの作成**は、カテゴリ定義の生成および、1 つまたは複数のビルトインの手法を使用した分類を指し、また**カテゴリ化**は、スコアリング、またはラベル付け、一意の識別子（名前/ID/値）を各レコードまたはドキュメントのカテゴリ定義に割り当てるプロセスのことを指します。

カテゴリ作成時、抽出されたコンセプトおよびタイプはカテゴリの構築ブロックとして使用されます。カテゴリを作成すると、カテゴリ定義の要素に一致するテキストが含まれる場合、レコードおよびドキュメントがカテゴリに割り当てられます。

SPSS Modeler Text Analytics には、自動カテゴリ作成手法がいくつか用意されており、ドキュメントまたはレコードを迅速にカテゴリ化することができます。使用できるそれぞれの手法は、特定の種類のデータおよび状況に適していますが、同じ分析で手法を組み合わせて、全範囲のドキュメントまたはレコードをキャプチャすると役に立つ場合があります。複数のカテゴリのコンセプトを表示したり、重複するカテゴリを見つけることができます。

日本語テキストのリソースの編集

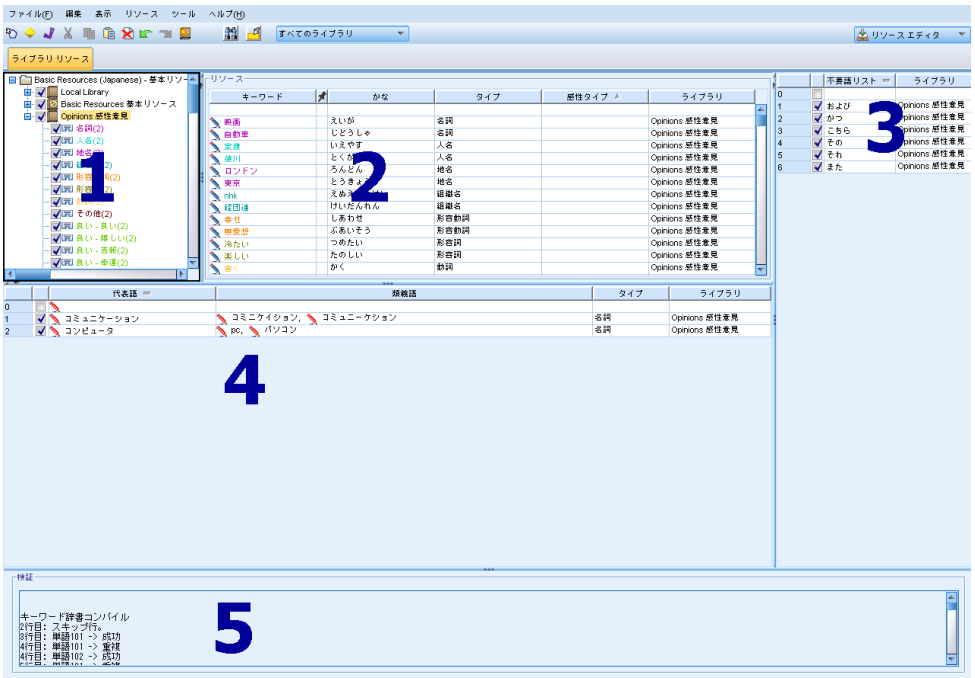
注:日本語テキスト展開はIBM® SPSS® Modeler Premiumで利用可能です。

IBM® SPSS® Modeler Text Analytics バージョン 14以降では、新しいテンプレートと、新しいテキスト分析パッケージ（TAP）が日本語で使用できます。キーワードを追加および編集してリソースに変更を加え、データに合

わせてカスタマイズすることができます。テキスト分析パッケージには、肯定的な感性、否定的な感性、コンテクスト/一般的な感性を示すカテゴリで構成されたカテゴリ セットも含まれています。

リソース エディタ および テンプレート エディタ でリソースの作業ができます。エディタはすべてのテキスト言語で同様の作業をしますが、日本語テキストでの作業については、次のような重要な違いがあります。

図 A-1
日本語テキストの リソース エディタ ウィンドウ



日本語テキストのリソースを編集する場合の主な相違点は次のとおりです。[ライブラリ リソース] タブの 4 つのメイン パネルの詳細は、エディタのインターフェイス p. 299 を参照してください。

1. [ライブラリ] パネル: 左上のこの領域は、他の言語と同じように機能します。ただし、新しいタイプを作成できない、タイプの名前を変更できないなど、いくつか異なる点があります。 詳細は、 16 章 p. 316 ライブラリの使用 を参照してください。

2. キーワード辞書の [キーワード] パネル ライブラリ ツリーの右側にあるこのパネルは、日本語テキストの場合、大きく異なります。キーワード名があるほか、かな名を追加でき、またキーワードに関連付けることができる 1 つまたは 2 つのタイプを選択できます。ただし、日本語以外の言語でできるように、キーワードの活用形あるいは日本語キーワードのマッチ オプションを割り当てることはできません。 詳細は、 p. 409 日本語のライブラリ ツリー、タイプ、キーワードのパネル を参照してください。

3. [類義語辞書] パネル: 日本語テキスト リソースには、リソースのすべての類義語を定義できる［類義語］タブがあります。［類義語］タブにもう 1 つ［タイプ］列があり、そこで入力した類義語のタイプを指定する必要があります。詳細は、 p. 417 日本語テキストの類義語辞書の使用 を参照してください。注:また、［オプションの要素］タブは日本語テキストの言語リソースには適用されないため、使用できません。

4. [不要語辞書] パネル: ワイルドカードの使用がサポートされていない点を除き、日本語テキスト リソースのこのパネルに異なる点はありません。

5. [検証] パネル: 日本語テキストの場合、抽出前にリソースを確認するための［検証］パネルがあります。日本語テキストから抽出する場合、抽出プロセスが始まる前に変更が検出されると、抽出エンジンは自動的にリソースを再コンパイルします。実行時のエラーを回避するため、抽出前にリソースを再コンパイルして検証し、発生したエラーを修正できます。 詳細は、 p. 418 日本語リソースの検証およびコンパイル を参照してください。

注: 日本語テキスト向けに編集可能なアドバンス リソースまたはテキスト リンク規則がないため、これらのタブは使用できません。

日本語のライブラリ ツリー、タイプ、キーワードのパネル

日本語辞書のライブラリおよびタイプの使用方法は、他の言語の場合と非常に似ています。 詳細は、 17 章 p. 332 キーワード辞書 を参照してください。

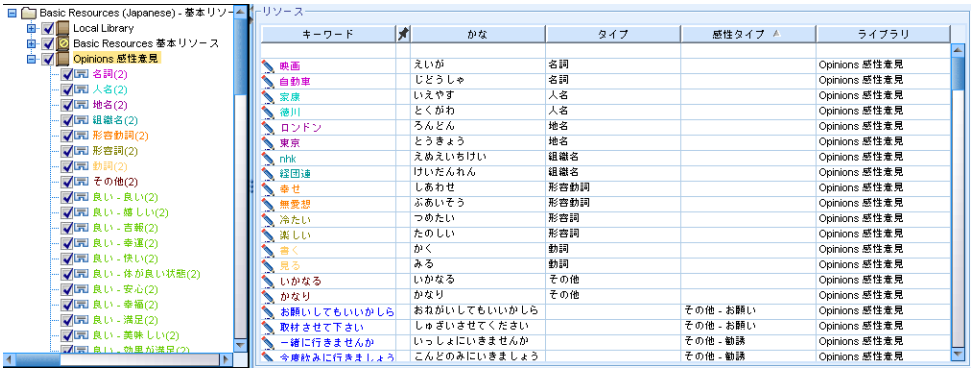
ただし、次のように異なる点がいくつかあります。

- 日本語テキスト リソースにタイプの異なるセットがある。 詳細は、 p. 412 日本語テキストで使用できるタイプ を参照してください。
- タイプの作成、または名前の変更はできないが、プロパティは編集できる。 詳細は、 p. 416 日本語のタイプのプロパティの編集 を参照してください。
- キーワードのかな名の指定、1 つのタイプまたは 2 番目の感性タイプへの割り当てなど、キーワードの追加および編集ができる。 詳細は、 p. 409 日本語のライブラリ ツリー、タイプ、キーワードのパネル を参照してください。

ライブラリ ツリー パネルには、キーワード辞書のほか、ライブラリが表示されます。左側のライブラリまたはタイプを選択すると、右側のキーワード パネルに、選択したライブラリまたはキーワード辞書のキーワードが表示されます。［キーワード］パネルで直接または［**キーワードを追加**］ダイアログ ボックスを使用して、キーワードをキーワード辞書に追加できます。追加するキーワードは、単語でも複合語でもかまいません。リストの一番上に空白の行があり、そこに新しいキーワードを追加できます。

キーワード辞書のキーワードを定義すると、デフォルトでは名詞と見なされ、自動的にタイプ <**名詞**> に割り当てられます。ただし、タイプを <**動詞**>、<**形容詞**>、<**地名**> など別の基本タイプに変更できます。抽出エンジンが、このキーワードが [**タイプ**] 列で割り当てられているタイプと同じ品詞として機能していることを認識した場合、キーワードはそのタイプに割り当てられ、抽出されます。また、キーワードを [**感性タイプ**] 列のいずれかの感性タイプに割り当てることができます。感性二次分析を使用すると、テキストには 2 回目の処理が行われ、キーワードを検索して感性タイプに割り当てようとします。さらに、感性タイプおよび基本タイプの 2 つを定義し、二次感性分析時に抽出エンジンでこのキーワードがいずれのタイプにも合致することが認識された場合、感性タイプが優先され、抽出結果パネルおよびテキスト リンク分析結果に表示されます。たとえば、動詞が <**動詞**> タイプとして抽出され、「愛されている」のような肯定的な種類のタイプとしても抽出された場合、感性のキャプチャが単なる品詞よりより重要である場合が多いため、このキーワードはインターフェイス上では肯定的なタイプに割り当てられているものとして表示されます。

図 A-2
日本語リソースの [ライブラリ] パネルおよび [キーワード] パネル



テーブル A-4
[キーワード] パネルの列の説明

列名	列の説明
キーワード	単語または複合語を入力します。キーワードが表示される色は、キーワードが保存または強制投入されるタイプの色によって異なります。[タイプのプロパティ] ダイアログ ボックスでタイプの色を変更できます。詳細は、 p. 416 日本語のタイプのプロパティの編集 を参照してください。 通常、キーワードは漢字で書かれますが、かなを組み合わせる場合もあります。**重要!** カタカナを使用した動詞の入力はサポートされていません。
強制	このセルに押しピンのアイコンをクリックして投入すると、抽出エンジンは、他のライブラリのこの同じキーワードの他の出現を無視します。詳細は、 17 章 p. 341 キーワードの強制 を参照してください。 すべての言語で同じように機能します。
かな	漢字のキーワード名の読みがなを入力します。

列名	列の説明
タイプ	キーワードを割り当てる基本タイプ名を選択します。 詳細は、 p. 412 日本語テキストで使用できるタイプ を参照してください。
感性タイプ	2 番目の分析が実行されると、キーワードを割り当てる感性タイプ名を選択します。 詳細は、 p. 412 日本語テキストで使用できるタイプ を参照してください。
ライブラリ	キーワードが格納されているライブラリを選択します。ライブラリ ツリー パネルでキーワードを別のタイプにドラッグ アンド ドロップして、そのライブラリを変更できます。

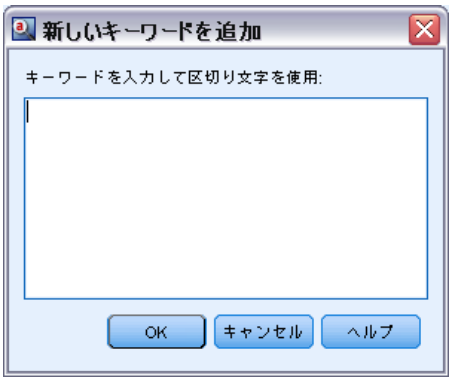
キーワード辞書に 1 つのキーワードを追加するには

- ライブラリ ツリー パネルで、キーワードを追加したいキーワード辞書を選択します。
- 中央のパネルのキーワード リストで、使用できる最初の空白セルにキーワードを入力し、このキーワードに必要なオプションを設定します。

キーワード辞書に複数のキーワードを追加するには

- ライブラリ ツリー パネルで、キーワードを追加したいキーワード辞書を選択します。
- メニューの [ツール] → [新規キーワード] を選択します。[新しいキーワードを追加] ダイアログ ボックスが開きます。

図 A-3
[新しいキーワードを追加] ダイアログ ボックス



- キーワードを入力するか、キーワードのセットを貼り付けて、選択したキーワード辞書に追加したいキーワードを入力します。複数のキーワードを入力する場合、[オプション] ダイアログで定義された区切り文字を使用してキーワードを区切るか、各キーワードを新しい行に追加するつようがあります。 詳細は、 8 章 p. 143 オプションの設定 を参照してください。

▶ [OK] をクリックすると、キーワードが辞書に追加されます。ダイアログボックスが閉じ、辞書に新しいキーワードが表示されます。

日本語テキストで使用できるタイプ

日本語リソースに新しいタイプを追加することはできませんが、日本語リソースからキーワードを追加および削除することができます。次の表に、現在使用できるタイプの一覧を示します。

注:日本語テキスト展開はIBM® SPSS® Modeler Premiumで利用可能です。

基本抽出のタイプ

抽出が開始されると、次のタイプが使用されます。

テーブル A-5
基本抽出のタイプ

タイプ	説明
名詞	「車」や「映画」など、物を示す単語人名、地名、組織名は別にカテゴリ化されます。
人名	「徳川」や「家康」など、特定の人物の名前に該当する名詞。「徳川家康」のように、名と姓を組み合わせても人名となります。
地名	「東京」やロンドンなど、特定の場所を示す名詞。
組織名	「経団連」など、特定の会社や組織を示す名詞。
形容動詞	物事の特徴または状況を説明する「静か」のような単語。「静かでない」や「静かなこと」のように句として使用することができます。
形容詞	物事の特徴または状況を説明する「楽しい」のような単語。「楽しくなる」や「楽しいこと」のように句として使用することができます。
動詞	動きまたは動作を説明する、タイプ I（子音語幹）動詞、タイプ II（母音語幹）動詞、不規則（サ行変格活用、カ行変格活用）動詞などの単語。
その他	「非常に」、「いかなる」、「そして」、「ありがとう」など、助動詞、連体詞、接続詞、感嘆詞などの語。

感性分析のタイプ

感性抽出に二次分析機能を選択した場合、8 つの基本タイプのほか、多くのタイプを取得します。

テーブル A-6
感性分析のタイプ

タイプ	説明
良い - 良い	「良い」と分類できる、一般的に肯定的なものの表現。
良い - 嬉しい	喜びの感動を生み出す望ましいイベントを説明。
良い - 吉報	相当な努力によってのみ可能な喜びのイベントを説明。
良い - 幸運	たまたま、またはまったくの偶然によってのみ可能な喜びのイベントを説明。
良い - 快い	喜びの心理的な感覚に切り替える状況または環境であることを示す表現。
良い - 体が良い状態	病気、怪我、疲労のない体の状態、または身体的な状況が改善している状態を説明。
良い - 安心	落ち着いて、傷ついたり被害を受けるリスクがないことを説明。
良い - 幸福	特に好ましい状況、またはある人の動作から愛情を、または誕生の状況にあることを説明。
良い - 満足	心が落ち着く望ましいイベントを説明。
良い - 美味しい	食べ物の味が良いことを示す。
良い - 効果が満足	あるものが、期待された効果をあげたことを示す。
良い - 感動	何かの重要性、意味または価値が特別に良いことを示す。
良い - 感謝	相手の動作を肯定的に認識することを示す。
良い - 祝福	ある人物の状況が望ましいという（話者にとってある程度受け入れられる）気持ちを表現。
良い - 喜び全般	話者にとってあまり関連がない肯定的なイベント。
良い - 楽しい	交友、娯楽、レクリエーションなどの活動を示すまたは予期する。
良い - 可笑しい	楽しい刺激を与えるユーモラスな性質があることを示す。
良い - 笑い	良いまたはユーモラスなものによって起こる笑いを示す。
良い - 期待	良いイベントが将来起こることを示す。
良い - 楽しみ全般	話者にとってあまり関連がない楽しいイベントまたは肯定的な活動/動作。
良い - 金額への賞賛	購入者の立場から、望ましい金銭的価値があることを示す。
良い - 対応が早い	サービスがタイムリーに提供または完了することを示す。
良い - 対応が親切	サービス提供者の態度または動作に配慮が行き届いていることを示す。
良い - 説明が良い	情報の種類/または品質、およびその提供の方法が適切であるということを示す。
良い - 対応への賞賛	サービスの提供者を褒める、上記以外の考え。
良い - 褒め・賞賛	あるものの性質、機能、動作を褒める、上記以外の考え。

タイプ	説明
良い – 好き	あるものを所有したいまたは近くにいたいという欲求を示す。
良い – 入会希望	あるグループの一員になりたいまたは一員のままでいたいという欲求を示す。
良い – 買いたい	あるものを得るためにお金を使いたいまたは使う予定があることを示す。
良い – 好評・人気	あるものを必要としているまたは褒めている人の数が目標を超えていることを示す。
良い – 売れた	あるものを購入した人々の有無または購入数または金額が目標を超えていることを示す。
悪い – 悪い	「悪い」と分類できる、一般的に否定的なものの表現。
悪い – 怒り	予定通りに物事が起こらなかった場合に感じる怒りの感情。
悪い – 批判	他者が適切な選択をしなかったことに対する考えを表す。
悪い – お叱り	他者をある人物の意図に従うよう強制させる言葉または動作。
悪い – 誹謗・中傷	他者のあまりに低い評価を示すために使用する言葉。
悪い – 軽蔑	他者の性格、能力、または性質が著しく書けているということを示す。
悪い – 恨み	他者によってもたらされる不利な点に対する報復または憤りを示す。
悪い – 嫌がらせ	コミュニケーションを抑制する目的で使用する言葉。
悪い – 不満	希望する者または状態を取得できなかったことによる、不快な感情。
悪い – 不味い	食べ物の味が悪いことを示す。
悪い – 効果が不満	あるものが、期待された効果をあげていないことを示す。
悪い – 金額が不満	購入者の立場から、あるものの金銭的価値が望ましいものでないことを示す。
悪い – 対応への不満	サービス提供者に過失があることを示す。
悪い – 対応が遅い	サービスが不適切な時に行われたこと、またはサービスがまだ行われていないことを示す。
悪い – 対応が不親切	サービス提供者の態度または動作によってもたらされる不快な感情を示す。
悪い – 説明が悪い	情報の種類/または品質、およびその提供の方法が不適切であるということを示す。
悪い – 返答なし	状況が満足いく場合であっても、サービス提供者が適切な回答を提供しなかったことを示す。
悪い – 不快	否定的な心理的感情に切り替える状況または環境であることを示す。
悪い – 怒り全般	上記以外の怒りの感情。話者の組織または会社によってもたらされる一般的な怒り、または言及された怒りによってもたらされた出来事の説明。
悪い – 悲しい	あるものを失ったまたは得られなかった場合の不快な感情。

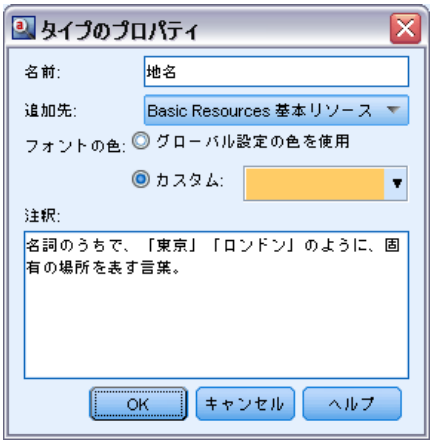
タイプ	説明
悪い – 凶報	十分な努力にかかわらず、ある目的を達成できなかったことを示す。
悪い – 不運	自分の過失ではなく、不幸な偶然または運によってもたらされた否定的な結果を示す。
悪い – ショック	予測できない、否定的なものまたは発生によってショックを受けたこと、または適切な回答が見つからないことを示す。
悪い – 残念	期待されることがそのようにできなかった場合にもたらされる悲しい感情。
悪い – 落胆	悲しく落胆した感情によって支配されている状態。
悪い – 諦め	話者または他者によってもたらされた否定的な状態が改善できないことを示す。
悪い – 後悔	代わりのものがあっても、過去に適切な選択をできなかったことに対する考えを示す。
悪い – 謝罪	他者を機ずつ行けたことに対する話者の認識を示す。
悪い – 淋しい	他者との関係が十分でないまたは、関係する人の数が少ないことを示す。
悪い – 哀れみ	他者の状況が話者よりとても悪いことに対する気持ちを示す。
悪い – 悩み	選択をする必要があるが、選択肢から選ぶことができないことを示す。
悪い – 困っている	行動が必要な状況に対応する効果的な方法がないことに対する気持ちを表す。
悪い – 苦しい	外的な要因または自分の誤りまたは間違いによって正常に行動できない、不快な心理的状況を示す。
悪い – 体が悪い状態	病気、怪我、疲労のある体の状態、または身体的な状況が改善していない状態を説明。
悪い – 不安	あることが望ましい状態で継続できないまたは期待通りにいかないということに対する感情を示す。
悪い – 恐怖	あるものによって、傷つく恐れがある状態を示す。
悪い – 悲しみ全般	指定されていない者に対する一般的な悲しみなど、上記以外の悲しみの感情。
悪い – 嫌い	あるものを遠ざけたいまたは、離れたいということを示す。
悪い – 退会希望	あるグループの一員であることを辞めたいまたは参加したくないという欲求を示す。
悪い – 買いたくない	あるものが欲しくないまたは言及したものに対して支払う予定がないことを示す。
悪い – 不評・不人気	あるものを好きな人々の数がある目標に達しなかったことまたは、そのものに対する否定的な感情を持つ人々が多くいることを示す。
悪い – 売れていない	あるものを購入した人々がいないまたは購入数または金額が目標に達していないことを示す。
その他 – 疑問	他人のより詳細な検討または検証が必要な情報を必要としていることを示す。
その他 – 問い合わせ	他人が持っている情報を要求していることを示す。

タイプ	説明
その他 – 要望	(他人に直接責任がある、または話者より順位が低い場合) 他者に問題を解決するよう命じることを示す。
その他 – 提案・忠告	(他人に直接責任がある、または話者より順位が低い場合) より良い行動を行うよう命じることを示す。
その他 – お願い	(他人に直接責任がない、または話者より順位が低い場合) 他者に何か実行するよう要望する表現。
その他 – 激励	他者を勇気づけることを示す表現、または動作を促す説明。
その他 – 勧誘	他者に、話者と一緒に何かを行うことを要求する表現。
その他 – 驚き	イベントが突然であることまたは尺度が、論理的な判断/理解を超えていることを示す表現。
評価なし – 評価なし	評価に値する表現がない。

日本語のタイプのプロパティの編集

日本語リソースでタイプを作成することはできませんが、タイプのプロパティを編集して表示することができます。また、マッチ オプションや活用形などのオプションは日本語テキストに適用されません。

図 A-4
日本語テキスト リソースの [タイプのプロパティ] ダイアログ ボックス



名前: キーワード辞書の名前。

追加先: 新しいキーワード辞書を作成するライブラリを指定します。

フォントの色: このフィールドを指定すると、インターフェイスでこのタイプの結果と他のタイプの結果とを区別できるようになります。[**グローバル設定の色を使用**] を選択すると、タイプのデフォルト色がこのキーワードに使用されます。このデフォルト色は、[オプション] ダイアログ ボックスで設定されます。詳細は、 8 章 p. 145 オプション : [表示] タブ を

参照してください。[カスタム] を選択すると、ドロップダウン リストから色を選択できます。

注釈: このフィールドはオプションで、任意のコメントまたは説明に使用できます。

タイプのプロパティを表示または編集するには

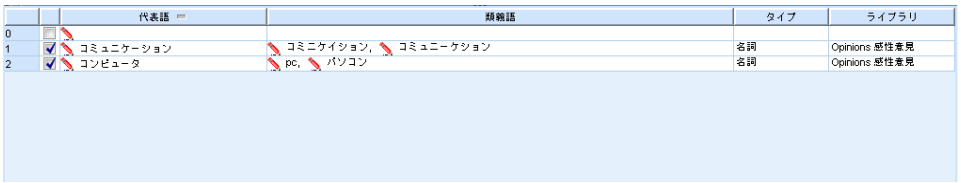
- プロパティを表示したいタイプを選択します。
- マウスを右クリックし、コンテキスト メニューから [タイプのプロパティ] をクリックします。[タイプのプロパティ] ダイアログ ボックスが開きます。
- 必要に応じて変更してください。
- [OK] をクリックすると、キーワード辞書への変更が保存されます。

日本語テキストの類義語辞書の使用

日本語テキストの場合、類義語辞書には、類義語を管理するタブ、[類義語] タブが 1 つだけあります。類義語とは、同じ意味を持つ複数の語を関連付けたものです。また、類義語はキーワードを略語とグループ化したり、一般的にスペルミスのある単語と正しいスペルの単語とをグループ化したりするために使用できます。

注:日本語テキスト展開はIBM® SPSS® Modeler Premiumで利用可能です。

図 A-5
日本語テキストの類義語エントリ



類義語定義はは、2 つの部分で構成されています。代表語は、抽出エンジンがすべての類義語キーワードをグループ化する基準となるキーワードです。この代表語が別の代表語の類義語として使用されていない限り、または不要語として登録されていない限り、[抽出結果] パネルに表示されるコンセプトとなります。類義語のリストは、代表語の下にグループ化されるキーワードです。

[類義語] タブで、テーブルの一番上の空白行に類義語定義を入力できます。まず代表語とその類義語を定義します。この定義を格納するライブラリも選択できます。抽出時、類義語のすべての出現を、最終的な抽出で代表語に基づいてグループ化します。 詳細は、 17 章 p. 337 キーワードを追加 を参照してください。

キーワード辞書を作成している場合、キーワードを入力し、そのキーワードに対して 3 つまたは 4 つの類義語が考えられます。この場合、類義語辞書にすべてのキーワードと代表語を入力し、類義語をドラッグすることができます。

重要! ワイルドカードおよび特殊文字は、日本語テキストの類義語ではサポートされていません。

類義語エントリを追加するには

- ［類義語］パネルの［類義語］タブに表示されたテーブルの一番上の空白行で、［代表語］列に代表語を入力します。入力した代表語が色つきで表示されます。この色は、キーワードが表示されるまたは強制されるタイプを示します。キーワードが黒で表示されている場合、キーワード辞書にはこのキーワードがないことを意味します。

- 代表語の右側の 2 番目のセルをクリックして、類義語のセットを入力します。［オプション］ダイアログ ボックスで定義したグローバル区切り文字を使用して、各エントリを区切ります。入力されたすべての類義語は、同じタイプでなければなりません。詳細は、 8 章 p. 143 オプションの設定 を参照してください。 入力したキーワードが色つきで表示されます。この色は、キーワードが出現するタイプを示します。キーワードが黒で表示されている場合、キーワード辞書にはこのキーワードがないことを意味します。

- 3 番目の列、［タイプ］列で、これらの類義語のタイプを指定します。ただし、代表語は、抽出時にタイプが割り当てられます。代表語がコンセプトとして抽出されない場合、抽出結果でこの列に表示されたタイプが代表語に割り当てられます。

- 最後のセルをクリックして、この類義語定義を格納するライブラリを選択します。

  注:ここでは、リソース エディタ ビューまたは テンプレート エディタでどのように変更を行うかについて説明します。［抽出結果］パネル、［データ］パネル、［カテゴリ］パネル、または他のビューの［クラスタ定義］ダイアログ ボックスで直接、この種類の調整を行うこともできます。 詳細は、 9 章 p. 168 抽出結果の調整 を参照してください。

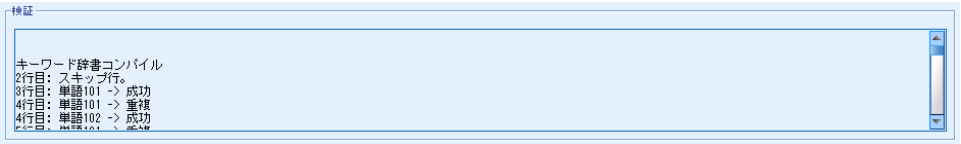
日本語リソースの検証およびコンパイル

日本語テキストの場合、抽出前にリソースを確認するための［検証］パネルがあります。日本語テキストの抽出プロセスが開始される前に変更が検出されると、抽出エンジンは自動的にリソースを再コンパイルします。抽出時にエラーが見つかると、処理を適切に完了できない場合があります。

コンパイル エラーを回避するために、 リソース エディタ または テンプレート エディタ で変更を行った後、リソースを検証してからコンパイルすることをお勧めします。エラー メッセージが表示された場合、修正して検証をもう一度行うことができます。

注:日本語テキスト展開はIBM® SPSS® Modeler Premiumで利用可能です。

図 A-6
日本語テキストの [検証] パネル



リソースを検証するには

- メニューの [**ツール**] → [**リソースを検証**] を選択します。[検証] パネルが開き、コンパイルおよびエラーのメッセージが表示されます。

日本語についてのその他の例外

ユーザー定義リソースを上書きする内部リソース

日本語テキストの場合、デフォルトのリソースには、コンパイル済みの内部基本リソースがいくつか含まれています。これらの内部リソースは編集できません。このため、リソース エディタ または テンプレート エディタ を使用して、変更または調整を行うことができます。ほぼすべての場合、リソースで定義したキーワード、類義語および不要語リストのエントリは、コンパイル済み内部リソースより優先されます。ただし、次の例の一部で示されているように、例外がいくつかあります。

- このようなキーワードを追加しても特定のタイプが抽出結果に影響を与えない場合があります。このようなケースは、データにいくつかの形態的要素、区切り記号、または記号を使用している長い文が含まれる場合に最も発生します。また、日本語テキスト リソースには、事前にコンパイルされている多くの一般的キーワードが含まれているため、特定の言語学的定義に常に強制投入される共通単語がいくつかあります。
- 抽出エンジンは、**ある**、**いる**、または**なる**の抽出を常に強制するため、これらのキーワードを除外することができない場合があります。

- キーワード**東京**のタイプを <**地名**> から <**名詞**> に変更することはできますが、キーワードのタイプをキーワード辞書を使用して <**地名**> から <**動詞**> または <**形容詞**> に変更しようとしても、抽出エンジンに無視されます。
- リソース エディタ または テンプレート エディタ で行った変更がある文からの抽出結果に影響を与える場合がありますが、抽出プロセスの最後に各文の共起語を参照するため、別の文には影響を与えません。

半角カタカナの表示問題

半角カタカナの文字は、抽出時、全角カタカナに内部的に変換されますが、インタラクティブ ワークベンチ セッションの［データ］パネルで表示される場合、元の半角カタカナ文字で表示されます（テキスト マイニング ノードのみ)。半角カタカナ文字は［データ］パネルで強調表示できません。この問題を回避するには、処理前にレコード全体を全角カタカナに変換します。

大文字および小文字の使用

アプリケーションに読み込む場合、アルファベットの大文字は小文字に一時的に変換されます。ただし、［データ］パネルには、元のテキストと同じ文字でテキストが表示されます。小文字および大文字は、この製品では同じように扱われます。

付録

B

注意

本情報は全世界で提供する製品およびサービスについて作成したものです。

本書に記載の製品、サービス、または機能がその他の国おいては提供されていない場合があります。お住まいの国で利用可能な製品、サービス、および機能については、現地 IBM の営業担当員にお尋ねください。本書で IBM 製品、プログラム、またはサービスに言及していても、その IBM 製品、プログラム、またはサービスのみが使用可能であることを意味するものではありません。これらに代えて、IBM の知的所有権を侵害することのない、機能的に同等の製品、プログラム、またはサービスを使用することができます。ただし、IBM 以外の製品とプログラムの操作またはサービスの評価および検証は、お客様の責任で行っていただきます。

IBM は、本書に記載されている内容に関して特許権（特許出願中のものを含む）を保有している場合があります。本書の提供は、お客様にこれらの特許権について実施権を許諾することを意味するものではありません。実施権についてのお問い合わせは、書面にて下記宛先にお送りください。

IBM Director of LicensingIBM CorporationNorth Castle DriveArmonk, NY 10504-1785U.S.A.

For license inquiries regarding double-byte character set (DBCS) information, contact the IBM Intellectual Property Department in your country or send inquiries, in writing, to:

Intellectual Property LicensingLegal and Intellectual Property LawIBM Japan Ltd.1623-14, Shimotsuruma, Yamato-shiKanagawa 242-8502 Japan

The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law: INTERNATIONAL BUSINESS MACHINES PROVIDES THIS PUBLICATION “AS IS” WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.Some states do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors.Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication.IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites.The materials at those Web sites are not part of the materials for this IBM product and use of those Web sites is at your own risk.

IBM は、お客様が提供するいかなる情報も、お客様に対してなんら義務も負うことのない、自ら適切と信ずる方法で、使用もしくは配布することができるものとします。

本プログラムのライセンス保持者で、、(i) 独自に作成したプログラムとその他のプログラム (本プログラムを含む) との間での情報交換、および (ii) 交換された情報の相互利用を可能にすることを目的として、本プログラムに関する情報を必要とする方は、下記に連絡してください。

IBM ソフトウェア グループAttention:Licensing233 S. Wacker Dr.Chicago, IL 60606USA

本プログラムに関する上記の情報は、適切な使用条件の下で使用することができますが、有償の場合もあります。

本書で説明されているライセンス・プログラムまたはその他のライセンス資料は、 IBM 所定のプログラム契約の契約条項、IBM プログラムのご使用条件、またはそれと同等の条項に基づいて、IBM より提供されます。

本書に記述された性能は、管理された環境下で判定されたものです。従って、異なる使用環境で取得された結果は大きく異なる可能性があります。一部の測定は開発段階のシステムで行われた場合もあり、これらの測定結果が一般的に入手可能なシステムと同じであることは保証されません。また、一部の測定値は外挿法により推定されています。実際の結果は異なる場合もあります。本書のユーザーは個々の環境で入手したデータを検証する必要があります。

IBM 以外の製品に関する情報は、その製品の供給者、出版物、もしくはその他の公に利用可能なソースから入手したものです。IBM は、それらの製品のテストは行っておりません。したがって、他社製品に関する実行性、互換性、またはその他の要求については確証できません。 IBM 以外の製品の性能に関する質問は、それらの製品の供給者にお願いします。

This information contains examples of data and reports used in daily business operations.To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products.All of these names are fictitious and any similarity to the names and addresses used by an actual business enterprise is entirely coincidental.

電子的に複製されたこの情報を表示する場合には、写真やカラーのイラストは表示されない場合があります。

商標

IBM, the IBM logo, ibm.com, and SPSS are trademarks of IBM Corporation, registered in many jurisdictions worldwide. A current list of IBM trademarks is available on the Web at http://www.ibm.com/legal/copytrade.hmtl.

Adobe、Adobe のロゴ、PostScript、および、PostScript のロゴは、Adobe Systems Incorporated のアメリカ合衆国、その他の国家、または両方における登録商標、または、商標です。

Intel、Intel ロゴ、Intel Inside、Intel Inside ロゴ、Intel Centrino、Intel Centrino ロゴ、Celeron、Intel Xeon、Intel SpeedStep、Itanium、Pentium は 米国およびその他の国における Intel Corporation または、その子会社の商標または登録商標です。

Linux は、Linus Torvalds 氏のアメリカ合衆国およびその他の国における登録商標です。

Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

UNIX is a registered trademark of The Open Group in the United States and other countries.

Java and all Java-based trademarks and logos are trademarks of Sun Microsystems, Inc. in the United States, other countries, or both.

他の製品またはサービスの名称は、IBM または、他の会社の商標である場合があります。

索引