IBM SPSS Modeler Text Analytics 18.1 — podręcznik użytkownika

Uwaga

Przed skorzystaniem z niniejszych informacji oraz produktu, którego one dotyczą, należy zapoznać się z informacjami zamieszczonymi w sekcji “Uwagi” na stronie 221.

Informacje o produkcie

Niniejsze wydanie publikacji dotyczy wersji 18.1, wydania 0, modyfikacji 0 produktu IBM SPSS Modeler Text Analytics oraz wszystkich następnych wersji i modyfikacji do czasu, aż w kolejnym wydaniu publikacji zostanie zawarta informacja o stosownej zmianie.

Spis treści

Przedmowa

Produkt IBM® SPSS Modeler Text Analytics oferuje wydajne możliwości analiz tekstu, które wykorzystują zaawansowane rozwiązania lingwistyczne oraz przetwarzanie języka naturalnego (NLP, Natural Language Processing) w celu szybkiego przetwarzania zróżnicowanych nieustrukturyzowanych danych tekstowych, a także wyodrębniania i porządkowania kluczowych pojęć z uzyskanego tekstu. Ponadto IBM SPSS Modeler Text Analytics może grupować te pojęcia w kategorie.

Około 80% danych przechowywanych w organizacji ma postać dokumentów tekstowych i są to na przykład raporty, strony WWW, wiadomości e-mail i notatki z centrów zgłoszeniowych. Tekst jest kluczowym czynnikiem, dzięki któremu organizacja może lepiej zrozumieć zachowania swoich klientów. System, który wykorzystuje NLP, może inteligentnie wyodrębniać pojęcia, a wśród nich frazy złożone. Ponadto znajomość języka bazowego umożliwia przypisywanie terminów do powiązanych z nimi grup, takich jak produkty, organizacje i osoby — na podstawie znaczeń i kontekstów. W rezultacie można szybko określić istotność informacji w odniesieniu do konkretnych potrzeb. Te wyodrębnione pojęcia i kategorie można łączyć z istniejącymi danymi ustrukturyzowanymi, takimi jak dane demograficzne, a następnie stosować w celu modelowania, korzystając z produktu IBM SPSS Modeler i zawartego w nim pełnego pakietu narzędzi do eksploracji danych, aby w rezultacie takiego połączenia podejmować lepsze decyzje przy zmniejszonej ilości zakłóceń.

Systemy lingwistyczne są zależne od ilości wiedzy — im więcej informacji zawierają ich słowniki, tym wyższa jest jakość wyników. Produkt IBM SPSS Modeler Text Analytics jest dostarczany z zestawem zasobów lingwistycznych, takich jak biblioteki, szablony, słowniki terminów i synonimów. Ten produkt pozwala na dalsze rozwijanie i dostosowywanie tych zasobów lingwistycznych w zależności od potrzeb. Precyzyjne dostosowywanie zasobów lingwistycznych to często proces interaktywny, niezbędny do dokładnego odczytywania i klasyfikowania pojęć. Produkt zawiera również niestandardowe szablony, biblioteki i słowniki dla konkretnych domen, takich jak CRM i genomika.

Informacje o programie IBM Business Analytics

Oprogramowanie IBM Business Analytics dostarcza kompletne, spójne i dokładne informacje, na których mogą polegać osoby decyzyjne chcąc polepszyć wyniki biznesowe. Wszechstronne portfolio obejmujące moduły analiza biznesowa, analiza prognostyczna, zarządzanie wynikami i strategiami finansowymi oraz aplikacje analityczne zapewnia jasny, natychmiastowy i pozwalający na podjęcie działań wgląd w bieżące wyniki oraz daje możliwość przewidywania przyszłych wyników. W połączeniu z licznymi rozwiązaniami branżowymi, sprawdzonymi praktykami i profesjonalnymi usługami organizacje o różnych rozmiarach mogą wspomagać najwyższą produktywność, w sposób pewny zautomatyzować decyzje i uzyskać lepsze wyniki.

Oprogramowanie IBM SPSS Predictive Analytics, będąc częścią tego portfolio, wspomaga organizacje w zakresie przewidywania przyszłych zdarzeń oraz pozwala proaktywnie wpływać na ten wgląd w celu wspomagania lepszych wyników finansowych. Klienci komercyjni, rządowi i uczelnie na całym świecie polegają na technologii IBM SPSS zapewniającej przewagę konkurencyjną przyciągającą, zatrzymującą i rozwijającą klientów, zmniejszając nieuczciwość i ryzyko. Wdrażając oprogramowanie IBM SPSS do swojej codziennej działalności, organizacje stają się przewidującymi przedsiębiorstwami, zdolnymi do zarządzania i automatyzacji decyzji w celu realizacji celów biznesowych i osiągnięcia mierzalnej przewagi konkurencyjnej. W celu uzyskania dalszych informacji lub skontaktowania się z przedstawicielem proszę wejść na stronę http://www.ibm.com/spss.

Wsparcie techniczne

Wsparcie techniczne jest dostępne w celu zapewnienia klientom obsługi technicznej. Klienci mogą się kontaktować z działem Wsparcia technicznego w celu uzyskania pomocy dotyczącej korzystania z produktów IBM Corp. lub pomocy w instalacji dla jednego z obsługiwanych środowisk sprzętowych. Aby skontaktować się z działem Wsparcia technicznego, wejdź na stronę internetową IBM Corp. pod adresem http://www.ibm.com/support. W przypadku prośby o pomoc należy przygotować swoje dane identyfikacyjne, dane swojej organizacji, a także dane dotyczące usług wsparcia.

Rozdział 1. O programie IBM SPSS Modeler Text Analytics

Produkt IBM SPSS Modeler Text Analytics oferuje wydajne możliwości analiz tekstu, które wykorzystują zaawansowane rozwiązania lingwistyczne oraz przetwarzanie języka naturalnego (NLP, Natural Language Processing) w celu szybkiego przetwarzania zróżnicowanych nieustrukturyzowanych danych tekstowych, a także wyodrębniania i porządkowania kluczowych pojęć z uzyskanego tekstu. Ponadto IBM SPSS Modeler Text Analytics może grupować te pojęcia w kategorie.

Około 80% danych przechowywanych w organizacji ma postać dokumentów tekstowych i są to na przykład raporty, strony WWW, wiadomości e-mail i notatki z centrów zgłoszeniowych. Tekst jest kluczowym czynnikiem, dzięki któremu organizacja może lepiej zrozumieć zachowania swoich klientów. System, który wykorzystuje NLP, może inteligentnie wyodrębniać pojęcia, a wśród nich frazy złożone. Ponadto znajomość języka bazowego umożliwia przypisywanie terminów do powiązanych z nimi grup, takich jak produkty, organizacje i osoby — na podstawie znaczeń i kontekstów. W rezultacie można szybko określić istotność informacji w odniesieniu do konkretnych potrzeb. Te wyodrębnione pojęcia i kategorie można łączyć z istniejącymi danymi ustrukturyzowanymi, takimi jak dane demograficzne, a następnie stosować w celu modelowania, korzystając z produktu IBM SPSS Modeler i zawartego w nim pełnego pakietu narzędzi do eksploracji danych, aby w rezultacie takiego połączenia podejmować lepsze decyzje przy zmniejszonej ilości zakłóceń.

Systemy lingwistyczne są zależne od ilości wiedzy — im więcej informacji zawierają ich słowniki, tym wyższa jest jakość wyników. Produkt IBM SPSS Modeler Text Analytics jest dostarczany z zestawem zasobów lingwistycznych, takich jak biblioteki, szablony, słowniki terminów i synonimów. Ten produkt pozwala na dalsze rozwijanie i dostosowywanie tych zasobów lingwistycznych w zależności od potrzeb. Precyzyjne dostosowywanie zasobów lingwistycznych to często proces interaktywny, niezbędny do dokładnego odczytywania i klasyfikowania pojęć. Produkt zawiera również niestandardowe szablony, biblioteki i słowniki dla konkretnych domen, takich jak CRM i genomika.

Wdrożenie. Strumienie eksploracji tekstu można wdrażać za pomocą produktu IBM SPSS Modeler Solution Publisher w celu oceniania danych nieustrukturyzowanych w czasie rzeczywistym. Możliwość wdrażania tych strumieni pozwala na pomyślne zaimplementowanie eksploracji tekstu w zamkniętej pętli. Przykład: organizacja może analizować notatki z rozmów wychodzących i przychodzących, stosując modele predykcyjne w celu zwiększenia dokładności przekazu marketingowego w czasie rzeczywistym.

Uwaga: Aby używać IBM SPSS Modeler Text Analytics z komponentem IBM SPSS Modeler Solution Publisher, dodaj katalog <katalog_instalacyjny>/ext/bin/spss.TMWBServer do zmiennej środowiskowej $LD_LIBRARY_PATH.

Aktualizowanie produktu IBM SPSS Modeler Text Analytics do wersji 18.1

Aktualizacja z poprzednich wersji produktu PASW Text Analytics lub Text Mining for Clementine.

Przed zainstalowaniem produktu IBM SPSS Modeler Text Analytics w wersji 18.1 należy zapisać i wyeksportować wszelkie protokoły TAP, szablony oraz biblioteki z bieżącej wersji, które mają być używane w nowej wersji. Zalecamy zapisanie tych plików w katalogu, który nie zostanie usunięty ani nadpisany podczas instalacji najnowszej wersji.

Po zainstalowaniu najnowszej wersji produktu IBM SPSS Modeler Text Analytics można załadować zapisany plik TAP, dodać zapisane biblioteki oraz zaimportować i załadować wszelkie zapisane szablony, aby następnie używać ich w najnowszej wersji.

Ważne: W przypadku deinstalacji bieżącej wersji bez zapisywania i wyeksportowania potrzebnych plików wszelkie efekty pracy z pakietami TAP, szablonami oraz bibliotekami publicznymi zostaną utracone i nie będą dostępne w produkcie IBM SPSS Modeler Text Analytics 18.1.

Informacje o eksploracji tekstu

Coraz większa ilość informacji jest dziś przechowywana w formatach nieustrukturyzowanych lub częściowo ustrukturyzowanych, takich jak wiadomości e-mail od klientów, notatki konsultantów infolinii, odpowiedzi na otwarte pytania ankietowe, kanały informacyjne, formularze WWW itd. Ta obfitość informacji sprawia, że wiele organizacji staje przed problemem ich zbierania, eksplorowania i wykorzystania.

Eksploracja tekstu to proces polegający na analizowaniu zbiorów materiałów tekstowych w celu wychwycenia w nich najważniejszych pojęć, tematów i motywów oraz ujawnienia ukrytych relacji i trendów bez uprzedniej znajomości konkretnych wyrazów lub terminów, których autorzy tekstu użyli do wyrażenia tych pojęć. Eksploracja tekstu niekiedy niesłusznie mylona jest z wyszukiwaniem informacji, jest jednak zupełnie innym procesem. Precyzyjne wyszukiwanie i niezawodne przechowywanie informacji jest olbrzymim wyzwaniem; równie ważnymi procesami jest wyodrębnianie wysokiej jakości treści, terminologii i relacji ukrytych w tych informacjach oraz zarządzanie nimi.

Eksploracja tekstu i eksploracja danych

Wynikiem lingwistycznej eksploracji każdego pojedynczego tekstu, np. artykułu, jest indeks pojęć oraz informacje o tych pojęciach. Te oczyszczone i ustrukturyzowane informacje można powiązać z innymi źródłami danych, by uzyskać odpowiedzi na takie pytania, jak:

- Które pojęcia występują razem?
- Z jakimi innymi pojęciami są powiązane?
- Jakie kategorie wyższego poziomu można utworzyć na podstawie wyodrębnionych informacji?
- Co można przewidzieć na podstawie pojęć lub kategorii?
- Jak można przewidzieć zachowania na podstawie pojęć lub kategorii?

Łączne zastosowanie technik eksploracji tekstu i eksploracji danych umożliwia bardziej pogłębioną analizę informacji niż operowanie wyłącznie na danych ustrukturyzowanych albo wyłącznie na danych nieustrukturyzowanych. Taki proces zwykle składa się z następujących etapów:

1. **Identyfikacja tekstu do eksploracji.** Przygotowanie tekstu do eksploracji. Jeśli tekst jest zapisany w wielu plikach — zapisanie plików w jednym miejscu. W przypadku baz danych — określenie, w których polach (zmiennych) znajduje się tekst.
2. **Eksploracja tekstu i wyodrębnienie danych ustrukturyzowanych.** Zastosowanie algorytmów eksploracji do tekstu źródłowego.
3. **Zbudowanie modeli pojęć i kategorii.** Zidentyfikowanie kluczowych pojęć i/lub uklasyfikowanie. Zwykle wynikiem eksploracji danych nieustrukturyzowanych jest bardzo duża liczba pojęć. Identyfikacja najlepszych pojęć i kategorii, które należałoby wykorzystać do oceny.
4. **Analiza danych ustrukturyzowanych.** Zastosowanie tradycyjnych technik eksploracji danych, takich jak tworzenie i analiza skupień, klasyfikacja i modelowanie predykcyjne do ujawnienia relacji między pojęciami. Scalenie wyodrębnionych pojęć z pozostałymi danymi ustrukturyzowanymi w celu prognozowania przyszłych zachowań na podstawie pojęć.

Analiza i kategoryzacja tekstu

Analiza tekstu, która jest jedną z postaci analizy jakościowej, polega na wyodrębnieniu użytecznych informacji z tekstu , aby możliwe było pogrupowanie kluczowych idei i pojęć zawartych w tekście w odpowiednią liczbę kategorii. Analizę tekstu można prowadzić na tekstach dowolnego rodzaju i dowolnej długości, jednak strategie analizy będą różne w zależności od charakteru tekstu.

Krótsze rekordy lub dokumenty najłatwiej poddają się kategoryzacji, ponieważ nie są tak bardzo złożone i zwykle zawierają mniej niejednoznacznych wyrazów i odpowiedzi. Na przykład jeśli poprosimy respondentów o wskazanie trzech ulubionych form spędzania wakacji w krótkiej odpowiedzi na pytanie otwarte, to możemy spodziewać się wielu krótkich odpowiedzi, takich jak *opalanie się na plaży*, *zwiedzanie parków narodowych* lub *nicnierobienie*. Z kolei dłuższe, otwarte odpowiedzi mogą być złożone i długie, zwłaszcza jeśli respondenci są wykształceni, zmotywowani i

mają dość czasu na wypełnienie kwestionariusza. Jeśli poprosimy respondentów o opisanie ich przekonań politycznych lub analizujemy blogi o tematyce politycznej, możemy spodziewać się długich komentarzy na przeróżne tematy oraz bardzo różnych stanowisk.

Możliwość wyodrębniania kluczowych pojęć i definiowania wartościowych analitycznie kategorii na podstawie dłuższych tekstów w bardzo krótkim czasie jest najważniejszą korzyścią ze stosowania produktu IBM SPSS Modeler Text Analytics. Korzyść ta wynika z zastosowania kombinacji zautomatyzowanych technik lingwistycznych i statystycznych w celu uzyskania najbardziej wiarygodnych wyników na każdym etapie procesu analizy tekstu.

Przetwarzanie lingwistyczne i przetwarzanie języka naturalnego (NLP)

Głównym problemem przy pracy z tekstem nieustrukturyzowanym jest brak standardowych reguł pisania tekstu w sposób zrozumiały dla komputera. Wypowiedzi językowe, a tym samym także znaczenia, bardzo różnią się między dokumentami i fragmentami tekstu. Jedynym sposobem na precyzyjne wyszukanie i uporządkowanie informacji w takich danych nieustrukturyzowanych jest analiza wypowiedzi i interpretacja ich znaczenia. Istnieje kilka zautomatyzowanych metod wyodrębniania pojęć z informacji nieustrukturyzowanych. Strategie te można ogólnie podzielić na dwie grupy: lingwistyczne i nielingwistyczne.

Niektóre organizacje próbowały stosować zautomatyzowane rozwiązania nielingwistyczne oparte na statystyce i sieciach neuronowych. Rozwiązania te, gdy zostaną zaimplementowane w systemie komputerowym, przeglądają i kategoryzują kluczowe pojęcia szybciej niż ludzie. Niestety, dokładność wyników uzyskiwanych za pomocą tych rozwiązań jest raczej niska. Większość systemów statystycznych po prostu zlicza wystąpienia wyrazów i oblicza ich statystyczne bliskości do pojęć pokrewnych. Systemy takie generują liczne wyniki bezwartościowe (tzw. szum) i nie wychwytują wyników, które powinny znaleźć (tzw. cisza).

Aby skompensować tę ograniczoną dokładność w niektórych rozwiązaniach stosuje się złożone reguły nielingwistyczne, które pomagają w odróżnianiu wyników istotnych od nieistotnych. Takie techniki nazywa się *eksploracją tekstu w oparciu o reguły*.

Natomiast *lingwistyczna eksploracja tekstu* polega na zastosowaniu technik przetwarzania języka naturalnego (NLP — natural language processing), czyli komputerowej analizy ludzkich wypowiedzi, do analizy wyrazów, fraz i składni lub struktury tekstu. System, który wykorzystuje NLP, może inteligentnie wyodrębniać pojęcia, a wśród nich frazy złożone. Co więcej, znajomość języka tekstu umożliwia klasyfikowanie pojęć w grupy pojęć pokrewnych, takich jak produkty, organizacje lub osoby, na podstawie znaczenia i kontekstu.

Lingwistyczna eksploracja tekstu znajduje znaczenia w tekście podobnie, jak robią to ludzie — rozpoznając różne formy wyrazów jako bliskoznaczne i analizując strukturę zdań będącą rusztowaniem, na którym opiera się interpretacja tekstu. Ta strategia jest równie szybka i ekonomiczna, jak systemy statystyczne, ale oferuje znacznie większą dokładność i wymaga mniejszego zaangażowania człowieka.

Aby zilustrować różnicę między statystyczną a lingwistyczną strategią wyodrębniania z tekstów we wszystkich językach oprócz japońskiego, zobaczmy, jakie wyniki wygeneruje każda z tych strategii w odpowiedzi na pytanie o powielanie dokumentów. Zarówno rozwiązanie statystyczne, jak i lingwistyczne musi rozwinąć wyraz powielanie, by uwzględnić jego synonimy, takie jak kopiowanie i reprodukcja. Nieuwzględnienie synonimów prowadziłoby do potencjalnego pominięcia ważnych informacji. Jeśli rozwiązanie statystyczne spróbuje wyszukać inne terminy o tym samym znaczeniu, to prawdopodobnie zwróci także termin birth, generując liczne nieistotne wyniki. Interpretacja języka pozwala pokonać niejednoznaczności, czyniąc z lingwistycznej eksploracji tekstu strategię z definicji bardziej niezawodną.

Dzięki zastosowaniu technik lingwistycznych analizator sentymentu może wyodrębniać wyrażenia o bardziej istotnym znaczeniu. Analiza i wychwytywanie emocji pozwala pokonać niejednoznaczności, czyniąc z lingwistycznej eksploracji tekstu strategię z definicji bardziej niezawodną.

Zrozumienie działania procesu wyodrębniania pomoże w podejmowaniu kluczowych decyzji dotyczących optymalizacji zasobów lingwistycznych (biblioteki, typy, synonimy itd.). Oto etapy procesu wyodrębniania:

- Przekształcenie danych źródłowych do formatu standardowego

- Identyfikacja terminów kandydackich
- Identyfikacja klas równoważności i integracja synonimów
- Przypisanie typów
- Indeksowanie, a następnie — w razie potrzeby — dopasowanie wzorców przy użyciu dodatkowego analizatora.

Etap 1. Przekształcenie danych źródłowych do formatu standardowego

W pierwszym etapie zaimportowane dane są przekształcane do jednolitego formatu, który może być używany do dalszej analizy. To przekształcenie odbywa się wewnętrznie i nie powoduje zmiany oryginalnych danych.

Etap 2. Identyfikacja terminów kandydackich

Ważne jest zrozumienie roli zasobów lingwistycznych w identyfikacji terminów kandydackich podczas wyodrębniania lingwistycznego. Zasoby lingwistyczne są używane przy każdym wyodrębnianiu. Mają postać szablonów, bibliotek i zasobów skompilowanych. Biblioteki zawierają listy wyrazów, relacji i inne informacje służące do definiowania i optymalizacji wyodrębniania. Skompilowanych zasobów nie można przeglądać ani edytować. Jednak pozostałe zasoby można edytować w edytorze Template Editor lub, jeśli pracujesz na interaktywnym pulpicie roboczym, w edytorze Resource Editor.

Zasoby skompilowane to podstawowe, wewnętrzne komponenty mechanizmu wyodrębniania w produkcie IBM SPSS Modeler Text Analytics . Do zasobów tych należy ogólny słownik zawierający listę form podstawowych z kodami części mowy (rzeczownik, czasownik, przymiotnik itd.).

Oprócz tych zasobów skompilowanych razem z produktem dostarczanych jest kilka bibliotek, które można wykorzystać jako dopełnienie definicji typów i pojęć zawartych w zasobach skompilowanych, a także jako źródła synonimów. Biblioteki te — oraz biblioteki utworzone samodzielnie przez użytkownika — składają się z kilku słowników. Są to: słowniki typów, słowniki synonimów oraz słowniki wykluczeń.

Po zaimportowaniu i przekształceniu danych mechanizm wyodrębniania rozpoczyna wykrywanie terminów kandydackich do wyodrębnienia. Terminy kandydackie to wyrazy lub grupy wyrazów identyfikujące pojęcia w tekście. W trakcie przetwarzania tekstu pojedyncze wyrazy (**terminy pojedyncze**) i złożenia wyrazów (**terminy wielowyrazowe**) są identyfikowane na podstawie wzorców części mowy. Następnie poprzez analizę powiązań opartych na sentymencie identyfikowane są kandydackie słowa kluczowe sentymentu.

Uwaga: Terminy we wspomnianym wyżej ogólnym słowniku skompilowanym są zbiorem wszystkich wyrazów, które prawdopodobnie byłyby nieinteresujące lub lingwistycznie niejednoznaczne jako terminy pojedyncze. Wyrazy te są wykluczane z wyników wyodrębniania terminów pojedynczych. Jednak są ponownie analizowane przy określaniu części mowy lub wyszukiwaniu dłuższych terminów wielowyrazowych.

Etap 3. Identyfikacja klas równoważności i integracja synonimów

Po zidentyfikowaniu terminów pojedynczych i terminów wielowyrazowych oprogramowanie, korzystając ze słownika normalizacji, tworzy klasy równoważności. Klasa równoważności jest podstawową formą frazy lub pojedynczą formą dwóch wariantów tej samej frazy. Aby określić, którego pojęcia użyć w klasie równoważności, mechanizm wyodrębniania stosuje poniższe reguły w kolejności, w jakiej zostały tutaj wymienione:

- Forma określona przez użytkownika w bibliotece.
- Najczęściej używana forma, zgodnie z prekompilowanymi zasobami.

Etap 4. Przypisanie typów

Następnie do wyodrębnionych pojęć przypisywane są typy. Typ jest semantyczną grupą pojęć. Na tym etapie używane są zarówno zasoby skompilowane, jak i biblioteki. Typy odzwierciedlają pojęcia poziomowe, kwalifikatory i określenia o wydźwięku pozytywnym i negatywnym, imiona, miejsca, organizacje i nie tylko. Więcej informacji zawiera temat “Słowniki typów” na stronie 177.

Zasoby tekstu japońskiego mają odrębny zestaw typów.

Systemy lingwistyczne są zależne od ilości wiedzy — im więcej informacji zawierają ich słowniki, tym wyższa jest jakość wyników. Odpowiednia modyfikacja zawartości słowników, np. definicji synonimów, może uprościć uzyskane wyniki. Często mamy tu do czynienia z procesem iteracyjnym, który jest niezbędny do precyzyjnego wyszukania pojęć. Zasadniczym elementem programu IBM SPSS Modeler Text Analytics jest mechanizm przetwarzania języka naturalnego (NLP).

Jak działa wyodrębnianie

W trakcie wyodrębniania kluczowych pojęć i koncepcji z odpowiedzi IBM SPSS Modeler Text Analytics przeprowadza lingwistyczną analizę tekstu. Ta strategia zapewnia szybkość i ekonomiczność typową dla systemów statystycznych. Ale oferuje znacznie wyższy poziom dokładności i wymaga mniejszego zaangażowania użytkownika. Lingwistyczna analiza tekstu jest oparta jest na przetwarzaniu języka naturalnego (spotyka się też nazwę: lingwistyka komputerowa).

Ważne: W przypadku tekstów w języku japońskim kroki procesu wyodrębniania są inne.

Zrozumienie działania procesu wyodrębniania pomoże w podejmowaniu kluczowych decyzji dotyczących optymalizacji zasobów lingwistycznych (biblioteki, typy, synonimy itd.). Oto etapy procesu wyodrębniania:

- Przekształcenie danych źródłowych do formatu standardowego
- Identyfikacja terminów kandydackich
- Identyfikacja klas równoważności i integracja synonimów
- Przypisanie typów
- Indeksowanie
- Dopasowywanie wzorców i wyodrębnianie zdarzeń

Etap 1. Przekształcenie danych źródłowych do formatu standardowego

W pierwszym etapie zaimportowane dane są przekształcane do jednolitego formatu, który może być używany do dalszej analizy. To przekształcenie odbywa się wewnętrznie i nie powoduje zmiany oryginalnych danych.

Etap 2. Identyfikacja terminów kandydackich

Ważne jest zrozumienie roli zasobów lingwistycznych w identyfikacji terminów kandydackich podczas wyodrębniania lingwistycznego. Zasoby lingwistyczne są używane przy każdym wyodrębnianiu. Mają postać szablonów, bibliotek i zasobów skompilowanych. Biblioteki zawierają listy wyrazów, relacji i inne informacje służące do definiowania i optymalizacji wyodrębniania. Skompilowanych zasobów nie można przeglądać ani edytować. Jednak pozostałe zasoby (szablony) można edytować w oknie Template Editor lub, jeśli aktywna jest sesja interaktywnego pulpitu roboczego, w oknie Resource Editor.

Zasoby skompilowane to podstawowe, wewnętrzne komponenty mechanizmu wyodrębniania w produkcie IBM SPSS Modeler Text Analytics. Do zasobów tych należy ogólny słownik zawierający listę form podstawowych z kodami części mowy (rzeczownik, czasownik, przymiotnik, przysłówek, imiesłów przymiotnikowy, spójnik, określnik lub przyimek). Do zasobów należą także zastrzeżone typy wbudowane, które umożliwiają przypisywanie wielu wyodrębnionych terminów do typów <Location> (miejsce), <Organization> (organizacja) lub <Person> (osoba). Więcej informacji zawiera temat “Typy wbudowane” na stronie 178.

Oprócz tych zasobów skompilowanych razem z produktem dostarczanych jest kilka bibliotek, które można wykorzystać jako dopełnienie definicji typów i pojęć zawartych w zasobach skompilowanych, a także jako źródła innych typów i synonimów. Biblioteki te — oraz biblioteki utworzone samodzielnie przez użytkownika — składają się z kilku słowników. Są to: słowniki typów, słowniki zastąpień (synonimów i elementów opcjonalnych) oraz słowniki wykluczeń. Więcej informacji zawiera Rozdział 15, “Praca z bibliotekami”, na stronie 167.

Po zaimportowaniu i przekształceniu danych mechanizm wyodrębniania rozpoczyna wykrywanie terminów kandydackich do wyodrębnienia. Terminy kandydackie to wyrazy lub grupy wyrazów identyfikujące pojęcia w tekście. Podczas przetwarzania tekstu pojedyncze wyrazy (*terminy pojedyncze*), których nie ma w zasobach skompilowanych, są traktowane jako terminy kandydackie. Kandydackie *terminy wielowyrazowe* są identyfikowane na podstawie wzorców części mowy. Na przykład termin wielowyrazowy sports car, który jest zgodny ze wzorcem części mowy "przymiotnik rzeczownik", składa się z dwóch elementów. Termin wielowyrazowy fast sports car, który jest zgodny ze wzorcem części mowy "przymiotnik przymiotnik rzeczownik", składa się z trzech elementów.

Uwaga: Terminy w słowniku reprezentują ogólne wymienione powyżej skompilowanej listy wszystkich słów, które są prawdopodobnie nieinteresujące lub lingwistycznie niejednoznaczny jako uniterms. Wyrazy te są wykluczane z wyników wyodrębniania terminów pojedynczych. Jednak są ponownie analizowane przy określaniu części mowy lub wyszukiwaniu dłuższych terminów wielowyrazowych.

Wreszcie, specjalny algorytm analizuje łańcuchy zapisane wielkimi literami, takie jak nazwy działów, co pozwala na wyodrębnienie takich specjalnych wzorców.

Etap 3. Identyfikacja klas równoważności i integracja synonimów

Po zidentyfikowaniu terminów pojedynczych i terminów wielowyrazowych oprogramowanie, korzystając z zestawu algorytmów, porównuje je i tworzy klasy równoważności. Klasa równoważności to podstawowa forma frazy lub jedna forma dwóch wariantów tej samej frazy. Frazy przypisuje się do klas równoważności, aby mieć pewność, że np. frazy president of the company i company president nie zostaną potraktowane jako różne pojęcia. Aby określić, którego pojęcia użyć w klasie równoważności, tj. czy terminem wiodącym ma być president of the company, czy company president, mechanizm wyodrębniania stosuje poniższe reguły w kolejności, w jakiej zostały tutaj wymienione:

- Forma określona przez użytkownika w bibliotece.
- Forma występująca najczęściej w całej treści tekstu.
- Najkrótsza forma występująca w całej treści tekstu (która zwykle odpowiada formie podstawowej).

Etap 4. Przypisanie typów

Następnie do wyodrębnionych pojęć przypisywane są typy. Typ jest semantyczną grupą pojęć. Na tym etapie używane są zarówno zasoby skompilowane, jak i biblioteki. Typy odzwierciedlają pojęcia poziomowe, kwalifikatory i określenia o wydźwięku pozytywnym i negatywnym, imiona, miejsca, organizacje i nie tylko. Użytkownik może zdefiniować dodatkowe typy. Więcej informacji zawiera temat “Słowniki typów” na stronie 177.

Krok 5. Indeksowanie

Cały zbiór rekordów lub dokumentów jest indeksowany, co polega na utworzeniu, dla każdej klasy równoważności, wskaźników między pozycjami w tekście a reprezentatywnymi terminami. Zakłada się przy tym, że wszystkie odmienione wystąpienia pojęcia kandydackiego są indeksowane jako forma podstawowa tego pojęcia. Obliczana jest globalna liczebność wystąpień każdej formy podstawowej.

Krok 6. Dopasowywanie wzorców i wyodrębnianie zdarzeń

IBM SPSS Modeler Text Analytics potrafi wykrywać nie tylko typy i pojęcia, lecz także relacje między nimi. W produkcie dostępnych jest kilka algorytmów i bibliotek, które umożliwiają wyodrębnianie relacji między typami i pojęciami. Są szczególnie przydatne, gdy interesują nas określone opinie (np. reakcje na produkt) lub powiązania odzwierciedlające relacje między osobami lub obiektami (na przykład powiązania między grupami politycznymi lub genomami).

Jak działa kategoryzacja

Podczas tworzenia modeli kategorii w programieIBM SPSS Modeler Text Analytics istnieje kilka technik, które można zastosować do tworzenia kategorii. Ponieważ każdy zbiór danych jest inny, liczba metod tworzenia kategorii i kolejność ich stosowania może się z czasem zmieniać. Jako że każdy użytkownik może inaczej interpretować te same wyniki, konieczne może być wypróbowanie różnych technik i wybranie tej, która przynosi najlepsze wyniki w analizie

konkretnych danych tekstowych. W programie IBM SPSS Modeler Text Analytics można tworzyć modele kategorii na interaktywnym pulpicie roboczym, który stwarza warunki do dalszej eksploracji i optymalizacji kategorii.

W niniejszym podręczniku **budowanie kategorii** oznacza generowanie definicji kategorii i klasyfikację przy użyciu jednej lub wielu wbudowanych technik, natomiast **kategoryzacja** oznacza ocenianie, czyli proces przypisywania unikalnych identyfikatorów (nazwa/identyfikator/wartość) do definicji kategorii dla każdego rekordu lub dokumentu.

W trakcie budowania kategorii wyodrębnione pojęcia i typy są używana jako elementy składowe kategorii. Podczas budowania kategorii rekordy lub dokumenty są automatycznie przypisywane do kategorii, jeśli zawierają tekst pasujący do elementu definicji tej kategorii.

IBM SPSS Modeler Text Analytics oferuje kilka zautomatyzowanych technik budowania kategorii, które pomagają w szybkiej kategoryzacji dokumentów lub rekordów.

Techniki grupowania

Każda z dostępnych technik dobrze nadaje się do pracy z określonymi rodzajami danych i w określonych warunkach, jednak często przydatna jest możliwość połączenia w jednej analizie różnych technik w celu wydobycia bogatszego zbioru informacji z dokumentów lub rekordów. Pojęcie może znaleźć się w więcej niż jednej kategorii, mogą też pojawić się kategorie nadmiarowe.

Concept Root Derivation. Te technika tworzy kategorie, wybierając jedno pojęcie i wyszukując pojęcia mu pokrewne poprzez analizę relacji morfologicznych lub wspólnych rdzeni między komponentami tych pojęć. Technika ta jest bardzo przydatna do rozpoznawania synonimicznych pojęć złożonych z więcej niż jednego wyrazu, ponieważ pojęcia w każdej wygenerowanej kategorii są synonimami lub są bardzo bliskie znaczeniowo. Technika ta działa z danymi o różnych długościach i generuje mniejszą liczbę bardziej zwartych kategorii. Na przykład pojęcie opportunities to advance zostałoby połączone w grupę z pojęciami opportunity for advancement i advancement opportunity. Więcej informacji zawiera temat “Wywodzenie rdzeni pojęć” na stronie 107. Ta opcja nie jest dostępna w przypadku tekstu japońskiego.

Semantic Network. Ta technika najpierw rozpoznaje możliwe sensy każdego pojęcia na podstawie obszernego indeksu relacji między wyrazami, a potem tworzy kategorie poprzez grupowanie pojęć pokrewnych. Sprawdza się najlepiej, gdy pojęcia są znane sieci semantycznej i nie są zbyt niejednoznaczne. Jest mniej użyteczna, gdy tekst zawiera terminologię specjalistyczną lub żargon nieznany sieci. Przykładowo pojęcie granny smith apple mogłoby zostać połączone w grupę z pojęciami gala apple i winesap apple, ponieważ podobnie jak „granny smith” oznaczają odmiany jabłoni. Natomiast pojęcie animal mogłoby zostać połączone w grupę z pojęciami cat i kangaroo, ponieważ są one hiponimami słowa animal (tj. zawężają jego znaczenie). W tej wersji technika ta jest dostępna tylko w odniesieniu do języka angielskiego. Więcej informacji zawiera temat “Sieci semantyczne” na stronie 109.

Concept Inclusion. Ta technika buduje kategorie, łącząc z w grupę z jednym pojęciem inne pojęcia złożone z wielu terminów (wielu wyrazów) w zależności od tego, czy zawierają one wyrazy będące podzbiorami czy nadzbiorami występującego w nim wyrazu. Na przykład pojęcie seat zostałoby połączone w grupę z pojęciami safety seat, seat belt i seat belt buckle. Więcej informacji zawiera temat “Włączanie pojęć” na stronie 108.

Co-occurrence. Ta technika tworzy kategorie na podstawie współwystąpień znalezionych w tekście. Koncepcja działania tej techniki opiera się na tym, że gdy pojęcia lub wzorce pojęć często występują razem w dokumentach i rekordach, takie współwystąpienia odzwierciedlają potencjalną relację, którą warto uwzględnić w definicjach kategorii. Istotne współwystąpienia wyrazów powodują utworzenie reguły współwystępowania, którą można wykorzystać jako deskryptor dla nowej podkategorii. Na przykład, jeśli wiele rekordów zawiera wyrazy price i availability (ale niewiele rekordów zawiera jeden z tych wyrazów bez drugiego), to pojęcia te zostaną połączone w regułę współwystępowania, (price & available) i przypisane na przykład do podkategorii kategorii price. Więcej informacji zawiera temat “Reguły współwystępowania” na stronie 110.

Minimalna ilość rekordów dokumentów. Aby móc lepiej ocenić, na ile interesujące są współwystąpienia, można zdefiniować minimalną liczbę dokumentów lub rekordów, jaką musi zawierać dane współwystąpienie, aby było używane jako deskryptor w kategorii.

Węzły produktu IBM SPSS Modeler Text Analytics

Wraz z wieloma standardowymi węzłami dostarczanymi z produktem IBM SPSS Modeler użytkownik ma do dyspozycji także węzły eksploracji tekstu, które umożliwiają zastosowanie zaawansowanych technik analizy tekstu w strumieniach. IBM SPSS Modeler Text Analytics oferuje kilka węzłów eksploracji tekstu. Węzły te są dostępne na karcie IBM SPSS Modeler Text Analytics palety węzłów.

Dostępne są następujące węzły:

- **Węzeł źródłowy File List** generuje listę nazw dokumentów jako dane wejściowe do procesu eksploracji tekstu. Możliwość ta jest przydatna, jeśli tekst znajduje się w dokumentach zewnętrznych, a nie w bazie danych lub innym pliku ustrukturyzowanym. Węzeł generuje jedną zmienną z jednym rekordem dla każdego dokumentu lub folderu na liście; zmienne te można wybrać jako dane wejściowe dla kolejnego węzła eksploracji tekstu. Więcej informacji zawiera temat “Węzeł File List” na stronie 11.
- **Węzeł źródłowy Web Feed** umożliwia odczyt tekstu z kanałów informacyjnych WWW, takich jak blogi, lub kanałów informacyjnych w formacie RSS lub HTML, a potem wykorzystanie tego tekstu w procesie eksploracji. Węzeł generuje jedną lub więcej zmiennych dla każdego rekordu znalezionego w kanałach informacyjnych. Zmienne te można wybrać jako dane wejściowe dla kolejnego węzła eksploracji tekstu. Więcej informacji zawiera temat “Węzeł Web Feed” na stronie 13.
- **Węzeł Language Identifier** skanuje tekst źródłowy w celu określenia, w którym języku naturalnym został napisany, a następnie umieszcza o tym informację w nowej zmiennej. Węzeł ten, wykorzystywany głównie podczas pracy z dużymi ilościami danych, okazuje się bardzo przydatny, gdy w źródłach danych występuje kilka języków, z których przetworzony ma być tylko jeden. Więcej informacji zawiera temat “Węzeł języka” na stronie 17.
- **Węzeł Test Mining** używa metod lingwistycznych, aby wyodrębnić kluczowe pojęcia z tekstu, tworzy kategorie na podstawie tych pojęć i innych danych i oferuje możliwość identyfikowania relacji i powiązań między pojęciami na podstawie znanych wzorców (jest to tzw. analiza powiązań w tekście). Węzeł ten może być używany do eksplorowania treści tekstowych lub do wygenerowania modelu pojęć albo modelu kategorii. Pojęcia i kategorie można łączyć z istniejącymi danymi ustrukturyzowanymi, takimi jak dane demograficzne, a następnie stosować w modelowaniu. Więcej informacji zawiera temat “Węzeł modelowania Text Mining” na stronie 20.
- **Węzeł Text Link Analysis** pozwala na wyodrębnianie pojęć, a także identyfikuje relacje między pojęciami na podstawie znanych wzorców w tekście. Wyodrębnianie wzorców umożliwia wykrywanie relacji między pojęciami, a także opinii lub kwalifikatorów skojarzonych z tymi pojęciami. Węzeł Text Link Analysis oferuje bardziej bezpośredni sposób identyfikowania i wyodrębniania wzorców z tekstu, a także może dodawać wynikowe wzorce do zbioru danych w strumieniu. Jednak analizę TLA można również prowadzić na interaktywnym pulpicie modelowania, w ramach węzła modelowania Text Mining. Więcej informacji zawiera temat “Węzeł Text Link Analysis” na stronie 47.
- Podczas eksploracji tekstu z dokumentów zewnętrznych **węzeł wynikowy Text Mining** umożliwia wygenerowania strony HTML zawierającej odsyłacze do dokumentów, z których zostały wyodrębnione pojęcia. Więcej informacji zawiera temat “Węzeł File Viewer” na stronie 55.

Zastosowania

Zasadniczo każdy, kto często musi przeglądać duże ilości dokumentów, aby zidentyfikować kluczowe elementy do dalszej eksploracji, może odnieść korzyści z zastosowania produktu IBM SPSS Modeler Text Analytics.

Oto niektóre obszary zastosowań produktu:

- **Badania naukowe i medyczne.** Eksploracja materiałów pomocniczych, takich jak raporty patentowe i artykuły w czasopismach specjalistycznych. Wykrywanie powiązań, które były dotychczas nieznane (np. powiązanie między lekarzem a określonym produktem), i wytyczanie na ich podstawie kierunków dalszej eksploracji. Przyspieszenie prac nad nowymi lekami. Badania genetyczne.
- **Analizy inwestycyjne.** Analiza codziennych raportów analitycznych, artykułów prasowych i informacji dla prasy publikowanych przez spółki w celu wykrycia kluczowych elementów strategii i zmian na rynku. Analiza trendów w zbiorach takich informacji ujawnia potencjalne problemy lub szanse stojące przed jedną firmą lub całą branżą w danym okresie.

- **Wykrywanie oszustw.** Wykrywanie anomalii i potencjalnych zagrożeń oszustwem w obszernych danych tekstowych w sektorze bankowości lub służby zdrowia.
- **Badania rynku.** Badania mające na celu rozpoznanie najważniejszych tematów w odpowiedziach respondentów na pytania otwarte.
- **Analiza bloków i kanałów WWW.** Eksploracja i budowanie modeli na podstawie kluczowych pojęć pojawiających się w kanałach z wiadomościami, blogach itd.
- **Zarządzanie relacjami z klientami (CRM).** Budowanie modeli na podstawie danych z wszystkich punktów interakcji z klientem, takich jak poczta elektroniczna, transakcje i ankiety.

Rozdział 2. Wczytywanie tekstu źródłowego

Dane do eksploracji tekstu mogą być zapisane w dowolnym standardowym formacie używanym przez program IBM SPSS Modeler, w tym także w bazach danych i innych formatach „tabelarycznych" złożonych z wierszy i kolumn, a także w formatach dokumentów, takich jak Microsoft Word, Adobe PDF lub HTML, które nie mają takiej struktury.

- Aby wczytywać w postaci tekstowej dokumenty nieposiadające standardowej struktury danych, zapisane w takich formatach, jak Microsoft Word, Microsoft Excel, i Microsoft PowerPoint, a także w formatach Adobe PDF, XML, HTML i innych, można użyć węzła File List do wygenerowania listy dokumentów lub folderów stanowiących dane wejściowe dla procesu eksploracji tekstu. Aby uzyskać więcej informacji, patrz "Węzeł File List".
- Aby wczytywać tekst z kanałów WWW, takich jak blogi lub kanały z aktualnościami w formacie RSS lub HTML, można wykorzystać węzeł Web Feed do formatowania danych z kanału WWW do postaci odpowiedniej dla procesu eksploracji tekstu. Aby uzyskać więcej informacji, patrz "Węzeł Web Feed" na stronie 13.
- Do wczytywania tekstu w dowolnym standardowym formacie danych używanym przez program SPSS Modeler, na przykład w formie bazy danych zawierającej jedno lub więcej pól tekstowych z uwagami klientów, można używać dowolnych standardowych węzłów źródłowych programu SPSS Modeler. Aby uzyskać więcej informacji, patrz dokumentację węzła SPSS Modeler.
- Podczas przetwarzania dużych ilości danych, które mogą zawierać tekst w wielu językach, należy użyć węzła języka, aby wskazać język używany w określonej zmiennej. Aby uzyskać więcej informacji, patrz "Węzeł języka" na stronie 17.

Węzeł File List

Aby wczytywać w postaci tekstowej dokumenty nieustrukturyzowane zapisane w takich formatach, jak Microsoft Word, Microsoft Excel i Microsoft PowerPoint, a także w formatach Adobe PDF, XML, HTML i innych, można użyć węzła File List do wygenerowania listy dokumentów lub folderów stanowiących dane wejściowe dla procesu eksploracji tekstu. Jest to konieczne, ponieważ nieustrukturyzowanych dokumentów tekstowych nie da się przedstawić w formie zmiennych i rekordów (wierszy i kolumn), tak jak innych danych, z których korzysta IBM SPSS Modeler.

Węzeł File List działa jako węzeł źródłowy.

Węzeł ten znajduje się na karcie IBM SPSS Modeler Text Analytics palety węzłów w oknie IBM SPSS Modeler. Więcej informacji zawiera temat "Węzły produktu IBM SPSS Modeler Text Analytics" na stronie 8.

Ważne: Wszelkie nazwy katalogów i plików zawierające znaki nieuwzględnione w standardzie kodowania obowiązującym na lokalnym komputerze nie są obsługiwane. Przy próbie wykonania strumienia zawierającego węzeł File List wszelkie nazwy plików lub katalogów zawierające takie znaki spowodują niepowodzenie wykonania. Taka sytuacja może wystąpić w przypadku nazw katalogów lub plików w językach obcych, na przykład nazw plików w języku japońskim na komputerze z francuskimi ustawieniami regionalnymi.

Local data support. Jeżeli masz połączenie ze zdalnym serwerem IBM SPSS Modeler Text Analytics Server i strumień zawierający węzeł File List, to dane powinny znajdować się na tym samym komputerze, co IBM SPSS Modeler Text Analytics Server , lub komputer serwera musi mieć dostęp do folderu, w którym przechowywane są dane źródłowe węzła File List.

Uwaga: Nie można używać węzła File List do oceny w ramach konfiguracji IBM SPSS Collaboration and Deployment Services - Scoring.

Węzeł File List: karta Settings

Na tej karcie można definiować katalogi, rozszerzenia plików i dane wejściowe dla węzła.

Uwaga: Mechanizm wyodrębniania w ramach eksploracji tekstu nie może przetwarzać plików Microsoft Office i Adobe PDF na platformach innych niż Microsoft Windows. Jednak pliki XML, HTML lub tekstowe zawsze mogą być przetwarzane.

Wszelkie nazwy katalogów i plików zawierające znaki nieuwzględnione w standardzie kodowania obowiązującym na lokalnym komputerze nie są obsługiwane. Przy próbie wykonania strumienia zawierającego węzeł File List wszelkie nazwy plików lub katalogów zawierające takie znaki spowodują niepowodzenie wykonania. Taka sytuacja może wystąpić w przypadku nazw katalogów lub plików w językach obcych, na przykład nazw plików w języku japońskim na komputerze z francuskimi ustawieniami regionalnymi.

Directory Określa katalog główny zawierający dokumenty, których listę chcesz utworzyć.

- **Include subdirectories** Określa, że przeszukiwane mają być także podkatalogi.

File type(s) to include in list: Można wybierać typy plików i rozszerzenia lub anulować ich wybór. Anulowanie wyboru rozszerzenia powoduje ignorowanie plików o tym rozszerzeniu. Można filtrować według następujących rozszerzeń:

Tabela 1. Filtry typów plików według rozszerzeń nazw.

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

Uwaga: Aby uzyskać więcej informacji, patrz “Węzeł File List” na stronie 11.

Jeśli masz pliki bez rozszerzeń lub z nazwami zakończonymi kropką (na przykład Plik01 lub Plik01.),to aby je wybrać, skorzystaj z opcji **No extension**.

Input encoding Jeśli zmienna wyjściowa będzie zawierać dokładny tekst, wybierz odpowiednią wartość z poniższej listy:

- Automatic (European)
- Automatic (Japanese)
- UTF-8
- UTF-16
- ISO-8859-1
- ISO-8859-2
- Windows-1250
- US ascii

Wyniki są prezentowane w formie dokumentów tekstowych UTF-8.

Ważne: Począwszy od wersji 14 opcja „List of directories” nie jest już dostępna i generowana jest tylko lista plików

Węzeł List Node: pozostałe karty

Karta Typy jest standardową kartą w węzłach programu IBM SPSS Modeler, podobnie jak karta Adnotacje.

Korzystanie z węzła File List w eksploracji tekstu

Węzeł File List jest używany, gdy dane tekstowe rezydują w zewnętrznych dokumentach nieustrukturyzowanych, w takich formatach, jak Microsoft Word, Microsoft Excel i Microsoft PowerPoint, a także w formatach Adobe PDF, XML, HTML i innych.

Załóżmy na przykład, że połączyliśmy węzeł File List z węzłem Text Mining, aby uzyskać tekst przechowywany w zewnętrznych dokumentach:

1. **Węzeł File List (karta Settings).** Najpierw dodaliśmy ten węzeł do strumienia, aby określić, gdzie są przechowywane dokumenty tekstowe. Wybraliśmy katalog zawierający wszystkie dokumenty, na których chcemy przeprowadzać eksplorację tekstu.
2. **Węzeł Text Mining (karta Fields).** Następnie dodaliśmy węzeł Text Mining i połączyliśmy go z węzłem File List. W tym węźle zdefiniowaliśmy format wejściowy, szablon zasobów i format wyjściowy. Wybraliśmy nazwę zmiennej wygenerowaną z węzła File List, zmienną tekstową oraz inne ustawienia. Więcej informacji zawiera temat “Korzystanie z węzła Text Mining w strumieniu” na stronie 30.

Więcej informacji o korzystaniu z węzła Text Mining zawiera sekcja “Węzeł modelowania Text Mining” na stronie 20.

Węzeł Web Feed

Węzeł Web Feed może być używany do przygotowywania danych z kanałów informacyjnych WWW na potrzeby eksploracji tekstu. Ten węzeł przyjmuje kanały informacyjne WWW w dwóch formatach:

- Format RSS. RSS to prosty standaryzowany format oparty na XML dla treści WWW. Adres URL dla tego formatu wskazuje na stronę, na której znajduje się zestaw dowiązanych artykułów, takich jak syndykowane źródła wiadomości i blogi. RSS jest formatem standaryzowanym, dlatego każdy dowiązany artykuł jest automatycznie identyfikowany i traktowany jako osobny rekord w wynikowym strumieniu danych. Nie są wymagane żadne dodatkowe dane wejściowe umożliwiające identyfikowanie istotnych danych tekstowych oraz rekordów z kanału informacyjnego, chyba że względem tekstu zastosowano technikę filtrowania.
- Format HTML. Na karcie Input można zdefiniować jeden lub większą liczbę adresów URL do stron HTML. Następnie na karcie Records należy zdefiniować znacznik początku rekordu, a także wskazać znaczniki ograniczające treść docelową i przypisać te znaczniki do wybranych pól wyjściowych (opis, tytuł, data modyfikacji itp.). Więcej informacji zawiera temat “Węzeł Web Feed: karta Records” na stronie 14.

Ważne! Jeśli informacje mają być pobierane z sieci WWW za pośrednictwem serwera proxy, należy w pliku net.properties włączyć serwer proxy zarówno dla klienta, jak i serwera IBM SPSS Modeler Text Analytics. Szczegółowe instrukcje postępowania znajdują się w samym pliku. Dotyczy to sytuacji, w której dostęp do sieci WWW odbywa się za pośrednictwem węzła Web Feed lub pobierana jest licencja na oprogramowanie SDL Software as a Service (SaaS), ponieważ odpowiednie połączenia przechodzą przez Java™. Wspomniany plik domyślnie znajduje się w lokalizacji *C:\Program Files\IBM\SPSS\Modeler\18.1\jre\lib\net.properties*.

Wynikiem tego węzła jest zestaw pól używanych do opisywania rekordów. Pole **Description** jest często używane, ponieważ zawiera większość zawartości tekstowej. Jednak użytkownik może być również zainteresowany innymi polami, na przykład krótkim opisem rekordu (pole **Short Desc**) lub tytułem rekordu (pole **Title**). Dowolne pola wyjściowe można wybrać jako dane wyjściowe dla następnego w kolejności węzła Text Mining.

Uwaga: Nie można używać węzła Web Feed do oceny w ramach konfiguracji produktu IBM SPSS Collaboration and Deployment Services - Scoring.

Węzeł ten znajduje się na karcie IBM SPSS Modeler Text Analytics palety węzłów w oknie IBM SPSS Modeler. Więcej informacji zawiera temat “Węzły produktu IBM SPSS Modeler Text Analytics” na stronie 8.

Węzeł Web Feed: karta Input

Karta Input jest używana do określania co najmniej jednego adresu WWW lub URL w celu przechwycenia danych tekstowych. W kontekście eksploracji tekstu można określać adresy URL dla kanałów informacyjnych zawierających dane tekstowe.

Ważne: W przypadku pracy z danymi innymi niż RSS lepszym rozwiązaniem może być użycie narzędzia do ekstrakcji danych ze stron WWW, takiego jak WebQL®, które posłuży do automatyzacji gromadzenia treści, a następnie tworzenia odniesień z tego narzędzia przy użyciu innego węzła źródłowego.

Można ustawić następujące parametry:

Enter or paste URLs. Do tego pola można wpisać lub wklejać adresy URL (co najmniej jeden). Jeśli wprowadzasz więcej niż jeden adres, wprowadź tylko jeden na wiersz i użyj klawisza **Enter/Return**, aby rozdzielić wiersze. Wprowadź pełny adres URL do pliku. W przypadku kanałów informacyjnych te adresy URL mogą przyjmować jeden z dwóch formatów:

- Format RSS. RSS to prosty standaryzowany format oparty na XML dla treści WWW. Adres URL dla tego formatu wskazuje na stronę, na której znajduje się zestaw dowiązanych artykułów, takich jak syndykowane źródła wiadomości i blogi. RSS jest formatem standaryzowanym, dlatego każdy dowiązany artykuł jest automatycznie identyfikowany i traktowany jako osobny rekord w wynikowym strumieniu danych. Nie są wymagane żadne dodatkowe dane wejściowe umożliwiające identyfikowanie istotnych danych tekstowych oraz rekordów z kanału informacyjnego, chyba że względem tekstu zastosowano technikę filtrowania.
- Format HTML. Na karcie Input można zdefiniować jeden lub większą liczbę adresów URL do stron HTML. Następnie na karcie Records należy zdefiniować znacznik początku rekordu, a także wskazać znaczniki ograniczające treść docelową i przypisać te znaczniki do wybranych pól wyjściowych (opis, tytuł, data modyfikacji itp.). W przypadku pracy z danymi innymi niż RSS lepszym rozwiązaniem może być użycie narzędzia do ekstrakcji danych ze stron WWW, takiego jak WebQL®, które posłuży do automatyzacji gromadzenia treści, a następnie tworzenia odniesień z tego narzędzia przy użyciu innego węzła źródłowego. Więcej informacji zawiera temat "Węzeł Web Feed: karta Records".

Number of most recent entries to read per URL. To pole określa maksymalną liczbę rekordów do odczytu dla każdego adresu URL określonego w polu, począwszy od pierwszego rekordu znalezionego w kanale informacyjnym. Ilość tekstu wpływa na szybkość przetwarzania podczas dalszego wyodrębniania w węźle Text Mining lub węźle Text Link Analysis.

Save and reuse previous web feeds when possible. Ta opcja umożliwia skanowanie kanałów informacyjnych WWW, a przetworzone wyniki są buforowane. Następnie po kolejnych wykonaniach strumienia, jeśli zawartość konkretnego kanału informacyjnego nie uległa zmianie lub jeśli kanał informacyjny jest niedostępny (na przykład z powodu braku dostępu do Internetu), wówczas w celu przyspieszenia przetwarzania używana jest zbuforowana wersja. Każda nowa treść wykryta w tych kanałach informacyjnych również jest buforowana przy następnym wykonaniu węzła.

- **Etykieta**. W przypadku wybrania opcji **Save and reuse previous web feeds when possible** należy określić nazwę etykiety dla wyników. Ta etykieta będzie używana do opisywania kanałów informacyjnych buforowanych na serwerze. Jeśli żadna etykieta nie zostanie określona lub nie zostanie rozpoznana, wówczas ponowne użycie nie będzie możliwe. Buforami kanałów informacyjnych WWW można zarządzać w tabeli sesji produktu IBM SPSS Text Analytics Administration Console zawartej w IBM SPSS Deployment Manager. Więcej informacji zawiera dokument Deployment Manager — Podręcznik użytkownika.

Węzeł Web Feed: karta Records

Karta Records służy do określania zawartości tekstowej kanałów informacyjnych innych niż RSS poprzez wskazywanie początku każdego rekordu, a także poprzez podawanie innych istotnych informacji dotyczących poszczególnych rekordów. Jeśli wiadomo, że kanał informacyjny inny niż RSS (HTML) zawiera tekst znajdujący się w wielu rekordach, wówczas na tej karcie należy wskazać znacznik początku rekordu, ponieważ w przeciwnym wypadku tekst będzie traktowany jako jeden rekord. Kanały RSS są standaryzowane i nie wymagają określania żadnych znaczników na tej karcie, ale nadal można wyświetlić podgląd treści na karcie Preview.

Ważne! W przypadku pracy z danymi innymi niż RSS lepszym rozwiązaniem może być użycie narzędzia do ekstrakcji danych ze stron WWW, takiego jak WebQL®, które posłuży do automatyzacji gromadzenia treści, a następnie tworzenia odniesień z tego narzędzia przy użyciu innego węzła źródłowego.

URL. Ta lista rozwijana zawiera adresy URL wprowadzone na karcie Input. Zawiera kanały informacyjne w formacie HTML i RSS. Jeśli adres URL jest zbyt długi i nie zmieści się na liście rozwijanej, zostanie automatycznie obcięty na środku, a obcięty tekst zostanie zastąpiony wielokropkiem, np. *http://www.ibm.com/example/start-of-address...rest-of-address/path.htm*.

- Jeśli **kanał informacyjny w formacie HTML** zawiera więcej niż jeden rekord (lub wpis), wówczas można zdefiniować znaczniki HTML zawierające dane odpowiadające polu widocznemu w tabeli. Można na przykład zdefiniować znacznik początkowy oznaczający, że rozpoczął się nowy rekord, znacznik daty modyfikacji albo nazwiska autora.
- **W przypadku kanałów informacyjnych w formacie RSS** nie pojawia się żadna zachęta do wprowadzania znaczników, ponieważ RSS jest formatem standaryzowanym. Jednak w razie potrzeby przykładowe wyniki można wyświetlić na karcie Preview. Wszystkie rozpoznane kanały informacyjne RSS poprzedza obraz logo RSS.

Karta Source. Na tej karcie można wyświetlić kod źródłowy dla dowolnych kanałów informacyjnych HTML. Ten kod nie jest dostępny do edycji. Za pomocą pola Find można odszukać konkretne znaczniki i informacje na tej stronie, które następnie można skopiować i wkleić do tabeli poniżej. W polu Find nie jest rozróżniana wielkość znaków i nie są znajdowane łańcuchy częściowe.

Karta Preview. Na tej karcie można sprawdzić to, w jaki sposób rekord będzie odczytywany przez węzeł Web Feed. Jest to szczególnie użyteczne w przypadku kanałów informacyjnych HTML, ponieważ umożliwia zmianę sposobu odczytu rekordu poprzez zdefiniowanie znaczników HTML w tabeli poniżej karty Preview.

Non-RSS record start tag. Ta opcja obowiązuje tylko w przypadku kanałów informacyjnych innych niż RSS. Jeśli kanał informacyjny HTML zawiera dużą ilość tekstu, który zostanie przez użytkownika podzielony na wiele rekordów, wówczas w tej opcji należy określić znacznik HTML wskazujący początek rekordu (na przykład artykuł lub wpis w blogu). Jeśli w przypadku kanału informacyjnego innego niż RSS żaden taki znacznik nie zostanie zdefiniowany, wówczas cała strona będzie traktowana jako jeden rekord, cała zawartość będzie dostępna w polu **Description**, a data wykonania węzła będzie używana jako data modyfikacji (**Modified Date**), a także jako data publikacji (**Published Date**).

Field table. Ta opcja obowiązuje tylko w przypadku kanałów informacyjnych innych niż RSS. W tej tabeli można podzielić zawartość tekstową na konkretne pola wynikowe. W tym celu należy wprowadzić znacznik początkowy dla dowolnego ze wstępnie zdefiniowanych pól wyjściowych. Należy wprowadzić tylko znacznik początkowy. Wszystkie operacje dopasowania są wykonywane poprzez analizę składni HTML i dopasowywanie zawartości tabeli do nazw i atrybutów znaczników znalezionych w kodzie HTML. Przyciski dostępne na dole mogą być używane do kopiowania zdefiniowanych znaczników i ich ponownego użycia w pozostałych kanałach informacyjnych.

Tabela 2. Możliwe pola wyjściowe dla kanałów informacyjnych innych niż RSS (formaty HTML)

Nazwa pola wyjściowego	Oczekiwana zawartość znacznika
Title	Znacznik ograniczający tytuł rekordu (opcjonalny).
Short Desc	Znacznik ograniczający krótki opis lub etykietę (opcjonalny).
Description	Znacznik ograniczający tekst główny. Jeśli to pole zostanie pozostawione puste, będzie zawierało całą pozostałą zawartość w znaczniku <body> (jeśli istnieje pojedynczy rekord) lub zawartość znalezioną w bieżącym rekordzie (jeśli określono ogranicznik rekordu).
Author	Znacznik ograniczający nazwisko autora tekstu (opcjonalny).
Contributors	Znacznik ograniczający nazwiska kontrybutorów (opcjonalny).
Published Date	Znacznik ograniczający datę opublikowania tekstu. Jeśli to pole pozostanie puste, wówczas będzie zawierało datę odczytu danych przez węzeł.
Modified Date	Znacznik ograniczający datę modyfikacji tekstu. Jeśli to pole pozostanie puste, wówczas będzie zawierało datę odczytu danych przez węzeł.

Jeśli użytkownik wprowadzi do tabeli znacznik, wówczas kanał informacyjny zostanie zeskanowany, przy czym ten znacznik będzie używany jako znacznik minimum, a nie jako znacznik dokładnej zgodności. Oznacza to, że w przypadku wprowadzenia znacznika do pola Title ten znacznik będzie zgodny z dowolnym znacznikiem w kanale informacyjnym, co obejmuje także znaczniki o określonych atrybutach (np.), przez co znacznik będzie równy znacznikowi głównemu ( ) oraz wszelkim znacznikom pochodnym, które zawierają atrybut i używają zawartości dla pola wyjściowego Title. Jeśli zostanie wprowadzony znacznik główny, wówczas wszelkie pozostałe atrybuty również zostaną uwzględnione.

Tabela 3. Przykłady znaczników HTML, które identyfikują tekst dla pól wyjściowych

Jeśli wprowadzisz:	Będzie zgodny z:	A także zgodny z:	Ale niezgodny z:
			każdym innym tagiem
<p class=”auth”>	<p class=”auth”>	<p color=”black” class=”auth” id=”85643”>	<p color=”black”>

Węzeł Web Feed: karta Content Filter

Karta Content Filter umożliwia stosowanie techniki filtrowania względem zawartości z kanału informacyjnego RSS. Ta karta nie ma zastosowania w przypadku kanałów informacyjnych HTML. Za pomocą tej karty można filtrować zawartość z kanału informacyjnego na podstawie tego, czy zawiera dużo tekstu w nagłówkach, stopkach, menu, w postaci reklam itp. Przy użyciu tej karty można usunąć z zawartości niepożądane znaczniki HTML, JavaScript, a także skrócić słowa lub wiersze z zawartości.

Content Filtering. Jeśli technika czyszczenia nie będzie używana, należy wybrać opcję **None**. W przeciwnym wypadku należy wybrać opcję **RSS Content Cleaner**.

Opcje RSS Content Cleaner. Jeśli zostanie wybrana opcja **RSS Content Cleaner**, wówczas można kasować wiersze na podstawie niektórych kryteriów. Wiersz jest ograniczony znacznikiem HTML, takim jak <p> i <li>, ale nie dotyczy to znaczników znajdujących się w wierszu, takich jak , <b> ani <font>. Należy zwrócić uwagę na to, że znaczniki
 są przetwarzane jako podziały wierszy.

- **Discard short lines**. Ta opcja powoduje ignorowanie wierszy zawierających zdefiniowaną w opcji minimalną liczbę słów (**minimum number of words**).
- **Discard lines with short words**. Ta opcja powoduje ignorowanie zawierających zdefiniowaną w opcji minimalną średnią liczbę słów (**minimum average word length**).
- **Discard lines with many single character words**. Ta opcja powoduje ignorowanie wierszy zawierających więcej słów jednoznakowych (**proportion of single character words**) niż ustalona proporcja.
- **Discard lines containing specific tags**. Ta opcja powoduje ignorowanie tekstu w wierszach, które zawierają dowolne znaczniki określone w polu.
- **Discard lines containing specific text**. Ta opcja powoduje ignorowanie wierszy, które zawierają dowolny tekst określony w polu.

Korzystanie z węzła Web Feed w eksploracji tekstu

Węzeł Web Feed może być używany do przygotowywania danych tekstowych z kanałów informacyjnych WWW na potrzeby eksploracji tekstu. Ten węzeł przyjmuje kanały informacyjne WWW w formacie HTML i RSS. Te kanały informacyjne pełnią rolę danych wejściowych dla procesu eksploracji tekstu (dla dalszego węzła Text Mining lub węzła Text Link Analysis).

W przypadku korzystania z węzła Web Feed należy w węźle Text Mining lub Text Link Analysis wybrać opcję Text field represents **actual text** w celu wskazania, że te kanały informacyjne są połączone z poszczególnymi artykułami lub wpisami w blogu.

Ważne! Jeśli informacje mają być pobierane z sieci WWW za pośrednictwem serwera proxy, należy w pliku net.properties włączyć serwer proxy zarówno dla klienta, jak i serwera IBM SPSS Modeler Text Analytics. Szczegółowe instrukcje postępowania znajdują się w samym pliku. Dotyczy to sytuacji, w której dostęp do sieci WWW odbywa się za pośrednictwem węzła Web Feed lub pobierana jest licencja na oprogramowanie SDL Software as a Service (SaaS), ponieważ odpowiednie połączenia przechodzą przez Java. Wspomniany plik domyślnie znajduje się w lokalizacji *C:\Program Files\IBM\SPSS\Modeler\18.1\jre\lib\net.properties*.

Przykład: węzeł Web Feed (kanał informacyjny RSS) z węzłem modelowania Text Mining

Załóżmy na przykład, że połączyliśmy węzeł Web Feed z węzłem Text Mining w celu przekazania danych tekstowych z kanału informacyjnego RSS do procesu eksploracji tekstu.

1. **Węzeł Web Feed (karta Input).** Najpierw dodaliśmy ten węzeł do strumienia, aby określić, gdzie znajduje się zawartość kanału informacyjnego i zweryfikować strukturę zawartości. Na pierwszej karcie udostępniliśmy adres URL do kanału informacyjnego RSS. Nasz przykład dotyczy kanału informacyjnego RSS, dlatego formatowanie jest już zdefiniowane i nie musimy wprowadzać żadnych dodatkowych zmian na karcie Records. W przypadku kanałów informacyjnych RSS dostępny jest opcjonalny algorytm filtrowania zawartości, jednak w tym przypadku nie był on stosowany.
2. **Węzeł Text Mining (karta Fields).** Następnie dodaliśmy i połączyliśmy węzeł Text Mining z węzłem Web Feed. Na tej karcie zdefiniowaliśmy wyjście pola tekstowego według węzła Web Feed. W tym przypadku chcieliśmy użyć pola **Description**. Ponadto wybraliśmy opcję Text field represents **actual text**, a także inne ustawienia.
3. **Węzeł Text Mining (karta Model).** Następnie na karcie Model wybraliśmy tryb budowania i zasoby. W tym przykładzie wybieramy zbudowanie modelu pojęć bezpośrednio z tego węzła przy użyciu domyślnego szablonu zasobów.

Więcej informacji o korzystaniu z węzła Text Mining zawiera sekcja “Węzeł modelowania Text Mining” na stronie 20.

Węzeł języka

Do identyfikacji języka naturalnego w polu tekstu w danych źródłowych można użyć węzła języka.

Wynikiem tego węzła jest pochodna zmienna, która zawiera wykryty kod języka.

Uwaga: Nie można użyć węzła języka do oceny w ramach w konfiguracji IBM SPSS Collaboration and Deployment Services - Scoring.

Węzeł ten znajduje się na karcie IBM SPSS Modeler Text Analytics palety węzłów w oknie IBM SPSS Modeler. Więcej informacji zawiera temat “Węzły produktu IBM SPSS Modeler Text Analytics” na stronie 8.

Węzeł języka: karta Settings

Na tej karcie można określić, w jaki sposób przedstawiać szczegóły języka dotyczące wybranej zmiennej tekstowej.

Text field Wybierz zmienną tekstową, dla której chcesz określić język.

Derive field name Wprowadź nazwę zmiennej pochodnej, która będzie zawierać kod wykrytego języka. Wartością domyślną jest *Language*.

Default value for when language cannot be identified Podaj nazwę zmiennej, która zostanie utworzona, jeśli język nie zostanie zidentyfikowany. Dostępne są następujące opcje:

- **Undefined** Jeśli wybrano tę opcję, wówczas zmienna pochodna nie zawiera żadnych wartości.
- **Supported** Jeśli wybrano tę opcję, wówczas można wybrać jeden z następujących obsługiwanych języków ISO:
  - Angielski (EN)
  - Niemiecki (DE)
  - Hiszpański (ES)
  - Francuski (FR)
  - Włoski (IT)
  - Japoński (JA)
  - Niderlandzki (NL)
  - Portugalski (PT)
- **Custom** Jeśli żaden z obsługiwanych języków nie jest odpowiedni, należy użyć tej opcji, aby określić, że powinna być używana wartość użytkownika. Zwykle mogą to być 2 litery kodu języka ISO, ale również dowolny łańcuch tekstowy, który jest wymagany.

Rozdział 3. Eksploracja w poszukiwaniu pojęć i kategorii

Węzeł Text Mining służy do generowania jednego z dwóch modeli użytkowych eksploracji tekstu:

- *Modele użytkowe pojęć* ujawniają i wyodrębniają pojęcia wyróżniające się wśród ustrukturyzowanych i nieustrukturyzowanych danych tekstowych.
- *Modele użytkowe kategorii* oceniają dokumenty oraz rekordy i przypisują je do kategorii, które są tworzone z wyodrębnionych pojęć (i wzorców).

Wyodrębnione pojęcia i wzorce, a także kategorie z modeli użytkowych, mogą być łączone z istniejącymi danymi ustrukturyzowanymi, takimi jak dane demograficzne, a następnie mogą być stosowane przy użyciu pełnego pakietu narzędzi IBM SPSS Modeler w celu podejmowania lepszych decyzji przy zmniejszonej ilości zakłóceń. Na przykład jeśli klienci często wymieniają problemy z logowaniem jako główne czynniki utrudniające wykonywanie zadań zarządzania kontami online, wówczas w modelach można uwzględnić "problemy z logowaniem".

Ponadto węzeł modelowania Text Mining jest w pełni zintegrowany w produkcie IBM SPSS Modeler, dzięki czemu strumienie eksploracji tekstu można wdrażać za pośrednictwem programu IBM SPSS Modeler Solution Publisher, który umożliwia ocenianie danych nieustrukturyzowanych w czasie rzeczywistym w aplikacjach, takich jak PredictiveCallCenter. Możliwość wdrażania tych strumieni pozwala na pomyślne implementacje eksploracji tekstu w zamkniętej pętli. Przykład: organizacja może analizować notatki z rozmów wychodzących i przychodzących, stosując modele predykcyjne w celu zwiększenia dokładności przekazu marketingowego w czasie rzeczywistym. Wykazano, że wykorzystywanie wyników modelu eksploracji tekstu w strumieniach poprawia dokładność predykcyjnych modeli danych.

Uwaga: Aby używać IBM SPSS Modeler Text Analytics z komponentem IBM SPSS Modeler Solution Publisher, dodaj katalog <katalog_instalacyjny>/ext/bin/spss.TMWBServer do zmiennej środowiskowej $LD_LIBRARY_PATH.

W produkcie IBM SPSS Modeler Text Analytics często odwołujemy się do wyodrębnionych pojęć i kategorii. Ważne jest zrozumienie znaczenia pojęć i kategorii, ponieważ mogą one ułatwiać podejmowanie bardziej świadomych decyzji podczas eksploracji i tworzenia modeli.

Pojęcia i modele użytkowe pojęć

Podczas procesu wyodrębniania dane tekstowe są skanowane i analizowane w celu identyfikowania interesujących lub istotnych pojedynczych słów, takich jak wybory lub pokój, a także fraz, takich jak wybory prezydenckie, wybory na prezydenta oraz traktaty pokojowe. Te słowa i frazy są zbiorczo określane jako *terminy*. Istotne terminy są wyodrębniane z wykorzystaniem zasobów lingwistycznych, a podobne terminy są grupowane pod terminem wiodącym, nazywanym **pojęciem**.

W ten sposób jedno pojęcie może reprezentować różne terminy w zależności od tekstu i zestawu wykorzystywanych zasobów językowych. Załóżmy na przykład, że mamy wyniki badania zadowolenia pracowników i wyodrębniliśmy pojęcie salary (wynagrodzenie). Załóżmy również, że przy wyszukiwaniu wszystkich rekordów związanych z pojęciem salary zauważyliśmy, że wyraz salary nie zawsze jest obecny w tekście — niektóre rekordy zawierały podobne terminy, takie jak wage, wages i salaries. Terminy te są zgrupowane pod pojęciem salary, ponieważ na podstawie reguł przetwarzania lub zasobów lingwistycznych mechanizm wyodrębniania uznał je za podobne lub stwierdził, że są synonimami. W tym przypadku wszystkie dokumenty lub rekordy zawierające którykolwiek z tych terminów będą traktowane, jak gdyby zawierały wyraz salary.

W celu sprawdzenia, jakie terminy są pogrupowane w ramach pojęcia, można eksplorować pojęcie w interaktywnym środowisku roboczym albo zapoznać się z tym, jakie synonimy zawiera model pojęcia. Więcej informacji zawiera temat "Terminy w modelach pojęć" na stronie 33.

Model użytkowy pojęcia zawiera zestaw pojęć, które mogą być używane w celu wskazywania rekordów lub dokumentów mogących również zawierać pojęcie (w tym dowolne z jego synonimów lub pogrupowanych terminów). Model pojęcia może być używany na dwa sposoby. Pierwszym jest eksploracja i analiza pojęć wykrytych w oryginalnym tekście źródłowym albo szybka identyfikacja interesujących dokumentów. Drugim jest zastosowanie tego modelu względem nowych rekordów lub dokumentów tekstowych w celu szybkiej identyfikacji tych samych pojęć kluczowych w nowych dokumentach/rekordach, np. wykrywanie w czasie rzeczywistym pojęć kluczowych w danych z notatników w centrum zgłoszeniowym.

Więcej informacji zawiera temat “Model użytkowy Text Mining: model pojęć” na stronie 30.

Kategorie i modele użytkowe kategorii

Możliwe jest tworzenie **kategorii**, które w gruncie rzeczy reprezentują bardziej ogólne pojęcia albo tematy przeznaczone do przechwytywania kluczowych idei, informacji i postaw wyrażonych w tekście. Każda kategoria obejmuje zestaw deskryptorów, takich jak *pojęcia*, *typy* i *reguły*. Razem te deskryptory służą do określania, czy konkretny rekord lub dokument należy do danej kategorii. Dokument lub rekord można przeskanować, aby sprawdzić, czy jakikolwiek fragment tekstu jest zgodny z deskryptorem. W przypadku znalezienia dopasowania dokument/rekord jest przypisywany do tej kategorii. Ten proces jest nazywany **klasyfikowaniem**.

Kategorie mogą być tworzone automatycznie przy wykorzystaniu dostępnego w produkcie, niezawodnego zestawu automatycznych technik; ręcznie z wykorzystaniem dodatkowych spostrzeżeń na temat danych albo przy użyciu obu tych metod. Możliwe jest także wczytanie zestawu wstępnie utworzonych kategorii z pakietu analizy tekstu, za pośrednictwem karty Model w tym węźle. Ręczne tworzenie kategorii lub korygowanie definicji kategorii jest możliwe tylko za pośrednictwem interaktywnego środowiska roboczego. Więcej informacji zawiera temat “Węzeł Text Mining: karta Model” na stronie 23.

Model użytkowy kategorii zawiera zestaw kategorii wraz z jego deskryptorami. Ten model może służyć do klasyfikowania zestawu dokumentów lub rekordów na podstawie tekstu zawartego w każdym z nich. Każdy dokument lub rekord jest czytany, a następnie przypisywany do każdej kategorii, dla której znalezione zostało dopasowanie z deskryptorem. W ten sposób dokument lub rekord może zostać przypisany do więcej niż jednej kategorii. Modele użytkowe kategorii mogą również służyć na przykład do zapoznawania się z kluczowymi ideami w odpowiedziach na otwarte pytania do ankiet albo w serii wpisów w blogu.

Więcej informacji zawiera temat “Model użytkowy Text Mining: model kategorii” na stronie 39.

Węzeł modelowania Text Mining

Węzeł Text Mining przy użyciu technik analizy lingwistycznej i analizy liczebności występowania wyodrębnia kluczowe pojęcia z tekstu i tworzy kategorie zawierające te pojęcia oraz inne dane. Węzeł ten można wykorzystać do eksploracji danych tekstowych lub do wygenerowania modelu użytkowego pojęć lub modelu użytkowego kategorii. Podczas wykonywania tego węzła modelowania wewnętrzny mechanizm lingwistyczny wyodrębnia i porządkuje pojęcia, wzorce i/lub kategorie, stosując metody przetwarzania języka naturalnego.

Korzystając z opcji **Generate directly**, można uruchomić węzeł Tex Mining, a po jego wykonaniu automatycznie wygenerować model użytkowy pojęć lub kategorii. Alternatywą jest zastosowanie bardziej interaktywnego, eksploracyjnego podejścia w trybie **Build interactively**, w którym można nie tylko wyodrębniać pojęcia, tworzyć kategorie i optymalizować zasoby lingwistyczne, lecz także prowadzić analizę powiązań w tekście i eksplorować skupienia. Więcej informacji zawiera temat “Węzeł Text Mining: karta Model” na stronie 23.

Węzeł ten znajduje się na karcie IBM SPSS Modeler Text Analytics palety węzłów w oknie IBM SPSS Modeler. Więcej informacji zawiera temat “Węzły produktu IBM SPSS Modeler Text Analytics” na stronie 8.

Wymagania. Węzły modelowania Text Mining przyjmują dane z węzła Web Feed, węzła File List i wszystkich standardowych węzłów źródłowych. Ten węzeł jest instalowany razem z produktem IBM SPSS Modeler Text Analytics i dostępny na palecie IBM SPSS Modeler Text Analytics.

Uwaga: Zastępuje on węzeł Text Extraction dla wszystkich użytkowników i węzeł Text Mining dla użytkowników japońskich, oferowany w poprzednich wersjach produktu Text Mining for Clementine. Jeśli masz starsze strumienie korzystające z tych węzłów lub modeli użytkowych, musisz przebudować je, stosując nowy węzeł Text Mining.

Węzeł Text Mining: karta Fields

Karta Fields służy do określenia ustawień dotyczących zmiennych zawierających dane, z których mają być wyodrębniane pojęcia. W przypadku pracy z obszernymi zbiorami danych warto rozważyć użycie przed tym węzłem węzła próby w celu skrócenia czasu przetwarzania. Więcej informacji zawiera temat “Wcześniejsze wybieranie próby w celu zaoszczędzenia czasu” na stronie 29.

Można ustawić następujące parametry:

ID field Wybierz zmienną zawierającą identyfikator rekordów tekstowych. Identyfikatory muszą być liczbami całkowitymi. Zmienna identyfikacyjna służy jako indeks dla poszczególnych rekordów tekstowych. Użyj zmiennej identyfikacyjnej, jeśli zmienna tekstowa zawiera tekst, który ma być eksplorowany.

Text field. Wybierz zmienną zawierającą tekst do eksploracji. Zmienna ta zależy od źródła danych.

Language field Wybierz zmienną zawierającą 2-literowy identyfikator języka ISO. Jeśli nie wybierzesz zmiennej, dla każdego dokumentu przyjęty zostanie język na podstawie dostarczonego szablonu.

Document type. Typ dokumentu określa strukturę tekstu. Wybierz jeden z następujących typów:

- **Full text.** Odpowiedni dla większości dokumentów i źródeł tekstowych. Podczas wyodrębniania przeglądany jest cały zbiór tekstu. Z tą opcją, w odróżnieniu od pozostałych, nie są związane żadne ustawienia dodatkowe.
- **Structured text.** Odpowiedni do analizy formularzy bibliograficznych, patentów i innych plików zawierających regularne struktury dające się rozpoznać i przeanalizować. Zastosowanie tego typu dokumentów wiąże się z pominięciem całości lub części procesu wyodrębniania. Umożliwia zdefiniowanie separatorów terminów, przypisywanie typów i określenie wymaganej minimalnej częstotliwości. W przypadku wybrania tej opcji należy kliknąć przycisk **Settings** i wprowadzić separatory tekstu w obszarze **Structured Text Formatting** okna dialogowego Document Settings. Więcej informacji zawiera temat “Ustawienia dokumentów na karcie Zmienne” na stronie 22.

Textual unity. Wybierz jeden z następujących trybów wyodrębniania:

- **Document mode.** Tryb odpowiedni do pracy z krótkimi i semantycznie jednorodnymi dokumentami, takimi jak artykuły prasowe.
- **Paragraph mode.** Odpowiedni do stron WWW i dokumentów bez znaczników. W procesie wyodrębniania dokumenty są semantycznie dzielone na podstawie takich cech wewnętrznych, jak wewnętrzne znaczniki i składnia. Wybranie tego trybu powoduje, że ocenianie odbywa się akapit po akapicie. Dlatego, na przykład, reguła apple & orange jest spełniona tylko wtedy, gdy apple i orange występują w tym samym akapicie.

  Uwaga: Z uwagi na sposób wyodrębniania tekstu z dokumentów PDF tryb **Paragraph mode** nie działa z takimi dokumentami. Wynika to z faktu, że podczas wyodrębniania pomijane są znaki powrotu karetki.

Paragraph mode settings. Ta opcja jest dostępna tylko wtedy, opcję zgodności tekstowej ustawiono na **Paragraph mode**. Określ progi liczby znaków obowiązujące we wszystkich procesach wyodrębniania. Rzeczywisty rozmiar jest zaokrąglany w górę lub w dół do najbliższej kropki. Aby mieć pewność, że asocjacje wyrazów wygenerowane z tekstu zawartego w zbiorze dokumentów są reprezentatywne, należy unikać określania zbyt małego rozmiaru wyodrębniania.

- **Minimum.** Określ minimalną liczbę znaków, jaka ma być używana w każdym wyodrębnianiu.
- **Maksimum.** Określ maksymalną liczbę znaków, jaka ma być używana w każdym wyodrębnianiu.

Partition mode Tryb dzielenia na podzbiory pozwala zdecydować, czy podział ma się odbywać na podstawie ustawień typu węzła, czy w inny sposób. Dzielenie na podzbiory polega na podzieleniu danych na próby uczące i testujące.

Ustawienia dokumentów na karcie Zmienne

Structured Text Formatting

Jeśli chcesz pominąć wszystkie lub niektóre etapy procesu wyodrębniania, ponieważ dane wejściowe są ustrukturyzowane lub chcesz określić reguły postępowania z tekstem, jako typ dokumentów wybierz **Structured text** i zadeklaruj pola lub znaczniki zawierające tekst w sekcji **Structured Text Formatting** okna dialogowego Document Settings. W procesie wyodrębniania terminy będą wywodzone wyłącznie z tekstu zawartego we wskazanych polach lub znacznikach (i ich znacznikach podrzędnych). Wszelkie niezadeklarowane pola i znaczniki będą ignorowane.

W pewnych kontekstach przetwarzanie lingwistyczne nie jest wymagane, a mechanizm wyodrębniania lingwistycznego można zastąpić jawnymi deklaracjami. W pliku bibliograficznym, w którym pola zawierające słowa kluczowe są oddzielone separatorami, takimi jak średniki (;) lub przecinki (,), wystarczy wyodrębnić łańcuch spomiędzy dwóch separatorów. Dlatego można zrezygnować z pełnego procesu wyodrębniania i zamiast niego zastosować specjalne reguły postępowania z tekstem, które będą obejmowały deklaracje separatorów między terminami, przypisywały typy do wyodrębnionego tekstu lub narzucały minimalną liczebność wystąpień przy wyodrębnianiu.

Deklarując elementy tekstu ustrukturyzowanego, należy przestrzegać następujących reguł:

- W jednym wierszu można zadeklarować tylko jedno pole, jeden znacznik lub jeden element. Zadeklarowane pola/znaczniki/elementy nie muszą być obecne w danych.
- W deklaracjach rozróżniana jest wielkość liter.
- Jeśli deklarujesz znacznik z atrybutami, na przykład <title id="1234">, i chcesz uwzględnić wszelkie warianty lub, w tym przypadku, wszystkie identyfikatory, dodaj znacznik bez atrybutu lub końcowego nawiasu (>), na przykład <title
- Dwukropek za nazwą pola lub znacznika sygnalizuje, że jest to tekst ustrukturyzowany. Dwukropek należy dopisać bezpośrednio za polem lub znacznikiem, ale przed ewentualnymi separatorami, typami lub wartościami częstości, na przykład author: lub <place>:.
- Aby wskazać, że pole lub znacznik zawiera wiele terminów rozdzielonych separatorem, zadeklaruj separator za dwukropkiem, na przykład: author:, lub <section>:;.
- Aby przypisać typ do zawartości znacznika, zadeklaruj nazwę typu za dwukropkiem i separatorem, na przykład: author:,Person lub <place>:;Location. Typy deklaruje się przy użyciu tych samych nazw, które używane są w oknie Resource Editor.
- Aby zdefiniować minimalną liczebność wystąpień pola lub znacznika, określ liczbę na końcu wiersza, na przykład: author:,Person1 lub <place>:;Location5. n jest liczebnością wystąpień, co oznacza, że termin znaleziony w polu lub znaczniku musi wystąpić co najmniej n razy w całym zbiorze dokumentów, aby został wyodrębniony. W takiej sytuacji wymagane jest także zdefiniowanie separatora.
- Jeśli znacznik zawiera dwukropek, to należy poprzedzić ten dwukropek ukośnikiem odwrotnym, aby deklaracja nie została zignorowana. Na przykład, jeśli jedno z pól nosi nazwę <topic:source>, to należy je zadeklarować jako <topic\:source>.

Dla celów ilustracyjnych załóżmy, że mamy w dokumentach następujące powtarzalne pola bibliograficzne:

```
author:Morel, Kawashima
abstract:W artykule opisano metody deklarowania pól.
publication:Dokumentacja eksploracji tekstu
datepub:March 2010
```

Gdybyśmy chcieli wyodrębniać tylko nazwiska autorów i streszczenia, ale ignorować resztę treści, moglibyśmy zadeklarować następujące pola:

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

Przykładowa deklaracja author:,Person1 oznacza, że zawartość pola nie powinna być przetwarzania lingwistycznie. Określa natomiast, że pole author zawiera więcej niż jedno nazwisko, nazwiska są oddzielone od siebie przecinkami i powinny być przypisywane do typu Person, a jeśli nazwisko występuje co najmniej raz w całym zbiorze dokumentów,

to powinno zostać wyodrębnione. Ponieważ pole abstract: jest wymienione bez żadnych dodatkowych deklaracji, to w procesie wyodrębniania będzie przeglądane i będzie podlegać standardowemu przetwarzaniu lingwistycznemu oraz przypisywaniu typów.

XML Text Formatting

Jeśli chcesz ograniczyć proces wyodrębniania tylko do tekstu wewnątrz określonych znaczników XML, użyj typu dokumentu **XML text** i zadeklaruj znaczniki zawierające tekst w sekcji **XML Text Formatting** okna dialogowego Document Settings. W procesie wyodrębniania terminy będą wywodzone wyłącznie z tekstu zawartego we wskazanych znacznikach i ich znacznikach podrzędnych.

Ważne! Jeśli chcesz pominąć proces wyodrębniania i określić reguły dotyczące separatorów między terminami, przypisać typy do wyodrębnianych terminów lub określić wymaganą minimalną liczebność wyodrębnianych terminów, użyj opcji **Structured text** opisanej poniżej.

Deklarując znaczniki w celu określenia formatu tekstu XML, należy przestrzegać następujących reguł:

- W jednym wierszu można zadeklarować tylko jeden znacznik XML.
- W elementach znaczników rozróżniana jest wielkość liter.
- Jeśli znacznik ma atrybuty, na przykład <title id="1234">, i chcesz uwzględnić wszelkie warianty lub, w tym przypadku, wszystkie identyfikatory, dodaj znacznik bez atrybutu lub końcowego nawiasu (>), na przykład <title

Dla celów ilustracyjnych załóżmy, że mamy następujący dokument XML:

```
<section>Kodeks drogowy
     <title id="01234">Znaki drogowe</title>
     <p>Znaki drogowe są przydatne.</p>
</section>
<p>Znajomość przepisów jest istotna.</p>
```

Na potrzeby przykładu zadeklarujemy następujące znaczniki:

```
<section>
<title
```

W tym przykładzie, ponieważ zadeklarowaliśmy znacznik <section>, tekst w tym znaczniku oraz znacznikach w nim zagnieżdżonych Znaki drogowe i Znaki drogowe są przydatne, będzie przeglądany w procesie wyodrębniania tekstu. Jednak tekst Znajomość przepisów jest istotna będzie ignorowany, bo znacznik <p> nie był zadeklarowany jawnie i nie jest zagnieżdżony w żadnym ze znaczników zadeklarowanych.

Węzeł Text Mining: karta Model

Na karcie Model określa się metodę budowania i ogólne ustawienia dotyczące wyników wykonania węzła.

Można ustawić następujące parametry:

Model name. Można wygenerować nazwę modelu, która będzie automatycznie oparta na zmiennej przewidywanej lub identyfikacyjnej (albo na typie modelu w przypadkach, gdy żadna taka zmienna nie jest określona), albo określić nazwę niestandardową.

Użyj danych podzielonych na podzbiory. Jeśli zdefiniowano zmienną dzielącą na podzbiory, ta opcja umożliwia użycie podczas budowania modelu wyłącznie danych z podzbioru uczącego.

Build mode Określa sposób generowania modeli użytkowych podczas wykonywania strumienia z węzłem Text Mining. Alternatywą jest zastosowanie bardziej interaktywnego, eksploracyjnego podejścia w trybie **Build interactively**, w którym można nie tylko wyodrębniać pojęcia, tworzyć kategorie i optymalizować zasoby lingwistyczne, lecz także prowadzić analizę powiązań w tekście i eksplorować skupienia.

- **Build interactively** Ta opcja powoduje, że podczas wykonywania strumienia uruchamiany jest interaktywny interfejs, w którym można wyodrębniać pojęcia i wzorce, eksplorować i optymalizować wyodrębnione wyniki, budować i optymalizować kategorie, optymalizować zasoby lingwistyczne (szablony, synonimy, typy, biblioteki itd.)oraz budować modele użytkowe kategorii. Więcej informacji zawiera temat “Build Interactively”.
- **Generate directly** Ta opcja powoduje, że podczas wykonywania strumienia automatycznie tworzony jest model. Utworzony model dodawany jest do palety Modele. Inaczej niż w przypadku pulpitu interaktywnego, tutaj w czasie wykonywania nie są potrzebne żadne dodatkowe działania użytkownika — wystarczy zdefiniować ustawienia w węźle. Po wybraniu tej opcji widoczne staną się opcje charakterystyczne dla modelu, za pośrednictwem których można określić typ modelu, który ma zostać wygenerowany. Więcej informacji zawiera temat “Generate Directly” na stronie 25.

Store large models in AS Jeśli istnieje połączenie z serwerem IBM SPSS Analytic Server, zaznacz tę opcję, aby zapisać modele zdalnie na serwerze.

Uwaga: Każdy model, który jest tworzony i przechowywany na serwerze, można oceniać tylko na danym serwerze. Aby wznowić sesję interaktywnego środowiska roboczego zawierającą taki model, niezbędne jest połączenie z oryginalnym serwerem, który został użyty do utworzenia sesji.

Copy resources from Podczas eksploracji tekstu wyodrębnianie prowadzone jest nie tylko na podstawie ustawień na karcie Expert, lecz również na podstawie zasobów lingwistycznych. Zasoby te są podstawą dla przetwarzania tekstu w trakcie wyodrębniania w celu uzyskania z niego pojęć, typów, a niekiedy także wzorców. Można skopiować zasoby do tego węzła z szablonu zasobów albo z pakietu analizy tekstu (TAP). Wybierz jeden z tych elementów i kliknij przycisk **Load**, aby określić pakiet lub szablon, z którego zostaną skopiowane zasoby. Podczas ładowania kopia zasobów zapisywana jest w węźle. Jeśli zatem w przyszłości zechcesz użyć zmienionego szablonu lub pakietu TAP, musisz go ponownie załadować — tutaj lub w sesji na interaktywnym pulpicie roboczym. Dla ułatwienia w węźle widoczna jest data i godzina skopiowania i załadowania zasobów. Więcej informacji zawiera temat “Kopiowanie zasobów z szablonów i pakietów TAP” na stronie 26.

Text language. Określa język tekstu poddawanego eksploracji. Zasoby skopiowane w węźle sterują wyświetlanymi opcjami dotyczącymi języka. Wybierz język, w którym zasoby zostały dostosowane.

Build Interactively

Na karcie Model węzła modelowania Text Mining można wybrać węzeł budowania dla modeli użytkowych. Wybranie opcji **Build interactively** spowoduje otwarcie interaktywnego interfejsu podczas wykonywania strumienia. Na takim interaktywnym pulpicie roboczym można:

- Wyodrębniać elementy i eksplorować wyniki eksploracji, w tym pojęcia i typy, aby odkrywać wydatne tendencje, koncepcje i trendy w danych tekstowych.
- Stosować różne metody budowania i uzupełniania kategorii na podstawie pojęć, typów, wzorców TLA i reguł, aby później możliwe było ocenianie dokumentów i rekordów w oparciu o te kategorie.
- Optymalizować zasoby lingwistyczne (szablony zasobów, biblioteki, słowniki, synonimy i inne zasoby), aby uzyskiwać coraz lepsze wyniki w procesie iteracyjnego wyodrębniania, analizowania i optymalizacji pojęć.
- Prowadzić analizę powiązań w tekście (TLA — text link analysis) i używać ujawnionych wzorców TLA do budowania lepszych modeli użytkowych kategorii. Węzeł Text Link Analysis nie oferuje takich opcji eksploracji ani możliwości modelowania.
- Generować skupienia w celu ujawniania nowych relacji i eksplorować relacje między pojęciami, typami, wzorcami i kategoriami w panelu Visualization.
- Generować zoptymalizowane modele użytkowe kategorii na palecie Modele w programie IBM SPSS Modeler i używać ich w innych strumieniach.

Uwaga: Nie można zbudować modelu interaktywnego w przypadku tworzenia zadania IBM SPSS Collaboration and Deployment Services.

Użyj sesji (kategorii, TLA, zasobów itp.) z ostatniej aktualizacji węzła. Podczas pracy z interaktywnym pulpitem roboczym można aktualizować węzeł danymi sesji (parametrami wyodrębniania, zasobami, definicjami kategorii itd.).

Opcja **Use session work** umożliwia ponowne uruchomienie interaktywnego pulpitu z zapisanymi danymi sesji. Ta opcja jest nieaktywna przy pierwszym użyciu węzła, ponieważ nie mogły być jeszcze zapisane żadne dane sesji. Aby dowiedzieć się, jak zaktualizować węzeł danymi sesji, by można było korzystać z tej opcji, patrz “Aktualizowanie węzłów modelowania i zapisywanie” na stronie 77.

Jeśli uruchomisz sesję *z tą opcją*, to ustawienia wyodrębniania, kategorie, zasoby i inne wyniki pracy z ostatniej sesji interaktywnego pulpitu zapisane w węźle będą dostępne w nowej sesji. Ponieważ ta opcja powoduje użycie zapisanych danych sesji, niektóre dane, takie jak zasoby skopiowane z szablonu i inne karty, są wyłączone i ignorowane. Jeśli jednak uruchomisz sesję *bez* tej opcji, to używana będzie zawartość węzła według bieżącego stanu, zatem praca wykonana wcześniej na pulpicie będzie niedostępna.

Uwaga: Jeśli zmienisz węzeł źródłowy strumienia już po zbuforowaniu wyników wyodrębniania za pomocą opcji **Use session work...**, trzeba będzie uruchomić nowe wyodrębnianie po uruchomieniu sesji interaktywnego pulpitu roboczego, aby móc korzystać ze zaktualizowanych wyników wyodrębniania.

Skip extraction and reuse cached data and results. Można ponownie wykorzystać wszelkie zbuforowane wyniki wyodrębniania i dane w sesji pracy z interaktywnym pulpitem roboczym. Ta opcja jest szczególnie użyteczna, gdy dla zaoszczędzenia czasu chcesz ponownie wykorzystać posiadane wyniki wyodrębniania, zamiast czekać na zakończenie nowego wyodrębniania. Aby można było użyć tej opcji, węzeł powinien być wcześniej zaktualizowany danymi z sesji pracy z interaktywnym pulpitem roboczym, a opcja **Keep the session work and cache text data with extraction results for reuse** musi być wybrana. Aby dowiedzieć się, jak zaktualizować węzeł danymi sesji, by można było korzystać z tej opcji, patrz “Aktualizowanie węzłów modelowania i zapisywanie” na stronie 77.

Begin session by. Wybierz opcję określającą widok i działanie inicjowane zaraz po uruchomieniu sesji pracy z interaktywnym pulpitem roboczym. Niezależnie od wybranego widoku początkowego można w trakcie sesji przełączać się na inne widoki.

- **Using extraction results to build categories.** Ta opcja powoduje uruchomienie interaktywnego pulpitu roboczego w widoku Categories i Concepts oraz, w razie potrzeby, przeprowadzenie wyodrębnienia. W tym widoku można tworzyć kategorie i wygenerować model kategorii. Można także przełączyć się na inny widok. Więcej informacji zawiera temat Rozdział 7, “Tryb pracy z interaktywnym pulpitem roboczym”, na stronie 67.
- **Exploring text link analysis (TLA) results.** Ta opcja powoduje, że po uruchomieniu pulpitu następuje wyodrębnienie i rozpoznanie relacji między pojęciami w tekście, na przykład opinii lub innych powiązań, w widoku Text Link Analysis. Aby móc korzystać z tej opcji i uzyskać wyniki, należy wybrać szablon lub pakiet analizy tekstu zawierający reguły wzorców TLA. Wyodrębnianie TLA z większych zbiorów danych może być czasochłonne. W takich sytuacjach warto rozważyć zastosowanie wcześniej węzła próby. Więcej informacji zawiera Rozdział 11, “Eksplorowanie analizy powiązań w tekście”, na stronie 141.
- **Analyzing co-word clusters.** Ta opcja powoduje uruchomienie pulpitu w widoku Clusters i zaktualizowanie wszelkich nieaktualnych wyników wyodrębniania. W widoku tym można przeprowadzić analizę skupień współwystępujących wyrazów, która wygeneruje zestaw skupień. Analiza skupień współwystępujących wyrazów to proces, który rozpoczyna się od analizy siły powiązania między dwoma pojęciami na podstawie ich współwystępowania w danym rekordzie lub dokumencie, a kończy się na grupowaniu silnie powiązanych pojęć w skupienia. Więcej informacji zawiera temat Rozdział 7, “Tryb pracy z interaktywnym pulpitem roboczym”, na stronie 67.

Generate Directly

Na karcie Model węzła modelowania Text Mining można wybrać węzeł budowania dla modeli użytkowych. Po wybraniu opcji **Generate directly** możesz ustawić opcje węzła, a potem po prostu uruchomić strumień. Wynikiem będzie model użytkowy pojęć, który zostanie od razu umieszczony na palecie Modele. Inaczej niż w przypadku pulpitu interaktywnego, tutaj w czasie wykonywania nie są potrzebne żadne dodatkowe działania użytkownika — wystarczy zdefiniować ustawienia liczebności w węźle.

Maximum number of concepts to include in model. Ta opcja, która ma zastosowanie wyłącznie przy automatycznym (nieinteraktywnym) budowaniu modelu, wskazuje, że ma zostać utworzony model pojęć. Określa także, że liczba pojęć w tym modelu nie powinna przekraczać podanej liczby maksymalnej.

- **Check concepts based on highest frequency. Top number of concepts.** Jest to liczba pojęć, które zostaną zaznaczone począwszy od pojęcia o największej częstości (liczebności wystąpień). W tym miejscu częstość oznacza liczbę wystąpień pojęcia (i wszystkich związanych z nim terminów) w całym zbiorze dokumentów/rekordów. Ta liczba może być wyższa od liczby rekordów, ponieważ jedno pojęcie może występować wielokrotnie w tym samym rekordzie.
- **Uncheck concepts that occur in too many records. Percentage of records.** Usuwa zaznaczenie pojęć, których wyrażona procentowo liczba wystąpień w rekordach jest większa od określonej. Ta opcja jest użyteczna do wykluczania pojęć, które występują często w tekście lub w każdym rekordzie, ale nie mają znaczenia w prowadzonej analizie.

Optimize for speed of scoring. Opcja ta jest domyślnie wybrana i powoduje, że utworzony model będzie zwarty i będzie szybko oceniał dane. Usunięcie zaznaczenia tej opcji spowoduje utworzenie znacznie większego modelu, który prowadzi ocenę znacznie wolniej. Jednak większy model gwarantuje, że oceny wyświetlone początkowo w wygenerowanym modelu pojęć będą takie same, jak oceny uzyskane w wyniku analizy tego samego tekstu za pomocą modelu użytkowego.

Kopiowanie zasobów z szablonów i pakietów TAP

Podczas eksploracji tekstu wyodrębnianie prowadzone jest nie tylko na podstawie ustawień na karcie Expert, lecz również na podstawie zasobów lingwistycznych. Zasoby te są podstawą dla przetwarzania tekstu w trakcie wyodrębniania w celu uzyskania z niego pojęć, typów, a niekiedy także wzorców. Można skopiować zasoby do tego węzła z *szablonu zasobów*, a pracując w węźle Text Mining, można także wybrać *pakiet analizy tekstu* (TAP).

Domyślnie w momencie umieszczania węzła w obszarze roboczym zasoby są kopiowane do niego z podstawowego szablonu właściwego dla języka objętego licencją na produkt. Jeśli masz licencje na wiele języków, szablon do automatycznego załadowania wybierany jest na podstawie pierwszego wybranego języka.

Podczas ładowania kopia wybranych zasobów zapisywana jest w węźle. Kopiowana jest jedynie zawartość szablonu lub pakietu TAP, nie jest natomiast tworzone powiązanie między szablonem lub pakietem TAP a węzłem. Oznacza to, że jeśli szablon lub pakiet TAP zostanie później zmieniony, to zmiany te nie będą automatycznie dostępne w węźle. Krótko mówiąc, zawsze używane są zasoby załadowane do węzła, chyba że ponownie załadujesz kopię szablonu lub pakietu TAP lub zaktualizujesz węzeł Text Mining i wybierzesz opcję **Use session work**. Więcej informacji można znaleźć w sekcji **Korzystanie z danych z sesji** w dalszej części tego tematu.

Wybieraj szablon lub pakiet TAP właściwy dla języka, w którym zapisane są dane tekstowe. Można używać tylko szablonów lub pakietów TAP właściwych dla języków, na które masz licencję. Jeśli chcesz analizować powiązania w tekście, musisz wybrać szablon zawierający wzorce TLA. Jeśli szablon zawiera wzorce TLA, w kolumnie TLA okna dialogowego Load Resource Template będzie widoczna odpowiednia ikona.

Uwaga: Nie można ładować pakietów TAP do węzła Text Link Analysis.

Szablony zasobów

Szablon zasobów to predefiniowany zestaw bibliotek i zaawansowanych zasobów lingwistycznych oraz nielingwistycznych, które zostały zoptymalizowane pod kątem konkretnej dziedziny lub zastosowania. Już w momencie umieszczania węzła modelowania Text Mining w obszarze roboczym strumienia jest do niego ładowana kopia zasobów z podstawowego szablonu, ale można zmienić szablon lub załadować pakiet analizy tekstu, wybierając opcję **Resource template** albo **Text analysis package** i klikając przycisk **Load**. W przypadku szablonów można następnie wybrać szablon w oknie dialogowym Load Resource Template.

Uwaga: Jeżeli żądany szablon nie figuruje na liście, ale masz na zapisaną lokalnie w komputerze jego wyeksportowaną kopię, to możesz ją teraz zaimportować. Możesz także wyeksportować szablon z tego okna dialogowego, aby udostępnić go innym użytkownikom. Więcej informacji zawiera temat “Importowanie i eksportowanie szablonów” na stronie 164.

Pakiety analizy tekstu (TAP)

Pakiet analizy tekstu (TAP — text analysis package) to predefiniowany zestaw bibliotek oraz zaawansowanych zasobów lingwistycznych i nielingwistycznych połączony w pakiet z jednym lub wieloma zestawami predefiniowanych kategorii. IBM SPSS Modeler Text Analytics oferuje kilka gotowych pakietów TAP dla tekstów w języku angielskim oraz japońskim, a każdy z tych pakietów jest zoptymalizowany pod kątem konkretnej dziedziny. Nie można edytować tych pakietów TAP, ale można użyć ich jako punktu wyjścia do budowania własnego modelu kategorii. Można także tworzyć własne pakiety TAP w sesji interaktywnej. Więcej informacji zawiera temat "Ładowanie pakietów analizy tekstu" na stronie 131.

Uwaga: Nie można ładować pakietów TAP do węzła Text Link Analysis.

Korzystanie z opcji „Use Session Work" (kata Model)

Wprawdzie zasoby są kopiowane do węzła na karcie Model, ale można również później wprowadzać interaktywnie zmiany w zasobach i ewentualnie zaktualizować węzeł modelowania z uwzględnieniem tych zmian. W takim przypadku należałoby wybrać opcję **Use session work** na karcie Model węzła modelowania Text Mining.

Wybranie opcji **Use session work** powoduje wyłączenie przycisku **Load** w oknie węzła, co sygnalizuje, że zasoby z interaktywnego pulpitu roboczego zostaną użyte zamiast zasobów załadowanych wcześniej do węzła.

Aby wprowadzić zmiany w zasobach po wybraniu opcji **Use session work**, można edytować lub podmieniać zasoby bezpośrednio w sesji pracy z interaktywnym pulpitem roboczym, korzystając z widoku Resource Editor. Więcej informacji zawiera temat "Aktualizowanie zasobów węzła po załadowaniu" na stronie 163.

Węzeł Text Mining: karta Expert

Karta Expert zawiera pewne zaawansowane parametry wpływające na sposób wyodrębniania tekstu i postępowania z nim. Parametry w tym oknie dialogowym sterują podstawowym przebiegiem, a także kilkoma funkcjami zaawansowanymi procesu wyodrębniania. Jednak stanowią tylko część wszystkich opcji dostępnych dla użytkownika. Istnieje także szereg zasobów lingwistycznych i opcji wpływających na wyniki wyodrębniania i sterowanych za pośrednictwem szablonu zasobów wybieranego na karcie Model. Więcej informacji zawiera temat "Węzeł Text Mining: karta Model" na stronie 23.

Uwaga: Cała ta karta jest wyłączona, jeśli na karcie Model wybrano tryb **Build interactively** i korzystanie z zapisanych informacji z interaktywnego pulpitu roboczego. Wówczas obowiązują ustawienia wyodrębniania z ostatniej zapisanej sesji pracy z pulpitem.

Informacje dotyczące języka angielskiego, francuskiego, hiszpańskiego, holenderskiego, niemieckiego, portugalskiego i włoskiego

Można ustawić następujące parametry wyodrębniania z tekstów w językach innych niż japoński, np. angielskim, hiszpańskim, francuskim, niemieckim itd.:

Uwaga: W dalszej części tego tematu znajdują się informacje o ustawieniach zaawansowanych dotyczących tekstu japońskiego.

Limit extraction to concepts with a global frequency of at least [n]. Określa, ile razy wyraz lub fraza musi wystąpić w tekście, aby został(a) wyodrębniony/-a. Na przykład wartość 5 powoduje, że wyodrębnianie będą tylko te wyrazy i frazy, które występują co najmniej pięć razy w całym zbiorze rekordów lub dokumentów.

W niektórych przypadkach zmiana tego limitu silnie wpływa na wyniki wyodrębniania, a w efekcie na utworzone kategorie. Załóżmy, że pracujemy nad danymi o sieci restauracji i opisywany limit jest ustawiony na 1. W tym przypadku w wynikach wyodrębniania mogą znaleźć się pojęcia: *pizza (1), thin pizza (2), spinach pizza (2)* i *favorite pizza (2)*. Jeśli jednak ograniczymy wyodrębnianie do pojęć występujących co najmniej 5 razy, powyższe trzy pojęcia nie zostaną wyodrębnione. W wynikach uzyskamy natomiast pojęcie *pizza (7)*, ponieważ *pizza* jest najprostszą formą, a wyraz ten istniał już jako potencjalny kandydat. W zależności od zawartości reszty tekstu liczność wystąpień może przekraczać siedem, jeśli w tekście wyraz pizza występuje jeszcze w innych frazach. Ponadto, jeśli pojęcie *spinach*

pizza było już deskryptorem kategorii, to zamiast niego może być konieczne dodanie deskryptora *pizza* w celu uwzględnienia wszystkich rekordów. Dlatego, jeśli kategorie zostały już wcześniej utworzone, limit wystąpień należy modyfikować ostrożnie.

Należy zauważyć, że opisywana opcja dotyczy tylko wyodrębniania; jeśli szablon zawiera terminy (a zwykle zawiera), a termin z szablonu zostanie znaleziony w tekście, to termin ten będzie indeksowany niezależnie od liczebności wystąpień.

Załóżmy na przykład, że szablon Basic Resources zawiera termin "los angeles" typu <Location> w bibliotece Core; jeśli dokument zawiera termin Los Angeles tylko raz, to Los Angeles zawsze znajdzie się na liście pojęć. Aby temu zapobiec, należy ustawić filtr w celu uwidocznienia pojęć występujących co najmniej tyle razy, ile wynosi limit określony w polu **Limit extraction to concepts with a global frequency of at least [n]**.

Accommodate punctuation errors. Ta opcja powoduje, że podczas wyodrębniania tekst zawierający błędy interpunkcyjne (na przykład nieprawidłowo użyte znaki interpunkcyjne) będzie tymczasowo normalizowany w celu poprawienia efektywności wyodrębniania pojęć. Ta opcja jest bardzo użyteczna, gdy mamy do czynienia z krótkimi tekstami o niskiej jakości (np. odpowiedziami na pytania otwarte w ankietach, wiadomościami e-mail i danymi z systemów CRM) lub gdy system zawiera wiele skrótów.

Accommodate spelling for a minimum word character length of [n] Ta opcja powoduje zastosowanie techniki grupowania rozmytego, która grupuje błędnie napisane lub podobne wyrazy pod jednym pojęciem. Algorytm grupowania rozmytego tymczasowo usuwa wszystkie samogłoski (z wyjątkiem pierwszej) oraz podwójne/potrójne spółgłoski z wyodrębnianych wyrazów, a następnie porównuje wyniki, by sprawdzić, czy są identyczne. Zatem wyrazy modeling i modelling zostałyby połączone w jedną grupę. Jeśli jednak każdy termin ma przypisany inny typ, z wyjątkiem typu <Unknown>, to grupowanie rozmyte nie będzie stosowane.

Można też określić minimalną liczbę znaków *rdzennych* wymaganą do zastosowania grupowania rozmytego. Liczba znaków rdzennych w terminie obliczana jest poprzez zsumowanie wszystkich znaków i odjęcie znaków tworzących przyrostki przy odmianie, a w wypadku terminów będących wyrazami złożonymi, także znaków tworzących określniki i przyimki. Na przykład termin exercises ma 8 znaków rdzennych w swojej postaci “exercise”, ponieważ litera *s* na końcu tworzy odmianę (w tym przypadku liczbę mnogą). Podobnie, apple sauce ma 10 znaków rdzennych (“apple sauce”), a manufacturing of cars ma 16 znaków rdzennych (“manufacturing car”). Ta metoda liczenia znaków jest stosowana tylko do sprawdzania, czy ma być przeprowadzane grupowanie rozmyte, ale nie wpływa na sposób dopasowywania wyrazów.

Uwaga: Jeśli później okaże się, że pewne wyrazy są grupowane nieprawidłowo, można wykluczyć konkretne pary wyrazów ze stosowania tej techniki, jawnie deklarując je w sekcji **Fuzzy Grouping: Exceptions** na karcie Advanced Resources. Więcej informacji zawiera temat “Grupowanie rozmyte” na stronie 191.

Extract uniterms Ta opcja wyodrębnia pojedyncze wyrazy (terminy pojedyncze), o ile tylko nie są już częścią wyrazu złożonego i są albo rzeczownikami, albo nierozpoznanymi częściami mowy.

Extract nonlinguistic entities Ta opcja wyodrębnia Obiekty nielingwistyczne, takie jak numery telefonów, numery ubezpieczenia społecznego, godziny, daty, waluty, cyfry, wartości procentowe, adresy e-mail i adresy HTTP. W sekcji **Nonlinguistic Entities: Configuration** karty Advanced Resources można uwzględniać i wykluczać określone typy obiektów nielingwistycznych. Wykluczenie zbędnych obiektów sprawi, że mechanizm wyodrębniania nie będzie marnował czasu na ich przetwarzanie. Więcej informacji zawiera temat “Konfiguracja” na stronie 195.

Uppercase algorithm Ta opcja wyodrębnia proste i złożone terminy, które nie figurują we wbudowanych słownikach, o ile pierwsza litera terminu jest wielka. Jest to dobry sposób na wyodrębnienie większości rzeczowników własnych.

Group partial and full person names together when possible Ta opcja grupuje imiona i nazwiska, które w tekście występują w różnych postaciach. Jest to użyteczne, ponieważ imiona i nazwiska często na początku tekstu przytaczane są w pełnym brzmieniu, ale później już występują tylko w wersji skróconej. W przypadku wybrania tej opcji program próbuje dopasować każdy pojedynczy termin typu <Unknown> do ostatniego wyrazu każdego terminu złożonego typu <Person> (osoba). Na przykład, jeśli znaleziony zostanie wyraz *nowak* o początkowo przypisanym typie

<Unknown>, to mechanizm wyodrębniania sprawdzi, czy jakiekolwiek terminy złożone typu <Person> zawierają jako ostatni wyraz właśnie *nowak*, na przykład *piotr nowak*. Ta opcja nie ma zastosowania do imion, ponieważ większość z nich nigdy nie jest wyodrębniana jako termin pojedynczy.

Maximum nonfunction word permutation Ta opcja określa maksymalną liczbę wyrazów niefunkcyjnych, które mogą być obecne, gdy stosowana jest technika permutacji. Technika permutacji grupuje podobne frazy różniące się tylko wyrazami niefunkcyjnymi (na przykład „of" lub „the"), niezależnie od odmiany. Załóżmy na przykład, że wartość ta jest ustawiona na maksymalnie dwa wyrazy i wyodrębniono zarówno termin company officials, jak i termin officials of the company. W tym przypadku oba terminy zostaną połączone w grupę, ponieważ po zignorowaniu wyrazów of the zostaną uznane za identyczne.

Use derivation when grouping multiterms Podczas przetwarzania wielkich zbiorów danych wybierz tę opcję, aby grupować terminy wielowyrazowe według reguł derywacji.

Uwaga: Aby możliwy był eksport wyników analizy powiązań w tekście, należy rozpocząć sesję z opcją **Exploring text link analysis results** i wybrać zasoby zawierające definicje TLA. Zawsze można wyodrębnić wyniki analizy TLA później, podczas sesji pracy z interaktywnym pulpitem roboczym, za pośrednictwem okna dialogowego Extraction Settings. Więcej informacji zawiera temat "Wyodrębnianie danych" na stronie 80.

Dla tekstu japońskiego

Opcje w tym oknie dotyczące języka japońskiego są inne niż dla pozostałych języków, co wynika z różnic w procesach wyodrębniania. Aby pracować z tekstem japońskim, należy ponadto na karcie Model tego węzła wybrać szablon lub pakiet analizy tekstu zoptymalizowany dla języka japońskiego. Więcej informacji zawiera temat "Kopiowanie zasobów z szablonów i pakietów TAP" na stronie 26.

Secondary Analysis. Po rozpoczęciu procesu wyodrębniania następuje wyodrębnienie podstawowych słów kluczowych na podstawie domyślnego zestawu typów. Gdy jednak wybierzesz dodatkowy analizator, możesz uzyskać większą liczbę pojęć lub pogłębione pojęcia, ponieważ mechanizm wyodrębniający będzie uwzględniał partykuły i czasowniki pomocnicze jako części pojęcia. W wynikach analizy sentymentu zostanie uwzględnionych wiele dodatkowych typów. Co więcej, wybranie dodatkowego analizatora umożliwi także wygenerowanie wyników analizy powiązań w tekście.

Uwaga: Użycie dodatkowego analizatora wydłuża proces wyodrębniania.

- **Dependency analysis**. Wybranie tej opcji powoduje uwzględnienie dodatkowych partykuł przy wyodrębnianiu podstawowych typów i słów kluczowych. Pozwala także uzyskać wzbogacone wzorce przy analizie powiązań w tekście (TLA).
- **Sentiment analysis**. Wybranie tego analizatora pozwala wyodrębnić dodatkowe pojęcia oraz, tam gdzie to możliwe, wyodrębnić wzorce powiązań TLA. Oprócz typów podstawowych można też korzystać z ponad 80 typów sentymentu. Te typy służą do ujawniania pojęć i wzorców w treściach wyrażających emocje, odczucia i opinie. Dostępne są trzy opcje, które służą do odpowiedniego ukierunkowania analizy sentymentu: **All sentiments**, **Representative sentiment only** i **Conclusions only**.
- **No secondary analyzer** Ta opcja wyłącza wszystkie dodatkowe analizatory. Jest ona ukryta, jeśli na karcie Model wybrano opcję **Exploring text link analysis (TLA) results**, ponieważ do uzyskania wyników TLA niezbędny jest dodatkowy analizator. Jeśli najpierw wybierzesz tę opcję, ale później wybierzesz opcję **Exploring text link analysis (TLA) results**, to w trakcie wykonywania strumienia zostanie zgłoszony błąd.

Wcześniejsze wybieranie próby w celu zaoszczędzenia czasu

Gdy dane mają dużą objętość, przetwarzanie może trwać wiele minut, a nawet godzin — zwłaszcza w trakcie sesji pracy z interaktywnym pulpitem roboczym. Im większa objętość danych, tym dłużej będą trwały procesy wyodrębniania i kategoryzacji. Aby pracować efektywniej, możesz przed węzłem Text Mining dodać węzły próby programu IBM SPSS Modeler. Węzeł próby pobiera losową próbę i umożliwia korzystanie z mniejszego podzbioru dokumentów lub rekordów podczas kilku pierwszych przebiegów.

Mniejsza próba jest często całkowicie wystarczająca do podjęcia decyzji o ewentualnych korektach zasobów, a nawet do utworzenia większości lub wszystkich kategorii. Gdy analiza mniejszego zbioru danych przyniesie zadowalające wyniki, można zastosować tę samą technikę do utworzenia kategorii na podstawie całego zbioru danych. Następnie można wyszukać dokumenty lub rekordy, które nie pasują do utworzonych kategorii, i wprowadzić niezbędne korekty.

Uwaga: Węzeł próby jest standardowym węzłem programu IBM SPSS Modeler.

Korzystanie z węzła Text Mining w strumieniu

Węzeł modelowania Test Mining służy do uzyskiwania dostępu do danych i wyodrębniania pojęć w ramach strumienia. Aby uzyskać dostęp do danych, można wykorzystać dowolny węzeł źródłowy, na przykład węzeł Baza danych, węzeł Plik zmiennych, węzeł Web Feed lub węzeł Plik kolumnowy. W przypadku tekstu, który znajduje się w dokumentach zewnętrznych, można użyć węzła File List.

Przykład 1: Użycie węzłów File List i Text Mining do bezpośredniego tworzenia modelu użytkowego pojęć

Poniższy przykład ilustruje zastosowanie węzła File List razem z węzłem modelowania Text Mining do wygenerowania modelu użytkowego pojęć. Aby uzyskać więcej informacji o korzystaniu z węzła File List, patrz “Węzeł File List” na stronie 11.

1. **Węzeł File List (karta Settings).** Najpierw dodaliśmy ten węzeł do strumienia, aby określić, gdzie są przechowywane dokumenty tekstowe. Wybraliśmy katalog zawierający wszystkie dokumenty, na których chcemy przeprowadzać eksplorację tekstu.
2. **Węzeł Text Mining (karta Fields).** Następnie dodaliśmy węzeł Text Mining i połączyliśmy go z węzłem File List. W tym węźle zdefiniowaliśmy format wejściowy, szablon zasobów i format wyjściowy. Wybraliśmy nazwę zmiennej wygenerowaną z węzła File List i wybraliśmy zmienną tekstową. Określiliśmy również inne ustawienia. Więcej informacji zawiera temat “Korzystanie z węzła Text Mining w strumieniu”.
3. **Węzeł Text Mining (karta Model)**. Teraz na karcie Model wybraliśmy tryb budowania, który pozwoli na wygenerowanie modelu użytkowego pojęć bezpośrednio z tego węzła. Można wybrać inny szablon zasobów lub zachować zasoby podstawowe.

Przykład 2: Użycie węzłów Plik programu Excel i Text Mining do interaktywnego budowania modelu kategorii

Ten przykład pokazuje, że węzeł Text Mining umożliwia również uruchomienie sesji pracy z interaktywnym pulpitem roboczym. Aby uzyskać więcej informacji o interaktywnym pulpicie roboczym, patrz Rozdział 7, “Tryb pracy z interaktywnym pulpitem roboczym”, na stronie 67.

1. **Węzeł źródłowy Plik programu Excel (karta Data).** Najpierw dodaliśmy ten węzeł do strumienia, aby określić, gdzie jest przechowywany tekst.
2. **Węzeł Text Mining (karta Fields).** Następnie dodaliśmy węzeł Text Mining i połączyliśmy go z węzłem Plik programu Excel. Na pierwszej karcie zdefiniowaliśmy format wejściowy. Wybraliśmy nazwę zmiennej z węzła źródłowego.
3. **Węzeł Text Mining (karta Model).** Następnie na karcie Model wybraliśmy opcję interaktywnego budowania modelu użytkowego kategorii oraz wykorzystania wyników wyodrębniania do automatycznego zbudowania kategorii. W tym przykładzie załadowaliśmy kopię zasobów i zestaw kategorii z pakietu analizy tekstu.
4. **Sesja pracy z interaktywnym pulpitem roboczym**. Następnie uruchomiliśmy strumień, co spowodowało otwarcie interaktywnego pulpitu roboczego. Po przeprowadzeniu wyodrębniania rozpoczęliśmy eksplorację danych i doskonalenie kategorii.

Model użytkowy Text Mining: model pojęć

Model użytkowy pojęć jest generowany po pomyślnym wykonaniu węzła Text Mining, jeśli na karcie Model wybrano opcję **Generate a model directly**. Model użytkowy pojęć służy do wykrywania w czasie rzeczywistym kluczowych pojęć w innych danych tekstowych, na przykład notatkach konsultantów telefonicznego centrum zgłoszeniowego.

Sam model użytkowy pojęć składa się z listy pojęć przypisanych do typów. Można wybrać dowolne z pojęć lub wszystkie pojęcia w modelu, by na ich podstawie oceniać inne dane. W trakcie wykonywania strumienia zawierającego model użytkowy Text Mining do danych są dodawane nowe pola zgodnie z trybem budowania, który jeszcze przed rozpoczęciem budowania modelu wybrano na karcie Model węzła modelowania Text Mining. Więcej informacji zawiera temat "Model pojęć: karta Model".

Jeśli model użytkowy został wygenerowany przy użyciu dokumentów przetłumaczonych, to ocena będzie wykonywana na języku docelowym tłumaczenia. Podobnie, jeśli model użytkowy został wygenerowany przy użyciu języka angielskiego, można w tym modelu użytkowym określić język tłumaczenia, ponieważ dokumenty będą tłumaczone na język angielski.

Wygenerowane modele użytkowe Text Mining umieszczane są na palecie modeli użytkowych (która znajduje się na karcie Modele w prawym górnym rogu okna IBM SPSS Modeler).

Wyświetlanie wyników

Aby wyświetlić informacje o modelu użytkowym, kliknij prawym przyciskiem myszy węzeł na palecie modeli użytkowych i wybierz polecenie **Browse** z menu kontekstowego (lub **Edit** w przypadku modeli w strumieniu).

Dodawanie modeli do strumieni

Aby dodać model użytkowy do strumienia, kliknij ikonę na palecie modeli użytkowych, a następnie kliknij obszar roboczy strumienia w miejscu, w którym chcesz umieścić węzeł. Zamiast tego można też kliknąć ikonę i z menu kontekstowego wybrać polecenie **Add to Stream**. Następnie połącz strumień z nowym węzłem. Można teraz przekazać do strumienia dane w celu wygenerowania predykcji.

Przestroga: Jeśli chcesz użyć modelu użytkowego oceniania do ponownego wygenerowania węzła modelowania zawierającego i model kategorii, i użyty szablon, zalecamy utworzenie, zamiast węzła modelowania, pakietu TAP i użycie go w sesji interaktywnej przed wygenerowaniem modelu użytkowego oceniania.

Model pojęć: karta Model

Na karcie Model modelu pojęć wyświetlana jest lista wyodrębnionych pojęć. Pojęcia są prezentowane w tabeli, w której każdy wiersz zawiera jedno pojęcie. Karta ta służy do wybierania pojęć, które będą używane do oceniania.

Uwaga: Jeśli wygenerowano model użytkowy kategorii, a nie pojęć, to na tej karcie będą wyświetlane inne informacje. Więcej informacji zawiera temat "Model użytkowy kategorii: karta Model" na stronie 39.

Domyślnie wszystkie pojęcia są wybrane jako używane do oceniania (pola wyboru w lewej skrajniej kolumnie). Zaznaczenie sygnalizuje, że pojęcie będzie używane podczas oceniania. Brak zaznaczenia sygnalizuje, że pojęcie będzie wykluczone z oceniania. Można zaznaczyć wiele wierszy naraz, wybierając je i klikając jedno z ich pól wyboru.

Dodatkowe informacje o każdym z pojęć znajdują się w następujących kolumnach:

Concept. Jest to wiodący wyodrębniony wyraz lub wiodąca wyodrębniona fraza. W niektórych przypadkach pojęcie reprezentuje termin będący jego nazwą oraz inne terminy powiązane z tym samym pojęciem. Aby sprawdzić, które terminy należą do pojęcia, wyświetl panel Underlying Term wewnątrz tej karty i wybierz pojęcie, aby u dołu okna dialogowego wyświetlić odpowiadające mu terminy. Więcej informacji zawiera temat "Terminy w modelach pojęć" na stronie 33.

Global. W tym miejscu globalna częstość oznacza liczbę wystąpień pojęcia (i wszystkich związanych z nim terminów) w całym zbiorze dokumentów/rekordów.

- **Bar chart.** Globalna częstość tego pojęcia w danych tekstowych przedstawiona w formie wykresu słupkowego. Kolor słupka oznacza typ, do którego pojęcie zostało przypisane. Dzięki temu łatwo można wizualnie rozróżnić typy.
- **%.** Globalna częstość tego pojęcia w danych tekstowych przedstawiona w formie procentowej.

- **N.** Faktyczna liczba wystąpień tego pojęcia w danych tekstowych.

Docs. Wartość Docs oznacza tutaj liczbę dokumentów lub rekordów, w których występuje pojęcie (i wszystkie jego terminy).

- **Bar chart.** Liczba dokumentów, w których występuje to pojęcie, przedstawiona w formie wykresu słupkowego. Kolor słupka oznacza typ, do którego pojęcie zostało przypisane. Dzięki temu łatwo można wizualnie rozróżnić typy.
- **%.** Liczba dokumentów, w których występuje to pojęcie, przedstawiona w formie procentowej.
- **N.** Faktyczna liczba dokumentów lub rekordów zawierających to pojęcie.

Typ. Typ, do którego pojęcie zostało przypisane. Dla każdego pojęcia kolumny Global i Docs wyświetlane są w kolorze oznaczającym jego typ. **Typ** jest semantyczną grupą pojęć. Więcej informacji zawiera temat “Słowniki typów” na stronie 177.

Praca z pojęciami

Kliknięcie komórki tabeli prawym przyciskiem myszy umożliwia wyświetlenie menu kontekstowego zawierającego następujące opcje:

- **Select All.** Wszystkie wiersze w tabeli zostaną wybrane.
- **Copy.** Wybrane pojęcia są kopiowane do schowka.
- **Copy With Fields** Wybrane pojęcia są kopiowane do schowka razem z nagłówkiem kolumny.
- **Check Selected.** Zaznacza wszystkie pola wyboru wybranych wierszy w tabeli, co powoduje, że odpowiednie pojęcia będą używane do oceniania.
- **Uncheck Selected.** Usuwa zaznaczenie wszystkich pól wyboru wybranych wierszy w tabeli.
- **Check All.** Zaznacza wszystkie pola wyboru w tabeli. W efekcie wszystkie pojęcia zostaną zastosowane w ostatecznych wynikach.
- **Uncheck All.** Usuwa zaznaczenie wszystkich pól wyboru w tabeli. Anulowanie zaznaczenia pojęcia spowoduje, że nie będzie ono stosowane w ostatecznych wynikach.
- **Include Concepts.** Powoduje wyświetlenie okna dialogowego Include Concepts. Więcej informacji zawiera temat “Opcje włączania pojęć do oceniania”.

Opcje włączania pojęć do oceniania

Aby szybko zaznaczać lub usuwać zaznaczenie pojęć, które mają być używane przy ocenieniu, kliknij przycisk **Include Concepts.**.

Rysunek 1. Przycisk Include Concepts na pasku narzędzi

Kliknięcie tego przycisku na pasku narzędzi spowoduje otwarcie okna dialogowego Include Concepts, w którym można wybrać pojęcia na podstawie reguł. W ocenianiu uwzględniane będą wszystkie pojęcia zaznaczone na karcie Model. Stosując regułę w tym podrzędnym oknie dialogowym, możesz zmienić wybór pojęć używanych podczas oceniania.

Do wyboru dostępne są następujące opcje:

Check concepts based on highest frequency. Top number of concepts. Jest to liczba pojęć, które zostaną zaznaczone począwszy od pojęcia o największej globalnej częstości (liczebności wystąpień). W tym miejscu częstość oznacza liczbę wystąpień pojęcia (i wszystkich związanych z nim terminów) w całym zbiorze dokumentów/rekordów. Ta liczba może być wyższa od liczby rekordów, ponieważ jedno pojęcie może występować wielokrotnie w tym samym rekordzie.

Check concepts based on document count. Minimum count. Jest to najmniejsza liczba dokumentów wymagana, by pojęcie zostało zaznaczone. Liczba dokumentów oznacza tutaj liczbę dokumentów/rekordów, w których występuje pojęcie (i wszystkie jego terminy).

Check concepts assigned to the type. Wybierz typ z listy rozwijanej, aby zaznaczyć wszystkie pojęcia przypisane do tego typu. Pojęcia są przypisywane do typów automatycznie w trakcie wyodrębniania. **Typ** jest semantyczną grupą pojęć. Typy odzwierciedlają pojęcia poziomowe, kwalifikatory i określenia o wydźwięku pozytywnym i negatywnym, kwalifikatory kontekstowe, imiona, miejsca, organizacje i nie tylko. Więcej informacji zawiera temat “Słowniki typów” na stronie 177.

Uncheck concepts that occur in too many records. Percentage of records. Usuwa zaznaczenie pojęć, których wyrażona procentowo liczba wystąpień w rekordach jest większa od określonej. Ta opcja jest użyteczna do wykluczania pojęć, które występują często w tekście lub w każdym rekordzie, ale nie mają znaczenia w prowadzonej analizie.

Uncheck concepts assigned to the type. Usuwa zaznaczenia pojęć, którym przypisany jest typ wybrany z listy rozwijanej.

Terminy w modelach pojęć

Można sprawdzić, jakie terminy są zdefiniowane w ramach pojęć zaznaczonych w tabeli. Klikając przycisk terminów pojęć na pasku narzędzi, można wyświetlić tabelę takich terminów w podzielonym panelu na dole okna dialogowego.

Do terminów zdefiniowanych w ramach pojęć (nazywanych też terminami bazowymi pojęć) należą synonimy zdefiniowane w zasobach lingwistycznych (niezależnie od tego, czy wystąpiły w tekście, czy nie), a także wszelkie znalezione w tekście terminy w liczbie mnogiej/pojedynczej, permutacje terminów, terminy wygenerowane w wyniku grupowania rozmytego itd.

Rysunek 2. Przycisk Display Underlying Terms na pasku narzędzi

Uwaga: Listy terminów w ramach pojęcia nie można edytować. Lista jest generowana na podstawie słownika zastąpień, definicji synonimów (w słowniku zastąpień), grupowania rozmytego i przy użyciu innych technik według danych zawartych w zasobach lingwistycznych. Aby wpłynąć na grupowanie terminów w pojęcia i sposób postępowania z terminami, należy bezpośrednio zmodyfikować zasoby (można je edytować w oknie Resource Editor w ramach interaktywnego pulpitu roboczego lub w oknie Template Editor, a potem ponownie załadować do węzła), a następnie ponownie wykonać strumień, aby uzyskać nowy model użytkowy ze zaktualizowanymi wynikami.

Kliknięcie komórki z terminem lub pojęciem prawym przyciskiem myszy umożliwia wyświetlenie menu kontekstowego zawierającego następujące opcje:

- **Copy.** Wybrana komórka zostanie skopiowana do schowka.
- **Copy With Fields.** Wybrana komórka zostanie skopiowana do schowka razem z nagłówkami kolumn.
- **Select All.** Wszystkie komórki w tabeli zostaną wybrane.

Model pojęć: karta Settings

Karta Settings służy do określania wartości zmiennej tekstowej w celu uwzględnienia nowych danych wejściowych, gdy zachodzi taka potrzeba. Jest ona także miejscem, w którym można zdefiniować model danych dla wyniku (tryb oceniania).

Uwaga: Ta karta jest dostępna tylko wtedy, gdy model użytkowy został umieszczony w obszarze roboczym. Nie jest wyświetlana, gdy to okno dialogowe zostanie otwarte bezpośrednio z palety Modele.

Tryb oceniania: Concepts as records

W tym trybie oceniania dla każdej pary pojęcie/dokument tworzony jest nowy rekord. Zwykle wyniki zawierają więcej rekordów niż dane wejściowe.

Oprócz zmiennych wejściowych do danych dodawane są także następujące zmienne:

Tabela 4. Zmienne wynikowe w trybie „Concepts as records”.

Zmienna	Opis
Concept	Zawiera nazwę pojęcia wyodrębnioną ze zmiennej z danymi tekstowymi.
Type	Zawiera typ pojęcia w postaci pełnej nazwy, na przykład *Location* lub *Person*. **Typ** jest semantyczną grupą pojęć. Więcej informacji zawiera temat “Słowniki typów” na stronie 177.
Count	Wyświetla liczbę wystąpień pojęcia (i jego terminów) w treści tekstu (rekordach/dokumentach).

Wybranie tej opcji powoduje wyłączenie wszystkich pozostałych opcji z wyjątkiem **Accommodate punctuation errors**.

Tryb oceniania: Concepts as fields

W modelach pojęć dla każdego rekordu wejściowego i pojęcia występującego w danym rekordzie/dokumencie tworzony jest nowy rekord. Zatem liczba rekordów wynikowych jest taka sama, jak w danych wejściowych. Jednak każdy rekord (wiersz) zawiera teraz po jednej nowej zmiennej (kolumnie) dla każdego pojęcia zaznaczonego na karcie Model. Wartość każdej zmiennej pojęcia zależy od tego, czy jako wartość zmiennej na tej karcie wybrano **Flags**, czy **Counts**.

Uwaga: Jeśli zbiór danych jest bardzo obszerny, np. jest bazą danych DB2, wybranie opcji **Concepts as fields** może spowodować problemy z przetwarzaniem. W takich przypadkach zaleca się korzystanie z opcji **Concepts as records**.

Field Values. Wybierz, czy nowa zmienna każdego pojęcia będzie zawierać liczebność, czy flagę.

- **Flags.** Ta opcja służy do generowania wynikowych flag z dwiema możliwymi wartościami: na przykład *Tak*/*Nie*, *True*/*False*, *T*/*F*, albo *1* i *2*. Typy składowania ustawiane są automatycznie odpowiednio do wybranej wartości. Na przykład, jeśli wprowadzisz liczbowe wartości flag, to automatycznie będą traktowane jako wartości całkowitoliczbowe. Flagi mogą być składowane jako łańcuchy, liczby całkowite, liczby rzeczywiste lub wartości data/czas. Wprowadź wartość flagi oznaczającą **True** i wartość oznaczającą **False**.
- **Counts.** Służy do generowania liczebności wystąpień pojęcia w danym rekordzie.

Rozszerzenie nazwy zmiennej. Określ rozszerzenie nazwy zmiennej. Nazwy zmiennych generowane są poprzez połączenie nazwy pojęcia z rozszerzeniem.

- **Add as.** Określ, gdzie do nazwy zmiennej powinno być dodawane rozszerzenie. Wybierz **Prefix**, aby dodawać rozszerzenie na początku łańcucha. Wybierz **Suffix**, aby dodawać rozszerzenie na końcu łańcucha.

Accommodate punctuation errors. Ta opcja powoduje, że podczas wyodrębniania tekst zawierający błędy interpunkcyjne (na przykład nieprawidłowo użyte znaki interpunkcyjne) będzie tymczasowo normalizowany w celu poprawienia efektywności wyodrębniania pojęć. Ta opcja jest bardzo użyteczna, gdy mamy do czynienia z krótkimi tekstami o niskiej jakości (np. odpowiedziami na pytania otwarte w ankietach, wiadomościami e-mail i danymi z systemów CRM) lub gdy system zawiera wiele skrótów.

Uwaga: Opcja **Accommodate punctuation errors** nie ma zastosowania podczas pracy z tekstem w języku japońskim.

Model pojęć: karta Fields

Karta Fields określa wartość zmiennej tekstowej w celu uwzględnienia nowych danych wejściowych, gdy zachodzi taka potrzeba.

Uwaga: Ta karta jest dostępna tylko wtedy, gdy model użytkowy został umieszczony w obszarze roboczym. Nie jest wyświetlana, gdy to okno dialogowe zostanie otwarte bezpośrednio z palety Modele.

Text field. Wybierz zmienną zawierającą tekst do eksploracji. Zmienna ta zależy od źródła danych.

Document type. Typ dokumentu określa strukturę tekstu. Wybierz jeden z następujących typów:

- **Full text.** Odpowiedni dla większości dokumentów i źródeł tekstowych. Podczas wyodrębniania przeglądany jest cały zbiór tekstu. Z tą opcją, w odróżnieniu od pozostałych, nie są związane żadne ustawienia dodatkowe.
- **Structured text.** Odpowiedni do analizy formularzy bibliograficznych, patentów i innych plików zawierających regularne struktury dające się rozpoznać i przeanalizować. Zastosowanie tego typu dokumentów wiąże się z pominięciem całości lub części procesu wyodrębniania. Umożliwia zdefiniowanie separatorów terminów, przypisywanie typów i określenie wymaganej minimalnej częstotliwości. W przypadku wybrania tej opcji należy kliknąć przycisk **Settings** i wprowadzić separatory tekstu w obszarze **Structured Text Formatting** okna dialogowego Document Settings. Więcej informacji zawiera temat “Ustawienia dokumentów na karcie Zmienne” na stronie 22.

Input encoding Ta opcja jest dostępna tylko wtedy, gdy wskazano, że zmienna tekstowa zawiera ścieżki (opcja **Pathnames to documents**). Określa ona domyślne kodowanie tekstu. We wszystkich językach z wyjątkiem japońskiego dokonywana jest konwersja z określonego lub rozpoznanego kodowania na ISO-8859-1. Zatem nawet jeśli określono inne kodowanie, mechanizm wyodrębniania przed przetworzeniem tekstu przekonwertuje go na kodowanie ISO-8859-1. Wszelkie znaki niewystępujące w definicji kodowania ISO-8859-1 zostaną przekształcone w spacje. W przypadku tekstu japońskiego można wybrać kilka opcji kodowania: SHIFT_JIS, EUC_JP, UTF-8 lub ISO-2022-JP.

Text language. Określa język tekstu poddawanego eksploracji. Jest to główny język wykryty podczas wyodrębniania. Użytkownicy zainteresowani zakupem licencji na obsługiwany język, do którego obecnie nie mają dostępu, powinni skontaktować się z przedstawicielem handlowym.

Model pojęć: karta Summary

Na karcie Summary prezentowane są informacje o samym modelu (folder *Analiza*), zmiennych używanych w modelu (folder *Zmienne*), ustawieniach używanych przy budowaniu modelu (folder *Ustawienia budowania*) i o uczeniu modelu (folder *Podsumowanie uczenia*).

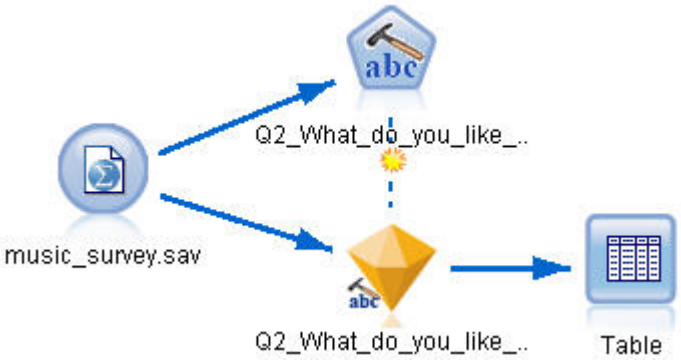
Gdy po raz pierwszy otworzysz węzeł modelowania, foldery na karcie Podsumowanie są zwinięte. Aby wyświetlić interesujące wyniki, skorzystaj z rozszerzanego elementu sterującego po lewej stronie folderu albo kliknij przycisk **Expand All**, co spowoduje wyświetlenie wszystkich wyników. Aby ukryć wyniki po ich przejrzeniu, użyj rozszerzanego elementu sterującego w celu zwinięcia konkretnego folderu albo kliknij przycisk **Collapse All**, aby zwinąć wszystkie foldery.

Korzystanie z modeli użytkowych pojęć w strumieniu

Korzystając z węzła modelowania Text Mining, można wygenerować model użytkowy pojęć lub model użytkowy kategorii (w ramach sesji pracy z interaktywnym pulpitem roboczym). Poniższy przykład ilustruje zastosowanie modelu pojęć w prostym strumieniu.

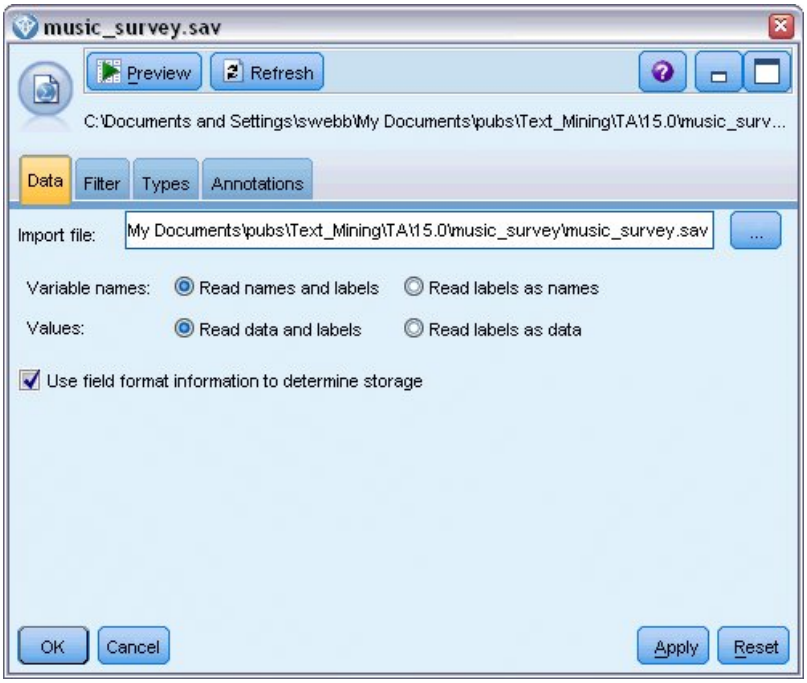
Przykład: węzeł Plik Statistics z modelem użytkowym pojęć.

Poniższy przykład ilustruje zastosowanie modelu użytkowego wygenerowanego przez węzeł Text Mining.



Rysunek 3. Przykładowy strumień: węzeł Plik Statistics z modelem użytkowym Text Mining

1. **Węzeł Plik Statistics (karta Data).** Najpierw dodaliśmy ten węzeł do strumienia, aby określić, gdzie są przechowywane dokumenty tekstowe.

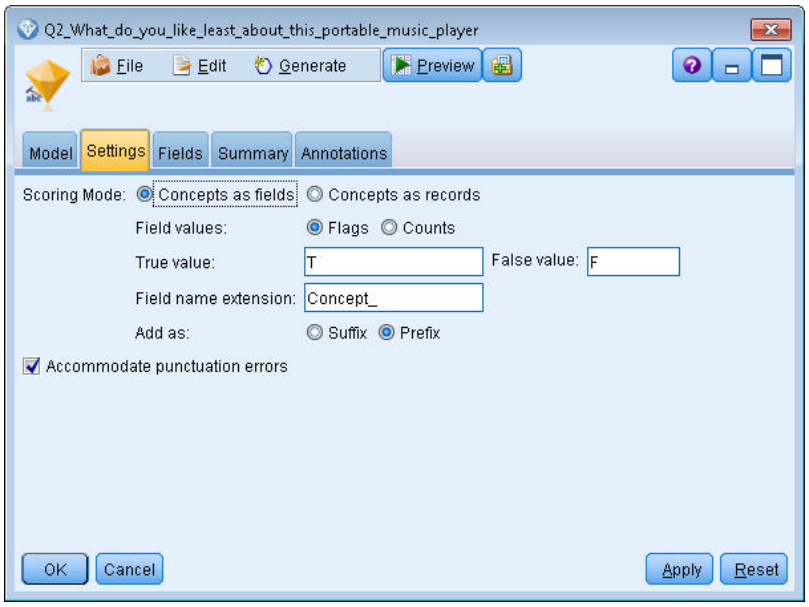


Rysunek 4. Okno dialogowe węzła Plik Statistics: karta Data

2. **Model użytkowy pojęć wygenerowany przez węzeł Text Mining (karta Model).** Następnie dodaliśmy węzeł modelu użytkowego pojęć i połączyliśmy go z węzłem Plik Statistics. Wybraliśmy pojęcia, które chcemy wykorzystać do oceniania danych.

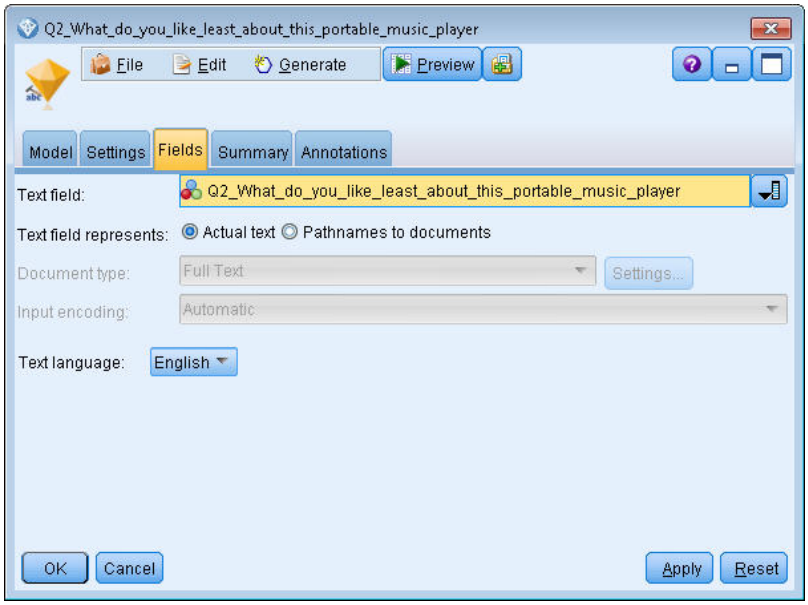
Rysunek 5. Okno dialogowe modelu użytkowego Text Mining: karta Model

3. **Model użytkowy pojęć wygenerowany przez węzeł Text Mining (karta Settings).** Następnie zdefiniowaliśmy format wyjściowy i wybraliśmy opcję *Concepts as fields*. W wynikach dla każdego pojęcia zaznaczonego na karcie Model zostanie utworzona jedna nowa zmienna. Nazwa każdej zmiennej zostanie skonstruowana z nazwy pojęcia i przedrostka "*Concept_*"



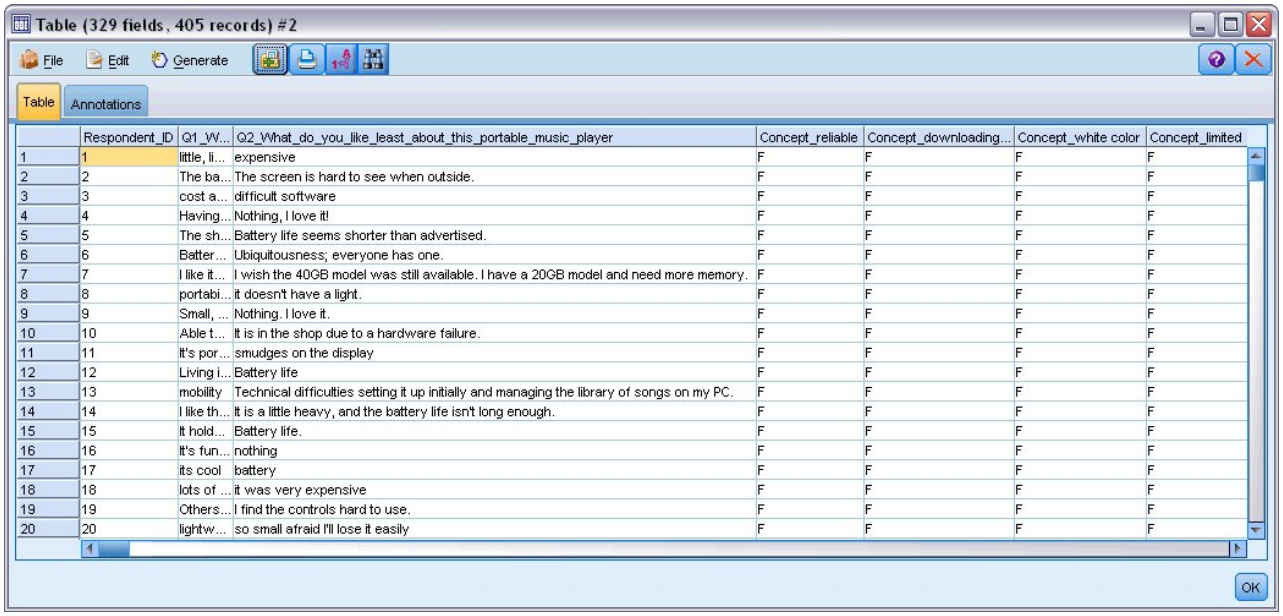
Rysunek 6. Okno dialogowe modelu użytkowego pojęć wygenerowanego przez węzeł Text Mining: karta Settings

4. **Model użytkowy pojęć wygenerowany przez węzeł Text Mining (karta Fields).** Następnie wybraliśmy zmienną tekstów **Q2_What_do_you_like_least_about_this_portable_music_player**, której nazwa pochodzi z węzła Plik Statistics. Ponadto wybraliśmy opcję **Text field represents: Actual text**.



Rysunek 7. Okno dialogowe modelu użytkowego pojęć wygenerowanego przez węzeł Text Mining: karta Fields

5. **Węzeł Tabela.** Następnie dołączyliśmy węzeł Tabela, aby wyświetlić wyniki i wykonać strumień. Na ekranie zostanie wyświetlona tabela wyników.



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

Rysunek 8. Tabela wyników przewinięta tak, by były widoczne flagi pojęć.

Model użytkowy Text Mining: model kategorii

Model użytkowy kategorii jest generowany przez węzeł Text Mining, gdy zażądasz wygenerowania go w interaktywnym pulpicie roboczym. Ten model użytkowy zawiera zestaw kategorii, których definicje składają się z pojęć, typów, wzorców TLA i/lub reguł kategorii. Model użytkowy służy do kategoryzacji odpowiedzi na ankiety, wpisów w blogach, innych kanałów WWW oraz wszelkich innych danych tekstowych.

Jeśli uruchomisz sesję pracy z interaktywnym pulpitem w węźle modelowania, możesz eksplorować wyniki wyodrębniania oraz zoptymalizować zasoby i kategorie przed wygenerowaniem modeli kategorii. W trakcie wykonywania strumienia zawierającego model użytkowy Text Mining do danych są dodawane nowe pola zgodnie z trybem budowania, który jeszcze przed rozpoczęciem budowania modelu wybrano na karcie Model węzła modelowania Text Mining. Więcej informacji zawiera temat “Model użytkowy kategorii: karta Model”.

Jeśli model użytkowy został wygenerowany przy użyciu dokumentów przetłumaczonych, to ocena będzie wykonywana na języku docelowym tłumaczenia. Podobnie, jeśli model użytkowy został wygenerowany przy użyciu języka angielskiego, można w tym modelu użytkowym określić język tłumaczenia, ponieważ dokumenty będą tłumaczone na język angielski.

Wygenerowane modele użytkowe Text Mining umieszczane są na palecie modeli użytkowych (która znajduje się na karcie Modele w prawym górnym rogu okna IBM SPSS Modeler).

Wyświetlanie wyników

Aby wyświetlić informacje o modelu użytkowym, kliknij prawym przyciskiem myszy węzeł na palecie modeli użytkowych i wybierz polecenie **Browse** z menu kontekstowego (lub **Edit** w przypadku modeli w strumieniu).

Dodawanie modeli do strumieni

Aby dodać model użytkowy do strumienia, kliknij ikonę na palecie modeli użytkowych, a następnie kliknij obszar roboczy strumienia w miejscu, w którym chcesz umieścić węzeł. Zamiast tego można też kliknąć ikonę i z menu kontekstowego wybrać polecenie **Add to Stream**. Następnie połącz strumień z nowym węzłem. Można teraz przekazać do strumienia dane w celu wygenerowania predykcji.

Przestroga: Jeśli chcesz użyć modelu użytkowego oceniania do ponownego wygenerowania węzła modelowania zawierającego i model kategorii, i użyty szablon, zalecamy utworzenie, zamiast węzła modelowania, pakietu TAP i użycie go w sesji interaktywnej przed wygenerowaniem modelu użytkowego oceniania.

Model użytkowy kategorii: karta Model

Karta Model modelu kategorii zawiera listę kategorii ujętych w modelu (po lewej stronie) i deskryptory wybranej kategorii (po prawej stronie). Każda kategoria złożona jest z szeregu deskryptorów. Deskryptory powiązane z wybraną kategorią są wyświetlane w tabeli. Deskryptorami mogą być pojęcia, reguły kategorii, typy i wzorce TLA. Wyświetlany jest także typ deskryptora, jak również przykłady ilustrujące, co takiego reprezentuje dany deskryptor.

Ta karta służy do wyboru kategorii, które mają być używane podczas oceniania. Model kategorii ocenia dokumenty i rekordy, przypisując je do kategorii. Jeśli w tekście lub terminach pojęć dokumentu lub rekordu występuje jeden lub więcej deskryptorów, to dokument lub rekord jest przypisywany do kategorii, do której należy ten deskryptor. Do terminów zdefiniowanych w ramach pojęć (nazywanych też terminami bazowymi pojęć) należą synonimy zdefiniowane w zasobach lingwistycznych (niezależnie od tego, czy wystąpiły w tekście, czy nie), a także wszelkie znalezione w tekście terminy w liczbie mnogiej/pojedynczej, permutacje terminów, terminy wygenerowane w wyniku grupowania rozmytego itd.

Uwaga: Jeśli wygenerowano model użytkowy pojęć, a nie kategorii, to na tej karcie będą wyświetlane inne informacje. Więcej informacji zawiera temat “Model pojęć: karta Model” na stronie 31.

Drzewo kategorii

Aby wyświetlić dodatkowe informacje o kategorii, wybierz ją — zostaną wyświetlone informacje o jej deskryptorach. Przy każdym deskryptorze dostępne są następujące informacje:

- Nazwa deskryptora (**Descriptor**). To pole zawiera ikonę oznaczającą rodzaj deskryptora oraz jego nazwę.

Tabela 5. Ikony deskryptorów

	Pojęcia		Wzorce TLA
	Typy		Reguły kategorii

- **Typ.** To pole zawiera nazwę typu deskryptora. Typy są zbiorami podobnych pojęć (grupami semantycznymi), takich jak nazwy organizacji lub produktów albo pozytywne opinie. Reguły nie są przypisywane do typów.
- **Details.** To pole zawiera listę elementów wchodzących w skład deskryptora. W zależności od listy dopasowań wyświetlana jest cała lista lub tylko jej część. Wynika to z ograniczeń rozmiaru okna dialogowego.

Wybieranie i kopiowanie kategorii

Domyślnie wszystkie kategorie najwyższego poziomu są wybrane jako używane do oceniania (pola wyboru w lewym panelu). Zaznaczenie sygnalizuje, że kategoria będzie używana podczas oceniania. Brak zaznaczenia sygnalizuje, że kategoria będzie wykluczona z oceniania. Można zaznaczyć wiele wierszy naraz, wybierając je i klikając jedno z ich pól wyboru. Ponadto, jeśli wybrana jest kategoria lub podkategoria, ale jedna z jej podkategorii nie jest wybrana, to pola wyboru będą miały niebieskie tło, które oznacza, że wybrana jest tylko część kategorii podrzędnych.

Kliknięcie kategorii w drzewie prawym przyciskiem myszy umożliwia wyświetlenie menu kontekstowego zawierającego następujące opcje:

- **Check Selected.** Zaznacza wszystkie pola wyboru wybranych wierszy w tabeli.
- **Uncheck Selected.** Usuwa zaznaczenie wszystkich pól wyboru wybranych wierszy w tabeli.
- **Check All.** Zaznacza wszystkie pola wyboru w tabeli. W efekcie wszystkie kategorie zostaną zastosowane w ostatecznych wynikach. Można też użyć odpowiedniej ikony zaznaczonego pola wyboru na pasku narzędzi.
- **Uncheck All.** Usuwa zaznaczenie wszystkich pól wyboru w tabeli. Usunięcie zaznaczenia kategorii spowoduje, że nie będzie ona stosowana w ostatecznych wynikach. Można też użyć odpowiedniej ikony pustego pola wyboru na pasku narzędzi.

Kliknięcie komórki tabeli prawym przyciskiem myszy umożliwia wyświetlenie menu kontekstowego zawierającego następujące opcje:

- **Copy.** Wybrane pojęcia są kopiowane do schowka.
- **Copy With Fields.** Wybrany deskryptor jest kopiowany do schowka razem z nagłówkami kolumn.
- **Select All.** Wszystkie wiersze w tabeli zostaną wybrane.

Model użytkowy kategorii: karta Settings

Karta Settings służy do określania wartości zmiennej tekstowej w celu uwzględnienia nowych danych wejściowych, gdy zachodzi taka potrzeba. Jest ona także miejscem, w którym można zdefiniować model danych dla wyniku (tryb oceniania).

Uwaga: Ta karta jest dostępna w oknie dialogowym węzła tylko wtedy, gdy model użytkowy został umieszczony w obszarze roboczym lub w strumieniu. Nie jest wyświetlana, gdy okno dialogowe modelu użytkowego zostanie otwarte bezpośrednio z palety Modele.

Tryb oceniania: Categories as fields

Zatem liczba rekordów wynikowych jest taka sama, jak w danych wejściowych. Jednak każdy rekord zawiera teraz po jednej nowej zmiennej dla każdej kategorii zaznaczonej na karcie Model. Dla każdej zmiennej wprowadź wartość flagi oznaczającą **True** i wartość oznaczającą **False**, na przykład *Tak*/*Nie*, *Prawda*/*Fałsz*, *T*/*F*, albo *1* i *2*. Typy składowania są ustawiane automatycznie odpowiednio do wybranej wartości. Na przykład, jeśli wprowadzisz liczbowe wartości flag, to automatycznie będą traktowane jako wartości całkowitoliczbowe. Flagi mogą być składowane jako łańcuchy, liczby całkowite, liczby rzeczywiste lub wartości data/czas.

Uwaga: Jeśli zbiór danych jest bardzo obszerny, np. jest bazą danych DB2, wybranie opcji **Categories as fields** może spowodować problemy z przetwarzaniem. W takich przypadkach zaleca się korzystanie z opcji **Categories as records**.

Rozszerzenie nazwy zmiennej. Można wybrać opcję określania przedrostka/przyrostka (rozszerzenia) nazwy zmiennej lub użyć kodów kategorii. Nazwy zmiennych generowane są poprzez połączenie nazwy kategorii z rozszerzeniem.

- **Add as.** Określ, gdzie do nazwy zmiennej powinno być dodawane rozszerzenie. Wybierz **Prefix**, aby dodawać rozszerzenie na początku łańcucha. Wybierz **Suffix**, aby dodawać rozszerzenie na końcu łańcucha.

If a subcategory is unselected. Ta opcja umożliwia określenie, w jaki sposób traktowane będą deskryptory należące do podkategorii niewybranych do oceny. Dostępne są dwie opcje.

- Opcja **Exclude its descriptors completely from scoring** powoduje, że deskryptory niezaznaczonych podkategorii będą ignorowane i nie będą używane podczas oceniania.
- Opcja **Aggregate descriptors with those in parent category** powoduje, że deskryptory niezaznaczonych podkategorii będą używane jako deskryptory kategorii nadrzędnej wobec tych podkategorii. Jeśli niezaznaczone są kategorie na kilku poziomach, deskryptory będą agregowane aż do pierwszej dostępnej kategorii nadrzędnej.

Accommodate punctuation errors. Ta opcja powoduje, że podczas wyodrębniania tekst zawierający błędy interpunkcyjne (na przykład nieprawidłowo użyte znaki interpunkcyjne) będzie tymczasowo normalizowany w celu poprawienia efektywności wyodrębniania pojęć. Ta opcja jest bardzo użyteczna, gdy mamy do czynienia z krótkimi tekstami o niskiej jakości (np. odpowiedziami na pytania otwarte w ankietach, wiadomościami e-mail i danymi z systemów CRM) lub gdy system zawiera wiele skrótów.

Uwaga: Opcja **Accommodate punctuation errors** nie ma zastosowania podczas pracy z tekstem w języku japońskim.

Tryb oceniania: Categories as records

W tym trybie oceniania dla każdej pary kategoria, dokument tworzony jest nowy rekord. Zwykle wyniki zawierają więcej rekordów niż dane wejściowe. Oprócz zmiennych wejściowych do danych dodawane są także zmienne wynikowe — zależnie od rodzaju modelu.

Tabela 6. Zmienne wynikowe w trybie „Categories as records”.

Nowa zmienna wynikowa	Opis
Category	Zawiera nazwę kategorii, do której dokument tekstowy został przypisany. Jeśli kategoria jest podkategorią innej, pełna ścieżka kategorii jest sterowana przez wartość wybraną w tym oknie dialogowym.

Values for hierarchical categories. Ta opcja określa, w jaki sposób nazwy podkategorii są wyświetlane w wynikach.

- **Full category path.** Ta opcja wyświetli nazwę kategorii oraz pełną ścieżkę kategorii nadrzędnych, w razie potrzeby używając ukośników do rozdzielania nazw kategorii od nazw podkategorii.
- **Short category path.** Ta opcja wyświetli tylko nazwę kategorii, ale użyje wielokropków do wyświetlenia liczby kategorii nadrzędnych dla określonej kategorii.
- **Bottom level category.** Ta opcja wyświetli tylko nazwę kategorii bez pełnej ścieżki i bez wyświetlonych kategorii nadrzędnych.

If a subcategory is unselected. Ta opcja umożliwia określenie, w jaki sposób traktowane będą deskryptory należące do podkategorii niewybranych do oceny. Dostępne są dwie opcje.

- Opcja **Exclude its descriptors completely from scoring** powoduje, że deskryptory niezaznaczonych podkategorii będą ignorowane i nie będą używane podczas oceniania.
- Opcja **Aggregate descriptors with those in parent category** powoduje, że deskryptory niezaznaczonych podkategorii będą używane jako deskryptory kategorii nadrzędnej wobec tych podkategorii. Jeśli niezaznaczone są kategorie na kilku poziomach, deskryptory będą agregowane aż do pierwszej dostępnej kategorii nadrzędnej.

Accommodate punctuation errors. Ta opcja powoduje, że podczas wyodrębniania tekst zawierający błędy interpunkcyjne (na przykład nieprawidłowo użyte znaki interpunkcyjne) będzie tymczasowo normalizowany w celu poprawienia efektywności wyodrębniania pojęć. Ta opcja jest bardzo użyteczna, gdy mamy do czynienia z krótkimi tekstami o niskiej jakości (np. odpowiedziami na pytania otwarte w ankietach, wiadomościami e-mail i danymi z systemów CRM) lub gdy system zawiera wiele skrótów.

Uwaga: Opcja **Accommodate punctuation errors** nie ma zastosowania podczas pracy z tekstem w języku japońskim.

Model użytkowy kategorii: karta Other

Karta Fields i karta Settings dla modelu użytkowego kategorii są takie same, jak dla modelu użytkowego pojęcia.

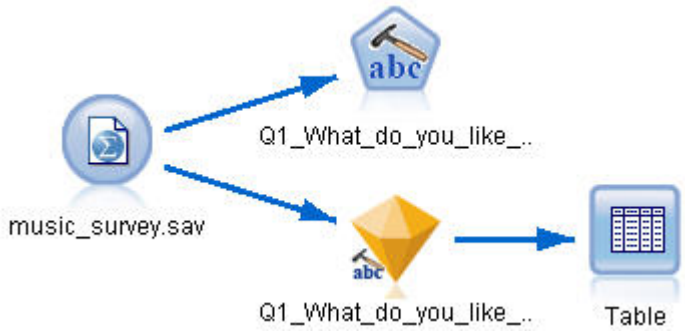
- Karta Fields. Więcej informacji zawiera temat “Model pojęć: karta Fields” na stronie 34.
- Karta Summary. Więcej informacji zawiera temat “Model pojęć: karta Summary” na stronie 35.

Używanie modeli użytkowych kategorii w strumieniu

Model użytkowy kategorii eksploracji tekstu jest generowany z sesji interaktywnego pulpitu roboczego. Można użyć tego modelu użytkowego w strumieniu.

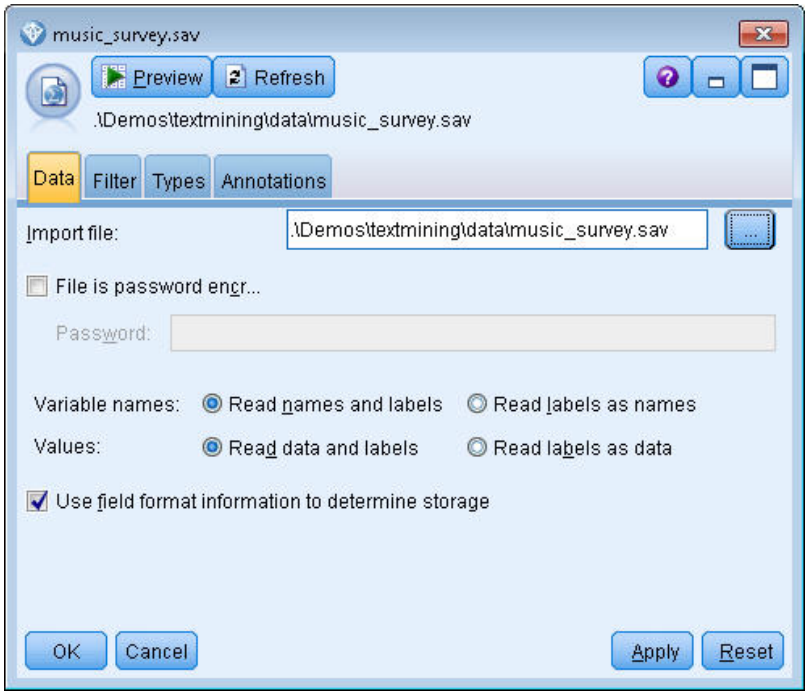
Przykład: węzeł Plik Statistics z modelem użytkowym kategorii

W poniższym przykładzie przedstawiono sposób użycia modeli użytkowych Text Mining.



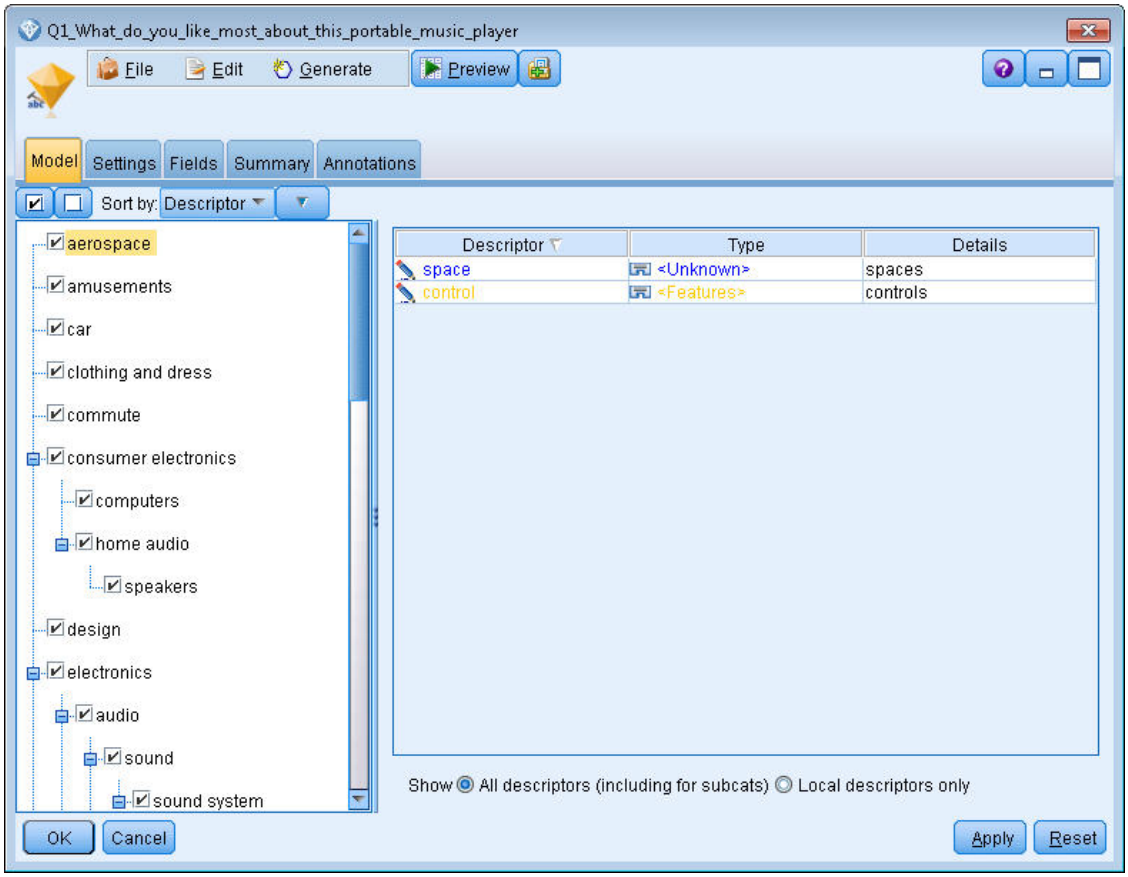
Rysunek 9. Przykładowy strumień: węzeł Plik Statistics z modelem użytkowym kategorii

1. **Węzeł Plik Statistics (karta Data).** Najpierw dodaliśmy ten węzeł do strumienia, aby określić, gdzie są przechowywane dokumenty tekstowe.



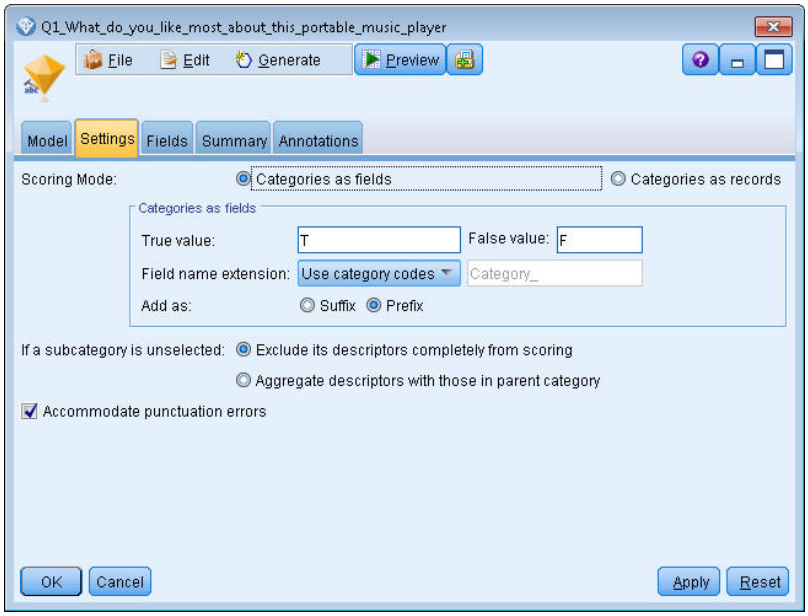
Rysunek 10. Okno dialogowe węzła Plik Statistics: karta Data

2. **Model użytkowy kategorii (karta Model).** Następnie dodaliśmy węzeł modelu użytkowego Kategoria i połączyliśmy go z węzłem Plik Statistics. Wybraliśmy kategorie, które chcieliśmy wykorzystać do oceniania danych.



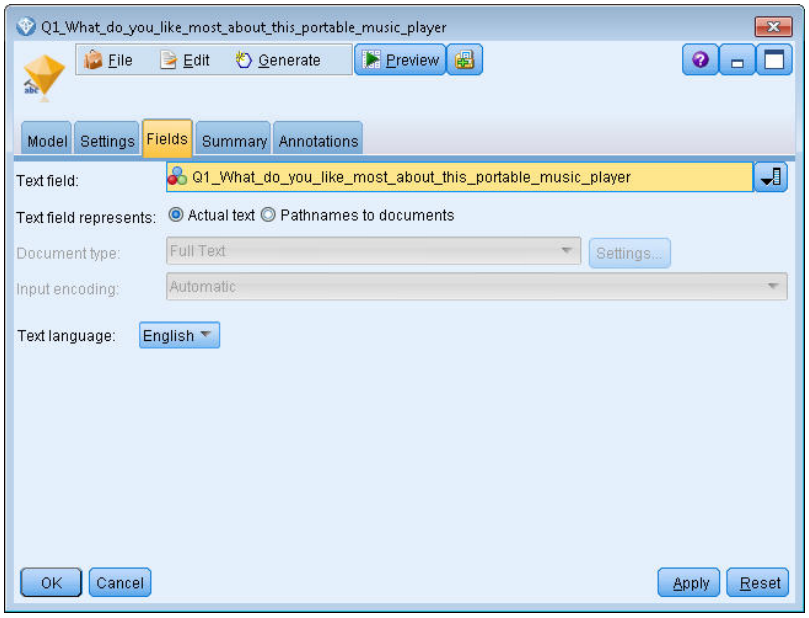
Rysunek 11. Okno dialogowe modelu użytkowego Text Mining: karta Model

3. **Model użytkowy Text Mining (karta Settings).** Następnie zdefiniowaliśmy format wyjściowy **Categories as fields**.



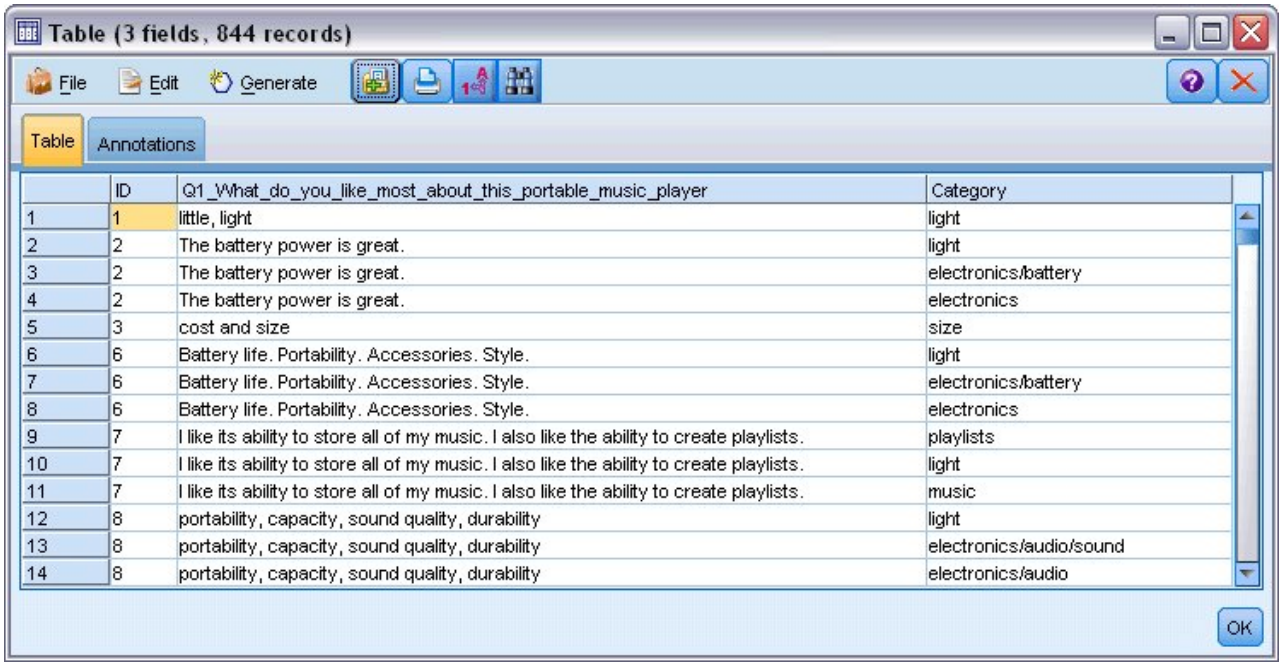
Rysunek 12. Okno dialogowe modelu użytkowego kategorii: karta Settings

4. **Model użytkowy kategorii (karta Fields).** Następnie wybraliśmy zmienną tekstową, czyli nazwę zmiennej z węzła Plik Statistics, i wybraliśmy opcję Text field represents **Actual text** oraz inne ustawienia.



Rysunek 13. Okno dialogowe modelu użytkowego Text Mining: karta Fields

5. **Węzeł Tabela.** Następnie dołączyliśmy węzeł Tabela, aby wyświetlić wyniki i wykonać strumień.



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

Rysunek 14. Wyniki w formie tabeli

Rozdział 4. Eksploracja w poszukiwaniu powiązań w tekście

Węzeł Text Link Analysis

Węzeł Text Link Analysis (TLA) uzupełnia wyodrębnianie pojęć o technikę dopasowywania wzorców w celu zidentyfikowania relacji między pojęciami w danych tekstowych na podstawie znanych wzorców. Te relacje mogą opisywać odczucia klientów wobec produktu, współpracę między firmami, a nawet relacje między genami lub substancjami leczniczymi.

Na przykład samo wyodrębnienie nazwy produktu naszego konkurenta może nie dostarczyć wartościowej wiedzy. Za pomocą tego węzła można jednak dowiedzieć się także, jakie odczucia klienci mają wobec tego produktu, jeśli tylko odpowiednie opinie kryją się w dostępnych danych. Relacje i powiązania są identyfikowane i wyodrębniane przez dopasowywanie znanych wzorców do danych tekstowych.

Można użyć reguł wzorców TLA dostępnych w niektórych szablonach zasobów dostarczanych razem z produktem IBM SPSS Modeler Text Analytics lub utworzyć/edytować własne. Reguły wzorców zawierają makra, listy wyrazów i odstępy między wyrazami, które składają się na zapytanie boolowskie lub regułę stosowaną wobec tekstu wejściowego. Za każdym razem, gdy reguła wzorca TLA znajdzie pasujący tekst, tekst ten może być wyodrębniony jako wynik analizy TLA i zrestrukturyzowany do postaci danych wynikowych. Więcej informacji zawiera Rozdział 18, “Informacje o regułach powiązań w tekście”, na stronie 201.

Węzeł Text Link Analysis oferuje bardziej bezpośredni sposób identyfikowania i wyodrębniania wzorców z tekstu, a także może dodawać wynikowe wzorce powiązań TLA do zbioru danych w strumieniu. Jednak węzeł analizy odsyłaczy tekstowych nie jest jedynym dostępnych sposobem analizy powiązań w tekście. Można również użyć interaktywnego pulpitu roboczego w węźle modelowania Text Mining.

W interaktywnym pulpicie roboczym można eksplorować wynikowe wzorce TLA i użyć ich jako deskryptorów kategorii i/lub wykorzystać do analizy zstępującej wyników i tworzenia wykresów. Więcej informacji zawiera Rozdział 11, “Eksplorowanie analizy powiązań w tekście”, na stronie 141. W istocie wyodrębnianie wyników TLA za pomocą węzła Text Mining jest doskonałym sposobem eksploracji i precyzyjnego dopasowywania szablonów do konkretnych danych — z myślą o późniejszym wykorzystaniu tych szablonów w węźle TLA.

Definicja wyników może składać się maksymalnie z 6 parametrów (części). Wynikowe wzorce w języku japońskim złożone są tylko z jednej lub dwóch części. Więcej informacji zawiera temat “Wyniki z węzła TLA” na stronie 50.

Węzeł ten znajduje się na karcie IBM SPSS Modeler Text Analytics palety węzłów w oknie IBM SPSS Modeler. Więcej informacji zawiera temat “Węzły produktu IBM SPSS Modeler Text Analytics” na stronie 8.

Wymagania. Węzeł Text Link Analysis akceptuje dane tekstowe wczytane do zmiennej za pomocą dowolnego ze standardowych węzłów źródłowych (Baza danych, Plik płaski itd.) lub, także w zmiennej, listę ścieżek do zewnętrznych dokumentów wygenerowaną przez węzeł File List lub Web Feed.

Mocne strony. Węzeł Text Link Analysis wychodzi poza proste wyodrębnianie pojęć, dostarczając informacji na temat relacji *między* pojęciami, a także o powiązanych z danymi opiniach i kwalifikatorach.

Węzeł Text Link Analysis: karta Fields

Karta Fields służy do określenia ustawień dotyczących zmiennych zawierających dane, z których mają być wyodrębniane pojęcia. Można ustawić następujące parametry:

Zmienna identyfikacyjna. Wybierz zmienną zawierającą identyfikator rekordów tekstowych. Identyfikatory muszą być liczbami całkowitymi. Zmienna identyfikacyjna służy jako indeks dla poszczególnych rekordów tekstowych. Użyj zmiennej identyfikacyjnej, jeśli zmienna tekstowa zawiera tekst, który ma być eksplorowany.

Text field. Wybierz zmienną zawierającą tekst do eksploracji. Zmienna ta zależy od źródła danych.

Language field. Wybierz zmienną zawierającą 2-literowy identyfikator języka ISO. Jeśli nie wybierzesz zmiennej, dla każdego dokumentu przyjęty zostanie język na podstawie dostarczonego szablonu.

Document type. Typ dokumentu określa strukturę tekstu. Wybierz jeden z następujących typów:

- **Full text.** Odpowiedni dla większości dokumentów i źródeł tekstowych. Podczas wyodrębniania przeglądany jest cały zbiór tekstu. Z tą opcją, w odróżnieniu od pozostałych, nie są związane żadne ustawienia dodatkowe.
- **Structured text.** Odpowiedni do analizy formularzy bibliograficznych, patentów i innych plików zawierających regularne struktury dające się rozpoznać i przeanalizować. Zastosowanie tego typu dokumentów wiąże się z pominięciem całości lub części procesu wyodrębniania. Umożliwia zdefiniowanie separatorów terminów, przypisywanie typów i określenie wymaganej minimalnej częstotliwości. W przypadku wybrania tej opcji należy kliknąć przycisk **Settings** i wprowadzić separatory tekstu w obszarze **Structured Text Formatting** okna dialogowego Document Settings. Więcej informacji zawiera temat “Ustawienia dokumentów na karcie Zmienne” na stronie 22.

Textual unity. Wybierz jeden z następujących trybów wyodrębniania:

- **Document mode.** Tryb odpowiedni do pracy z krótkimi i semantycznie jednorodnymi dokumentami, takimi jak artykuły prasowe.
- **Paragraph mode.** Odpowiedni do stron WWW i dokumentów bez znaczników. W procesie wyodrębniania dokumenty są semantycznie dzielone na podstawie takich cech wewnętrznych, jak wewnętrzne znaczniki i składnia. Wybranie tego trybu powoduje, że ocenianie odbywa się akapit po akapicie. Dlatego, na przykład, reguła apple & orange jest spełniona tylko wtedy, gdy apple i orange występują w tym samym akapicie.

  Uwaga: Z uwagi na sposób wyodrębniania tekstu z dokumentów PDF tryb **Paragraph mode** nie działa z takimi dokumentami. Wynika to z faktu, że podczas wyodrębniania pomijane są znaki powrotu karetki.

Paragraph mode settings. Ta opcja jest dostępna tylko wtedy, opcję zgodności tekstowej ustawiono na **Paragraph mode**. Określ progi liczby znaków obowiązujące we wszystkich procesach wyodrębniania. Rzeczywisty rozmiar jest zaokrąglany w górę lub w dół do najbliższej kropki. Aby mieć pewność, że asocjacje wyrazów wygenerowane z tekstu zawartego w zbiorze dokumentów są reprezentatywne, należy unikać określania zbyt małego rozmiaru wyodrębniania.

- **Minimum.** Określ minimalną liczbę znaków, jaka ma być używana w każdym wyodrębnianiu.
- **Maksimum.** Określ maksymalną liczbę znaków, jaka ma być używana w każdym wyodrębnianiu.

Copy resources from. Podczas eksploracji tekstu wyodrębnianie prowadzone jest nie tylko na podstawie ustawień na karcie Expert, lecz również na podstawie zasobów lingwistycznych. Zasoby te są podstawą dla przetwarzania tekstu w trakcie wyodrębniania w celu uzyskania z niego pojęć, typów, a także wzorców TLA. Zasoby można kopiować do tego węzła z szablonu zasobu.

Szablon zasobów to predefiniowany zestaw bibliotek i zaawansowanych zasobów lingwistycznych oraz nielingwistycznych, które zostały zoptymalizowane pod kątem konkretnej dziedziny lub zastosowania. Zasoby te są podstawą dla przetwarzania danych w trakcie wyodrębniania. Kliknij opcję **Load** i wybierz szablon, z którego chcesz skopiować zasoby.

Szablony są ładowane zaraz po wybraniu, a nie podczas wykonywania strumienia. Podczas ładowania kopia zasobów zapisywana jest w danym węźle. Jeśli zatem w przyszłości zechcesz użyć zmienionego szablonu lub pakietu, musisz go ponownie załadować. Więcej informacji zawiera temat “Kopiowanie zasobów z szablonów i pakietów TAP” na stronie 26.

Text language. Określa język tekstu poddawanego eksploracji. Zasoby skopiowane w węźle sterują wyświetlanymi opcjami dotyczącymi języka. Wybierz język, w którym zasoby zostały dostosowane.

Węzeł Text Link Analysis: karta Expert

W tym węźle automatycznie włączane jest wyodrębnianie wzorców analizy powiązań w tekście (TLA). Karta Expert zawiera pewne zaawansowane parametry wpływające na sposób wyodrębniania tekstu i postępowania z nim. Parametry w tym oknie dialogowym sterują podstawowym przebiegiem, a także kilkoma funkcjami zaawansowanymi procesu wyodrębniania. Istnieje także szereg zasobów lingwistycznych i opcji wpływających na wyniki wyodrębniania i sterowanych za pośrednictwem wybranego szablonu zasobów.

Informacje dotyczące języka angielskiego, francuskiego, hiszpańskiego, holenderskiego, niemieckiego, portugalskiego i włoskiego

Accommodate punctuation errors. Ta opcja powoduje, że podczas wyodrębniania tekst zawierający błędy interpunkcyjne (na przykład nieprawidłowo użyte znaki interpunkcyjne) będzie tymczasowo normalizowany w celu poprawienia efektywności wyodrębniania pojęć. Ta opcja jest bardzo użyteczna, gdy mamy do czynienia z krótkimi tekstami o niskiej jakości (np. odpowiedziami na pytania otwarte w ankietach, wiadomościami e-mail i danymi z systemów CRM) lub gdy system zawiera wiele skrótów.

Accommodate spelling for a minimum word character length of [n] Ta opcja powoduje zastosowanie techniki grupowania rozmytego, która grupuje błędnie napisane lub podobne wyrazy pod jednym pojęciem. Algorytm grupowania rozmytego tymczasowo usuwa wszystkie samogłoski (z wyjątkiem pierwszej) oraz podwójne/potrójne spółgłoski z wyodrębnianych wyrazów, a następnie porównuje wyniki, by sprawdzić, czy są identyczne. Zatem wyrazy modeling i modelling zostałyby połączone w jedną grupę. Jeśli jednak każdy termin ma przypisany inny typ, z wyjątkiem typu <Unknown>, to grupowanie rozmyte nie będzie stosowane.

Można też określić minimalną liczbę znaków *rdzennych* wymaganą do zastosowania grupowania rozmytego. Liczba znaków rdzennych w terminie obliczana jest poprzez zsumowanie wszystkich znaków i odjęcie znaków tworzących przyrostki przy odmianie, a w wypadku terminów będących wyrazami złożonymi, także znaków tworzących określniki i przyimki. Na przykład termin exercises ma 8 znaków rdzennych w swojej postaci “exercise”, ponieważ litera *s* na końcu tworzy odmianę (w tym przypadku liczbę mnogą). Podobnie, apple sauce ma 10 znaków rdzennych (“apple sauce”), a manufacturing of cars ma 16 znaków rdzennych (“manufacturing car”). Ta metoda liczenia znaków jest stosowana tylko do sprawdzania, czy ma być przeprowadzane grupowanie rozmyte, ale nie wpływa na sposób dopasowywania wyrazów.

Uwaga: Jeśli później okaże się, że pewne wyrazy są grupowane nieprawidłowo, można wykluczyć konkretne pary wyrazów ze stosowania tej techniki, jawnie deklarując je w sekcji **Fuzzy Grouping: Exceptions** na karcie Advanced Resources. Więcej informacji zawiera temat “Grupowanie rozmyte” na stronie 191.

Extract uniterms Ta opcja wyodrębnia pojedyncze wyrazy (terminy pojedyncze), o ile tylko nie są już częścią wyrazu złożonego i są albo rzeczownikami, albo nierozpoznanymi częściami mowy.

Extract nonlinguistic entities Ta opcja wyodrębnia Obiekty nielingwistyczne, takie jak numery telefonów, numery ubezpieczenia społecznego, godziny, daty, waluty, cyfry, wartości procentowe, adresy e-mail i adresy HTTP. W sekcji **Nonlinguistic Entities: Configuration** karty Advanced Resources można uwzględniać i wykluczać określone typy obiektów nielingwistycznych. Wykluczenie zbędnych obiektów sprawi, że mechanizm wyodrębniania nie będzie marnował czasu na ich przetwarzanie. Więcej informacji zawiera temat “Konfiguracja” na stronie 195.

Uppercase algorithm Ta opcja wyodrębnia proste i złożone terminy, które nie figurują we wbudowanych słownikach, o ile pierwsza litera terminu jest wielka. Jest to dobry sposób na wyodrębnienie większości rzeczowników własnych.

Group partial and full person names together when possible Ta opcja grupuje imiona i nazwiska, które w tekście występują w różnych postaciach. Jest to użyteczne, ponieważ imiona i nazwiska często na początku tekstu przytaczane są w pełnym brzmieniu, ale później już występują tylko w wersji skróconej. W przypadku wybrania tej opcji program próbuje dopasować każdy pojedynczy termin typu <Unknown> do ostatniego wyrazu każdego terminu złożonego typu <Person> (osoba). Na przykład, jeśli znaleziony zostanie wyraz *nowak* o początkowo przypisanym typie <Unknown>, to mechanizm wyodrębniania sprawdzi, czy jakiekolwiek terminy złożone typu <Person> zawierają

jako ostatni wyraz właśnie *nowak*, na przykład *piotr nowak*. Ta opcja nie ma zastosowania do imion, ponieważ większość z nich nigdy nie jest wyodrębniana jako termin pojedynczy.

Maximum nonfunction word permutation Ta opcja określa maksymalną liczbę wyrazów niefunkcyjnych, które mogą być obecne, gdy stosowana jest technika permutacji. Technika permutacji grupuje podobne frazy różniące się tylko wyrazami niefunkcyjnymi (na przykład „of" lub „the"), niezależnie od odmiany. Załóżmy na przykład, że wartość ta jest ustawiona na maksymalnie dwa wyrazy i wyodrębniono zarówno termin company officials, jak i termin officials of the company. W tym przypadku oba terminy zostaną połączone w grupę, ponieważ po zignorowaniu wyrazów of the zostaną uznane za identyczne.

Use derivation when grouping multiterms Podczas przetwarzania wielkich zbiorów danych wybierz tę opcję, aby grupować terminy wielowyrazowe według reguł derywacji.

Dla tekstu japońskiego

W przypadku języka japońskiego można wybrać analizator dodatkowy.

Secondary Analysis. Po rozpoczęciu procesu wyodrębniania następuje wyodrębnienie podstawowych słów kluczowych na podstawie domyślnego zestawu typów. Gdy jednak wybierzesz dodatkowy analizator, możesz uzyskać większą liczbę pojęć lub pogłębione pojęcia, ponieważ mechanizm wyodrębniający będzie uwzględniał partykuły i czasowniki pomocnicze jako części pojęcia. W wynikach analizy sentymentu zostanie uwzględnionych wiele dodatkowych typów. Co więcej, wybranie dodatkowego analizatora umożliwi także wygenerowanie wyników analizy powiązań w tekście.

Uwaga: Użycie dodatkowego analizatora wydłuża proces wyodrębniania.

- **Dependency analysis**. Wybranie tej opcji powoduje uwzględnienie dodatkowych partykuł przy wyodrębnianiu podstawowych typów i słów kluczowych. Pozwala także uzyskać wzbogacone wzorce przy analizie powiązań w tekście (TLA).
- **Sentiment analysis**. Wybranie tego analizatora pozwala wyodrębnić dodatkowe pojęcia oraz, tam gdzie to możliwe, wyodrębnić wzorce powiązań TLA. Oprócz typów podstawowych można też korzystać z ponad 80 typów sentymentu. Te typy służą do ujawniania pojęć i wzorców w treściach wyrażających emocje, odczucia i opinie. Dostępne są trzy opcje, które służą do odpowiedniego ukierunkowania analizy sentymentu: **All sentiments**, **Representative sentiment only** i **Conclusions only**.

Wyniki z węzła TLA

Po uruchomieniu węzła Text Link Analysis dane są restrukturyzowane. Ważne jest, aby zrozumieć sposób restrukturyzacji danych w procesie eksploracji tekstu. Jeśli chcesz użyć innej struktury do eksploracji danych, możesz użyć węzłów na palecie Zmienne. Na przykład, jeśli masz dane, w których każdy wiersz reprezentuje jeden rekord tekstowy, to zostanie utworzony jeden wiersz dla każdego wzorca wykrytego w źródłowych danych tekstowych. Każdy wiersz wyników zawiera 15 zmiennych:

- Sześć zmiennych (**Concept#**, na przykład **Concept1**, **Concept2**, ... i **Concept6**) reprezentuje wszystkie pojęcia znalezione w procesie dopasowywania wzorców.
- Sześć zmiennych (**Type#**, such as **Type1**, **Type2**, ... i **Type6**) reprezentuje typy pojęć.
- **Rule Name** reprezentuje nazwę reguły TLA użytej do dopasowania tekstu i wygenerowania wyniku.
- Zmienna o nazwie zmiennej identyfikacyjnej określonej w węźle i reprezentująca rekord lub identyfikator dokumentu przeniesiony wprost z danych wejściowych.
- **Matched Text** przedstawia część danych tekstowych w pierwotnym rekordzie lub dokumencie, która została dopasowana do wzorca TLA.

Uwaga: Reguły TLA dla tekstu japońskiego generują wyniki złożone tylko z jednej lub dwóch części.

Uwaga: Wszystkie istniejące już strumienie zawierające węzeł Text Link Analysis z wersji wcześniejszych niż 5.0 mogą nie być w pełni wykonywalne do momentu zaktualizowania węzłów. Niektóre udoskonalenia wprowadzone w nowszych wersjach produktu IBM SPSS Modeler wymagają zastąpienia starszych węzłów nowszymi wersjami, które są bardziej elastyczne we wdrażaniu i zaawansowane.

Możliwe jest również automatyczne tłumaczenie niektórych języków. Ta funkcja umożliwia eksplorowanie dokumentów w języku, którego użytkownik nie zna. Aby użyć funkcji tłumaczenia, musisz mieć dostęp do odpowiedniego oprogramowania SaaS firmy SDL. Więcej informacji zawiera temat Ustawienia tłumaczenia.

Buforowanie wyników analizy TLA

Buforowanie wyników analizy powiązań w tekście powoduje zapisywanie tych wyników w pamięci podręcznej strumienia. Aby uniknąć powtarzania wyodrębniania wyników analizy TLA za każdym razem, gdy strumień jest wykonywany, wybierz węzeł Text Link Analysis i wybierz z menu opcje **Edit > Node > Cache > Enable**. Przy następnym wykonaniu strumienia wyniki zostaną zbuforowane w węźle. Przy ikonie węzła wyświetlany jest symbol graficzny dokumentu, którego kolor zmienia się z białego na zielony, gdy pamięć podręczna jest wypełniona. Pamięć podręczna jest zachowywana przez cały czas trwania sesji. Aby zachować pamięć podręczną na inny dzień (gdy strumień zostanie zamknięty i ponownie otwarty), wybierz węzeł, a następnie wybierz z menu opcje **Edit > Node > Cache > Save Cache**. Przy następnym otwarciu strumienia można będzie załadować dane z pamięci podręcznej zamiast ponownie uruchamiać tłumaczenie.

Można też zapisać lub włączyć pamięć podręczną węzła, klikając węzeł prawym przyciskiem myszy i wybierając opcję **Cache** z menu kontekstowego.

Korzystanie z węzła Text Link Analysis w strumieniu

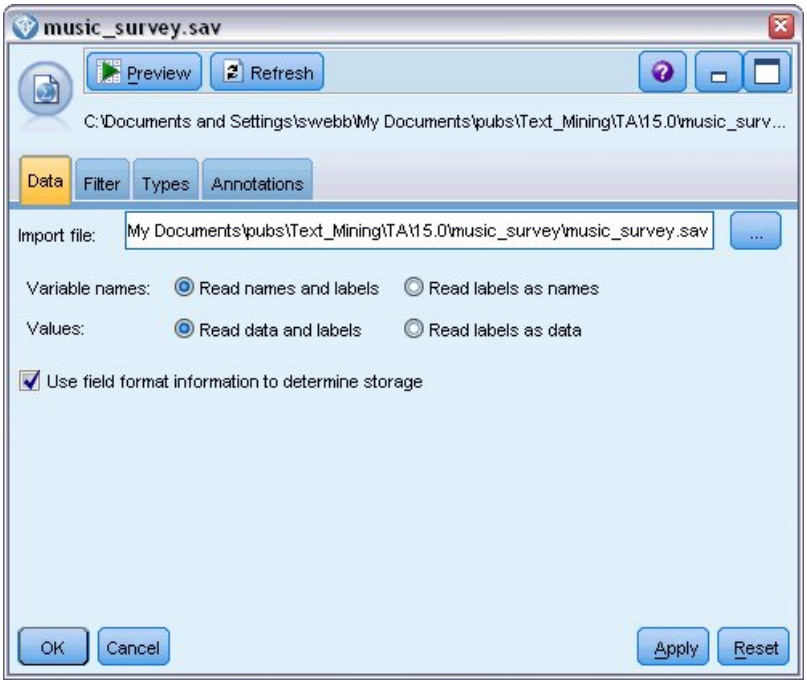
Węzeł Text Link Analysis jest używany w strumieniu do uzyskiwania dostępu do danych i wyodrębniają pojęć. Aby uzyskać dostęp do danych, można wykorzystać dowolny węzeł źródłowy,

Przykład: węzeł Plik Statistics z węzłem Text Link Analysis

W poniższym przykładzie przedstawiono sposób użycia węzła Text Link Analysis.

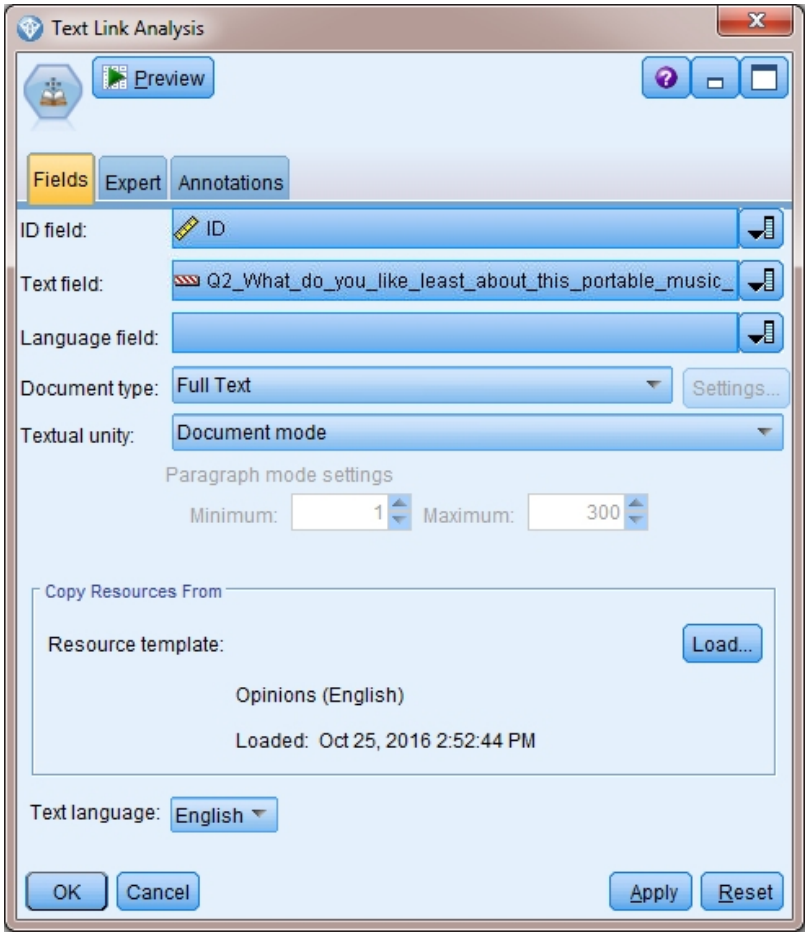
Rysunek 15. Przykład: węzeł Plik Statistics z węzłem Text Link Analysis

1. **Węzeł Plik Statistics (karta Data).** Najpierw dodaliśmy ten węzeł do strumienia, aby określić, gdzie jest przechowywany tekst.



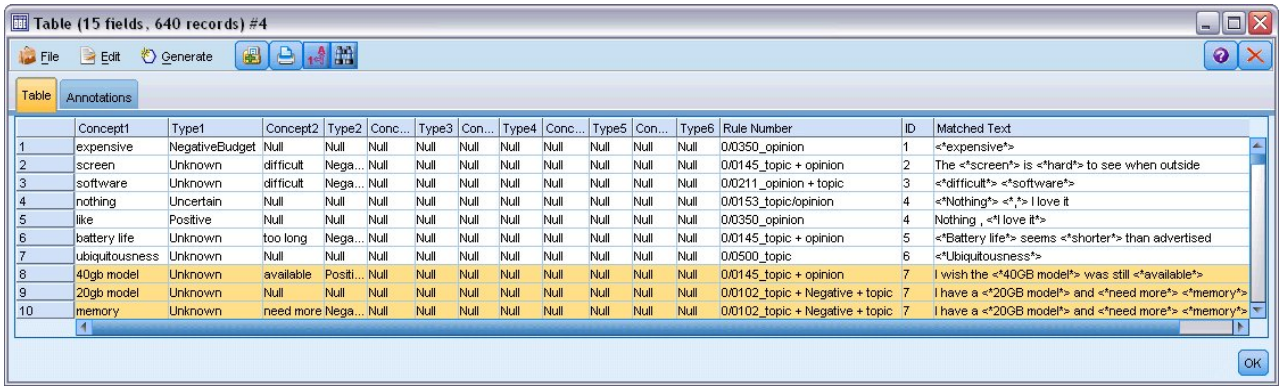
Rysunek 16. Okno dialogowe węzła Plik Statistics: karta Data

2. **Węzeł Text Link Analysis (karta Fields).** Następnie podłączyliśmy ten węzeł do strumienia w celu wyodrębnienia pojęć dla dalszego modelowania lub przeglądania. Określiliśmy zmienną identyfikacyjną i nazwę zmiennej tekstowej zawierającej dane, a także inne ustawienia.



Rysunek 17. Okno dialogowe węzła Text Link Analysis: karta Fields

3. **Węzeł Tabela.** Na koniec dołączyliśmy węzeł Tabela, aby wyświetlić pojęcia, które zostały wyodrębnione z naszych dokumentów tekstowych. W tabeli znajdują się wynikowe wzorce TLA znalezione w danych po wykonaniu tego strumienia z węzłem Text Link Analysis. Niektóre wyniki zawierają tylko jedno dopasowane pojęcie/typ. Inne wyniki są bardziej złożone i zawierają kilka typów i pojęć. Dodatkowo, w wyniku przepuszczenia danych przez węzeł Text Link Analysis i wyodrębnienia pojęć, zmieniło się kilka aspektów danych. Oryginalne dane użyte w tym przykładzie składały się z 8 zmiennych i 405 rekordów. Po wykonaniu węzła Text Link Analysis mamy 15 zmiennych 640 rekordów. Istnieje po jednym wierszu dla każdego wynikowego wzorca TLA. Na przykład ID 7 zmienia się w trzy wiersze, ponieważ wyodrębniono trzy wynikowe wzorce TLA. Można użyć węzła łączenia, aby połączyć te wyniki z oryginalnymi danymi.



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

Rysunek 18. Węzeł wynikowy Tabela

Rozdział 5. Przeglądanie tekstu ze źródła zewnętrznego

Węzeł File Viewer

Eksplorując zbiór dokumentów, można przekazać pełne nazwy ścieżek plików bezpośrednio do węzłów Text Mining. Jednak w przypadku skierowania wyników do węzła Tabela widoczna będzie tylko pełna ścieżka dokumentu zamiast tekstu, który ten dokument zawiera. Węzeł File Viewer może być używany zamiast węzła Tabela i umożliwia dostęp do rzeczywistego tekstu w obrębie każdego z dokumentów bez konieczności scalania ich wszystkich razem w jednym pliku.

Węzeł File Viewer pomaga lepiej zrozumieć wyniki z ekstrakcji tekstu, umożliwiając dostęp do tekstu źródłowego (nieprzetłumaczonego), z którego zostały wyodrębnione pojęcia; bez tego węzła byłby on niedostępny w strumieniu. Ten węzeł dodaje się do strumienia za węzłem File List, aby uzyskać listę odsyłaczy do wszystkich plików.

Wynikiem tego węzła jest okno z listą wszystkich dokumentów, które zostały odczytane i wykorzystane do wyodrębnienia pojęć. W tym oknie można kliknąć odpowiednią ikonę na pasku narzędzi, aby w zewnętrznej przeglądarce wywołać raport z listą nazw dokumentów przekształconych w odsyłacze hipertekstowe. Można kliknąć taki odsyłacz, aby otworzyć odpowiedni dokument. Więcej informacji zawiera temat “Korzystanie z węzła File Viewer” na stronie 56.

Węzeł ten znajduje się na karcie IBM SPSS Modeler Text Analytics palety węzłów w oknie IBM SPSS Modeler. Więcej informacji zawiera temat “Węzły produktu IBM SPSS Modeler Text Analytics” na stronie 8.

Uwaga: Jeśli użytkownik pracuje w trybie klient-serwer i węzły File Viewer są częścią strumienia, zbiory dokumentów muszą być przechowywane w katalogu serwera WWW na serwerze. Ponieważ węzeł wynikowy Text Mining tworzy listę dokumentów przechowywanych w katalogu serwera WWW, ustawienia zabezpieczeń serwera WWW wpływają na uprawnienia dostępu do tych dokumentów.

Ustawienia węzła File Viewer

Można określić następujące ustawienia węzła File Viewer.

Document field. Wybierz zmienną należącą do danych, która zawiera pełną nazwę i ścieżkę dokumentów, które mają być wyświetlane.

Title for generated HTML page. Utwórz tytuł, który ma być wyświetlany w górnej części strony zawierającej listę dokumentów.

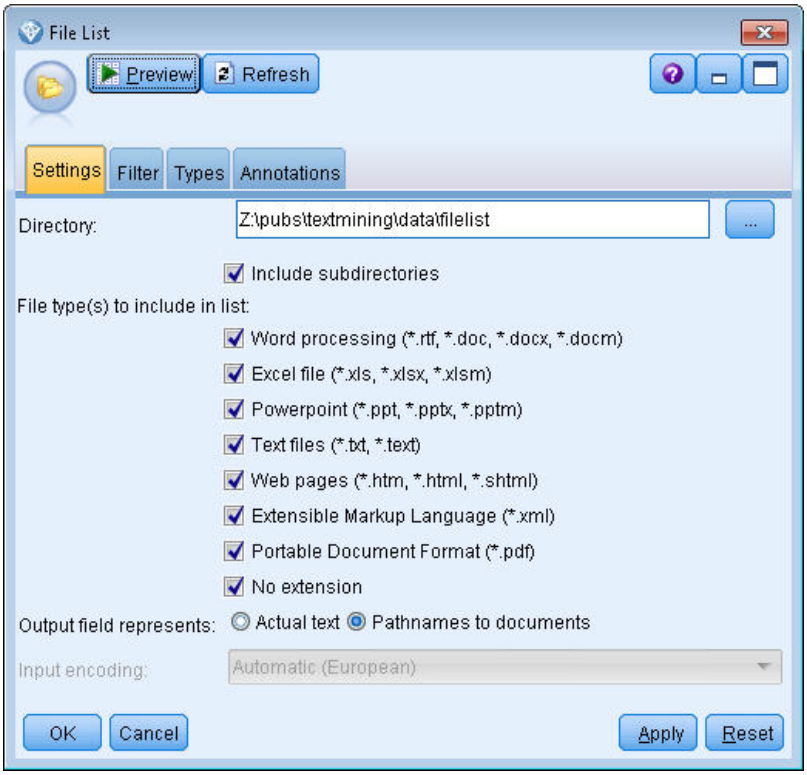
Korzystanie z węzła File Viewer

Poniższy przykład ilustruje zastosowanie węzła File Viewer.

Przykład: węzeł File List i węzeł File Viewer

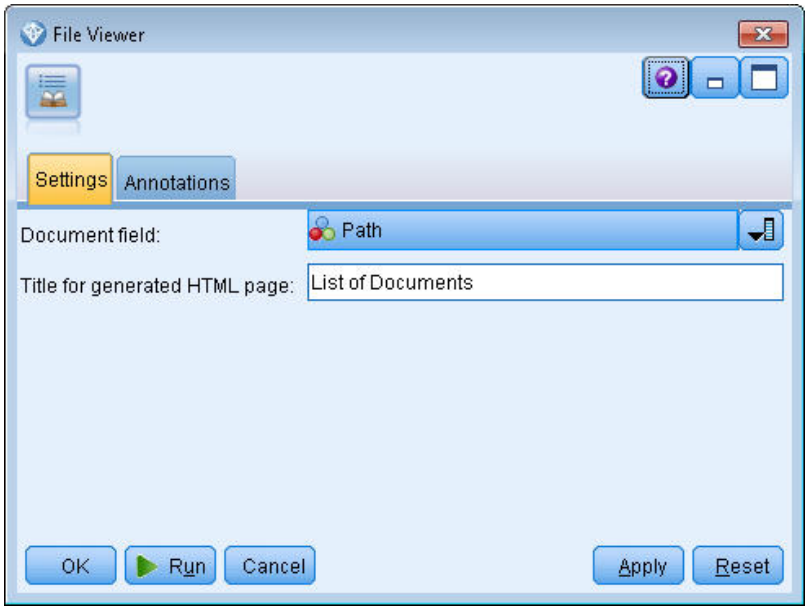
Rysunek 19. Strumień ilustrujący użycie węzła File Viewer

1. **Węzeł File List (karta Settings).** Najpierw dodaliśmy ten węzeł, aby określić, gdzie są przechowywane dokumenty tekstowe.



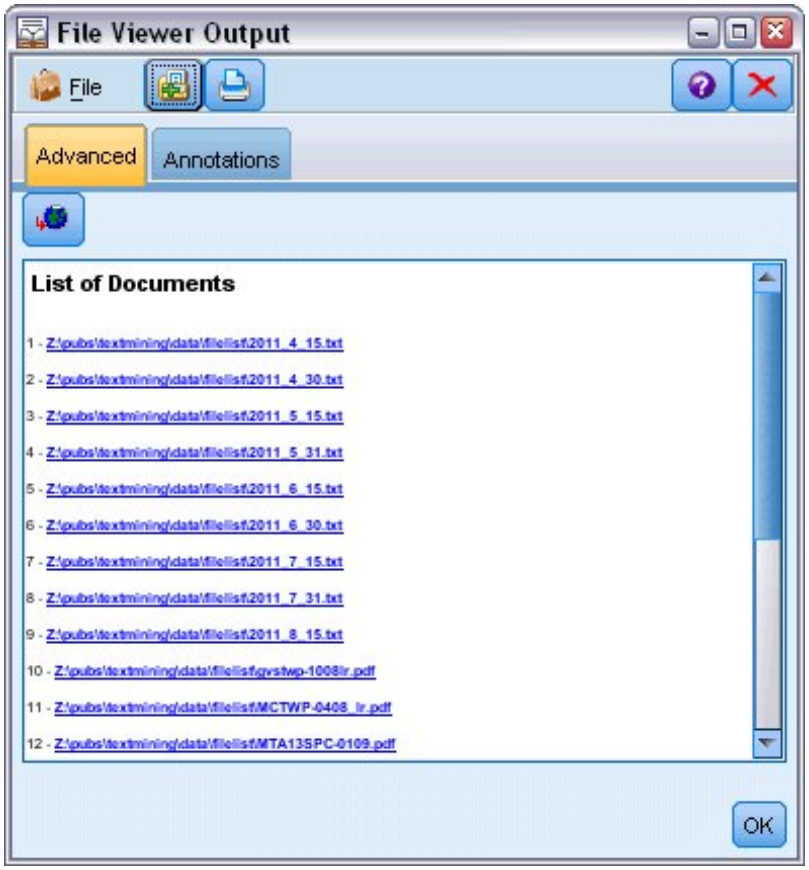
Rysunek 20. Okno dialogowe węzła File List: karta Settings

2. **Węzeł File Viewer (karta Settings).** Następnie dołączyliśmy węzeł File Viewer, aby utworzyć listę dokumentów w formacie HTML.



Rysunek 21. Okno dialogowe węzła File Viewer: karta Settings

3. **Okno dialogowe File Viewer Output.** Następnie uruchomiliśmy strumień, który wygenerował listę dokumentów w nowym oknie.



Rysunek 22. File Viewer Output

4. Aby wyświetlić dokumenty, kliknęliśmy na pasku narzędzi przycisk kuli ziemskiej z czerwoną strzałką. Spowodowało to otwarcie listy odsyłaczy do dokumentów w przeglądarce.

Rozdział 6. Właściwości węzłów używane w skryptach

IBM SPSS Modeler udostępnia język skryptowy, który umożliwia wykonywanie strumieni z poziomu wiersza komend. W tym miejscu można zapoznać się z informacjami na temat właściwości węzłów dostarczonych z produktem IBM SPSS Modeler Text Analytics. Więcej informacji na temat standardowego zestawu węzłów dostarczonych z produktem IBM SPSS Modeler zawiera Podręcznik tworzenia skryptów i automatyzacji.

Węzeł File List: filelistnode

Właściwości przedstawione w tabeli poniżej można stosować do tworzenia skryptów. Sam węzeł nosi nazwę filelistnode.

Tabela 7. Właściwości węzła File List używane w skryptach

Właściwości skryptów	Typ danych
path	*string*
recurse	*flag*
word_processing	*flag*
excel_file	*flag*
powerpoint_file	*flag*
text_file	*flag*
web_page	*flag*
xml_file	*flag*
pdf_file	*flag*
no_extension	*flag*

Uwaga: Parametr „Create list” nie jest już dostępny, a wszelkie skrypty zawierające tę opcję zostaną automatycznie przekształcone w skrypty generujące pliki (Files).

Węzeł Web Feed: webfeednode

Właściwości przedstawione w tabeli poniżej można stosować do tworzenia skryptów. Sam węzeł nosi nazwę webfeednode.

Tabela 8. Właściwości węzła Web Feed używane w skryptach

Właściwości skryptów	Typ danych	Opis właściwości
urls	*string1 string2 ...stringn*	Każdy adres URL jest określony w strukturze listy. Lista adresów URL jest rozdzielana znakami “\n”
recent_entries	*flag*	
limit_entries	*integer*	Liczba najnowszych wpisów do odczytu na adres URL.
use_previous	*flag*	Umożliwia zapisanie i ponowne użycie buforu Web Feed.
use_previous_label	*string*	Nazwa zapisanego bufora WWW.
start_record	*string*	Znacznik początku inny niż RSS.
url *n* .title	*string*	Dla każdego adresu URL na liście należy określić jeden w tej właściwości. Pierwszym będzie url1.title, przy czym numer jest zgodny z pozycją na liście adresów URL. Jest to znacznik początkowy zawierający tytuł treści.

Tabela 8. Właściwości węzła Web Feed używane w skryptach (kontynuacja)

Właściwości skryptów	Typ danych	Opis właściwości
url *n* .short_description	*string*	Tak samo, jak w przypadku url *n* .title.
url *n* .short_description	*string*	Tak samo, jak w przypadku url *n* .title.
url *n* .authors	*string*	Tak samo, jak w przypadku url *n* .title.
url *n* .contributors	*string*	Tak samo, jak w przypadku url *n* .title.
url *n* .published_date	*string*	Tak samo, jak w przypadku url *n* .title.
url *n* .modified_date	*string*	Tak samo, jak w przypadku url *n* .title.
html_alg	None HTMLCleaner	Metoda filtrowania zawartości.
discard_lines	*flag*	Powoduje odrzucenie krótkich wierszy. Używane z min_words
min_words	*integer*	Minimalna liczba słów.
discard_words	*flag*	Powoduje odrzucenie krótkich wierszy. Używane z min_avg_len
min_avg_len	*integer*	
discard_scw	*flag*	Powoduje odrzucenie wierszy zawierających wiele słów jednoznakowych. Używane z max_scw
max_scw	*integer*	Maksymalna proporcja od 0 do 100 procent słów jednoznakowych na wiersz
discard_tags	*flag*	Powoduje odrzucenie wierszy zawierających konkretne znaczniki.
tags	*string*	W przypadku słów specjalnych wymagane jest zastosowanie ukośnika odwrotnego (\) jako znaku zmiany znaczenia.
discard_spec_words	*flag*	Powoduje odrzucenie wierszy zawierających określone łańcuchy.
words	*string*	W przypadku słów specjalnych wymagane jest zastosowanie ukośnika odwrotnego (\) jako znaku zmiany znaczenia.

Węzeł języka: languageidentifier

Właściwości przedstawione w tabeli poniżej można stosować do tworzenia skryptów. Sam węzeł nosi nazwę languageidentifier.

Tabela 9. Właściwości węzła języka używane w skryptach

Właściwości skryptów	Typ danych	Opis właściwości
text	*field*	
language_field_name	*string*	Nazwa zmiennej, która jest generowana w wyniku.
unidentified_language_value	Undefined Supported Custom	Wartość domyślna, która ma być używana, gdy język nie zostanie zidentyfikowany.
unidentified_language_supported	en de es fr it ja nl pt	Kod Iso. Dostępny tylko wtedy, gdy atrybut unidentified_language_value ma wartość Supported.

Tabela 9. Właściwości węzła języka używane w skryptach (kontynuacja)

Właściwości skryptów	Typ danych	Opis właściwości
unidentified_language_custom	*string*	Dostępny tylko wtedy, gdy atrybut unidentified_language_value ma wartość Custom.

Węzeł Text Mining: TextMiningWorkbench

Można użyć następujących parametrów, aby zdefiniować lub zaktualizować węzeł za pomocą skryptu. Sam węzeł nosi nazwę TextMiningWorkbench.

Ważne: Nie można określić innego szablonu zasobu za pośrednictwem skryptów. Jeśli uważasz, że potrzebujesz szablonu, należy go wybrać w oknie dialogowym węzła.

Tabela 10. Właściwości węzła modelowania Text Mining używane w skryptach

Właściwości skryptów	Typ danych	Opis właściwości
text	*field*	
method	ReadText ReadPath	
docType	*integer*	Możliwe wartości to (0,1,2), gdzie 0 = pełnotekstowe, 1 = tekst ustrukturyzowany, a 2 = XML
encoding	Automatic "UTF-8" "UTF-16" "ISO-8859-1" "US-ASCII" "CP850" "EUC-JP" "SHIFT-JIS" "ISO2022-JP"	Należy zwrócić uwagę, że wartości ze znakami specjalnymi, takie jak "UTF-8", powinny być ujęte w cudzysłów, aby znaki nie zostały potraktowane jak operatory matematyczne.
unity	*integer*	Możliwe wartości to (0,1), gdzie 0 = tryb akapitów, a 1 = tryb dokumentów
para_min	*integer*	
para_max	*integer*	
mtag	*string*	Zawiera wszystkie ustawienia mtag (z okna dialogowego ustawień dla plików XML)
mclef	*string*	Zawiera wszystkie ustawienia mclef (z okna dialogowego ustawień dla plików tekstowych ustrukturyzowanych)
partition	*field*	
custom_field	*flag*	Wskazuje, czy zostanie określona zmienna dzieląca na podzbiory.
use_model_name	*flag*	
model_name	*string*	
use_partitioned_data	*flag*	Jeśli zdefiniowana jest zmienna dzieląca na podzbiory, do budowania modelu używane są wyłącznie dane uczące.
model_output_type	Interactive Model	Interactive powoduje generowanie modelu kategorii. Model powoduje generowanie modelu pojęć.

Tabela 10. Właściwości węzła modelowania Text Mining używane w skryptach (kontynuacja)

Właściwości skryptów	Typ danych	Opis właściwości
use_interactive_info	*flag*	Tylko do tworzenia interaktywnego w sesji pulpitu roboczego
reuse_extraction_results	*flag*	Tylko do tworzenia interaktywnego w sesji pulpitu roboczego
interactive_view	Categories TLA Clusters	Tylko do tworzenia interaktywnego w sesji pulpitu roboczego
extract_top	*integer*	Ten parametr jest używany, gdy model_type = Concept
use_check_top	*flag*	
check_top	*integer*	
use_uncheck_top	*flag*	
uncheck_top	*integer*	
język	de en es fr it ja nl pt	
frequency_limit	*integer*	Nieaktualna w wersji 14.0.
concept_count_limit	*integer*	Ogranicza wyodrębnianie do pojęć, których globalna liczebność równa jest co najmniej tej wartości. Niedostępna w przypadku tekstu japońskiego
fix_punctuation	*flag*	Niedostępna w przypadku tekstu japońskiego
fix_spelling	*flag*	Niedostępna w przypadku tekstu japońskiego
spelling_limit	*integer*	Niedostępna w przypadku tekstu japońskiego
extract_uniterm	*flag*	Niedostępna w przypadku tekstu japońskiego
extract_nonlinguistic	*flag*	Niedostępna w przypadku tekstu japońskiego
upper_case	*flag*	Niedostępna w przypadku tekstu japońskiego
group_names	*flag*	Niedostępna w przypadku tekstu japońskiego
permutation	*integer*	Maksymalna liczba permutacji wyrazu niefunkcyjnego (domyślnie 3). Niedostępna w przypadku tekstu japońskiego.
jp_algorithmset Tylko wnioski Tylko reprezentatywne Wszystkie sentymenty	0 1 2	Dotyczy tylko wyodrębniania z tekstu japońskiego. 0 = Dodatkowe wyodrębnianie informacji o sentymencie 1 = Wyodrębnianie zależności 2 = Bez dodatkowej analizy
jp_algorithm_sense_mode	0 1 2	Dotyczy tylko wyodrębniania z tekstu japońskiego. 0 = Tylko wnioski 2 = Tylko reprezentatywne 3 = Wszystkie sentymenty

Model użytkowy Text Mining: TMWBModelApplier

Właściwości przedstawione w tabeli poniżej można stosować do tworzenia skryptów. Model użytkowy nosi nazwę TMWBModelApplier.

Tabela 11. Właściwości modelu użytkowego Text Mining

Właściwości skryptów	Typ danych	Opis właściwości
scoring_mode	Fields Records	
field_values	Flags Counts	Ta opcja nie jest dostępna w modelu użytkowym kategorii. Dla właściwości Flags ustaw wartość TRUE lub FALSE.
true_value	*string*	Za pośrednictwem właściwości Flags zdefiniuj wartość oznaczającą true.
false_value	*string*	Za pośrednictwem właściwości Flags zdefiniuj wartość oznaczającą false.
extension_concept	*string*	Określ rozszerzenie nazwy zmiennej. Nazwy zmiennych generowane są poprzez połączenie nazwy pojęcia z rozszerzeniem. Określ miejsce, w którym ma zostać umieszczone to rozszerzenie, przy użyciu wartości add_as.
extension_category	*string*	Rozszerzenie nazwy zmiennej. Można wybrać opcję określania przedrostka/przyrostka (rozszerzenia) nazwy zmiennej lub użyć kodów kategorii. Nazwy zmiennych generowane są poprzez połączenie nazwy kategorii z rozszerzeniem. Określ miejsce, w którym ma zostać umieszczone to rozszerzenie, przy użyciu wartości add_as.
add_as	Suffix Prefix	
fix_punctuation	*flag*	
excluded_subcategories_descriptors	RollUpToParent Ignore	Tylko w przypadku modeli kategorii. Jeśli podkategoria nie jest wybrana. Ta opcja umożliwia określenie, w jaki sposób traktowane będą deskryptory należące do podkategorii niewybranych do oceny. Dostępne są dwie opcje. • Ignore. Opcja Exclude its descriptors completely from scoring powoduje, że deskryptory niezaznaczonych podkategorii będą ignorowane i nie będą używane podczas oceniania. • RollUpToParent. Opcja Aggregate descriptors with those in parent category powoduje, że deskryptory niezaznaczonych podkategorii będą używane jako deskryptory kategorii nadrzędnej wobec tych podkategorii. Jeśli niezaznaczone są kategorie na kilku poziomach, deskryptory będą agregowane aż do pierwszej dostępnej kategorii nadrzędnej
check_model	*flag*	Nieaktualna w wersji 14
text	*field*	
method	ReadText ReadPath	
docType	*integer*	Możliwe wartości to (0,1,2), gdzie 0 = pełnotekstowe, 1 = tekst ustrukturyzowany, a 2 = XML

Tabela 11. Właściwości modelu użytkowego Text Mining (kontynuacja)

Właściwości skryptów	Typ danych	Opis właściwości
encoding	Automatic "UTF-8" "UTF-16" "ISO-8859-1" "US-ASCII" "CP850" "EUC-JP" "SHIFT-JIS" "ISO2022-JP"	Należy zwrócić uwagę, że wartości ze znakami specjalnymi, takie jak "UTF-8", powinny być ujęte w cudzysłów, aby znaki nie zostały potraktowane jak operatory matematyczne.
język	de en es fr it ja nl pt	

Węzeł Text Link Analysis: textlinkanalysis

Można użyć parametrów z poniższej tabeli, aby zdefiniować lub zaktualizować węzeł za pomocą skryptu. Sam węzeł nosi nazwę textlinkanalysis.

Ważne! Nie można określić szablonu zasobów za pośrednictwem skryptu. Aby wybrać szablon, należy zrobić to w oknie dialogowym węzła.

Tabela 12. Właściwości węzła Text Link Analysis (TLA) używane w skryptach

Właściwości skryptów	Typ danych	Opis właściwości
id_field	*field*	
text	*field*	
method	ReadText ReadPath	
docType	*integer*	Możliwe wartości to (0,1,2), gdzie 0 = pełnotekstowe, 1 = tekst ustrukturyzowany, a 2 = XML
encoding	Automatic "UTF-8" "UTF-16" "ISO-8859-1" "US-ASCII" "CP850" "EUC-JP" "SHIFT-JIS" "ISO2022-JP"	Należy zwrócić uwagę, że wartości ze znakami specjalnymi, takie jak "UTF-8", powinny być ujęte w cudzysłów, aby znaki nie zostały potraktowane jak operatory matematyczne.
unity	*integer*	Możliwe wartości to (0,1), gdzie 0 = tryb akapitów, a 1 = tryb dokumentów
para_min	*integer*	
para_max	*integer*	
mtag	*string*	Zawiera wszystkie ustawienia mtag (z okna dialogowego ustawień dla plików XML)
mclef	*string*	Zawiera wszystkie ustawienia mclef (z okna dialogowego ustawień dla plików tekstowych ustrukturyzowanych)

Tabela 12. Właściwości węzła Text Link Analysis (TLA) używane w skryptach (kontynuacja)

Właściwości skryptów	Typ danych	Opis właściwości
język	de en es fr it ja nl pt	
concept_count_limit	*integer*	Ogranicza wyodrębnianie do pojęć, których globalna liczebność równa jest co najmniej tej wartości. Niedostępna w przypadku tekstu japońskiego
fix_punctuation	*flag*	Niedostępna w przypadku tekstu japońskiego
fix_spelling	*flag*	Niedostępna w przypadku tekstu japońskiego
spelling_limit	*integer*	Niedostępna w przypadku tekstu japońskiego
extract_uniterm	*flag*	Niedostępna w przypadku tekstu japońskiego
extract_nonlinguistic	*flag*	Niedostępna w przypadku tekstu japońskiego
upper_case	*flag*	Niedostępna w przypadku tekstu japońskiego
group_names	*flag*	Niedostępna w przypadku tekstu japońskiego
permutation	*integer*	Maksymalna liczba permutacji wyrazu niefunkcyjnego (domyślnie 3). Niedostępna w przypadku tekstu japońskiego.
jp_algorithmset Tylko wnioski Tylko reprezentatywne Wszystkie sentymenty	0 1 2	Dotyczy tylko wyodrębniania z tekstu japońskiego. 0 = Dodatkowe wyodrębnianie informacji o sentymencie 1 = Wyodrębnianie zależności 2 = Bez dodatkowej analizy
jp_algorithm_sense_mode	0 1 2	Dotyczy tylko wyodrębniania z tekstu japońskiego. 0 = Tylko wnioski 2 = Tylko reprezentatywne 3 = Wszystkie sentymenty

Rozdział 7. Tryb pracy z interaktywnym pulpitem roboczym

Z poziomu węzła modelowania Text Mining można uruchomić sesję interaktywnego pulpitu roboczego w trakcie wykonywania strumienia. Na tym pulpicie można wyodrębniać kluczowe pojęcia z danych tekstowych, budować kategorie i eksplorować wzorce analizy powiązań w tekście oraz skupienia, a także generować modele kategorii. W tym rozdziale omawiamy interfejs pulpitu roboczego z perspektywy wysokopoziomowej oraz przedstawiamy najważniejsze elementy, z którymi pracują użytkownicy, takie jak:

- **Wyniki wyodrębniania.** Po przeprowadzeniu wyodrębniania są to kluczowe wyrazy i frazy wyodrębnione z danych tekstowych, nazywane także *pojęciami*. Pojęcia te są pogrupowane w *typy*. Korzystając z tych pojęć i typów, można eksplorować dane i tworzyć kategorie. Elementami tymi zarządza się w widoku **Categories and Concepts**.
- **Kategorie.** Korzystając z deskryptorów (takich jak wyniki wyodrębniania, wzorce i reguły) jako definicji, można ręcznie lub automatycznie utworzyć zestaw kategorii, do których będą przypisywane rekordy w zależności od tego, czy zawierają część definicji kategorii, czy nie. Elementami tymi zarządza się w widoku **Categories and Concepts**.
- **Skupienia.** *Skupienia* są to grupy pojęć, między którymi wykryto powiązane świadczące o istnieniu relacji. Pojęcia są grupowane przy użyciu złożonego algorytmu, który bierze pod uwagę, między innymi, jak często dwa pojęcia występują razem, a jak często występują osobno. Elementami tymi zarządza się w widoku **Clusters**. Pojęcia tworzące skupienie można też dodać do kategorii.
- **Wzorce analizy powiązań w tekście.** Jeśli zasoby lingwistyczne zawierają reguły wzorców powiązań w tekście lub próbujesz używać szablonu zasobów, który zawiera już jakieś reguły TLA, możesz wyodrębniać wzorce z danych tekstowych. Wzorce te pomagają w ujawnianiu interesujących relacji między pojęciami w danych. Można także użyć ich jako deskryptorów w definicjach kategorii. Zarządza się nimi w widoku **Text Link Analysis**. W przypadku tekstu w języku japońskim, należy wybrać dodatkowy analizator i włączyć wyodrębnianie TLA
- **Zasoby lingwistyczne.** Proces wyodrębniania realizowany jest w oparciu o zestaw parametrów i definicji lingwistycznych, które rządzą sposobem wyodrębniania pojęć z tekstu. Tymi parametrami i definicjami zarządza się w ramach szablonów i bibliotek w widoku **Resource Editor**.

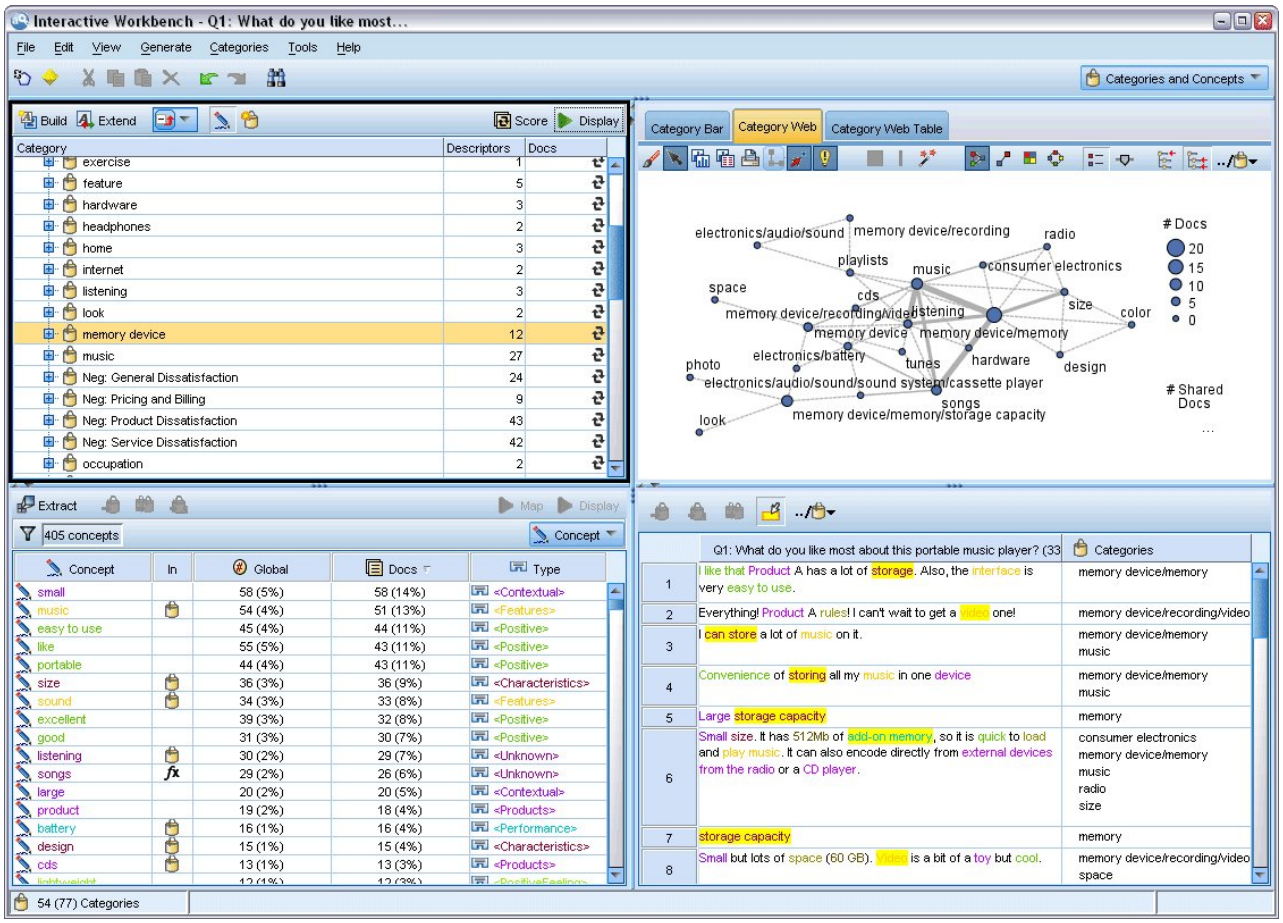
Potencjalne problemy dotyczące interaktywnego pulpitu roboczego

- Wiele interaktywnych pulpitów roboczych może spowalniać działanie programu. Podczas pracy z interaktywnym pulpitem roboczym SPSS Modeler Text Analytics i SPSS Modeler współużytkują wspólny mechanizm środowiska wykonawczego Java. W zależności od liczby otwartych interaktywnych pulpitów roboczych podczas sesji programu SPSS Modeler pamięć systemowa może spowalniać działanie aplikacji, nawet w przypadku uruchamiania i zamykania tej samej sesji. Efekt jest szczególnie wyraźny gdy użytkownik pracuje z dużą ilością danych lub gdy komputer dysponuje mniejszą ilością pamięci RAM niż zalecane 4 GB. Jeśli komputer wolno reaguje, zaleca się zapisać wszystkie zmiany, zamknąć program SPSS Modeler i ponownie uruchomić aplikację. uruchamianie produktu SPSS Modeler Text Analytics na komputerze z mniejszą ilością pamięci niż zalecana, zwłaszcza podczas pracy z dużymi zestawami danych lub przez dłuższy czas, może spowodować wyczerpanie zasobów pamięciowych środowiska wykonawczego Java i wyłączenie systemu. W przypadku pracy z dużą ilością danych zdecydowanie zaleca się rozbudowę pamięci do zalecanego lub wyższego poziomu (albo wykorzystanie produktuSPSS Modeler Text Analytics Server).
- Klientowi programu SPSS Modeler może zabraknąć pamięci w przypadku uruchomienia wielu interaktywnych pulpitów roboczych programu SPSS Modeler Text Analytics bez restartowania aplikacji. Należy monitorować użycie pamięci w wierszu statusu i, jeśli jej poziom jest niski, zamknąć i ponownie otworzyć klienta programu SPSS Modeler.

Widok Categories and Concepts

Interfejs aplikacji składa się z kilku różnych widoków. Widok Categories and Concepts to okno, w którym można tworzyć i eksplorować kategorie, a także eksplorować i optymalizować wyniki wyodrębniania. *Kategorie* są to grupy ściśle pokrewnych pojęć i wzorców, do których dokumenty i rekordy są przypisywane w procesie oceniania. Natomiast *pojęcia* są to najbardziej podstawowe wyniki wyodrębniania, których można używać jako deskryptorów będących

elementami składowymi kategorii.



Rysunek 23. Widok Categories and Concepts

Widok Categories and Concepts jest podzielony na cztery panele, a każdy z nich można ukrywać lub uwidaczniać, wybierając jego nazwę z menu View. Więcej informacji zawiera Rozdział 9, “Kategoryzacja danych tekstowych”, na stronie 93.

Panel Categories

Ten znajdujący się w lewym górnym rogu obszar przedstawia tabelę, w której można zarządzać wszystkimi zbudowanymi kategoriami. Po zakończeniu wyodrębniania pojęć i typów z danych tekstowych można rozpocząć automatyczne tworzenie kategorii za pomocą takich technik, jak sieci semantyczne, włączanie pojęć itp., albo ręcznie. Jeśli klikniesz dwukrotnie nazwę kategorii, zostanie otwarte okno dialogowe Category Definitions z listą wszystkich deskryptorów (takich jak pojęcia, typy i reguły), które składają się na definicję kategorii. Więcej informacji zawiera temat Rozdział 9, “Kategoryzacja danych tekstowych”, na stronie 93. Nie wszystkie techniki automatyczne są dostępne dla wszystkich języków.

Po wybraniu wiersza w panelu w panelach Data i Visualization można wyświetlić informacje na temat odpowiednich dokumentów/rekordów lub deskryptorów.

Panel Extraction Results

Ten obszar znajdujący się w lewym dolnym rogu przedstawia wyniki wyodrębniania. Po uruchomieniu wyodrębniania mechanizm wyodrębniania odczytuje dane tekstowe, identyfikuje istotne pojęcia i przypisuje każdemu z nich typ. *Pojęcia* to wyrazy lub frazy wyodrębnione z danych tekstowych. *Typy* są semantycznymi zbiorami pojęć zapisanymi w

słownikach typów. Po zakończeniu wyodrębniania pojęcia i typy wyświetlane są w różnych kolorach na panelu Extraction Results. Więcej informacji zawiera “Wyniki wyodrębniania: pojęcia i typy” na stronie 79.

Zestaw terminów bazowych danego pojęcia można wyświetlić, zatrzymując wskaźnik myszy nad nazwą pojęcia. Spowoduje to wyświetlenie podpowiedzi z nazwą pojęcia i maksymalnie kilkoma wierszami terminów zgrupowanych pod tym pojęciem. Do tych bazowych terminów należą synonimy zdefiniowane w zasobach lingwistycznych (niezależnie od tego, czy wystąpiły w tekście, czy nie), a także wszelkie wyodrębnione terminy w liczbie mnogiej/pojedynczej, permutacje terminów, terminy wygenerowane w wyniku grupowania rozmytego itd. Można skopiować te terminy bazowe, by wyświetlić ich kompletną listę. Należy w tym celu kliknąć nazwę pojęcia prawym przyciskiem myszy i wybrać opcję z menu kontekstowego.

Eksploracja tekstu jest procesem iteracyjnym, w którym wyniki wyodrębniania są przeglądane z uwzględnieniem kontekstu danych tekstowych, optymalizowane w celu wygenerowania nowych wyników i ponownie oceniane. Wyniki wyodrębniania można optymalizować, modyfikując zasoby lingwistyczne. Tę optymalizację można przeprowadzić bezpośrednio w panelu Extraction Results lub Data, ale także bezpośrednio w widoku Resource Editor. Więcej informacji zawiera temat “Widok Resource Editor” na stronie 73.

Uwaga: Jeśli w widocznym panelu mieści się więcej wyników, można użyć przycisków znajdujących się w dolnej części panelu, aby przeglądać wyniki, lub wprowadzić numer strony i przejść do niej.

Panel Visualization

Ten obszar, który znajduje się w prawym górnym rogu, z różnych perspektyw przedstawia wspólne elementy kategoryzacji dokumentów/rekordów. Każdy wykres zawiera podobne informacje, ale zaprezentowane w inny sposób i na innym poziomie szczegółowości. Wykresy umożliwiają analizowanie wyników kategoryzacji i pomagają w optymalizacji kategorii lub raportów. Na przykład na wykresie można wykryć kategorie zbyt podobne (na przykład mające więcej niż 75% wspólnych dokumentów lub rekordów) lub zbyt odmienne. Zawartość wykresu odpowiada wyborom dokonanym w innych panelach. Więcej informacji zawiera temat “Wykresy i tabele kategorii” na stronie 147.

Panel Data

Panel Data znajduje się w prawym dolnym rogu. Jest on tabelą zawierającą dokumenty lub rekordy odpowiadające elementom wybranym w innym obszarze widoku. W panelu Data widoczny jest tekst zależny od tego, co wybrano. Po dokonaniu wyboru kliknij przycisk **Display**, aby zapełnić panel Data odpowiednim tekstem.

Jeśli w innym panelu dokonano wyboru, to w odpowiednich dokumentach lub rekordach pojęcia będą wyróżnione kolorem, aby łatwiej było je odszukać w tekście. Można także zatrzymać wskaźnik myszy nad elementem oznaczonym kolorem, aby wyświetlić podpowiedź z nazwą pojęcia, pod którym dany termin został wyodrębniony, oraz typem, do którego został przypisany. Więcej informacji zawiera temat “Panel Data” na stronie 101.

Wyszukiwanie i znajdowanie w widoku Categories and Concepts

W niektórych przypadkach konieczne jest szybkie wyszukanie informacji w określonej sekcji. Za pomocą narzędzi wyszukiwania można wprowadzić łańcuch do wyszukania i zdefiniować inne kryteria wyszukiwania, takie jak rozróżnianie wielkości liter lub kierunek wyszukiwania. Następnie można wybrać panel, w którym ma być prowadzone wyszukiwanie.

Aby użyć funkcji Find

1. W widoku Categories and Concepts wybierz z menu opcje **Edit > Find**. Nad panelem Categories i panelem Visualization pojawi się pasek narzędzi wyszukiwania.
2. W polu tekstowym wprowadź łańcuch wyrazów, który chcesz wyszukać. Za pomocą przycisków na pasku narzędzi można sterować uwzględnianiem wielkości liter, dopasowywaniem częściowym i kierunkiem wyszukiwania.
3. W pasku narzędzi kliknij nazwę panelu, w którym chcesz wyszukiwać. W przypadku znalezienia dopasowania tekst w oknie zostanie podświetlony.
4. Aby wyszukać następne dopasowanie, należy kliknąć nazwę panelu ponownie.

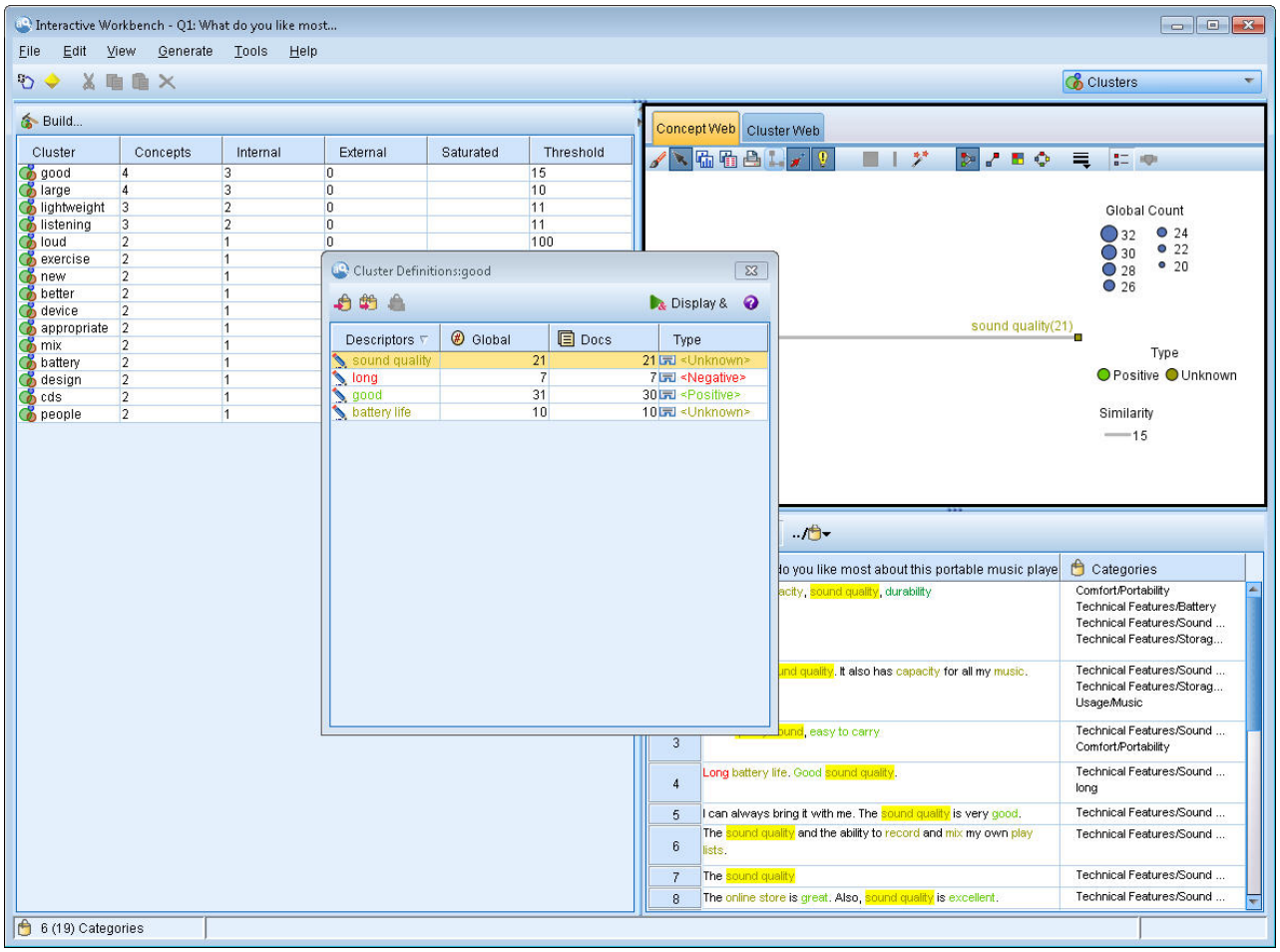
Widok Clusters

W widoku Clusters można budować i eksplorować skupienia znalezione w danych tekstowych. *Skupienia* są grupami pojęć wygenerowanymi przez algorytmy tworzenia skupień na podstawie tego, jak często pojęcia występują w tekście i jak często występują razem. Skupienia służą do grupowania pojęć współwystępujących, natomiast kategorie służą do grupowania dokumentów lub rekordów na podstawie dopasowania zawartego w nich tekstu do deskryptorów (pojęć, reguł, wzorców) dla każdej kategorii.

Skupienie tym lepiej ujawnia interesujące relacje między pojęciami, im częściej pojęcia w skupieniu występują razem i im rzadziej występują z innymi pojęciami. Dwa pojęcia współwystępują, gdy oba (lub gdy ich synonimy lub terminy) występują tym samym dokumencie lub rekordzie. Więcej informacji zawiera temat Rozdział 10, “Analiza skupień”, na stronie 135.

Możesz budować skupienia i eksplorować je na wykresach ułatwiających wykrywanie relacji między pojęciami, których znalezienie bez pomocy skupień byłoby zbyt czasochłonne. Wprawdzie nie można dodawać całych skupień do kategorii, jednak można dodawać pojęcia ze skupienia do kategorii, korzystając z okna dialogowego Cluster Definitions. Więcej informacji zawiera temat “Definicje skupień” na stronie 139.

Możliwe jest wprowadzanie zmian w ustawieniach tworzenia skupień. Zmiany ustawień wpłyną na wyniki. Więcej informacji zawiera temat “Tworzenie skupień” na stronie 136.



Rysunek 24. Widok Clusters

Widok Clusters jest podzielony na trzy panele, a każdy z nich można ukrywać lub uwidaczniać, wybierając jego nazwę z menu View. Zwykle widoczny jest tylko panel Clusters i panel Visualization.

Panel Clusters

Ten panel znajdujący się po lewej stronie, przedstawia skupienia ujawnione w danych tekstowych. Możesz wygenerować wyniki tworzenia skupień, klikając przycisk **Build**. Algorytm tworzy skupienia, próbując zidentyfikować pojęcia, które często występują razem.

W trakcie nowego wyodrębniania wyniki tworzenia skupień są kasowane i konieczne jest odbudowanie skupień w celu uzyskania aktualnych wyników. Można zmienić pewne ustawienia budowania skupień, takie jak maksymalna liczba skupień, maksymalna liczba pojęć w skupieniu lub maksymalna liczba powiązań z pojęciami zewnętrznymi. Więcej informacji zawiera temat “Eksplorowanie skupień” na stronie 138.

Panel Visualization

Ten panel znajduje się w prawym górnym rogu i udostępnia dwa ujęcia skupień: wykres sieciowy pojęć i wykres sieciowy skupień. Jeśli panel nie jest widoczny, można uzyskać do niego dostęp z menu View (**View>Visualization**). W zależności od wyboru dokonanego w panelu skupień można wyświetlić odpowiednie interakcje między skupieniami lub wewnątrz skupień. Wyniki są prezentowane w wielu formatach:

- **Concept Web.** Ten wykres przedstawia wszystkie pojęcia w wybranych skupieniach, a także pojęcia powiązane spoza skupienia.
- **Cluster Web.** Wykres sieciowy przedstawiający powiązania z wybranych skupień do innych skupień oraz powiązania między tymi innymi skupieniami.

Uwaga: Aby można było wyświetlić wykres sieciowy skupień, należy wcześniej stworzyć skupienia z powiązaniami zewnętrznymi. Powiązania zewnętrzne są to powiązania między parami pojęć w odrębnych skupieniach (jedno pojęcie w jednym skupieniu, a drugie pojęcie w innym). Więcej informacji zawiera temat “Wykresy skupień” na stronie 149.

Panel Data

Panel Data znajduje się w prawym dolnym rogu i jest domyślnie ukryty. Nie można wyświetlić żadnych wyników na panelu Data z panelu Clusters, ponieważ skupienia obejmują wiele dokumentów/rekordów, przez co dane nie niosą żadnych interesujących informacji. Można jednak wyświetlić dane odpowiadające dokonanemu wyborowi, korzystając z okna dialogowego Cluster Definitions. W panelu Data widoczny jest tekst zależny od tego, co wybrano w tym oknie dialogowym. Po dokonaniu wyboru kliknij przycisk **Display &**, aby wypełnić panel Data dokumentami lub aktami zawierającymi łącznie wszystkie pojęcia.

W odpowiednich dokumentach lub rekordach pojęcia będą wyróżnione kolorem, aby łatwiej było je odszukać w tekście. Można także zatrzymać wskaźnik myszy nad elementem oznaczonym kolorem, aby wyświetlić nazwę pojęcia, pod którym dany termin został wyodrębniony, oraz typ, do którego został przypisany. Panel Data może zawierać wiele kolumn, ale kolumna zmiennej tekstowej jest zawsze widoczna. Ma nazwę taką samą, jak zmienna tekstowa użyta podczas wyodrębniania, albo jak nazwa dokumentu, jeśli dane tekstowe znajdują się w wielu różnych plikach. Dostępne są także inne kolumny. Więcej informacji zawiera temat “Panel Data” na stronie 101.

Widok Text Link Analysis

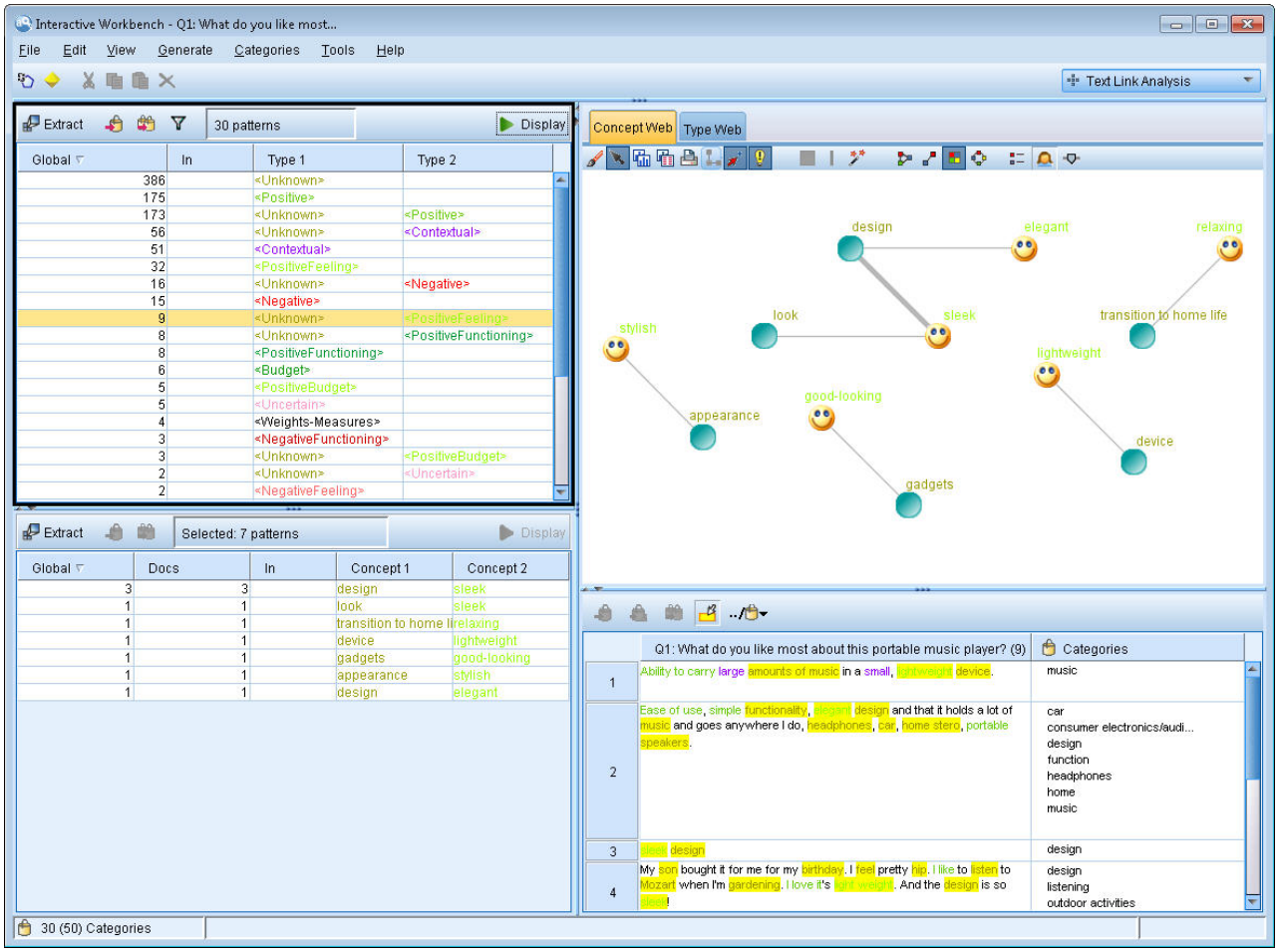
W widoku Text Link Analysis można budować i eksplorować wzorce analizy powiązań w tekście znalezione w danych tekstowych. Analiza powiązań w tekście (TLA) jest to technika dopasowywania wzorców, która umożliwia definiowanie reguł i porównywanie ich z faktycznie wyodrębnionymi pojęciami i relacjami występującymi w tekście.

Wzorce są najbardziej użyteczne wtedy, gdy próbujemy poznać relacje między pojęciami lub opinie na określony temat. Do niektórych zastosowań tej techniki należy wyodrębnianie opinii o produkcie z danych ankietowych, wyszukiwanie relacji między genami w artykułach naukowych z dziedziny medycyny lub ujawnianie relacji między osobami i miejscami zawartych w informacjach wywiadowczych.

Po wyodrębnieniu wzorców TLA można przeglądać je w panelach Data lub Visualization, a nawet dodawać je do kategorii w widoku Categories and Concepts. Aby możliwe było wyodrębnienie wyników TLA, w używanym

szablonie zasobów lub w używanych bibliotekach, muszą być zdefiniowane reguły wzorców TLA. Więcej informacji zawiera Rozdział 18, “Informacje o regułach powiązań w tekście”, na stronie 201.

W tym widoku prezentowane są wyniki wyodrębniania wzorców TLA. Jeśli nie przeprowadzono jeszcze wyodrębniania, należy użyć przycisku **Extract** i wybrać opcję włączenia wyodrębniania wzorców .



Rysunek 25. Widok Text Link Analysis

Widok Text Link Analysis jest podzielony na cztery panele, a każdy z nich można ukrywać lub uwidaczniać, wybierając jego nazwę z menu View. Więcej informacji zawiera Rozdział 11, “Eksplorowanie analizy powiązań w tekście”, na stronie 141.

Panele Type Patterns i Concept Patterns

Po lewej stronie panele Type Patterns i Concept Patterns to dwa połączone panele, w których można eksplorować i wybierać wyniki wzorca TLA. Wzorce składają się z szeregów do sześciu typów lub sześciu pojęć. Należy zwrócić uwagę, że dla tekstu japońskiego wzorce są szeregami tylko do jednego lub dwóch typów lub pojęć. Reguła wzorca TLA zdefiniowana w zasobach lingwistycznych dyktuje stopień skomplikowania wyników wzorca. Więcej informacji zawiera temat Rozdział 18, “Informacje o regułach powiązań w tekście”, na stronie 201.

Wyniki wzorca są najpierw grupowane na poziomie typu, a następnie dzielone na wzorce pojęć. Z tego powodu istnieją dwa różne panele wyników: Type Patterns (po lewej stronie na górze) i Concept Patterns (po lewej stronie na dole).

- **Type Patterns.** Panel Type Patterns przedstawia wyodrębnione wzorce składające się z dwóch lub większej liczby powiązanych typów odpowiadających regule wzorca TLA. Wzorce typu są przedstawione jako <Organization> + <Location> + <Positive>, co może zapewniać pozytywną opinię o organizacji w określonej lokalizacji.

- **Concept Patterns.** Panel Concept Patterns prezentuje wyodrębnione wzorce na poziomie pojęć dla wszystkich wzorców wybranych w tym momencie w znajdującym się wyżej panelu Type Patterns. Wzorce pojęć mają strukturę, taką jak np. hotel + paris + wonderful.

Tak jak w przypadku wyników wyodrębniania w widoku Categories and Concepts możesz tu przejrzeć wyniki. Jeśli chcesz doprecyzować typy i pojęcia, z których składają się te wzorce, można to zrobić w panelu Extraction Results widoku Categories and Concepts lub bezpośrednio w narzędziu Resource Editor, a następnie wyodrębnić ponownie wzorce.

Panel Visualization

Ten panel znajduje się w prawym górnym rogu widoku Text Link Analysis i prezentuje wykres sieciowy wybranych wzorców jako wzorce typu lub wzorce pojęć. Jeśli panel nie jest widoczny, można uzyskać do niego dostęp z menu View (**View > Visualization**). W zależności od tego, co wybrano w innych panelach, można wyświetlić powiązane interakcje między dokumentami/rekordami i wzorcami.

Wyniki są prezentowane w wielu formatach:

- **Concept Graph.** Ten wykres przedstawia wszystkie pojęcia w wybranych wzorcach. Szerokość linii i wielkości węzłów (jeśli nie są pokazane ikony typu) na wykresie pojęć przedstawia liczbę globalnych wystąpień w wybranej tabeli.
- **Type Graph.** Ten wykres przedstawia wszystkie typy w wybranych wzorcach. Szerokość linii i wielkości węzłów (jeśli nie są pokazane ikony typu) na wykresie przedstawia liczbę globalnych wystąpień w wybranej tabeli. Węzły są przedstawiane w określonym kolorze lub w formie ikony.

Więcej informacji zawiera temat “Wykresy analizy powiązań w tekście” na stronie 150.

Panel Data

Panel Data znajduje się w prawym dolnym rogu. Jest on tabelą zawierającą dokumenty lub rekordy odpowiadające elementom wybranym w innym obszarze widoku. W panelu Data widoczny jest tekst zależny od tego, co wybrano. Po dokonaniu wyboru kliknij przycisk **Display**, aby zapełnić panel Data odpowiednim tekstem.

Jeśli w innym panelu dokonano wyboru, to w odpowiednich dokumentach lub rekordach pojęcia będą wyróżnione kolorem, aby łatwiej było je odszukać w tekście. Można także zatrzymać wskaźnik myszy nad elementem oznaczonym kolorem, aby wyświetlić podpowiedź z nazwą pojęcia, pod którym dany termin został wyodrębniony, oraz typem, do którego został przypisany. Więcej informacji zawiera temat “Panel Data” na stronie 101.

Widok Resource Editor

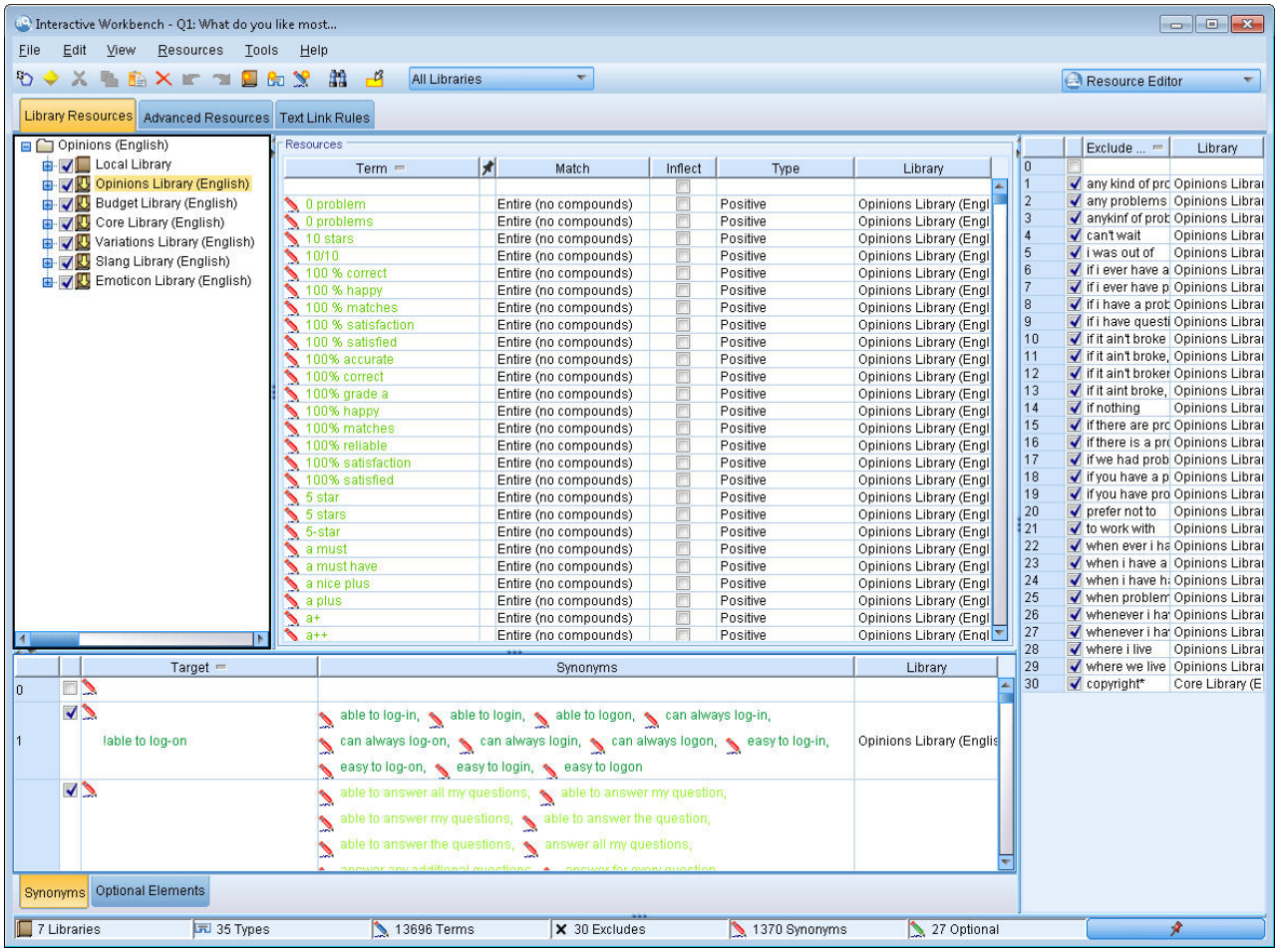
IBM SPSS Modeler Text Analytics szybko i dokładnie wyodrębnia kluczowe pojęcia z danych tekstowych, korzystając z elastycznego mechanizmu wyodrębniania. Działanie tego mechanizmu w dużej mierze zależy od zasobów lingwistycznych, które sterują analizą i interpretacją obszernych nieustrukturyzowanych danych tekstowych.

Widok Resource Editor stanowi środowisko przeglądania i optymalizacji zasobów lingwistycznych służących do wydobywania pojęć, grupowania ich w typy, wykrywania wzorców w danych tekstowych i realizacji wielu innych zadań. IBM SPSS Modeler Text Analytics oferuje kilka wstępnie skonfigurowanych szablonów źródłowych. Ponadto w przypadku niektórych języków można korzystać z zasobów w pakietach analizy tekstu. Więcej informacji zawiera temat “Korzystanie z pakietów analizy tekstu (TAP)” na stronie 129.

Ponieważ zasoby te nie zawsze są idealnie dopasowane do kontekstu danych, użytkownik może w oknie Resource Editor tworzyć i edytować własne zasoby właściwe dla konkretnego kontekstu lub domeny, a także zarządzać tymi zasobami. Więcej informacji zawiera Rozdział 15, “Praca z bibliotekami”, na stronie 167.

Aby uprościć proces optymalizacji zasobów lingwistycznych, można wykonywać typowe zadania słownikowe bezpośrednio z widoku Categories and Concepts, korzystając z menu kontekstowych w panelach Extraction Results i Data. Więcej informacji zawiera temat “Optymalizacja wyników wyodrębniania” na stronie 87.

Uwaga: Interfejs dla zasobów przystosowanych do tekstu japońskiego jest nieco inny.



Rysunek 26. Widok Resource Editor

Operacje wykonywane przez użytkowników w widoku Resource Editor związane są z zarządzaniem i optymalizacją zasobów lingwistycznych. Zasoby te są przechowywane w postaci szablonów i bibliotek. Widok Resource Editor jest podzielony na cztery części: panel drzewa bibliotek (Library Tree), panel słownika typów (Type Dictionary), panel słownika zastąpień (Substitution Dictionary) i panel słownika wykluczeń (Exclude Dictionary).

Uwaga: Więcej informacji zawiera temat “Interfejs edytora” na stronie 158.

Określanie opcji

Ogólne opcje programu IBM SPSS Modeler Text Analytics można ustawić w oknie dialogowym Options. To okno dialogowe zawiera następujące karty:

- **Session.** Ta karta zawiera ogólne opcje i separatory.
- **Display.** Ta karta zawiera opcje kolorów używanych w interfejsie.
- **Sounds.** Ta karta zawiera opcje dotyczące sygnałów dźwiękowych.

Aby edytować opcje

1. Z menu wybierz opcje **Tools > Options**. Zostanie otwarte okno dialogowe Options.
2. Wybierz kartę zawierającą informacje, które chcesz zmienić.
3. Zmień dowolne opcje.
4. Kliknij przycisk **OK**, aby zapisać zmiany.

Okno dialogowe Options: karta Session

Na tej karcie można wybrać niektóre ustawienia podstawowe.

Data Pane and Category Graph Display. Ta opcja wpływa na sposób prezentacji danych w panelu Data oraz w panelu Visualization w ramach widoku Categories and Concepts.

- **Display limit for Data pane and Category Web.** Ta opcja określa maksymalną liczbę dokumentów, jaka ma być wyświetlana w panelu Data oraz na wykresach i diagramach w widoku Categories and Concepts.
- **Show categories for documents/records at Display time.** Gdy ta opcja jest wybrana, dokumenty lub rekordy są oceniane po każdym kliknięciu przycisku Display, zatem wszelkie kategorie, do których one należą, mogą być wyświetlone w kolumnie Categories panelu Data oraz na wykresach kategorii. W niektórych przypadkach, zwłaszcza gdy zbiór danych jest obszerny, celowe może być wyłączenie tej opcji w celu istotnego przyspieszenia wyświetlania danych i wykresów.

Add to Category from Data Pane. Ta opcja wpływa na to, jakie elementy będą dodawane do kategorii podczas dodawania dokumentów i rekordów z panelu Data.

- **In Categories and Concepts view, copy.** Dodanie dokumentu lub rekordu z panelu Data w tym widoku spowoduje skopiowanie albo tylko pojęć (**Concepts only**), abo zarówno pojęć, jak i wzorców (**Concepts and Patterns**).
- **In Text Link Analysis view, copy.** Dodanie dokumentu lub rekordu z panelu Data w tym widoku spowoduje skopiowanie albo tylko wzorców (**Patterns only**), abo zarówno pojęć, jak i wzorców (**Concepts and Patterns**).

Resource Editor delimiter. Wybierz znak, jaki ma być używany w charakterze ogranicznika (separatora) przy wprowadzaniu takich elementów, jak pojęcia, synonimy i elementy opcjonalne, w widoku Resource Editor.

Okno dialogowe Options: karta Display

Na tej karcie można edytować opcje wpływające na ogólny wygląd i zachowanie aplikacji oraz kolory używane do rozróżnienia elementów.

Uwaga: Aby przełączyć produkt na wygląd klasyczny lub wygląd jednej z wcześniejszych wersji, otwórz okno dialogowe Opcje użytkownika z menu Narzędzia głównego okna programu IBM SPSS Modeler.

Custom Colors. Edytuj kolory elementów wyświetlanych na ekranie. Dla każdego z elementów w tabeli można zmienić kolor. Aby określić kolor niestandardowy, kliknij obszar koloru na prawo od elementu, który chcesz zmienić, a następnie wybierz kolor z listy rozwijanej.

- **Non-extracted text.** Dane tekstowe, które nie zostały wyodrębnione, ale są widoczne w panelu Data.
- **Highlight background.** Kolor tła zaznaczenia podczas wybierania elementów w panelach lub zaznaczania tekstu w panelu Data.
- **Extraction needed background.** Kolor tła, który w panelach Extraction Results, Patterns i Clusters sygnalizuje dokonanie zmian w bibliotekach i potrzebę przeprowadzenia wyodrębniania.
- **Category feedback background.** Kolor tła kategorii, które pojawia się po operacji.
- **Default type.** Domyślny kolor dla typów i pojęć wyświetlanych w panelu Data i panelu Extraction Results. Ten kolor zostanie zastosowany do wszystkich typów niestandardowych utworzonych w edytorze Resource Editor. Można zmienić ten kolor domyślny dla danego słownika typu użytkownika, edytując właściwości tego słownika typów w edytorze Resource Editor. Więcej informacji zawiera temat “Tworzenie typów” na stronie 179.
- **Striped table 1.** Pierwszy z dwóch kolorów używanych na zmianę w tabeli w oknie dialogowym Edit Forced concepts w celu odróżnienia zestawów wierszy.

- **Striped table 2.** Drugi z dwóch kolorów używanych na zmianę w tabeli w oknie dialogowym Edit Forced concepts w celu odróżnienia zestawów wierszy.

Uwaga: Jeśli klikniesz przycisk **Reset to Defaults** wszystkie opcje w oknie dialogowym zostaną przywrócone do stanu z chwili instalacji produktu.

Okno dialogowe Options: karta Sounds

Na tej karcie można edytować opcje wpływające na dźwięki. W obszarze Sound Events można określić dźwięki, które mają być używane do powiadamiania o zdarzeniach. Dostępnych jest szereg różnych dźwięków. Użyj przycisków z wielokropkiem (**...**), aby wyszukać i wybrać dźwięk. Pliki *.wav* z dźwiękami dla programu IBM SPSS Modeler Text Analytics są przechowywane w podkatalogu *media* w katalogu instalacyjnym. Jeśli chcesz zrezygnować z wszelkich dźwięków, wybierz opcję **Mute All Sounds**. Dźwięki są wyciszone domyślnie.

Uwaga: Jeśli klikniesz przycisk **Reset to Defaults** wszystkie opcje w oknie dialogowym zostaną przywrócone do stanu z chwili instalacji produktu.

Ustawienia pomocy w programie Microsoft Internet Explorer

Ustawienia programu Microsoft Internet Explorer

Większość funkcji pomocy tej aplikacji działa w oparciu o przeglądarkę Microsoft Internet Explorer. Niektóre wersje przeglądarki Internet Explorer (w tym wersja dostarczona z systemem Microsoft Windows XP z dodatkiem Service Pack 2) domyślnie blokują „zawartość aktywną" w oknach przeglądarki na komputerze lokalnym. Ustawienie domyślne może powodować blokowanie niektórych treści pomocy. Aby cała treść pomocy była dostępna, można zmienić domyślne ustawienie w programie Internet Explorer.

1. Z menu programu Internet Explorer wybierz:

   Narzędzia > Opcje internetowe...
2. Kliknij kartę **Zaawansowane**.
3. Przewiń do sekcji **Zabezpieczenia**.
4. Zaznacz opcję **Zezwalaj zawartości aktywnej na działanie w plikach na moim komputerze**.

Generowanie modeli użytkowych i węzłów modelowania

W sesji interaktywnej możesz wykorzystać efekty dotychczasowej pracy do wygenerowania:

- **Węzła modelowania eksploracji tekstu.** Węzeł modelowania wygenerowany z interaktywnego pulpitu roboczego to węzeł Text Mining, którego ustawienia i opcje będą odzwierciedlać ustawienia i opcje zachowane w otwartej sesji interaktywnej. Może być przydatny, gdy nie masz już pierwotnego węzła Text Mining lub gdy chcesz utworzyć nową wersję. Więcej informacji zawiera Rozdział 3, "Eksploracja w poszukiwaniu pojęć i kategorii", na stronie 19.
- **Model użytkowy kategorii.** Model użytkowy wygenerowany z interaktywnego pulpitu roboczego jest modelem użytkowym kategorii. Aby móc wygenerować model użytkowy kategorii, musisz mieć co najmniej jedną kategorię w widoku Categories and Concepts. Więcej informacji zawiera temat "Model użytkowy Text Mining: model kategorii" na stronie 39.

Aby wygenerować węzeł modelowania eksploracją tekstu

1. Z menu wybierz opcje **Generate > Generate Modeling Node**. Węzeł modelowania Text Mining zostanie dodany do obszaru roboczego ze wszystkimi ustawieniami wybranymi obecnie w sesji interaktywnej. Nazwa węzła zostanie utworzona na podstawie nazwy zmiennej tekstowej.

Aby wygenerować model użytkowy kategorii

1. Z menu wybierz opcje **Generate > Generate Model**. Model użytkowy generowany jest bezpośrednio na palecie Model z nazwą domyślną

Aktualizowanie węzłów modelowania i zapisywanie

Zaleca się, aby podczas pracy w sesji interaktywnej od czasu do czasu aktualizować węzeł modelowania w celu zapisania zmian. Węzeł modelowania należy również aktualizować po zakończeniu pracy w sesji interaktywnej, by zapisać wyniki pracy. Aktualizacja węzła modelowania powoduje zapisanie zawartości sesji interaktywnej w węźle Text Mining, z którego ta sesja została uruchomiona. Nie powoduje to zamknięcia okna wyników.

Ważne! Ta aktualizacja nie powoduje zapisania strumienia. Chcąc zapisać strumień, należy zrobić to w oknie głównym IBM SPSS Modeler po zmianie węzła modelowania.

Aby zaktualizować węzeł modelowania

1. Z menu wybierz opcje **File > Update Modeling Node**. Węzeł modelowania zostanie zaktualizowany z uwzględnieniem ustawień tworzenia i wyodrębniania oraz opcji i kategorii użytkownika.

Zamykanie i kończenie sesji

Po zakończeniu pracy w danej sesji można ją opuścić na trzy różne sposoby:

- **Save.** Ta opcja pozwala najpierw zapisać efekty pracy z powrotem do macierzystego węzła modelowania na potrzeby przyszłych sesji, a także opublikować biblioteki do ponownego wykorzystania w innych sesjach. Więcej informacji zawiera temat “Współużytkowanie bibliotek” na stronie 172. Po zapisaniu okno sesji zostanie zamknięte, a sesja zostanie usunięta z menedżera wyników w oknie IBM SPSS Modeler.
- **Exit.** Ta opcja spowoduje odrzucenie wszystkich niezapisanych zmian, zamknięcie okna sesji i usunięcie sesji z menedżera wyników w oknie IBM SPSS Modeler. Aby zwolnić pamięć, zaleca się zapisywanie wszystkich istotnych prac i zamykanie sesji.
- **Close.** Ta opcja nie spowoduje zapisania ani odrzucenia efektów pracy. Zamyka okno sesji, ale sesja będzie nadal działać. Można otworzyć okno sesji ponownie, wybierając tę sesję w menedżerze wyników w oknie IBM SPSS Modeler.

Aby zamknąć sesję pulpitu roboczego

1. Z menu wybierz opcje **File > Close**.

Ułatwienia dostępu z użyciem klawiatury

Interfejs interaktywnego pulpitu roboczego udostępnia skróty klawiaturowe, które ułatwiają dostęp do funkcji produktu. Na najbardziej podstawowym poziomie klawisz Alt w połączeniu z innymi klawiszami (np. kombinacja Alt+F rozwija menu Plik) służy do aktywacji menu poszczególnych okien, natomiast klawisz Tab do przechodzenia do kolejnych elementów sterujących w oknie dialogowym. W tej sekcji przedstawiono podstawowe skróty klawiaturowe stanowiące alternatywną formę nawigacji. W interfejsie programu IBM SPSS Modeler dostępne są także inne skróty klawiaturowe.

Tabela 13. Ogólne skróty klawiaturowe

Klawisz skrótu	Funkcja
Ctrl+1	Wyświetlenie pierwszej karty na panelu podzielonym na karty.
Ctrl+2	Wyświetlenie drugiej karty na panelu podzielonym na karty.
Ctrl+A	Zaznaczenie wszystkich elementów w aktywnym panelu.
Ctrl+C	Skopiowanie zaznaczonego tekstu do schowka.
Ctrl+E	Rozpoczęcie wyodrębniania w widokach Categories and Concepts oraz Text Link Analysis.
Ctrl+F	Wyświetlenie paska narzędzi Find w oknie Resource Editor/Template Editor, jeśli nie jest jeszcze widoczny, i uaktywnienie go.
Ctrl+I	W widoku Categories and Concepts uruchomienie okna dialogowego Category Definitions dla wybranej kategorii. W widoku Cluster uruchomienie okna dialogowego Cluster Definitions dla wybranej kategorii.
Ctrl+R	Otwarcie okna dialogowego Add Terms w oknie Resource Editor/Template Editor.

Tabela 13. Ogólne skróty klawiaturowe (kontynuacja)

Klawisz skrótu	Funkcja
Ctrl+T	Otwarcie okna dialogowego Type Properties w celu utworzenia nowego typu w oknie Resource Editor/Template Editor.
Ctrl+V	Wklejenie zawartości schowka.
Ctrl+X	Wycięcie zaznaczonych elementów z okna Resource Editor/Template Editor.
Ctrl+Y	Powtórzenie ostatniej czynności w widoku.
Ctrl+Z	Cofnięcie ostatniej czynności w widoku.
F1	Wyświetlenie pomocy, a w oknie dialogowym — wyświetlenie pomocy kontekstowej do elementu.
F2	Włączanie/wyłączanie trybu edycji komórek tabeli.
F6	Przenoszenie aktywności między głównymi panelami w aktywnym widoku.
F8	Przeniesienie aktywności do pasków dzielących panele w celu umożliwienia zmiany rozmiaru.
F10	Rozwinięcie głównego menu Plik.
strzałka w górę, strzałka w dół	Zmiana rozmiaru panelu w pionie po wybraniu paska dzielącego.
strzałka w lewo, strzałka w prawo	Zmiana rozmiaru panelu w poziomie po wybraniu paska dzielącego.
Home, End	Maksymalne pomniejszenie lub powiększenie panelu po wybraniu paska dzielącego.
Tabulator	Przechodzenie naprzód pomiędzy elementami w oknie, panelu lub oknie dialogowym.
Shift+F10	Wyświetlenie menu kontekstowego elementu.
Shift+Tab	Przechodzenie wstecz pomiędzy elementami w oknie lub oknie dialogowym.
Shift+strzałka	Zaznaczanie znaków w polu edycji w trybie edycji (F2).
Ctrl+Tab	Przeniesienie aktywności do następnego głównego obszaru w oknie.
Shift+Ctrl+Tab	Przeniesienie aktywności do poprzedniego głównego obszaru w oknie.

Skróty w oknach dialogowych

Dostępnych jest kilka skrótów klawiaturowych i klawiszy lektora ekranowego przydatnych podczas korzystania z okien dialogowych. Po wejściu do okna dialogowego konieczne może być naciśnięcie klawisza Tab w celu ustawienia aktywności na pierwszym elemencie sterującym i zainicjowaniu lektora ekranowego. Pełna lista specjalnych skrótów klawiaturowych i skrótów do obsługi lektora ekranowego znajduje się w poniższej tabeli.

Tabela 14. Skróty w oknach dialogowych

Klawisz skrótu	Funkcja
Tabulator	Przechodzenie naprzód pomiędzy elementami w oknie lub oknie dialogowym.
Ctrl+Tab	Przejście naprzód z pola tekstowego do następnego elementu.
Shift+Tab	Przechodzenie wstecz pomiędzy elementami w oknie lub oknie dialogowym.
Shift+Ctrl+Tab	Przejście wstecz z pola tekstowego do poprzedniego elementu.
spacja	Wybór aktywnego elementu sterującego lub przycisku.
Esc	Anulowanie zmian i zamknięcie okna dialogowego.
Wprowadzanie	Zatwierdzenie zmian i zamknięcie okna dialogowego (równoważne naciśnięciu przycisku OK). Jeśli aktywne jest pole tekstowe, należy najpierw nacisnąć kombinację klawiszy Ctrl+Tab, aby z niego wyjść.

Rozdział 8. Wyodrębnianie pojęć i typów

Po każdym uruchomieniu strumienia, który otwiera interaktywny pulpit roboczy, przeprowadzane jest automatycznie wyodrębnianie z danych tekstowych w strumieniu. Wynikiem tego wyodrębniania jest zbiór pojęć, typów oraz — jeśli zasoby lingwistyczne obejmują wzorce TLA — także wynikowe wzorce. Można wyświetlać pojęcia i typy oraz pracować z nimi w panelu Extraction Results. Więcej informacji zawiera temat “Jak działa wyodrębnianie” na stronie 5.

Jeśli chcesz zoptymalizować wyniki wyodrębniania, możesz zmodyfikować zasoby lingwistyczne i ponowić wyodrębnianie. Więcej informacji zawiera temat “Optymalizacja wyników wyodrębniania” na stronie 87. Proces wyodrębniania przebiega w oparciu o zasoby oraz parametry określone w oknie dialogowym Extract. To one dyktują sposób wyodrębniania i organizacji wyników. Wyniki wyodrębniania można wykorzystać do zdefiniowania większości, a nawet wszystkich definicji kategorii.

Wyniki wyodrębniania: pojęcia i typy

W trakcie wyodrębniania całe dane tekstowe są przeglądane, a istotne pojęcia są rozpoznawane, wyodrębniane i przypisywane do typów. Po zakończeniu wyodrębniania wyniki tego procesu pojawiają się na panelu Extraction Results, który znajduje się w lewym dolnym rogu widoku Categories and Concepts. Po pierwszym uruchomieniu sesji szablon zasobu lingwistycznego wybranego w węźle używany jest do wyodrębnienia i zorganizowania tych pojęć i typów.

Uwaga: Jeśli w widocznym panelu mieści się więcej wyników, można użyć przycisków znajdujących się w dolnej części panelu, aby przeglądać wyniki, lub wprowadzić numer strony i przejść do niej.

Wyodrębnione pojęcia, typy i wzorce TLA są łącznie nazywane **wynikami wyodrębniania** i pełnią rolę deskryptorów lub elementów składowych kategorii. Pojęć, typów i wzorców można też używać w regułach kategorii. Ponadto techniki automatyczne korzystają z pojęć i typów przy konstruowaniu kategorii.

Eksploracja tekstu jest procesem iteracyjnym, w którym wyniki wyodrębniania są przeglądane z uwzględnieniem kontekstu danych tekstowych, optymalizowane w celu wygenerowania nowych wyników i ponownie oceniane. Po przeprowadzeniu wyodrębniania należy przejrzeć wyniki i wprowadzić niezbędne zmiany poprzez zmodyfikowanie zasobów lingwistycznych. Zasoby można po części optymalizować bezpośrednio w panelu Extraction Results, panelu Data, oknie dialogowym Category Definitions oraz oknie dialogowym Cluster Definitions. Więcej informacji zawiera temat “Optymalizacja wyników wyodrębniania” na stronie 87. Można także robić to bezpośrednio w widoku Resource Editor. Więcej informacji zawiera temat “Widok Resource Editor” na stronie 73.

Po optymalizacji można powtórzyć wyodrębnianie, aby zobaczyć nowe wyniki. Optymalizacja wyników wyodrębniania od samego początku gwarantuje, że przy każdym ponownym wyodrębnieniu uzyska się identyczne wynikowe definicje kategorii — idealnie dopasowane do kontekstu danych. W ten sposób dokumenty/rekordy zostaną przypisane do definicji kategorii w sposób bardziej dokładny i powtarzalny.

Pojęcia

W procesie wyodrębniania dane tekstowe są przeglądane i analizowane w celu wykrycia w tekście interesujących lub istotnych pojedynczych słów (takich jak wybory lub pokój) i fraz wielowyrazowych (takich jak wybory prezydenckie, wybory prezydenta lub traktaty pokojowe). Te słowa i frazy są zbiorczo określane jako *terminy*. Na podstawie zasobów lingwistycznych istotne terminy są wyodrębniane, a następnie podobne terminy są grupowane pod jednym terminem wiodącym, który nazywamy **pojęciem**.

Zestaw terminów bazowych danego pojęcia można wyświetlić, zatrzymując wskaźnik myszy nad nazwą pojęcia. Spowoduje to wyświetlenie podpowiedzi z nazwą pojęcia i maksymalnie kilkoma wierszami terminów zgrupowanych pod tym pojęciem. Do tych bazowych terminów należą synonimy zdefiniowane w zasobach lingwistycznych

(niezależnie od tego, czy wystąpiły w tekście, czy nie), a także wszelkie wyodrębnione terminy w liczbie mnogiej/pojedynczej, permutacje terminów, terminy wygenerowane w wyniku grupowania rozmytego itd. Można skopiować te terminy bazowe, by wyświetlić ich kompletną listę. Należy w tym celu kliknąć nazwę pojęcia prawym przyciskiem myszy i wybrać opcję z menu kontekstowego.

Domyślnie pojęcia są zapisane małymi literami i posortowane malejąco według liczby wystąpień w dokumentach (kolumna Doc.) . Każdemu wyodrębnionemu pojęciu przypisywany jest typ. Typy pomagają w grupowaniu podobnych pojęć. Pojęcia są oznaczone kolorami zgodnie ze swoimi typami. Kolory definiuje się we właściwościach typu w oknie Resource Editor. Więcej informacji zawiera temat “Słowniki typów” na stronie 177.

Gdy jakieś pojęcie, typ lub wzorzec używany jest w definicji kategorii, w dającej się sortować kolumnie **In** widoczna jest ikona .

Typy

Typy grupują pojęcia pod względem semantycznym. Każdemu wyodrębnionemu pojęciu przypisywany jest typ. Typy pomagają w grupowaniu podobnych pojęć. W produkcie IBM SPSS Modeler Text Analytics dostępnych jest od razu kilka gotowych typów, takich jak <Location> (miejsce), <Organization> (organizacja), <Person> (osoba), <Positive> (pozytywne), <Negative> (negatywne) itd. Na przykład typ <Location> grupuje słowa kluczowe związane z położeniem geograficznym i określające miejsca. Ten typ byłby przypisywany do takich pojęć, jak warszawa, londyn i tokio. W większości języków pojęcia nieznalezione w żadnym słowniku typu, ale wyodrębnione z tekstu, otrzymują automatycznie przypisywany typ <Unknown>Więcej informacji zawiera temat “Typy wbudowane” na stronie 178.

W widoku Type wyodrębnione typy są domyślnie wyświetlane w kolejności malejącej według częstotliwości globalnej. Typy są oznaczone różnymi kolorami, aby było je łatwiej rozróżniać. Kolory są zapisane we właściwościach typu. Więcej informacji zawiera temat “Tworzenie typów” na stronie 179. Można także tworzyć własne typy.

Wzorce

Z danych tekstowych mogą być także wyodrębniane wzorce. Jednak jest do tego potrzebna (w oknie Resource Editor) biblioteka zawierająca reguły analizy powiązań w tekście (TLA — Text Link Analysis (TLA). Należy także wybrać opcję wyodrębniania wzorców w ustawieniach węzła IBM SPSS Modeler Text Analytics lub w oknie dialogowym Extract za pomocą opcji **Enable Text Link Analysis pattern extraction**. Więcej informacji zawiera Rozdział 11, “Eksplorowanie analizy powiązań w tekście”, na stronie 141.

Wyodrębnianie danych

Kiedy wymagane jest wyodrębnianie, panel Extraction Results staje się żółty i poniżej paska narzędzi w tym panelu pojawia się komunikat **Press Extract Button to Extract Concepts**.

wyodrębnianie może być konieczne, jeśli nie ma jeszcze wyników wyodrębniania, wprowadzono zmiany do zasobów lingwistycznych i wymagana jest aktualizacja wyników wyodrębniania lub otwarto ponownie sesję, gdzie nie zapisano wyników wyodrębniania (**Tools > Options**).

Uwaga: Jeśli zmienisz węzeł źródłowy strumienia już po zbuforowaniu wyników wyodrębniania za pomocą opcji **Use session work...**, trzeba będzie uruchomić nowe wyodrębnianie po uruchomieniu sesji interaktywnego pulpitu roboczego, aby móc korzystać ze zaktualizowanych wyników wyodrębniania.

Po uruchomieniu wyodrębniania pojawia się wskaźnik postępu informujący o statusie wyodrębniania. W tym czasie mechanizm wyodrębniania odczytuje wszystkie dane tekstowe i identyfikuje powiązane terminy i wzorce oraz wyodrębnia je i przypisuje do typu. Następnie mechanizm próbuje grupować terminy-synonimy pod jednym terminem nazywanym pojęciem. Kiedy proces się zakończy, wynikowe pojęcia, typy i wzorce pojawiają się w panelu Extraction Results.

Proces wyodrębniania powoduje powstanie zestawu pojęć i typów, jak również wzorców TLA (Text Link Analysis), jeśli są włączone. Można przeglądać i pracować z tymi pojęciami i typami w panelu Extraction Results widoku Categories and Concepts. Jeśli wyodrębniono wzorce TLA, widać je w widoku Text Link Analysis.

Uwaga: Istnieje relacja pomiędzy wielkością zbioru danych i czasem, jaki jest wymagany do zakończenia procesu wyodrębniania. Zawsze można rozważyć wstawienie węzła Sample we wcześniejszej części strumienia lub optymalizację konfiguracji komputera.

Aby wyodrębnić dane

1. W menu wybierz kolejno następujące opcje **Tools > Extract**. Lub kliknij przycisk **Extract** na pasku.
2. Jeśli ustawiono wyświetlanie okna dialogowego Extraction Settings, można w nim wprowadzić zmiany. Dalej w tym temacie opisano wszystkie ustawienia.
3. Kliknij przycisk **Extract**, aby rozpocząć proces wyodrębniania. Po rozpoczęciu wyodrębniania otwiera się okno dialogowe postępu. Po wyodrębnianiu wyniki pojawiają się w panelu Extraction Results. Domyślnie pojęcia są zapisane małymi literami i posortowane malejąco według liczby wystąpień w dokumentach (kolumna Doc.) .

Można przejrzeć wyniki, używając opcji paska narzędzi, aby posortować wyniki inaczej, odfiltrować wyniki lub przełączyć na inny widok (pojęcia lub typy). Można również doprecyzować wyniki wyodrębniania, pracując z zasobami lingwistycznymi. Więcej informacji zawiera temat “Optymalizacja wyników wyodrębniania” na stronie 87.

Potencjalne problemy związane z wyodrębnieniem

Wiele interaktywnych pulpitów roboczych może spowalniać działanie programu. Podczas pracy z interaktywnym pulpitem roboczym SPSS Modeler Text Analytics i SPSS Modeler współużytkują wspólny mechanizm środowiska wykonawczego Java. W zależności od liczby otwartych interaktywnych pulpitów roboczych podczas sesji programu SPSS Modeler, nawet w przypadku uruchamiania i zamykania tej samej sesji, pamięć systemowa może spowalniać działanie aplikacji. Efekt jest szczególnie wyraźny gdy użytkownik pracuje z dużą ilością danych lub gdy komputer dysponuje mniejszą ilością pamięci RAM niż zalecane 4 GB. Jeśli komputer wolno reaguje, zaleca się zapisać wszystkie zmiany, zamknąć program SPSS Modeler i ponownie uruchomić aplikację. uruchamianie produktu SPSS Modeler Text Analytics na komputerze z mniejszą ilością pamięci niż zalecana, zwłaszcza podczas pracy z dużymi zestawami danych lub przez dłuższy czas, może spowodować wyczerpanie zasobów pamięciowych środowiska wykonawczego Java i wyłączenie systemu. W przypadku pracy z dużą ilością danych zdecydowanie zaleca się rozbudowę pamięci do zalecanego lub wyższego poziomu (albo wykorzystanie produktuSPSS Modeler Text Analytics Server).

Informacje dotyczące języka angielskiego, francuskiego, hiszpańskiego, holenderskiego, niemieckiego, portugalskiego i włoskiego

Okno dialogowe Extraction Settings zawiera podstawowe opcje wyodrębniania.

Enable Text Link Analysis pattern extraction. Określa, czy chcesz wyodrębnić wzorce TLA z danych tekstowych. To ustawienie zakłada również, że istnieją reguły wzorców TLA w jednej z bibliotek narzędzia Resource Editor. Ta opcja może znacznie wydłużyć czas wyodrębniania. Więcej informacji zawiera Rozdział 11, “Eksplorowanie analizy powiązań w tekście”, na stronie 141.

Accommodate punctuation errors. Ta opcja powoduje, że podczas wyodrębniania tekst zawierający błędy interpunkcyjne (na przykład nieprawidłowo użyte znaki interpunkcyjne) będzie tymczasowo normalizowany w celu poprawienia efektywności wyodrębniania pojęć. Ta opcja jest bardzo użyteczna, gdy mamy do czynienia z krótkimi tekstami o niskiej jakości (np. odpowiedziami na pytania otwarte w ankietach, wiadomościami e-mail i danymi z systemów CRM) lub gdy system zawiera wiele skrótów.

Accommodate spelling for a minimum word character length of [n] Ta opcja powoduje zastosowanie techniki grupowania rozmytego, która grupuje błędnie napisane lub podobne wyrazy pod jednym pojęciem. Algorytm grupowania rozmytego tymczasowo usuwa wszystkie samogłoski (z wyjątkiem pierwszej) oraz podwójne/potrójne spółgłoski z wyodrębnianych wyrazów, a następnie porównuje wyniki, by sprawdzić, czy są identyczne. Zatem wyrazy

modeling i modelling zostałyby połączone w jedną grupę. Jeśli jednak każdy termin ma przypisany inny typ, z wyjątkiem typu <Unknown>, to grupowanie rozmyte nie będzie stosowane.

Można też określić minimalną liczbę znaków *rdzennych* wymaganą do zastosowania grupowania rozmytego. Liczba znaków rdzennych w terminie obliczana jest poprzez zsumowanie wszystkich znaków i odjęcie znaków tworzących przyrostki przy odmianie, a w wypadku terminów będących wyrazami złożonymi, także znaków tworzących określniki i przyimki. Na przykład termin exercises ma 8 znaków rdzennych w swojej postaci "exercise", ponieważ litera *s* na końcu tworzy odmianę (w tym przypadku liczbę mnogą). Podobnie, apple sauce ma 10 znaków rdzennych ("apple sauce"), a manufacturing of cars ma 16 znaków rdzennych ("manufacturing car"). Ta metoda liczenia znaków jest stosowana tylko do sprawdzania, czy ma być przeprowadzane grupowanie rozmyte, ale nie wpływa na sposób dopasowywania wyrazów.

Uwaga: Jeśli później okaże się, że pewne wyrazy są grupowane nieprawidłowo, można wykluczyć konkretne pary wyrazów ze stosowania tej techniki, jawnie deklarując je w sekcji **Fuzzy Grouping: Exceptions** na karcie Advanced Resources. Więcej informacji zawiera temat "Grupowanie rozmyte" na stronie 191.

Extract uniterms Ta opcja wyodrębnia pojedyncze wyrazy (terminy pojedyncze), o ile tylko nie są już częścią wyrazu złożonego i są albo rzeczownikami, albo nierozpoznanymi częściami mowy.

Extract nonlinguistic entities Ta opcja wyodrębnia Obiekty nielingwistyczne, takie jak numery telefonów, numery ubezpieczenia społecznego, godziny, daty, waluty, cyfry, wartości procentowe, adresy e-mail i adresy HTTP. W sekcji **Nonlinguistic Entities: Configuration** karty Advanced Resources można uwzględniać i wykluczać określone typy obiektów nielingwistycznych. Wykluczenie zbędnych obiektów sprawi, że mechanizm wyodrębniania nie będzie marnował czasu na ich przetwarzanie. Więcej informacji zawiera temat "Konfiguracja" na stronie 195.

Uppercase algorithm Ta opcja wyodrębnia proste i złożone terminy, które nie figurują we wbudowanych słownikach, o ile pierwsza litera terminu jest wielka. Jest to dobry sposób na wyodrębnienie większości rzeczowników własnych.

Group partial and full person names together when possible Ta opcja grupuje imiona i nazwiska, które w tekście występują w różnych postaciach. Jest to użyteczne, ponieważ imiona i nazwiska często na początku tekstu przytaczane są w pełnym brzmieniu, ale później już występują tylko w wersji skróconej. W przypadku wybrania tej opcji program próbuje dopasować każdy pojedynczy termin typu <Unknown> do ostatniego wyrazu każdego terminu złożonego typu <Person> (osoba). Na przykład, jeśli znaleziony zostanie wyraz *nowak* o początkowo przypisanym typie <Unknown>, to mechanizm wyodrębniania sprawdzi, czy jakiekolwiek terminy złożone typu <Person> zawierają jako ostatni wyraz właśnie *nowak*, na przykład *piotr nowak*. Ta opcja nie ma zastosowania do imion, ponieważ większość z nich nigdy nie jest wyodrębniana jako termin pojedynczy.

Maximum nonfunction word permutation Ta opcja określa maksymalną liczbę wyrazów niefunkcyjnych, które mogą być obecne, gdy stosowana jest technika permutacji. Technika permutacji grupuje podobne frazy różniące się tylko wyrazami niefunkcyjnymi (na przykład „of" lub „the"), niezależnie od odmiany. Załóżmy na przykład, że wartość ta jest ustawiona na maksymalnie dwa wyrazy i wyodrębniono zarówno termin company officials, jak i termin officials of the company. W tym przypadku oba terminy zostaną połączone w grupę, ponieważ po zignorowaniu wyrazów of the zostaną uznane za identyczne.

Use derivation when grouping multiterms Podczas przetwarzania wielkich zbiorów danych wybierz tę opcję, aby grupować terminy wielowyrazowe według reguł derywacji.

Index Option for Concept Map Określa, że w momencie wyodrębniania ma zostać zbudowany indeks mapy, aby można było szybko narysować później mapy pojęć. Aby edytować ustawienia indeksu, kliknij przycisk **Settings**. Więcej informacji zawiera temat "Budowanie indeksów map pojęć" na stronie 86.

Always show this dialog before starting an extraction Określ, czy chcesz wyświetlać okno dialogowe Extraction Settings za każdym razem, gdy wykonywane jest wyodrębnianie, czy nigdy nie chcesz wyświetlać tego okna, tylko po przejściu do menu Tools lub czy chcesz widzieć pytanie, czy chcesz edytować ustawienia wyodrębniania za każdym razem, gdy wykonywane jest wyodrębnianie.

Dla tekstu japońskiego

Okno dialogowe Extraction Settings zawiera podstawowe opcje wyodrębniania dla języka japońskiego. Domyślnie ustawienia wybrane w oknie dialogowym są takie same, jak wybrane na karcie Expert węzła modelowania Text Mining. W celu pracy z tekstem japońskim należy użyć tekstu jako danych wejściowych, jak również wybrać szablon języka japońskiego lub pakiet analizy tekstu na karcie Model węzła Text Mining. Więcej informacji zawiera temat “Kopiowanie zasobów z szablonów i pakietów TAP” na stronie 26.

Secondary Analysis. Po rozpoczęciu procesu wyodrębniania następuje wyodrębnienie podstawowych słów kluczowych na podstawie domyślnego zestawu typów. Gdy jednak wybierzesz dodatkowy analizator, możesz uzyskać większą liczbę pojęć lub pogłębione pojęcia, ponieważ mechanizm wyodrębniający będzie uwzględniał partykuły i czasowniki pomocnicze jako części pojęcia. W wynikach analizy sentymentu zostanie uwzględnionych wiele dodatkowych typów. Co więcej, wybranie dodatkowego analizatora umożliwi także wygenerowanie wyników analizy powiązań w tekście.

Uwaga: Użycie dodatkowego analizatora wydłuża proces wyodrębniania.

- **Dependency analysis**. Wybranie tej opcji powoduje uwzględnienie dodatkowych partykuł przy wyodrębnianiu podstawowych typów i słów kluczowych. Pozwala także uzyskać wzbogacone wzorce przy analizie powiązań w tekście (TLA).
- **Sentiment analysis**. Wybranie tego analizatora pozwala wyodrębnić dodatkowe pojęcia oraz, tam gdzie to możliwe, wyodrębnić wzorce powiązań TLA. Oprócz typów podstawowych można też korzystać z ponad 80 typów sentymentu. Te typy służą do ujawniania pojęć i wzorców w treściach wyrażających emocje, odczucia i opinie. Dostępne są trzy opcje, które służą do odpowiedniego ukierunkowania analizy sentymentu: **All sentiments**, **Representative sentiment only** i **Conclusions only**.
- **No secondary analyzer** Ta opcja wyłącza wszystkie dodatkowe analizatory. Nie można wybrać tej opcji, jeśli zaznaczono opcję **Enable Text Link Analysis pattern extraction**, ponieważ do uzyskania wyników TLA niezbędny jest dodatkowy analizator.

Enable Text Link Analysis pattern extraction Określa, czy chcesz wyodrębnić wzorce TLA z danych tekstowych. To ustawienie zakłada również, że istnieją reguły wzorców TLA w jednej z bibliotek narzędzia Resource Editor. Ta opcja może znacznie wydłużyć czas wyodrębniania. Dodatkowo należy wybrać dodatkowy analizator, aby wyodrębnić wyniki wzorca TLA. Więcej informacji zawiera Rozdział 11, “Eksplorowanie analizy powiązań w tekście”, na stronie 141.

Filtrowanie wyników wyodrębniania

Podczas pracy na bardzo obszernych zbiorach danych proces wyodrębniania może wygenerować miliony wyników. Wielu użytkownikom tak obszerne wyniki utrudnią lub uniemożliwią analizę. Dlatego, aby skupić się na najbardziej interesujących wynikach, można je odfiltrować za pomocą okna dialogowego Filter dostępnego w panelu Extraction Results.

Należy pamiętać, że wszystkie ustawienia wybrane w oknie dialogowym Filter są używane łącznie do filtrowania wyników wyodrębniania dostępnych na potrzeby tworzenia kategorii.

Filter by Frequency Można odfiltrować wyniki tak, by były wyświetlane tylko wyniki o określonej liczebności globalnej lub liczebności dokumentów.

- **Global frequency** to łączna liczba wystąpień pojęcia w całym zbiorze dokumentów i rekordów. Podana jest w kolumnie **Global**.
- **Document frequency** to łączna liczba dokumentów lub rekordów, w których występuje pojęcie. Podana jest w kolumnie **Docs**.

Na przykład, jeśli pojęcie nato występuje 800 razy w 500 rekordach, to globalna liczebność tego pojęcia wynosiłaby 800, a liczebność dokumentów wynosiłaby 500.

And by Type Można odfiltrować wyniki tak, by wyświetlane były tylko wyniki należące do określonego typu. Można wybrać wszystkie typy lub tylko określone typy.

And by Match Text Można odfiltrować wyniki tak, by wyświetlane były tylko wyniki zgodne ze zdefiniowaną tutaj regułą. Wprowadź zbiór znaków, które mają być wyszukiwane, w polu **Match text** , a następnie wybierz warunek dopasowania.

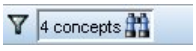
Tabela 15. Warunki dopasowywania tekstu

Warunek	Opis
Contains	Tekst zostanie dopasowany, jeśli łańcuch wystąpi w dowolnym miejscu. (Opcja domyślna)
Starts with	Tekst zostanie dopasowany, jeśli pojęcie lub nazwa typu zaczyna się od określonego tekstu.
Ends with	Tekst zostanie dopasowany, jeśli pojęcie lub nazwa typu kończy się określonym tekstem.
Exact match	Cały tekst musi być dopasowany do pojęcia lub nazwy typu.

Wyniki wyświetlane w panelu Extraction Result

Oto przykłady wyników (w języku angielskim) wyświetlanych na pasku narzędzi panelu Extraction Results w zależności od wybranych filtrów.

Tabela 16. Przykłady wyników filtrowania

Wyniki filtrowania	Opis
	Pasek narzędzi zawiera liczbę wyników. Ponieważ nie zastosowano filtru dopasowującego tekst i nie osiągnięto maksimum, nie są wyświetlane dodatkowe ikony.
	Na pasku narzędzi widać, że liczba wyników została ograniczona do maksimum określonego dla filtru, które w tym przypadku wynosi 300. Fioletowa ikona oznacza, że osiągnięto maksymalną liczbę pojęć. Zatrzymaj wskaźnik nad ikoną, aby wyświetlić więcej informacji.
	Na pasku narzędzi widać, że liczba wyników została ograniczona przez filtr dopasowujący tekst. Sygnalizuje to ikona lupy.

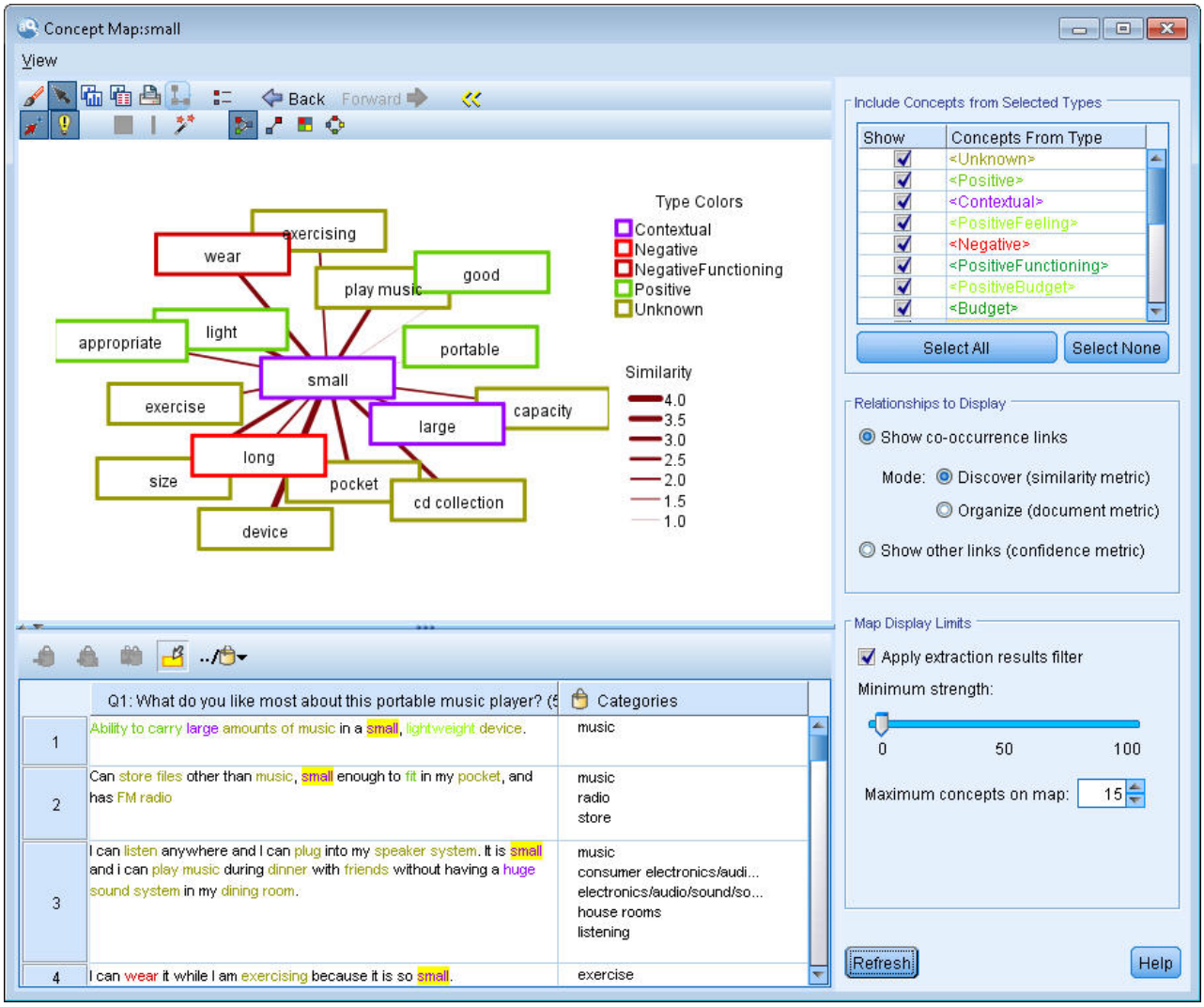
Aby filtrować wyniki

1. Z menu wybierz opcje **Tools > Filter**. Zostanie otwarte okno dialogowe Filter.
2. Wybierz i doprecyzuj filtry, których chcesz używać.
3. Kliknij przycisk **OK**, aby zastosować filtry i wyświetlić nowe wyniki w panelu Extraction Results.

Eksplorowanie map pojęć

Można utworzyć mapę pojęć, aby poznać sposób, w jaki pojęcia są ze sobą połączone relacjami. Wybierając jedno pojęcie i klikając opcję **Map**, można otworzyć okno mapy pojęć, w którym można eksplorować zestaw pojęć pokrewnych wybranemu pojęciu. Można odfiltrować pojęcia, które są wyświetlane, edytując ustawienia, takie jak typy do uwzględnienia, rodzaje relacji do wyszukania itd.

Ważne: Zanim będzie można utworzyć mapę, należy wygenerować indeks. Ten proces może potrwać kilka minut. Jednak po wygenerowaniu indeksu nie trzeba generować go ponownie do czasu ponownego wyodrębnienia. Jeśli indeks ma być generowany automatycznie po każdym wyodrębnianiu, wybierz odpowiednią opcję w ustawieniach wyodrębniania. Więcej informacji zawiera temat “Wyodrębnianie danych” na stronie 80.



Rysunek 27. Mapa pojęć dla wybranego pojęcia

Aby wyświetlić mapę pojęć

1. W panelu Extraction Results wybierz jedno pojęcie.
2. Na pasku narzędzi tego panelu kliknij przycisk **Map**. Jeśli indeks mapy został już wygenerowany, mapa pojęć otworzy się w osobnym oknie dialogowym. Jeśli indeks mapy nie został wygenerowany lub jest nieaktualny, musi zostać odbudowany. Ten proces może potrwać kilka minut.
3. Kliknij po mapie, aby ją eksplorować. Jeśli dwukrotnie klikniesz powiązane pojęcie, mapa zostanie odświeżona i pojawią się pojęcia zewnętrzne z tym, które zostało kliknięte.
4. Pasek narzędzi udostępnia podstawowe narzędzia do pracy z mapą, takie jak powrót do poprzedniej mapy, filtrowanie powiązań według siły relacji, a także otwieranie okna dialogowego filtrowania, w którym można określić typy pojęć, które będą wyświetlane, a także rodzaj relacji, które mają być przedstawiane. Drugi wiersz paska narzędzi zawiera narzędzia edycji wykresu. Więcej informacji zawiera temat “Używanie pasków narzędzi i palet wykresów” na stronie 151.
5. Jeśli znajdowane powiązania są niezadowalające, przejrzyj ustawienia tej mapy po jej prawej stronie.

Ustawienia mapy: Include Concepts from Selected Types

Tylko pojęcia, które należą do typów wybranych w tabeli, są pokazywane na mapie. Aby ukryć pojęcia określonego typu, anuluj wybór tego typu w tabeli.

Ustawienia mapy: Relationships to Display

Show co-occurrence links. Aby wyświetlić powiązania współwystąpień, wybierz tryb. Tryb wpływa na sposób obliczania siły powiązań.

- *Discover (similarity metric)*. W przypadku tej metryki siła powiązań jest obliczana przy użyciu bardziej złożonego algorytmu, który uwzględnia, jak często dwa pojęcia występują osobno, a także, jak często występują razem. Duża siła oznacza, że para pojęć występuje częściej razem niż pojęcia te występują osobno. W poniższym wzorze wszystkie wartości zmiennopozycyjne są przekształcane w liczby całkowite.

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

Rysunek 28. Wzór na współczynnik podobieństwa

W tym wzorze C_I to liczba dokumentów lub rekordów, w których występuje pojęcie I.

C_J to liczba dokumentów lub rekordów, w których występuje pojęcie J.

C_{IJ} to liczba dokumentów lub rekordów, w których para pojęć I i J współwystępuje w zestawie dokumentów.

- *Organize (document metric)*. Siła połączeń dla tej metryki jest obliczana na podstawie surowej liczby współwystąpień. Ogólnie, im większą liczebność mają dwa pojęcia, tym bardziej prawdopodobne jest, że czasami będą występować razem. Duża siła oznacza, że pojęcia często występują razem.

Show other links (confidence metric). Można nakazać wyświetlanie innych powiązań: semantycznych, wywodzenia (morfologicznych) lub włączających (składniowych). Ich siła zależy od liczby kroków, która dzieli jedno pojęcie od drugiego. Możliwości te pomagają w optymalizacji zasobów, a zwłaszcza synonimów, oraz usuwaniu niejednoznaczności. Krótkie opisy poszczególnych technik grupowania zawiera temat “Zaawansowane ustawienia językowe” na stronie 105

Uwaga: Należy pamiętać, że jeśli ich nie wybrano podczas budowania indeksu lub jeśli nie znaleziono relacji, to nie zostanie wyświetlona żadna informacja. Więcej informacji zawiera temat “Budowanie indeksów map pojęć”.

Ustawienia mapy: Map Display Limits

Apply extraction results filter. Jeśli nie chcesz używać wszystkich pojęć, można użyć filtru wyników wyodrębniania. Następnie należy wybrać tę opcję, a IBM SPSS Modeler Text Analytics wyszuka pojęcia pokrewne w przefiltrowanym zestawie. Więcej informacji zawiera temat “Filtrowanie wyników wyodrębniania” na stronie 83.

Minimum strength. Tutaj ustaw minimalną siłę powiązania. Wszystkie pojęcia pokrewne o sile relacji niższej niż ta granica zostaną ukryte na mapie.

Maximum concepts on map. Określ maksymalną liczbę relacji do wyświetlenia na mapie.

Budowanie indeksów map pojęć

Zanim będzie można utworzyć mapę, należy wygenerować indeks relacji między pojęciami. Za każdym razem, gdy użytkownik tworzy mapę pojęć, IBM SPSS Modeler Text Analytics odwołuje się do tego indeksu. Można wybrać, które relacje będą indeksowane, określając techniki w tym oknie dialogowym.

Grouping techniques. Wybierz jedną lub więcej technik. Krótki opis każdej z tych technik zawiera sekcja “Informacje o technikach lingwistycznych” na stronie 107. Nie wszystkie techniki są dostępne dla wszystkich języków tekstu.

Prevent pairing of specific concepts. Zaznacz to pole wyboru, aby w wynikach nie grupować lub nie łączyć w pary dwóch pojęć. Aby tworzyć pary pojęć lub nimi zarządzać, kliknij przycisk **Manage Pairs**. Więcej informacji zawiera temat “Zarządzanie parami wyjątków powiązań” na stronie 106.

Budowanie indeksu może potrwać kilka minut. Jednak po wygenerowaniu indeksu nie trzeba generować go ponownie do czasu ponownego wyodrębnienia, chyba że chcesz zmienić ustawienia, by uwzględnić więcej relacji. Jeśli indeks ma być generowany automatycznie po każdym wyodrębnianiu, wybierz odpowiednią opcję w ustawieniach wyodrębniania. Więcej informacji zawiera temat “Wyodrębnianie danych” na stronie 80.

Optymalizacja wyników wyodrębniania

Wyodrębnianie jest procesem iteracyjnym, w którym możliwe jest wyodrębnianie pojęć, przeglądanie wyników, wprowadzanie zmian i ponowne wyodrębnianie w celu zaktualizowania wyników. Ponieważ warunkiem powodzenia eksploracji tekstu i kategoryzacji jest dokładność i ciągłość, optymalizacja wyników już na samym początku procesu zagwarantuje, że każde ponowne wyodrębnianie przyniesie dokładnie te same wyniki w definicjach kategorii. W ten sposób rekordy i dokumenty zostaną przypisane do kategorii w sposób bardziej dokładny i powtarzalny.

Wyniki wyodrębniania są elementami składowymi kategorii. Podczas tworzenia kategorii na podstawie tych wyników wyodrębniania rekordy i dokumenty są automatycznie przypisywane do kategorii, jeśli zawierają tekst pasujący do jednego lub wielu deskryptorów tej kategorii. Mimo że można rozpocząć kategoryzację przed zoptymalizowaniem zasobów lingwistycznych, lepiej jest co najmniej raz przejrzeć wyniki wyodrębniania, zanim przystąpi się do tworzenia kategorii.

Wstępna weryfikacja wyników może ujawnić elementy, które mechanizm wyodrębniania powinien potraktować inaczej niż zrobił to przy pierwszym wyodrębnianiu. Rozważmy następujące przykłady:

- **Nierozpoznane synonimy.** Załóżmy, że znaleźliśmy kilka pojęć, które uznajemy za synonimy, na przykład smart, intelligent, bright i knowledgeable, a wszystkie one figurują jako odrębne pojęcia na liście wyników. Możemy utworzyć definicję synonimu, w której intelligent, bright i knowledgeable zostaną zgrupowanej pod docelowym pojęciem smart. W ten sposób wszystkie te pojęcia zostaną zgrupowane z pojęciem smart, a globalna liczebność będzie wyższa. Więcej informacji zawiera temat “Dodawanie synonimów” na stronie 88.
- **Pojęcie z błędnie przypisanym typem.** Załóżmy, że pojęcia w wynikach wyodrębniania są przypisane do jednego typu, ale chcemy je przypisać do innego. Inny przykład: wyobraźmy sobie, że w wynikach wyodrębniania znajdujemy 15 pojęć oznaczających warzywa i chcemy, aby wszystkie były przypisane do nowego typu o nazwie <Vegetable>. W większości języków pojęcia nieznalezione w żadnym słowniku typu, ale wyodrębnione z tekstu, otrzymują automatycznie przypisywany typ <Unknown> Możliwe jest dodawanie pojęć do typów. Więcej informacji zawiera temat “Dodawanie pojęć do typów” na stronie 89.
- **Pojęcia nieistotne.** Załóżmy, że znaleźliśmy wyodrębnione pojęcie o bardzo dużej liczebności, tj. występujące w bardzo wielu rekordach lub dokumentach. Uznajemy jednak, że to pojęcie jest nieistotne dla naszej analizy. Możemy je wykluczyć z wyodrębniania. Więcej informacji zawiera temat “Wykluczanie pojęć z wyodrębniania” na stronie 90.
- **Nieprawidłowe dopasowania.** Załóżmy, że przeglądamy rekordy lub dokumenty zawierające określone pojęcie i odkrywamy, że dwa wyrazy zostały błędnie zgrupowane (na przykład faculty i facility). Takie dopasowanie może być skutkiem działania wewnętrznego algorytmu grupowania rozmytego, który tymczasowo ignoruje ciągi dwóch lub trzech spółgłosek i samogłosek, aby pogrupować często spotykane błędne formy zapisu tego samego wyrazu. Możemy dodać te dwa wyrazy do listy par, które nie powinny być grupowane. Więcej informacji zawiera temat “Grupowanie rozmyte” na stronie 191. Grupowanie rozmyte nie jest dostępne w przypadku tekstu japońskiego.
- **Niewyodrębnione pojęcia.** Załóżmy, że spodziewamy się znaleźć w wynikach wyodrębniania określone pojęcia, ale przeglądając tekst w rekordach lub dokumentach zauważamy, że niektóre wyrazy lub frazy nie zostały wyodrębnione. Często są to nieinteresujące nas czasowniki lub przymiotniki. Jednak często chcemy użyć wyrazu, który nie został wyodrębniony, w definicji jednej z kategorii. Aby wyodrębnić pojęcie, można wymusić wpisanie terminu do słownika typu. Więcej informacji zawiera temat “Wymuszanie wyodrębniania wyrazów” na stronie 90.

Wiele z tych zmian można wprowadzić bezpośrednio na panelu Extraction Results, panelu Data, w oknie dialogowym Category Definitions lub w oknie dialogowym Cluster Definitions poprzez wybieranie elementów i klikanie prawym przyciskiem myszy w celu wywołania menu kontekstowego.

Po prowadzeniu zmian kolor tła panelu zmieni się, wskazując, że trzeba ponownie przeprowadzić wyodrębnianie. Więcej informacji zawiera temat “Wyodrębnianie danych” na stronie 80. W przypadku pracy z obszernymi zbiorami danych bardziej efektywnym rozwiązaniem może okazać się ponawianie wyodrębniania po wprowadzeniu kilku zmian, a nie po każdej zmianie.

Uwaga: Cały zbiór edytowalnych zasobów lingwistycznych wykorzystany do wygenerowania wyników wyodrębniania można wyświetlić w widoku Resource Editor (View > Resource Editor). Zasoby te mają postać bibliotek i słowników. Można modyfikować pojęcia i typy bezpośrednio w bibliotekach i słownikach. Więcej informacji zawiera Rozdział 15, “Praca z bibliotekami”, na stronie 167.

Dodawanie synonimów

Synonimy kojarzą dwa lub większą liczbę wyrazów o tym samym znaczeniu. Można również użyć synonimów do grupowania terminów z ich skrótami lub do grupowania najczęściej występujących błędnie zapisanych form terminu z jego poprawną formą. Użycie synonimów zwiększa częstość występowania pojęcia docelowego, co ułatwia odkrywanie podobnych informacji wyrażonych w danych w różny sposób.

Szablony zasobów lingwistycznych i biblioteki dostarczone z produktem zawierają wiele predefiniowanych synonimów. Jeśli jednak odkryjesz nierozpoznane synonimy, możesz je zdefiniować, aby zostały rozpoznane przy następnym wyodrębnianiu.

Pierwszym krokiem jest wybór pojęcia docelowego, czyli wiodącego. *Pojęcie docelowe* to wyraz lub fraza, pod którym/którą chcesz zgrupować wszystkie synonimy w ostatecznych wynikach. W trakcie wyodrębniania synonimy są grupowane pod tym pojęciem docelowym. Drugim krokiem jest identyfikacja wszystkich synonimów tego pojęcia. Wszystkie synonimy w ostatecznych wynikach wyodrębniania zostaną zastąpione pojęciem docelowym. Aby termin mógł być synonimem, musi zostać wyodrębniony. Jednak pojęcie docelowe nie musi być koniecznie wyodrębnione, aby miało miejsce zastąpienie. Na przykład, jeśli termin intelligent ma być zastępowany terminem smart, to intelligent jest synonimem, a smart jest pojęciem docelowym.

Gdy utworzysz nową definicję synonimu, do słownika zostanie dodane nowe pojęcie docelowe. Następnie musisz dodać synonimy do pojęcia docelowego. Każde utworzenie lub zmiana synonimu jest rejestrowana w słownikach synonimów w edytorze Resource Editor. Jeśli chcesz przejrzeć całą zawartość tych słowników synonimów lub wprowadzić dużą liczbę zmian, lepszym wyjściem może być praca bezpośrednio w oknie Resource Editor. Więcej informacji zawiera temat “Słowniki zastąpień/synonimów” na stronie 184.

Wszystkie nowe synonimy są automatycznie zapisywane w pierwszej bibliotece wymienionej w drzewie bibliotek edytora Resource Editor; domyślnie jest to biblioteka o nazwie *Local Library*.

Uwaga: Jeśli nie możesz znaleźć definicji synonimu za pośrednictwem menu kontekstowych lub bezpośrednio w oknie Resource Editor, to być może doszło do dopasowania w wyniku wewnętrznego grupowania rozmytego. Więcej informacji zawiera temat “Grupowanie rozmyte” na stronie 191.

Aby utworzyć nowy synonim

1. W panelu Extraction Results, panelu Data, oknie dialogowym Category Definitions lub oknie dialogowym Cluster Definitions wybierz pojęcia, dla których chcesz utworzyć nowy synonim.
2. Z menu wybierz opcje **Edit > Add to Synonym > New**. Zostanie otwarte okno dialogowe Create Synonym.
3. Wprowadź pojęcie docelowe w polu tekstowym Target. Jest to pojęcie, pod którym zostaną zgrupowane wszystkie synonimy.
4. Aby dodać więcej synonimów, wprowadź je w polu listy Synonyms. Oddzielaj synonimy separatorem globalnym. Więcej informacji zawiera temat “Okno dialogowe Options: karta Session” na stronie 75.
5. W przypadku pracy z tekstem w języku japońskim należy przypisać typ do tych synonimów przez wybranie nazwy typu w polu **Synonyms from type**. Jednak termin docelowy ma typ przypisany podczas wyodrębniania. Jeśli jednak termin docelowy nie został wyodrębniony jako pojęcie, to w wynikach wyodrębniania zostanie mu przypisany typ podany w tej kolumnie.

6. Kliknij przycisk **OK**, aby zastosować zmiany. Okno dialogowe zostanie zamknięte, a kolor tła panelu Extraction Results zmieni się, wskazując, że trzeba ponownie przeprowadzić wyodrębnianie, aby zobaczyć wprowadzone zmiany. Jeśli chcesz wprowadzić kilka zmian, wprowadź je wszystkie przed ponownym wyodrębnianiem.

Aby uzupełnić synonim

1. W panelu Extraction Results, panelu Data, oknie dialogowym Category Definitions lub oknie dialogowym Cluster Definitions wybierz pojęcia, które chcesz dodać do istniejącej definicji synonimu.
2. Z menu wybierz opcje **Edit > Add to Synonym**. Menu zawiera zestaw synonimów, a ostatnio utworzony znajduje się na początku listy. Wybierz nazwę synonimu, do którego chcesz dodać wybrane pojęcia. Jeśli wyświetlana jest nazwa synonimu, którego szukasz, wybierz go, a wybrane pojęcia zostaną dodane do tej definicji synonimu. Jeśli żądany synonim nie jest widoczny, wybierz opcję **More**, aby wyświetlić okno dialogowe All Synonyms.
3. W oknie dialogowym All Synonyms można posortować listę według kolejności naturalnej (kolejności tworzenia) lub w porządku rosnącym lub malejącym. Wybierz nazwę synonimu, do którego chcesz dodać wybrane pojęcia, a następnie kliknij przycisk **OK**. Okno dialogowe zostanie zamknięte, a pojęcia zostaną dodane do definicji synonimu.

Dodawanie pojęć do typów

W trakcie wyodrębniania wyodrębnione pojęcia są przypisywane do typów w celu zgrupowania terminów, które mają ze sobą coś wspólnego. IBM SPSS Modeler Text Analytics jest dostarczany z wieloma typami wbudowanymi. Więcej informacji zawiera temat “Typy wbudowane” na stronie 178. W większości języków pojęcia nieznalezione w żadnym słowniku typu, ale wyodrębnione z tekstu, otrzymują automatycznie przypisywany typ <Unknown>

Przeglądając wyniki, możesz dostrzec pojęcia, którym został przypisany niewłaściwy typ. Bywa, że pewna grupa wyrazów powinna mieć swój własny typ. W tych przypadkach celowe byłoby przypisanie pojęć do innego typu lub utworzenie zupełnie nowego typu. Nie można tworzyć nowych typów dla języka japońskiego.

Załóżmy na przykład, że pracujemy z danymi z ankiety dotyczącej samochodów i interesuje nas kategoryzacja według różnych obszarów pojazdu. Możemy utworzyć typ <Dashboard>, który będzie grupował wszystkie pojęcia związane ze wskaźnikami, przyciskami itp. na deskach rozdzielczych pojazdów. Następnie moglibyśmy przypisać do tego nowego typu takie pojęcia, jak gas gauge, heater, radio i odometer.

Inny przykład: załóżmy że pracujemy z danymi ankietowymi dotyczącymi uczelni wyższych, a mechanizm wyodrębniania przypisał nazwę uczelni Johns Hopkins (chodzi o uniwersytet) do typu <Person> zamiast do typu <Organization>. W takim przypadku możemy dodać to pojęcie do typu <Organization>.

Za każdym razem, gdy tworzysz typ lub dodajesz pojęcia do listy terminów typu zmiany te są rejestrowane w słownikach typów należących do bibliotek zasobów lingwistycznych w edytorze Resource Editor. Jeśli chcesz przejrzeć zawartości tych bibliotek lub wprowadzić dużą liczbę zmian, lepszym wyjściem może być praca bezpośrednio w oknie Resource Editor. Więcej informacji zawiera temat “Dodawanie terminów” na stronie 180.

Aby dodać pojęcie do typu

1. W panelu Extraction Results, panelu Data, oknie dialogowym Category Definitions lub oknie dialogowym Cluster Definitions wybierz pojęcia, które chcesz dodać do istniejącego typu.
2. Kliknij prawym przyciskiem myszy, aby otworzyć menu kontekstowe.
3. Z menu wybierz opcje **Edit > Add to Type**. Menu zawiera zestaw typów, a ostatnio utworzony znajduje się na początku listy. Wybierz nazwę typu, do którego chcesz dodać wybrane pojęcia. Jeśli wyświetlana jest nazwa typu, którego szukasz, wybierz go, a wybrane pojęcia zostaną dodane do tego typu. Jeśli żądany typ nie jest widoczny, wybierz opcję **More**, aby wyświetlić okno dialogowe All Types.
4. W oknie dialogowym All Synonyms można posortować listę według kolejności naturalnej (tworzenia) lub w porządku rosnącym lub malejącym. Wybierz nazwę typu, do którego chcesz dodać wybrane pojęcia, a następnie kliknij przycisk **OK**. Okno dialogowe zostanie zamknięte, a pojęcia zostaną dodane jako terminy do typu.

Uwaga: W przypadku tekstu japońskiego zdarzają się sytuacje, w których zmiana typu terminu nie wpłynie na to, jaki typ zostanie mu ostatecznie przypisany w wynikach wyodrębniania. Dzieje się tak, że ponieważ w przypadku niektórych podstawowych terminów słowniki wewnętrzne mają pierwszeństwo.

Aby utworzyć nowy typ

1. W panelu Extraction Results, panelu Data, oknie dialogowym Category Definitions lub oknie dialogowym Cluster Definitions wybierz pojęcia, dla których chcesz utworzyć nowy typ.
2. Z menu wybierz opcje **Edit > Add to Type > New**. Zostanie otwarte okno dialogowe Type Properties.
3. Wprowadź nową nazwę dla tego typu w polu tekstowym Name i wprowadź zmiany do innych pól. Więcej informacji zawiera temat “Tworzenie typów” na stronie 179.
4. Kliknij przycisk **OK**, aby zastosować zmiany. Okno dialogowe zostanie zamknięte, a kolor tła panelu Extraction Results zmieni się, wskazując, że trzeba ponownie przeprowadzić wyodrębnianie, aby zobaczyć wprowadzone zmiany. Jeśli chcesz wprowadzić kilka zmian, wprowadź je wszystkie przed ponownym wyodrębnianiem.

Wykluczanie pojęć z wyodrębniania

Przeglądając wyniki, możesz trafić na pojęcia, które nie powinny być wyodrębniane ani wykorzystywane przez techniki zautomatyzowanego budowania kategorii. W niektórych przypadkach pojęcia te mają bardzo dużą liczebność, ale są całkowicie nieistotne dla analizy. Wówczas można oznaczyć pojęcie jako wykluczone z ostatecznych wyników wyodrębniania. Zwykle do tego słownika dodaje się wyrazy wypełniające lub frazy używane w tekście dla zapewnienia jego ciągłości, ale nie wnoszące nic ważnego, które mogłyby tylko zaciemnić wyniki wyodrębniania. Dodając te pojęcia do słownika wykluczeń powodujesz, że nie będą nigdy wyodrębniane.

Wykluczenie pojęcia powoduje, że wszystkie warianty wykluczonego pojęcia znikną z wyników następnego wyodrębniania. Jeśli wykluczone pojęcie jest już deskryptorem jakiejś kategorii, to po ponownym wyodrębnianiu pozostanie w tej kategorii z licznością zerową.

Wykluczenia są rejestrowane w słowniku wykluczeń w edytorze Resource Editor. Jeśli chcesz przejrzeć wszystkie definicje wykluczeń i edytować je bezpośrednio, lepszym wyjście może być praca bezpośrednio w oknie Resource Editor. Więcej informacji zawiera temat “Słowniki wykluczeń” na stronie 187.

Uwaga: W przypadku tekstu japońskiego istnieją sytuacje, w których wykluczenie terminu lub typu nie odniesie skutku. Dzieje się tak, że ponieważ w przypadku niektórych podstawowych terminów słowniki wewnętrzne z zasobów japońskich mają pierwszeństwo.

Aby wykluczyć pojęcia

1. W panelu Extraction Results, panelu Data, oknie dialogowym Category Definitions lub oknie dialogowym Cluster Definitions wybierz pojęcia, które chcesz wykluczyć z wyodrębniania.
2. Kliknij prawym przyciskiem myszy, aby otworzyć menu kontekstowe.
3. Wybierz opcję **Exclude from Extraction**. Pojęcie zostanie dodane do słownika wykluczeń w edytorze Resource Editor, a kolor tła panelu Extraction Results zmieni się, wskazując, że trzeba ponownie przeprowadzić wyodrębnianie, aby zobaczyć wprowadzone zmiany. Jeśli chcesz wprowadzić kilka zmian, wprowadź je wszystkie przed ponownym wyodrębnianiem.

Uwaga: Wszystkie wykluczone terminy są automatycznie zapisywane w pierwszej bibliotece wymienionej w drzewie bibliotek edytora Resource Editor; domyślnie jest to biblioteka o nazwie *Local Library*.

Wymuszanie wyodrębniania wyrazów

Przeglądając dane tekstowe w panelu Data po zakończeniu wyodrębniania, możesz zauważyć wyrazy lub frazy, które nie były wyodrębnione. Często są to nieinteresujące nas czasowniki lub przymiotniki. Jednak często chcemy użyć wyrazu, który nie został wyodrębniony, w definicji jednej z kategorii.

Jeśli chcesz wyodrębnić te wyrazy i frazy, możesz wymusić wpisanie terminu do słownika typu. Więcej informacji zawiera temat “Wymuszanie terminów” na stronie 183.

Ważne! Oznaczenie terminu w słowniku jako wymuszonego nie zawsze gwarantuje pożądany skutek. Nawet jeśli jawnie dodasz termin do słownika, może on nie pojawić się w wynikach ponownego wyodrębniania lub pojawi się, ale nieco inaczej niż został zadeklarowany. Choć takie sytuacje są rzadkością, mogą wystąpić, jeśli wyraz lub frazę wyodrębniono już wcześniej jako część dłuższej frazy. Aby temu zapobiec, w słowniku typu dla tego terminu wybierz opcję dopasowywania **Entire (no compounds)**. Więcej informacji zawiera temat “Dodawanie terminów” na stronie 180.

Rozdział 9. Kategoryzacja danych tekstowych

W widoku Categories and Concepts można tworzyć **kategorie**, które zasadniczo reprezentują bardziej ogólne pojęcia albo tematy i pozwolą wychwycić kluczowe idee, informacje i postawy wyrażone w tekście.

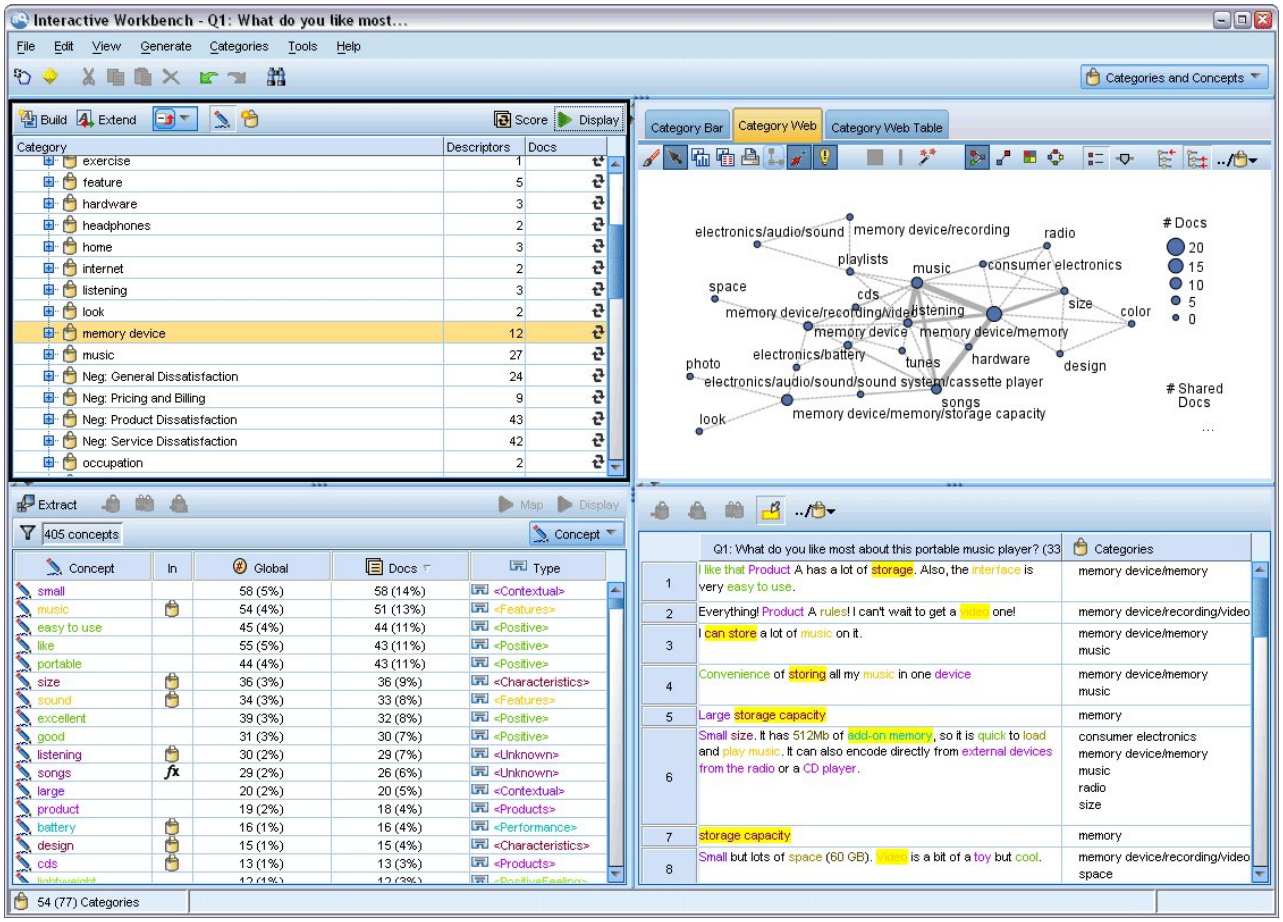
Począwszy od wersji 14 produktu IBM SPSS Modeler Text Analytics, kategorie mogą również mieć strukturę hierarchiczną, co oznacza, że mogą zawierać podkategorie, a te podkategorie mogą zawierać własne podkategorie itd. Istnieje możliwość zaimportowania predefiniowanych struktur kategorii, uprzednio zwanych ramkami kodu, z hierarchiczną strukturą kategorii, a także tworzenia kategorii hierarchicznych bezpośrednio w programie.

W rezultacie kategorie hierarchiczne umożliwiają zbudowanie struktury drzewa z jedną lub większą liczbą podkategorii w celu bardziej dokładnego grupowania elementów, takich jak różne pojęcia lub obszary tematyczne. Oto prosty przykład związany ze spędzaniem wolnego czasu: odpowiedź na pytanie *What activity would you like to do if you had more time?* może przynieść kategorie najwyższego poziomu *sports*, *art and craft*, *fishing* itd.; na poziomie o jeden niższym niż *sports* możemy mieć kategorie odpowiadające różnym grupom dyscyplin: *ball games*, *water-related* itd.

Kategoria składa się z zestawu deskryptorów, takich jak *pojęcia*, *typy*, *wzorce* i *reguły kategorii*. Razem te deskryptory służą do określania, czy konkretny dokument lub rekord należy do danej kategorii. Tekst w dokumencie lub rekordzie można przeszukać, aby sprawdzić, czy jakikolwiek jego fragment pasuje do deskryptora. Jeśli zostanie znalezione dopasowanie, dokument/rekord jest przypisywany do tej kategorii. Ten proces jest nazywany **klasyfikowaniem**.

Można opracowywać, budować i wizualnie eksplorować kategorie, korzystając z danych prezentowanych w czterech panelach widoku Categories and Concepts, z których każdy może być ukryty lub wyświetlany poprzez wybranie odpowiedniej nazwy panelu z menu View.

- **Panel Categories.** Ten panel służy do tworzenia kategorii i zarządzania nimi. Więcej informacji zawiera temat “Panel Categories” na stronie 94.
- **Panel Extraction Results.** Ten panel służy do eksploracji i pracy z wyodrębnionymi pojęciami i typami. Więcej informacji zawiera temat “Wyniki wyodrębniania: pojęcia i typy” na stronie 79.
- **Panel Visualization.** W tym panelu możliwe jest wizualne eksplorowanie kategorii i ich interakcji. Więcej informacji zawiera temat “Wykresy i tabele kategorii” na stronie 147.
- **Panel Data.** Ten panel służy do eksplorowania i przeglądania tekstu zawartego w dokumentach i rekordach odpowiadających wyborom dokonanym w tym panelu. Więcej informacji zawiera temat “Panel Data” na stronie 101.



Rysunek 29. Widok Categories and Concepts

Można zacząć pracę od zestawu kategorii zaimportowanego z pakietu analizy tekstu (TAP) lub zaimportować kategorie z predefiniowanego pliku kategorii, można także tworzyć własne kategorie. Kategorie mogą być tworzone automatycznie przy wykorzystaniu dostępnego w produkcie, wszechstronnego zestawu technik automatycznych, które używają wyników wyodrębniania (pojęć, typów i wzorców) do generowania kategorii i ich deskryptorów. Kategorie mogą być również tworzone ręcznie na postawie dodatkowych spostrzeżeń użytkownika. Jednak ręczne tworzenie kategorii lub ich optymalizowanie jest możliwe tylko w interaktywnym pulpicie roboczym. Więcej informacji zawiera temat “Węzeł Text Mining: karta Model” na stronie 23. Definicje kategorii można tworzyć ręcznie, przeciągając i upuszczając wyniki wyodrębniania do kategorii. Można wzbogacać te kategorie lub puste kategorie poprzez dodawanie reguł kategorii, przy użyciu własnych predefiniowanych kategorii lub poprzez zastosowanie kombinacji tych metod.

Każda z dostępnych technik dobrze nadaje się do pracy z określonymi rodzajami danych i w określonych warunkach, jednak często przydatna jest możliwość połączenia w jednej analizie różnych technik w celu wydobycia bogatszego zbioru informacji z dokumentów lub rekordów. W trakcie kategoryzacji użytkownik może też dostrzec inne zmiany, które warto byłoby wprowadzić w zasobach lingwistycznych.

Panel Categories

Panel Categories to obszar, w którym można tworzyć kategorie i zarządzać nimi. Ten panel znajduje się w lewym górnym rogu widoku Categories and Concepts. Po zakończeniu wyodrębniania pojęć i typów z danych tekstowych można rozpocząć automatyczne tworzenie kategorii za pomocą takich technik, jak włączenie pojęcia, współwystąpienia itp., albo ręcznie. Więcej informacji zawiera temat “Budowanie kategorii” na stronie 103.

Za każdym razem, gdy kategoria jest tworzona lub aktualizowana, dokumenty lub rekordy można ocenić, klikając przycisk **Score** w celu sprawdzenia, czy jakikolwiek fragment tekstu jest zgodny z deskryptorem w danej kategorii.

Jeśli zostanie znalezione dopasowanie, dokument lub rekord jest przypisywany do tej kategorii. W rezultacie większość (jeśli nie sto procent) dokumentów lub rekordów zostaje przypisana do kategorii na podstawie deskryptorów w kategoriach.

Uwaga: Jeśli w widocznym panelu mieści się więcej kategorii, można użyć przycisków znajdujących się w dolnej części panelu, aby przeglądać kategorie, lub wprowadzić numer strony i przejść do niej.

Tabela drzewa kategorii

Tabela drzewa w tym panelu prezentuje zestaw kategorii, podkategorii i deskryptorów. Drzewo zawiera również kilka kolumn prezentujących informacje o każdym elemencie drzewa. Do wyświetlania mogą być dostępne następujące kolumny:

- **Code** Wyświetla wartość kodu dla każdej kategorii. Ta kolumna jest domyślnie ukryta. Tę kolumnę można wyświetlić za pomocą menu: **View > Categories Pane**.
- **Category.** Zawiera drzewo kategorii z nazwami kategorii i podkategorii. Dodatkowo, jeśli zostanie kliknięta ikona deskryptorów na pasku narzędzi, zostanie wyświetlony także zestaw deskryptorów.
- **Descriptors.** Liczba deskryptorów, które składają się na definicję. Ta liczba nie zawiera liczby deskryptorów w podkategoriach. Liczba nie jest podana, gdy nazwa deskryptora jest widoczna w kolumnie **Categories**. Możesz wyświetlać lub ukrywać deskryptory w drzewie samodzielnie za pośrednictwem menu: **View > Categories Pane > All Descriptors**.
- **Docs** Po zakończeniu oceniania ta kolumna zawiera liczbę dokumentów lub rekordów, które są skategoryzowane w kategorii i wszystkich jej podkategoriach. Jeśli więc 5 rekordów odpowiada kategorii najwyższego poziomu na podstawie deskryptorów, a 7 innych rekordów odpowiada podkategorii na podstawie deskryptorów, to łączna liczba dokumentów dla kategorii najwyższego poziomu jest sumą obu tych liczb (12). Jeśli jednak ten sam rekord pasuje do kategorii najwyższego poziomu i jej podkategorii, wówczas liczba wynosiłaby 11.

Jeśli nie istnieją żadne kategorie, tabela nadal zawiera dwa wiersze. Najwyższy wiersz o nazwie **All Documents**, zawiera łączną liczbę dokumentów lub rekordów. Drugi wiersz o nazwie **Uncategorized** zawiera liczbę dokumentów/rekordów, które mają jeszcze zostać sklasyfikowane.

Dla każdej kategorii w panelu mała żółta ikona poprzedza nazwę kategorii. Jeśli klikniesz dwukrotnie kategorię lub wybierzesz opcję **View > Category Definitions** w menu, zostanie otwarte okno dialogowe Category Definitions ze wszystkimi elementami, nazywanymi *deskryptorami*, które składają się na definicję kategorii, takimi jak pojęcia, typy, wzorce i reguły kategorii. Więcej informacji zawiera temat “Informacje o kategoriach” na stronie 100. Domyślnie tabela drzewa kategorii nie prezentuje deskryptorów w kategoriach. Jeśli chcesz wyświetlić deskryptory bezpośrednio w drzewie, a nie w oknie dialogowym Category Definitions, kliknij przycisk przełącznika z ołówkiem na pasku narzędzi. Jeśli ten przełącznik jest wybrany, można rozwinąć drzewo, aby wyświetlić także deskryptory.

Ocenianie kategorii

Kolumna **Docs**. w tabeli drzewa kategorii wyświetla liczbę dokumentów lub rekordów, które są przypisane do tej konkretnej kategorii. Jeśli liczby są nieaktualne lub nie są obliczone, w kolumnie tej wyświetlana jest ikona. Można kliknąć przycisk **Score** na pasku narzędzi panelu, aby ponownie obliczyć liczbę dokumentów. Należy pamiętać, że w przypadku pracy z dużymi zbiorami danych proces oceniania może zająć dużo czasu.

Wybieranie kategorii w drzewie

Dokonując wyboru w drzewie, można wybierać tylko kategorie równorzędne — tj. jeśli wybierzesz kategorie najwyższego poziomu, nie możesz wybrać podkategorii. A jeśli wybierzesz 2 podkategorie danej kategorii, nie możesz równocześnie wybrać podkategorii innej kategorii. Próba wybrania nieciągłego zestawu kategorii spowoduje utratę poprzedniego wyboru.

Wyświetlanie paneli Data i Visualization

Po wybraniu wiersza w tabeli można kliknąć przycisk **Display**, aby odświeżyć panele Visualization i Data z uwzględnieniem informacji wynikających z dokonanego wyboru. Jeśli panel nie jest widoczny, kliknięcie przycisku **Display** spowoduje wyświetlenie panelu.

Optymalizacja kategorii

Kategoryzacja za pierwszym razem może nie przynieść idealnych wyników i mogą pojawić się kategorie, które zechcesz usunąć lub połączyć z innymi kategoriami. Może również okazać się — po dokonaniu przeglądu wyników wyodrębniania — że istnieją kategorie, które nie zostały utworzone, a byłyby przydatne. W takiej sytuacji można wprowadzić ręczne zmiany do wyników, aby lepiej dostosować je do konkretnego kontekstu. Więcej informacji zawiera temat “Edytowanie i optymalizacja kategorii” na stronie 132.

Metody i strategie tworzenia kategorii

Jeśli jeszcze nie przeprowadzono wyodrębniania lub wyniki wyodrębniania są nieaktualne, próba użycia jednej z technik budowania lub uzupełniania kategorii spowoduje wyświetlenie monitu o przeprowadzenie wyodrębniania. Po zastosowaniu wybranej techniki pojęcia i typy pogrupowane w kategorię nadal mogą być wykorzystywane do tworzenia kategorii przy użyciu innych technik. Oznacza to, że jedno pojęcie może wystąpić w wielu kategoriach, chyba że wprost zabroniono ich ponownego wykorzystania.

Przestrzeganie poniższych wskazówek pomoże w tworzeniu optymalnych kategorii:

- **Metody tworzenia kategorii**
- **Strategie tworzenia kategorii**
- **Porady dotyczące tworzenia kategorii**

Metody tworzenia kategorii

Ponieważ każdy zestaw danych jest inny, liczba metod tworzenia kategorii i kolejność ich stosowania może się z czasem zmieniać. Ponadto, ponieważ także cele eksploracji różnych zbiorów danych mogą być różne, niekiedy konieczne jest eksperymentowanie z różnymi metodami w celu wybrania tej, która przyniesie najlepsze rezultaty analizy konkretnych danych tekstowych. Żadna z technik automatycznych nie zapewni idealnej kategoryzacji danych; dlatego zalecamy wypróbowanie różnych technik automatycznych i stosowanie tych, które najlepiej sprawdzają się przy pracy z konkretnymi danymi.

Oprócz zastosowania pakietów analizy tekstu (TAP, **.tap*) z gotowymi zestawami kategorii można także kategoryzować odpowiedzi przy użyciu dowolnej kombinacji następujących metod.

- **Techniki budowania automatycznego.** Dostępnych jest kilka opcji automatycznego tworzenia kategorii na podstawie analizy lingwistycznej lub analizy liczebności wystąpień. Więcej informacji zawiera temat “Budowanie kategorii” na stronie 103.
- **Techniki uzupełniania automatycznego.** Dostępnych jest kilka technik lingwistycznych umożliwiających uzupełnianie istniejących kategorii o nowe lub ulepszone deskryptory, tak aby wychwytywały więcej rekordów. Więcej informacji zawiera temat “Uzupełnianie kategorii” na stronie 112.
- **Techniki ręczne.** Istnieje kilka technik ręcznych, takich jak przeciąganie i upuszczanie. Więcej informacji zawiera temat “Ręczne tworzenie kategorii” na stronie 115.

Strategie tworzenia kategorii

Poniższa lista strategii w żadnej mierze nie nie jest wyczerpująca, ale daje ogólne pojęcie o możliwych podejściach do procesu budowania kategorii.

- Gdy definiujesz węzeł Text Mining, wybierz zestaw kategorii z pakietu analizy tekstu (TAP), aby na początku analizy mieć już do dyspozycji pewne gotowe kategorie. Możliwe, że kategorie te wystarczą do zadowalającej kategoryzacji tekstu. Jeśli jednak chcesz dodać więcej kategorii, możesz edytować ustawienia budowania kategorii

(**Categories > Build Settings**). Otwórz okno dialogowe **Advanced Settings: Linguistics** i wybierz opcję danych wejściowych do tworzenia kategorii **Unused extraction results**, a następnie zbuduj dodatkowe kategorie.

- Gdy definiujesz węzeł, wybierz zestaw kategorii z pakietu TAP w widoku Categories and Concepts na interaktywnym pulpicie roboczym. Następnie przeciągaj i upuszczaj nieużywane pojęcia lub wzorce do kategorii, tak jak uznasz to za stosowne. Teraz uzupełnij istniejące i zmodyfikowane właśnie kategorie (**Categories > Extend Categories**), aby uzyskać więcej deskryptorów związanych z istniejącymi deskryptorami kategorii.
- Zbuduj kategorie automatycznie, korzystając z zaawansowanych ustawień lingwistycznych (**Categories > Build Categories**). Następnie ręcznie zoptymalizuj kategorie, usuwając deskryptory, usuwając kategorie lub scalając podobne kategorie aż do uzyskania zadowalających definicji kategorii. Ponadto, jeśli pierwotnie kategorie zbudowano **bez** użycia opcji **Generalize with wildcards where possible**, możesz podjąć próbę automatycznego uproszczenia kategorii za pomocą opcji **Generalize**.
- Zaimportuj predefiniowany plik kategorii z bardzo opisowymi nazwami kategorii i/lub adnotacjami. Ponadto, jeśli pierwotnie przeprowadzano import **bez** opcji importowania lub generowania deskryptorów na podstawie nazw kategorii, możesz później przejść do okna dialogowego Extend Categories i wybrać opcję **Extend empty categories with descriptors generated from the category name** . Następnie ponownie uzupełnij te kategorie, ale tym razem korzystaj z technik grupowania.
- Ręcznie utwórz pierwszy zestaw kategorii, sortując pojęcia lub wzorce pojęć według liczebności, a następnie przeciągając i upuszczając najbardziej interesujące w panelu Categories. Po zapisaniu pierwszego zestawu kategorii użyj funkcji Extend (**Categories > Extend Categories**) w celu uzupełnienia i zoptymalizowania wszystkich wybranych kategorii, aby uwzględniały inne powiązane deskryptory i w efekcie znajdowały więcej rekordów.

Zalecamy, aby po zastosowaniu tych technik dokonać przeglądu wynikowych kategorii i ręcznie wprowadzić drobne korekty, usunąć wszelkie błędy w klasyfikacji lub dodać pominięte rekordy lub wyrazy. Ponadto, ponieważ użycie kilku różnych technik może doprowadzić do powstania nadmiarowych kategorii, można także scalić lub usunąć zbędne kategorie. Więcej informacji zawiera temat “Edytowanie i optymalizacja kategorii” na stronie 132.

Porady dotyczące tworzenia kategorii

Aby ułatwić użytkownikom tworzenie optymalnych kategorii, przedstawiamy pewne wskazówki, które mogą pomóc w podejmowaniu decyzji o wyborze strategii działania.

Wskazówki dotyczące współczynnika kategorie/dokumenty

Kategorie, do których dokumenty i rekordy są przypisywane, rzadko wykluczają się wzajemnie w jakościowej analizie tekstu. Jest to spowodowany co najmniej dwiema przyczynami:

- Po pierwsze, z reguły im dłuższy tekst w dokumencie lub rekordzie, tym bardziej jednoznaczne idee i opinie wyraża. Dlatego prawdopodobieństwo, że dokument lub rekord zostanie przypisany do wielu kategorii, jest znacznie większe.
- Po drugie, często istnieją różne sposoby pogrupowania i interpretacji dokumentów lub rekordów tekstowych, a sposoby te nie są logicznie odrębne. W przypadku ankiety z pytaniem otwartym o przekonania polityczne respondenta moglibyśmy utworzyć takie kategorie, jak *Liberal* i *Conservative* albo *Republican* i *Democrat*, a także bardziej szczegółowe kategorie, takie jak *Socially Liberal*, *Fiscally Conservative* itd. Te kategorie nie muszą być wzajemnie wykluczające się i dokładne.

Wskazówki dotyczące liczby kategorii do utworzenia

Kategorie powinny wynikać bezpośrednio z danych — widząc coś ciekawego w danych, można utworzyć kategorię reprezentującą tę informację. W ogólnym przypadku nie ma zalecanego górnego limitu liczby kategorii, które można utworzyć. Istnieje możliwość utworzenia zbyt wielu kategorii, by dało się z nimi pracować. Obowiązują dwie zasady:

- **Liczebność kategorii.** Aby kategoria była użyteczna, musi zawierać minimalną liczbę dokumentów lub rekordów. Jeden lub dwa dokumenty mogą zawierać coś bardzo intrygującego, ale jeśli są to tylko nieliczne z 1000 dokumentów, to zawarte w nich informacje mogą nie być wystarczająco częste w populacji, by były praktycznie użyteczne.

- **Złożoność.** Im więcej kategorii utworzysz, tym więcej musisz przejrzeć i podsumować po zakończeniu analizy. Jednak zbyt wiele kategorii może tylko skomplikować pracę, nie wnosząc użytecznych szczegółów.

Niestety, nie ma reguł pozwalających określić, ile kategorii to „zbyt wiele", ani jaka jest minimalna liczba rekordów w kategorii. Takich ustaleń trzeba dokonać na podstawie konkretnej sytuacji i potrzeb.

Mamy jednak kilka wskazówek ułatwiających rozpoczęcie pracy. Pomimo tego, że liczba kategorii nie powinna być zbyt duża, na wczesnych etapach analizy lepiej jest mieć za dużo niż za mało kategorii. Łatwiej jest grupować kategorie, które są względnie podobne do siebie, niż dzielić obserwacje między nowe kategorie, dlatego strategia pracy od większej do mniejszej liczby kategorii jest zwykle najlepsza. Ze względu na iteracyjny charakter eksploracji tekstu oraz łatwość, z jaką można ją prowadzić za pomocą tego programu oprogramowania, stworzenie na początku większej liczby kategorii jest w pełni akceptowalne.

Wybór najlepszych deskryptorów

Poniżej przedstawiamy wytyczne dotyczące wybierania i optymalnego tworzenia deskryptorów kategorii (pojęć, typów, wzorców TLA i reguł kategorii). Deskryptory są elementami składowymi kategorii. Gdy niektóre lub wszystkie elementy tekstu w dokumencie lub rekordzie pasują do deskryptora, ten dokument lub rekord jest przypisywany do kategorii zawierającej deskryptor.

Jeśli deskryptor nie odpowiada wyodrębnionemu pojęciu lub wzorcowi, nie zostanie on dopasowany do żadnych dokumentów ani rekordów. Dlatego należy używać pojęć, typów, wzorców i reguł kategorii tak, jak opisano w poniższych akapitach.

Ponieważ pojęcia reprezentują nie tylko same siebie, lecz również zestaw powiązanych terminów, np. formy pojedyncze/mnogie, synonimy i odmiany pisowni, to wyłącznie samo pojęcie powinno być deskryptorem lub częścią deskryptora. Aby dowiedzieć się więcej na temat terminów danego pojęcia, należy kliknąć nazwę pojęcia w panelu Extraction Results w widoku Categories and Concepts. Po zatrzymaniu wskaźnika myszy nad nazwą pojęcia pojawi się podpowiedź z terminami znalezionymi w tekście podczas ostatniego wyodrębniania. Nie wszystkie pojęcia zawierają terminy. Na przykład, jeśli car i vehicle są synonimami, ale car wyodrębniono jako pojęcie zawierające termin vehicle, to w deskryptorze należy użyć tylko pojęcia car, ponieważ zostanie ono dopasowane automatycznie także do dokumentów lub rekordów zawierających termin vehicle.

Pojęcia i typy jako deskryptory

Użyj pojęcia jako deskryptora, jeśli chcesz znaleźć wszystkie dokumenty lub rekordy zawierające pojęcie (lub dowolne z jego terminów). W tym przypadku stosowanie bardziej złożonych reguł kategorii nie jest konieczne, ponieważ dosłowna nazwa pojęcia jest wystarczająca. Należy pamiętać, że w przypadku użycia zasobów, które wyodrębniają opinie, niekiedy pojęcia mogą zmienić się podczas wyodrębniania wzorców TLA w celu wychwycenia rzeczywistego sensu zdania (w następnej sekcji zamieszczono przykład analizy TLA).

Na przykład na podstawie odpowiedzi respondentów opisujących ich ulubione owoce, takich jak „*Apple and pineapple are the best*", mogłyby zostać wyodrębnione wyrazy apple i pineapple. Jeśli dodasz pojęcie apple jako deskryptor do kategorii, wszystkie odpowiedzi z pojęciem apple (lub dowolnym z jego terminów) będą przypisywane do tej kategorii.

Jeśli jednak chcesz wiedzieć, w których odpowiedziach pojęcie *apple* występuje w jakimkolwiek kontekście, możesz napisać np. regułę kategorii * apple *, która wychwyci także odpowiedzi zawierające takie pojęcia, jak apple, apple sauce lub french apple tart.

Można również wychwycić wszystkie dokumenty lub rekordy, które zawierają pojęcia tego samego typu, podając bezpośrednio typ jako deskryptor (np. <Fruit>). Należy pamiętać, że nie można używać znaku * z typami.

Więcej informacji zawiera temat "Wyniki wyodrębniania: pojęcia i typy" na stronie 79.

Wzorce analizy powiązań w tekście (TLA) jako deskryptory

Zastosowanie wzorca TLA jako deskryptora jest celowe, gdy chcemy wychwytywać zniuansowane idee wyrażone nie wprost. Podczas wyodrębniania TLA tekst jest przetwarzany zdanie po zdaniu lub klauzula po klauzuli, a nie całościowo (jako cały dokument lub rekord). Jednoczesne uwzględnienie wszystkich części zdania umożliwia zidentyfikowanie np. opinii, relacji między dwoma elementami lub negacji, a tym samym trafniejsze wychwycenie rzeczywistego sensu zdania. Istnieje możliwość użycia wzorców pojęć lub wzorców typów jako deskryptorów. Więcej informacji zawiera temat “Wzorce typów i pojęć” na stronie 143.

Jeśli na przykład mamy tekst "*the room was not that clean*", to mogłyby zostać wyodrębnione pojęcie: room i clean. Jeśli jednak włączone jest wyodrębnianie TLA, to analiza TLA może wykryć, że wartość clean została użyta w kontekście negacji i w rzeczywistości oznacza not clean, co jest synonimem pojęcia dirty. Widzimy, że użycie samego pojęcia clean jako deskryptora spowodowałoby wychwycenie tego tekstu, ale także innych dokumentów lub rekordów, w których mowa o czystości. Z tego powodu lepszym rozwiązaniem może być użycie wzorca pojęcia TLA z pojęciem wynikowym dirty, ponieważ ono także pasuje do naszego tekstu, a prawdopodobnie jest bardziej odpowiednim deskryptorem.

Reguły biznesowe kategorii jako deskryptory

Reguły kategorii są instrukcjami, które automatycznie klasyfikują dokumenty lub rekordy w kategorie na podstawie wyrażenia logicznego, używając wyodrębnionych pojęć, typów i wzorców oraz operatorów boolowskich. Na przykład, można napisać wyrażenie, które oznacza *uwzględnij w tej kategorii wszystkie rekordy zawierające wyodrębnione pojęcie* embassy, *ale nie* argentina.

Można zapisywać i stosować reguły kategorii jako deskryptory w kategorii, aby wyrazić kilka różnych idei. Służą do tego operatory boolowskie &, | i !() . Szczegółowe informacje o składni reguł oraz sposobie ich pisania i edytowania zawiera sekcja “Korzystanie z reguł kategorii” na stronie 116.

- Użycie reguł kategorii z operatorem boolowskim & (AND) ułatwia znajdowanie dokumentów lub rekordów, w których występują co najmniej 2 pojęcia. Aby nastąpiło dopasowanie do kategorii, dwa lub więcej pojęć połączonych operatorami & nie musi występować w tym samym zdaniu lub frazie, ale może wystąpić w dowolnym miejscu tego samego dokumentu lub rekordu. Na przykład, jeśli jako deskryptor utworzymy regułę kategorii food & cheap, to znajdzie ona rekord zawierający tekst „*the food was pretty expensive, but the rooms were cheap*”, mimo że to nie rzeczownik food został określony mianem cheap. Dopasowanie wynika po prostu z obecności wyrazów food i cheap.
- Używając jako deskryptora reguły kategorii z operatorem boolowskim !() (NOT), można znaleźć dokumenty lub rekordy, w którym niektóre elementy występują, a inne nie. W ten sposób można uniknąć grupowania informacji, które wydają się powiązane na podstawie wyrazów, ale nie kontekstu. Na przykład, jeśli utworzysz jako deskryptor regułę kategorii <Organization> & !(ibm), to znajdzie ona tekst *SPSS Inc. was a company founded in 1967*, ale nie tekst *the software company was acquired by IBM.*.
- Użycie reguł kategorii z operatorem boolowskim| (OR) ułatwia znajdowanie dokumentów lub rekordów, w których występuje jedno z kilku pojęć i lub jeden z kilku typów. Na przykład, jeśli utworzysz jako deskryptor regułę kategorii (personnel|staff|team|coworkers) & bad, to znajdzie ona każdy dokument lub rekord, w którym dowolny z tych rzeczowników występuje razem z pojęciem bad.
- Użyj typów w regułach kategorii, aby uczynić je bardziej ogólnymi potencjalnie łatwiejszymi do wdrożenia. Załóżmy, że operujesz na danych dotyczących hotelu i chcesz dowiedzieć się, co klienci myślą o personelu hotelu. Do terminów powiązanych z tym zagadnieniem należą np. receptionist, waiter, waitress, reception desk, front desk itd. W takim przypadku możesz utworzyć nowy typ o nazwie <HotelStaff> i dodać wszystkie powyższe terminy do tego typu. Wprawdzie można utworzyć jedną regułę kategorii dla każdego rodzaju personelu, na przykład [* waitress * & nice], [* desk * & friendly], [* receptionist * & accommodating], można też utworzyć bardziej ogólną regułę kategorii w oparciu o typ more generic <HotelStaff>, aby wychwytywać wszystkie pozytywne opinie o personelu hotelu: [<HotelStaff> & <Positive>].

Uwaga: Stosując wzorce TLA w regułach kategorii, można używać zarówno +, jak i &. Więcej informacji zawiera temat “Używanie wzorców TLA w regułach kategorii” na stronie 118.

Przykłady różnic między dopasowywaniem pojęć, wzorców TLA i reguł kategorii

Poniższy przykład ilustruje wpływ użycia pojęcia, reguły i wzorca TLA jako deskryptora na kategoryzację dokumentów lub rekordów. Załóżmy, że masz 5 poniższych rekordów.

- A: *"awesome restaurant staff, excellent food and rooms comfortable and clean."*
- B: *"restaurant personnel was awful, but rooms were clean."*
- C: *"Comfortable, clean rooms."*
- D: *"My room was not that clean."*
- E: *"Clean."*

Ponieważ rekordy zawierają słowo *clean* i chcesz wychwycić tę informację, możesz utworzyć jeden z deskryptorów przedstawionych w poniższej tabeli. W zależności od idei, którą chcesz wychwycić, możesz wybierać deskryptory generujące różne wyniki.

Tabela 17. W jaki sposób przykładowe rekordy zostały dopasowane do deskryptorów.

Deskryptor	W	P	Z	D	E	Objaśnienie
clean	*match*	*match*	*match*	*match*	*match*	Deskryptor jest wyodrębnionym pojęciem. Każdy rekord zawierał pojęcie clean, nawet rekord D, ponieważ bez analizy TLA nie wiadomo, że „*not clean*" oznacza w istocie dirty.
clean + .	-	-	-	-	*match*	Deskryptor jest wzorcem TLA reprezentującym samo pojęcie clean. Zostanie dopasowany tylko do rekordu, w którym clean wyodrębniono bez powiązanego pojęcia podczas analizy TLA.
[clean]	*match*	*match*	*match*	-	*match*	Deskryptor jest regułą kategorii szukającą reguły TLA zawierającej pojęcie clean (samo lub w połączeniu z innymi). Dopasowane zostały wszystkie rekordy, których wyniki TLA zawierały clean, niezależnie od tego, czy pojęcie clean było powiązane z innym pojęciem, takim jak room, i niezależnie od tego, na której było pozycji.

Informacje o kategoriach

Kategorie dotyczą grupy blisko powiązanych pojęć, opinii lub postaw. Aby kategoria była przydatna, powinna umożliwiać łatwe opisanie za pomocą krótkiego zwrotu lub etykiety, która oddaje jej znaczenie.

Na przykład: jeśli analizowane są odpowiedzi w ankiecie przeprowadzonej wśród konsumentów dotyczącej nowego proszku do prania, można utworzyć kategorię z etykietą *odor*, która zawiera wszystkie odpowiedzi opisujące zapach produktu. Jednak taka kategoria nie rozróżniłaby pomiędzy osobami, które uważały zapach za miły i takimi, które uważały przeciwnie. Ponieważ produkt IBM SPSS Modeler Text Analytics może wyodrębniać opinie, kiedy używane są odpowiednie zasoby, można utworzyć dwie inne kategorie, aby identyfikować respondentów, którym *podobał się zapach* i respondentów, którym *nie podobał się zapach*.

Kategorie można utworzyć i obsługiwać w panelu Categories w lewym górnym rogu widoku Categories and Concepts . Każda kategoria jest definiowana przez jeden lub wiele deskryptorów. **Deskryptory** to pojęcia, typy i wzorce, jak również reguły kategorii, które zostały użyte do zdefiniowania kategorii.

Aby zobaczyć deskryptory, które tworzą daną kategorię, kliknij ikonę ołówka na pasku narzędzi panelu Categories, a następnie rozwiń drzewo, aby zobaczyć deskryptory. Można też wybrać kategorię i otworzyć okno dialogowe Category Definitions (**View > Category Definitions**).

Kiedy budujesz kategorie automatycznie, używając technik budowania kategorii, takich jak włączenie pojęcia, techniki będą używać pojęć i typów jako deskryptorów, aby tworzyć kategorie. Jeśli wyodrębniasz wzorce TLA, możesz również dodawać wzorce lub części tych wzorców jako deskryptory kategorii. Więcej informacji zawiera Rozdział 11, "Eksplorowanie analizy powiązań w tekście", na stronie 141. Jeśli budujesz grupy, możesz dodać pojęcia w grupie do nowych lub istniejących kategorii. Można też ręcznie utworzyć reguły kategorii, aby używać ich jako deskryptorów w swoich kategoriach. Więcej informacji zawiera temat "Korzystanie z reguł kategorii" na stronie 116.

Właściwości reguły

Oprócz deskryptorów kategorie mają także właściwości, które można edytować, aby zmieniać nazwy kategorii, dodawać etykiety, dodawać adnotacje.

Istnieją następujące właściwości:

- **Name.** Ta nazwa jest domyślnie wyświetlana w drzewie. Kategoriom tworzonym przy użyciu technik automatycznych nazwy nadawane są automatycznie.
- **Label.** Etykiety umożliwiają utworzenie bardziej czytelnych opisów kategorii, które można wykorzystać w innych programach, tabelach lub na wykresach. Wybranie opcji wyświetlania etykiet powoduje, że kategoria będzie w interfejsie oznaczona tą etykietą.
- **Code.** Numer kodowy odpowiadający wartości kodowej tej kategorii. .
- **Annotation.** W tym polu do każdej kategorii można dodać krótki opis. Gdy kategoria jest generowana za pomocą okna Build Categories, do tej adnotacji automatycznie dodawana jest notatka. Można też dodać do adnotacji przykładowy tekst bezpośrednio z panelu Data, zaznaczając tekst i wybierając z menu opcję **Categories > Add to Annotation**.

Panel Data

Podczas tworzenia kategorii może wystąpić potrzeba przejrzenia niektórych danych tekstowych, z którym pracujesz. Na przykład, jeśli utworzysz kategorię, do której należy 640 dokumentów, interesująca może być treść tych dokumentów i faktyczne brzmienie tekstu. Istnieje możliwość przeglądania rekordów lub dokumentów w panelu Data, który znajduje się w prawym dolnym rogu. Jeśli domyślnie panel ten nie jest wyświetlony, z menu wybierz kolejno opcje **View > Panes > Data**.

W panelu Data wyświetlany jest jeden wiersz na każdy dokument lub rekord odpowiadający wyborom dokonanym na panelu Categories, panelu Extraction Results lub oknie dialogowym Category Definitions, przy czym liczba wierszy nie może przekroczyć określonego limitu. Domyślnie liczba dokumentów lub rekordów wyświetlanych w panelu Data jest ograniczona, aby możliwe było szybsze wyświetlanie danych. Można jednak zmienić tę liczbę w oknie dialogowym Options. W przypadku bardzo dużych zbiorów danych szybkość wyświetlania może być zwiększona przez wyłączenie opcji, aby wyświetlić kategorie. Więcej informacji zawiera temat "Okno dialogowe Options: karta Session" na stronie 75.

Uwaga: Jeśli w widocznym panelu mieści się więcej rekordów, można użyć przycisków znajdujących się w dolnej części panelu, aby przeglądać rekordy, lub wprowadzić numer strony i przejść do niej.

Wyświetlanie i odświeżanie panelu Data

Panel Data nie jest odświeżany automatycznie, ponieważ w przypadku większych zbiorów danych automatyczne odświeżanie mogłoby być czasochłonne. Oznacza to, że za każdym razem, gdy dokonasz wyboru w innym panelu w tym widoku lub w oknie dialogowym Category Definitions, musisz kliknąć przycisk **Display**, aby odświeżyć zawartość panelu Data.

Dokumenty lub rekordy tekstowe

Jeśli dane tekstowe mają postać rekordów, a tekst jest stosunkowo krótki, to zmienna tekstowa w panelu Data zawiera całe dane tekstowe. Jednak podczas pracy z rekordami i większymi zbiorami danych w kolumnie zmiennej tekstowej wyświetlany jest krótki fragment tekstu, a w panelu Text Preview po prawej stronie wyświetlana jest większa część lub

całość tekstu z rekordu zaznaczonego w tabeli. Jeśli dane tekstowe mają postać odrębnych dokumentów, to w panelu Data wyświetlana jest nazwa pliku dokumentu. Po wybraniu dokumentu otwierany jest panel Text Preview z tekstem zaznaczonego dokumentu.

Kolory i wyróżnienia

W wyświetlanych danych pojęcia i deskryptory znalezione w dokumentach lub rekordach są wyróżniane kolorami, aby łatwiej było je odszukać w tekście. Kolory odpowiadają typom pojęć. Można także zatrzymać wskaźnik myszy nad elementem oznaczonym kolorem, aby wyświetlić nazwę pojęcia, pod którym dany termin został wyodrębniony, oraz typ, do którego został przypisany. Tekst niewyodrębniony jest wyświetlany kolorem czarnym. Te niewyodrębnione wyrazy są zwykle spójnikami (*and* lub *with*), zaimkami (*me* lub *they*) oraz czasownikami (*is*, *have* lub *take*).

Kolumny panelu Data

Kolumna zmiennej tekstowej jest widoczna zawsze, ale można też wyświetlić inne kolumny. Aby wyświetlić inne kolumny, wybierz z menu opcje **View > Data Pane**, a następnie wybierz kolumnę, którą chcesz wyświetlić w panelu Data. Do wyświetlania mogą być dostępne następujące kolumny:

- **"Text field name" (#)/Documents** Dodaje kolumnę z danymi tekstowymi, z których wyodrębnione były pojęcia i typy. Jeśli dane są zawarte w dokumentach, kolumna nosi nazwę Dokumenty i widoczna jest tylko nazwa dokumentu lub pełna ścieżka. Właściwy tekst z dokumentów jest widoczny w panelu Text Preview. Po nazwie tej kolumny podana jest w nawiasach liczba wierszy w panelu Data. Niekiedy nie będą wyświetlane wszystkie dokumenty lub rekordy. Wynika to z limitu ustawionego w oknie dialogowym Options, który wprowadza się w celu przyspieszenia ładowania. Po osiągnięciu limitu po liczbie pojawi się dopisek **- Max**. Więcej informacji zawiera temat “Okno dialogowe Options: karta Session” na stronie 75.
- **Categories** Zawiera listę wszystkich kategorii, do których należy rekord. Gdy ta kolumna jest widoczna, odświeżanie panelu Data w celu wyświetlenia najbardziej aktualnych informacji może trwać nieco dłużej.
- **Relevance Rank** Zawiera rangi poszczególnych rekordów w jednej kategorii. Ranga informuje o tym, jak dobrze dany rekord pasuje do kategorii na tle innych rekordów w tej samej kategorii. Aby wyświetlić rangę, wybierz kategorię w panelu Categories (lewym górnym). Więcej informacji zawiera temat “Istotność kategorii”.
- **Category Count** Zawiera liczbę kategorii, do których należy rekord.

Istotność kategorii

Aby budować lepsze kategorie, możesz przeglądać istotność dokumentów lub rekordów w każdej kategorii, a także istotność wszystkich kategorii, do których należy dokument lub rekord.

Istotność przypisania kategorii do rekordu

Za każdym razem, gdy dokument lub rekord pojawia się w panelu Data, wszystkie kategorie, do których należy, są wymienione w kolumnie Categories. Gdy dokument lub rekord należy do wielu kategorii, kategorie w tej kolumnie są wyświetlane w kolejności od najbardziej do najmniej istotnych dopasowań. Pierwsza kategoria na liście to z założenia ta, która najlepiej pasuje do tego dokumentu lub rekordu. Więcej informacji zawiera temat “Panel Data” na stronie 101.

Istotności przypisania rekordu do kategorii

Po wybraniu kategorii można sprawdzić istotność każdego z rekordów w kolumnie Relevance Rank w panelu Data. Ten ranking istotności określa, na ile dobrze dokument lub rekord pasuje do wybranej kategorii w porównaniu z innymi rekordami w tej kategorii. Aby wyświetlić ranking rekordów dla pojedynczej kategorii, należy wybrać tę kategorię w panelu Categories (lewy górny panel), a ranga dokumentu lub rekordu pojawi się w kolumnie. Ta kolumna domyślnie niewidoczna, ale można ją wyświetlić. Więcej informacji zawiera temat “Panel Data” na stronie 101.

Im niższa liczbowa ranga rekordu, tym lepsze dopasowanie lub większa istotność tego rekordu w wybranej kategorii, zatem 1 oznacza najlepsze dopasowanie. Jeśli więcej niż jeden rekord ma taką samą istotność, każda pojawia się z tą

samą rangą i znakiem równości (=), co oznacza, że istotności są równe. Na przykład mogą pojawić się rangi 1=, 1=, 3, 4 itd., co będzie oznaczać, że istnieją dwa rekordy uznane za najlepiej dopasowane do tej kategorii.

Wskazówka: Można dodać tekst z najbardziej istotnego rekordu do adnotacji kategorii jako jej opis. Tekst możesz dodać bezpośrednio w panelu Data, zaznaczając tekst i wybierając z menu opcję **Categories > Add to Annotation**.

Budowanie kategorii

Choć można korzystać z kategorii zawartych w pakiecie analizy tekstu, można również utworzyć kategorie automatycznie, korzystając z szeregu technik lingwistycznych i opartych na określaniu liczebności wystąpień. W oknie dialogowym Build Categories Settings można zastosować zautomatyzowane techniki lingwistyczne i oparte na określaniu liczebności, aby utworzyć kategorie na podstawie pojęć lub wzorców pojęć.

W ogólnym wypadku kategoria może składać się z różnych rodzajów deskryptorów (typów, pojęć, wzorców TLA, reguł kategorii). Podczas tworzenia kategorii za pomocą technik zautomatyzowanych wynikowe kategorie otrzymują nazwy na podstawie pojęć lub wzorców pojęć (w zależności od wybranych danych wejściowych), a każda kategoria zawiera zestaw deskryptorów. Te deskryptory mogą mieć postać reguł, kategorii lub pojęć i uwzględniać wszystkie pojęcia pokrewne wykryte za pomocą technik.

Po zbudowaniu kategorii można dowiedzieć się wiele na ich temat, przeglądając je w panelu Categories i eksplorując je na wykresach i w tabelach. Następnie można zastosować techniki ręczne, aby wprowadzić drobne korekty, usunąć wszelkie błędy w klasyfikacji lub dodać pominięte rekordy lub wyrazy. Po zastosowaniu wybranej techniki pojęcia, typy i wzorce pogrupowane w kategorię nadal mogą być wykorzystywane do tworzenia kategorii przy użyciu innych technik. Ponadto, ponieważ użycie kilku różnych technik może doprowadzić do powstania nadmiarowych lub niewłaściwych kategorii, można także scalić lub usunąć zbędne kategorie. Więcej informacji zawiera temat “Edytowanie i optymalizacja kategorii” na stronie 132.

Ważne! We wcześniejszych wersjach produktu reguły współwystępowania i synonimów były ujmowane w nawiasy kwadratowe. W tej wersji nawiasy kwadratowe oznaczają wzorzec wynikowy analizy powiązań w tekście. Natomiast reguły współwystępowania i synonimów są ujmowane w nawiasy okrągłe, na przykład (speaker systems|speakers).

Aby tworzyć kategorie

1. Z menu wybierz opcje **Categories > Build Categories**. O ile nie wyłączono wszystkich monitów, zostanie wyświetlony komunikat.
2. Wybierz, czy chcesz utworzyć kategorie teraz, czy najpierw edytować ustawienia.

- Kliknij przycisk **Build Now**, aby rozpocząć tworzenie kategorii przy użyciu bieżących ustawień. Ustawienia wybrane domyślnie są często wystarczające do rozpoczęcia procesu klasyfikowania. Proces tworzenia kategorii rozpocznie się i zostanie wyświetlone okno dialogowe postępu.
- Kliknij przycisk **Edit**, aby przejrzeć i zmodyfikować ustawienia budowania.

Uwaga: Maksymalna liczba kategorii, które mogą być wyświetlane, wynosi 10 000. Osiągnięcie lub przekroczenie tej liczby spowoduje wyświetlenie ostrzeżenia. W takiej sytuacji należy zmienić opcje tworzenia lub uzupełniania kategorii, aby ograniczyć liczbę tworzonych kategorii.

Wejścia

Kategorie są tworzone na podstawie deskryptorów wywiedzionych ze wzorców typów lub z typów. W tej tabeli można wybrać pojedyncze typy obiektów lub wzorce, które mają zostać uwzględnione w procesie tworzenia kategorii.

Type patterns. Jeśli wybierzesz wzorce typów, kategorie będą budowane na podstawie wzorców, a nie samych typów lub pojęć. Dzięki temu skategoryzowane zostaną rekordy lub dokumenty zawierające wzorzec pojęcia należący do wybranego wzorca typu. Zatem, jeśli w tabeli wybierzesz wzorzec typu <Budget> i <Positive>, to mogą zostać wygenerowane takie kategorie, jak cost & <Positive> lub rates & excellent.

W przypadku zastosowania wzorców typów jako danych wejściowych dla automatycznego budowania kategorii w niektórych przypadkach techniki mogą budować strukturę kategorii na różne, alternatywne sposoby. Z technicznego punktu widzenia nie ma jedynie słusznej struktury kategorii, jednak niektóre struktury kategorii mogą być lepiej niż inne dostosowane do konkretnych potrzeb analitycznych. Aby wpłynąć na uzyskiwane wyniki, można wyznaczyć preferowany typ. Wszystkie kategorie najwyższego poziomu tworzone będą na podstawie pojęcia należącego do wybranego tutaj typu (i żadnych innych typów). Każda podkategoria będzie zawierać wzorzec powiązań w tekście utworzony na podstawie tego typu. Gdy wybierzesz ten typ w polu **Structure categories by pattern type:**, tabela zostanie zaktualizowana, aby wyświetlić tylko wzorce mające zastosowanie, tj. zawierające wybrany typ. Zwykle wstępnie wybrany będzie typ <Unknown>. Skutkuje to wybraniem wszystkich wzorców zawierających typ <Unknown> (w przypadku tekstów w językach innych niż japoński). . Tabela przedstawia typy w porządku malejącym, rozpoczynając od typu z największą liczbą rekordów lub dokumentów (**Doc.** count).

Types. Gdy wybierzesz typy, kategorie zostaną zbudowane z pojęć należących do wybranych typów. Jeśli zatem w tabeli wybierzesz typ <Budget>, mogą zostać wygenerowane takie kategorie, jak cost lub price, ponieważ cost i price są pojęciami przypisanymi do typu <Budget>.

Domyślnie wybrane są tylko typy, które wychwycą większość rekordów lub dokumentów. Ten wstępny wybór pozwala szybko skupić się na najbardziej interesujących typach i uniknąć budowania kategorii nieinteresujących. Tabela przedstawia typy w porządku malejącym, rozpoczynając od typu z największą liczbą rekordów lub dokumentów (**Doc.** count). Typy z biblioteki Opinions domyślnie nie są wybrane w tabeli typów.

Wybrane ustawienia wejściowe mają wpływ na uzyskiwane kategorie. Jeśli wybierzesz opcję użycia typów jako danych wejściowych, w wynikach uwypuklone będą jednoznacznie pokrewne pojęcia. Na przykład przy budowaniu kategorii na podstawie typów mogłaby zostać utworzona kategoria Fruit obejmująca pojęcia apple, pear, citrus fruits, orange itd. Jeśli jednak wybierzesz opcję Type Patterns i wzorzec <Unknown> + <Positive>, to możesz uzyskać kategorię fruit + <Positive> z jednym lub dwoma rodzajami owoców, na przykład fruit + tasty i apple + good. Ten drugi wynik zawiera tylko 2 wzorce pojęć, ponieważ w innych wystąpieniach owoce nie zawsze są pozytywnie kwalifikowane. Taki rezultat może wystarczyć do analizy bieżących danych tekstowych, jednak w badaniach obejmujących różne zestawy dokumentów celowe może okazać się dodanie innych deskryptorów, takich jak citrus fruit + positive lub przypadków użycia. Zastosowanie typów jako jedynych kryteriów wejściowych pomoże wyszukać wszystkie możliwe owoce.

Techniki

Ponieważ każdy zbiór danych jest inny, liczba metod tworzenia kategorii i kolejność ich stosowania może się z czasem zmieniać. Jako że różne zbiory danych mogą być eksplorowane w różnych celach i pod różnym kątem, konieczne może być wypróbowanie różnych technik i wybranie tej, która przynosi najlepsze wyniki w analizie konkretnych danych tekstowych.

Nie trzeba być ekspertem, by korzystać z tych technik. Domyślnie wybrane są najczęściej stosowane i przeciętne ustawienia. Oznacza to, że możesz pominąć okna ustawień zaawansowanych i przejść bezpośrednio do budowania kategorii. Podobnie, jeśli wprowadzisz tutaj zmiany, nie musisz za każdym razem wracać do okna dialogowego ustawień, ponieważ zawsze zachowywane są najnowsze ustawienia.

Wybierz techniki lingwistyczne lub oparte na liczebności wystąpień i kliknij przycisk Advanced Settings, aby wyświetlić ustawienia dla wybranych technik. Żadna z technik automatycznych nie zapewni idealnej kategoryzacji danych; dlatego zalecamy wypróbowanie różnych technik automatycznych i stosowanie tych, które najlepiej sprawdzają się przy pracy z konkretnymi danymi.Nie można jednocześnie stosować technik lingwistycznych i opartych na liczebności.

- **Zaawansowane techniki lingwistyczne**. Aby uzyskać więcej informacji, patrz “Zaawansowane ustawienia językowe” na stronie 105.
- **Zaawansowane techniki oparte na liczebności wystąpień**. Aby uzyskać więcej informacji, patrz “Zaawansowane ustawienia liczebności” na stronie 111.

Zaawansowane ustawienia językowe

Podczas tworzenia kategorii można wybierać spośród wielu zaawansowanych technik lingwistycznych budowania kategorii, takich jak *wywodzenie rdzenia pojęcia*, (technika niedostępna dla języka japońskiego) , *włączanie pojęć*, *sieci semantyczne* (tylko dla tekstów w języku angielskim) i *reguły współwystępowania*. Techniki te mogą być używane indywidualnie lub łącznie do tworzenia kategorii.

Należy jednak pamiętać, że ponieważ każdy zbiór danych jest inny, liczba metod tworzenia kategorii i kolejność ich stosowania może się z czasem zmieniać. Jako że różne zbiory danych mogą być eksplorowane w różnych celach i pod różnym kątem, konieczne może być wypróbowanie różnych technik i wybranie tej, która przynosi najlepsze wyniki w analizie konkretnych danych tekstowych. Żadna z technik automatycznych nie zapewni idealnej kategoryzacji danych; dlatego zalecamy wypróbowanie różnych technik automatycznych i stosowanie tych, które najlepiej sprawdzają się przy pracy z konkretnymi danymi.

W oknie dialogowym Advanced Settings: Linguistics dostępne są następujące obszary i pola:

Dane wejściowe i wyjściowe

Category input Wybierz, na podstawie czego będą tworzone kategorie:

- **Unused extraction results**. Ta opcja powoduje budowanie kategorii z wyników wyodrębniania, które nie są używane w żadnych istniejących kategoriach. Minimalizuje to tendencję do dopasowywania tych samych rekordów do wielu kategorii i ogranicza liczbę generowanych kategorii.
- **All extraction results**. Ta opcja powoduje budowanie kategorii przy użyciu dowolnych wyników wyodrębniania. Taki sposób postępowania jest najbardziej użyteczny, gdy nie istnieją jeszcze kategorie lub jeśli istnieje niewiele kategorii.

Category output Wybierz ogólną strukturę kategorii, które zostaną utworzone:

- **Hierarchical with subcategories**. Ta opcja umożliwia tworzenie podkategorii i pod-podkategorii. Można ustawić głębokość hierarchii, wybierając maksymalną liczbę poziomów (pole **Maximum levels created**), które mogą być utworzone. Jeśli wybierzesz opcję 3, kategorie będą mogły zawierać podkategorie i te podkategorie także będą mogły mieć podkategorie.
- **Flat categories (single level only)**. Ta opcja dopuszcza zbudowanie tylko jednego poziomu kategorii, co oznacza, że nie będą generowane podkategorie.

Techniki grupowania

Każda z dostępnych technik dobrze nadaje się do pracy z określonymi rodzajami danych i w określonych warunkach, jednak często przydatna jest możliwość połączenia w jednej analizie różnych technik w celu wydobycia bogatszego zbioru informacji z dokumentów lub rekordów. Pojęcie może znaleźć się w więcej niż jednej kategorii, mogą też pojawić się kategorie nadmiarowe.

Concept Inclusion. Ta technika buduje kategorie, łącząc z w grupę z jednym pojęciem inne pojęcia złożone z wielu terminów (wielu wyrazów) w zależności od tego, czy zawierają one wyrazy będące podzbiorami czy nadzbiorami występującego w nim wyrazu. Na przykład pojęcie seat zostałoby połączone w grupę z pojęciami safety seat, seat belt i seat belt buckle. Więcej informacji zawiera temat “Włączanie pojęć” na stronie 108.

Semantic Network. Ta technika najpierw rozpoznaje możliwe sensy każdego pojęcia na podstawie obszernego indeksu relacji między wyrazami, a potem tworzy kategorie poprzez grupowanie pojęć pokrewnych. Sprawdza się najlepiej, gdy pojęcia są znane sieci semantycznej i nie są zbyt niejednoznaczne. Jest mniej użyteczna, gdy tekst zawiera terminologię specjalistyczną lub żargon nieznany sieci. Przykładowo pojęcie granny smith apple mogłoby zostać połączone w grupę z pojęciami gala apple i winesap apple, ponieważ podobnie jak „granny smith” oznaczają odmiany jabłoni. Natomiast pojęcie animal mogłoby zostać połączone w grupę z pojęciami cat i kangaroo, ponieważ są one hiponimami słowa animal (tj. zawężają jego znaczenie). W tej wersji technika ta jest dostępna tylko w odniesieniu do języka angielskiego. Więcej informacji zawiera temat “Sieci semantyczne” na stronie 109.

Uwaga: Opcja **Maximum search distance** jest dostępna tylko wtedy, gdy zaznaczono opcję **Semantic Network**.

Maximum search distance Wybierz, jak daleko ma być prowadzone wyszukiwanie, zanim wygenerowane zostaną kategorie. Im mniejsza wartość, tym mniej wyników zostanie wygenerowanych, jednak wyniki te będą mniej zaszumione i z większym prawdopodobieństwem będą istotnie powiązane ze sobą nawzajem. Im większa wartość, tym więcej wyników zostanie wygenerowanych, ale wyniki te mogą być mniej wiarygodne lub istotne. Choć opcja ta obowiązuje globalnie we wszystkich technikach, jej działanie jest najbardziej odczuwalne w sieciach semantycznych i analizie współwystąpień.

Prevent pairing of specific concepts. Zaznacz to pole wyboru, aby w wynikach nie grupować lub nie łączyć w pary dwóch pojęć. Aby tworzyć pary pojęć lub nimi zarządzać, kliknij przycisk **Manage Pairs...**. Więcej informacji zawiera temat "Zarządzanie parami wyjątków powiązań".

Generalize with wildcards where possible Wybierz tę opcję, aby umożliwić generowanie ogólnych reguł w kategoriach przy użyciu wieloznacznego symbolu gwiazdki. Na przykład, zamiast tworzyć wiele deskryptorów, takich jak [apple tart + .] i [apple sauce + .], program może wygenerować ogólny deskryptor [apple * + .]. Generalizacja przy użyciu znaków wieloznacznych prowadzi często do uzyskania dokładnie takiej samej liczby rekordów lub dokumentów, jak poprzednio. Jednak zaletą tej opcji jest zmniejszenie liczby i uproszczenie deskryptorów kategorii. Ponadto opcja ta umożliwia potencjalnie klasyfikowanie większej liczby nowych rekordów lub dokumentów za pomocą tych samych kategorii (na przykład w badaniach podłużnych).

Inne opcje budowania kategorii

Oprócz wybierania technik grupowania można zmodyfikować kilka innych opcji budowania:

Maximum number of top level categories created. Użyj tej opcji, aby ograniczyć liczbę kategorii, które mogą być wygenerowane po kliknięciu przycisku Build Categories. W niektórych przypadkach lepsze wyniki uzyskuje się, wpisując tutaj wysoką wartość, a potem usuwając nieinteresujące kategorie.

Minimum number of descriptors and/or subcategories per category. Użyj tej opcji, aby określić minimalną liczbę deskryptorów i podkategorii, jaką musi zawierać kategoria, aby została utworzona. Ta opcja pomaga ograniczyć liczbę kategorii, które nie wychwytują istotnej liczby rekordów lub dokumentów.

Allow descriptors to appear in more than one category Ta opcja dopuszcza użycie deskryptorów w więcej niż jednej kategorii. Opcja ta jest zwykle wybrana, ponieważ elementy często i w naturalny sposób pasują do dwóch lub więcej kategorii, dlatego dopuszczenie takiej możliwości zwykle prowadzi do uzyskania kategorii wyższej jakości. Jeśli ta opcja nie zostanie wybrana, może wystąpić mniejsze nakładanie się rekordów w różnych kategoriach, co niekiedy jest pożądane (w zależności od typu danych). Jednak w przypadku większości typów danych ograniczanie deskryptorów do pojedynczej kategorii zwykle powoduje utratę jakości kategorii lub mniejszego pokrycia danych kategoriami. Załóżmy na przykład, że mamy pojęcie car seat manufacturer. Gdy opisywana opcja jest wybrana, to pojęcie może występować w jednej kategorii na podstawie tekstu car seat, a w innej na podstawie manufacturer. Jeśli jednak ta opcja nie jest wybrana, mimo że nadal można uzyskać obie kategorie, pojęcie car seat manufacturer będzie ujęte tylko jako deskryptor w kategorii, do której pasuje najlepiej na podstawie kilku czynników, w tym liczby rekordów, w których występują terminy car seat i manufacturer (wystąpienia zliczane są osobno dla każdego terminu).

Resolve duplicate category names by Wybierz, w jaki sposób postępować z nowymi kategoriami lub podkategoriami, których nazwy byłyby takie same, jak nazwy istniejących kategorii. Można połączyć nowe kategorie (i ich deskryptory) z istniejącymi kategoriami o tej samej nazwie. Albo można pominąć tworzenie kategorii, jeśli zostanie znaleziona istniejąca kategoria o tej samej nazwie.

Zarządzanie parami wyjątków powiązań

Podczas tworzenia kategorii, grupowania i mapowania pojęć wewnętrzne algorytmy grupują wyrazy na podstawie znanych związków. Aby nie dopuścić do powiązania dwóch pojęć w parę, można włączyć odpowiednią funkcję w oknie dialogowym **Build Categories Advanced Settings**, **Build Clusters** i **Concept Map Index Settings**, a następnie kliknąć przycisk **Manage Pairs**.

W wyświetlonym oknie dialogowym **Manage Link Exceptions** można dodawać, edytować lub usuwać pary pojęć. Wprowadź jedną parę na wiersz. Wprowadzenie pary w tym miejscu uniemożliwi utworzenie takiej pary podczas budowania lub uzupełniania kategorii, grupowania i mapowania pojęć. Wprowadź wyrazy dokładnie w takiej postaci, w jakiej mają być uwzględniane, na przykład wyraz z akcentami nie jest równy wersji bez akcentów.

Na przykład, jeśli chcesz mieć pewność, że hot dog i dog nie zostaną zgrupowane, dodaj tę parę w osobnym wierszu tabeli.

Informacje o technikach lingwistycznych

Podczas tworzenia lub uzupełniania kategorii można wybierać spośród wielu zaawansowanych technik lingwistycznych budowania kategorii, takich jak *wywodzenie rdzenia pojęcia*, (technika niedostępna dla języka japońskiego), *włączanie pojęć*, *sieci semantyczne* (tylko dla tekstów w języku angielskim) i *reguły współwystępowania*. Techniki te mogą być używane indywidualnie lub łącznie do tworzenia kategorii.

Nie trzeba być ekspertem, by korzystać z tych technik. Domyślnie wybrane są najczęściej stosowane i przeciętne ustawienia. Jeśli chcesz, możesz pominąć to okno dialogowe zaawansowanych ustawień i przejść bezpośrednio do tworzenia lub uzupełniania własnych kategorii. Podobnie, jeśli wprowadzisz tutaj zmiany, nie musisz za każdym razem wracać do okna dialogowego ustawień, ponieważ zawsze zachowywane są najnowsze ustawienia.

Należy jednak pamiętać, że ponieważ każdy zbiór danych jest inny, liczba metod i kolejność ich stosowania może się z czasem zmieniać. Jako że różne zbiory danych mogą być eksplorowane w różnych celach i pod różnym kątem, konieczne może być wypróbowanie różnych technik i wybranie tej, która przynosi najlepsze wyniki w analizie konkretnych danych tekstowych. Żadna z technik automatycznych nie zapewni idealnej kategoryzacji danych; dlatego zalecamy wypróbowanie różnych technik automatycznych i stosowanie tych, które najlepiej sprawdzają się przy pracy z konkretnymi danymi.

Oto główne techniki lingwistyczne automatycznego budowania kategorii:

- **Concept root derivation.** Te technika tworzy kategorie, wybierając jedno pojęcie i wyszukując pojęcia mu pokrewne poprzez analizę relacji morfologicznych lub wspólnych rdzeni między komponentami tych pojęć. Więcej informacji zawiera temat “Wywodzenie rdzeni pojęć”. Ta opcja nie jest dostępna w przypadku tekstu japońskiego.
- **Concept inclusion.** Te technika tworzy kategorie, wybierając jedno pojęcie i wyszukując pojęcia, w których jest ono zawarte. Więcej informacji zawiera temat “Włączanie pojęć” na stronie 108.
- **Semantic network.** Ta technika najpierw rozpoznaje możliwe sensy każdego pojęcia na podstawie obszernego indeksu relacji między wyrazami, a potem tworzy kategorie poprzez grupowanie pojęć pokrewnych. Więcej informacji zawiera temat “Sieci semantyczne” na stronie 109. Ta opcja jest dostępna tylko w przypadku tekstu w języku angielskim.
- **Co-occurrence.** Ta metoda tworzy reguły współwystępowania, których można użyć do utworzenia nowej kategorii, uzupełnienia kategorii lub jako kryterium wejściowego dla innej techniki kategoryzacji. Więcej informacji zawiera temat “Reguły współwystępowania” na stronie 110.

Wywodzenie rdzeni pojęć

Uwaga: Ta technika nie jest dostępna dla tekstu w języku japońskim.

Technika wywodzenia rdzenia pojęcia tworzy kategorie, wybierając jedno pojęcie i wyszukując pojęcia mu pokrewne poprzez analizę relacji morfologicznych między komponentami tych pojęć. Komponent jest wyrazem. Technika ta próbuje grupować pojęcia, analizując końcówki (przyrostki) każdego komponentu w pojęciu i znajdując inne pojęcia, które mogą być z tych komponentów wywiedzione. Zakłada się bowiem, gdy słowa pochodzą od siebie, to prawdopodobnie są bliskie znaczeniowo. W celu identyfikacji końcówek używane są reguły wewnętrzne właściwe dla danego języka. Na przykład pojęcie opportunities to advance zostałoby połączone w grupę z pojęciami opportunity for advancement i advancement opportunity.

Wywodzenie rdzeni pojęć można zastosować do dowolnego tekstu. Samodzielnie technika ta tworzy względnie mało kategorii, a każda kategoria zawiera zwykle mało pojęć. Pojęcia w każdej kategorii są synonimami lub pojęciami

pokrewnymi sytuacyjnie. Algorytm ten może być pomocny nawet w przypadku ręcznego budowania kategorii; synonimy znalezione przez algorytm mogą być synonimami właśnie tych pojęć, którymi użytkownik jest szczególnie zainteresowany.

Uwaga: Można zapobiec grupowaniu pojęć, określając je jawnie. Więcej informacji zawiera temat “Zarządzanie parami wyjątków powiązań” na stronie 106.

Komponentyzacja i usuwanie odmiany terminów

Podczas wywodzenia rdzeni pojęć terminy w pierwszej kolejności są dzielone na komponenty (wyrazy), a następnie komponenty te są pozbawiane odmiany. Podczas stosowania tej techniki pojęcia i powiązane z nimi terminy są ładowane i dzielone na komponenty na podstawie separatorów, takich jak spacje, łączniki i apostrofy. Na przykład termin system administrator jest dzielony na składniki {administrator, system}.

Jednak niektóre części oryginalnego terminu mogą nie zostać wykorzystane. Są to tzw. wyrazy ignorowane. W języku angielskim do komponentów ignorowanych mogą należeć wyrazy: a, and, as, by, for, from, in, of, on, or, the, to i with.

Na przykład termin examination of the data złożony jest z komponentów {data, examination}, ponieważ wyrazy of i the są ignorowane. Ponadto zbiór komponentów jest nieuporządkowany. W ten sposób następujące trzy składniki mogą być równoważne: cough relief for child, child relief from a cough i relief of child cough, ponieważ wszystkie mają ten sam zbiór komponentów {child, cough, relief}. Gdy para terminów zostanie zidentyfikowana jako równoważna, odpowiednie pojęcia są łączone w celu utworzenia nowego pojęcia, które odwołuje się do wszystkich terminów.

Ponadto, ponieważ komponenty składnika mogą być odmieniane automatycznie, wewnętrznie stosowane reguł języka w celu zidentyfikowania terminów równoważnych niezależnie od wariantów odmiany, np. liczby mnogiej. W ten sposób terminy level of support i support levels mogą być identyfikowane jako równoważne, ponieważ pojedyncza forma rzeczownika po usunięciu odmiany brzmi level.

Działanie techniki wywodzenia rdzenia pojęcia

Po komponentyzacji terminów i pozbawieniu jej odmiany (patrz poprzednia sekcja) algorytm wywodzenia rdzeni pojęć analizuje końcówki (przyrostki) komponentów, aby znaleźć rdzeń komponentu, a następnie grupuje pojęcia z innymi pojęciami, które mają takie same lub podobne rdzenie. Końcówki są identyfikowane przy użyciu zestawu reguł lingwistycznych charakterystycznych dla języka tekstu. Na przykład w języku angielskim istnieje reguła mówiąca, że termin pojęcia kończący się przyrostkiem ical może być wywiedziony z pojęcia o tym samym rdzeniu i końcówce ic. Przy użyciu tej reguły (po usunięciu odmiany) algorytm będzie mógł zgrupować pojęcia epidemiologic study i epidemiological studies.

Ponieważ terminy są już podzielone na komponenty, a komponenty do pominięcia (na przykład in i of) zostały już zidentyfikowane, algorytm wywodzenia rdzeni pojęć potrafi także zgrupować pojęcia studies in epidemiology i epidemiological studies.

Zestaw reguł wywodzenia rdzeni pojęć został wybrany w taki sposób, że większość pojęć grupowanych według tego algorytmu stanowi synonimy: pojęcia epidemiologic studies, epidemiological studies i studies in epidemiology są terminami równoważnymi. Aby zwiększyć kompletność wyników wprowadzono pewne reguły wywodzenia, które umożliwiają algorytmowi grupowanie pojęć pokrewnych sytuacyjnie. Na przykład algorytm może zgrupować pojęcia empire builder i empire building.

Włączanie pojęć

Technika włączenia pojęć buduje kategorie, identyfikując pojęcia zawarte w innych pojęciach za pomocą algorytmów szeregu leksykalnego. Zakłada się, że jeśli wyrazy w pojęciu są podzbiorem innego pojęcia, to między tymi pojęciami istnieje relacja semantyczna. Włączanie jest zaawansowaną techniką, która może być używana z dowolnym typem tekstu.

Ta technika działa dobrze w połączeniu z sieciami semantycznymi, ale mogą być też używane osobno. Włączanie pojęć może również przynieść lepsze wyniki, gdy dokumenty lub rekordy zawierają wiele terminów branżowych lub żargonowych. Zwłaszcza, jeśli słowniki zostały wcześniej zoptymalizowane tak, by specjalistyczne terminy były wyodrębniane i odpowiednio grupowane (z synonimami).

Działanie techniki włączania pojęć

Zanim algorytm włączania pojęć zostanie zastosowany, składniki są komponentyzowane i pozbawiane odmiany. Więcej informacji zawiera temat “Wywodzenie rdzeni pojęć” na stronie 107. Następny algorytm włączania pojęć analizuje zbiory komponentów. Dla każdego zbioru komponentów algorytm wyszukuje kolejny zbiór komponentów, który jest podzbiorem pierwszego zbioru.

Na przykład, jeśli masz pojęcie continental breakfast ze zbiorem komponentów {breakfast, continental}, a także pojęcie breakfast ze zbiorem komponentów {breakfast}, algorytm uzna continental breakfast za jeden z rodzajów pojęcia breakfast i zgrupuje oba pojęcia.

A oto bardziej rozbudowany przykład: jeśli na panelu Extraction Results masz pojęcie seat i zastosujesz opisywany algorytm, to do tej samej kategorii trafią także pojęcia safety seat, leather seat, seat belt, seat belt buckle, infant seat carrier i car seat laws.

Ponieważ terminy są już podzielone na komponenty, a komponenty do pominięcia (na przykład in i of) zostały już zidentyfikowane, algorytm włączania pojęć potrafi zorientować się, że advanced spanish course zawiera pojęcie course in spanish.

Uwaga: Można zapobiec grupowaniu pojęć, określając je jawnie. Więcej informacji zawiera temat “Zarządzanie parami wyjątków powiązań” na stronie 106.

Sieci semantyczne

W tej wersji produktu technika sieci semantycznych jest dostępna tylko w przypadku tekstów w języku angielskim.

Ta metoda tworzy kategorie przy użyciu wbudowanej sieci relacji między wyrazami. Z tego powodu ta technika może generować bardzo dobre wyniki, gdy terminy są konkretne i nie są zbyt niejednoznaczne. Jednak nie należy oczekiwać, że technika ta znajdzie wiele powiązań między wysoce technicznymi/specjalistycznymi pojęciami. W przypadku takich pojęć może okazać się, że techniki włączania pojęć i wywodzenia rdzeni pojęć będą bardziej użyteczne.

Działanie techniki sieci semantycznych

Istotą techniki sieci semantycznych jest wykorzystanie znanych relacji między wyrazami do tworzenia kategorii synonimów lub hiponimów. Z **hiponimem** mamy do czynienia, gdy jedno pojęcie jest rodzajem drugiego pojęcia, tak że istnieje relacja hierarchiczna, znana również jako relacja ISA. Na przykład, jeśli animal jest pojęciem, to cat i kangaroo są hiponimami animal, ponieważ są rodzajami zwierząt.

Oprócz synonimów i hiponimów technika sieci semantycznej analizuje również część powiązań i całe powiązania między pojęciami typu <Location>. Na przykład technika ta zgrupuje pojęcia normandy, provence i france w jednej kategorii, ponieważ Normandia i Prowansja są regionami Francji.

Działanie techniki sieci semantycznych rozpoczyna się od identyfikacji możliwych znaczeń poszczególnych terminów w sieci semantycznej. Gdy pojęcia zostaną zidentyfikowane jako synonimy lub hiponimy, są grupowane w jedną kategorię. Na przykład opisywana technika utworzyłaby jedną kategorię zawierającą trzy pojęcia: eating apple, dessert apple i granny smith, ponieważ sieć semantyczna zawiera informacje o tym, że: 1) dessert apple jest synonimem eating apple, 2) granny smith jest rodzajem eating apple (zatem jest hiponimem eating apple).

Wiele pojęć, zwłaszcza jednowyrazowych, ma charakter niejednoznaczny, jeśli rozpatruje się je osobno. Na przykład pojęcie buffet może oznaczać rodzaj posiłku lub mebel. Jeśli zbiór pojęć zawiera pojęcia meal, furniture i buffet, to

algorytm musi wybrać pomiędzy zgrupowaniem pojęcia buffet z pojęciem meal albo furniture. Należy pamiętać, że w niektórych przypadkach wybory dokonane przez algorytm mogą nie być odpowiednie w kontekście konkretnego zbioru rekordów lub dokumentów.

W przypadku niektórych typów danych technika sieci semantycznych działa lepiej niż włączanie pojęć. Obie te techniki rozpoznają, że apple pie jest rodzajem pie, ale tylko technika sieci semantycznej rozpozna, że tart również jest rodzajem pie.

Sieci semantyczne będą działać w połączeniu z innymi technikami. Załóżmy na przykład, że wybrano zarówno technikę sieci semantycznych, jak i włączania pojęć, a sieć semantyczna zgrupowała pojęcie teacher z pojęciem tutor (ponieważ tutor jest rodzajem teacher). Algorytm włączania może zgrupować pojęcie graduate tutor z pojęciem tutor, a w rezultacie współpracy dwóch algorytmów powstanie kategoria wynikowa zawierająca wszystkie trzy pojęcia: tutor, graduate tutor i teacher.

Opcje techniki sieci semantycznych

Istnieje szereg dodatkowych ustawień, które mogą być interesujące dla użytkownika tej techniki.

- Zmień wartość **Maximum search distance**. Wybierz, jak daleko ma być prowadzone wyszukiwanie, zanim wygenerowane zostaną kategorie. Im mniejsza wartość, tym mniej wyników zostanie wygenerowanych, jednak wyniki te będą mniej zaszumione i z większym prawdopodobieństwem będą istotnie powiązane ze sobą nawzajem. Im większa wartość, tym więcej wyników zostanie wygenerowanych, ale wyniki te mogą być mniej wiarygodne lub istotne.

Na przykład, w zależności od odległości, algorytm algorytm przeszukuje pojęcia począwszy od Danish pastry do coffee roll (pojęcie nadrzędne), potem bun (pojęcie nadrzędne pojęcia nadrzędnego), a potem bread.

Zmniejszenie odległości wyszukiwania umożliwia tworzenie mniejszych kategorii, które mogą być łatwiejsze w praktycznym zastosowaniu, a także bywa przydatne, gdy generowane kategorie są zbyt obszerne lub grupują zbyt wiele pojęć.

Ważne! Dodatkowo zaleca się, aby w przypadku stosowania tej techniki nie używać opcji **Accommodate spelling errors for a minimum root character limit of** (zdefiniowanej na karcie Expert węzła lub w oknie dialogowym Extract) dla grupowania rozmytego, ponieważ niektóre grupy mogą mieć bardzo niekorzystny wpływ na wyniki.

Reguły współwystępowania

Reguły współwystępowania umożliwiają wykrywanie i grupowanie pojęć, które są silnie powiązane w obrębie zbioru dokumentów lub rekordów. Koncepcja działania tej techniki zasadza się na tym, że gdy pojęcia często występują razem w dokumentach i rekordach, takie współwystąpienia odzwierciedlają potencjalną relację, którą warto uwzględnić w definicjach kategorii. Ta metoda tworzy reguły współwystępowania, których można użyć do utworzenia nowej kategorii, uzupełnienia kategorii lub jako kryterium wejściowego dla innej techniki kategoryzacji. Dwa pojęcia silnie współwystępują, jeśli często występują razem w pewnym zbiorze rekordów i rzadko występują oddzielnie w pozostałych rekordach. Ta metoda może przynieść dobre wyniki w przypadku dużych zbiorów danych obejmujących co najmniej kilkaset dokumentów lub rekordów.

Na przykład, jeśli wiele rekordów zawiera wyrazy price i availability, to pojęcia te mogą być zgrupowane w regułę współwystępowania (price & available). Inny przykład: jeśli pojęcia peanut butter, jelly, sandwich występują częściej razem niż osobno, to zostaną zgrupowane w jedną regułę współwystępowania (peanut butter & jelly & sandwich).

Ważne! We wcześniejszych wersjach produktu reguły współwystępowania i synonimów były ujmowane w nawiasy kwadratowe. W tej wersji nawiasy kwadratowe oznaczają wzorzec wynikowy analizy powiązań w tekście. Natomiast reguły współwystępowania i synonimów są ujmowane w nawiasy okrągłe, na przykład (speaker systems|speakers).

Działanie reguł współwystępowania

Ta technika przegląda dokumenty lub rekordy w poszukiwaniu dwóch lub większej liczby pojęć, które mają tendencje do łącznego występowania. Dwa pojęcia silnie współwystępują, jeśli często występują razem w pewnym zbiorze dokumentów lub rekordów i rzadko występują oddzielnie w pozostałych dokumentach lub rekordach.

Gdy znalezione zostaną pojęcia współwystępujące, tworzona jest reguła kategorii. Reguły takie składają się z dwóch lub większej liczby pojęć połączonych za pomocą operatora boolowskiego &. Reguły są instrukcjami logicznymi, które będą automatycznie klasyfikować dokument lub rekord do kategorii, jeśli zbiór pojęć w regule w całości współwystępuje w tym dokumencie lub rekordzie.

Opcje reguł współwystępowania

Jeśli używana jest technika reguł współwystępowania, można zoptymalizować kilka ustawień, które mają wpływ na końcowe reguły:

- Zmień wartość **Maximum search distance**. Wybierz, jak daleko technika ma wyszukać współwystąpień. W przypadku zwiększenia odległości wyszukiwania zmniejsza się minimalna wartość podobieństwa wymagana do stwierdzenia obecności współwystąpienia. W rezultacie może powstać wiele reguł współwystępowania, ale reguły o niskiej wartości podobieństwa będą często mało istotne. W przypadku zmniejszenia odległości wyszukiwania zwiększa się minimalna wymagana wartość podobieństwa; w rezultacie powstanie mniej reguł, ale o potencjalnie większej istotności.
- **Minimum number of documents.** Minimalna liczba rekordów lub dokumentów , które muszą zawierać daną parę pojęć, aby była ona traktowana jako współwystąpienie; im mniejsza jest ta wartość, tym łatwiej znaleźć współwystąpienia. Zwiększenie tej wartości powoduje wygenerowanie mniejszej liczby współwystąpień, które będą jednak bardziej istotne. Na przykład załóżmy, że pojęcia "apple" "pear" występują razem w 2 rekordach (a żadne z tych pojęć nie występuje w żadnych innych rekordach). Opcja **Minimum number of documents.** jest ustawiona na 2 (wartość domyślna), technika współwystąpień utworzy regułę kategorii (apple and pear). Jeśli wartość zostanie zwiększona do 3, reguła nie zostanie utworzona.

Uwaga: W przypadku małych zbiorów danych (< 1000 odpowiedzi) znalezienie jakichkolwiek współwystąpień z ustawieniami domyślnymi może być niemożliwe. W takim przypadku spróbuj zwiększyć odległość wyszukiwania.

Uwaga: Można zapobiec grupowaniu pojęć, określając je jawnie. Więcej informacji zawiera temat “Zarządzanie parami wyjątków powiązań” na stronie 106.

Zaawansowane ustawienia liczebności

Istnieje możliwość utworzenia kategorii na podstawie prostej i mechanicznej techniki zliczania wystąpień. Za pomocą tej techniki można zbudować jedną kategorię dla każdego elementu (typu, pojęcia lub wzorca), który występuje co najmniej w określonej liczbie rekordów lub dokumentów. Dodatkowo można utworzyć pojedynczą kategorię dla wszystkich elementów o mniejszej liczebności. Pod pojęciem liczebności rozumiemy liczbę rekordów lub dokumentów zawierających wyodrębnione pojęcie (lub jego synonimy), typ lub wzorzec, a nie łączną liczbę wystąpień w całym tekście.

Grupowanie często używanych elementów może przynieść interesujące wyniki, ponieważ często ujawnia typowe lub istotne odpowiedzi. Technika ta jest bardzo przydatna do analizy niewykorzystanych wyników wyodrębniania już po zastosowaniu innych technik. Inne zastosowanie polega na użyciu tej techniki od razu po wyodrębnieniu, gdy nie istnieją jeszcze inne kategorie, ręcznym usunięciu z wyników nieinteresujących kategorii, a następnie uzupełnieniu pozostałych kategorii, aby pasowały do jeszcze większej liczby rekordów lub dokumentów. Więcej informacji zawiera temat “Uzupełnianie kategorii” na stronie 112.

Zamiast używać tej techniki, można posortować pojęcia lub wzorce pojęć malejąco według liczby rekordów lub dokumentów w panelu Extraction Results, a następnie przeciągnąć i upuścić pozycje z początku uzyskanej listy do panelu Categories w celu utworzenia odpowiednich kategorii.

W oknie dialogowym Advanced Settings: Frequencies dostępne są następujące pola

Generate category descriptors at. Wybierz rodzaj kryteriów wejściowych dla deskryptorów. Więcej informacji zawiera temat “Budowanie kategorii” na stronie 103.

- **Concepts level**. Wybranie tej opcji oznacza, że używane będą liczebności pojęć lub wzorców pojęć. Pojęcia będą używane, jeśli jako kryteria wejściowe dla tworzenia kategorii wybrano typy, a wzorce pojęć będą używane, jeśli wybrano wzorce typów. Z reguły zastosowanie tej techniki na poziomie pojęć spowoduje wygenerowanie bardziej szczegółowych wyników, ponieważ pojęcia i wzorce pojęć reprezentują niższy poziom pomiaru.
- **Types level**. Wybranie tej opcji oznacza, że używane będą liczebności typów lub wzorców typów. Typy będą używane, jeśli jako kryteria wejściowe dla tworzenia kategorii wybrano typy, a wzorce typów będą używane, jeśli wybrano wzorce typów. Stosowanie tej techniki na poziomie typów pozwala uzyskać szybki przegląd rodzajów informacji obecnych w danych.

Minimum doc. count for items to have their own category. Ta opcja służy do tworzenia kategorii z często występujących pozycji. Ta opcja ogranicza wyniki do tych kategorii zawierających deskryptor, które wystąpiły w co najmniej X rekordów lub dokumentów, gdzie X jest wartością wprowadzoną w tej opcji.

Group all remaining items into a category called. Ta opcja umożliwia grupowanie wszystkich rzadko występujących pojęć i typów w pojedynczą kategorię „zbiorczą” o nazwie wybranej przez użytkownika. Domyślnie ta kategoria ma nazwę *Other*.

Category input. Wybierz grupę, do której chcesz zastosować techniki:

- **Unused extraction results**. Ta opcja powoduje budowanie kategorii z wyników wyodrębniania, które nie są używane w żadnych istniejących kategoriach. Minimalizuje to tendencję do dopasowywania tych samych rekordów do wielu kategorii i ogranicza liczbę generowanych kategorii.
- **All extraction results**. Ta opcja powoduje budowanie kategorii przy użyciu dowolnych wyników wyodrębniania. Taki sposób postępowania jest najbardziej użyteczny, gdy nie istnieją jeszcze kategorie lub jeśli istnieje niewiele kategorii.

Resolve duplicate category names by. Wybierz, w jaki sposób postępować z nowymi kategoriami lub podkategoriami, których nazwy byłyby takie same, jak nazwy istniejących kategorii. Można połączyć nowe kategorie (i ich deskryptory) z istniejącymi kategoriami o tej samej nazwie. Albo można pominąć tworzenie kategorii, jeśli zostanie znaleziona istniejąca kategoria o tej samej nazwie.

Uzupełnianie kategorii

Uzupełnianie to proces polegający na automatycznym dodawaniu i doskonaleniu deskryptorów w celu „powiększenia” istniejących kategorii. Celem jest uzyskanie lepszych kategorii wychwytujących pokrewne rekordy lub dokumenty, które nie były początkowo do nich przypisane.

Wybrane techniki automatycznego grupowania próbują zidentyfikować pojęcia, wzorce TLA i reguły kategorii związane z istniejącymi deskryptorami kategorii. Te nowe pojęcia, wzorce i reguły kategorii są następnie dodawane jako nowe deskryptory lub dodawane do istniejących deskryptorów. Techniki grupowania używane do uzupełniania kategorii to: *wywodzenie rdzenia pojęcia* (technika niedostępna dla języka japońskiego), *włączanie pojęcia*, *sieci semantyczne* (tylko dla języka angielskiego) i *reguły współwystępowania*. Metoda **Extend empty categories with descriptors generated from the category name** generuje deskryptory na podstawie wyrazów występujących w nazwach kategorii, dlatego im bardziej opisowe nazwy kategorii, tym lepsze wyniki.

Uwaga: Techniki oparte na liczebności nie mogą być stosowane do uzupełniania kategorii.

Uzupełnianie jest doskonałym sposobem interaktywnej optymalizacji kategorii. Poniżej znajdują się przykłady sytuacji, w których można potencjalnie uzupełnić kategorię:

- Po zakończeniu przeciągania/upuszczania wzorców pojęć w celu utworzenia kategorii w panelu Categories.
- Po ręcznym utworzeniu kategorii i dodaniu prostych reguł i deskryptorów.
- Po zaimportowaniu pliku predefiniowanych kategorii z bardzo opisowymi nazwami kategorii.
- Po zoptymalizowaniu kategorii pochodzących z pakietu TAP, jeśli wybrano opcję .

Tę samą kategorię można uzupełniać wielokrotnie. Na przykład, jeśli zaimportowano plik predefiniowanych kategorii z bardzo opisowymi nazwami, można najpierw uzupełnić kategorie za pomocą opcji **Extend empty categories with descriptors generated from the category name**, aby uzyskać pierwszy zestaw deskryptorów, a następnie ponownie uzupełnić te kategorie. Jednak w innych przypadkach wielokrotne uzupełnianie może skutkować powstaniem zbyt ogólnej kategorii, jeśli deskryptory będą z każdym uzupełnieniem coraz szersze i szersze. Ponieważ w technikach budowania i uzupełniania stosowane są podobne algorytmy, uzupełnienie kategorii bezpośrednio po ich zbudowaniu prawdopodobnie nie dostarczy bardziej interesujących wyników.

Wskazówki:

- Jeśli po próbie uzupełnienia kategorii okaże się, że wyniki są niezadowalające, zawsze można cofnąć operację (**Edit > Undo**) od razu po zakończeniu uzupełniania.
- W procesie uzupełniania w jednej kategorii mogą powstać dwie lub większa liczba reguł pasujących dokładnie do tego samego zestawu dokumentów, ponieważ reguły są tworzone niezależnie od siebie. W razie potrzeby można przejrzeć kategorie i usunąć nadmiarowość poprzez ręczną edycję opisu kategorii. Więcej informacji zawiera temat “Edytowanie deskryptorów kategorii” na stronie 133.

Aby uzupełnić kategorie

1. W panelu Categories wybierz kategorie, które chcesz uzupełnić.
2. Z menu wybierz opcje **Categories > Extend Categories**. O ile nie wyłączono wszystkich monitów, zostanie wyświetlony komunikat.
3. Wybierz, czy chcesz utworzyć kategorie teraz, czy najpierw edytować ustawienia.

- Kliknij przycisk **Extend Now**, aby rozpocząć uzupełnianie kategorii przy użyciu bieżących ustawień. Proces rozpocznie się, a następnie zostanie wyświetlone okno dialogowe postępu.
- Kliknij przycisk **Edit**, aby przejrzeć i zmienić ustawienia.

Po próbie uzupełnienia wszystkie kategorie, dla których znaleziono nowe deskryptory, są oznaczone przez słowo **Extended** w panelu Categories, dzięki czemu można je szybko zidentyfikować. Słowo Extended pozostaje widoczne do chwili, aż ponownie uzupełnisz kategorię, zmienisz ją w inny sposób lub skasujesz to oznaczenie za pomocą menu kontekstowego.

Uwaga: Maksymalna liczba kategorii, które mogą być wyświetlane, wynosi 10 000. Osiągnięcie lub przekroczenie tej liczby spowoduje wyświetlenie ostrzeżenia. W takiej sytuacji należy zmienić opcje tworzenia lub uzupełniania kategorii, aby ograniczyć liczbę tworzonych kategorii.

Każda z dostępnych technik dobrze nadaje się do pracy z określonymi rodzajami danych i w określonych warunkach, jednak często przydatna jest możliwość połączenia w jednej analizie różnych technik w celu wydobycia bogatszego zbioru informacji z dokumentów lub rekordów. W interaktywnym pulpicie roboczym pojęcia i typy pogrupowane w kategorię nadal mogą być wykorzystywane do tworzenia kategorii. Pojęcie może znaleźć się w więcej niż jednej kategorii, mogą też pojawić się kategorie nadmiarowe.

W oknie dialogowym Extend Categories: Settings dostępne są następujące obszary i pola:

Zastępowanie średnią. Wybierz kryterium wejściowe, które będzie używane do uzupełniania kategorii:

- **Unused extraction results**. Ta opcja powoduje budowanie kategorii z wyników wyodrębniania, które nie są używane w żadnych istniejących kategoriach. Minimalizuje to tendencję do dopasowywania tych samych rekordów do wielu kategorii i ogranicza liczbę generowanych kategorii.
- **All extraction results**. Ta opcja powoduje budowanie kategorii przy użyciu dowolnych wyników wyodrębniania. Taki sposób postępowania jest najbardziej użyteczny, gdy nie istnieją jeszcze kategorie lub jeśli istnieje niewiele kategorii.

Techniki grupowania

Krótki opis każdej z tych technik zawiera temat “Zaawansowane ustawienia językowe” na stronie 105. Do tych technik należą:

- **Concept root derivation** *(opcja niedostępna w przypadku tekstu japońskiego)*
- **Semantic network** (*tylko dla tekstu angielskiego i nieużywana, jeśli wybrana jest opcja Generalize only.*)
- **Concept inclusion**
- **Co-occurrence** i podopcja **Minimum number of docs**

Niektóre typy są trwale wyłączone z techniki sieci semantycznych, ponieważ te typy nie będą generować istotnych wyników. Do typów tych należą: <Positive>, <Negative>, <IP>, inne typy nielingwistyczne itd.

Maximum search distance Wybierz, jak daleko ma być prowadzone wyszukiwanie, zanim wygenerowane zostaną kategorie. Im mniejsza wartość, tym mniej wyników zostanie wygenerowanych, jednak wyniki te będą mniej zaszumione i z większym prawdopodobieństwem będą istotnie powiązane ze sobą nawzajem. Im większa wartość, tym więcej wyników zostanie wygenerowanych, ale wyniki te mogą być mniej wiarygodne lub istotne. Choć opcja ta obowiązuje globalnie we wszystkich technikach, jej działanie jest najbardziej odczuwalne w sieciach semantycznych i analizie współwystąpień.

Prevent pairing of specific concepts. Zaznacz to pole wyboru, aby w wynikach nie grupować lub nie łączyć w pary dwóch pojęć. Aby tworzyć pary pojęć lub nimi zarządzać, kliknij przycisk **Manage Pairs..**. Więcej informacji zawiera temat “Zarządzanie parami wyjątków powiązań” na stronie 106.

Tam, gdzie to możliwe: Wybierz, czy tylko uzupełnić, czy uogólnić deskryptory za pomocą znaków wieloznacznych, czy jedno i drugie.

- **Extend and generalize.** Ta opcja spowoduje uzupełnienie wybranych kategorii, a następnie uogólnienie deskryptorów. Jeśli wybierzesz opcję uogólnienia, produkt utworzy ogólne reguły kategorii w kategoriach przy użyciu wieloznacznego symbolu gwiazdki. Na przykład, zamiast tworzyć wiele deskryptorów, takich jak [apple tart + .] i [apple sauce + .], program może wygenerować ogólny deskryptor [apple * + .]. Generalizacja przy użyciu znaków wieloznacznych prowadzi często do uzyskania dokładnie takiej samej liczby rekordów lub dokumentów, jak poprzednio. Jednak zaletą tej opcji jest zmniejszenie liczby i uproszczenie deskryptorów kategorii. Ponadto opcja ta umożliwia potencjalnie klasyfikowanie większej liczby nowych rekordów lub dokumentów za pomocą tych samych kategorii (na przykład w badaniach podłużnych).
- **Extend only**. Ta opcja spowoduje uzupełnienie kategorii bez uogólniania. Celowe może być wybranie najpierw opcji **Extend only** dla kategorii utworzonych ręcznie, a potem ponownie uzupełnienie tych samych kategorii przy użyciu opcji **Extend and generalize**.
- **Generalize only.** Ta opcja spowoduje uogólnienie deskryptorów bez uzupełniania kategorii w inny sposób.

  Uwaga: Wybranie tej opcji powoduje wyłączenie opcji **Semantic network**, ponieważ opcja **Semantic network** jest dostępna tylko w przypadku, gdy opis ma zostać uzupełniony.

Inne opcje uzupełniania kategorii

Oprócz wybierania technik do zastosowania można zmienić dowolne z następujących opcji:

Maximum number of items to extend a descriptor by. W przypadku uzupełniania deskryptora o elementy (pojęcia, typy i inne wyrażenia) zdefiniuj maksymalną liczbę elementów, które można dodać do jednego deskryptora. Jeśli ten limit zostanie ustawiony na 10, to do istniejącego deskryptora zostanie dodanych nie więcej niż 10 elementów. Jeśli istnieje więcej niż 10 elementów do dodania, technika przerwie dodawanie nowych elementów po dziesiątym elemencie. W ten sposób można skrócić listę deskryptorów, ale nie ma gwarancji, że najbardziej interesujące elementy zostaną użyte w pierwszej kolejności. Preferowane może być ograniczenie wielkości uzupełnienia bez pogorszenia jakości dzięki opcji **Generalize with wildcards where possible**. Ta opcja dotyczy tylko deskryptorów, które zawierają boolowskie operatory & (I) lub ! (NIE).

Also extend subcategories. Ta opcja powoduje również uzupełnienie wszystkich kategorii poniżej wybranych.

Extend empty categories with descriptors generated from the category name. Ta metoda ma zastosowanie tylko do pustych kategorii, które mają 0 deskryptorów. Jeśli kategoria zawiera już deskryptory, nie zostanie w ten sposób uzupełniona. Ta opcja służy do automatycznego tworzenia deskryptorów dla każdej kategorii na podstawie wyrazów,

które tworzą nazwę kategorii. Nazwa kategorii jest przeszukiwana w celu sprawdzenia, czy wyrazy w nazwie pasują do wyodrębnionych pojęć. Jeśli jakieś pojęcie zostanie rozpoznane, zostanie użyte do znalezienia zgodnych wzorców pojęcia, które razem z nim staną się deskryptorami kategorii. Ta opcja daje najlepsze wyniki, gdy nazwy kategorii są długie i opisowe. Jest to szybka metoda generowania deskryptorów kategorii, które z kolei umożliwiają kategorii wychwytywanie rekordów zawierających te deskryptory. Ta opcja jest użyteczna w przypadku importowania kategorii z innego miejsca lub podczas ręcznego tworzenia kategorii z długimi nazwami opisowymi.

Generate descriptors as. Ta opcja ma zastosowanie tylko wtedy, gdy poprzednia opcja jest wybrana.

- **Concepts.** Wybierz tę opcję, aby wygenerować wynikowe deskryptory w formie pojęć, niezależnie od tego, czy zostały one wyodrębnione z tekstu źródłowego.
- **Patterns.** Wybierz tę opcję, aby wygenerować wynikowe deskryptory w formie pojęć, niezależnie od tego, czy wynikowe deskryptory lub jakiekolwiek inne deskryptory zostały wyodrębnione z tekstu źródłowego.

Ręczne tworzenie kategorii

Kategorie można tworzyć, używając zautomatyzowanych technik tworzenia kategorii oraz edytora reguł, a dodatkowo można tworzyć kategorie ręcznie. Istnieją następujące metody ręczne:

- Tworzenie pustej kategorii, do której elementy są dodawane pojedynczo. Więcej informacji zawiera temat “Tworzenie nowych kategorii lub zmienianie nazw kategorii”.
- Przeciąganie terminów, typów i wzorców do panelu kategorii. Więcej informacji zawiera temat “Tworzenie kategorii za pomocą metody przeciągania i upuszczania”.

Tworzenie nowych kategorii lub zmienianie nazw kategorii

Można utworzyć puste kategorie w celu dodawania do nich pojęć i typów. Można również zmieniać nazwy kategorii.

Aby utworzyć nową pustą kategorię

1. Przejdź do panelu Categories.
2. Z menu wybierz kolejno opcje **Categories > Create Empty Category**. Otworzy się okno dialogowe Category Properties.
3. Wprowadź nazwę dla tej kategorii w polu Name.
4. Kliknij przycisk **OK**, aby zatwierdzić nazwę i zamknąć okno dialogowe. Okno dialogowe zamyka się i w panelu pojawia się nowa nazwa kategorii.

Możesz teraz rozpocząć dodawanie do tej kategorii. Więcej informacji zawiera temat “Dodawanie deskryptorów do kategorii” na stronie 132.

Aby zmienić nazwę kategorii

1. Wybierz kategorię i wybierz opcje **Categories > Rename Category**. Otworzy się okno dialogowe Category Properties.
2. Wprowadź nową nazwę dla tej kategorii w polu Name.
3. Kliknij przycisk **OK**, aby zatwierdzić nazwę i zamknąć okno dialogowe. Okno dialogowe zamyka się i w panelu pojawia się nowa nazwa kategorii.

Tworzenie kategorii za pomocą metody przeciągania i upuszczania

Technika przeciągania i upuszczania jest techniką ręczną, a nie bazującą na algorytmach. Można utworzyć kategorie w panelu Categories, przeciągając:

- Wyodrębnione pojęcia, typy lub wzorce z panelu Extraction Results do panelu Categories.
- Wyodrębnione pojęcia w panelu Data do panelu Categories.
- Całe wiersze z panelu Data do panelu Categories. Spowoduje to utworzenie kategorii składającej się ze wszystkich wyodrębnionych pojęć i wzorców zawartych w tym wierszu.

Uwaga: Panel Extraction Results obsługuje wybór wielu elementów na raz, aby ułatwić użytkownikowi przeciąganie i upuszczanie wielu elementów.

Ważne! Z panelu Data nie można przeciągać i upuszczać pojęć, które nie zostały wyodrębnione z tekstu. Jeśli chcesz wymusić wyodrębnienie pojęcia, o którym wiesz, że występuje w danych, dodaj to pojęcie do typu. Następnie uruchom ponownie proces wyodrębniania. Nowe wyniki wyodrębniania będą zawierać pojęcie, który właśnie zostało dodane. Następnie można go użyć w kategorii. Więcej informacji zawiera temat “Dodawanie pojęć do typów” na stronie 89.

Aby utworzyć kategorie za pomocą metody przeciągania i upuszczania:

1. Na panelu Extraction Results lub Data wybierz jedno lub kilka pojęć, wzorców, typów, rekordów lub rekordów cząstkowych.
2. Przytrzymując wciśnięty przycisk myszy, przeciągnij element do istniejącej kategorii lub do obszaru panelu, aby utworzyć nową kategorię.
3. Po osiągnięciu tego obszaru, w którym chcesz upuścić element, zwolnij przycisk myszy. Element zostanie dodany do panelu Categories. Kategorie, które zostały zmodyfikowane, są wyświetlane ze specjalnym kolorem tła. Jest to tak zwane **tło informacyjne kategorii**. Więcej informacji zawiera temat “Określanie opcji” na stronie 74.

Uwaga: Wynikowej kategorii została automatycznie przypisana nazwa. Możesz zmienić tę nazwę. Więcej informacji zawiera temat “Tworzenie nowych kategorii lub zmienianie nazw kategorii” na stronie 115.

Aby sprawdzić, które rekordy są przypisane do kategorii, wybierz kategorię w panelu Categories. Panel danych zostanie automatycznie odświeżony i wyświetli wszystkie rekordy dla tej kategorii.

Korzystanie z reguł kategorii

Kategorie można tworzyć na wiele sposobów. Jeden z tych sposobów jest zdefiniowanie reguł kategorii wyrażających idee. Reguły kategorii są instrukcjami, które automatycznie klasyfikują dokumenty lub rekordy w kategorie na podstawie wyrażenia logicznego, używając wyodrębnionych pojęć, typów i wzorców oraz operatorów boolowskich. Na przykład, można napisać wyrażenie, które oznacza *uwzględnij w tej kategorii wszystkie rekordy zawierające wyodrębnione pojęcie* embassy, *ale nie* argentina.

Niektóre reguły kategorii są tworzone automatycznie podczas tworzenia kategorii przy użyciu techniki grupowania, takiej jak *analiza współwystąpień* i *wywodzenie rdzeni pojęć* (**Categories > Build Settings > Advanced Settings: Linguistics**), jednak można również utworzyć reguły kategorii ręcznie w edytorze reguł, na podstawie własnej interpretacji danych i kontekstu. Każda reguła jest przypisana do jednej kategorii, dzięki czemu każdy dokument lub rekord pasujący do reguły jest przypisywany do tej kategorii.

Reguły kategorii sprzyjają jakości i produktywności eksploracji tekstu i dalszej analizy ilościowej, ponieważ pozwalają na bardziej precyzyjną kategoryzację odpowiedzi. Doświadczenie i wiedza biznesowa użytkownika pozwala często na pogłębioną interpretację danych i kontekstu. Można wykorzystać tę interpretację, aby przekształcić posiadaną wiedzę w reguły kategorii w celu kategoryzowania dokumentów lub rekordów z jeszcze większą efektywnością i dokładnością. Istotą takich reguł jest łączenie wyodrębnionych elementów za pomocą operatorów boolowskich.

Możliwość tworzenia reguł zwiększa dokładność, efektywność i produktywność kodowania, pozwalając połączyć wiedzę biznesową z technologią wyodrębniania.

Uwaga: Aby zapoznać się z przykładami znajdowania tekstu przez reguły, patrz “Przykłady reguł kategorii” na stronie 122

Składnia reguł kategorii

Niektóre reguły kategorii są tworzone automatycznie podczas tworzenia kategorii przy użyciu techniki grupowania, takiej jak *analiza współwystąpień* i *wywodzenie rdzeni pojęć* (**Categories> Build Settings > Advanced Settings:**

Linguistics), jednak można również utworzyć reguły kategorii ręcznie w edytorze reguł. Każda reguła jest deskryptorem jednej kategorii, dlatego każdy dokument lub rekord pasujący do reguły jest automatycznie przypisywany do tej kategorii.

Uwaga: Aby zapoznać się z przykładami znajdowania tekstu przez reguły, patrz “Przykłady reguł kategorii” na stronie 122

Chcąc utworzyć lub edytować regułę, należy ją otworzyć w edytorze reguł. Można dodawać pojęcia, typy lub wzorce oraz używać symboli wieloznacznych, aby uelastycznić dopasowanie. Używając wyodrębnionych pojęć, typów i wzorców, można skorzystać z funkcji wyszukiwania wszystkich pojęć pokrewnych.

Ważne! Zalecamy przeciąganie i upuszczanie pojęć bezpośrednio z panelu Extraction Results, Text Link Analysis lub Data do edytora reguł albo dodawanie ich za pośrednictwem menu kontekstowego, co pozwoli uniknąć najprostszych pomyłek.

Rozpoznanie pojęcia, typu lub wzorca jest sygnalizowane ikoną wyświetlaną obok tekstu.

Tabela 18. Ikony wyodrębniania

Ikona	Opis
	Wyodrębnione pojęcie
	Wyodrębniony typ
	Wyodrębniony wzorzec

Składnia i operatory reguł

Poniższa tabela zawiera znaki używane do definiowania składni reguł. Znaków tych używa się wraz z pojęciami, typami i wzorcami.

Tabela 19. Obsługiwana składnia

Znakowa	Opis
&	Boolowskie „i”. Na przykład a & b zawiera a *i* b ; oto konkretne przykłady: - invasion & united states - 2016 & olympics - good & apple
|	Operator boolowski „lub” oznacza sumę logiczną, czyli do dopasowania wystarczy obecność jednego elementu. Na przykład a | b zawiera a *lub* b lub oba te elementy: - attack | france - condominium | apartment
!()	Boolowskie „nie”. Na przykład !(a) nie zawiera a: !(good & hotel), assassination & !(austria) lub !(gold) & !(copper)
*	Symbol wieloznaczny reprezentujący od jednego znaku do całego wyrazu, w zależności od tego, jak został użyty. Więcej informacji zawiera temat “Używanie symboli wieloznacznych w regułach kategorii” na stronie 120.
()	Ogranicznik wyrażenia. Wartość wyrażenia w nawiasach jest wyznaczana w pierwszej kolejności.
+	Łącznik wzorców służący do utworzenia wzorców o określonej kolejności. Musi być używany razem z nawiasami kwadratowymi. Więcej informacji zawiera temat “Używanie wzorców TLA w regułach kategorii” na stronie 118.

Tabela 19. Obsługiwana składnia (kontynuacja)

Znakowa	Opis
[]	Ogranicznik wzorca jest wymagany, jeśli w regule kategorii chcesz dopasowywać pojęcia na podstawie wyodrębnionego wzorca TLA. Zawartość nawiasów kwadratowych jest traktowana jako wzorzec TLA i nigdy nie będzie dopasowywana do pojęć lub typów na podstawie prostego współwystępowania. Jeśli nie wyodrębniono danego wzorca TLA, to dopasowanie nie będzie możliwe. Więcej informacji zawiera temat "Używanie wzorców TLA w regułach kategorii". Nie należy używać nawiasów kwadratowych, chcąc dopasować pojęcia i typy. *Uwaga*: W starszych wersjach współwystąpienia i reguły synonimów wygenerowane za pomocą technik budowania kategorii były ujęte w nawiasy kwadratowe. W przypadku wszystkich nowych wersji nawiasy kwadratowe wskazują na obecność wzorca TLA. Zamiast tego reguły tworzone przez technikę współwystąpień i synonimów są ujmowane w nawiasy okrągłe, na przykład (speaker systems|speakers).

Operatory & i | są przemienne, zatem a & b = b & a i a | b = b | a.

Poprzedzanie znaków ukośnikiem odwrotnym

Jeśli pojęcie zawiera znak, który jest również elementem składni, należy umieścić ukośnik odwrotny przed tym znakiem, by reguła została poprawnie zinterpretowana. Znak ukośnika odwrotnego (\) jest używany do zmiany znaczenia znaków, które bez niego mają specjalne znaczenie. W przypadku przeciągania i upuszczania elementów do edytora ukośniki odwrotne są dodawane automatycznie.

Następujące znaki składni reguł muszą być poprzedzone ukośnikiem odwrotnym, jeśli mają być one traktowane dosłownie:

```
& ! | + < > ( ) [ ] *
```

Ponieważ pojecie r&d zawiera operator "and" (&), wymagane jest użycie ukośnika lewego w edytorze reguł, np.: r\&d.

Używanie wzorców TLA w regułach kategorii

Wzorce analizy powiązań w tekście mogą być jawnie zdefiniowane w regułach kategorii, co pozwala uzyskać jeszcze bardziej konkretne i kontekstowe wyniki. Definiując wzorzec w regule kategorii, pomijasz wyniki prostego wyodrębniania i żądasz dopasowywania dokumentów i rekordów wyodrębnionych wynikowych wzorców analizy powiązań w tekście.

Ważne! Aby dopasowywać dokumenty na podstawie wzorców TLA w regułach kategorii, należy wcześniej uruchomić wyodrębnianie z włączoną analizą powiązań w tekście. Reguła kategorii będzie szukać dopasowań znalezionych podczas tego procesu. Jeśli nie wybrano opcji eksplorowania wyników TLA na karcie Model węzła Text Mining, można włączyć wyodrębnianie TLA w ustawieniach wyodrębniania w sesji interaktywnej, a następnie ponownie przeprowadzić wyodrębnianie. Więcej informacji zawiera temat "Wyodrębnianie danych" na stronie 80.

Ujmowanie w nawiasy kwadratowe. Wzorzec TLA musi być ujęty w nawiasy kwadratowe [ ], jeśli jest używany wewnątrz reguły kategorii. Ogranicznik wzorca jest wymagany, jeśli chcesz dopasowywać pojęcia na podstawie wyodrębnionego wzorca TLA. Ponieważ reguły kategorii mogą zawierać typy, pojęcia lub wzorce, nawiasy wskazują, że ich zawartość jest wyodrębnionym wzorcem TLA. Jeśli nie wyodrębniono danego wzorca TLA, to dopasowanie nie będzie możliwe. Jeśli w panelu Categories widzisz wzorzec bez nawiasów, np. apple + good, to prawdopodobnie wzorzec został dodany bezpośrednio do kategorii poza edytorem reguł kategorii. Na przykład, jeśli dodasz wzorzec pojęcia bezpośrednio do kategorii z widoku analizy powiązań w tekście, to nie będzie on wyświetlany w nawiasach kwadratowych. Jednak używając wzorzec w obrębie reguły kategorii, należy ująć wzorzec w nawiasy kwadratowe, na przykład [banana + !(good)].

Używanie znaku + we wzorcach. W programie IBM SPSS Modeler Text Analytics wzorzec może składać się maksymalnie z 6 części. Aby wskazać, że kolejność jest istotna, połącz elementy znakiem +, na przykład [company1 + acquired + company2]. W tym przypadku kolejność jest istotna, ponieważ powoduje zmianę znaczenia — która spółka była przejmującą, a która przejmowaną. Kolejność nie jest określana przez strukturę zdania, ale przez strukturę

wynikowego wzorca TLA. Na przykład, jeśli masz tekst "*I love Paris*" i chcesz wyodrębnić jego znaczenie, wzorzec TLA prawdopodobnie będzie miał postać [paris + like] lub [<Location> + <Positive>], a nie [<Positive> + <Location>], ponieważ w domyślnych zasobach służących do analizy opinii zwykle opinia znajduje się na 2. miejscu we wzorcu 2-częściowym. Dlatego celowe może być użycie wzorca bezpośrednio jako deskryptora kategorii. Jeśli jednak chcesz użyć wzorca jako część złożonej instrukcji, zwróć szczególną uwagę na kolejność elementów w obrębie wzorców w widoku Text Link Analysis, ponieważ kolejność ma duży wpływ na możliwość znalezienia dopasowania.

Załóżmy na przykład, że masz dwa teksty: "I *like pineapple*" i "I hate *pineapple*. However, I *like* strawberries". Wyrażenie like & pineapple pasuje do obu tekstów, bo jest wyrażeniem opisującym pojęcie, a nie regułą TLA (nie jest umieszczone w nawiasach kwadratowych). Wyrażenie pineapple + like pasuje tylko do tekstu "I *like pineapple*", ponieważ w drugim tekście wyraz *like* jest powiązany z wyrazem *strawberries*.

Grupowanie przy użyciu wzorców. Można uprościć reguły za pomocą własnych wzorców. Załóżmy, że chcesz wychwycić następujące trzy wyrażenia: cayenne peppers + like, chili peppers + like i peppers + like. Można je pogrupować w regule pojedynczej kategorii, na przykład [* peppers & like]. Jeśli masz także wyrażenie hot peppers + good, możesz pogrupować wszystkie cztery za pomocą takiej reguły, jak [* peppers + <Positive>].

Kolejność we wzorcach. W celu lepszego zorganizowania wyników reguły analizy powiązań w tekście dostarczone w szablonach instalowanych z produktem próbują generować proste wzorce w tej samej kolejności, bez względu na kolejność wyrazów w zdaniu. Na przykład, jeśli masz rekord zawierający tekst "*Good presentations.*" i inny rekord zawierający tekst "*the presentations were good*", to oba teksty pasują do tej samej reguły i zostaną na liście wynikowych wzorców pojęć wygenerowane w tej samej kolejności presentation + good, a nie jako dwa wzorce presentation + good oraz good + presentation. A we wzorcu dwuczęściowym, takim jak w przykładzie, pojęcia przypisane do typów z biblioteki Opinions będą domyślnie prezentowane w wynikach na końcu, na przykład apple + bad.

Tabela 20. Korzystanie ze składni wzorców i operatorów boolowskich

Wyrażenie	**Pasuje do dokumentu lub rekordu, który**
[ ]	Zawiera dowolny wzorzec TLA. Ogranicznik wzorca jest wymagany *w regułach kategorii*, jeśli chcesz dopasowywać na podstawie wyodrębnionego wzorca TLA. Zawartość nawiasów kwadratowych jest traktowana jako wzorzec TLA, a nie jako proste pojęcie lub typ. Jeśli nie wyodrębniono danego wzorca TLA, to dopasowanie nie będzie możliwe. Aby utworzyć regułę, która nie zawiera żadnych wzorców, można użyć składni !([ ]).
[a]	Zawiera wzorzec, którego co najmniej jednym elementem na dowolnej pozycji jest a. Na przykład [deal] pasuje do [deal + good] i do samego [deal + .]
[a + b]	Zawiera wzorzec pojęcia. Na przykład [deal + good]. *Uwaga:* Jeśli chcesz wychwycić tylko ten wzorzec bez dodawania żadnych innych elementów, zalecamy dodanie wzorca bezpośrednio do kategorii, a nie tworzenie reguły z jego użyciem.
[a + b + c]	Zawiera wzorzec pojęcia. Znak + oznacza, że kolejność dopasowywanych elementów jest istotna. Na przykład [company1 + acquired + company2].
[<A> + <B>]	Zawiera dowolny wzorzec typu <A> na pierwszej pozycji i typu <B> na drugiej pozycji; istnieją dokładnie dwie pozycje. Znak + oznacza, że kolejność dopasowywanych elementów jest istotna. Na przykład [<Budget> + <Negative>]. *Uwaga:* Jeśli chcesz wychwycić tylko ten wzorzec bez dodawania żadnych innych elementów, zalecamy dodanie wzorca bezpośrednio do kategorii, a nie tworzenie reguły z jego użyciem.
[<A> & <B>]	Zawiera dowolny wzorzec typu <A> i typu <B>. Na przykład [<Budget> & <Negative>]. Ten wzorzec TLA nigdy nie zostanie wyodrębniony; jednak zapisany jako taki wzorzec w istocie równy jest [<Budget> + <Negative>]|[<Negative> + <Budget>]. Kolejność dopasowywanych elementów jest nieistotna. Ponadto inne elementy mogą nie występować we wzorcu, ale muszą co najmniej być typu <Budget> i <Negative>.

Tabela 20. Korzystanie ze składni wzorców i operatorów boolowskich (kontynuacja)

Wyrażenie	Pasuje do dokumentu lub rekordu, który
[a + .]	Zawiera wzorzec, w którym a jest jedynym pojęciem, a pozostałe pozycje tego wzorca są puste. Na przykład [deal + .] pasuje do wzorca pojęcia, w którym jedynym wynikiem jest pojęcie deal. Jeśli dodano pojęcie deal jako deskryptora kategorii, zwrócone zostałyby wszystkie rekordy z pojęciem deal, w tym pozytywne opinie o tym pojęciu. Jednak wzorzec [deal + .] pasuje tylko do rekordów reprezentujących pojęcie deal, ale nie inne relacje lub opinie, więc nie pasuje na przykład do deal + fantastic. *Uwaga:* Jeśli chcesz wychwycić tylko ten wzorzec bez dodawania żadnych innych elementów, zalecamy dodanie wzorca bezpośrednio do kategorii, a nie tworzenie reguły z jego użyciem.
[<A> + <>]	Zawiera wzorzec, w którym <A> jest jedynym typem. Na przykład [<Budget> + <>] pasuje do wzorca, którego jedynym wynikiem jest pojęcie typu <Budget>. *Uwaga:* Można użyć <> do określenia typu pustego, ale wyłącznie za symbolem + we wzorcu typu, na przykład [<Budget> + <>], a nie np. we wzorcu [price + <>]. *Uwaga:* Jeśli chcesz wychwycić tylko ten wzorzec bez dodawania żadnych innych elementów, zalecamy dodanie wzorca bezpośrednio do kategorii, a nie tworzenie reguły z jego użyciem.
[a + !(b)]	Zawiera co najmniej jeden wzorzec, który obejmuje pojęcie a, ale nie obejmuje pojęcia b. Musi obejmować co najmniej jeden wzorzec. Na przykład [price + !(high)] lub w przypadku typów: [!(<Fruit>|<Vegetable>) + <Positive>]
!([<A> & <B>])	Nie zawiera określonego wzorca. Na przykład !([<Budget> & <Negative>]).

Uwaga: Aby zapoznać się z przykładami znajdowania tekstu przez reguły, patrz “Przykłady reguł kategorii” na stronie 122

Używanie symboli wieloznacznych w regułach kategorii

Do pojęć w regułach można dodawać symbole wieloznaczne, aby reguły pasowały do szerszego zbioru pojęć. Symbol wieloznaczny gwiazdki (*) można umieścić przed i/lub po wyrazie, aby wskazać sposób, w jaki pojęcia mogą być dopasowywane. Wyróżnia się dwa zastosowania symboli wieloznacznych:

- **Symbole wieloznaczne dołączane**. Te znaki wstawiane są od razu po łańcuchu znaków lub przed nim i nie są oddzielone od niego spacją. Na przykład operat* pasuje do *operat*, *operate*, *operates*, *operations*, *operational* i tak dalej.
- **Symbole wieloznaczne zastępujące wyrazy**. Takie symbole wieloznaczne (gwiazdki) poprzedzają pojęcie lub następują po nim, ale są oddzielone od niego spacją. Na przykład * operation pasuje do *operation*, *surgical operation*, *post operation* i tak dalej. Ponadto symbol wieloznaczny zastępujący wyraz może być używany razem z dołączanym symbolem wieloznacznym, na przykład * operat* * pasuje do *operation*, *surgical operation*, *telephone operator*, *operatic aria* i tak dalej. Jak widać w ostatnim przykładzie, zaleca się ostrożne stosowanie symboli wieloznacznych, aby „sieć” nie została zarzucona zbyt szeroko i nie wychwyciła niepożądanych wyników.

Wyjątki!

- Symbol wieloznaczny nie może być stosowany autonomicznie. Na przykład niedopuszczalny jest zapis (apple | *).
- Symbol wieloznaczny nie może być używany do dopasowywania nazw typów. <Negative*> nie pasuje do żadnej nazwy typu.
- Nie można wykluczyć wybranych typów z dopasowywania do pojęć znalezionych za pomocą symboli wieloznacznych. Typ jest przypisywany do pojęcia automatycznie.
- Symbol wieloznaczny nie może znajdować się pośrodku sekwencji wyrazów, czy to na końcu lub początku wyrazu (open* account), czy to jako autonomiczny komponent (open * account). Nie można używać symboli wieloznacznych w nazwach typów. Na przykład reguła w postaci wyraz* wyraz, taki jak apple* recipe, nie pasuje do applesauce recipe ani w ogóle do żadnego tekstu. Jednak , apple* * pasuje do *applesauce recipe*, *apple pie*, *apple*

i tak dalej. Oto inny przykład: wyraz * wyraz, na przykład apple * toast, nie pasuje do *apple cinnamon toast* ani do żadnego innego tekstu, ponieważ gwiazdka pojawia się między dwoma wyrazami. Jednak apple * pasuje do *apple cinnamon toast*, *apple*, *apple pie* i tak dalej.

Tabela 21. Stosowanie symboli wieloznacznych

Wyrażenie	**Pasuje do dokumentu lub rekordu, który**
*apple	Obejmuje pojęcie, które kończy się podanymi literami, ale może mieć dowolną liczbę liter w przedrostku. Na przykład: *apple kończy się literami *apple*, ale może mieć przedrostek: - apple - pineapple - crabapple
apple*	Obejmuje pojęcie, które rozpoczyna się od podanych liter, ale może mieć dowolną liczbę liter jako przyrostek. Na przykład: apple* zaczyna się od liter *apple*, ale może mieć przyrostek lub go nie mieć: - apple - applesauce - applejack Na przykład reguła apple* & !(pear* | quince), który obejmuje pojęcie rozpoczynające się od liter apple, ale nie pojęcie rozpoczynające się od liter *pear* i nie pojęcie quince, NIE pasuje do: apple & quince ale pasuje do: - applesauce - apple & orange
product	Obejmuje pojęcie, które zawiera podane litery product, ale może mieć dowolną liczbę liter jako przedrostek i/lub przyrostek. Na przykład *product* pasuje do: - product - byproduct - unproductive
* loan	Obejmuje pojęcie zawierające wyraz loan, ale może być złożone z innym wyrazem go poprzedzającym. Na przykład * loan pasuje do: - loan - car loan - home equity loan Na przykład [* delivery + <Negative>] obejmuje pojęcie kończące się wyrazem delivery na pierwszej pozycji i pojęcie o typie <Negative> na drugiej pozycji. Taka reguła pasuje na przykład do pojęć: - package delivery + slow - overnight delivery + late
event *	Obejmuje pojęcie zawierające wyraz event, ale może być złożone z innym wyrazem następującym po tym wyrazie. Na przykład event * pasuje do: - event - event location - event planning committee

Tabela 21. Stosowanie symboli wieloznacznych (kontynuacja)

Wyrażenie	Pasuje do dokumentu lub rekordu, który
* apple *	Obejmuje pojęcie, które może zaczynać się od dowolnego wyrazu, po którym następuje wyraz apple i potencjalnie następny wyraz. * oznacza 0 lub n wystąpień, więc taka reguła pasuje też do apple. Na przykład * apple * pasuje do: - gala applesauce - granny smith apple crumble - famous apple pie - apple Na przykład reguła [* reservation* * + <Positive>], która obejmuje pojęcie zawierające wyraz reservation (na dowolnej pozycji wewnątrz tego pojęcia) na pierwszej pozycji oraz pojęcie typu <Positive> na drugiej pozycji pasowałaby do: - reservation system + good - online reservation + good

Uwaga: Aby zapoznać się z przykładami znajdowania tekstu przez reguły, patrz “Przykłady reguł kategorii”

Przykłady reguł kategorii

W celu zademonstrowania, w jaki sposób reguły są dopasowywane do rekordów w zależności od składni, prezentujemy poniższy przykład.

Rekordy przykładowe

Załóżmy, że mamy dwa rekordy:

- **Rekord A**: *„when I checked my wallet, I saw I was missing 5 dollars.”*
- **Rekord B**: *„$5 was found at the picnic area, but the blanket was missing.”*

Następujące dwie tabele przedstawiają oczekiwane wyniki wyodrębniania pojęć, typów oraz wzorców pojęć i typów.

Pojęcia i typy wyodrębnione z przykładowych rekordów

Tabela 22. Pojęcia i typy wyodrębnione z przykładowych rekordów

Wyodrębnione pojęcie	Typ pojęcia
wallet	<Unknown>
missing	<Negative>
USD5	<Currency>
blanket	<Unknown>
picnic area	<Unknown>

Wzorce TLA wyodrębnione z przykładowych rekordów

Tabela 23. Wynikowe wzorce TLA wyodrębnione z przykładowych rekordów

Wyodrębnione wzorce pojęć	Wyodrębnione wzorce typów	Z rekordu
picnic area + .	<Unknown> + <>	Rekord B
wallet + .	<Unknown> + <>	Rekord A
blanket + missing	<Unknown> + <Negative>	Rekord B
USD5 + .	<Currency> + <>	Rekord B

Tabela 23. Wynikowe wzorce TLA wyodrębnione z przykładowych rekordów (kontynuacja)

Wyodrębnione wzorce pojęć	Wyodrębnione wzorce typów	Z rekordu
USD5 + missing	<Currency> + <Negative>	Rekord A

Możliwe dopasowania generowane przez reguły kategorii

Poniższa tabela zawiera przykłady składni, które można wprowadzić w edytorze reguł kategorii. Nie wszystkie podane tutaj reguły działają i nie wszystkie pasują do tych samych rekordów. Oto, jak różnice w zapisie wpływają na wyniki dopasowywania rekordów.

Tabela 24. Przykładowe reguły

Składnia reguły	Wynik
USD5 & missing	Pasuje do rekordów A i B, ponieważ oba zawierają wyodrębnione pojęcie missing i wyodrębnione pojęcie USD5. Jest równoważna: (USD5 & missing)
missing & USD5	Pasuje do rekordów A i B, ponieważ oba zawierają wyodrębnione pojęcie missing i wyodrębnione pojęcie USD5. Jest równoważna: (missing & USD5)
missing & <Currency>	Pasuje do rekordów A i B, ponieważ oba zawierają wyodrębnione pojęcie missing i pojęcie typu <Currency>. Jest równoważna: (missing & <Currency>)
<Currency> & missing	Pasuje do rekordów A i B, ponieważ oba zawierają wyodrębnione pojęcie missing i pojęcie typu <Currency>. Jest równoważna: (<Currency> & missing)
[USD5 + missing]	Pasuje do rekordu A, ale **nie** do B, ponieważ rekord B nie wygenerował żadnych wynikowych wzorców TLA zawierających pojęcia USD5 + missing (patrz poprzednia tabela). Jest to odpowiednik wynikowego wzorca TLA: USD5 + missing
[missing + USD5]	Nie pasuje ani do rekordu A, ani do B, ponieważ żaden z wyodrębnionych wzorców TLA (patrz poprzednia tabela) nie pasuje do podanej tutaj kolejności (missing na pierwszej pozycji). Jest to odpowiednik wynikowego wzorca TLA: USD5 + missing
[missing & USD5]	Pasuje do A, ale **nie** do B, ponieważ z rekordu B nie wyodrębniono takiego wzorca TLA. Znak & wskazuje, że kolejność jest nieistotna; dlatego ta reguła szuka zarówno [missing + USD5], jak i [USD5 + missing]. Pasuje tylko [USD5 + missing] z rekordu A.
[missing + <Currency>]	Nie pasuje ani do rekordu A, ani do B, ponieważ żaden wyodrębniony wzorzec TLA nie pasuje do tej kolejności. Tutaj brak odpowiednika, ponieważ wyniki TLA oparte są na terminach (USD5 + missing) lub typach (<Currency> + <Negative>), ale nie na kombinacjach pojęć i typów.
[<Currency> + <Negative>]	Pasuje do rekordu A, ale nie do B, ponieważ z rekordu B nie wyodrębniono żadnego wzorca TLA. Jest to odpowiednik wynikowego wzorca TLA: <Currency> + <Negative>
[<Negative> + <Currency>]	Nie pasuje ani do rekordu A, ani do B, ponieważ żaden wyodrębniony wzorzec TLA nie pasuje do tej kolejności. W szablonie Opinions znalezienie *tematu* z *opinią* powoduje, że *temat* (<Currency>) zajmuje domyślnie pierwszą pozycję, a *opinia* (<Negative>) zajmuje drugą pozycję.

Tworzenie reguł kategorii

Aby tworzyć lub edytować regułę, należy otworzyć ją w edytorze reguł. Można dodawać pojęcia, typy lub wzorce oraz używać symboli wieloznacznych, aby uelastycznić dopasowanie. Używanie rozpoznanych pojęć, typów i wzorców jest

dobrym wyborem, ponieważ umożliwia znajdowanie wszystkich pojęć pokrewnych. Na przykład użycie pojęcia spowoduje, że reguła będzie także znajdowała wszystkie powiązane z nim terminy, formy w liczbie mnogiej i synonimy. Podobnie użycie typu powoduje, że reguła wychwytuje wszystkie pojęcia przypisane do tego typu.

Istniejącą regułę można otworzyć w edytorze reguł, klikając prawym przyciskiem myszy nazwę kategorii i wybierając opcję **Create Rule**.

W edytorze można używać menu kontekstowych, techniki przeciągania i upuszczania lub ręcznie wprowadzać pojęcia, typy i wzorce. Elementy te można łączyć za pomocą operatorów boolowskich. (&, !(), |) i ujmować w nawiasy, formułując w ten sposób wyrażenia opisujące reguły. Zalecamy przeciąganie i upuszczanie pojęć bezpośrednio z panelu Extraction Results lub panelu Data do edytora reguł, co pozwoli uniknąć najprostszych pomyłek. Aby uniknąć błędów, należy ściśle przestrzegać składni reguł. Więcej informacji zawiera temat “Składnia reguł kategorii” na stronie 116.

Uwaga: Aby zapoznać się z przykładami znajdowania tekstu przez reguły, patrz “Przykłady reguł kategorii” na stronie 122.

Aby utworzyć regułę

1. Jeśli jeszcze nie zostały wyodrębnione żadne dane lub wyniki wyodrębniania są nieaktualne, przeprowadź teraz wyodrębnianie. Więcej informacji zawiera temat “Wyodrębnianie danych” na stronie 80.

   Uwaga: Jeśli filtr wyodrębniania spowoduje, że nie będą widoczne żadne pojęcia, to przy próbie utworzenia lub edytowania reguły kategorii zostanie wyświetlony komunikat o błędzie. Aby temu zapobiec, zmodyfikuj filtr wyodrębniania tak, by pojęcia były dostępne.
2. W panelu Categories wybierz kategorię, do której chcesz dodać regułę.
3. Z menu wybierz kolejno opcje **Categories > Create Rule**. W oknie zostanie otwarty panel edytora reguł kategorii.
4. W polu Rule Name wprowadź nazwę reguły. Jeśli nie podasz nazwy, za nazwę automatycznie przyjęte zostanie wyrażenie. Nazwę reguły możesz później zmienić.
5. W większym polu tekstowym wyrażenia można:
   - Bezpośrednio wprowadzać tekst lub przeciągać go z innych paneli. Używać tylko wyodrębnionych pojęć, typów i wzorców. Na przykład, jeśli wprowadzisz wyraz cats, ale na panelu Extraction Results figuruje on tylko w liczbie pojedynczej (cat), to edytor nie rozpozna wyrazu cats. W takiej sytuacji możliwe jest, że forma pojedyncza automatycznie uwzględnia także liczbę mnogą. Jeśli nie, to można użyć symbolu wieloznacznego. Więcej informacji zawiera temat “Składnia reguł kategorii” na stronie 116.
   - Wybierz pojęcia, typy lub wzorce, które chcesz dodać do reguły, i skorzystaj z menu.
   - Dodaj operatory boolowskie, aby powiązać elementy reguły. Za pomocą przycisków na pasku narzędzi możesz dodać do reguły operator „i” **&**, operator „lub” **|**, operator „nie” **!()**, nawiasy okrągłe **()** i nawiasy kwadratowe oznaczające wzorzec **[ ]**.
6. Kliknij przycisk **Test Rule**, aby sprawdzić, czy reguła jest prawidłowo zdefiniowana. Więcej informacji zawiera temat “Składnia reguł kategorii” na stronie 116. Obok napisu **Test result** pojawi się liczba znalezionych dokumentów lub rekordów. Na prawo od tego napisu widoczna jest liczba elementów rozpoznanych przez regułę oraz ewentualne komunikaty o błędach. Jeśli ikona obok typu, wzorca lub pojęcia jest opatrzona czerwonym znakiem zapytania, oznacza to, że dany element nie pasuje do żadnych znanych wyodrębnionych elementów. W takim przypadku reguła nie znajdzie żadnych rekordów.
7. Aby przetestować część reguły, zaznacz tę część i kliknij przycisk **Test Selection**.
8. W razie napotkania problemów wprowadź niezbędne zmiany i ponownie przetestuj regułę.
9. Po zakończeniu kliknij przycisk **Save & Close**, aby ponownie zapisać regułę i zamknąć edytor. Nazwa nowej reguły pojawi się w kategorii.

Edytowanie i usuwanie reguł

Po utworzeniu i zapisaniu reguły można ją w dowolnej chwili poddać edycji. Więcej informacji zawiera temat “Składnia reguł kategorii” na stronie 116.

Jeśli reguła nie jest już potrzebna, można ją usunąć.

Aby edytować reguły

1. Wybierz regułę w tabeli Descriptors w oknie dialogowym Category Definitions.
2. Z menu wybierz kolejno opcje **Categories > Edit Rule** lub kliknij dwukrotnie nazwę reguły. Zostanie otwarte okno edytora z wybraną regułą.
3. Wprowadź niezbędne zmiany w regule, korzystając z wyników wyodrębniania i przycisków na pasku narzędzi.
4. Ponownie przetestuj regułę, aby upewnić się, że zwraca oczekiwane wyniki.
5. Kliknij przycisk **Save & Close**, aby ponownie zapisać regułę i zamknąć edytor.

Aby usunąć regułę

1. Wybierz regułę w tabeli Descriptors w oknie dialogowym Category Definitions.
2. Z menu wybierz kolejno opcje **Edit > Delete**. Reguła zostanie usunięta z kategorii.

Importowanie i eksportowanie predefiniowanych kategorii

Jeśli własne kategorie są zapisane w programie Microsoft Excel (*.*xls*, *.*xlsx*), można je zaimportować do produktu IBM SPSS Modeler Text Analytics .

Można również eksportować kategorie, które są otwarte w projekcie interaktywnego pulpitu roboczego do formatu plików Microsoft Excel (*.*xls*, *.*xlsx*). Podczas eksportowania kategorii można zadecydować, aby uwzględnić lub wyłączyć niektóre informacje dodatkowe, takie jak deskryptory i oceny. Więcej informacji zawiera temat “Eksportowanie kategorii” na stronie 129.

Jeśli twoje predefiniowane kategorie nie mają kodów lub gdy potrzebujesz nowych kodów, można automatycznie wygenerować nowy zbiór kodów dla zestawu kategorii w panelu Categories, wybierając kolejno opcje **Categories > Manage Categories > Autogenerate Codes** z menu. Spowoduje to usunięcie istniejących kodów i automatyczne ich ponumerowanie.

Importowanie predefiniowanych kategorii

Możliwe jest importowanie predefiniowanych kategorii do produktu IBM SPSS Modeler Text Analytics . Przed importowaniem upewnij się, że plik predefiniowanej kategorii znajduje się w pliku Microsoft Excel (*.*xls*, *.*xlsx*) i ma strukturę jednego z formatów pomocniczych. Można ustawić, aby produkt automatycznie wykrywał format za użytkownika. Obsługiwane są następujące formaty:

- **Flat list format**: Więcej informacji można znaleźć w temacie “Płaski format listy” na stronie 126.
- **Compact format**: Więcej informacji można znaleźć w temacie “Format kompaktowy” na stronie 127.
- **Indented format**: Więcej informacji można znaleźć w temacie “Format wcięty” na stronie 128.

Aby importować predefiniowane kategorie

1. W menu interaktywnego pulpitu roboczego wybierz kolejno opcje **Categories > Manage Categories > Import Predefined Categories**. Wyświetlony zostaje kreator Import Predefined Categories.
2. Z listy rozwijanej Look In wybierz napęd i folder, w którym znajduje się plik.
3. Wybierz plik z listy. Nazwa pliku pojawia się w polu tekstowym File Name.
4. Z listy wybierz arkusz zawierający predefiniowane kategorie. Nazwa arkusza pojawia się w polu Worksheet.
5. Aby rozpocząć wybieranie formatu danych, kliknij przycisk **Next**.
6. Wybierz format dla swojego pliku lub wybierz opcję, aby pozwolić produktowi na próbę automatycznego wykrycia formatu. Automatyczne wykrywanie działa najlepiej dla najbardziej powszechnych formatów.
   - **Flat list format**: Więcej informacji można znaleźć w temacie “Płaski format listy” na stronie 126.
   - **Compact format**: Więcej informacji można znaleźć w temacie “Format kompaktowy” na stronie 127.
   - **Indented format**: Więcej informacji można znaleźć w temacie “Format wcięty” na stronie 128.

7. Aby zdefiniować dodatkowe opcje importowania, kliknij przycisk **Next**. Jeśli zadecydujesz, aby automatycznie wykryć format, przejdziesz do końcowego kroku.
8. Jeśli jeden lub więcej wierszy zawiera nagłówki kolumn lub inne informacje zewnętrzne, w opcji **Start import at row** wybierz numer wiersza, od którego chcesz rozpocząć importowanie. Na przykład: jeśli nazwy kategorii rozpoczynają się w wierszu 7, musisz wprowadzić liczbę 7 dla tej opcji, aby poprawnie zaimportować plik.
9. Jeśli plik zawiera kody kategorii, zaznacz opcję **Contains category codes**. Pomaga to kreatorowi poprawnie rozpoznać dane.
10. Sprawdź oznaczone kolorami komórki i legendę, aby upewnić się, że dane zostały poprawnie zidentyfikowane. Wszelkie błędy wykryte w pliku wyświetlone są na czerwono i odwołanie do nich wyświetlane jest poniżej tabeli podglądu formatu. Jeśli wybrano nieprawidłowy format, cofnij i wybierz inny. Jeśli chcesz wprowadzić poprawki w pliku, wprowadź zmiany i uruchom ponownie kreatora, wybierając znów plik. Przed zakończeniem działania kreatora należy poprawić wszystkie błędy.
11. Aby sprawdzić zestaw kategorii i podkategorii, które zostaną zaimportowane i aby zdefiniować, w jaki sposób tworzyć deskryptory dla tych kategorii, kliknij przycisk **Next**.
12. Przejrzyj zestaw kategorii, który będzie zaimportowany do tabeli. Jeśli nie są widoczne słowa kluczowe oczekiwane jako deskryptory, możliwe, że nie zostały rozpoznane podczas importu. Upewnij się, że mają poprawny przedrostek i pojawiają się we właściwej komórce.
13. Wybierz, jak obsługiwać istniejące wcześniej kategorie w sesji.
    - **Replace all existing categories**. Ta opcja czyści wszystkie istniejące kategorie, a następnie tylko nowo zaimportowane kategorie są używane na ich miejsce.
    - **Append to existing categories**. Ta opcja zaimportuje kategorie i scali wszelkie wspólne kategorie z istniejącymi kategoriami. Podczas dodawania elementów do istniejących kategorii należy określić, w jaki sposób obsługiwane będą duplikaty. Jedną z możliwości (opcja: **Merge**) jest scalanie importowanych kategorii z istniejącymi kategoriami, jeśli mają wspólną nazwę kategorii. Kolejną możliwością (opcja: **Exclude from import**) jest zakazanie importowania kategorii, jeśli istnieje już taka o tej samej nazwie.
14. **Import keywords as descriptors** to opcja importowania słów kluczowych zidentyfikowanych w danych jako deskryptorów dla powiązanej kategorii.
15. **Extend categories by deriving descriptors** to opcja, która wygeneruje deskryptory ze słów, które reprezentują nazwę kategorii lub podkategorii i/lub słów, które tworzą adnotacje. Jeśli słowa odpowiadają wyodrębnionym wynikom, są dodawane jako deskryptory do kategorii. Opcja zapewnia najlepsze wyniki, gdy nazwy kategorii lub adnotacje są długie i opisowe. Jest to szybka metoda wygenerowania deskryptorów kategorii, która pozwala na przechwycenie rekordów zawierających te deskryptory.
    - Pole **From** pozwala na wybranie, z jakiego tekstu tworzone będą deskryptory: z nazw lub kategorii i podkategorii, czy słów w adnotacjach lub z obu opcji.
    - Pole **As** pozwala na wybranie, aby tworzyć te deskryptory w formie pojęć lub wzorców TLA. Jeśli wyodrębnianie TLA nie miało miejsca, opcje **wzorców** są wyłączone w tym kreatorze.
16. Aby zaimportować wstępnie zdefiniowane kategorie do panelu Categories, kliknij przycisk **Finish**.

Płaski format listy

Na płaskim formacie listy istnieje tylko górny poziom kategorii bez hierarchii, co oznacza brak podkategorii lub podsieci. Nazwy kategorii znajdują się w pojedynczej kolumnie.

Plik w tym formacie może zawierać następujące informacje:

- Opcjonalna kolumna **codes** zawiera wartości numeryczne, które jednoznacznie identyfikują każdą kategorię. Jeśli określisz, że plik danych zawiera kody (opcja **Contains category codes** w kroku **Content Settings**), to kolumna zawierająca unikalne kody dla każdej kategorii musi występować w komórce bezpośrednio po lewej stronie nazwy kategorii. Jeśli dane nie zawierają kodów, ale chcesz wygenerować kody później, zawsze możesz to zrobić (**Categories > Manage Categories > Autogenerate Codes**).
- *Wymagana* kolumna **category names** zawiera wszystkie nazwy kategorii. Ta kolumna jest wymagana do importowania przy użyciu tego formatu.
- Opcjonalne **adnotacje** w komórce bezpośrednio po prawej stronie nazwy kategorii. Adnotacja zawiera tekst opisujący kategorie/podkategorie.

- Opcjonalne **słowa kluczowe** można zaimportować jako deskryptory dla kategorii. Aby słowa kluczowe zostały rozpoznane, muszą występować w komórce bezpośrednio pod powiązaną nazwą kategorii/podkategorii, a lista słów kluczowych musi być poprzedzona znakiem podkreślenia (_), np.: _firearms, weapons / guns. Komórka słów kluczowych może zawierać jedno lub wiele słów używanych do opisania każdej kategorii. Te słowa zostaną zaimportowane jako deskryptory lub zignorowane w zależności od tego, co określono w ostatnim kroku kreatora. Później deskryptory są porównywane z wyodrębnionymi wynikami z tekstu. Jeśli znaleziono dopasowanie, to ten rekord lub dokument jest oceniany w kategorii zawierającej ten deskryptor.

Tabela 25. Płaski format listy z kodami, słowami kluczowymi i adnotacjami

Kolumna A	**Kolumna B**	**Kolumna C**
Kod kategorii (*opcjonalny*)	Nazwa kategorii	Adnotacja
	_Deskryptor/lista słów kluczowych (*opcjonalnie*)	

Format kompaktowy

Format kompaktowy ma podobną strukturę do płaskiego formatu listy z taką różnicą, że format kompaktowy jest używany z kategoriami hierarchicznymi. Dlatego też wymagana jest kolumna poziomu kodu, aby zdefiniować poziom hierarchii każdej kategorii i podkategorii.

Plik w tym formacie może zawierać następujące informacje:

- *Wymagana* kolumna **code level** zawiera numery, które wskazują pozycję hierarchiczną dla kolejnych informacji w tym wierszu. Na przykład: jeśli określono wartości 1, 2 lub 3 i istnieją zarówno kategorie, jak i podkategorie, to 1 oznacza kategorie, 2 podkategorie, a 3 pod-podkategorie. Jeśli istnieją tylko kategorie i podkategorie, 1 oznacza kategorie, a 2 podkategorie. I tak dalej, aż do wymaganej głębokości kategorii.
- Opcjonalna kolumna **codes** zawiera wartości, które jednoznacznie identyfikują każdą kategorię. Jeśli określisz, że plik danych zawiera kody (opcja **Contains category codes** w kroku **Content Settings**), to kolumna zawierająca unikalne kody dla każdej kategorii musi występować w komórce bezpośrednio po lewej stronie nazwy kategorii. Jeśli dane nie zawierają kodów, ale chcesz wygenerować kody później, zawsze możesz to zrobić (**Categories > Manage Categories > Autogenerate Codes**).
- *Wymagana* kolumna **category names** zawiera wszystkie nazwy kategorii i podkategorii. Ta kolumna jest wymagana do importowania przy użyciu tego formatu.
- Opcjonalne **adnotacje** w komórce bezpośrednio po prawej stronie nazwy kategorii. Adnotacja zawiera tekst opisujący kategorie/podkategorie.
- Opcjonalne **słowa kluczowe** można zaimportować jako deskryptory dla kategorii. Aby słowa kluczowe zostały rozpoznane, muszą występować w komórce bezpośrednio pod powiązaną nazwą kategorii/podkategorii, a lista słów kluczowych musi być poprzedzona znakiem podkreślenia (_), np.: _firearms, weapons / guns. Komórka słów kluczowych może zawierać jedno lub wiele słów używanych do opisania każdej kategorii. Te słowa zostaną zaimportowane jako deskryptory lub zignorowane w zależności od tego, co określono w ostatnim kroku kreatora. Później deskryptory są porównywane z wyodrębnionymi wynikami z tekstu. Jeśli znaleziono dopasowanie, to ten rekord lub dokument jest oceniany w kategorii zawierającej ten deskryptor.

Tabela 26. Przykład formatu kompaktowego z kodami

Kolumna A	**Kolumna B**	**Kolumna C**
Poziom kodów hierarchicznych	Kod kategorii (*opcjonalny*)	Nazwa kategorii
Poziom kodów hierarchicznych	Kod podkategorii (*opcjonalny*)	Nazwa podkategorii

Tabela 27. Przykład formatu kompaktowego bez kodów

Kolumna A	**Kolumna B**
Poziom kodów hierarchicznych	Nazwa kategorii
Poziom kodów hierarchicznych	Nazwa podkategorii

Format wcięty

We wciętym formacie pliku zawartość jest hierarchiczna, co oznacza, że zawiera kategorie i jeden lub wiele poziomów podkategorii. Dodatkowo jego struktura jest wcięta, aby oznaczyć tę hierarchię. Każdy wiersz w pliku zawiera kategorię lub podkategorię, ale podkategorie są wcięte w stosunku do kategorii, a ewentualne pod-podkategorie są wcięte w stosunku do podkategorii itd. Można ręcznie utworzyć tę strukturę w programie Microsoft Excel lub użyć struktury eksportowanej z innego produktu i zapisanej w formacie aplikacji Microsoft Excel.

- **Kody kategorii najwyższego poziomu i nazwy kategorii** zajmują odpowiednio kolumny A i B. Jeśli kody nie występują, nazwa kategorii znajduje się w kolumnie A.
- **Kody podkategorii i nazwy podkategorii** zajmują odpowiednio kolumny B i C. Jeśli kody nie występują, nazwa podkategorii znajduje się w kolumnie B. Podkategoria jest członkiem kategorii. Nie mogą istnieć podkategorie, jeśli nie istnieją kategorie najwyższego poziomu.

Tabela 28. Wcięta struktura z kodami

Kolumna A	**Kolumna B**	**Kolumna C**	**Kolumna D**
Kod kategorii (*opcjonalny*)	Nazwa kategorii		
	Kod podkategorii (*opcjonalny*)	Nazwa podkategorii	
		Kod pod-podkategorii (*opcjonalny*)	Nazwa pod-podkategorii

Tabela 29. Wcięta struktura bez kodów

Kolumna A	**Kolumna B**	**Kolumna C**
Nazwa kategorii		
	Nazwa podkategorii	
		Nazwa pod-podkategorii

Plik w tym formacie może zawierać następujące informacje:

- Opcjonalne **kody** muszą być wartościami, które jednoznacznie identyfikują każdą kategorię lub podkategorię. Jeśli określisz, że plik danych zawiera kody (opcja **Contains category codes** w kroku **Content Settings**), to unikalny kod dla każdej kategorii lub podkategorii musi występować w komórce bezpośrednio po lewej stronie nazwy kategorii/podkategorii. Jeśli dane nie zawierają kodów, ale chcesz wygenerować kody później, zawsze możesz to zrobić (**Categories > Manage Categories > Autogenerate Codes**).
- *Wymagana* **nazwa** dla każdej kategorii i podkategorii. Podkategorie muszą być wcięte względem kategorii o jedną kolumnę w prawo w osobnym wierszu.
- Opcjonalne **adnotacje** w komórce bezpośrednio po prawej stronie nazwy kategorii. Adnotacja zawiera tekst opisujący kategorie/podkategorie.
- Opcjonalne **słowa kluczowe** można zaimportować jako deskryptory dla kategorii. Aby słowa kluczowe zostały rozpoznane, muszą występować w komórce bezpośrednio pod powiązaną nazwą kategorii/podkategorii, a lista słów kluczowych musi być poprzedzona znakiem podkreślenia (_), np.: _firearms, weapons / guns. Komórka słów kluczowych może zawierać jedno lub wiele słów używanych do opisania każdej kategorii. Te słowa zostaną zaimportowane jako deskryptory lub zignorowane w zależności od tego, co określono w ostatnim kroku kreatora. Później deskryptory są porównywane z wyodrębnionymi wynikami z tekstu. Jeśli znaleziono dopasowanie, to ten rekord lub dokument jest oceniany w kategorii zawierającej ten deskryptor.

Ważne! Jeśli używasz kodu na jednym poziomie, musisz uwzględnić kod dla każdej kategorii i podkategorii. W przeciwnym razie proces importowania zakończy się niepowodzeniem.

Eksportowanie kategorii

Można również eksportować kategorie, które są otwarte w projekcie interaktywnego pulpitu roboczego do formatu plików Microsoft Excel (*.*xls*, *.*xlsx*). Dane, które będą eksportowane pochodzą głównie z bieżącej zawartości panelu Categories lub z właściwości kategorii. Dlatego też zalecamy ponowne wykonanie oceny, jeśli planujesz eksportować wartość oceny **Docs** .

Tabela 30. Opcje eksportu kategorii

Zawsze eksportowane	Eksportowane opcjonalnie
• Kody kategorii, jeśli istnieją • Nazwy kategorii (i podkategorii) • Poziomy kodów, jeśli istnieją (format *Flat*/*Compact*) • Nagłówki kolumn (format *Flat*/*Compact*)	• Oceny **Docs.** . • Adnotacje kategorii • Nazwy deskryptorów • Liczebności deskryptorów

Ważne! Kiedy eksportowane są deskryptory, są one przekształcane na łańcuchy tekstowe z przedrostkiem w formie podkreślenia. Przy ponownym importowaniu do produktu zostaje utracona zdolność do rozróżnienia pomiędzy deskryptorami, które są wzorcami, regułami kategorii i zwykłymi pojęciami. Jeśli zamierzasz użyć ponownie tych kategorii w tym produkcie, stanowczo zalecamy utworzenie pliku TAP (text analysis package), ponieważ format TAP zachowa wszystkie deskryptory w zdefiniowanej obecnie formie, jak również wszystkie kategorie, kody oraz używane zasoby lingwistyczne. Pliki TAP mogą być używane w programach IBM SPSS Modeler Text Analytics i IBM SPSS Text Analytics for Surveys. Więcej informacji zawiera temat “Korzystanie z pakietów analizy tekstu (TAP)”.

Aby eksportować predefiniowane kategorie

1. W menu interaktywnego pulpitu roboczego wybierz kolejno opcje **Categories > Manage Categories > Export Categories**. Wyświetlony zostaje kreator Export Categories.
2. Wybierz lokalizację i wprowadź nazwę pliku, który będzie eksportowany.
3. W polu tekstowym File Name wprowadź nazwę pliku wyników.
4. Aby wybrać format, do którego będą eksportowane dane kategorii, kliknij przycisk **Next**.
5. Wybierz format spośród następujących:
   - **Flat or Compact list format**: Więcej informacji można znaleźć w temacie “Płaski format listy” na stronie 126. Listy Flat nie zawierają podkategorii. Więcej informacji zawiera temat “Format kompaktowy” na stronie 127. Listy Compact zawierają kategorie hierarchiczne.
   - **Indented format**: Więcej informacji można znaleźć w temacie “Format wcięty” na stronie 128.
6. Aby wybrać treści do eksportowania i aby przejrzeć proponowane dane, kliknij przycisk **Next**.
7. Przejrzyj treści dla eksportowanego pliku.
8. Zaznacz lub usuń zaznaczenie ustawień dodatkowych treści, które mają być eksportowane, takie jak **Annotations** lub **Descriptor names**.
9. Aby eksportować kategorie, kliknij przycisk **Finish**.

Korzystanie z pakietów analizy tekstu (TAP)

Pakiet analizy tekstu (TAP — text analysis package) pełni rolę szablonu kategoryzacji odpowiedzi tekstowych. Korzystając z pakietu TAP, można łatwo przeprowadzić kategoryzację danych tekstowych przy minimalnym zaangażowaniu użytkownika. Pakiet zawiera bowiem gotowe zestawy kategorii i zasoby lingwistyczne potrzebne do szybkiego i automatycznego zakodowania dużej liczby rekordów. Przy użyciu zasobów lingwistycznych dane tekstowe są analizowane i eksplorowane w celu wyodrębnienia kluczowych pojęć. Na podstawie kluczowych pojęć i wzorców znalezionych w tekście można przypisać rekordy do kategorii z zestawu kategorii wybranego w pakiecie TAP. Można utworzyć własny pakiet TAP lub zmodyfikować istniejący.

W skład pakietu TAP wchodzą następujące elementy:

- **Zestaw(y) kategorii**. Zestaw kategorii zasadniczo składa się z predefiniowanych kategorii, kodów kategorii, deskryptorów poszczególnych kategorii oraz nazwy całego zestawu. Deskryptory są elementami lingwistycznymi (pojęciami, typami, wzorcami i regułami), takimi jak termin *cheap* lub wzorzec *good price*. Deskryptory używane są do definiowania kategorii: gdy tekst pasuje do któregokolwiek z deskryptorów kategorii, dokument lub rekord jest umieszczany w tej kategorii.
- **Zasoby lingwistyczne**. Zasoby lingwistyczne to zbiór bibliotek i zaawansowanych zasobów zoptymalizowanych tak, by wyodrębniały kluczowe pojęcia i wzorce. Te wyodrębnione pojęcia i wzorce są z kolei używane jako deskryptory umożliwiające umieszczanie rekordów we właściwych kategoriach należących do zestawu kategorii.

Można utworzyć własny pakiet TAP, zmodyfikować istniejący lub korzystać z gotowych pakietów.

Po wybraniu pakietu TAP i zestawu kategorii, SPSS Modeler Text Analytics może przeprowadzić wyodrębnianie i skategoryzować rekordy.

Uwaga: Pakietów TAP utworzonych w produktach IBM SPSS Text Analytics for Surveys i SPSS Modeler Text Analytics można używać zamiennie. Należy jednak pamiętać, że reguły oceniania w systemie mogą się różnić w produkcie SPSS Modeler Text Analytics , w zależności od tego, czy należy załadować pakiet analizy tekstu (TAP) z produktu SPSS Modeler Text Analytics bezpośrednio, lub czy można załadować TAP z produktu IBM SPSS Text Analytics for Surveys. Zalecane jest użycie pakietów TAP, które są wykonywane w produkcie SPSS Modeler Text Analytics . Pakiety TAP wykonane w produkcie IBM SPSS Text Analytics for Surveys mogą być tworzone za pomocą innej wersji zasobów lingwistycznych.

Tworzenie pakietów analizy tekstu

Zawsze, gdy istnieje sesja z co najmniej jedną kategorią i jakimiś zasobami, można utworzyć pakiet analizy tekstu (TAP) z zawartości otwartej sesji pracy z interaktywnym pulpitem roboczym. Zestaw kategorii i deskryptorów (pojęć, typów, reguł lub wyników wzorców TLA) można przekształcić w pakiet TAP razem z wszystkimi zasobami lingwistycznymi otwartymi w edytorze zasobów.

Widoczny jest język, dla którego zasoby zostały utworzone. Język jest ustawiany na karcie Advanced Resources w widoku Template Editor lub Resource Editor.

Aby utworzyć pakiet analizy tekstu

1. Z menu wybierz kolejno następujące pozycje **File > Text Analysis Packages > Make Package**. Zostanie wyświetlone okno dialogowe Make Package.
2. Przejdź do katalogu, w którym zostanie zapisany pakiet TAP. Domyślnie pakiety są zapisywane w podkatalogu \TAP katalogu instalacyjnego produktu.
3. Do pola **File Name** wprowadź nazwę dla pakietu TAP.
4. Wprowadź etykietę do pola **Package Label**. Po wprowadzeniu nazwy pola ta nazwa automatycznie pojawia się jako etykieta, ale tę etykietę można zmienić.
5. W celu wykluczenia zestawu kategorii z pakietu TAP należy usunąć zaznaczenie pola wyboru **Include**. Dzięki temu kategoria nie zostanie dodana do pakietu. Domyślnie pakiet TAP zawiera jeden zestaw kategorii na pytanie. W pakiecie TAP musi istnieć co najmniej jeden zestaw kategorii
6. Zmień nazwy dowolnych zestawów kategorii. Kolumna **New Category Set** domyślnie zawiera nazwy ogólne wygenerowane poprzez dodanie przedrostka Cat_ do nazwy zmiennej tekstowej. Pojedyncze kliknięcie w komórce powoduje, że nazwa staje się dostępna do edycji. Naciśnięcie klawisza Enter lub kliknięcie w innym miejscu powoduje zastosowanie zmienionej nazwy. Jeśli zmienisz nazwę zestawu kategorii, wówczas nazwa zostanie zmieniona tylko w pakiecie TAP, ale nazwa zmiennej w otwartej sesji nie zostanie zmieniona.
7. W razie potrzeby zmień kolejność zestawów kategorii, używając klawiszy strzałek po prawej stronie tabeli zestawu kategorii.
8. Kliknij opcję **Save**, aby utworzyć pakiet analizy tekstu. Okno dialogowe zostanie zamknięte.

Ładowanie pakietów analizy tekstu

Podczas konfigurowania węzła modelowania eksploracji tekstu należy określić zasoby, które będą używane podczas wyodrębniania. Zamiast wybierać szablon zasobów można wybrać pakiet analizy tekstu (TAP), aby skopiować do węzła nie tylko jego zasoby, ale również zestaw kategorii.

Pakiety TAP są najbardziej interesujące w przypadku interaktywnego tworzenia modelu kategorii, ponieważ można używać zestawu kategorii jako punktu początkowego do kategoryzacji. Podczas tworzenia strumienia uruchamiana jest sesja pracy z interaktywnym pulpitem roboczym i ten zestaw kategorii pojawia się w panelu Categories. Dzięki temu użytkownik może natychmiast oceniać dokumenty i rekordy przy użyciu tych kategorii, a następnie modyfikować, budować i rozszerzać te kategorie aż spełnią potrzeby. Więcej informacji zawiera temat “Metody i strategie tworzenia kategorii” na stronie 96.

Począwszy od wersji 14 kliknięcie opcji **Load** i wybranie pakietu TAP pozwala również wyświetlić język, dla którego zdefiniowano zasoby w tym pakiecie TAP.

Aby załadować pakiet analizy tekstu

1. Przeprowadź edycję węzła modelowania Text Mining.
2. Na karcie Models wybierz opcję *Text analysis package* w sekcji **Copy Resources From**.
3. Kliknij przycisk **Load**. Zostanie otwarte okno dialogowe Load Text Analysis Package.
4. Wyszukaj lokalizację pakietu TAP zawierającego zasoby i zestaw kategorii, który zamierzasz skopiować do węzła. Domyślnie pakiety są zapisywane w podkatalogu \TAP katalogu instalacyjnego produktu.
5. Do pola **File Name** wprowadź nazwę dla pakietu TAP. Etykieta zostanie wyświetlona automatycznie.
6. Wybierz zestaw kategorii, z którego chcesz skorzystać. Jest to zestaw kategorii, który pojawi się w sesji pracy z interaktywnym pulpitem. Następnie możliwe będzie modyfikowanie i korygowanie tych kategorii ręcznie albo przy użyciu opcji kategorii Build lub Extend.
7. Kliknij przycisk **Load**, aby skopiować do węzła zawartość pakietu analizy tekstu. Okno dialogowe zostanie zamknięte. Gdy pakiet TAP zostanie załadowany, jego kopia zostanie umieszczona w węźle; dzięki temu wszelkie zmiany dotyczące zasobów i kategorii nie zostaną uwzględnione w pakiecie TAP, chyba że zostanie on zaktualizowany jawnie, a następnie ponownie załadowany.

Aktualizowanie pakietów analizy tekstu

Jeśli wprowadzisz poprawki do zestawu kategorii, zasobów lingwistycznych lub utworzysz zupełnie nowy zestaw kategorii, wówczas możesz zaktualizować pakiet analizy tekstu (TAP), co ułatwi późniejsze korzystanie z tych poprawek. Aby to zrobić, musisz przejść do otwartej sesji zawierającej informacje, jakie zamierzasz umieścić w pakiecie TAP. Podczas aktualizowania możesz dołączać zestawy kategorii, zastępować zasoby, zmieniać etykietę pakietu, a także zmieniać nazwę/kolejność zestawów kategorii.

Aby zaktualizować pakiet analizy tekstu

1. Z menu wybierz kolejno następujące pozycje **File > Text Analysis Packages > Update Package**. Zostanie wyświetlone okno dialogowe Update Package.
2. Przejdź do katalogu zawierającego pakiet analizy tekstu, który zamierzasz zaktualizować.
3. Do pola **File Name** wprowadź nazwę dla pakietu TAP.
4. Aby zastąpić zasoby lingwistyczne w pakiecie TAP na zasoby z bieżącej sesji, wybierz opcję **Replace the resources in this package with those in the open session**. Zwykle aktualizowanie zasobów lingwistycznych ma sens, ponieważ te zasoby były używane do wyodrębniania kluczowych pojęć i wzorców używanych do tworzenia definicji kategorii. Korzystanie z najnowszych zasobów lingwistycznych zapewnia uzyskanie najlepszych rezultatów klasyfikowania rekordów. Jeśli ta opcja nie zostanie wybrana, wówczas zasoby lingwistyczne znajdujące się już w pakiecie pozostaną niezmienione.
5. Aby zaktualizować tylko zasoby lingwistyczne, wybierz opcję **Replace the resources in this package with those in the open session**, a następnie wybierz tylko bieżące zestawy kategorii, które znajdują się już w TAP.

6. W celu dołączenia nowych zestawów kategorii z otwartej sesji do pakietu TAP zaznacz pole wyboru dla każdej kategorii, którą zamierzasz dodać. Możesz dodać jeden zestaw, wiele zestawów albo możesz w ogóle nie dodawać zestawów kategorii.
7. W celu usunięcia zestawów kategorii z pakietu TAP usuń zaznaczenie pola wyboru **Include**. Możesz usunąć zestaw kategorii, który znajduje się już w pakiecie TAP, ponieważ dodajesz zestaw poprawiony. W tym celu usuń zaznaczenie pola wyboru **Include** odpowiedniego zestawu kategorii w kolumnie Current Category Set. W pakiecie TAP musi istnieć co najmniej jeden zestaw kategorii
8. W razie potrzeby zmień nazwy zestawów kategorii. Pojedyncze kliknięcie w komórce powoduje, że nazwa staje się dostępna do edycji. Naciśnięcie klawisza Enter lub kliknięcie w innym miejscu powoduje zastosowanie zmienionej nazwy. Jeśli zmienisz nazwę zestawu kategorii, wówczas nazwa zostanie zmieniona tylko w pakiecie TAP, ale nazwa zmiennej w otwartej sesji nie zostanie zmieniona. Jeśli dwa zestawy kategorii mają tę samą nazwę, nazwy będą widoczne w kolorze czerwonym, aż do czasu skorygowania zduplikowanej nazwy.
9. W celu utworzenia nowego pakietu z zawartością sesji scaloną z zawartością wybranego pakietu TP kliknij opcję **Save As New**. Zostanie otwarte okno dialogowe Save As Text Analysis Package. Zapoznaj się z poniższymi instrukcjami.
10. Kliknij opcję **Update**, aby zapisać zmiany wprowadzone do wybranego pakietu TAP.

Aby zapisać pakiet analizy tekstu

1. Przejdź do katalogu, w którym zostanie zapisany plik TAP. Domyślnie pliki TAP są zapisywane w podkatalogu TAP katalogu instalacyjnego.
2. Do pola File Name wprowadź nazwę dla pliku TAP.
3. Wprowadź etykietę do pola Package label. Po wprowadzeniu nazwy pola ta nazwa automatycznie pojawia się jako etykieta. Tę etykietę można zmienić. Etykieta jest wymagana.
4. Kliknij opcję **Save**, aby utworzyć nowy pakiet.

Edytowanie i optymalizacja kategorii

Po utworzeniu kategorii zawsze trzeba je sprawdzić i wprowadzić pewne korekty. Oprócz optymalizacji zasobów lingwistycznych należy przejrzeć kategorie, poszukując możliwości połączenia lub wyczyszczenia ich definicji, a także sprawdzić niektóre skategoryzowane dokumenty lub rekordy. Można także przejrzeć dokumenty lub rekordy w kategorii i wprowadzić korekty w taki sposób, by kategorie wychwytywały niuanse i odrębne idee.

Można wykorzystać wbudowane zautomatyzowane techniki budowania kategorii; jednak prawdopodobnie konieczne będzie ręczne skorygowanie efektów działania tych technik. Po zastosowaniu jednej lub większej liczby technik w oknie pojawi się szereg nowych kategorii. Można teraz przeglądać dane w kategorii i wprowadzać korekty do momentu, aż definicje kategorii będą zadowalające. Więcej informacji zawiera temat “Informacje o kategoriach” na stronie 100.

Poniżej wymieniono niektóre możliwości optymalizacji kategorii, z których większość opisano na kolejnych stronach:

Dodawanie deskryptorów do kategorii

Po użyciu technik zautomatyzowanych najprawdopodobniej nadal część wyników wyodrębniania nie zostanie wykorzystana w żadnej z definicji kategorii. Należy przejrzeć tę listę w panelu Extraction Results. Jeśli znajdziesz elementy, które powinny zostać przeniesione do kategorii, możesz je dodać do istniejącej lub nowej kategorii.

Aby dodać pojęcie lub typ do kategorii

1. Z paneli Extraction Results i Data wybierz elementy, które chcesz dodać do nowej lub istniejącej kategorii.
2. Z menu wybierz opcje **Categories > Add to Category**. W oknie dialogowym All Categories zostanie wyświetlony zestaw kategorii. Wybierz kategorię, do której chcesz dodać wybrane elementy. Jeśli chcesz dodać elementy do nowej kategorii, wybierz opcję **New Category**. Nowa kategoria zostanie wyświetlona w panelu Categories, pod nazwą pierwszego wybranego elementu.

Edytowanie deskryptorów kategorii

Po utworzeniu kategorii można otworzyć każdą kategorię, aby zobaczyć wszystkie deskryptory, które tworzą jej definicję. W oknie dialogowym Category Definitions można wprowadzić szereg zmian w deskryptorach kategorii. Ponadto, jeśli kategorie są wyświetlane w drzewie kategorii, można także pracować z nimi w tym oknie.

Aby edytować kategorię

1. Wybierz kategorię, która ma być edytowana, w panelu Categories.
2. Z menu wybierz opcje **View > Category Definitions**. Zostanie otwarte okno dialogowe Category Definitions.
3. Wybierz deskryptor, który chcesz edytować, a następnie kliknij odpowiedni przycisk na pasku narzędzi.

Poniższa tabela zawiera opisy przycisków na pasku narzędzi, których można używać do edytowania definicji kategorii.

Tabela 31. Przyciski paska narzędzi i ich opisy.

Ikony	Opis
	Usuwa wybrane deskryptory z kategorii.
	Przenosi wybrane deskryptory do nowej lub istniejącej kategorii.
	Przenosi wybrane deskryptory w postaci reguły kategorii & do kategorii. Więcej informacji zawiera temat “Korzystanie z reguł kategorii” na stronie 116.
	Powoduje przeniesienie każdego z wybranych deskryptorów do osobnej nowej kategorii.
 **Display**	Aktualizuje zawartość panelu Data i Visualization na podstawie wybranych deskryptorów.

Przenoszenie kategorii

Jeśli chcesz umieścić kategorię w innej istniejącej kategorii lub przenieść deskryptory do innej kategorii, można przenieść kategorię lub deskryptory.

Aby przenieść kategorię

1. W panelu Categories wybierz kategorie , które chcesz przenieść do innej kategorii.
2. Z menu wybierz opcje **Categories > Move to Category**. Menu zawiera zestaw kategorii, a ostatnio utworzona kategoria znajduje się na początku listy. Wybierz nazwę kategorii, do której chcesz przenieść wybrane pojęcia.

- Jeśli widzisz poszukiwaną nazwę, wybierz ją, a wybrane elementy zostaną dodane do tej kategorii.
- Jeśli nazwa nie jest widoczna, wybierz opcję **More**, aby wyświetlić okno dialogowe All Categories, a następnie wybierz kategorię z listy.

Spłaszczanie kategorii

W przypadku struktury hierarchicznej obejmującej kategorie i podkategorie można zdecydować się na spłaszczenie struktury. Spłaszczenie kategorii polega na tym, że wszystkie deskryptory w podkategoriach tej kategorii są przenoszone do wybranej kategorii, a puste podkategorie są usuwane. W ten sposób wszystkie dokumenty, które pasowały do podkategorii, są teraz przypisane do wybranej kategorii.

Aby spłaszczyć kategorię

1. W panelu Categories wybierz kategorię (najwyższego poziomu lub podkategorię), którą chcesz spłaszczyć.

2. Z menu wybierz opcje **Categories > Flatten Categories**. Podkategorie zostaną usunięte i deskryptory zostaną przeniesione do wybranej kategorii.

Scalenie lub łączenie kategorii

Jeśli chcesz połączyć dwie lub więcej istniejących kategorii w nową kategorię, możesz je scalić. Podczas scalania kategorii tworzona jest nowa kategoria o nazwie ogólnej. Wszystkie pojęcia, typy i wzorce używane w deskryptorach kategorii są przenoszone do tej nowej kategorii. Można później zmienić nazwę tej kategorii, edytując jej właściwości.

Aby scalić kategorię lub część kategorii

1. W panelu Categories wybierz elementy, które mają zostać scalone.
2. Z menu wybierz opcje **Categories > Merge Categories**. Zostanie wyświetlone okno dialogowe Category Properties, w którym można wprowadzić nazwę nowo utworzonej kategorii. Wybrane kategorie zostaną scalone w nową kategorię jako podkategorie.

Usuwanie kategorii

Jeśli nie chcesz już zachowywać jakiejś kategorii, możesz ją usunąć.

Aby usunąć kategorię

1. W panelu Categories wybierz kategorię lub kategorie, które chcesz usunąć.
2. Z menu wybierz kolejno opcje **Edit > Delete**.

Rozdział 10. Analiza skupień

Skupienia pojęć można tworzyć i eksplorować w widoku Clusters (**View > Clusters**). *Skupienie* jest grupą pokrewnych pojęć wygenerowaną przez algorytmy tworzenia skupień na podstawie liczebności występowania tych pojęć w zestawie dokumentów/rekordów oraz liczebności ich łącznych wystąpień w tym samym dokumencie (czyli tak zwanych *współwystąpień*). Każde pojęcie w skupieniu współwystępuje z co najmniej jednym innym pojęciem w tym samym skupieniu. Skupienia służą do grupowania pojęć współwystępujących, natomiast kategorie służą do grupowania dokumentów lub rekordów na podstawie dopasowania zawartego w nich tekstu do deskryptorów (pojęć, reguł, wzorców) dla każdej kategorii.

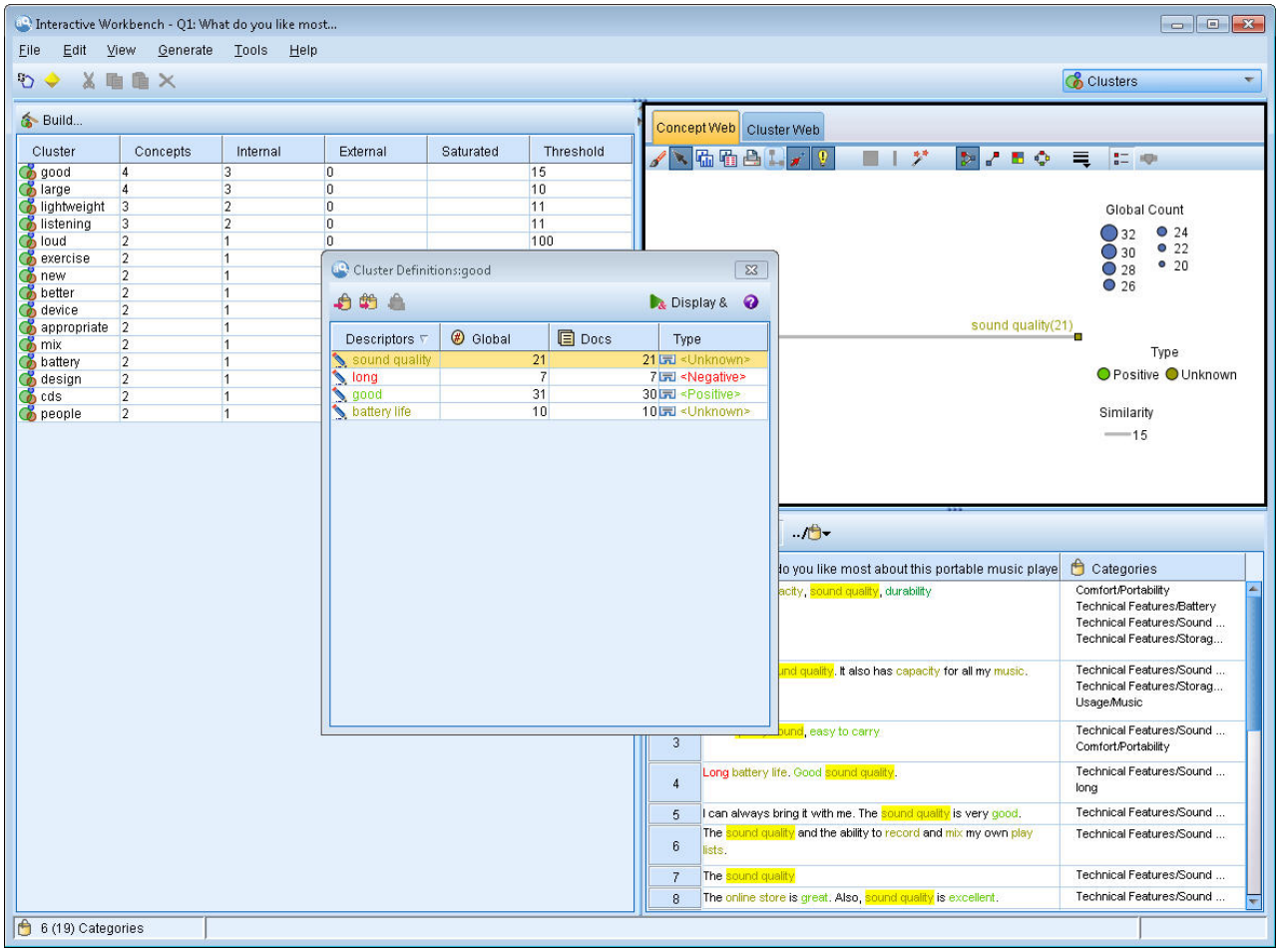
Dobre skupienie to takie, w którym pojęcia są silnie wzajemnie powiązane i często współwystępują, a także mają niewiele powiązań z pojęciami i innymi skupieniami. Podczas pracy z obszernymi zbiorami danych ta technika może spowodować istotne wydłużenie czasu przetwarzania.

Tworzenie skupień rozpoczyna się od analizy zestawu pojęć i wyszukania pojęć, które często współwystępują w dokumentach. Dwa pojęcia, które współwystępują w dokumencie, uznawane są za parę pojęć. Następnie proces tworzenia skupień wyznacza *podobieństwo* każdej pary pojęć, porównując liczbę dokumentów, w której para występuje łącznie, z liczbą dokumentów, w których występuje każde z pojęć. Więcej informacji zawiera temat “Obliczanie wartości powiązań na podstawie podobieństwa” na stronie 137.

Na koniec podobne pojęcia są grupowane w skupienia w drodze agregacji, a algorytm uwzględnia ich wartości powiązań i ustawienia zdefiniowane w oknie dialogowym Build Clusters. Poprzez agregację rozumiemy dodawanie pojęć do skupień lub włączanie mniejszych skupień do większych tak długo, aż skupienie będzie nasycone. Skupienie jest *nasycone*, gdy dodatkowe scalanie pojęć lub mniejszych skupień spowodowałoby przekroczenie przez skupienie ustawień określonych w oknie dialogowym Build Clusters (chodzi o liczbę pojęć, wewnętrzne powiązania lub zewnętrzne powiązania). Skupienie przyjmuje nazwę tego z zawartych w nim pojęć, które ma najwyższą łączną liczbę powiązań z innymi pojęciami w tym skupieniu.

Na koniec nie wszystkie pary pojęć trafiają do jednego skupienia, ponieważ mogą istnieć silniejsze powiązania w innym skupieniu lub nasycenie uniemożliwi scalenie ich skupień. Dlatego istnieją zarówno powiązania wewnętrzne, jak i zewnętrzne.

- *Powiązania wewnętrzne* są to powiązania między parami pojęć w obrębie skupienia. Nie wszystkie pojęcia w skupieniu są ze sobą powiązane. Jednak każde pojęcie jest powiązane z co najmniej jednym innym pojęciem w tym samym skupieniu.
- *Powiązania zewnętrzne* są to powiązania między parami pojęć w odrębnych skupieniach (jedno pojęcie w jednym skupieniu, a drugie pojęcie w innym).



Rysunek 30. Widok Clusters

Widok Clusters jest podzielony na trzy panele, a każdy z nich można ukrywać lub uwidaczniać, wybierając jego nazwę z menu View:

- **Clusters pane** W tym panelu można tworzyć powiązania i zarządzać nimi. Więcej informacji zawiera temat “Eksplorowanie skupień” na stronie 138.
- **Visualization pane** W tym panelu możliwe jest wizualne eksplorowanie skupień i ich interakcji. Więcej informacji zawiera temat “Wykresy skupień” na stronie 149.
- **Data pane** Można eksplorować i przeglądać tekst zawarty w dokumentach i rekordach odpowiadających wyborom dokonanym w oknie dialogowym Cluster Definitions. Więcej informacji zawiera temat “Definicje skupień” na stronie 139.

Tworzenie skupień

Po uzyskaniu po raz pierwszy dostępu do widoku Clusters żadne skupienia nie są widoczne. Można utworzyć skupienia za pomocą menu (**Tools > Build Clusters**) lub klikając przycisk **Build...** na pasku narzędzi. Ta czynność otwiera okno dialogowe Build Clusters, w którym można zdefiniować ustawienia i limity dla budowy skupień.

Uwaga: Gdy wyniki wyodrębniania nie będą już zgodne z zasobami, ten panel stanie się żółty, podobnie jak Panel Extraction Results. Możesz ponowić wyodrębnianie, aby uzyskać najnowsze wyniki wyodrębniania, a żółty kolor zniknie. Jednak po każdym wyodrębnianiu panel Clusters jest czyszczony i konieczne jest odbudowanie skupień. Podobnie, skupienia nie są zachowywane między sesjami.

W oknie dialogowym Build Clusters dostępne są następujące obszary i pola:

Wejścia

Inputs table Skupienia są tworzone na podstawie deskryptorów wywodzonych z określonych typów. W tej tabeli można wybrać typy, które mają zostać uwzględnione w procesie budowania. Typy, które wychwytują najwięcej rekordów lub dokumentów, są wstępnie zaznaczone.

Concepts to cluster: Wybierz metodę wyboru pojęć, które mają być używane do tworzenia skupień. Zmniejszając liczbę pojęć, można przyspieszyć proces tworzenia skupień. Istnieje możliwość utworzenia skupień przy użyciu określonej liczby najlepszych pojęć, odsetka najlepszych pojęć lub wszystkich pojęć:

- **Number based on doc. count** Jeśli wybrano opcję **Top number of concepts**, wprowadź liczbę pojęć, które mają być uwzględniane przy tworzeniu skupień. Wybierane są pojęcia o największej liczbie dokumentów. Liczba dokumentów to liczba dokumentów lub rekordów, w których występuje pojęcie. Maksymalna wartość to 150 000.
- **Percentage based on doc. count** Po wybraniu **Top percentage of concepts** wpisz procent pojęć, które mają być uwzględniane na potrzeby grupowania. Wybierane są pojęcia o największej liczbie dokumentów.

Ograniczenie liczby skupień wyjściowych

Maximum number of clusters to create Ta wartość jest maksymalną liczbą skupień do wygenerowania i wyświetlenia w panelu Clusters. W trakcie tworzenia skupień nasycone skupienia są prezentowane przed nienasyconymi i dlatego wiele utworzonych skupień będzie nasyconych. Aby wyświetlić więcej skupień nienasyconych, można zmienić to ustawienie na wartość większą niż liczba nasyconych skupień.

Maximum concepts in a cluster Ta wartość jest maksymalną liczbą pojęć, jaką może zawierać skupienie.

Minimum concepts in a cluster Ta wartość jest minimalną liczbą pojęć, jaka musi być powiązana, aby powstało skupienie.

Maximum number of internal links Ta wartość jest maksymalną liczbą pojęć, jaką może zawierać skupienie. Powiązania wewnętrzne są to powiązania między parami pojęć w obrębie skupienia.

Maximum number of external links Ta wartość jest maksymalną liczbą powiązań z pojęciami poza skupieniem. Powiązania zewnętrzne są to powiązania między parami pojęć w odrębnych skupieniach.

Minimum link value Najmniejsza wartość powiązania akceptowana dla pary pojęć, które mają być uwzględniane przy tworzeniu skupień. Wartość powiązania jest obliczana za pomocą formuły podobieństwa. Więcej informacji zawiera temat “Obliczanie wartości powiązań na podstawie podobieństwa”.

Prevent pairing of specific concepts. Zaznacz to pole wyboru, aby w wynikach nie grupować lub nie łączyć w pary dwóch pojęć. Aby tworzyć pary pojęć lub nimi zarządzać, kliknij przycisk **Manage Pairs**. Więcej informacji zawiera temat “Zarządzanie parami wyjątków powiązań” na stronie 106.

Obliczanie wartości powiązań na podstawie podobieństwa

Sama tylko liczba dokumentów, w jakiej para pojęć współwystępuje, nie mówi nic o podobieństwie tych pojęć. W takich przypadkach przydatna jest wartość podobieństwa. Wartość powiązania na podstawie podobieństwa wyznacza się, porównując liczebność dokumentów ze współwystąpieniami do liczebności dokumentów dla każdego pojęcia uczestniczącego w relacji. Jednostką miary podobieństwa jest liczba dokumentów, w których występuje pojęcie lub para pojęć. Pojęcie lub para pojęć „występuje” w dokumencie, jeśli zawarte/zawarta jest *co najmniej* w tym dokumencie. Wybierając odpowiednią opcję można spowodować, że grubość linii na wykresie pojęć będzie odzwierciedlała wartość powiązania na podstawie podobieństwa.

Algorytm ujawnia najsilniejsze relacje, tj. takie, w których tendencja do współwystępowania pojęć w danych jest znacznie silniejsza niż tendencja do ich pojedynczego występowania. Wewnętrznie algorytm oblicza współczynnik podobieństwa z zakresu od 0 do 1, gdzie 1 oznacza, że dwa pojęcia zawsze występują razem, a nigdy osobno. Wynikowy współczynnik podobieństwa jest mnożony przez 100 i zaokrąglany do najbliższej liczby całkowitej.

Współczynnik podobieństwa jest obliczany za pomocą wzoru przedstawionego na poniższym rysunku.

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

Rysunek 31. Wzór na współczynnik podobieństwa

Gdzie:

- C_I to liczba dokumentów lub rekordów, w których występuje pojęcie I.
- C_J to liczba dokumentów lub rekordów, w których występuje pojęcie J.
- C_{IJ} to liczba dokumentów lub rekordów, w których para pojęć I i J współwystępuje w zestawie dokumentów.

Załóżmy na przykład, że mamy 5000 dokumentów. Niech I oraz J będą wyodrębnionymi pojęciami, a IJ niech będzie współwystąpieniem pary I i J. W następującej tabeli zaproponowano dwa scenariusze obliczania współczynnika i wartości powiązania.

Tabela 32. Przykład liczebności pojęć

Pojęcie/para	**Scenariusz A**	**Scenariusz B**
Pojęcie: I	Występuje w 20 dokumentach	Występuje w 30 dokumentach
Pojęcie: J	Występuje w 20 dokumentach	Występuje w 60 dokumentach
Para pojęć: IJ	Współwystępuje w 20 dokumentach	Współwystępuje w 20 dokumentach
Współczynnik podobieństwa	1	0,22222
Wartość powiązania na podstawie podobieństwa	100	22

W scenariuszu A pojęcia I oraz J, a także para IJ występują w 20 dokumentach, co przekłada się na współczynnik podobieństwa 1, który oznacza, że pojęcia te zawsze występują razem. Wartość powiązania dla tej pary wynosi 100.

W scenariuszu B pojęcie I występuje w 30 dokumentach, a pojęcie J występuje w 60 dokumentach, ale para IJ występuje tylko w 20 dokumentach. W rezultacie współczynnik podobieństwa wynosi 0,22222. Wartość powiązania dla tej pary zostanie zaokrąglona w dół do 22.

Eksplorowanie skupień

Po zakończeniu tworzenia skupień można wyświetlić zestaw wyników w panelu Clusters. Dla każdego skupienia dostępne są następujące informacje w tabeli:

- **Grupowanie.** Jest to nazwa skupienia. Skupieniom nadawane są nazwy na podstawie pojęcia z największą liczbą powiązań wewnętrznych.
- **Concepts.** Jest to liczba pojęć w skupieniu. Więcej informacji zawiera temat “Definicje skupień” na stronie 139.
- **Internal.** Jest to liczba powiązań wewnętrznych w skupieniu. Powiązania wewnętrzne są to powiązania między parami pojęć w obrębie skupienia.
- **External.** Jest to liczba powiązań zewnętrznych w skupieniu. Powiązania zewnętrzne są to powiązania między parami pojęć należących do różnych skupień.
- **Sat.** Obecność symbolu wskazuje, że skupienie mogłoby być większe, ale przekroczyło co najmniej jeden limit i proces rozbudowy tego skupienia został przerwany, a skupienie jest *nasycone*. Na końcu procesu tworzenia skupień nasycone skupienia są prezentowane przed nienasyconymi, dlatego wiele utworzonych skupień będzie nasyconych. Aby wyświetlić więcej skupień nienasyconych, można zmienić ustawienie **Maximum number of clusters to create** na wartość większą niż liczba nasyconych skupień lub zmniejszyć ustawienie **Minimum link value**. Więcej informacji zawiera temat “Tworzenie skupień” na stronie 136.

- **Threshold.** Dla wszystkich par współwystępujących w skupieniu jest to najniższa w skupieniu wartość powiązania opartego na podobieństwie. Więcej informacji zawiera temat “Obliczanie wartości powiązań na podstawie podobieństwa” na stronie 137. Skupienie z wysokim progiem oznacza, że pojęcia w tym skupieniu mają wyższe ogólne podobieństwo i są bliżej spokrewnione niż pojęcia w skupieniu o niższym progu.

Aby uzyskać więcej informacji o skupieniu, można je wybrać, a wówczas na panelu wizualizacji po prawej stronie pojawią się dwa wykresy służące do jego eksploracji. Więcej informacji zawiera temat “Wykresy skupień” na stronie 149. Można także wyciąć zawartość tabeli i wkleić ją do innej aplikacji.

Gdy wyniki wyodrębniania nie będą już zgodne z zasobami, ten panel stanie się żółty, podobnie jak Panel Extraction Results. Możesz ponowić wyodrębnianie, aby uzyskać najnowsze wyniki wyodrębniania, a żółty kolor zniknie. Jednak po każdym wyodrębnianiu panel Clusters jest czyszczony i konieczne jest odbudowanie skupień. Podobnie, skupienia nie są zachowywane między sesjami.

Definicje skupień

Można wyświetlić wszystkie pojęcia należące do skupienia, wybierając to skupienie w panelu Clusters i otwierając okno dialogowe Cluster Definitions (**View > Cluster Definitions**).

Wszystkie pojęcia z wybranego skupienia pojawią się w oknie dialogowym Cluster Definitions. Jeśli w oknie dialogowym Cluster Definitions zaznaczysz jedno lub więcej pojęć i klikniesz przycisk **Display &**, na panelu Data zostaną wyświetlone wszystkie rekordy lub dokumenty, w których *wszystkie zaznaczone pojęcia występują wspólnie*. Jednak na panelu Data nie będą wyświetlane rekordy ani dokumenty po wybraniu skupienia w panelu Clusters. Aby uzyskać ogólne informacje o panelu Data, patrz “Panel Data” na stronie 101.

Wybranie pojęcia w tym oknie dialogowym wpływa także na wykres sieciowy pojęć. Więcej informacji zawiera temat “Wykresy skupień” na stronie 149. Podobnie, jeśli w oknie dialogowym Cluster Definitions zaznaczysz jedno lub więcej pojęć, na panelu Visualization zostaną wyświetlone wszystkie powiązania zewnętrzne i wewnętrzne tych pojęć.

Opisy kolumn

Ikony umożliwiają łatwą identyfikację deskryptorów.

Tabela 33. Kolumny i ikony deskryptorów

Kolumny	Opis
Descriptors	Nazwa pojęcia.
 **Global**	Wyświetla, ile razy ten deskryptor pojawia się w całym zbiorze danych, wartość nazywana również liczebnością globalną.
 **Docs**	Liczba dokumentów lub rekordów, w których ten deskryptor występuje, nazywana również liczebnością dokumentów.
Typ	Wyświetla typ lub typy, do których należy dany deskryptor. Jeśli deskryptor jest regułą kategorii, żadna nazwa typu nie jest wyświetlana w tej kolumnie.

Działania na pasku narzędzi

W tym oknie dialogowym można też wybrać jedno lub wiele pojęć, które mają być użyte w kategorii. Istnieje kilka sposobów wykonania tej czynności, ale najbardziej interesująca jest możliwość wybrania pojęć współwystępujących w skupieniu i dodania ich jako reguły kategorii. Więcej informacji zawiera temat “Reguły współwystępowania” na stronie 110. Można używać przycisków paska narzędzi, aby dodawać pojęcia do kategorii.

Tabela 34. Przyciski na pasku narzędzi służące do dodawania pojęć do kategorii

Ikony	Opis
	Dodaj wybrane pojęcia do nowej lub istniejącej kategorii.
	Dodaj wybrane pojęcia w postaci reguły kategorii & do nowej lub istniejącej kategorii. Więcej informacji zawiera temat “Korzystanie z reguł kategorii” na stronie 116.
	Dodaj każde z wybranych pojęć jako osobną nową kategorię.
	Aktualizuje zawartość panelu Data i Visualization na podstawie wybranych deskryptorów.

Uwaga: Można także dodać pojęcia do typu, jako synonimy, lub wykluczyć elementy, używając menu kontekstowych.

Rozdział 11. Eksplorowanie analizy powiązań w tekście

W widoku Text Link Analysis (TLA) można eksplorować wzorce będące wynikami analizy powiązań w tekście. Analiza powiązań w tekście jest to technika dopasowywania wzorców, która umożliwia definiowanie reguł wzorców i porównywanie ich z faktycznie wyodrębnionymi pojęciami i relacjami występującymi w tekście.

Na przykład samo wyodrębnienie pojęć dotyczących organizacji może nie dostarczyć wartościowej wiedzy. Korzystając z analizy TLA, możemy także uzyskać informacje o powiązaniach między tą organizacją a innymi organizacjami lub osobami wewnątrz organizacji. Można też wykorzystać analizę TLA do wyodrębniania opinii o produktach, a w niektórych językach także relacji między genami.

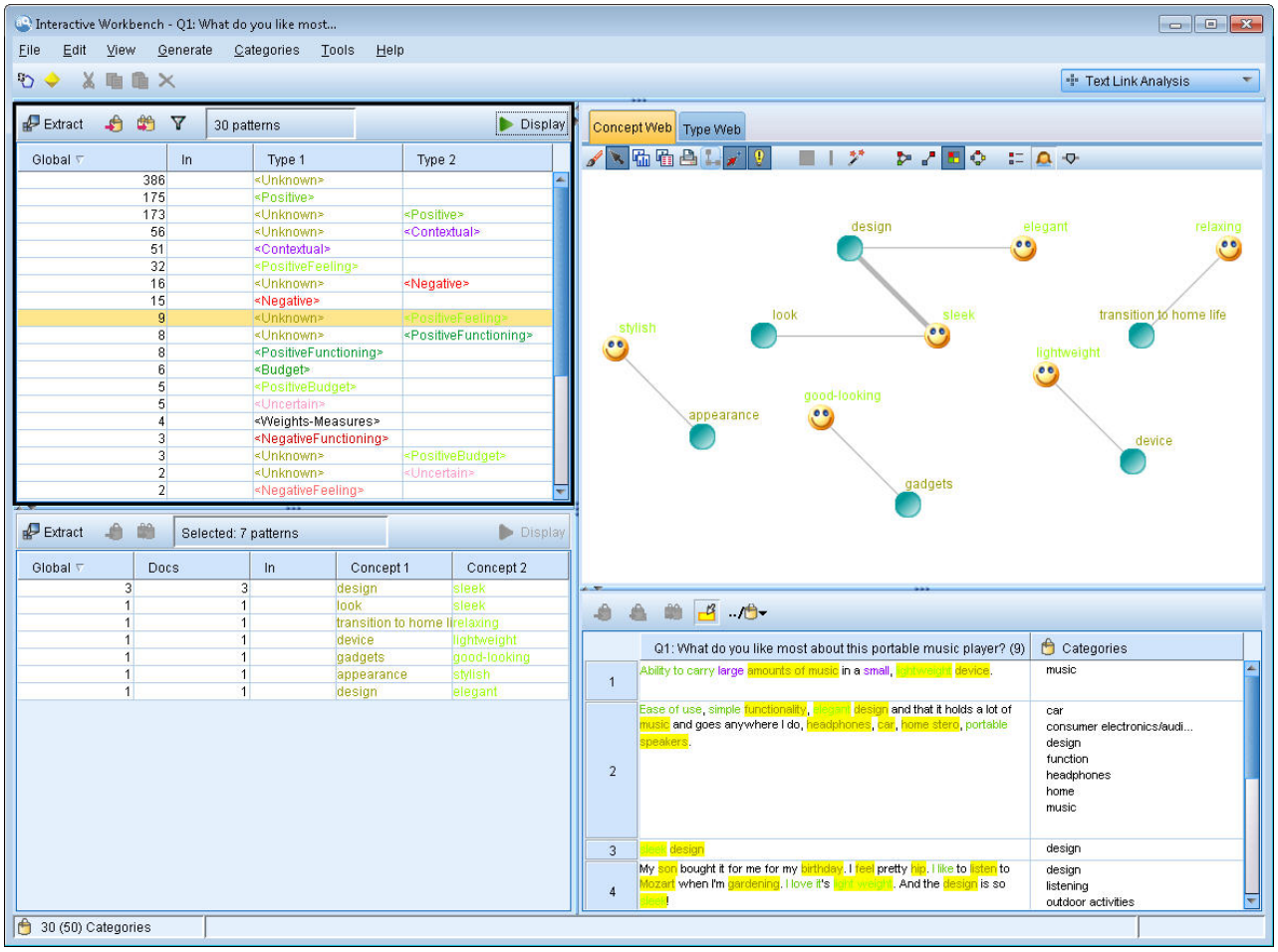
Wynikowe wzorce TLA można przeglądać na panelach Type i Concept Patterns w widoku Text Link Analysis. Więcej informacji zawiera temat “Wzorce typów i pojęć” na stronie 143. Dalsza eksploracja wyników jest możliwa w panelach Data i Visualization w tym widoku. Prawdopodobnie najważniejsza jest możliwość dodawania ich do kategorii.

Jeśli ta czynność nie została jeszcze wykonana, możesz kliknąć przycisk **Extract** i wybrać opcję **Enable Text Link Analysis pattern extraction** w oknie dialogowym Extract Settings. Więcej informacji zawiera temat “Wyodrębnianie wynikowych wzorców TLA” na stronie 142.

Aby możliwe było wyodrębnienie wynikowych wzorców TLA, w używanym szablonie zasobów lub w używanych bibliotekach muszą być zdefiniowane reguły wzorców TLA. Można używać wzorców TLA z niektórych szablonów zasobów dostarczanych razem z produktem IBM SPSS Modeler Text Analytics. Rodzaj relacji i wzorców, jakie można wyodrębniać, zależy wyłącznie od reguł TLA zdefiniowanych w zasobach. Możesz definiować własne reguły TLA dla wszystkich języków *z wyjątkiem* japońskiego. Wzorce składają się z makr, list wyrazów i odstępów między wyrazami, które z kolei składają się na zapytanie boolowskie lub regułę stosowaną wobec tekstu wejściowego. Więcej informacji zawiera Rozdział 18, “Informacje o regułach powiązań w tekście”, na stronie 201.

Za każdym razem, gdy reguła wzorca TLA znajdzie pasujący tekst, tekst ten może być wyodrębniony jako wzorzec i zrestrukturyzowany do postaci danych wynikowych. Wyniki są widoczne w panelach widoku Text Link Analysis. Każdy z paneli można ukrywać lub uwidaczniać, wybierając jego nazwę z menu View:

- **Panele Type i Concept Patterns.** W tych panelach można tworzyć i eksplorować wzorce. Więcej informacji zawiera temat “Wzorce typów i pojęć” na stronie 143.
- **Panel Visualization.** W tym panelu można wizualnie eksplorować interakcje między pojęciami i typami we wzorcach. Więcej informacji zawiera temat “Wykresy analizy powiązań w tekście” na stronie 150.
- **Panel Data.** Ten panel służy do eksplorowania i przeglądania tekstu zawartego w dokumentach i rekordach odpowiadających wyborom dokonanym w innym panelu. Więcej informacji zawiera temat “Panel Data” na stronie 145.



Rysunek 32. Widok Text Link Analysis

Wyodrębnianie wynikowych wzorców TLA

Proces wyodrębniania powoduje powstanie zestawu pojęć i typów, jak również wzorców TLA (Text Link Analysis), jeśli są włączone. Wyodrębnione wzorce są widoczne w widoku Text Link Analysis. Jeśli wyniki wyodrębniania nie są zsynchronizowane zasobami, panel Patterns przyjmuje kolor żółty, co sygnalizuje że ponowne wyodrębnienie dałoby inne wyniki.

Należy wybrać opcję wyodrębniania wzorców w ustawieniach węzła lub w oknie dialogowym Extract za pomocą opcji **Enable Text Link Analysis pattern extraction**. Więcej informacji zawiera temat “Wyodrębnianie danych” na stronie 80.

Uwaga: Istnieje relacja pomiędzy wielkością zbioru danych i czasem, jaki jest wymagany do zakończenia procesu wyodrębniania. Zobacz instrukcje dotyczące instalacji, aby uzyskać informacje o statystykach wydajności i zaleceniach. Zawsze można rozważyć wstawienie węzła Sample we wcześniejszej części strumienia lub optymalizację konfiguracji komputera.

Aby wyodrębnić dane

1. Z menu wybierz opcje **Tools > Extract**. Lub kliknij przycisk **Extract** na pasku.

2. Zmień dowolne z opcji, których chcesz użyć. Pamiętaj, że opcja **Enable Text Link Analysis pattern extraction** musi być wybrana na tej karcie, a szablon musi zawierać reguły TLA. W przeciwnym razie nie będzie możliwe wyodrębnianie wynikowych wzorców TLA. Więcej informacji zawiera temat “Wyodrębnianie danych” na stronie 80.
3. Kliknij przycisk **Extract**, aby rozpocząć proces wyodrębniania.

Po rozpoczęciu wyodrębniania otwiera się okno dialogowe postępu. Jeśli chcesz przerwać wyodrębnianie, kliknij przycisk **Cancel**. Po zakończeniu wyodrębniania okno dialogowe zostanie zamknięte, a wyniki pojawią się w panelu. Więcej informacji zawiera temat “Wzorce typów i pojęć”.

Wzorce typów i pojęć

Wzorzec składa się z dwóch części: kombinacji pojęć i wbudowanych typów. Wzorce są najbardziej użyteczne wtedy, gdy próbujemy poznać opinie na określony temat lub relacje między pojęciami. Na przykład samo wyodrębnienie nazwy produktu naszego konkurenta może nie dostarczyć wartościowej wiedzy. Możemy jednak przyjrzeć się wyodrębnionym wzorcom, aby sprawdzić, czy dokument lub rekord zawiera wyrażenia oceniające produkt jako dobry, zły lub drogi.

Wzorce mogą składać się maksymalnie z sześciu typów lub sześciu pojęć. Dlatego wiersze w obu panelach wzorców podzielone są na sześć pozycji. Każda pozycja odpowiada konkretnej pozycji elementu w regule wzorca TLA zdefiniowanej w zasobach lingwistycznych. W interaktywnym pulpicie roboczym pozycje bez wartości nie są wyświetlane w tabeli. Na przykład, jeśli najdłuższy wzorzec wynikowy zawiera nie więcej niż cztery pozycje, to dwie ostatnie nie są wyświetlane. Więcej informacji zawiera temat Rozdział 18, “Informacje o regułach powiązań w tekście”, na stronie 201.

Wzorce są najpierw grupowane na poziomie typu, a następnie dzielone na wzorce pojęć. Z tego powodu istnieją dwa różne panele wyników: **Type Patterns** (po lewej stronie na górze) i **Concept Patterns** (po lewej stronie na dole). Aby wyświetlić wszystkie zwrócone wzorce pojęć, wybierz wszystkie wzorce typów. W dolnym panelu pojęć pojawią się wtedy wszystkie wzorce pojęć aż do maksymalnej rangi (zdefiniowanej w oknie dialogowym Filter).

Type Patterns Ten panel przedstawia wynikowe wzorce składające się z jednego lub wielu powiązanych typów i pasujące do reguły TLA. Wzorce typu są przedstawione jako <Organization> + <Location> + <Positive>, co może zapewniać pozytywną opinię o organizacji w określonej lokalizacji. Składnia jest następująca

```
<Typ1> + <Typ2> + <Typ3> + <Typ4> + <Typ5> + <Typ6>
```

Concept Patterns Panel ten prezentuje wynikowe wzorce na poziomie pojęć dla wszystkich wzorców typów wybranych w tym momencie w znajdującym się wyżej panelu Type Patterns. Wzorce pojęć mają np. taką strukturę: hotel + paris + wonderful. Składnia jest następująca

```
pojęcie1 + pojęcie2 + pojęcie3 + pojęcie4 + pojęcie5 + pojęcie6
```

Gdy wynikowe wzorce zajmują mniej niż sześć pozycji (maksymalną liczbę), wyświetlane są tylko potrzebne pozycje (lub kolumny). Puste pozycje między zapełnionymi pozycjami są odrzucane, zatem wzorzec <Typ1>+<>+<Typ2>+<>+<>+<> może być przedstawiony jako <Typ1>+<Typ3>. W przypadku wzorca pojęć: pojęcie1+.+pojęcie2 (gdzie . oznacza wartość null).

Tak jak w przypadku wyników wyodrębniania w widoku Categories and Concepts możesz tu przejrzeć wyniki. Jeśli chcesz doprecyzować typy i pojęcia, z których składają się te wzorce, można to zrobić w panelu Extraction Results widoku Categories and Concepts lub bezpośrednio w narzędziu Resource Editor, a następnie wyodrębnić ponownie wzorce. Gdy pojęcie, typ lub wzorzec jest używany w definicji kategorii lub jest częścią reguły, w kolumnie **In** tabeli Pattern lub Extraction Results wyświetlana jest ikona kategorii lub ikony.

Uwaga: Jeśli w widocznym panelu mieści się więcej wyników, można użyć przycisków znajdujących się w dolnej części panelu, aby przeglądać wyniki, lub wprowadzić numer strony i przejść do niej.

Filtrowanie wyników analizy TLA

Podczas pracy na bardzo obszernych zbiorach danych proces wyodrębniania może wygenerować miliony wyników. Wielu użytkownikom tak obszerne wyniki utrudnią lub uniemożliwią analizę. Możesz jednak odfiltrować wyniki, aby skupić się na najbardziej interesujących. Zmieniając ustawienia w oknie dialogowym Filter, możesz ograniczyć zbiór wyświetlanych wzorców. Wszystkie te ustawienia są używane łącznie.

W widoku TLA okno dialogowe Filter zawiera następujące obszary i pola.

Filter by Frequency Można odfiltrować wyniki tak, by były wyświetlane tylko wyniki o określonej liczebności globalnej lub liczebności dokumentów.

- **Global frequency** to łączna liczba wystąpień wzorca w całym zbiorze dokumentów i rekordów. Podana jest w kolumnie **Global**.
- **Document frequency** to łączna liczba dokumentów lub rekordów, w których występuje wzorzec. Podana jest w kolumnie **Docs**.

Na przykład, jeśli wzorzec występuje 300 razy w 500 rekordach, to globalna liczebność tego wzorca wynosiłaby 300, a liczebność dokumentów wynosiłaby 500.

And by Match Text Można odfiltrować wyniki tak, by wyświetlane były tylko wyniki zgodne ze zdefiniowaną tutaj regułą. W polu **Match text** wprowadź zbiór znaków, który ma być dopasowywany, po czym wybierz, czy ten tekst ma być wyszukiwany w pojęciach, czy w nazwach typów, wskazując numer pozycji lub wszystkie pozycje. Następnie wybierz warunek stosowania dopasowania (nie trzeba używać nawiasów kątowych do oznaczania początku i końca nazwy typu). Z listy rozwijanej wybierz **And** albo **Or**, aby reguła dopasowywała oba wyrażenia lub tylko jedno z nich, po czym zdefiniuj drugie wyrażenie dopasowujące w taki sam sposób, jak pierwsze.

Tabela 35. Warunki dopasowywania tekstu

Warunek	Opis
Contains	Tekst zostanie dopasowany, jeśli łańcuch wystąpi w dowolnym miejscu. (Opcja domyślna)
Starts with	Tekst zostanie dopasowany, jeśli pojęcie lub nazwa typu zaczyna się od określonego tekstu.
Ends with	Tekst zostanie dopasowany, jeśli pojęcie lub nazwa typu kończy się określonym tekstem.
Exact Match	Cały tekst musi być dopasowany do pojęcia lub nazwy typu.

Wyniki wyświetlane w panelu Patterns

Załóżmy że używamy angielskiej wersji oprogramowania; oto niektóre przykłady wyników na pasku narzędzi panelu Patterns w zależności od ustawień filtrowania.

Rysunek 33. 1. przykład filtrowania wyników

W tym przykładzie na pasku narzędzi widzimy, że liczba zwróconych wzorców została ograniczona z uwagi na maksymalną rangę określoną w filtrze. Fioletowa ikona oznacza, że osiągnięto maksymalną liczbę wzorców. Zatrzymaj wskaźnik nad ikoną, aby wyświetlić więcej informacji. Patrz wcześniejsze objaśnienia do filtru **And by Rank**.

Rysunek 34. 2. przykład filtrowania wyników

W tym przykładzie na pasku narzędzi widać, że liczba wyników została ograniczona przez filtr dopasowujący tekst (ikona lupy). Można zatrzymać wskaźnik nad ikoną, aby wyświetlić dopasowywany tekst.

Aby filtrować wyniki

1. Z menu wybierz opcje **Tools > Filter**. Zostanie otwarte okno dialogowe Filter.
2. Wybierz i doprecyzuj filtry, których chcesz używać.
3. Kliknij przycisk **OK**, aby zastosować filtry i wyświetlić nowe wyniki.

Panel Data

Podczas wyodrębniania i eksplorowania wzorców analizy powiązań w tekście trzeba czasem przejrzeć źródłowe dane. Na przykład po to, aby zobaczyć treść rekordów, w których wykryto grupę wzorców. Istnieje możliwość przeglądania rekordów lub dokumentów w panelu Data, który znajduje się w prawym dolnym rogu. Jeśli domyślnie panel ten nie jest wyświetlony, z menu wybierz kolejno opcje **View > Panes > Data**.

W panelu Data wyświetlany jest jeden wiersz na każdy dokument lub rekord odpowiadający wyborom dokonanym w widoku, przy czym liczba wierszy nie może przekroczyć określonego limitu. Domyślnie liczba dokumentów lub rekordów wyświetlanych w panelu Data jest ograniczona, aby możliwe było szybsze wyświetlanie danych. Można jednak zmienić tę liczbę w oknie dialogowym Options. Więcej informacji zawiera temat “Okno dialogowe Options: karta Session” na stronie 75.

Uwaga: Jeśli w widocznym panelu mieści się więcej wyników, można użyć przycisków znajdujących się w dolnej części panelu, aby przeglądać wyniki, lub wprowadzić numer strony i przejść do niej.

Wyświetlanie i odświeżanie panelu Data

Panel Data nie jest odświeżany automatycznie, ponieważ w przypadku większych zbiorów danych automatyczne odświeżanie mogłoby być czasochłonne. Zatem za każdym razem, gdy wybierzesz w tym widoku wzorzec typu lub pojęcia, możesz kliknąć przycisk **Display**, aby odświeżyć zawartość panelu Data.

Dokumenty lub rekordy tekstowe

Jeśli dane tekstowe mają postać rekordów, a tekst jest stosunkowo krótki, to zmienna tekstowa w panelu Data zawiera całe dane tekstowe. Jednak podczas pracy z rekordami i większymi zbiorami danych w kolumnie zmiennej tekstowej wyświetlany jest krótki fragment tekstu, a w panelu Text Preview po prawej stronie wyświetlana jest większa część lub całość tekstu z rekordu zaznaczonego w tabeli. Jeśli dane tekstowe mają postać odrębnych dokumentów, to w panelu Data wyświetlana jest nazwa pliku dokumentu. Po wybraniu dokumentu otwierany jest panel Text Preview z tekstem zaznaczonego dokumentu.

Kolory i wyróżnienia

W wyświetlanych danych pojęcia i deskryptory znalezione w dokumentach lub rekordach są wyróżniane kolorami, aby łatwiej było je odszukać w tekście. Kolory odpowiadają typom pojęć. Można także zatrzymać wskaźnik myszy nad elementem oznaczonym kolorem, aby wyświetlić nazwę pojęcia, pod którym dany termin został wyodrębniony, oraz typ, do którego został przypisany. Tekst niewyodrębniony jest wyświetlany kolorem czarnym. Te niewyodrębnione wyrazy są zwykle spójnikami (*and* lub *with*), zaimkami (*me* lub *they*) oraz czasownikami (*is*, *have* lub *take*).

Kolumny panelu Data

Kolumna zmiennej tekstowej jest widoczna zawsze, ale można też wyświetlić inne kolumny. Aby wyświetlić inne kolumny, wybierz z menu opcje **View > Data Pane**, a następnie wybierz kolumnę, którą chcesz wyświetlić w panelu Data. Do wyświetlania mogą być dostępne następujące kolumny:

- **"Text field name" (#)/Documents** Dodaje kolumnę z danymi tekstowymi, z których wyodrębnione były pojęcia i typy. Jeśli dane są zawarte w dokumentach, kolumna nosi nazwę Dokumenty i widoczna jest tylko nazwa dokumentu lub pełna ścieżka. Właściwy tekst z dokumentów jest widoczny w panelu Text Preview. Po nazwie tej kolumny podana jest w nawiasach liczba wierszy w panelu Data. Niekiedy nie będą wyświetlane wszystkie dokumenty lub rekordy. Wynika to z limitu ustawionego w oknie dialogowym Options, który wprowadza się w celu

przyspieszenia ładowania. Po osiągnięciu limitu po liczbie pojawi się dopisek **- Max**. Więcej informacji zawiera temat “Okno dialogowe Options: karta Session” na stronie 75.

- **Categories** Zawiera listę wszystkich kategorii, do których należy rekord. Gdy ta kolumna jest widoczna, odświeżanie panelu Data w celu wyświetlenia najbardziej aktualnych informacji może trwać nieco dłużej.
- **Relevance Rank** Zawiera rangi poszczególnych rekordów w jednej kategorii. Ranga informuje o tym, jak dobrze dany rekord pasuje do kategorii na tle innych rekordów w tej samej kategorii. Aby wyświetlić rangę, wybierz kategorię w panelu Categories (lewym górnym). Więcej informacji zawiera temat “Istotność kategorii” na stronie 102.
- **Category Count** Zawiera liczbę kategorii, do których należy rekord.

Rozdział 12. Wykresy wizualizacji

Widok Categories and Concepts, widok Clusters i widok Text Link Analysis wszystkie mają panel wizualizacji w prawym górnym rogu okna. Można użyć tego panelu, aby wizualnie eksplorować dane. Dostępne są następujące wykresy.

- **Widok Categories and Concepts.** Ten widok ma trzy wykresy: *Category Bar*, *Category Web* i *Category Web Table*. W tym widoku wykresy są aktualizowane tylko po kliknięciu opcji **Display**. Więcej informacji zawiera temat "Wykresy i tabele kategorii".
- **Widok Clusters.** Ten wykres zawiera dwa wykresy sieciowe: *Concept Web Graph* i *Cluster Web Graph*. Więcej informacji zawiera temat "Wykresy skupień" na stronie 149.
- **Widok Text Link Analysis.** Ten wykres zawiera dwa wykresy sieciowe: *Concept Web Graph* i *Type Web Graph*. Więcej informacji zawiera temat "Wykresy analizy powiązań w tekście" na stronie 150.

Więcej informacji o wszystkich ogólnych paskach narzędzi i paletach używanych do edycji wykresów można znaleźć w sekcji dotyczącej edycji wykresów pomocy online lub w pliku *ModelerSPOnodes.pdf*, który jest dostępny jako część pobranego produktu.

Wykresy i tabele kategorii

Podczas budowania kategorii należy dokładnie weryfikować ich definicje, zawarte w nich dokumenty lub rekordy oraz nakładanie się kategorii. W panelu wizualizacji można przedstawiać kategorie z różnej perspektywy. Panel Visualization znajduje się w prawym górnym rogu widoku Categories and Concepts . Jeśli panel nie jest jeszcze widoczny, można uzyskać do niego dostęp z menu View (**View > Panes > Visualization**).

W tym widoku panel wizualizacji oferuje trzy sposoby prezentacji wspólnych elementów kategoryzacji dokumentów lub rekordów . Diagramy i wykresy wyświetlane w tym panelu umożliwiają analizowanie wyników kategoryzacji i pomagają w optymalizacji kategorii lub raportów. Podczas optymalizacji kategorii można używać tego panelu do weryfikowania definicji kategorii, by wykryć kategorie zbyt podobne (na przykład mające więcej niż 75% wspólnych dokumentów lub rekordów) lub zbyt odmienne. Jeśli dwie kategorie są zbyt podobne, celowe może być połączenie ich w jedną. Można też zoptymalizować definicje kategorii, usuwając niektóre deskryptory z jednej lub drugiej kategorii.

W zależności od wyboru dokonanego w panelu Extraction Results, panelu Categories lub oknie dialogowym Category Definitions, można przeglądać interakcje między dokumentami/rekordami i kategoriami na każdej z kart tego panelu. Każda karta zawiera podobne informacje, ale zaprezentowane w inny sposób i na innym poziomie szczegółowości. Aby jednak odświeżyć wykres z uwzględnieniem obecnie dokonanych wyborów, kliknij przycisk **Display** na pasku narzędzi panelu lub okna dialogowego, w którym tego wyboru dokonano.

Panel Visualization w widoku Categories and Concepts udostępnia następujące wykresy i diagramy:

- **Category Bar Chart.** Tabela z wykresem słupkowym prezentującym pokrywanie się dokumentów/rekordów odpowiadających dokonanemu wyborowi oraz związanych z nimi kategorii. Na wykresie słupkowym widoczne są także stosunki dokumentów/rekordów w kategoriach do łącznej liczby dokumentów/rekordów. Więcej informacji zawiera temat "Wykres słupkowy kategorii" na stronie 148.
- **Category Web Graph.** Ten wykres przedstawia pokrywanie się dokumentów/rekordów dla kategorii, do których dokumenty/rekordy należą, zgodnie z wyborami dokonanymi w innych panelach. Więcej informacji zawiera temat "Wykres sieciowy kategorii" na stronie 148.
- **Category Web Table.** Ta tabela przedstawia te same informacje, które widoczne są na karcie Category Web, ale w postaci tabelarycznej. Tabela zawiera trzy kolumny, które można sortować, klikając ich nagłówki. Więcej informacji zawiera temat "Tabela sieciowa kategorii" na stronie 148.

Więcej informacji zawiera temat Rozdział 9, "Kategoryzacja danych tekstowych", na stronie 93.

Wykres słupkowy kategorii

Ta karta zawiera tabelę i wykres słupkowy przedstawiający nakładanie się dokumentów/rekordów odpowiadających bieżącemu wyborowi i powiązanym kategoriom. Na wykresie słupkowym widoczne są także stosunki dokumentów/rekordów w kategoriach do łącznej liczby dokumentów/rekordów . Nie można edytować układu tego wykresu. Można jednak posortować kolumny, klikając nagłówki kolumn.

Wykres zawiera następujące kolumny:

- **Category.** Ta kolumna zawiera wybranych nazwy kategorii. Domyślnie jako pierwsza wyświetlana jest najczęściej występująca kategoria.
- **Bar.** Ta kolumna pokazuje, w formie wizualnej, stosunek liczby dokumentów lub rekordów w danej kategorii do łącznej liczby dokumentów lub rekordów.
- **Selection %.** W tej kolumnie przedstawiona jest wartość procentowa wyznaczona na podstawie stosunku łącznej liczby dokumentów lub rekordów w kategorii do łącznej liczby dokumentów lub rekordów w wyborze.
- **Docs.** . Ta kolumna przedstawia liczbę dokumentów lub rekordów w wyborze dla danej kategorii.

Wykres sieciowy kategorii

Na tej karcie wyświetlany jest wykres sieciowy kategorii. Ten wykres przedstawia pokrywanie się dokumentów lub rekordów dla kategorii, do których należą dokumenty lub rekordy, zgodnie z wyborami dokonanymi w innych panelach. Jeśli istnieją etykiety kategorii, to będą widoczne na wykresie. Można wybrać układ wykresu (sieciowy, kołowy, skierowany lub siatkowy) za pomocą przycisków na pasku narzędzi tego panelu.

W sieci każdy węzeł reprezentuje jedną kategorię. Za pomocą myszy można wybierać i przesuwać węzły w panelu. Wielkość węzła odzwierciedla wielkość względną wyznaczoną na podstawie liczny dokumentów lub rekordów tej kategorii w bieżącym wyborze. Grubość i kolor linii między dwiema kategoriami oznacza liczbę wspólnych dokumentów lub rekordów w tych kategoriach. Jeśli zatrzymasz wskaźnik myszy nad węzłem w trybie eksploracji, pojawi się podpowiedź z nazwą (lub etykietą) kategorii oraz łączną liczbą dokumentów lub rekordów w kategorii.

Uwaga: Domyślnie wykresy działają w trybie eksploracji, w którym można przesuwać węzły. Można jednak przełączyć się do trybu edycji, aby edytować układy wykresów, w tym zmieniać kolory i czcionki, legendę itp Aby uzyskać więcej informacji, patrz “Używanie pasków narzędzi i palet wykresów” na stronie 151.

W przypadku skopiowania danych wykresu za pomocą przycisku **Copy Visualization Data** i wklejenia ich do arkusza kalkulacyjnego lub edytora tekstu dane zostaną zaprezentowane w nagłówkach kolumn takich jak V1 czy V2 — aż do V7. Kolumny te zawierają następujące informacje:

- **V1, V2** Wartości te odpowiadają współrzędnym ekranu (odpowiednio X i Y).
- **V3, V5** Wyświetla kategorię pojęcia.
- **Size, V6** Wyświetla liczbę dokumentów, w których znaleziono pojęcia.
- **V7** Obecnie nieużywana.

Tabela sieciowa kategorii

Ta tabela przedstawia te same informacje, które wyświetlane są na karcie Category Web, ale w postaci tabelarycznej. Tabela zawiera trzy kolumny, które można sortować, klikając ich nagłówki:

- **Count.** Ta kolumna zawiera liczbę wspólnych dokumentów lub rekordów między dwiema kategoriami.
- **Category 1.** Ta kolumna zawiera nazwę pierwszej kategorii, po której następuje łączna liczba zawartych w niej dokumentów lub rekordów ujęta w nawiasy.
- **Category 2.** Ta kolumna zawiera nazwę drugiej kategorii, po której następuje łączna liczba zawartych w niej dokumentów lub rekordów ujęta w nawiasy.

Wykresy skupień

Po zbudowaniu skupień można eksplorować je wizualnie na wykresach sieciowych w panelu Visualization. Panel wizualizacji udostępnia dwa ujęcia skupień: wykres sieciowy pojęć i wykres sieciowy skupień. Wykresy sieciowe w tym panelu mogą być używane do analizy wyników tworzenia skupień i pomagają w ujawnieniu niektórych pojęć i reguł, które warto byłoby dodać do kategorii. Panel Visualization znajduje się w prawym górnym rogu widoku Clusters. Jeśli nie jest widoczny, można uzyskać dostęp do niego z menu View (**View > Panes > Visualization**). Wybierając skupienie w panelu Clusters, można automatycznie wyświetlić odpowiednie wykresy w panelu Visualization.

Uwaga: Domyślnie wykresy działają w trybie interaktywnym/wyboru, w którym można przesuwać węzły. Można jednak edytować układy wykresów w trybie edycji, w tym zmieniać kolory i czcionki, legendę itp Więcej informacji zawiera temat “Używanie pasków narzędzi i palet wykresów” na stronie 151.

Widok Clusters zawiera dwa wykresy sieciowe.

- **Concept Web Graph**. Ten wykres przedstawia wszystkie pojęcia w wybranych skupieniach, a także pojęcia powiązane spoza skupienia. Ten wykres pozwala zobaczyć, w jaki sposób pojęcia w skupieniu są powiązane, a także zorientować się w powiązaniach zewnętrznych. Więcej informacji zawiera temat “Wykres sieciowy pojęć”.
- **Wykres sieciowy skupień**. Ten wykres przedstawia wybrane skupienia ze wszystkimi powiązaniami zewnętrznymi wybranymi skupieniami pokazanymi jako linie kropkowane. Więcej informacji zawiera temat “Wykres sieciowy skupień” na stronie 150.

Więcej informacji zawiera temat Rozdział 10, “Analiza skupień”, na stronie 135.

Wykres sieciowy pojęć

Ta karta zawiera wykres sieciowy przedstawiający wszystkie pojęcia w wybranych skupieniach, a także pojęcia powiązane spoza skupienia. Ten wykres pozwala zobaczyć, w jaki sposób pojęcia w skupieniu są powiązane, a także zorientować się w powiązaniach zewnętrznych. Każde pojęcie w skupieniu jest reprezentowana jako węzeł o kolorze odzwierciedlającym typ. Więcej informacji zawiera temat “Tworzenie typów” na stronie 179.

Zostaną narysowane wewnętrzne powiązania między pojęciami w skupieniu, a grubość linii każdego powiązania będzie bezpośrednio odzwierciedlać liczbę dokumentów ze współwystąpieniami każdej pary pojęć albo wartość powiązania na podstawie podobieństwa, w zależności od wyboru dokonanego na pasku narzędzi wykresu. Widoczne są także powiązania zewnętrzne między pojęciami ze skupienia a pojęciami spoza skupienia.

Jeśli w oknie dialogowym Cluster Definitions zostaną wybrane pojęcia, to na wykresie sieciowym pojęć zostaną wyświetlone te pojęcia oraz ich powiązania zewnętrzne i wewnętrzne. Wszelkie powiązania między innymi pojęciami, które nie obejmują żadnego z zaznaczonych, nie będą widoczne na wykresie.

Uwaga: Domyślnie wykresy działają w trybie interaktywnym/wyboru, w którym można przesuwać węzły. Można jednak edytować układy wykresów w trybie edycji, w tym zmieniać kolory i czcionki, legendę itp Aby uzyskać więcej informacji, patrz “Używanie pasków narzędzi i palet wykresów” na stronie 151.

W przypadku skopiowania danych wykresu za pomocą przycisku **Copy Visualization Data** i wklejenia ich do arkusza kalkulacyjnego lub edytora tekstu dane zostaną zaprezentowane w nagłówkach kolumn takich jak V1 czy V2 — aż do V7. Kolumny te zawierają następujące informacje:

- **V1, V2** Wartości te odpowiadają współrzędnym ekranu (odpowiednio X i Y).
- **V3, V6** Wyświetla typ pojęcia.
- **V4, V5** Wyświetla etykietę pojęcia.
- **V7** Obecnie nieużywana.

Wykres sieciowy skupień

Ta karta zawiera wykres sieciowy przedstawiający wybrane skupienia. Powiązania zewnętrzne między wybranymi skupieniami, a także wszelkie powiązania między innymi skupieniami są pokazane jako linie kropkowane. W widoku sieciowym skupień każdy węzeł reprezentuje całe skupienie, a grubości linii narysowanych między skupieniami reprezentuje liczbę powiązań zewnętrznych pomiędzy dwoma skupieniami.

Ważne! Aby można było wyświetlić wykres sieciowy skupień, należy wcześniej stworzyć skupienia z powiązaniami zewnętrznymi. Powiązania zewnętrzne są to powiązania między parami pojęć w odrębnych skupieniach (jedno pojęcie w jednym skupieniu, a drugie pojęcie w innym).

Na przykład załóżmy, że mamy dwa skupienia. Skupienie A zawiera trzy pojęcia: A1, A2 i A3. Skupienie B zawiera dwa pojęcia: B1 u B2. Powiązane są następujące pojęcia: A1-A2, A1-A3, A2-B1 (zewnętrzne), A2-B2 (zewnętrzne), A1-B2 (zewnętrzne) i B1-B2. Oznacza to, że na wykresie sieciowym skupień grubość linii będzie reprezentować trzy powiązania zewnętrzne.

Uwaga: Domyślnie wykresy działają w trybie interaktywnym/wyboru, w którym można przesuwać węzły. Można jednak edytować układy wykresów w trybie edycji, w tym zmieniać kolory i czcionki, legendę itp Więcej informacji zawiera temat “Używanie pasków narzędzi i palet wykresów” na stronie 151.

Wykresy analizy powiązań w tekście

Po wyodrębnieniu wzorców analizy powiązań w tekście (TLA) można eksplorować je wizualnie na wykresach sieciowych w panelu Visualization. Panel wizualizacji udostępnia dwa ujęcia wzorców TLA: wykres sieciowy (wzorców) pojęć i wykres sieciowy (wzorców) typów. Wykresy sieciowe na tym panelu umożliwiają zwizualizowanie wzorców. Panel Visualization znajduje się w prawym górnym rogu widoku Text Link Analysis. Jeśli nie jest widoczny, można uzyskać dostęp do niego z menu View (**View > Panes > Visualization**). Jeśli nie dokonano wyboru, obszar wykresu jest pusty.

Uwaga: Domyślnie wykresy działają w trybie interaktywnym/wyboru, w którym można przesuwać węzły. Można jednak edytować układy wykresów w trybie edycji, w tym zmieniać kolory i czcionki, legendę itp Więcej informacji zawiera temat “Używanie pasków narzędzi i palet wykresów” na stronie 151.

Widok Text Link Analysis zawiera dwa wykresy sieciowe.

- **Concept Web Graph**. Ten wykres przedstawia wszystkie pojęcia w wybranych wzorcach. Szerokość linii i wielkości węzłów (jeśli nie są pokazane ikony typu) na wykresie pojęć przedstawia liczbę globalnych wystąpień w wybranej tabeli. Więcej informacji zawiera temat “Wykres sieciowy pojęć”.
- **Type Web Graph**. Ten wykres przedstawia wszystkie typy w wybranych wzorcach. Szerokość linii i wielkości węzłów (jeśli nie są pokazane ikony typu) na wykresie przedstawia liczbę globalnych wystąpień w wybranej tabeli. Węzły są przedstawiane w określonym kolorze lub w formie ikony. Więcej informacji zawiera temat “Wykres sieciowy typu”.

Więcej informacji zawiera temat Rozdział 11, “Eksplorowanie analizy powiązań w tekście”, na stronie 141.

Wykres sieciowy pojęć

Ten wykres przedstawia wszystkie pojęcia reprezentowane w obecnym wyborze. Na przykład, jeśli wybrano wzorzec typu z trzema pasującymi wzorcami pojęć, to na wykresie będą widoczne trzy zestawy powiązanych pojęć. Szerokość linii i wielkości węzłów na wykresie pojęć odzwierciedlają globalne liczebności. Wykres w formie wizualnej przedstawia te same informacje, które zostały wybrane w panelach wzorców. Typy pojęć są oznaczone kolorem lub ikoną, w zależności od wyboru dokonanego na pasku narzędzi wykresów. Więcej informacji zawiera temat “Używanie pasków narzędzi i palet wykresów” na stronie 151.

Wykres sieciowy typu

Ten wykres przedstawia wszystkie wzorce typów reprezentowane w obecnym wyborze. Na przykład, jeśli wybrano dwa wzorce pojęć, ten wykres przedstawiał będzie po jednym węźle na każdy typ obecny w wybranych wzorcach oraz

powiązania między typami w tym samym wzorcu. Szerokości linii i wielkości węzłów odzwierciedlają globalne liczebności dla zbioru. Wykres w formie wizualnej przedstawia te same informacje, które zostały wybrane w panelach wzorców. Typy identyfikowane są nie tylko przez nazwy typów widoczne na wykresie, lecz także kolorami lub ikonami, w zależności od wyboru dokonanego na pasku narzędzi wykresów. Więcej informacji zawiera temat "Używanie pasków narzędzi i palet wykresów".

Używanie pasków narzędzi i palet wykresów

Dla każdego wykresu istnieje pasek narzędzi, który umożliwia szybki dostęp do niektórych często używanych palet, na których można wykonać różne czynności na wykresach. Każdy widok (Categories and Concepts, Clusters i Text Link Analysis) ma trochę inny pasek narzędzi. Można wybrać pomiędzy trybem widoku *Explore* i trybem widoku *Edit*.

Tryb eksploracji umożliwia analityczną eksplorację danych i wartości reprezentowanych przez wizualizację, a tryb edycji umożliwia zmianę wyglądu i układu wizualizacji. Na przykład można zmienić czcionki i kolory tak, aby pasowały do oficjalnej stylistyki danej organizacji. Aby wybrać ten tryb, z menu wybierz **View > Visualization Pane > Edit Mode** (lub kliknij ikonę paska narzędzi).

W trybie edycji dostępnych jest kilka pasków narzędzi, które mają wpływ na różne aspekty układu wizualizacji. Jeśli któreś z tych pasków nie są wykorzystywane, można je ukryć. Dzięki temu zwiększa się w oknie dialogowym ilość miejsca dostępnego dla wykresu. Aby zaznaczyć lub usunąć zaznaczenie pasków narzędzi, kliknij odpowiednią nazwę paska narzędzi lub palety w menu View.

Więcej informacji o wszystkich ogólnych paskach narzędzi i paletach używanych do edycji wykresów można znaleźć w sekcji dotyczącej edycji wizualizacji pomocy online lub w pliku *ModelerSPOnodes.pdf*, który jest dostępny jako część pobranego produktu.

Tabela 36. Przyciski paska narzędzi Text Analytics.

Przycisk/lista	Opis
	Włącza tryb Edit. Przełącz do trybu Edit, aby zmienić wygląd wykresu, np. aby powiększyć czcionkę, dopasować kolory do korporacyjnych wytycznych stylu lub usunąć etykiety i legendy.
	Włącza tryb Explore. Domyślnie tryb Explore jest włączony, co oznacza, że można przesuwać i przeciągać węzły na wykresie, jak również umieszczać kursor myszy na obiektami wykresu, aby wyświetlić dodatkowe informacje podpowiedzi.
	Wybierz typ wyświetlania sieciowego w widoku Categories and Concepts, jak również w widoku Text Link Analysis. • **Circle Layout** Układ ogólny, który można zastosować względem dowolnego wykresu. Rozmieszcza wykres, zakładając, że łącza są nieukierunkowane i traktuje wszystkie węzły tak samo. Węzły są umieszczane tylko na obwodzie koła. • **Network Layout** Układ ogólny, który można zastosować względem dowolnego wykresu. Rozmieszcza wykres, zakładając, że łącza są nieukierunkowane i traktuje wszystkie węzły tak samo. Węzły są umieszczane swobodnie w ramach układu. • **Directed Layout** Układ, który można używać tylko dla wykresów kierunkowych. Ten układ tworzy struktury przypominające drzewa: od węzłów głównych do węzłów-liści, które są organizowane kolorami. Dane hierarchiczne są zazwyczaj dobrze wyświetlane w tym układzie. • **Grid Layout** Układ ogólny, który można zastosować względem dowolnego wykresu. Rozmieszcza wykres, zakładając, że łącza są nieukierunkowane i traktuje wszystkie węzły tak samo. Węzły są umieszczane tylko na punktach siatki w ramach obszaru.
	Reprezentacja wielkości łączy. Wybierz, co oznacza grubość linii na wykresie. Dotyczy to tylko widoku Clusters. Wykres sieciowy Clusters pokazuje tylko łącza pomiędzy grupami. Można wybrać pomiędzy: • **Similarity** Grubość wskazuje na liczbę zewnętrznych łączy pomiędzy dwiema grupami • **Co-occurrence** Grubość wskazuje liczbę dokumentów, w których występuje współwystępowanie deskryptorów.

Tabela 36. Przyciski paska narzędzi Text Analytics (kontynuacja).

Przycisk/lista	Opis
	Przycisk przełącznika, który po naciśnięciu wyświetla legendę. Jeśli przycisk nie jest naciśnięty, legenda nie jest wyświetlana.
	Przycisk przełącznika, który po naciśnięciu wyświetla ikony typu na wykresie zamiast używać kolorów typów. Dotyczy to tylko widoku Text Link Analysis.
	Przycisk przełącznika, który po naciśnięciu wyświetla suwak łączy pod wykresem. Można filtrować wyniki, przesuwając strzałkę.
	Będzie wyświetlać wykres dla najwyższego wybranego poziomu kategorii zamiast ich podkategorii.
	Będzie wyświetlać wykres dla najniższego poziomu wybranych kategorii.
	Ta opcja określa, w jaki sposób nazwy podkategorii są wyświetlane w wynikach. • **Full category path** Ta opcja wyświetli nazwę kategorii oraz pełną ścieżkę kategorii nadrzędnych, w razie potrzeby używając ukośników do rozdzielania nazw kategorii od nazw podkategorii. • **Short category path** Ta opcja wyświetli tylko nazwę kategorii, ale użyje wielokropków do wyświetlenia liczby kategorii nadrzędnych dla określonej kategorii. • **Bottom level category** Ta opcja wyświetli tylko nazwę kategorii bez pełnej ścieżki i bez wyświetlonych kategorii nadrzędnych.

Rozdział 13. Edytor zasobów sesji

IBM SPSS Modeler Text Analytics umożliwia szybkie i dokładne wyodrębnianie kluczowych pojęć z danych tekstowych. Ten proces wyodrębniania opiera się w dużej mierze na zasobach lingwistycznych, od których zależy sposób wyodrębniania informacji z danych tekstowych. Domyślnie te zasoby pochodzą z szablonów zasobów.

IBM SPSS Modeler Text Analytics jest dostarczany z zestawem specjalistycznych **szablonów zasobów**, które zawierają zestaw zasobów lingwistycznych i nielingwistycznych w postaci bibliotek i zasobów zaawansowanych. Określają one, w jaki sposób dane mają być obsługiwane i wyodrębniane. Więcej informacji zawiera Rozdział 14, “Szablony i zasoby”, na stronie 157.

W oknie dialogowym węzła można załadować kopię zasobów z szablonu do węzła. W razie potrzeby w interaktywnym pulpicie roboczym można dostosować te zasoby do konkretnych danych tego węzła. W interaktywnym pulpicie roboczym można również pracować z zasobami w widoku Resource Editor. Po każdym uruchomieniu interaktywnego pulpitu roboczego przeprowadzane jest wyodrębnianie przy użyciu zasobów załadowanych w oknie dialogowym węzła, chyba w węźle są zapisane zbuforowane dane i wyniki wyodrębniania.

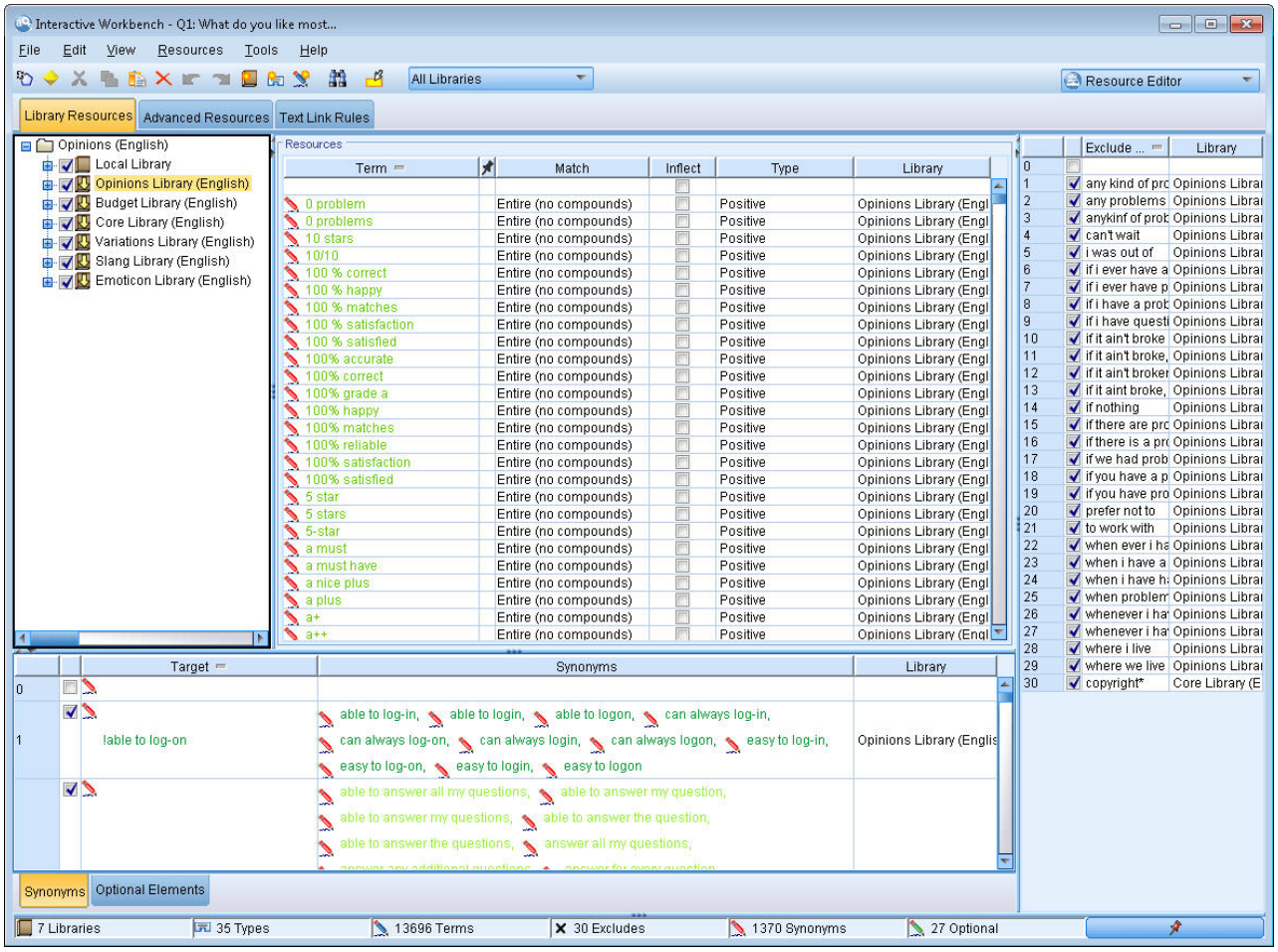
Edytowanie zasobów w oknie Resource Editor

Resource Editor zapewnia dostęp do zestawu zasobów używanego do wygenerowania wyników wyodrębniania (pojęć, typów i wzorców) dla sesji interaktywnej środowiska roboczego. Ten edytor jest bardzo podobny do edytora Template Editor, z tą różnicą, że w oknie Resource Editor edytuje się zasoby dla bieżącej sesji. Po zakończeniu pracy nad zasobami i wykonywania innych czynności można zaktualizować węzeł modelowania, aby zapisać efekty pracy i ewentualnie później wrócić do nich w kolejnej sesji pracy z interaktywnym pulpitem roboczym. Więcej informacji zawiera temat “Aktualizowanie węzłów modelowania i zapisywanie” na stronie 77.

Jeśli chcesz pracować bezpośrednio na szablonach służących do ładowania zasobów do węzłów, zaleca się użycie edytora Template Editor. Wiele czynności, które można wykonywać w edytorze Resource Editor, wykonuje się tak samo, jak w edytorze Template Editor. Oto niektóre z nich:

- **Praca z bibliotekami**. Więcej informacji zawiera temat Rozdział 15, “Praca z bibliotekami”, na stronie 167.
- **Tworzenie słowników typów**. Więcej informacji zawiera temat “Tworzenie typów” na stronie 179.
- **Dodawanie terminów do słowników**. Więcej informacji zawiera temat “Dodawanie terminów” na stronie 180.
- **Tworzenie synonimów**. Więcej informacji zawiera temat “Definiowanie synonimów” na stronie 185.
- **Importowanie i eksportowanie szablonów**. Więcej informacji zawiera temat “Importowanie i eksportowanie szablonów” na stronie 164.
- **Publikowanie bibliotek**. Więcej informacji zawiera temat “Publikowanie bibliotek” na stronie 173.

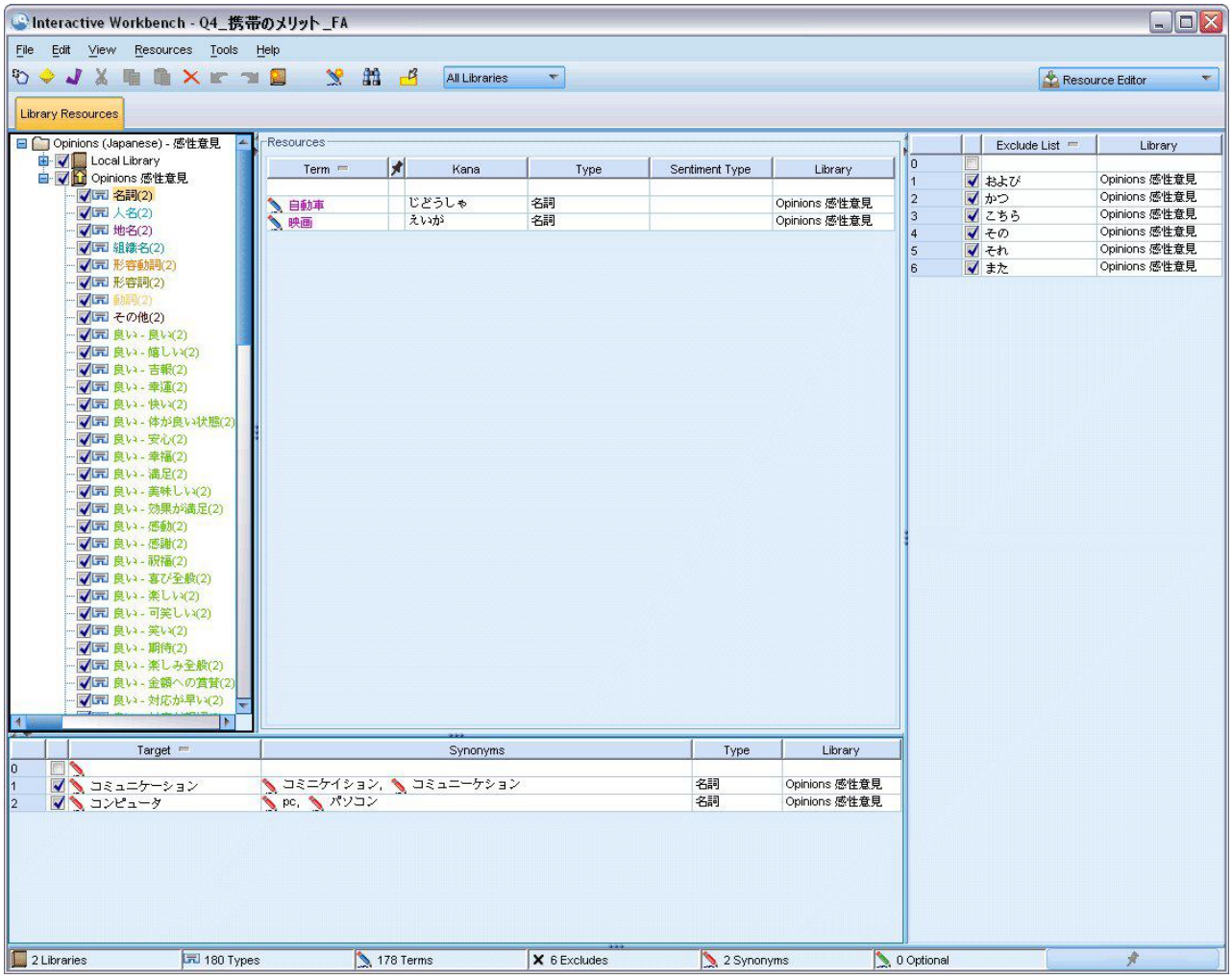
Informacje dotyczące języka angielskiego, francuskiego, hiszpańskiego, holenderskiego, niemieckiego, portugalskiego i włoskiego



Rysunek 35. Widok Resource Editor dla języków innych niż japoński

Dla tekstu japońskiego

Interfejs edytora dla tekstu w języku japońskim różni się od edytorów w innych językach.



Rysunek 36. Widok Resource Editor dla tekstu w języku japońskim

Tworzenie i modyfikowanie szablonów

Za każdym razem, gdy zmieniasz zasoby i chcesz te zmienione zasoby wykorzystać przyszłości, możesz zapisać je jako szablon. W takim przypadku można zapisać zasoby pod nazwą istniejącego szablonu lub pod nową nazwą. Następnie za każdym razem, gdy załadujesz ten szablon, uzyskasz dostęp do tych samych zasobów. Więcej informacji zawiera temat “Kopiowanie zasobów z szablonów i pakietów TAP” na stronie 26.

Uwaga: Możliwe jest także publikowanie i udostępnianie bibliotek. Więcej informacji zawiera temat “Współużytkowanie bibliotek” na stronie 172.

Aby utworzyć (lub zmodyfikować) szablon

1. Z menu w widoku Resource Editor wybierz opcje **Resources> Make Resource Template**. Zostanie otwarte okno dialogowe Make Resource Template.
2. Wprowadź nową nazwę w polu Template Name, jeśli chcesz utworzyć nowy szablon. Wybierz szablon w tabeli, jeśli chcesz zastąpić istniejący szablon aktualnie załadowanymi zasobami.
3. Kliknij przycisk **Save**, aby utworzyć szablon.

Ważne! Ponieważ szablony są ładowane po ich wybraniu w węźle, a nie po uruchomieniu strumienia, koniecznie załaduj szablon we wszystkich innych węzłach, w których jest używany, jeśli chcesz uwzględnić najnowsze zmiany. Więcej informacji zawiera temat “Aktualizowanie zasobów węzła po załadowaniu” na stronie 163.

Przełączanie się między szablonami zasobów

Jeśli chcesz zastąpić zasoby obecnie załadowane w tej sesji kopią zasobów z innego szablonu, możesz przełączyć się na te inne zasoby. Spowoduje to nadpisanie obecnie załadowanych zasobóww sesji. W przypadku przełączania się na inne zasoby w celu skorzystania z predefiniowanych reguł wzorców TLA, koniecznie wybierz szablon, który ma odpowiednie oznaczenie w kolumnie TLA.

Ważne! Nie można przełączyć się z szablonu japońskiego na szablon niejapoński lub odwrotnie.

Możliwość przełączania się między szablonami jest szczególnie przydatna, jeśli chcesz odtworzyć efekty pracy w sesji (kategorie, wzorce i zasoby), ale chcesz załadować zaktualizowaną kopię zasobów z szablonu bez utraty innych efektów pracy w sesji. Możesz wybrać szablon, którego zawartość chcesz skopiować do edytora Resource Editor, a następnie kliknąć przycisk **OK.** Spowoduje to zastąpienie zasobów w bieżącym sesji. Po zakończeniu sesji koniecznie zaktualizuj węzeł modelowania, jeśli chcesz zachować te zmiany do następnego uruchomienia interaktywnego pulpitu roboczego.

Uwaga: Jeśli przełączysz się na zawartość innego szablonu podczas sesji interaktywnej, nazwa szablonu wymieniona w węźle nadal będzie nazwą ostatniego załadowanego i skopiowanego szablonu. Aby skorzystać z tych zasobów lub innych efektów pracy interaktywnej, zaktualizuj węzeł modelowania przed zakończeniem sesji, a następnie wybierz opcję **Use session work** w węźle. Więcej informacji zawiera temat “Aktualizowanie węzłów modelowania i zapisywanie” na stronie 77.

Aby przełączyć się na inne zasoby

1. Z menu w widoku Resource Editor wybierz opcje **Resources> Switch Resource Templates**. Zostanie otwarte okno dialogowe Switch Resources.
2. W tabeli wybierz szablon, którego chcesz użyć.
3. Kliknij przycisk **OK**, aby odrzucić zasoby obecnie załadowane i zamiast nich załadować kopię zasobów z wybranego szablonu. Jeśli wprowadzisz zmiany w zasobach i chcesz zapisać biblioteki do wykorzystania w przyszłości, możesz opublikować, zaktualizować i udostępnić je przed przełączeniem. Więcej informacji zawiera temat “Współużytkowanie bibliotek” na stronie 172.

Rozdział 14. Szablony i zasoby

IBM SPSS Modeler Text Analytics umożliwia szybkie i dokładne wyodrębnianie kluczowych pojęć z danych tekstowych. Ten proces wyodrębniania opiera się w dużej mierze na zasobach lingwistycznych, od których zależy sposób wyodrębniania informacji z danych tekstowych. Więcej informacji zawiera temat “Jak działa wyodrębnianie” na stronie 5. Można zoptymalizować te zasoby w widoku Resource Editor.

Razem z produktem instalowany jest zestaw gotowych specjalistycznych zasobów. Zostały one opracowane w toku wieloletnich badań i optymalizacji pod kątem specyfiki konkretnych języków i zastosowań. Ponieważ dostarczone zasoby nie zawsze są idealnie dopasowane do kontekstu danych, można edytować gotowe szablony zasobów, a także tworzyć własne biblioteki precyzyjnie dostosowane do danych organizacji. Zasoby te są dostarczane w różnych formach i mogą być używane w sesji pracy z programem . Zasoby znajdują się w:

- **Szablonach zasobów**. Szablony składają się z bibliotek, typów i niektórych zasobów zaawansowanych, które wspólnie tworzą wyspecjalizowany zestaw zasobów dostosowany do konkretnej dziedziny lub kontekstu, takiego jak opinie o produktach.
- **Pakietach analizy tekstu (TAP)**. Pakiety TAP wiążą szablon zasobów z jednym lub wieloma wyspecjalizowanymi zestawami kategorii wygenerowanych przy użyciu zasobów z tego szablonu. Dzięki temu kategorie i zasoby są przechowywane razem i mogą być wielokrotnie wykorzystywane. Więcej informacji zawiera temat “Korzystanie z pakietów analizy tekstu (TAP)” na stronie 129.
- **Bibliotekach**. Biblioteki są używane jako elementy składowe dla pakietów TAP i szablonów. Można je również dodawać pojedynczo do zasobów w sesji. Każda biblioteka składa się z kilku słowników używanych do definiowania typów, synonimów i list wykluczeń oraz zarządzania tymi elementami. Choć biblioteki są również dostarczane indywidualnie, są spakowane w szablonach i pakietach TAP. Więcej informacji zawiera temat Rozdział 15, “Praca z bibliotekami”, na stronie 167.

Uwaga: Podczas wyodrębniania używane są również pewne skompilowane zasoby wewnętrzne. Te skompilowane zasoby zawierają dużą liczbę definicji dopełniających typy w bibliotece Core. Tych skompilowanych zasobów nie można edytować.

Resource Editor zapewnia dostęp do zestawu zasobów używanych do generowania wyników wyodrębniania (pojęć, typów i wzorców). W edytorze Resource Editor można wykonywać szereg różnych zadań, takich jak:

- **Praca z bibliotekami**. Więcej informacji zawiera temat Rozdział 15, “Praca z bibliotekami”, na stronie 167.
- **Tworzenie słowników typów**. Więcej informacji zawiera temat “Tworzenie typów” na stronie 179.
- **Dodawanie terminów do słowników**. Więcej informacji zawiera temat “Dodawanie terminów” na stronie 180.
- **Tworzenie synonimów**. Więcej informacji zawiera temat “Definiowanie synonimów” na stronie 185.
- **Aktualizowanie zasobów w pakiecie TAP**. Więcej informacji zawiera temat “Aktualizowanie pakietów analizy tekstu” na stronie 131.
- **Tworzenie szablonów**. Więcej informacji zawiera temat “Tworzenie i modyfikowanie szablonów” na stronie 155.
- **Importowanie i eksportowanie szablonów**. Więcej informacji zawiera temat “Importowanie i eksportowanie szablonów” na stronie 164.
- **Publikowanie bibliotek**. Więcej informacji zawiera temat “Publikowanie bibliotek” na stronie 173.

Template Editor a Resource Editor

Istnieją dwie główne metody pracy z szablonami, bibliotekami i ich zasobami oraz edytowania ich. Na zasobach lingwistycznych można operować w edytorze Template Editor lub w edytorze Resource Editor.

Template Editor

Template Editor umożliwia tworzenie i edytowanie szablonów zasobów bez użycia interaktywnego pulpitu roboczego i niezależnie od konkretnego węzła lub strumienia. Ten edytor służy do tworzenia lub edytowania szablonów zasobów przed załadowaniem ich do węzła Text Link Analysis i węzła modelowania Text Mining.

Template Editor jest dostępny z głównego paska narzędzi produktu IBM SPSS Modeler, za pośrednictwem menu **Tools> Text Analytics Template Editor**.

Resource Editor

Resource Editor jest dostępny w interaktywnym pulpicie roboczym, pozwala pracować z zasobami w kontekście konkretnego węzła i zbioru danych. Po dodaniu do strumienia węzła modelowania Text Mining można załadować kopię zawartości szablonu zasobów lub kopię pakietu analizy tekstu (zestawy kategorii *i* zasoby), aby określić sposób wyodrębniania tekstu do eksploracji. W przypadku uruchomienia interaktywnego pulpitu roboczego oprócz tworzenia kategorii, wyodrębniania wzorców TLA i tworzenia modeli kategorii można również zoptymalizować zasoby pod kątem danych sesji w zintegrowanym widoku Resource Editor. Więcej informacji zawiera temat “Edytowanie zasobów w oknie Resource Editor” na stronie 153.

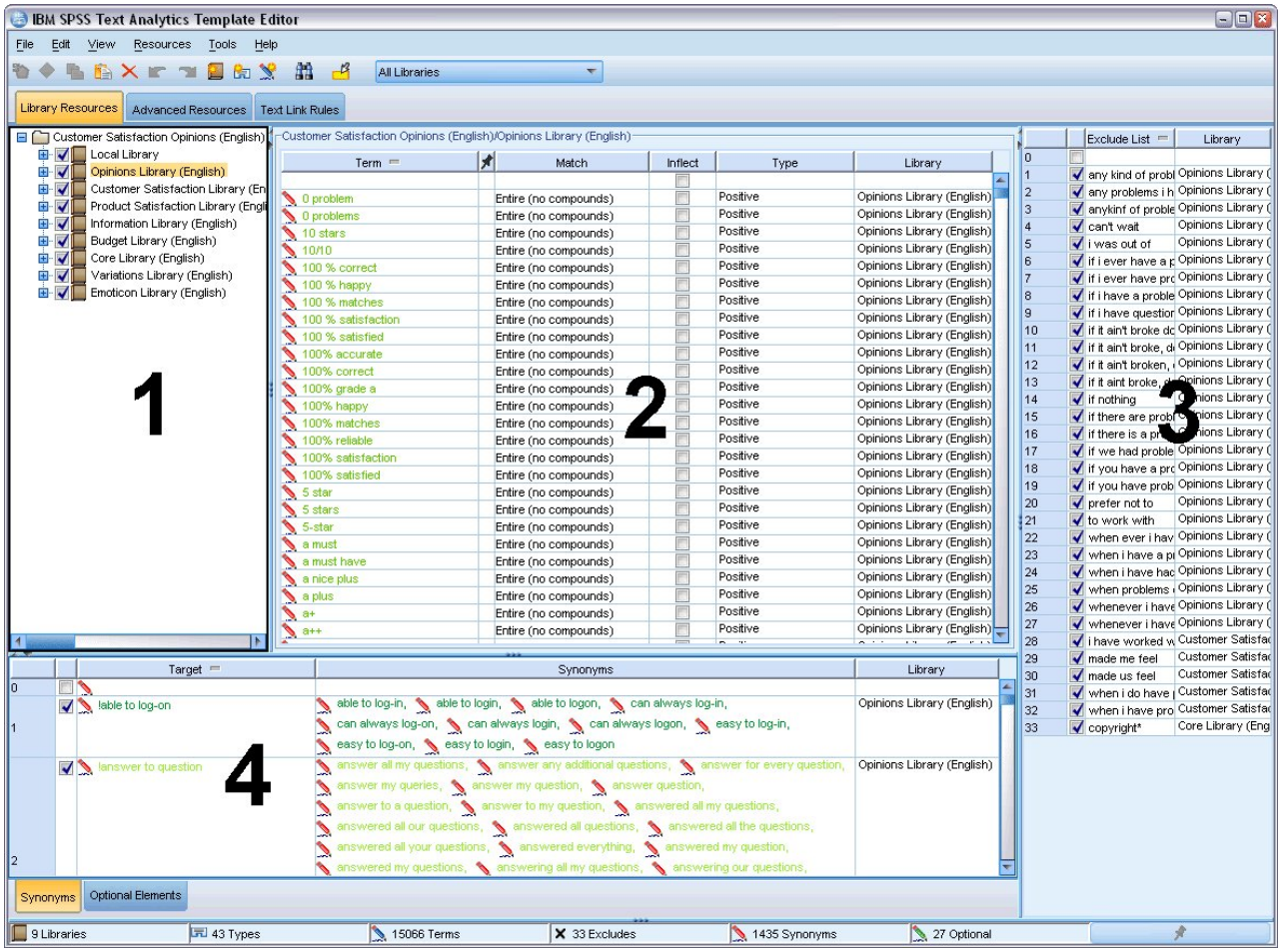
Za każdym razem, gdy operujesz na zasobach w interaktywnym pulpicie roboczym, zmiany te mają zastosowanie tylko w bieżącej sesji interaktywnej. Jeśli chcesz zapisać swoją pracę (zasoby, kategorie, wzorce itp.), aby kontynuować pracę podczas kolejnej sesji, musisz zaktualizować węzeł modelowania. Więcej informacji zawiera temat “Aktualizowanie węzłów modelowania i zapisywanie” na stronie 77.

Jeśli chcesz zapisać zmiany z powrotem do oryginalnego szablonu, którego zawartość została skopiowana do węzła modelowania, tak aby ta aktualizacja szablonu mogła zostać załadowana do innych węzłów, możesz utworzyć szablon z zasobów. Więcej informacji zawiera temat “Tworzenie i modyfikowanie szablonów” na stronie 155.

Interfejs edytora

Operacje wykonywane w widoku Template Editor lub Resource Editor związane są z zarządzaniem i optymalizacją zasobów lingwistycznych. Zasoby te są przechowywane w postaci szablonów i bibliotek. Więcej informacji zawiera temat “Słowniki typów” na stronie 177.

Karta Library Resources



Rysunek 37. Edytor szablonów eksploracji tekstu

Interfejs jest podzielony na cztery części w następujący sposób:

1. Panel drzewa bibliotek. Znajduje się w lewym górnym rogu i zawiera drzewo bibliotek. W drzewie tym można włączać i wyłączać biblioteki oraz filtrować widoki w innych panelach, wybierając bibliotekę w drzewie. Za pośrednictwem menu kontekstowych można wykonywać w tym drzewie wiele różnych operacji. Po rozwinięciu biblioteki w drzewie widoczne stają się typy, które ta biblioteka zawiera. Możesz także filtrować tę listę za pośrednictwem menu **View**, jeśli chcesz pracować tylko z jedną konkretną biblioteką.

2. Listy terminów na panelu słownika typów. Panel ten znajduje się na prawo od drzewa bibliotek i zawiera listy terminów słowników typów dla bibliotek wybranych w drzewie. **Słownik typu** jest zbiorem terminów pogrupowanych pod jedną etykietą lub nazwą typu. Odczytując dane tekstowe, mechanizm wyodrębniania porównuje wyrazy napotkane w tekście z terminami w słownikach typów. Jeśli wyodrębnione pojęcie figuruje jako termin w słowniku typów, to jest mu przypisywana nazwa typu właściwa dla słownika. Słownik typu można postrzegać jako słownik terminów, które mają ze sobą coś wspólnego. Na przykład typ <Location> w bibliotece Core zawiera takie pojęcia, jak new orleans, great britain i new york. Wszystkie te terminy reprezentują lokalizacje geograficzne. Jedna biblioteka może zawierać jeden lub więcej słowników typów. Więcej informacji zawiera temat “Słowniki typów” na stronie 177.

3. Panel słownika wykluczeń. Panel ten znajduje się po prawej stronie i zawiera zbiór terminów, które zostaną wykluczone z ostatecznych wyników wyodrębniania. Terminy figurujące w słowniku wykluczeń nie pojawiają się w panelu Extraction Results. Wykluczone terminy mogą być zapisane w bibliotece wybranej przez użytkownika. Jednak na panelu słownika wykluczeń wyświetlane są wszystkie wykluczone terminy dla wszystkich bibliotek widocznych w drzewie biblioteki. Więcej informacji zawiera temat “Słowniki wykluczeń” na stronie 187.

4. Panel słownika zastąpień. Panel ten znajduje się w lewym dolnym rogu i na osobnych kartach zawiera synonimy oraz elementy opcjonalne. Synonimy i elementy opcjonalne pomagają w grupowaniu podobnych terminów pod jednym pojęciem wiodącym lub docelowym w ostatecznych wynikach wyodrębniania. Słownik ten może zawierać znane synonimy oraz synonimy i elementy zdefiniowane przez użytkownika, jak również często spotykane błędnie zapisane formy terminów połączone w pary z terminami zapisanymi poprawnie. Definicje synonimów i elementy opcjonalne mogą być zapisane w bibliotece wybranej przez użytkownika. Jednak na panelu słownika zastąpień wyświetlane są pozycje ze wszystkich bibliotek widocznych w drzewie biblioteki. Wprawdzie na panelu tym wyświetlane są wszystkie synonimy i elementy opcjonalne ze wszystkich bibliotek, zastąpienia dla wszystkich bibliotek z drzewa są widoczne w tym panelu łącznie. Jedna biblioteka może zawierać tylko jeden słownik zastąpień. Więcej informacji zawiera temat “Słowniki zastąpień/synonimów” na stronie 184. Należy zwrócić uwagę, że karta Optional Elements nie ma zastosowania do zasobów lingwistycznych właściwych dla języka japońskiego.

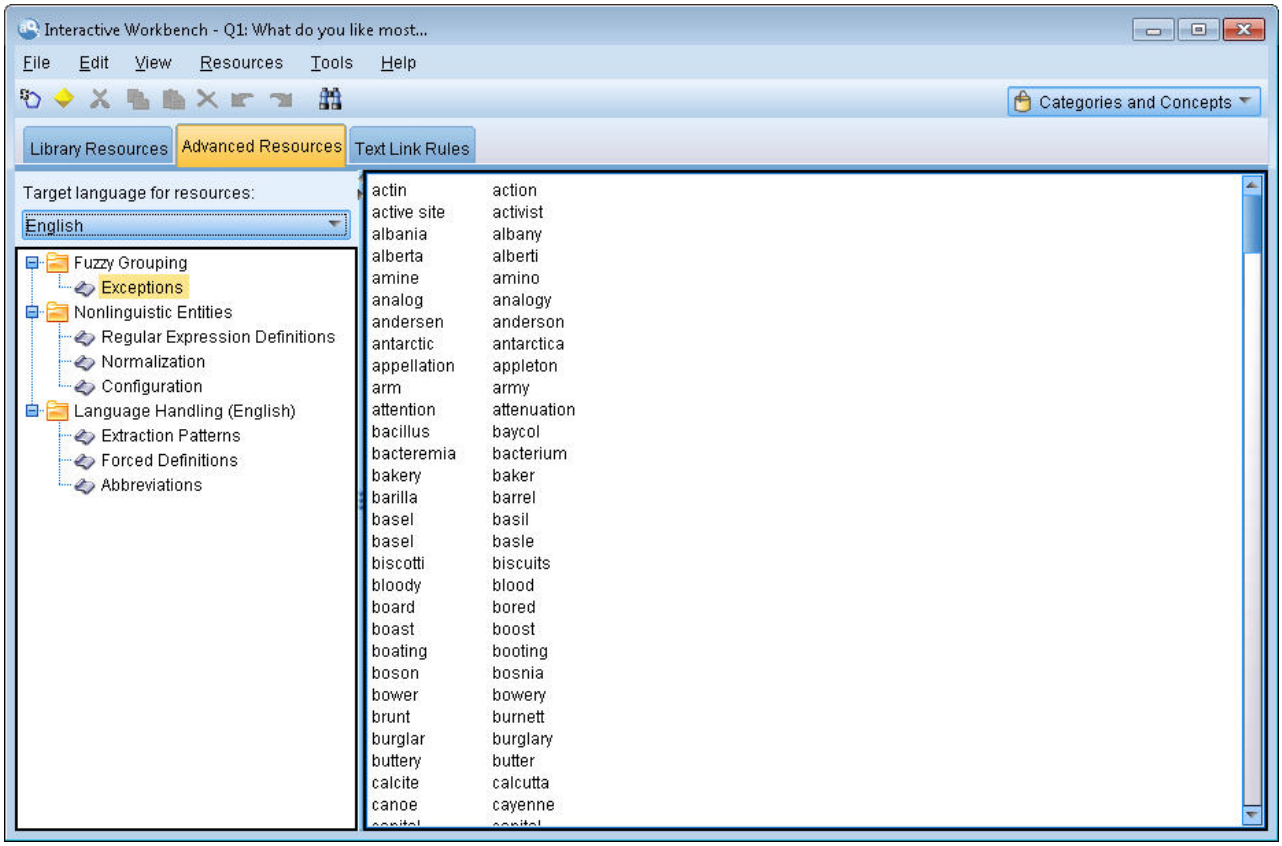
Uwagi:

- Jeśli chcesz, aby były widoczne tylko informacje dotyczące jednej biblioteki, możesz zmienić widok biblioteki, korzystając z listy rozwijanej na pasku narzędzi. Zawiera ona ogólny wpis **All Libraries** (Wszystkie biblioteki) oraz dodatkowe wpisy poszczególnych bibliotek. Więcej informacji zawiera temat “Przeglądanie bibliotek” na stronie 170.
- Interfejs edytora tekstu w języku japońskim różni się od edytora w innych językach.

Karta Advanced Resources

Zaawansowane zasoby są dostępne na drugiej karcie w widoku edytora. Możesz przeglądać i edytować zasoby zaawansowane na tej karcie. Więcej informacji zawiera Rozdział 17, “Informacje o zasobach zaawansowanych”, na stronie 189.

Ważne! Ta karta nie jest dostępna w przypadku zasobów dostosowanych dla tekstów w języku japońskim.

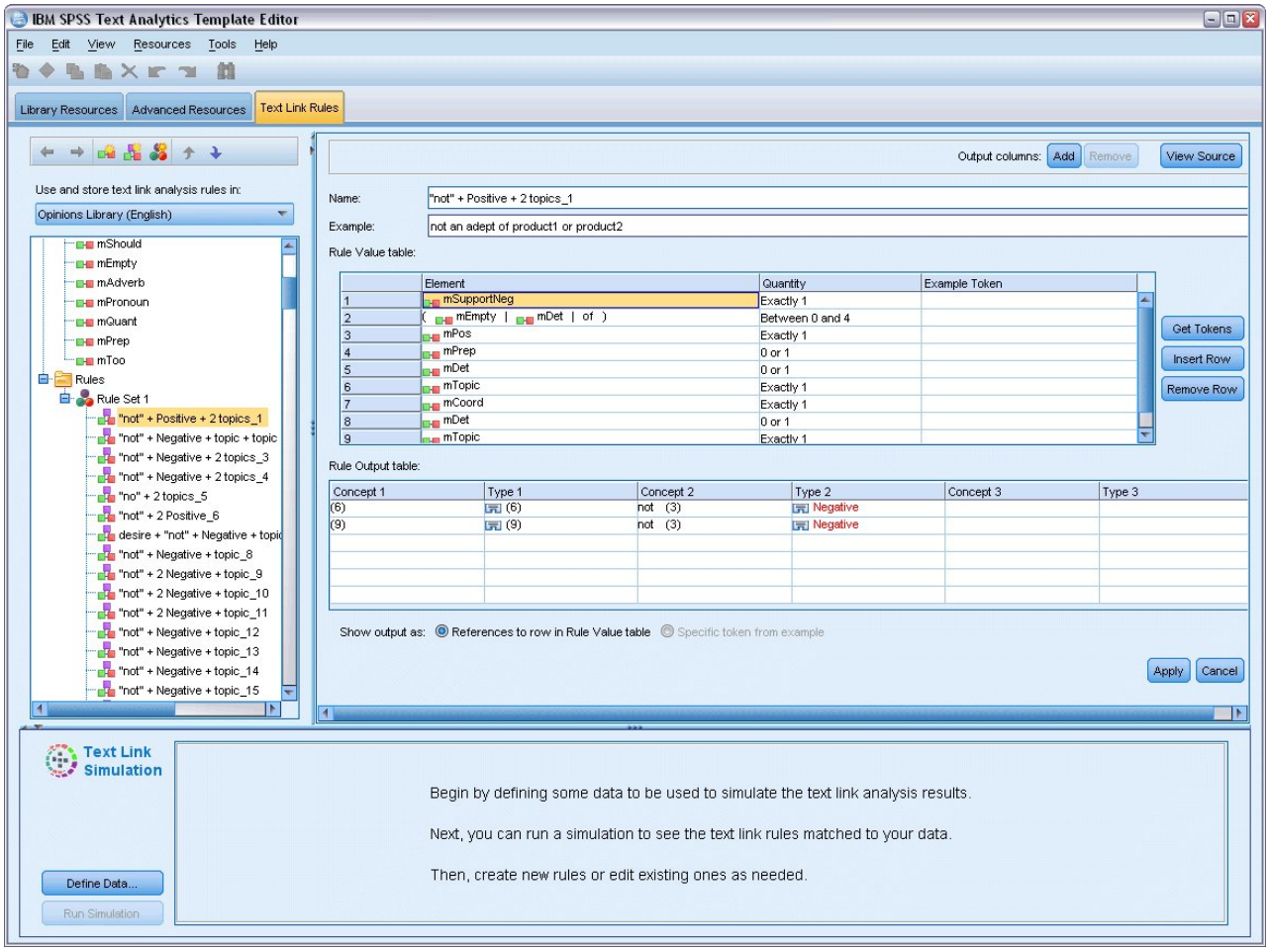


Rysunek 38. Edytor szablonów eksploracji tekstu - karta Advanced Resources

Karta Text Link Rules

Począwszy od wersji 14 reguły analizy powiązań w tekście są edytowalne na odrębnej karcie w widoku edytora. Możesz pracować w edytorze reguł, tworzyć własne reguły, a nawet uruchomić symulację, aby zobaczyć, w jaki sposób reguły wpłyną na wyniki analizy TLA. Więcej informacji zawiera Rozdział 18, “Informacje o regułach powiązań w tekście”, na stronie 201.

Ważne! Ta karta nie jest dostępna w przypadku zasobów dostosowanych dla tekstów w języku japońskim.



Rysunek 39. Edytor szablonów eksploracji tekstu — karta Text Link Rules

Otwieranie szablonów

Po uruchomieniu edytora Template Editor zostanie wyświetlona zachęta do otwarcia szablonu. Można też otworzyć szablon z menu File. Jeśli potrzebujesz szablonu, który zawiera reguły analizy powiązań w tekście (TLA), wybierz szablon z ikoną w kolumnie TLA. Język, w którym szablon został utworzony, jest wyświetlany w kolumnie Language.

Jeśli chcesz zaimportować szablon, który nie jest wyświetlany w tabeli, lub jeśli chcesz wyeksportować szablon, możesz użyć przycisków w oknie dialogowym Open Template. Więcej informacji zawiera temat “Importowanie i eksportowanie szablonów” na stronie 164.

Aby otworzyć szablon

1. Z menu w oknie Template Editor wybierz opcje **File > Open Resource Template**. Zostanie otwarte okno dialogowe Open Resource Template.
2. W tabeli wybierz szablon, którego chcesz użyć.
3. Kliknij przycisk **OK**, aby otworzyć ten szablon. Jeśli masz obecnie otwarty inny szablon w edytorze, kliknięcie przycisku OK spowoduje odrzucenie otwartego szablonu i wyświetlenie szablonu wybranego w oknie dialogowym. Jeśli wprowadzisz zmiany w zasobach i chcesz zapisać biblioteki do wykorzystania w przyszłości, możesz opublikować, zaktualizować i udostępnić je przed otwarciem innego szablonu. Więcej informacji zawiera temat “Współużytkowanie bibliotek” na stronie 172.

Zapisywanie szablonów

W edytorze Template Editor można zapisać zmiany wprowadzone do szablonu. Można zapisać je pod nazwą istniejącego szablonu lub pod nową nazwą.

Jeśli wprowadzisz zmiany w szablonie, który został już wcześniej załadowany do węzła, musisz ponownie załadować zawartość szablonu do węzła, aby uwzględnić najnowsze zmiany. Więcej informacji zawiera temat “Kopiowanie zasobów z szablonów i pakietów TAP” na stronie 26.

Lub, jeśli na karcie Model węzła Text Mining wybrana jest opcja **Use saved interactive work**, co oznacza, że używane są zasoby z poprzedniej sesji pracy z interaktywnym pulpitem roboczym, musisz przełączyć się na zasoby tego szablonu w ramach sesji interaktywnej. Więcej informacji zawiera temat “Przełączanie się między szablonami zasobów” na stronie 156.

Uwaga: Możliwe jest także publikowanie i udostępnianie bibliotek. Więcej informacji zawiera temat “Współużytkowanie bibliotek” na stronie 172.

Aby zapisać szablon

1. Z menu w oknie Template Editor, wybierz opcje **File > Save Resource Template**. Zostanie otwarte okno dialogowe Save Resource Template.
2. Wprowadź nową nazwę w polu Template name, jeśli chcesz zapisać ten szablon jako nowy. Wybierz szablon w tabeli, jeśli chcesz zastąpić istniejący szablon aktualnie załadowanymi zasobami.
3. W razie potrzeby wprowadź opis, aby w tabeli był widoczny komentarz lub adnotacja.
4. Kliknij przycisk **Save**, aby zapisać szablon.

Ważne! Ponieważ zasoby z szablonów i pakietów TAP są ładowane/kopiowane do węzła, należy zaktualizować zasoby, ponownie je ładując, jeśli wprowadzono zmiany w szablonie i zmiany te mają być wykorzystane w istniejącym strumieniu. Więcej informacji zawiera temat “Aktualizowanie zasobów węzła po załadowaniu”.

Aktualizowanie zasobów węzła po załadowaniu

Domyślnie, gdy użytkownik dodaje węzeł do strumienia, zestaw zasobów z domyślnego szablonu zostaje załadowany i osadzony w węźle. A w przypadku zmiany szablonu lub użycia pakietu TAP zawartość nadpisuje dotychczasowe zasoby. Ponieważ szablony i pakiety TAP nie są powiązane z węzłem bezpośrednio, wszelkie wprowadzone zmiany w szablonie lub pakiecie TAP nie są automatycznie dostępne w istniejącym węźle. Aby skorzystać z tych zmian, należy zaktualizować zasoby w węźle. Zasoby można aktualizować na dwa sposoby.

Metoda 1: Ponownie załadowanie zasobów w karcie Model

Jeśli chcesz zaktualizować zasoby w węźle przy użyciu nowego lub zaktualizowanego szablonu lub pakietu TAP, możesz ponownie załadować zasoby na karcie Model węzła. Ponowne załadowanie spowoduje zastąpienie kopii zasobów w węźle bardziej aktualną kopią. Dla wygody użytkownika czas i data aktualizacji zostaną wyświetlone na karcie Model wraz z nazwą szablonu źródłowego. Więcej informacji zawiera temat “Kopiowanie zasobów z szablonów i pakietów TAP” na stronie 26.

Jednak w przypadku pracy na interaktywnym pulpicie roboczym w węźle modelowania Text Mining i wybrania opcji **Use session work** na karcie Model używane będą zapisane efekty pracy i zasoby sesji, a przycisk **Load** będzie wyłączony. Wynika to z faktu, że w przeszłości podczas sesji interaktywnej wybrano opcję **Update Modeling Node** i zachowano kategorie, zasoby i inne efekty pracy. W tym przypadku, jeśli chcesz zmienić lub zaktualizować te zasoby, możesz skorzystać z następnej metody, czyli przełączenia się między zasobami w edytorze Resource Editor.

Metoda 2: Przełączanie się między zasobami w edytorze Resource Editor

Za każdym razem, gdy chcesz użyć innych zasobów podczas sesji interaktywnej, możesz podmienić te zasoby za pomocą okna dialogowego Switch Resources. Jest to szczególnie użyteczne, jeśli chcesz ponownie wykorzystać

istniejące definicje kategorii, ale zastąpić zasoby. W takim przypadku można wybrać opcję **Use session work** na karcie Model węzła modelowania Text Mining. W ten sposób można zablokować możliwość ponownego załadowania szablonu w oknie dialogowym węzła, a zamiast tego zachować ustawienia i zmiany dokonane podczas sesji. Następnie można uruchomić sesję interaktywnego pulpitu roboczego, uruchamiając strumień i przełączając się na inne zasoby w edytorze Resource Editor. Więcej informacji zawiera temat “Przełączanie się między szablonami zasobów” na stronie 156.

W celu zachowania efektów pracy, w tym zasobów, do wykorzystania w przyszłych sesjach należy zaktualizować węzeł modelowania z poziomu interaktywnego pulpitu roboczego, aby zasoby (i inne dane) zostały zapisane z powrotem do węzła. Więcej informacji zawiera temat “Aktualizowanie węzłów modelowania i zapisywanie” na stronie 77.

Uwaga: Jeśli przełączysz się na zawartość innego szablonu podczas sesji interaktywnej, nazwa szablonu wymieniona w węźle nadal będzie nazwą ostatniego załadowanego i skopiowanego szablonu. Aby skorzystać z tych zasobów lub innych efektów pracy interaktywnej, zaktualizuj węzeł modelowania przed zakończeniem sesji.

Zarządzanie szablonami

Niekiedy trzeba wykonać pewne pewne podstawowe czynności związane z zarządzaniem szablonami, takie jak zmienianie nazw szablonów, importowanie i eksportowanie szablonów lub usuwanie przestarzałych szablonów. Czynności te wykonuje się w oknie dialogowym Manage Templates. Importowanie i eksportowanie szablonów umożliwia współużytkowanie szablonów z innymi użytkownikami. Więcej informacji zawiera temat “Importowanie i eksportowanie szablonów”.

Uwaga: Nie można zmienić nazwy lub usunąć szablonu, który został zainstalowany (lub dostarczony) razem z produktem. Jeśli chcesz zmienić nazwę takiego szablonu, możesz otworzyć zainstalowany szablon i utworzyć nowy o dowolnej nazwie. Możesz usuwać swoje własne szablony. Jeśli jednak spróbujesz usunąć szablon dostarczony z produktem, to zostanie on przywrócony do pierwotnego stanu z chwili instalacji.

Aby zmienić nazwę szablonu

1. Z menu wybierz opcje **Resources > Manage Resource Templates**. Zostanie otwarte okno dialogowe Manage Templates.
2. Wybierz szablon, którego nazwę chcesz zmienić, i kliknij opcję **Rename**. Pole nazwy staje się polem dostępnym do edycji w tabeli.
3. Wpisz nową nazwę, a następnie naciśnij klawisz Enter. Zostanie otwarte okno dialogowe z potwierdzeniem.
4. Jeśli zgadzasz się na zmianę nazwy, kliknij przycisk **Yes**. Jeśli nie, kliknij przycisk **No**.

Aby usunąć szablon

1. Z menu wybierz opcje **Resources > Manage Resource Templates**. Zostanie otwarte okno dialogowe Manage Templates.
2. W oknie dialogowym Manage Templates wybierz szablon, który chcesz usunąć.
3. Kliknij przycisk **Delete**. Zostanie otwarte okno dialogowe z potwierdzeniem.
4. Kliknij przycisk **Yes**, aby usunąć, lub kliknij przycisk **No**, aby anulować żądanie. Jeśli zostanie kliknięty przycisk **Yes**, szablon zostanie usunięty.

Importowanie i eksportowanie szablonów

Szablony można współużytkować z innymi użytkownikami lub komputerami przez importowanie i eksportowanie ich. Szablony są przechowywane w wewnętrznej bazie danych, ale mogą być eksportowane jako pliki **.lrt* na dysk twardy.

Ponieważ istnieją okoliczności, w których konieczne może być importowanie lub eksportowanie szablonów, odpowiednie funkcje udostępniono w kilku oknach dialogowych. Wymieniono je poniżej.

- Okno dialogowe Open Template w edytorze Template Editor

- Okno dialogowe Load Resources w węźle modelowania Text Mining i węźle Text Link Analysis.
- Okno dialogowe Manage Templates w edytorach Template Editor i Resource Editor.

Aby zaimportować szablon

1. W oknie dialogowym kliknij przycisk **Import**. Zostanie otwarte okno dialogowe Import Template.
2. Wybierz plik szablonu zasobów (*.*lrt*) do zaimportowania i kliknij przycisk **Import**. Można zapisać zaimportowany szablon pod inną nazwą lub nadpisać istniejący. Okno dialogowe zostanie zamknięte, a szablon zostanie wyświetlony w tabeli.

Aby wyeksportować szablon

1. W oknie dialogowym wybierz szablon, który chcesz wyeksportować, a następnie kliknij przycisk **Export**. Zostanie otwarte okno dialogowe Select Directory.
2. Wybierz katalog, do którego chcesz wyeksportować bibliotekę, i kliknij przycisk **Export**. To okno dialogowe zostanie zamknięte, a szablon zostanie wyeksportowany do pliku o rozszerzeniu *.*lrt*.

Wychodzenie z edytora Template Editor

Po zakończeniu pracy w edytorze Template Editor można zapisać pracę i wyjść z edytora.

Aby wyjść z edytora Template Editor

1. Z menu wybierz opcje **File > Close**. Zostanie otwarte okno dialogowe Save and Close.
2. Wybierz opcję **Save changes to template**, aby zapisać otwarty szablon przed zamknięciem edytora.
3. Wybierz opcję **Publish libraries**, jeśli chcesz opublikować którekolwiek z bibliotek w otwartym szablonie przed zamknięciem edytora. Jeśli ta opcja zostanie wybrana, pojawi się zachęta do wybrania bibliotek do opublikowania. Więcej informacji zawiera temat “Publikowanie bibliotek” na stronie 173.

Tworzenie kopii zapasowej zasobów

Ze względów bezpieczeństwa wskazane może być tworzenie kopii zapasowych zasobów.

Ważne! Podczas odtwarzania cała zawartość zasobów zostanie wyczyszczona, a w produkcie będzie dostępna jedynie zawartość pliku kopii zapasowej. Dotyczy to także efektów bieżącej pracy.

Uwaga: Możliwe jest tylko tworzenie i odtwarzanie kopii zapasowych do tej samej wersji głównej oprogramowania. Na przykład, jeśli kopia zapasowa została utworzona w wersji 15, to nie można jej odtworzyć w wersji 16.

Aby utworzyć kopię zapasową zasobów

1. Z menu wybierz kolejno opcje **Resources > Backup Tools > Backup Resources**. Zostanie otwarte okno dialogowe Backup.
2. Wprowadź nazwę pliku kopii zapasowej, a następnie kliknij przycisk **Save**. Okno dialogowe zostanie zamknięte, a plik kopii zapasowej zostanie utworzony.

Aby odtworzyć zasoby

1. Z menu wybierz kolejno opcje **Resources > Backup Tools > Restore Resources**. Pojawi się alert informujący o tym, że odtwarzanie spowoduje zastąpienie bieżącej zawartości bazy danych.
2. Kliknij przycisk **Yes**, aby kontynuować. Zostanie otwarte okno dialogowe.
3. Wybierz plik kopii zapasowej, którą chcesz odtworzyć, i kliknij przycisk **Open**. Okno dialogowe zostanie zamknięte, a zasoby zostaną odtworzone w aplikacji.

Importowanie plików zasobów

Jeśli wprowadzono zmiany bezpośrednio w plikach zasobów poza produktem, to można zaimportować te pliki do wybranej biblioteki, wybierając tę bibliotekę i inicjując operację importu. Importując słownik, można też zaimportować wszystkie obsługiwane pliki do określonej otwartej biblioteki. Importować można wyłącznie pliki **.txt*.

Ważne! Pliki .txt w języku japońskim muszą być zapisane z kodowaniem UTF8. Ponadto nie można importować list wykluczeń dla języka japońskiego.

Każdy zaimportowany plik musi zawierać po jednym wpisie w każdym wierszu, a jeśli zawartość jest ustrukturyzowana jako:

- lista wyrazów lub fraz (po jednej na wiersz), plik jest importowany jako lista terminów do słownika typu, przy czym słownik typu przyjmuje nazwę pliku bez rozszerzenia.
- lista terminów w postaci *term1* <TAB> *term2*, wówczas importowana jest jako lista synonimów, gdzie *term1* jest terminem podstawowym, a *term2* jest terminem docelowym.

Aby zaimportować jeden plik zasobów

1. Z menu wybierz kolejno opcje **Resources > Import Files > Import Single File**. Zostanie otwarte okno dialogowe Import File.
2. Wybierz plik, który chcesz zaimportować, i kliknij przycisk **Import**. Zawartość pliku zostanie przekształcona do formatu wewnętrznego i dodana do biblioteki.

Aby zaimportować wszystkie pliki z katalogu

1. Z menu wybierz kolejno opcje **Resources > Import Files > Import Entire Directory**. Zostanie otwarte okno dialogowe Import Directory.
2. Na liście **Import** wybierz bibliotekę, do której chcesz zaimportować wszystkie pliki zasobów. Wybranie opcji **Default** spowoduje utworzenie nowej biblioteki o nazwie wskazanego katalogu.
3. Wybierz katalog, z którego mają być importowane pliki. Zawartość podkatalogów nie będzie odczytywana.
4. Kliknij przycisk **Import**. Okno dialogowe zostanie zamknięte, a zawartość importowanych plików zasobów pojawi się w edytorze w postaci słowników i plików zasobów zaawansowanych.

Rozdział 15. Praca z bibliotekami

Zasoby używane do wyodrębniania i grupowania terminów występujących danych tekstowych zawsze zawierają co najmniej jedną bibliotekę. Zestaw bibliotek widoczny jest drzewie bibliotek znajdującym się w lewej górnej części okna Template Editor i Resource Editor. Biblioteki składają się z trzech rodzajów słowników: typów, zastąpień i wykluczeń. Więcej informacji zawiera Rozdział 16, “Informacje o słownikach w bibliotekach”, na stronie 177.

Wybrany szablon zasobów lub zasoby z wybranego pakietu TAP zawierają kilka bibliotek, dzięki którym można natychmiast rozpocząć wyodrębnianie pojęć z danych tekstowych. Można jednak utworzyć własne biblioteki oraz publikować je do ponownego wykorzystania. Więcej informacji zawiera temat “Publikowanie bibliotek” na stronie 173.

Załóżmy na przykład, że często pracujemy z danymi tekstowymi z branży motoryzacyjnej. Po przeanalizowaniu danych decydujemy, że chcemy utworzyć zasoby dostosowane do naszej branży, aby uwzględnić specjalistyczne słownictwo lub żargon. Za pomocą edytora Template Editor można utworzyć nowy szablon, a w nim bibliotekę do wyodrębnienia i grupowania terminów właściwych dla branży motoryzacyjnej. Ponieważ informacje z tej biblioteki będą wykorzystywane wielokrotnie, należy opublikować bibliotekę w centralnym repozytorium, do którego można uzyskać dostęp w oknie dialogowym **Manage Libraries**, aby można ją było ponownie wykorzystać w innych sesjach strumienia .

Załóżmy, że interesuje nasz także pogrupowanie składników, które są specyficzne dla różnych poddziedzin, takich jak urządzenia elektroniczne, silniki, układy chłodzenia, lub nawet konkretnego producenta lub rynku. Możemy utworzyć bibliotekę dla każdej grupy, a następnie opublikować biblioteki w taki sposób, aby mogły być używane z wieloma zbiorami danych tekstowych. W ten sposób można dodać biblioteki, które najlepiej odpowiadają kontekstowi danych tekstowych.

Uwaga: Dodatkowe zasoby mogą być konfigurowane i zarządzane na karcie Advanced Resources. Niektóre mają zastosowanie do wszystkich bibliotek i zarządzania obiektami nielingwistycznymi, wyjątkami grupowania rozmytego i tak dalej. Ponadto na karcie Text Link Rules można edytować reguły wzorców analizy powiązań w tekście, które są charakterystyczne dla biblioteki. Więcej informacji zawiera Rozdział 17, “Informacje o zasobach zaawansowanych”, na stronie 189.

Biblioteki dostarczone z produktem

Domyślnie pewne biblioteki są instalowane razem z produktem IBM SPSS Modeler Text Analytics. Można użyć tych wstępnie sformatowanych bibliotek, aby uzyskać dostęp do tysięcy predefiniowanych terminów i synonimów, a także wielu różnych typów. Dostarczone biblioteki są zoptymalizowane pod kątem kilku różnych dziedzin i są dostępne w kilku różnych językach.

.

Istnieje wiele bibliotek, ale najczęściej używane są następujące:

- **Biblioteka lokalna.** Służy do przechowywania słowników zdefiniowanych przez użytkownika. Jest ona pustą biblioteką dodaną domyślnie do wszystkich zasobów. Zawiera też pusty słownik typów. Jest to szczególnie przydatne podczas wprowadzania zmian lub udoskonaleń do zasobów bezpośrednio (na przykład poprzez dodanie wyrazu do typu) z widoku Categories and Concepts, widoku Clusters i widoku Text Link Analysis . W tym przypadku te zmiany i udoskonalenia są automatycznie zapisywane w pierwszej bibliotece wymienionej w drzewie bibliotek edytora Resource Editor; domyślnie jest to biblioteka o nazwie *Local Library*. Nie można opublikować tej biblioteki, ponieważ jest ona charakterystyczna dla danych sesji . Jeśli chcesz opublikować jej zawartość, musisz zmienić jej nazwę.
- **Biblioteka Core.** Używana w większości przypadków, ponieważ zawiera pięć wbudowanych typów podstawowych reprezentujących osoby, miejsca, organizacje, produkty i elementy nieznane. Choć niekiedy w słownikach typów

biblioteki Core wymienione są tylko nieliczne terminy, typy te są w istocie tylko dopełnieniem wszechstronnych typów zdefiniowanych w wewnętrznych zasobach skompilowanych, które zostały dostarczone z produktem do eksploracji tekstu. Te wewnętrzne skompilowane zasoby zawierają tysiące terminów każdego typu. Z tego powodu, nawet jeśli terminu nie ma na liście słownika typu, termin ten może zostać wyodrębniony i przypisany do tupu podstawowego z biblioteki Core. Wyjaśnia to, dlaczego imiona, takie jak *George*, mogą być wyodrębniane i przypisywane do typu <Person>, mimo że w słowniku typu <Person> biblioteki Core figuruje tylko imię *John*. Podobnie, jeśli nie uwzględnisz biblioteki Core, wyniki wyodrębniania mogą nadal zawierać te typy, ponieważ mechanizm wyodrębniania nadal będzie używał zawierających je skompilowanych zasobów wewnętrznych.

- **Biblioteka opinii Opinions.** Najczęściej służy do wyodrębniania opinii i sentymentu z danych tekstowych. Biblioteka ta zawiera tysiące wyrazów oznaczających postawy, kwalifikatory i preferencje, które — gdy używane są łącznie z innymi terminami — wyrażają opinie o temacie. Ta biblioteka zawiera szereg wbudowanych typów, synonimów i wykluczeń. Zawiera także obszerny zestaw reguł wzorców używanych na potrzeby analizy powiązań w tekście. Aby można było korzystać z reguł analizy powiązań w tekście zawartych w tej bibliotece oraz z wynikowych wzorców generowanych przez te reguły, biblioteka ta musi być określona na karcie Text Link Rules. Więcej informacji zawiera Rozdział 18, “Informacje o regułach powiązań w tekście”, na stronie 201.
- **Biblioteka Budget.** Służy do wyodrębniania składników odnoszących się do kosztów czegoś. Biblioteka ta zawiera wiele wyrazów i fraz będących przymiotnikami, kwalifikatorami i wyrażającymi oceny dotyczące cen i jakości.
- **Biblioteka Variations.** Służy do uwzględniania przypadków, w których pewne warianty językowe wymagają zdefiniowania synonimów, by były prawidłowo grupowane. Ta biblioteka zawiera tylko definicje synonimów.

Mimo że niektóre z bibliotek dostarczanych poza zasobami mają zawartość zbliżoną do niektórych szablonów, szablony zostały precyzyjnie dopasowane do konkretnych zastosowań i zawierają dodatkowe zaawansowane zasoby. Zalecamy, aby podjąć próbę użycia szablonu zaprojektowanego dla rodzaju danych tekstowych, które mają być analizowane, a potem wprowadzać zmiany w tych zasobach, a nie dodawać poszczególne biblioteki do ogólnego szablonu.

Także pewne skompilowane zasoby są dostarczane z produktem IBM SPSS Modeler Text Analytics. Są one zawsze używane podczas procesu wyodrębniania i zawierają dużą liczbę definicji uzupełniających wbudowane słowniki typów w domyślnych bibliotekach. Ponieważ te zasoby są kompilowane, nie mogą być wyświetlane ani edytowane. Można jednak wymusić wpisanie terminu o typie określonym w zasobach skompilowanych do dowolnego innego słownika. Więcej informacji zawiera temat “Wymuszanie terminów” na stronie 183.

Tworzenie bibliotek

Można utworzyć dowolną liczbę bibliotek. Po utworzeniu nowej biblioteki można rozpocząć tworzenie słowników typów w tej bibliotece, a następnie wprowadzić terminy, synonimy i wykluczenia.

Aby utworzyć bibliotekę

1. Z menu wybierz opcje **Resources > New Library**. Zostanie otwarte okno dialogowe Library Properties .
2. Wprowadź nazwę biblioteki w polu tekstowym Name.
3. W razie potrzeby wpisz komentarz w polu tekstowym Annotation.
4. Kliknij przycisk **Publish**, jeśli chcesz opublikować tę bibliotekę teraz, przed wprowadzeniem jakichkolwiek zmian. Więcej informacji zawiera temat “Współużytkowanie bibliotek” na stronie 172. Można również opublikować bibliotekę później w dowolnym czasie.
5. Kliknij przycisk **OK**, aby utworzyć bibliotekę. Okno dialogowe zostanie zamknięte, a biblioteka zostanie wyświetlona w widoku drzewa. Jeśli rozwiniesz biblioteki w drzewie, zobaczysz, że biblioteka zawiera utworzony automatycznie pusty słownik typu. Można od razu rozpocząć dodawanie do niego terminów. Więcej informacji zawiera temat “Dodawanie terminów” na stronie 180.

Dodawanie bibliotek publicznych

Jeśli chcesz ponownie wykorzystać bibliotekę z innej sesji , możesz dodać ją do bieżących zasobów, o ile tylko jest biblioteką publiczną. **Biblioteka publiczna** to biblioteka, która została opublikowana. Więcej informacji zawiera temat “Publikowanie bibliotek” na stronie 173.

Ważne! Nie można dodać biblioteki japońskiej do zasobów innego języka i odwrotnie.

Gdy dodajesz bibliotekę publiczną, **lokalna** jej kopia jest osadzana w danych sesji. Możesz wprowadzić zmiany w takiej bibliotece; jednak musisz ponownie opublikować tę wersję publiczną, jeśli chcesz udostępnić zmiany w innych sesjach lub projektach.

Podczas dodawania biblioteki publicznej może pojawić się okno dialogowe Resolve Conflicts, jeśli zostaną wykryte konflikty między terminami i typami w jednej bibliotece a terminami i typami w pozostałych bibliotekach lokalnych. Użytkownik musi rozstrzygnąć te konflikty lub zaakceptować proponowane rozwiązania w celu zakończenia operacji. Więcej informacji zawiera temat “Rozstrzyganie konfliktów” na stronie 174.

Uwaga: Jeśli zawsze aktualizujesz biblioteki podczas uruchamiania interaktywnego pulpitu roboczego lub publikujesz biblioteki przy zamykaniu pulpitu , ryzyko rozsynchronizowania bibliotek będzie mniejsze. Więcej informacji zawiera temat “Współużytkowanie bibliotek” na stronie 172.

Aby dodać bibliotekę

1. Z menu wybierz opcje **Resources > Add Library**. Zostanie otwarte okno dialogowe Add Library.
2. Wybierz bibliotekę lub biblioteki z listy.
3. Kliknij przycisk **Add**. Jeśli występują konflikty między nowo dodanymi bibliotekami a jakimikolwiek bibliotekami, które istniały, pojawi się monit o rozstrzygnięcie konfliktu lub wprowadzenie zmian, aby możliwe było zakończenie operacji. Więcej informacji zawiera temat “Rozstrzyganie konfliktów” na stronie 174.

Znajdowanie terminów i typów

Można wyszukiwać w różnych oknach w edytorze przy użyciu funkcji Find. W edytorze można wybrać z menu opcję **Edit > Find**, co spowoduje wyświetlenie paska narzędzi wyszukiwania. Można użyć tego paska narzędzi, aby znajdować pojedyncze wystąpienia. Klikając ponownie przycisk **Find**, można znaleźć kolejne wystąpienia wyszukiwanego terminu.

Podczas wyszukiwania edytor przeszukuje tylko bibliotekę lub biblioteki wymienione na liście rozwijanej wyszukiwania na pasku narzędzi znajdowania. Jeśli wybrana jest opcja **All Libraries**, program przeszuka wszystkie elementy w edytorze.

Wyszukiwanie rozpoczyna się w obszarze, który jest aktywny. Wyszukiwanie jest kontynuowane w kolejnych sekcjach, wraca na początek i jest przerywane po powrocie do aktywnej komórki. Można odwrócić kolejność wyszukiwania za pomocą strzałek kierunkowych. Można również wybrać, czy w wyszukiwaniu będzie rozróżniana wielkość liter.

Aby znaleźć łańcuchy w widoku

1. Z menu wybierz opcje **Edit > Find**. Wyświetlony zostaje pasek narzędzi wyszukiwania.
2. Wprowadź łańcuch, który chcesz wyszukać.
3. Kliknij przycisk **Find**, aby rozpocząć wyszukiwanie. Zostanie wyróżnione następne wystąpienie terminu lub typu.
4. Klikaj przycisk ponownie, aby przechodzić od wystąpienia do wystąpienia.

Używanie znaku gwiazdki w terminach

Używanie znaku gwiazdki (*) w terminach jest szczególnie przydatne podczas pracy z językiem aglutynacyjnym, w którym nowe wyrazy powstają poprzez połączenie innych wyrazów bez wstawiania spacji. Na przykład w języku niemieckim wyraz *Übernachtungspreis* składa się z następujących elementów: *Übernachtung* + *s* + *Preis*.

Przykładowo, podczas wyszukiwania w terminach wyrazu preis* w typie Budget zostanie ono dopasowane do wyodrębnionych pojęć, takich jak preiserhöhung. W ten sam sposób wyraz *preis zostanie dopasowany do wyrazu Übernachtung, a *preis* do Übernachtungspreiserhöhung.

Przeglądanie bibliotek

Można wyświetlić zawartość jednej biblioteki lub wszystkich bibliotek. Ta opcja może być przydatna podczas pracy z wieloma bibliotekami lub gdy chcesz przejrzeć zawartość konkretnej biblioteki przed jej opublikowaniem. Zmiana widoku wpływa tylko na widoczność elementów na karcie Library Resources, ale nie wyklucza bibliotek z użytku podczas wyodrębniania. Więcej informacji zawiera temat “Wyłączanie bibliotek lokalnych” na stronie 171.

Domyślnym widokiem jest widok **All Libraries**, który zawiera wszystkie biblioteki w drzewie i ich zawartość w innych panelach. Wybór ten można zmienić za pomocą listy rozwijanej na pasku narzędzi lub za pomocą opcji menu (**View > Libraries**). Gdy przeglądana jest jedna biblioteka, wszystkie elementy z innych bibliotek są usuwane z widoku, ale nadal są odczytywane podczas wyodrębniania.

Aby zmienić widok bibliotek

1. Z menu na karcie Library Resources wybierz opcje **View > Libraries**. Zostanie otwarte menu ze wszystkimi lokalnymi bibliotekami.
2. Wybierz bibliotekę, które chcesz wyświetlić, lub wybierz opcję **All Libraries**, aby wyświetlić zawartość wszystkich bibliotek. Zawartość widoku zostanie odfiltrowana zgodnie z dokonanym wyborem.

Zarządzanie bibliotekami lokalnymi

Biblioteki lokalne są to biblioteki rezydujące wewnątrz sesji pracy z interaktywnym pulpitem roboczym lub wewnątrz szablonu, co odróżnia je od bibliotek publicznych. Więcej informacji zawiera temat “Zarządzanie bibliotekami publicznymi” na stronie 171. Do prostych operacji administracyjnych na bibliotekach lokalnych należą: zmiana nazwy, wyłączenie lub usunięcie biblioteki lokalnej.

Zmiana nazw bibliotek lokalnych

Można zmieniać nazwy bibliotek lokalnych. Zmiana nazwy biblioteki lokalnej spowoduje usunięcie jej powiązania z wersją publiczną, jeśli wersja publiczna istnieje. Oznacza to, że kolejne zmiany nie będą wspólne z wersją publiczną. Można ponownie opublikować tę bibliotekę lokalną pod nową nazwą. Oznacza to również, że nie będzie można zaktualizować oryginalnej wersji publicznej z uwzględnieniem wszystkich zmian, które zostaną wprowadzone do tej wersji lokalnej.

Uwaga: Nie można zmienić nazwy biblioteki publicznej.

1. Z menu wybierz opcje **Edit > Library Properties**. Zostanie otwarte okno dialogowe Library Properties.

Aby zmienić nazwę biblioteki lokalnej

1. W widoku drzewa wybierz bibliotekę, której nazwę chcesz zmienić.
2. Wprowadź nową nazwę biblioteki w polu tekstowym Name.
3. Kliknij przycisk **OK**, aby zaakceptować nową nazwę biblioteki. Okno dialogowe zostanie zamknięte, a nazwa biblioteki zostanie zaktualizowana w widoku drzewa.

Wyłączanie bibliotek lokalnych

Jeśli chcesz tymczasowo wyłączyć bibliotekę z procesu wyodrębniania, możesz usunąć zaznaczenie pola wyboru po lewej stronie nazwy biblioteki w widoku drzewa. Sygnalizuje to, że chcesz zachować bibliotekę, ale jej treść ma być pomijana podczas wykrywania konfliktów i podczas wyodrębniania.

Aby wyłączyć bibliotekę

1. W panelu drzewa bibliotek wybierz bibliotekę, którą chcesz wyłączyć.
2. Naciśnij klawisz spacji. Pole wyboru po lewej stronie nazwy nie jest zaznaczone.

Usuwanie bibliotek lokalnych

Można usunąć bibliotekę bez usuwania jej wersji publicznej i odwrotnie. Usuwanie lokalnej biblioteki spowoduje usunięcie biblioteki i całej jej zawartości z sesji . Usunięcie lokalnej wersji biblioteki nie powoduje usunięcia tej biblioteki z innych sesji, ani usunięcia wersji publicznej. Więcej informacji zawiera temat “Zarządzanie bibliotekami publicznymi”.

Aby usunąć bibliotekę lokalną

1. W widoku drzewa wybierz bibliotekę, która ma zostać usunięta.
2. Z menu wybierz opcje **Edit > Delete**, aby usunąć bibliotekę. Biblioteka zostanie usunięta.
3. Jeśli ta biblioteka nigdy nie była opublikowana, pojawi się komunikat z pytaniem, czy chcesz usunąć, czy zachować tę bibliotekę. Kliknij przycisk **Delete**, aby kontynuować, albo **Keep**, jeśli chcesz zachować tę bibliotekę.

Uwaga: Musi pozostać zawsze co najmniej jedna biblioteka.

Zarządzanie bibliotekami publicznymi

W celu ponownego wykorzystania bibliotek lokalnych można je publikować, a następnie pracować z nimi i je wyświetlić za pośrednictwem okna dialogowego Manage Libraries (**Resources > Manage Libraries**). Więcej informacji zawiera temat “Współużytkowanie bibliotek” na stronie 172. Do prostych operacji administracyjnych na bibliotekach publicznych należą: importowanie, eksportowanie lub usunięcie biblioteki publicznej. Nie można zmienić nazwy biblioteki publicznej.

Importowanie bibliotek publicznych

1. W oknie dialogowym Manage Libraries kliknij przycisk **Import...**. Zostanie otwarte okno dialogowe Import Library.
2. Wybierz plik biblioteki (**.lib*), który ma zostać zaimportowany, a jeśli chcesz dodać tę bibliotekę także lokalnie, wybierz opcję **Add library to current project**.
3. Kliknij przycisk **Import**. Okno dialogowe zostanie zamknięte. Jeśli biblioteka publiczna o takiej nazwie już istnieje, pojawi się prośba o zmianę nazwy biblioteki, która jest importowana, albo zgodę na nadpisanie bieżącej biblioteki publicznej.

Eksportowanie bibliotek publicznych

Można eksportować biblioteki publiczne w formacie *.lib*, dzięki czemu można je współużytkować.

1. W oknie dialogowym Manage Libraries wybierz na liście biblioteki, które chcesz wyeksportować.
2. Kliknij przycisk **Export**. Zostanie otwarte okno dialogowe Select Directory.
3. Wybierz katalog, do którego chcesz wyeksportować bibliotekę, i kliknij przycisk **Export**. Okno dialogowe zostanie zamknięte, a plik biblioteki (**.lib*) zostanie wyeksportowany.

Usuwanie bibliotek publicznych

Można usunąć bibliotekę lokalną bez usuwania wersji publicznej biblioteki i odwrotnie. Jeśli jednak biblioteka zostanie usunięta w tym oknie dialogowym, nie można jej będzie dodawać zasobów sesji, dopóki lokalna wersja nie zostanie opublikowana ponownie.

Jeśli usuniesz bibliotekę, która została zainstalowana z produktem, odtworzona zostanie pierwotnie zainstalowana wersja.

1. W oknie dialogowym Manage Libraries wybierz bibliotekę, która ma zostać usunięta. Listę można posortować, klikając odpowiedni nagłówek.
2. Kliknij przycisk **Delete**, aby usunąć bibliotekę. IBM SPSS Modeler Text Analytics sprawdza, czy lokalna wersja biblioteki jest taka sama, jak biblioteka publiczna. Jeśli tak, biblioteka jest usuwana bez alertu. Jeśli wersje biblioteki różnią się, zostanie otwarty alert z pytaniem, czy chcesz zachować, czy usunąć wersję publiczną.

Współużytkowanie bibliotek

Biblioteki umożliwiają wykorzystanie tych samych zasobów w różnych sesjach pracy z interaktywnym pulpitem roboczym . Biblioteki mogą znajdować się w dwóch stanach lub wersjach. Biblioteki edytowalne w edytorze i będące częścią sesji pracy z interaktywnym pulpitem roboczym są nazywane **bibliotekami lokalnymi**. Podczas pracy z interaktywnym pulpitem roboczym możesz na przykład wprowadzić liczne zmiany w bibliotece *Vegetables*. Jeśli uznasz, że te zmiany będą przydatne również do pracy z innymi danymi, możesz udostępnić zmienione zasoby, tworząc **publiczną** wersję biblioteki *Vegetables*. Biblioteka publiczna, jak sama nazwa wskazuje, jest dostępna dla wszelkich innych zasobów w dowolnych sesjach pracy z interaktywnym pulpitem roboczym.

Można wyświetlać biblioteki publiczne w oknie dialogowym Manage Libraries. Jeśli istnieje wersja publiczna biblioteki, można ją dodać do zasobów w innych kontekstach, dzięki czemu niestandardowe zasoby lingwistycznych mogą być współużytkowane.

Biblioteki dostarczone z produktem są od początku bibliotekami publicznymi. Możliwe jest edytowanie zasobów w tych bibliotekach, a następnie utworzenie nowej wersji publicznej. Te nowe wersje będą dostępne w innych sesjach pracy z interaktywnym pulpitem roboczym.

W miarę kontynuowania pracy z bibliotekami i wprowadzania zmian wersje bibliotek będą traciły wzajemną synchronizację. W niektórych przypadkach wersja lokalna może być nowsza od wersji publicznej, a w innych przypadkach wersja publiczna może być nowsza od wersji lokalnej. Zarówno wersja publiczna, jak i wersja lokalna może zawierać zmiany, których druga wersja nie zawiera. Taka sytuacja wystąpi, jeśli wersja publiczna zostanie zaktualizowana z poziomu innej sesji pracy z interaktywnym pulpitem roboczym. Jeśli wersje bibliotek rozsynchronizują się, można zsynchronizować je ponownie. Synchronizowanie wersji bibliotek polega na ponownym opublikowaniu i/lub zaktualizowaniu bibliotek lokalnych.

Przy każdym uruchomieniu lub zamknięciu interaktywnego pulpitu roboczego pojawi się monit o zsynchronizowanie wszystkich bibliotek, które wymagają aktualizacji lub ponownej publikacji. Dodatkowo można łatwo ustalić stan synchronizacji z biblioteki lokalnej za pomocą ikony znajdującej się obok nazwy biblioteki w widoku drzewa lub w oknie dialogowym Library Properties. Możesz również zrobić to w dowolnym momencie za pomocą opcji menu. Poniższa tabela zawiera opis pięciu możliwych stanów i powiązanych z nimi ikon.

Tabela 37. Stany synchronizacji biblioteki lokalnej.

Ikona	Opis statusu biblioteki lokalnej
	Nieopublikowana — biblioteka lokalna nie została nigdy opublikowana.
	Zsynchronizowana — lokalna i publiczna wersja biblioteki są identyczne. Dotyczy to również *biblioteki lokalnej*, której nie można opublikować, ponieważ z założenia zawiera tylko zasoby charakterystyczne dla sesji.
	Nieaktualna — wersja publiczna biblioteki jest nowsza od wersji lokalnej. Można zaktualizować wersję lokalną z uwzględnieniem zmian.

Tabela 37. Stany synchronizacji biblioteki lokalnej (kontynuacja).

Ikona	Opis statusu biblioteki lokalnej
	Nowsza — wersja lokalna biblioteki jest nowsza niż wersja publiczna. Można ponownie opublikować lokalną wersję do wersji publicznej.
	Rozsynchronizowane — i biblioteka publiczna, i lokalna, zawiera zmiany, których nie ma w drugiej bibliotece. Musisz zdecydować, czy zaktualizować, czy opublikować bibliotekę lokalną. W przypadku aktualizacji zostaną utracone wszystkie zmiany wprowadzone od czasu ostatniej aktualizacji lub publikacji. W przypadku publikowania zmiany w wersji publicznej zostaną nadpisane.

Uwaga: Jeśli zawsze aktualizujesz biblioteki podczas uruchamiania interaktywnego pulpitu roboczego lub publikujesz biblioteki przy zamykaniu pulpituotwierania lub publikujesz przy zamykaniu projektu, ryzyko rozsynchronizowania bibliotek będzie mniejsze.

Możesz w dowolnej chwili ponownie opublikować bibliotekę, gdy uznasz, że wprowadzone zmiany byłyby korzystne także w innych zawierających ją strumieniach . Jeśli zmiany byłyby korzystne w innych strumieniach, można następnie zaktualizować lokalne wersje tych strumieni. W ten sposób można tworzyć strumienie dla każdego kontekstu lub każdej dziedziny związanej z danymi, tworząc nowe biblioteki i/lub dodając do zasobów dowolną liczbę bibliotek publicznych.

Jeśli publiczna wersja biblioteki jest współużytkowana, istnieje większa szansa, że pojawią się różnice między lokalną a publiczną wersją. Za każdym razem, gdy uruchamiasz lub zamykasz i publikujesz z sesji pracy z interaktywnym pulpitem roboczym bądź otwierasz lub zamykasz z poziomu formularza szablonu Template Editor , wyświetlany jest komunikat umożliwiający opublikowanie i/lub zaktualizowanie bibliotek, których wersje nie są zsynchronizowane z wersjami w oknie dialogowym Manage Libraries. Jeśli publiczna wersja biblioteki jest nowsza niż wersja lokalna, pojawi się okno dialogowe z pytaniem, czy chcesz przeprowadzić aktualizację. Możesz zachować wersję lokalną bez zmian zamiast aktualizować ją na podstawie wersji publicznej albo scalić aktualizacje do biblioteki lokalnej.

Publikowanie bibliotek

Jeśli określona biblioteka nie była publikowana, to publikowanie polega na utworzeniu publicznej kopii biblioteki lokalnej w bazie danych. W przypadku ponownego publikowania biblioteki zawartość biblioteki lokalnej zastąpi istniejącą zawartość wersji publicznej. Po ponownej publikacji można zaktualizować tę bibliotekę w dowolnych innych sesjach strumienia, aby ich wersje lokalne były zsynchronizowane z wersją publiczną Mimo że możliwe jest opublikowanie biblioteki, wersja lokalna zawsze przechowywana jest w sesji .

Ważne! Jeśli wprowadzisz zmiany w bibliotece lokalnej, a w międzyczasie wersja publiczna biblioteki zostanie również zmieniona, biblioteka jest traktowana jako rozsynchronizowana. Zaleca się rozpoczęcie od zaktualizowania wersji lokalnej z uwzględnieniem zmian publicznych, wprowadzenie żądanych zmian, a następnie ponowne opublikowanie wersji lokalnej, aby obie wersje były identyczne. Jeśli najpierw wprowadzisz zmiany i opublikujesz bibliotekę, spowoduje nadpisanie zmian wprowadzonych w wersji publicznej.

Aby opublikować biblioteki lokalne w bazie danych

1. Z menu wybierz opcje **Resources > Publish Libraries**. Zostanie otwarte okno dialogowe Publish Libraries, a wszystkie biblioteki wymagające opublikowania będą domyślnie wybrane.
2. Zaznacz pola wyboru po lewej stronie bibliotek, które chcesz opublikować.
3. Kliknij przycisk **Publish**, aby opublikować biblioteki w bazie danych zarządzania bibliotekami.

Aktualizowanie bibliotek

Gdy uruchamiasz lub zamykasz sesję pracy z interaktywnym pulpitem roboczym, możesz zaktualizować lub opublikować biblioteki, które są rozsynchronizowane z ich wersjami publicznymi. Jeśli publiczna wersja biblioteki jest nowsza niż wersja lokalna, pojawi się okno dialogowe z pytaniem, czy chcesz przeprowadzić aktualizację. Możesz zachować wersję lokalną bez zmian zamiast aktualizować ją na podstawie wersji publicznej albo zastąpić wersję lokalną wersją publiczną. Jeśli publiczna wersja biblioteki jest nowsza od wersji lokalnej, możesz zaktualizować wersję

lokalną, aby zsynchronizować jej zawartość z zawartością wersji publicznej. Aktualizacja polega na uwzględnieniu w wersji lokalnej wszelkich zmian znalezionych w wersji publicznej.

Uwaga: Jeśli zawsze aktualizujesz biblioteki podczas uruchamiania interaktywnego pulpitu roboczego lub publikujesz biblioteki przy zamykaniu pulpitu otwierania lub publikujesz przy zamykaniu projektu, ryzyko rozsynchronizowania bibliotek będzie mniejsze. Więcej informacji zawiera temat “Współużytkowanie bibliotek” na stronie 172.

Aby zaktualizować biblioteki lokalne

1. Z menu wybierz opcje **Resources > Update Libraries**. Zostanie otwarte okno dialogowe Update Libraries, a wszystkie biblioteki wymagające aktualizacji będą domyślnie wybrane.
2. Zaznacz pola wyboru po lewej stronie bibliotek, które chcesz opublikować.
3. Kliknij przycisk **Update**, aby zaktualizować biblioteki lokalne.

Rozstrzyganie konfliktów

Konflikty między biblioteką lokalną a publiczną

Za każdym razem, gdy uruchamiasz sesję strumienia, IBM SPSS Modeler Text Analytics porównuje biblioteki lokalne z wymienionymi w oknie dialogowym Manage Libraries. Jeśli jakiekolwiek biblioteki lokalne w sesji nie są zsynchronizowane z wersjami opublikowanymi, otwierane jest okno dialogowe Library Synchronization Warning. Można wybrać spośród następujących opcji, aby zdecydować, które wersje bibliotek mają być używane:

- **All libraries local to file.** Ta opcja zachowuje wszystkie biblioteki lokalne bez zmian. Zawsze można ponownie opublikować lub zaktualizować je później.
- **All published libraries on this machine.** Ta opcja spowoduje zastąpienie widocznych bibliotek lokalnych wersjami z bazy danych.
- **All more recent libraries.** Ta opcja spowoduje zastąpienie starszych bibliotek lokalnych nowszymi wersjami publicznymi z bazy danych.
- **Inne.** Ta opcja umożliwia ręczne wybranie wersji w tabeli.

Konflikty między terminami objętymi wymuszeniem

Gdy dodajesz bibliotekę publiczną lub aktualizujesz bibliotekę lokalną, mogą zostać wykryte terminy i typy kolidujące z terminami i typami w innych bibliotekach należących do zasobów albo będące duplikatami. W takim przypadku pojawi się prośba o zweryfikowanie zgłoszonych konfliktów lub wprowadzenie zmian przed dokończeniem operacji w oknie dialogowym Edit Forced Terms. Więcej informacji zawiera temat “Wymuszanie terminów” na stronie 183.

Okno dialogowe Edit Forced Terms zawiera wszystkie pary kolidujących terminów lub typów. Aby ułatwić rozróżnianie par, wyświetlane są one na tle w naprzemiennych kolorach. Te kolory można zmieniać w oknie dialogowym Options. Więcej informacji zawiera temat “Okno dialogowe Options: karta Display” na stronie 75. Okno dialogowe Edit Forced Terms zawiera dwie karty:

- **Duplicates.** Ta karta zawiera zduplikowane terminy znalezione w bibliotekach. Jeśli za terminem wyświetlana jest ikona pinezki, oznacza to, że to wystąpienie terminu objęte jest wymuszeniem. Czarny znak X oznacza, że to wystąpienie terminu zostanie pominięte podczas wyodrębniania, ponieważ zostało objęte wymuszeniem w innym miejscu.
- **User Defined.** Ta karta zawiera listę terminów, które objęto ręcznym wymuszeniem na panelu słownika typu, a nie w wyniku rozstrzygnięcia konfliktów.

Uwaga: Okno dialogowe Edit Forced Terms otwiera się po zakończeniu dodawania lub aktualizacji biblioteki. Anulowanie tego okna dialogowego nie jest równoznaczne z anulowaniem aktualizacji lub dodawania biblioteki.

Aby rozstrzygnąć konflikty

1. W oknie dialogowym Edit Forced Terms wybierz przełącznik w kolumnie Use dla terminu, który chcesz objąć wymuszeniem.

2. Po zakończeniu kliknij przycisk **OK**, aby zastosować wymuszenia terminów i zamknąć okno dialogowe. Kliknięcie przycisku **Cancel** spowoduje anulowanie zmian wprowadzonych w tym oknie dialogowym.

Rozdział 16. Informacje o słownikach w bibliotekach

Zasoby używane do wyodrębniania danych tekstowych są przechowywane w postaci szablonów i bibliotek. Biblioteka może składać się z trzech słowników.

- **Słownik typu** zawiera zbiór terminów pogrupowanych pod jedną etykietą lub nazwą typu. Odczytując dane tekstowe, mechanizm wyodrębniania porównuje wyrazy napotkane w tekście z terminami w słownikach typów. W trakcie wyodrębniania terminy i synonimy w odmienionych formach są grupowane pod jednym terminem docelowym nazywanym pojęciem. Wyodrębnione pojęcia są przypisywane do słownika typu, w którym figurują jako terminy. Słownikami typów można zarządzać w lewym górnym i środkowym panelu edytora — w panelu drzewa bibliotek i panelu terminów. Więcej informacji zawiera temat “Słowniki typów”.
- **Słownik zastąpień** zawiera zbiór wyrazów zdefiniowanych jako synonimy lub elementy opcjonalne służących do grupowania podobnych terminów pod jednym terminem docelowym, nazywanym pojęciem. Słownikami zastąpień można zarządzać w lewym dolnym panelu edytora na karcie Synonyms i na karcie Optional. Więcej informacji zawiera temat “Słowniki zastąpień/synonimów” na stronie 184.
- **Słownik wykluczeń** zawiera zbiór terminów i typów, które będą usuwane z ostatecznych wyników wyodrębniania. Słownikami wykluczeń można zarządzać w prawym panelu edytora. Więcej informacji zawiera temat “Słowniki wykluczeń” na stronie 187.

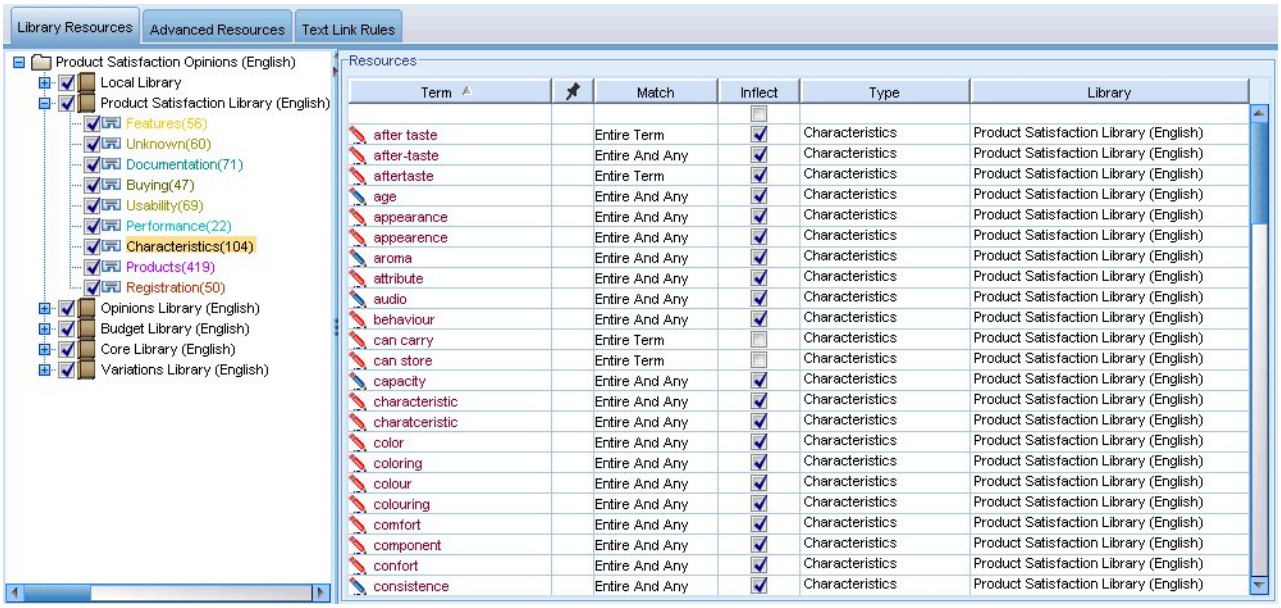
Więcej informacji zawiera temat Rozdział 15, “Praca z bibliotekami”, na stronie 167.

Słowniki typów

Słownik typu składa się z nazwy lub etykiety typu oraz listy terminów. Słownikami typów zarządza się w lewym górnym i środkowym panelu karty Library Resources w edytorze. Dostęp do tego widoku można uzyskać, wybierając opcję **View > Resource Editor** w menu, jeśli jest aktywna sesja interaktywnego pulpitu roboczego. Jeśli nie jest, można edytować słowniki konkretnego szablonu w oknie Template Editor.

Odczytując dane tekstowe, mechanizm wyodrębniania porównuje wyrazy napotkane w tekście z terminami w słownikach typów. Terminy są to wyrazy lub frazy figurujące w słownikach typów w zasobach lingwistycznych.

Gdy wyraz pasuje do terminu, jest przypisywany do nazwy typu właściwej dla tego terminu. W miarę odczytywania zasobów w trakcie wyodrębniania terminy, które zostały znalezione w tekście, przechodzą przez kilka kroków przetwarzania, zanim staną się pojęciami w panelu Extraction Results. Jeśli mechanizm wyodrębniania uzna kilka terminów należących do tego samego słownika typu za synonimy, to zostaną one zgrupowane pod najczęściej występującym terminem, który nazywamy *pojęciem*. Na przykład terminy question i query mogłyby ostatecznie pojawić się w wynikach jako pojęcie question.



Rysunek 40. Panele drzewa bibliotek i terminów

Lista słowników typów jest wyświetlana w panelu drzewa bibliotek po lewej stronie. Zawartość każdego typu słownika jest wyświetlana w panelu środkowym. Słowniki typu są czymś więcej niż tylko listami terminów. Sposób dopasowywania wyrazów i fraz w danych tekstowych do terminów zdefiniowanych w słownikach typów zależy od wybranej opcji dopasowywania. **Opcja dopasowywania** określa, w jaki sposób termin zostanie zakotwiczony względem kandydackiego (potencjalnego) wyrazu lub frazy w danych tekstowych. Więcej informacji zawiera temat “Dodawanie terminów” na stronie 180.

Uwaga: Nie wszystkie opcje, takie jak opcja dopasowywania i formy odmienione, mają zastosowanie w tekście japońskim.

Dodatkowo można uzupełniać terminy w słowniku typu, określając, czy mają być automatycznie generowane i dodawane ich odmienione formy. Generując formy odmienione, można automatycznie dodać formy liczby mnogiej do terminów w liczbie pojedynczej, formy w liczbie pojedynczej do terminów liczby mnogiej, a także dodać przymiotniki do słownika typu. Więcej informacji zawiera temat “Dodawanie terminów” na stronie 180.

Uwaga: W większości języków pojęcia nieznalezione w żadnym słowniku typu, ale wyodrębnione z tekstu, otrzymują automatycznie przypisywany typ <Unknown>

Używanie znaku gwiazdki w terminach

Używanie znaku gwiazdki (*) w terminach jest szczególnie przydatne podczas pracy z językiem aglutynacyjnym, w którym nowe wyrazy powstają poprzez połączenie innych wyrazów bez wstawiania spacji. Na przykład w języku niemieckim wyraz *Übernachtungspreis* składa się z następujących elementów: *Übernachtung* + *s* + *Preis*.

Przykładowo, podczas wyszukiwania w terminach wyrazu preis* w typie Budget zostanie ono dopasowane do wyodrębnionych pojęć, takich jak preiserhöhung. W ten sam sposób wyraz *preis zostanie dopasowany do wyrazu Übernachtung, a *preis* do Übernachtungspreiserhöhung.

Typy wbudowane

IBM SPSS Modeler Text Analytics jest dostarczany z zestawem zasobów lingwistycznych w formie bibliotek i skompilowanych zasobów. Dostarczone biblioteki zawierają zestaw wbudowanych słowników typów, takich jak <Location>, <Organization>, <Person> i <Product>.

Uwaga: Zestaw domyślnych wbudowanych typów jest inny w przypadku języka japońskiego.

Mechanizm wyodrębniania używa tych słowników typów do przypisywania typów do pojęć, np. typu <Location> do pojęcia paris. Mimo że w gotowych słownikach typów zdefiniowano dużą liczbę terminów, nie uwzględniają one wszystkich możliwości. W związku z tym słowniki można uzupełniać lub tworzyć własne słowniki. Aby zapoznać się z opisem zawartości określonego słownika typu dostarczonego z produktem, przeczytaj adnotację w oknie dialogowym Type Properties. Wybierz typ w drzewie i wybierz opcje **Edit > Properties** z menu kontekstowego.

Uwaga: Oprócz bibliotek dostarczonych z produktem również zasoby skompilowane (z których również korzysta mechanizm wyodrębniania) zawierają duża liczbę definicji komplementarnych wobec wbudowanych słowników. Jednak zawartość tych zasobów nie jest widoczna w produkcie. Można jednak wymusić wpisanie terminu o typie określonym w zasobach skompilowanych do dowolnego innego słownika. Więcej informacji zawiera temat “Wymuszanie terminów” na stronie 183.

Tworzenie typów

Można tworzyć słowniki typów, które pomagają w grupowaniu podobnych terminów. Gdy podczas wyodrębniania zostanie napotkany termin występujący w tym słowniku, zostanie przypisany do typu i wyodrębniony pod nazwą pojęcia. Gdy tworzysz bibliotekę, tworzony jest w niej od razu pusty słownik typu, do którego od razu można zacząć wprowadzać terminy.

Ważne!: Nie można tworzyć nowych typów dla zasobów języka japońskiego.

Jeśli analizujesz tekst o żywności i chcesz pogrupować terminy związane z warzywami, możesz utworzyć własny słownik typu <Vegetables>. Możesz do niego dodać wyrazy carrot, broccoli i spinach, jeśli uważasz, że są to ważne terminy, które wystąpią w tekście. Jeśli podczas wyodrębniania zostanie napotkany którykolwiek z tych terminów, to zostanie wyodrębniony jako pojęcie i przypisany do typu <Vegetables>.

Nie musisz definiować każdej formy wyrazu lub wyrażenia, ponieważ istnieje możliwość automatycznego generowania form odmienionych terminów. Wybranie tej opcji spowoduje, że mechanizm wyodrębniania automatycznie będzie rozpoznawał formy pojedyncze i mnogie oraz inne formy należące do tego typu. Ta opcja jest szczególnie przydatna, gdy typ zawiera głównie rzeczowniki, ponieważ mało prawdopodobne jest, by odmienione formy czasowników lub przymiotników były interesujące dla analityka.

Okno dialogowe Type Properties zawiera następujące pola.

Nazwa. Nazwa nadawana tworzonemu słownikowi typu. Zaleca się, aby nie używać spacji w nazwach typów, zwłaszcza jeśli co najmniej dwie nazwy typu rozpoczynają się od tego samego wyrazu.

Uwaga: Istnieją ograniczenia dotyczące nazw typów i użycia symboli. Na przykład nie należy używać symboli, takich jak "@" lub "!" w nazwach.

Default match. Domyślny domyślnego dopasowania instruuje mechanizm wyodrębniania co do sposobu dopasowywania tego terminu do danych tekstowych. Każdorazowo po dodaniu terminu do tego słownika typu jest to atrybut dopasowania automatycznie przypisywany do tego terminu. Zawsze można zmienić sposób dopasowania ręcznie na liście terminów. Dostępne są opcje: **Entire Term**, **Start**, **End**, **Any**, **Start or End**, **Entire and Start**, **Entire and End**, **Entire and (Start or End)** i **Entire (no compounds)**. Więcej informacji zawiera temat “Dodawanie terminów” na stronie 180. Ta opcja nie ma zastosowania do zasobów japońskich.

Add to. To pole wskazuje bibliotekę, w której zostanie utworzony nowy słownik typów.

Generate inflected forms by default. Ta opcja nakazuje mechanizmowi wyodrębniania wykorzystanie analizy morfologii gramatycznej do wychwytywania i grupowania podobnych form terminów, które zostały dodane do tego słownika, np. form pojedynczych i mnogich. Ta opcja jest szczególnie użyteczna, gdy typ zawiera przede wszystkim rzeczowniki. Po wybraniu tej opcji wszystkie nowe terminy dodawane do tego typu będą automatycznie miały włączoną tę opcję, ale można ją zmienić ręcznie na liście. Ta opcja nie ma zastosowania do zasobów japońskich.

Font color. To pole umożliwia odróżnienie wyników konkretnego typu od innych w interfejsie. Wybranie opcji **Use parent color** spowoduje, że dla tego słownika typów będzie używany domyślny kolor typu. Ten kolor domyślny jest ustawiony w oknie dialogowym opcji. Więcej informacji zawiera temat “Okno dialogowe Options: karta Display” na stronie 75. Jeśli wybrano opcję **Custom**, wybierz kolor z listy rozwijanej.

Annotation. To pole jest opcjonalne i może być używane w przypadku dowolnych komentarzy i opisów.

Aby utworzyć słownik typu

1. Wybierz bibliotekę, w której chcesz utworzyć nowy słownik typu.
2. Z menu wybierz opcje **Tools > New Type**. Zostanie otwarte okno dialogowe Type Properties.
3. Wprowadź nazwę typu słownika w polu tekstowym **Name** i wybierz żądane opcje.
4. Kliknij przycisk **OK**, aby utworzyć słownik typu. Nowy typ pojawi się w panelu drzewa bibliotek i w panelu środkowym. Można od razu rozpocząć dodawanie terminów. Aby uzyskać więcej informacji, patrz “Dodawanie terminów”.

Uwaga: Ta instrukcja dotyczy wprowadzania zmian w widoku Resource Editor lub oknie Template Editor. Należy pamiętać, że optymalizację można też prowadzić bezpośrednio z panelu Extraction Results, panelu Data, panelu Categories lub okna dialogowego Cluster Definitions w innych widokach. Więcej informacji zawiera temat “Optymalizacja wyników wyodrębniania” na stronie 87.

Dodawanie terminów

Panel drzewa biblioteki przedstawia biblioteki i można go rozwinąć, aby wyświetlić słowniki typów zawarte w tych bibliotekach. W środkowym panelu znajduje się lista terminów zawartych w wybranej bibliotece lub słowniku typu, w zależności od wyboru dokonanego w drzewie.

Ważne! Terminy dla zasobów w języku japońskim definiuje się w inny sposób.

W edytorze Resource Editor można dodawać terminy do słownika typów bezpośrednio w panelu terminu lub za pomocą okna dialogowego Add New Terms. Dodawane terminy mogą być pojedynczymi wyrazami lub terminami złożonymi. Na początku listy zawsze dostępny jest pusty wiersz, w którym można dodać nowy termin.

Uwaga: Ta instrukcja dotyczy wprowadzania zmian w widoku Resource Editor lub oknie Template Editor. Należy pamiętać, że optymalizację można też prowadzić bezpośrednio z panelu Extraction Results, panelu Data, panelu Categories lub okna dialogowego Cluster Definitions w innych widokach. Więcej informacji zawiera temat “Optymalizacja wyników wyodrębniania” na stronie 87.

Kolumna Term

W tej kolumnie należy wprowadzić pojedynczy wyraz lub termin złożony. Kolor terminu zależy od koloru typu, do którego termin jest przypisany (automatycznie lub przez wymuszenie). Można zmienić kolory typów w oknie dialogowym Type Properties. Więcej informacji zawiera temat “Tworzenie typów” na stronie 179.

Kolumna Force

Kliknięcie i umieszczenie ikony pinezki w tej komórce informuje mechanizm wyodrębniania, że ma zignorować wszelkie pozostałe wystąpienia tego samego terminu w innych bibliotekach. Więcej informacji zawiera temat “Wymuszanie terminów” na stronie 183.

Kolumna Match

W tej kolumnie wybierz opcję dopasowywania, aby poinstruować mechanizm wyodrębniania co do sposobu dopasowywania tego terminu do danych tekstowych. Przykłady zamieszczono w tabeli. Wartość domyślną można

zmienić, edytując właściwości typu. Więcej informacji zawiera temat “Tworzenie typów” na stronie 179. Z menu wybierz opcje **Edit > Change Match**. Poniżej znajdują się podstawowe opcje dopasowania; możliwe są także ich kombinacje:

- **Start**. Jeśli termin w słowniku pasuje do pierwszego wyrazu w pojęciu wyodrębnionym z tekstu, zostanie mu przypisany ten typ. Na przykład, jeśli wprowadzisz apple, to znalezione zostanie wyrażenie apple tart.
- **End**. Jeśli termin w słowniku pasuje do ostatniego wyrazu w pojęciu wyodrębnionym z tekstu, zostanie mu przypisany ten typ. Na przykład, jeśli wprowadzisz apple, to znalezione zostanie wyrażenie cider apple.
- **Any**. Jeśli termin w słowniku pasuje do dowolnego wyrazu w pojęciu wyodrębnionym z tekstu, zostanie mu przypisany ten typ. Na przykład, jeśli wprowadzisz apple, to opcja **Any** spowoduje przypisanie tego samego typu pojęciom apple tart, cider apple i cider apple tart.
- **Entire Term**. Jeśli całe pojęcie wyodrębnione z tekstu dokładnie pasuje do terminu ze słownika, to zostanie mu przypisany ten typ. Dodanie terminu z opcją **Entire term**, **Entire and Start**, **Entire and End**, **Entire and Any** lub **Entire (no compounds)** wymusza wyodrębnienie terminu.

  Ponadto, ponieważ typ <Person> wyodrębnia tylko imiona/nazwiska dwuczęściowe (np. *edith piaf* lub *mohandas gandhi*), celowe bywa jawne dodanie imion do tego słownika typu, jeśli chcesz, by wyodrębniane były imiona figurujące w tekście bez nazwisk. Na przykład, jeśli chcesz wychwycić wszystkie wystąpienia *edith* jako imiona osoby, dodaj edith do typu <Person> z opcją **Entire term** lub **Entire and Start**.
- **Entire (no compounds)**. Jeśli całe pojęcie wyodrębnione z tekstu dokładnie pasuje do terminu ze słownika, to zostanie mu przypisany ten typ, a wyodrębnianie zostanie zatrzymane, aby termin nie został dopasowany do dłuższego złożenia. Na przykład, jeśli wprowadzisz apple, opcja **Entire (no compound)** spowoduje przypisanie do typu terminu apple, ale nie spowoduje wyodrębnienia złożenia apple sauce, chyba że jest wymuszone gdzie indziej.

Analizując poniższą tabelę, załóżmy że termin apple figuruje w słowniku typu. W tabeli przedstawiono, które pojęcia byłyby wyodrębnione i opatrzone typem, gdyby zostały znalezione w tekście przy obowiązywaniu różnych opcji dopasowywania.

Tabela 38. Przykłady dopasowywania.

Opcje dopasowywania dla terminu:  **apple**	**Wyodrębnione pojęcia**			
	apple	**apple** *tart*	*ripe* **apple**	*homemade* **apple** *tart*
Entire Term	✔			
Start		✔		
End			✔	
Start or End		✔	✔	
Entire and Start	✔	✔		
Entire and End	✔		✔	

Tabela 38. Przykłady dopasowywania (kontynuacja).

Opcje dopasowywania dla terminu:  **apple**	**Wyodrębnione pojęcia**			
	apple	**apple** *tart*	*ripe* **apple**	*homemade* **apple** *tart*
Entire and (Start or End)	✔	✔	✔	
Any		✔	✔	✔
Entire and Any	✔	✔	✔	✔
Entire (no compounds)	✔	*nigdy nie wyodrębnione*	*nigdy nie wyodrębnione*	*nigdy nie wyodrębnione*

Kolumna Inflect

W tej kolumnie wybierz, czy mechanizm wyodrębniania powinien generować odmienione formy terminu podczas wyodrębniania, tak aby były grupowane ze sobą. Domyślna wartość dla tej kolumny jest zdefiniowana we właściwościach typu (okno Type Properties), ale można ją zmieniać dla indywidualnych przypadków w tej kolumnie. Z menu wybierz opcje **Edit > Change Inflection**.

Kolumna Type

W tej kolumnie należy wybrać słownik typu z listy rozwijanej. Lista typów jest filtrowana na podstawie wyboru dokonanego w panelu drzewa bibliotek. Pierwszy typ w liście to zawsze domyślny typ wybrany w oknie drzewa bibliotek. Z menu wybierz opcje **Edit > Change Type**.

Kolumna Library

W tej kolumnie podana jest nazwa biblioteki, w której zapisany jest termin. Możesz przeciągnąć i upuścić termin do innego typu w panelu drzewa bibliotek, aby przenieść go do innej biblioteki.

Aby dodać pojedynczy termin do słownika typu

1. W panelu drzewa bibliotek wybierz słownik typu, do którego ma zostać dodany termin.
2. Na liście terminów w panelu środkowym wpisz termin w pierwszej dostępnej pustej komórce i ustaw opcje dla tego terminu.

Aby dodać wiele terminów do słownika typu

1. W panelu drzewa bibliotek wybierz słownik typu, do którego chcesz dodać terminy.
2. Z menu wybierz opcje **Tools > New Terms**. Zostanie otwarte okno dialogowe Add New Terms.
3. Wprowadź terminy, które chcesz dodać do wybranego słownika typu, wpisując terminy lub wklejając zestaw terminów. Aby wprowadzić wiele terminów, należy oddzielić je za pomocą ogranicznika, który jest zdefiniowany w oknie dialogowym Options, lub dodać każdy termin w osobnym wierszu. Więcej informacji zawiera temat “Określanie opcji” na stronie 74.
4. Kliknij przycisk **OK**, aby dodać terminy do słownika. Dla tej biblioteki typu automatycznie wybrana została domyślna opcja dopasowywania. Okno dialogowe zostanie zamknięte, a nowe terminy pojawią się w słowniku.

Wymuszanie terminów

Jeśli chcesz, aby termin był przypisywany do konkretnego typu, możesz dodać go do odpowiedniego słownika typu. Jeśli jednak istnieje wiele terminów tej samej nazwie, mechanizm wyodrębniania musi wiedzieć, który typ ma być używany. Oznacza to, że pojawi się prośba o wybranie, który typ ma być używany. Funkcja ta jest nazywana **wymuszaniem** typu terminu. Ta opcja jest najbardziej użyteczna, gdy chcesz przesłonić przypisanie typu przyjęte w skompilowanym słowniku (wewnętrznym i nieedytowalnym). Ogólnie rzecz biorąc, zaleca się unikanie zduplikowanych terminów.

Wymuszenie nie spowoduje *usunięcia* pozostałych wystąpień terminu, zostaną one jedynie zignorowane przez mechanizm wyodrębniania. Można później zmienić wybór wystąpienia, które powinno być używane, przez wymuszenie lub anulowanie wymuszenia terminu. Może być też konieczne wymuszenie przypisania terminu do słownika typu podczas dodawania biblioteki publicznej lub aktualizacji biblioteki publicznej.

Można sprawdzić, które terminy są objęte wymuszeniem lub ignorowane, w kolumnie Force (druga kolumna w panelu terminu). Jeśli wyświetlana jest ikona pinezki, oznacza to, że to wystąpienie terminu objęte jest wymuszeniem. Czarny znak X oznacza, że to wystąpienie terminu zostanie pominięte podczas wyodrębniania, ponieważ zostało objęte wymuszeniem w innym miejscu. Dodatkowo, w przypadku wymuszenia typu terminu, termin zostanie wyświetlony w kolorze wymuszonego typu. Oznacza to, że jeśli dla terminu należącego zarówno do typu Typ 1, jak i Typ 2, wymuszono Typ 1, to wszystkie wystąpienia tego terminu w oknie będą wyświetlane w kolorze właściwym dla typu Typ 1.

Można dwukrotnie kliknąć ikonę w celu zmiany statusu. Jeśli termin występuje w innym miejscu, zostanie otwarte okno dialogowe Resolve Conflict i należy zdecydować, które wystąpienie ma być używane.

Zmiana nazw typów

Można zmienić nazwę słownika typu lub zmienić inne ustawienia słownika, edytując właściwości typu.

Ważne! Zaleca się, aby nie używać spacji w nazwach typów, zwłaszcza jeśli co najmniej dwie nazwy typu rozpoczynają się od tego samego wyrazu. Zalecamy również, aby nie zmieniać nazw typów w bibliotekach Core lub Opinions, ani nie zmieniać ich domyślnych atrybutów dopasowania.

Aby zmienić nazwę typu

1. W panelu drzewa bibliotek wybierz słownik typu, którego nazwę chcesz zmienić.
2. Kliknij prawym przyciskiem myszy i wybierz opcję **Type Properties** z menu kontekstowego. Zostanie otwarte okno dialogowe Type Properties.
3. Wprowadź nową nazwę słownika typu w polu tekstowym Name.
4. Kliknij przycisk **OK**, aby zaakceptować nową nazwę. Nowa nazwa typu będzie widoczna w panelu drzewa bibliotek.

Przenoszenie typów

Można przeciągnąć słownik typu do innej lokalizacji w bibliotece lub do innej biblioteki w drzewie.

Aby zmienić położenie typu w obrębie biblioteki

1. W panelu drzewa bibliotek wybierz słownik typu, który chcesz przenieść.
2. Z menu wybierz opcje **Edit > Move Up**, aby przenieść słownik typu o jedną pozycję w górę w panelu drzewa bibliotek, albo opcję **Edit > Move Down**, aby przenieść atrybut w dół o jedną pozycję.

Aby przenieść typ do innej biblioteki

1. W panelu drzewa bibliotek wybierz słownik typu, który chcesz przenieść.
2. Kliknij prawym przyciskiem myszy i wybierz opcję **Type Properties** z menu kontekstowego. Zostanie otwarte okno dialogowe Type Properties. (Można również przeciągnąć i upuścić typ do innej biblioteki).
3. W polu listy Add To wybierz bibliotekę, do której chcesz przenieść słownik typu.

4. Kliknij przycisk **OK**. Okno dialogowe zostanie zamknięte, a typ będzie teraz w bibliotece, która została wybrana.

Wyłączanie i usuwanie typów

Aby tymczasowo usunąć słownik typu, można go wyłączyć, usuwając zaznaczenie pola wyboru wyboru z lewej strony nazwy słownika w panelu drzewa bibliotek. Sygnalizuje to, że chcesz zachować słownik w bibliotece, ale jego zawartość ma być ignorowania podczas wykrywania konfliktów i podczas wyodrębniania.

Można również trwale usunąć słownik typu z biblioteki.

Aby wyłączyć słownik typu

1. W panelu drzewa bibliotek wybierz słownik typu, który chcesz wyłączyć.
2. Naciśnij klawisz spacji. Pole wyboru po lewej stronie nazwy typu nie jest teraz zaznaczone.

Aby usunąć słownik typu

1. W panelu drzewa bibliotek wybierz słownik typu, który chcesz usunąć.
2. Z menu wybierz opcje **Edit > Delete**, aby usunąć słownik typu.

Słowniki zastąpień/synonimów

Słownik zastąpień to zbiór terminów, który pomaga w grupowaniu podobnych terminów pod jednym terminem docelowym. Słownikami zastąpień zarządza się w dolnym panelu karty Library Resources. Dostęp do tego widoku można uzyskać, wybierając opcję **View > Resource Editor** w menu, jeśli jest aktywna sesja interaktywnego pulpitu roboczego. Jeśli nie jest, można edytować słowniki konkretnego szablonu w oknie Template Editor.

Można w tym słowniku zdefiniować dwa rodzaje zastąpień: **synonimy** i **elementy opcjonalne**. Można klikać karty w tym panelu, aby przełączać się między nimi.

Po przeprowadzeniu wyodrębniania może się okazać, że kilka pojęć wynikowych to w istocie synonimy lub odmienione formy innych pojęć. Poprzez identyfikację elementów opcjonalnych i synonimów można zmusić mechanizm wyodrębniania, aby odwzorował je na jeden termin docelowy.

Zastępowanie przy użyciu synonimów i elementów opcjonalnych zmniejsza liczbę pojęć w panelu Extraction Results, ponieważ łączy je w bardziej istotne, reprezentatywne pojęcia o wyższej liczbie wystąpień .

Uwaga: W przypadku zasobów w języku japońskim opcjonalne elementy nie mają zastosowania i nie są dostępne. Ponadto synonimy są obsługiwane odmiennie w przypadku tekstu w języku japońskim.

Synonimy

Synonimy kojarzą dwa lub większą liczbę wyrazów o tym samym znaczeniu. Można również użyć synonimów do grupowania terminów z ich skrótami lub do grupowania najczęściej występujących błędnie zapisanych form terminu z jego poprawną formą. Synonimy można definiować na karcie Synonyms.

Definicja synonimu jest złożona z dwóch części. Pierwszy termin to termin docelowy (**Target**). To pod nim mają być grupowane wszystkie synonimy. Jeśli ten termin docelowy nie jest używany jako synonim innego terminu docelowego lub wykluczony, to prawdopodobnie stanie się pojęciem i pojawi się na panelu Extraction Results. Na drugim miejscu znajduje się lista synonimów, które zostaną zgrupowane pod terminem docelowym.

Na przykład, jeśli chcesz, aby wartość automobile była zastępowana przez wartość vehicle, to wartość automobile jest synonimem, a vehicle jest terminem docelowym.

Można wprowadzić dowolne słowa w kolumnie **Synonim**, ale jeśli wyraz nie zostanie znaleziony podczas wyodrębniania i termin miał opcję dopasowania Entire, to zastąpienie nie będzie możliwe. Jednak termin docelowy nie musi być koniecznie wyodrębniony, aby synonimy zostały zgrupowane pod tym terminem.

Elementy opcjonalne

Elementy opcjonalne to wyrazy opcjonalne w terminie złożonym, które mogą być ignorowane podczas wyodrębniania, tak aby podobne terminy były traktowane jako równoważne, nawet jeśli mają nieznacznie różne postacie w tekście. Elementy opcjonalne są pojedynczymi wyrazami, których usunięcie z terminu złożonego może doprowadzić do dopasowania z innym terminem. Te pojedyncze wyrazy mogą znajdować się w dowolnym miejscu w obrębie terminu złożonego na początku, w środku lub na końcu. Opcjonalne elementy można definiować na karcie Optional.

Na przykład, aby zgrupować terminy ibm i ibm corp, należy zadeklarować corp jako element opcjonalny. Inny przykład: jeśli określisz termin access jako element opcjonalny, a podczas wyodrębniania zostaną znalezione terminy internet access speed i internet speed, to zostaną zgrupowane pod tym z dwóch terminów, który występuje częściej.

Uwaga: W przypadku zasobów dla tekstu japońskiego nie ma karty Optional Elements, ponieważ elementy opcjonalne nie są stosowane.

Definiowanie synonimów

Na karcie Synonyms można wprowadzić definicję synonimów w pustym wierszu w górnej części tabeli. Rozpocznij od definiowania terminu docelowego i jego synonimów. Można również wybrać bibliotekę, w której chcesz zapisać tę definicję. Podczas wyodrębniania wszystkie wystąpienia synonimów zostaną pogrupowane pod terminem docelowym w ostatecznym wyodrębnieniu. Więcej informacji zawiera temat “Dodawanie terminów” na stronie 180.

Na przykład, jeśli dane tekstowe zawierają dużo informacji o tematyce telekomunikacyjnej, mogą w nich wystąpić terminy: cellular phone, wireless phone i mobile phone. W omawianym przykładzie celowe może być zdefiniowanie terminów cellular i mobile jako synonimów wireless. Jeśli zdefiniujesz te synonimy, to każde wyodrębnione wystąpienie cellular phone i mobile phone będzie traktowane jako ten sam termin, co wireless phone, i terminy te będą występowały łącznie na liście.

Tworząc bibliotekę typu, możesz wprowadzić termin, a potem zastanowić się nad trzema lub czterema synonimami tego terminu. Wszystkie te terminy synonimiczne i termin docelowy wprowadź następnie do słownika zastąpień, po czym przeciągnij synonimy.

Uwaga: Synonimy są obsługiwane inaczej w przypadku tekstu w języku japońskim.

Zastępowanie synonimów działa także w odniesieniu do form odmienionych (np. liczby mnogiej) synonimów. W zależności od kontekstu celowe może być wprowadzenie ograniczeń w zastępowaniu terminów. Niektóre znaki umożliwiają ograniczenie zasięgu przetwarzania synonimów:

- **Wykrzyknik (!).** Gdy wykrzyknik bezpośrednio poprzedza synonim (!synonim), to zakazuje zastępowania odmienionych form synonimu terminem docelowym. Jednak wykrzyknik bezpośrednio poprzedzający termin docelowy !termin docelowy oznacza, że żadna część złożonego terminu docelowego ani żaden jego wariant nie powinny być zastępowane.
- **Gwiazdka (*).** Gwiazdka umieszczona bezpośrednio za synonimem (synonim*) oznacza, że ten wyraz ma być zastępowany przez termin docelowy. Na przykład, jeśli zdefiniowano manage* jako synonim, a management jako termin docelowy, to associate managers zostanie zastąpione terminem docelowym associate management. Można także dodać spację i gwiazdkę za wyrazem (synonim *), na przykład internet *. Jeśli zdefiniowano termin docelowy jako internet, a synonimy jako internet * * i web *, to internet access card i web portal zostaną zastąpione terminem internet. W tym słowniku wyraz ani łańcuch nie może zaczynać się od wieloznacznej gwiazdki.
- **Daszek (^).** Daszek i spacja poprzedzające synonim (^ synonim) oznaczają, że grupowanie synonimów ma być stosowane tylko wtedy, gdy termin rozpoczyna się od synonimu. Na przykład, jeśli zdefiniowano ^ wage jako synonim, a income jako termin docelowy i oba te terminy zostaną wyodrębnione, to zostaną zgrupowane pod terminem income. Jeśli jednak wyodrębnione zostaną terminy minimum wage i income, to nie zostaną zgrupowane, ponieważ minimum wage nie zaczyna się od wage. Między tym symbolem a synonimem musi znajdować się spacja.

- **Znak dolara ($).** Spacja i znak dolara następujące po synonimie (synonim $) oznaczają, że grupowanie synonimów ma być stosowane tylko wtedy, gdy termin kończy się synonimem. Na przykład, jeśli zdefiniowano cash $ jako synonim, a money jako termin docelowy i oba te terminy zostaną wyodrębnione, to zostaną zgrupowane pod terminem money. Jeśli jednak wyodrębnione zostaną terminy cash cow i money, to nie zostaną zgrupowane, ponieważ cash cow nie kończy się wyrazem cash. Między tym symbolem a synonimem musi znajdować się spacja.
- **Daszek (^) i znak dolara ($).** Jeśli daszek i znak dolara zostaną użyte łącznie, (^ synonim $), to termin pasuje do synonimu tylko wtedy, gdy jest dokładnie z nim zgodny. Oznacza to, że aby nastąpiło grupowanie synonimu z terminem, przed i po synonimie nie mogą występować żadne wyrazy. Na przykład można zdefiniować ^ van $ jako synonim, a truck jako termin docelowy, aby tylko termin van był grupowany z terminem truck, ale by nazwisko marie van guerin pozostało niezmienione. Ponadto, gdy zdefiniujesz synonim z użyciem daszka i znaku dolara, to wystąpienie tego wyrazu w tekście źródłowym spowoduje automatyczne wyodrębnienie synonimu.

Uwaga: Te znaki specjalne i symbole wieloznaczne nie są obsługiwane dla tekstu japońskiego.

Aby dodać wpis synonimu

1. Gdy widoczny będzie panel zastąpień, kliknij kartę **Synonyms** w lewym dolnym rogu.
2. W pustym wierszu na górze tabeli wprowadź termin docelowy w kolumnie Target. Wprowadzony termin docelowy pojawi się w kolorze. Ten kolor oznacza typ terminu (przypisany lub wymuszony). Jeśli termin jest wyświetlany w kolorze czarnym, oznacza to, że nie figuruje w żadnym słowniku typu.
3. Kliknij w drugiej komórce po prawej stronie terminu docelowego i wprowadź zestaw synonimów. Oddziel pozycje za pomocą separatora globalnego zdefiniowanego w oknie dialogowym Options. Więcej informacji zawiera temat “Określanie opcji” na stronie 74. Wprowadzane terminy są wyświetlane w kolorach. Ten kolor reprezentuje typ terminu. Jeśli termin jest wyświetlany w kolorze czarnym, oznacza to, że nie figuruje w żadnym słowniku typu.
4. Kliknij ostatnią komórkę, aby wybrać bibliotekę, w której ma zostać zapisana ta definicja synonimu.

Uwaga: Ta instrukcja dotyczy wprowadzania zmian w widoku Resource Editor lub oknie Template Editor. Należy pamiętać, że optymalizację można też prowadzić bezpośrednio z panelu Extraction Results, panelu Data, panelu Categories lub okna dialogowego Cluster Definitions w innych widokach. Więcej informacji zawiera temat “Optymalizacja wyników wyodrębniania” na stronie 87.

Definiowanie elementów opcjonalnych

W zakładce Optional można zdefiniować opcjonalne elementy dla dowolnej biblioteki. Te elementy są grupowane dla każdej biblioteki. Gdy biblioteka zostanie dodana do panelu drzewa bibliotek, na karcie Optional pojawia się pusty wiersz elementu opcjonalnego.

Wszystkie wpisy są automatycznie przekształcane w małe litery. Mechanizm wyodrębniania dopasowuje wpisy do wyrazów zapisanych w tekście zarówno małymi, jak i wielkimi literami.

Uwaga: W przypadku zasobów w języku japońskim opcjonalne elementy nie mają zastosowania i nie są dostępne.

Uwaga: Terminy rozdziela się za pomocą separatora zdefiniowanego w oknie dialogowym Options. Więcej informacji zawiera temat “Określanie opcji” na stronie 74. Jeśli element opcjonalny, który wprowadzasz, zawiera ten sam separator, należy poprzedzić go ukośnikiem odwrotnym.

Aby dodać wpis

1. Gdy widoczny będzie panel zastąpień, kliknij kartę Optional w lewym dolnym rogu edytora.
2. W kolumnie Optional Elements kliknij komórkę właściwą dla biblioteki, do której chcesz dodać wpis.
3. Wprowadź element opcjonalny. Oddziel pozycje za pomocą separatora globalnego zdefiniowanego w oknie dialogowym Options. Więcej informacji zawiera temat “Określanie opcji” na stronie 74.

Wyłączanie i usuwanie zastąpień

Można tymczasowo usunąć wpis, wyłączając go w słowniku. Wyłączenie wpisu spowoduje, że wpis będzie ignorowany w czasie wyodrębniania.

Można również usunąć niepotrzebne lub nieaktualne wpisy ze słownika zastąpień.

Aby wyłączyć wpis

1. W słowniku wybierz wpis, który chcesz wyłączyć.
2. Naciśnij klawisz spacji. Pole wyboru po lewej stronie wpisu będzie teraz niezaznaczone.

Uwaga: Można także usunąć zaznaczenie pola wyboru po lewej stronie wpisu, aby go wyłączyć.

Aby usunąć wpis synonimu

1. W słowniku wybierz wpis, który chcesz usunąć.
2. Z menu wybierz opcje **Edit > Delete** lub naciśnij klawisz **Delete** na klawiaturze. Wpis zostanie usunięty ze słownika.

Aby usunąć wpis elementu opcjonalnego

1. W słowniku kliknij dwukrotnie wpis, który chcesz usunąć.
2. Ręcznie usuń termin.
3. Naciśnij klawisz Enter, aby zastosować zmianę.

Słowniki wykluczeń

Słownik wykluczeń to lista wyrazów, fraz oraz fragmentów łańcuchów. Wszelkie terminy pasujące do wpisu ze słownika wykluczeń lub zawierające taki wpis będą ignorowane lub wykluczane z wyodrębniania. Słownikami wykluczeń zarządza się w prawym panelu edytora. Zwykle do tego słownika dodaje się wyrazy wypełniające lub frazy używane w tekście dla zapewnienia jego ciągłości, ale nie wnoszące nic ważnego, które mogłyby tylko zaciemnić wyniki wyodrębniania. Dodając te terminy do słownika wykluczeń powodujesz, że nie będą nigdy wyodrębniane.

Słownikami typów zarządza się w prawym górnym panelu karty Library Resources w edytorze. Dostęp do tego widoku można uzyskać, wybierając opcję **View > Resource Editor** w menu, jeśli jest aktywna sesja interaktywnego pulpitu roboczego. Jeśli nie jest, można edytować słowniki konkretnego szablonu w oknie Template Editor.

W słowniku wykluczeń można wprowadzić wyraz, frazę lub część łańcucha w pustym wierszu w górnej części tabeli. Do słownika wykluczeń można dodawać łańcuchy znaków obejmujące jeden lub więcej wyrazów, a nawet części wyrazów, posługując się gwiazdką jako symbolem wieloznacznym. Wpisy zadeklarowane w słowniku wykluczeń uniemożliwią wyodrębnienie związanych z nimi terminów. Jeśli wpis jest zadeklarowany także w innym miejscu interfejsu, na przykład w słowniku typu, to w innych słownikach będzie przekreślony. Łańcuch nie musi występować w danych tekstowych ani być zadeklarowany w jakimikolwiek słowniku typy, aby był stosowany.

Uwaga: Jeśli do słownika wykluczeń dodasz pojęcie, które jest także terminem docelowym synonimu, to wykluczony zostanie zarówno termin docelowy, jak i wszystkie jego synonimy. Więcej informacji zawiera temat “Definiowanie synonimów” na stronie 185.

Korzystanie z symboli wieloznacznych (*)

We wszystkich językach z wyjątkiem japońskiego można użyć symbolu wieloznacznego gwiazdki, aby zasygnalizować, że wykluczany wpis ma być traktowany jako łańcuch częściowy. Wszelkie terminy znalezione przez mechanizm wyodrębniania zawierające wyraz, który rozpoczyna się lub kończy łańcuchem wprowadzonym do słownika wykluczeń, zostaną wykluczone z ostatecznego wyodrębniania. Istnieją jednak dwa przypadki, w których zastosowanie symbolu wieloznacznego nie jest dozwolone.

- Łącznik (-) poprzedzony gwiazdką, na przykład *-

- Apostrof (') poprzedzony gwiazdką, na przykład **'s*

Tabela 39. Przykłady wpisów wykluczania.

Pozycja	Przykład	Wyniki
wyraz	*next*	Nie będą wyodrębniane żadne pojęcia (ani ich terminy) zawierające wyraz *next*.
fraza	*for example*	Nie będą wyodrębniane żadne pojęcia (ani ich terminy) zawierające frazę *for example*.
cząstkowe	*copyright**	Ten wpis wyklucza wszelkie pojęcia (i ich terminy) pasujące do odmian wyrazu *copyright* lub je zawierające, na przykład copyrighted, copyrighting, copyrights lub copyright 2010.
cząstkowe	**ware*	Ten wpis wyklucza wszelkie pojęcia (i ich terminy) pasujące do odmian wyrazu *ware* lub je zawierające, na przykład freeware, shareware, software, hardware, beware lub silverware.

Aby dodać wpisy

1. W pustym wierszu na górze tabeli wprowadź termin. Wprowadzony termin zostanie wyświetlony w kolorze. Ten kolor reprezentuje typ terminu. Jeśli termin jest wyświetlany w kolorze czarnym, oznacza to, że nie figuruje w żadnym słowniku typu.

Aby wyłączyć wpisy

Można tymczasowo usunąć wpis, wyłączając go w słowniku wykluczeń. Wyłączenie wpisu spowoduje, że wpis będzie ignorowany w czasie wyodrębniania.

1. W słowniku wykluczeń wybierz wpis, który chcesz wyłączyć.
2. Naciśnij klawisz spacji. Pole wyboru po lewej stronie wpisu będzie teraz niezaznaczone.

Uwaga: Można także usunąć zaznaczenie pola wyboru po lewej stronie wpisu, aby go wyłączyć.

Aby usunąć wpisy

Można usunąć niepotrzebne lub nieaktualne wpisy ze słownika wykluczeń.

1. W słowniku wykluczeń wybierz wpis, który chcesz usunąć.
2. Z menu wybierz kolejno opcje **Edit > Delete**. Wpis zostanie usunięty ze słownika.

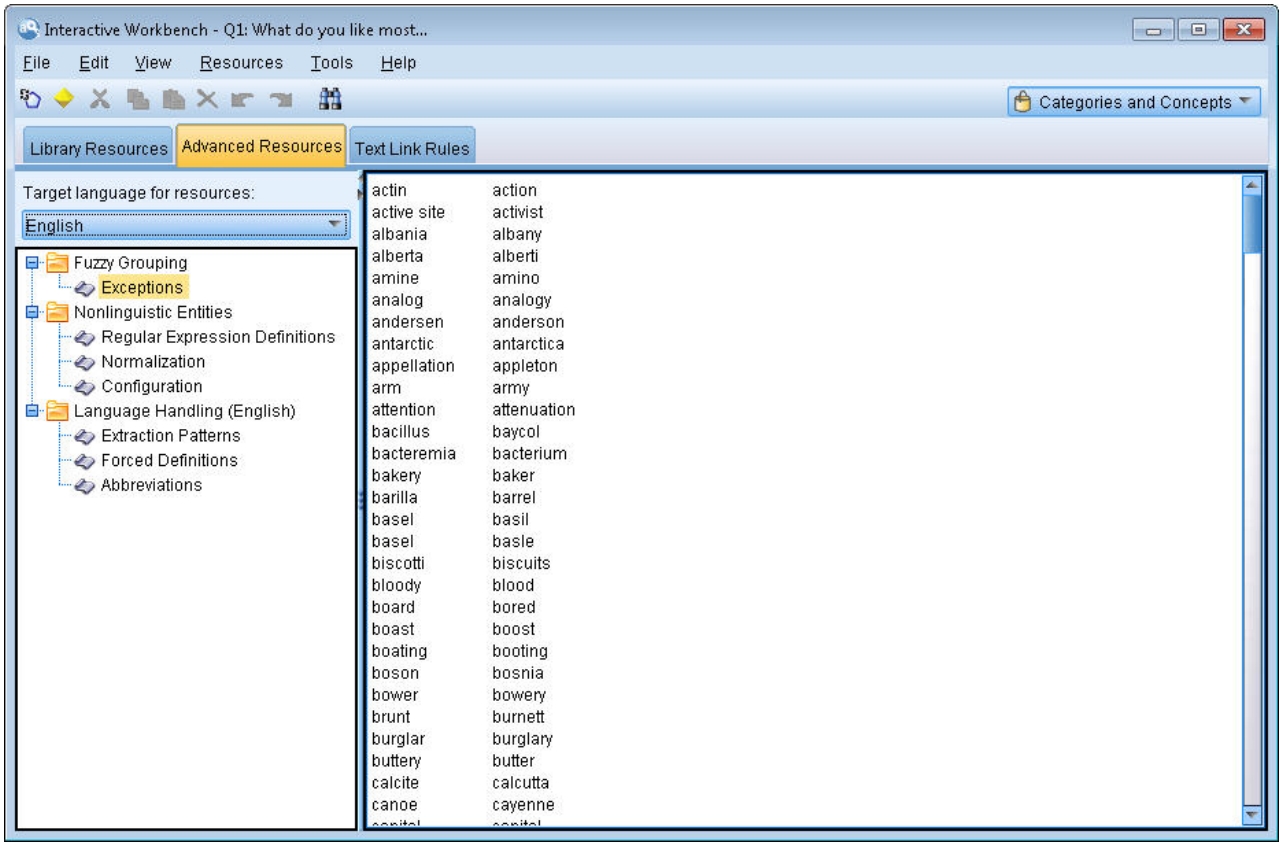
Rozdział 17. Informacje o zasobach zaawansowanych

W produkcie można korzystać ze słowników typów, wykluczeń i podstawień, a ponadto można pracować z różnymi ustawieniami zasobów zaawansowanych, takimi jak ustawienia grupowania rozmytego, a także definicje typów nielingwistycznych. Z tymi zasobami można pracować na karcie zasobów zaawansowanych w widoku Template Editor lub widoku Resource Editor.

Ważne: Ta karta nie jest dostępna w przypadku zasobów dostosowanych dla tekstów w języku japońskim.

Po przejściu do karty zasobów zaawansowanych można edytować następujące informacje:

- **Język docelowy dla zasobów.** Określa język, w którym zasoby będą tworzone i dostosowywane. Więcej informacji zawiera temat “Język docelowy dla zasobów” na stronie 191.
- **Grupowanie rozmyte (wyjątki).** Umożliwia wykluczanie par słów z algorytmu grupowania rozmytego (poprawki błędów pisowni). Więcej informacji zawiera temat “Grupowanie rozmyte” na stronie 191.
- **Obiekty nielingwistyczne.** Określanie obiektów nielingwistycznych, które będą wyodrębniane lub nie. Ponadto określanie wyrażeń regularnych i reguł normalizacji stosowanych podczas ich wyodrębniania. Więcej informacji zawiera temat “Obiekty nielingwistyczne” na stronie 192.
- **Obsługa języków.** Deklarowanie sposobów tworzenia struktur zdań (wzorce wyodrębniania i definicje wymuszone), a także określanie sposobów używania skrótów dla wybranego języka. Więcej informacji zawiera temat “Obsługa języków” na stronie 196.



Rysunek 41. Edytor szablonów eksploracji tekstu - karta Advanced Resources

Uwaga: Za pomocą paska narzędzi Znajdź i zamień można szybko znajdować informacje, a także wprowadzać spójne zmiany w sekcji. Aby uzyskać więcej informacji, patrz “Zastępowanie”.

Aby edytować zasoby zaawansowane

1. Odszukaj i wybierz zasoby, które zamierzasz edytować. Zawartość pojawi się w panelu po prawej stronie.
2. Użyj przycisków menu lub na pasku narzędzi, aby wyciąć, skopiować lub wkleić zawartość.
3. Przeprowadź edycję plików wybranych do zmiany, korzystając z reguł formatowania z tej sekcji. Zmiany zostaną zapisane bezpośrednio po ich wprowadzeniu. W celu przywrócenia poprzedniego stanu można użyć strzałek cofania i ponawiania na pasku narzędzi.

Znajdowanie

W niektórych przypadkach konieczne jest szybkie wyszukanie informacji w określonej sekcji. Na przykład podczas analizy powiązań w tekście konieczne może być operowanie na setkach makr i definicji wzorców. Korzystając z funkcji Find, można szybko znaleźć określoną regułę. Do wyszukiwania informacji w określonej sekcji można użyć paska narzędzi Find.

Aby użyć funkcji Find

1. Odszukaj i wybierz sekcję zasobu, którą chcesz przeszukać. Zawartość pojawi się w prawym panelu edytora.
2. Z menu wybierz opcje **Edit > Find**. W górnej części okna dialogowego Edit Advanced Resources pojawi się pasek narzędzi Find.
3. W polu tekstowym wprowadź łańcuch wyrazów, który chcesz wyszukać. Za pomocą przycisków na pasku narzędzi można sterować uwzględnianiem wielkości liter, dopasowywaniem częściowym i kierunkiem wyszukiwania.
4. Kliknij przycisk **Find**, aby rozpocząć wyszukiwanie. W przypadku znalezienia dopasowania tekst w oknie zostanie podświetlony.
5. Ponownie kliknij przycisk **Find**, aby wyszukać następne dopasowanie.

Uwaga: Podczas pracy na karcie Text Link Rules opcja Find jest dostępna tylko podczas przeglądania kodu źródłowego.

Zastępowanie

W niektórych przypadkach konieczne jest wprowadzenie masowych zmian w zasobach zaawansowanych. Funkcja Replace umożliwia takie powtarzalne modyfikacje.

Aby użyć funkcji Replace

1. Odszukaj i wybierz sekcję zasobu, w której chcesz szukać i zastępować. Zawartość pojawi się w prawym panelu edytora.
2. Z menu wybierz kolejno opcje **Edit > Replace**. Zostanie otwarte okno dialogowe Replace.
3. W polu tekstowym **Find what** wprowadź łańcuch wyrazów, który chcesz wyszukać.
4. W polu tekstowym **Replace with** wprowadź łańcuch, którym chcesz zastąpić znaleziony tekst.
5. Zaznacz opcję **Match whole word only**, jeśli chcesz znajdować i zastępować tylko całe wyrazy.
6. Zaznacz opcję **Match case**, jeśli chcesz znajdować i zastępować tylko wyrazy, w których dokładnie zgadza się wielkość liter.
7. Kliknij przycisk **Find Next**, aby znaleźć dopasowanie. W przypadku znalezienia dopasowania tekst w oknie zostanie podświetlony. Jeśli nie chcesz zastępować tego konkretnego wystąpienia, klikaj ponownie przycisk **Find Next** aż do znalezienia dopasowania, które chcesz zastąpić.
8. Kliknij przycisk **Replace**, aby zastąpić wybrane dopasowanie.
9. Kliknij przycisk **Replace All**, aby zastąpić wszystkie dopasowania w zaznaczonej sekcji. Zostanie otwarty komunikat z liczbą przeprowadzonych zamian.
10. Po zakończeniu zastępowania kliknij przycisk **Close**. Okno dialogowe zostanie zamknięte.

Uwaga: W razie popełnienia pomyłki przy zastępowaniu można cofnąć operację zastąpienia, zamykając okno dialogowe i wybierając polecenie **Edit > Undo** z menu. Należy wykonać tę czynność po jednym razie dla każdej zmiany, która ma zostać wycofana.

Język docelowy dla zasobów

Zasoby tworzone są z myślą o konkretnym języku tekstu. Język, dla którego zoptymalizowane są te zasoby, jest zdefiniowany na karcie Advanced Resources. W razie potrzeby można przełączyć się na inny język, wybierając go w polu kombi **Target language for resources**. Ponadto wskazany tutaj język będzie widoczny jako język wszelkich pakietów analizy tekstu utworzonych przy użyciu tych zasobów.

Ważne: Potrzeba zmiany języka w zasobach występuje bardzo rzadko. Zmiana taka może spowodować problemy, jeśli zasób przestanie pasować do języka wyodrębniania. Choć możliwość ta jest rzadko wykorzystywana, można zmienić język, jeśli planuje się używanie opcji języka ALL podczas wyodrębniania, ponieważ oczekuje się, że tekst będzie zapisany w więcej niż jednym języku. Zmieniając język, można np. uzyskać dostęp do wzorców wyodrębniania, skrótów i definicji wymuszonych dla dodatkowego języka. Jednak przed opublikowaniem lub zapisaniem dokonanych zmian w zasobach lub uruchomieniem kolejnego wyodrębniania należy przywrócić język główny, który ma być wyodrębniany.

Grupowanie rozmyte

Wybranie opcji **Accommodate spelling for a minimum root character limit of** w węźle Text Mining i ustawieniach wyodrębniania jest równoznaczne z włączeniem algorytmu grupowania rozmytego.

Grupowanie rozmyte pomaga w grupowaniu wyrazów, które często są zapisywane błędnie lub z drobnymi różnicami, poprzez tymczasowe usunięcie wszystkich samogłosek (z wyjątkiem pierwszej) i podwójnych lub potrójnych spółgłosek z wyodrębnionych wyrazów, a następnie porównanie ich w celu sprawdzenia, czy rezultat jest identyczny. W trakcie wyodrębniania funkcja grupowania rozmytego jest stosowana do wyodrębnionych terminów, a wyniki są porównywane w celu wykrycia ewentualnych dopasowań. W razie znalezienia dopasowań pierwotne terminy są grupowane na ostatecznej liście wyodrębnionych terminów. Grupowane są pod terminem, który najczęściej występuje w danych.

Uwaga: Jeśli dwa porównywane terminy przypisane są do różnych typów, z wyjątkiem typu <Unknown>, to technika grupowania rozmytego nie będzie stosowana do tej pary terminów. Innymi słowy, aby technika ta była stosowana, terminy muszą należeć do tego samego typu lub do typu <Unknown>.

Jeśli po włączeniu opisywanej funkcji zorientujesz się, że dwa wyrazy o podobnej pisowni zostały nieprawidłowo zgrupowane, możesz wykluczyć je z grupowania rozmytego. Można to zrobić wprowadzając nieprawidłowo dopasowane pary w sekcji Exceptions na karcie Advanced Resources. Więcej informacji zawiera Rozdział 17, “Informacje o zasobach zaawansowanych”, na stronie 189.

Następujący przykład ilustruje działanie grupowania rozmytego. Jeśli grupowanie rozmyte jest włączone, poniższe wyrazy będą traktowane jako identyczne i dopasowywane w następujący sposób:

```
color  -> colr                mountain -> montn
colour -> colr                montana  -> montn

modeling  -> modlng           furniture -> furntr
modelling -> modlng           furnature -> furntr
```

W powyższym przykładzie prawdopodobnie pożądane byłoby uniknięcie zgrupowania wyrazów mountain i montana. Można je zatem wprowadzić w sekcji Exceptions, w następujący sposób:

```
mountain      montana
```

Ważne: W niektórych przypadkach wyjątki grupowania rozmytego nie zapobiegają łączeniu 2 wyrazów w parę, ponieważ stosowane są określone reguły synonimów. W takim przypadku można spróbować wprowadzić synonimy z

zastosowaniem wykrzyknika (!) jako symbolu wieloznacznego, aby zapobiec traktowaniu wyrazów jak synonimów w wynikach. Aby uzyskać więcej informacji, patrz “Definiowanie synonimów” na stronie 185.

Reguły formatowania wyjątków grupowania rozmytego

- W jednym wierszu definiuj tylko jedną parę stanowiącą wyjątek.
- Używaj wyrazów prostych lub złożonych.
- W wyrazach używaj tylko małych liter. Wyrazy z wielkimi literami będą ignorowane.
- Wyrazy w każdej parze oddzielaj znakiem TAB.

Obiekty nielingwistyczne

Podczas pracy z niektórymi typami danych szczególnie interesujące jest wyodrębnianie dat, numerów ubezpieczenia społecznego, wartości procentowych lub innych obiektów nielingwistycznych. Obiekty te są jawnie zadeklarowane w pliku konfiguracyjnym, w którym można je aktywować lub dezaktywować. Więcej informacji zawiera temat “Konfiguracja” na stronie 195. W celu optymalizacji wyników działania mechanizmu wyodrębniania dane nielingwistyczne są normalizowane tak, by były grupowane zgodnie z predefiniowanymi formatami. Więcej informacji zawiera temat “Normalizacja” na stronie 195.

Uwaga: Wyodrębnianie obiektów nielingwistycznych można włączać i wyłączać w ustawieniach wyodrębniania.

Dostępne obiekty nielingwistyczne

Możliwe jest wyodrębnianie obiektów nielingwistycznych zawartych w tabeli. Nazwa typu jest ujęta w nawiasy.

Tabela 40. Obiekty nielingwistyczne, które mogą być wyodrębniane

Adresy	(<Address>)
Aminokwasy	(<Aminoacid>)
Waluty	(<Currency>)
Daty	(<Date>)
Opóźnienie	(<Delay>)
Cyfry	(<Digit>)
Adresy e-mail	(<email>)
Adresy HTTP/URL	(<url>)
Adres IP	(<IP>
Organizacje	(<Organization>)
Procenty	(<Percent>)
Produkty	(<Product>)
Białka	(<Gene>)
Numery telefonów	(<PhoneNumber>)
Godziny	(<Time>)
Numery ubezpieczenia społecznego w USA	(<SocialSecurityNumber>)
Wagi i miary	(<Weights-Measures>)

Czyszczenie tekstu przed przetwarzaniem

Zanim rozpocznie się wyodrębnianie obiektów nielingwistycznych, tekst wejściowy jest czyszczony. W tym kroku wprowadzane są następujące tymczasowe zmiany umożliwiające rozpoznanie i właściwe wyodrębnienie obiektów nielingwistycznych:

- Każda sekwencja dwóch lub więcej spacji zamieniana jest na jedną spację.
- Znaki tabulacji są zastępowane spacjami.
- Pojedyncze znaki końca wiersza lub ciągi takich znaków są zastępowane spacją, a wiele ciągów znaku końca wiersza zamienianych jest na oznaczenie końca akapitu. Koniec wiersza może być oznaczony znakiem powrotu karetki (CR), znakiem wysuwu wiersza (LF) albo oboma takimi znakami.
- Znaczniki HTML i XML są tymczasowo usuwane i ignorowane.

Definicje wyrażeń regularnych

W przypadku wyodrębniania obiektów nielingwistycznych niekiedy konieczne jest zmodyfikowanie lub uzupełnienie definicji wyrażeń regularnych. Służy do tego sekcja **Regular Expression Definitions** na karcie Advanced Resources. Więcej informacji zawiera Rozdział 17, “Informacje o zasobach zaawansowanych”, na stronie 189.

Plik jest podzielony na odrębne sekcje. Pierwsza sekcja nosi nazwę [macros]. Oprócz tej sekcji istnieją także dodatkowe sekcje — po jednej dla każdego obiektu nielingwistycznego. Do tego pliku można dodawać nowe sekcje. W każdej sekcji reguły są ponumerowane (*regexp1*, *regexp2* i tak dalej). Reguły muszą mieć kolejne numery od 1 do *n*. Każda przerwa w numeracji spowoduje całkowite zawieszenie przetwarzania pliku.

W pewnych przypadkach obiekt jest zależny od języka. Obiekt uznaje się za zależny od języka, jeśli parametr języka w pliku konfiguracyjnym jest dla tego obiektu różny od 0. Więcej informacji zawiera temat “Konfiguracja” na stronie 195. Gdy obiekt jest zależny od języka, nazwę sekcji należy poprzedzić nazwą języka, na przykład [english/PhoneNumber]. Ta sekcja zawierać będzie reguły mające zastosowanie tylko do angielskich numerów telefonów, gdy obiekt PhoneNumber będzie miał przypisany język numer 2.

Ważne! Jeśli po zmodyfikowaniu tego lub innego pliku w edytorze mechanizm wyodrębniania przestanie działać zgodnie z oczekiwaniami, można skorzystać z opcji **Reset to Original** na pasku narzędzi, aby przywrócić pierwotną standardową zawartość pliku. Przy modyfikacji tego pliku niezbędny jest pewien poziom wiedzy na temat wyrażeń regularnych. Jeśli potrzebna jest dodatkowa pomoc, należy skontaktować się z firmą IBM Corp.

Znaki specjalne . [] {} () \ * + ? | ^ $

Wszystkie znaki dopasowywane są do samych siebie, z wyjątkiem następujących znaków specjalnych, które pełnią szczególne funkcje w wyrażeniach: .[{()*+?|^$. Aby użyć któregokolwiek z tych znaków dosłownie, należy poprzedzić go w definicji ukośnikiem odwrotnym (\).

Załóżmy, że próbujemy wyodrębniać adresy WWW. W takich obiektach znak kropki jest bardzo ważny, dlatego należy poprzedzać go ukośnikiem odwrotnym, na przykład:

```
www\.[a-z]+\.[a-z]+
```

Operatory powtórzeń i kwantyfikatory ? + * {}

Aby uzyskać bardziej elastyczne definicje, można w wyrażeniach regularnych stosować kilka standardowych symboli wieloznacznych. Są to symbole: * ? +

- *Gwiazdka* * oznacza *zero lub więcej wystąpień* poprzedzającego ją łańcucha. Na przykład ab*c pasuje do "*ac*", "*abc*", "*abbbc*" itd.
- *Znak plus* + oznacza *jedno lub więcej wystąpień* poprzedzającego go łańcucha. Na przykład ab+c pasuje do "*abc*", "*abbc*", "*abbbc*", ale nie do "*ac*".
- *Znak zapytania ?* oznacza *zero wystąpień lub jedno wystąpienie* poprzedzającego go łańcucha. Na przykład modell?ing pasuje do "*modeling*" i "*modeling*".
- *Odgraniczenie powtórzenia nawiasami {}* oznacza granice powtórzenia Na przykład

  [0-9]{n} pasuje do dowolnej cyfry powtórzonej dokładnie *n* razy. Na przykład [0-9]{4} pasuje do “*1998*”, ale nie do “*33*”, ani nie do “*19983*”.

  [0-9]{n,} pasuje do dowolnej cyfry powtórzonej *n lub więcej* razy. Na przykład [0-9]{3,} pasuje do “*199*” lub “*1998*”, ale nie do “*19*”.

[0-9]{n,m} pasuje do dowolnej cyfry powtórzonej *od n do m razy włącznie*. Na przykład [0-9]{3,5} pasuje do "*199*", "*1998*" lub "*19983*", ale nie do "*19*", ani nie do "*199835*".

Opcjonalne spacje i łączniki

W niektórych przypadkach konieczne jest umieszczenie w definicji opcjonalnej spacji. Na przykład, gdybyśmy chcieli wyodrębniać nazwy walut, takie jak "*uruguayan pesos*", "*uruguayan peso*", "*uruguay pesos*", "*uruguay peso*", "*pesos*" lub "*peso*", konieczne byłoby uwzględnienie faktu, że mogą one składać się z dwóch wyrazów rozdzielonych spacją. W takim przypadku definicję należy zapisać w postaci (uruguayan |uruguay)?pesos?. Ponieważ po wyrazie *uruguayan* lub *uruguay* w kombinacji z *pesos*/*peso* występuje spacja, opcjonalną spację należy umieścić w sekwencji opcjonalnej (uruguayan |uruguay). Gdyby spacja nie była umieszczona w sekwencji opcjonalnej, np. (uruguayan|uruguay)? pesos?, to wyrażenie nie pasowałoby do "*pesos*" lub "*peso*", ponieważ spacja traktowana byłaby jak znak wymagany.

Jeśli interesuje nas lista elementów obejmująca znak łącznika (-), to łącznik musi być umieszczony na końcu listy. Na przykład, jeśli szukamy przecinka (,) lub łącznika (-), to definicja powinna zawierać listę [,-], a nie [-,].

Kolejność łańcuchów na listach i w makrach

Należy zawsze definiować najdłuższą sekwencję przed krótszymi. W przeciwnym razie najdłuższa sekwencja nigdy nie zostanie odczytana, ponieważ dopasowana zostanie ta krótsza. Na przykład, gdybyśmy szukali łańcucha "*billion*" lub "*bill*", to wyraz "*billion*" musi być zdefiniowany przed "*bill*". A zatem, na przykład, (billion|bill), a nie (bill|billion). Dotyczy to także makr, ponieważ makra są listami łańcuchów.

Kolejność reguł w sekcji definicji

Definiuj jedną regułę w jednym wierszu. W każdej sekcji reguły są ponumerowane (*regexp1*, *regexp2* i tak dalej). Reguły muszą mieć kolejne numery od 1 do *n*. Każda przerwa w numeracji spowoduje całkowite zawieszenie przetwarzania pliku. Aby dezaktywować wpis, umieść symbol # na początku każdego wiersza definicji danego wyrażenia regularnego. Aby aktywować wpis, usuń symbol # z początku wiersza.

W każdej sekcji najbardziej zawężające reguły muszą być zdefiniowane przed najbardziej ogólnymi, aby przetwarzanie odbywało się prawidłowo. Na przykład, jeśli szukamy daty w postaci "*miesiąc rok*" i w postaci "*miesiąc*", to reguła "*miesiąc rok*" musi być zdefiniowana przed regułą "*miesiąc*". Oto właściwa definicja:

```
#@# January 1932
regexp1=$(MONTH),? [0-9]{4}

#@# January
regexp2=$(MONTH)
```

and not

```
#@# January
regexp1=$(MONTH)

#@# January 1932
regexp2=$(MONTH),? [0-9]{4}
```

Korzystanie z makr w regułach

Jeśli w kilku regułach ma być używana ta sama sekwencja, można skorzystać z marka. Wówczas ewentualną zmianę definicji sekwencji trzeba będzie wprowadzić tylko raz, a nie we wszystkich regułach, które się do niej odwołują. Załóżmy na przykład, że mamy następujące makro:

```
MONTH=((january|february|march|april|june|july|august|september|october|
november|december)|(jan|feb|mar|apr|may|jun|jul|aug|sep|oct|nov|dec)(\.)?)
```

Każde odwołanie do nazwy makra musi być ujęte w $(), na przykład: regexp1=$(MONTH)

Wszystkie makra muszą być zdefiniowane w sekcji [macros].

Normalizacja

Podczas wyodrębniania obiektów nielingwistycznych napotkane obiekty są normalizowane w taki sposób, by zostały pogrupowane zgodnie z predefiniowanymi formatami. Na przykład symbole walut i równoważne im oznaczenia słowne są traktowane jako tożsame. Wpisy używane do normalizacji są wymienione w sekcji **Normalization** na karcie Advanced Resources. Więcej informacji zawiera Rozdział 17, “Informacje o zasobach zaawansowanych”, na stronie 189. Plik jest podzielony na odrębne sekcje.

Ważne! Ten plik jest przeznaczony tylko dla użytkowników zaawansowanych. Jest mało prawdopodobne, że wystąpi potrzeba jego modyfikacji. Jeśli potrzebna jest dodatkowa pomoc, należy skontaktować się z firmą IBM Corp.

Reguły formatowania normalizacji

- W jednym wierszu umieszczaj tylko jeden wpis normalizacyjny.
- Ściśle przestrzegaj podziału pliku na sekcje. Nie można dodawać nowych sekcji.
- Aby dezaktywować wpis, umieść symbol # na początku odpowiedniego wiersza. Aby aktywować wpis, usuń symbol # z początku wiersza.

Daty zapisane w języku angielskim a normalizacja

Domyślnie daty w szablonie dla języka angielskiego są rozpoznawane w formacie amerykańskim, tj.: miesiąc, data, rok. Aby zmienić format na dzień, miesiąc, rok, należy dezaktywować wiersz „format:US” (dodając # na początku wiersza) i aktywować wiersz „format:UK” (usuwając # z początku tego wiersza).

Konfiguracja

W pliku konfiguracji obiektów nielingwistycznych można aktywować i dezaktywować typy obiektów nielingwistycznych, które mają być wyodrębniane. Dezaktywacja niepotrzebnych obiektów może przyspieszyć przetwarzanie. Aktywacji i dezaktywacji dokonuje się w sekcji **Configuration** na karcie Advanced Resources. Więcej informacji zawiera Rozdział 17, “Informacje o zasobach zaawansowanych”, na stronie 189. Jeśli włączone jest wyodrębnianie obiektów nielingwistycznych, to mechanizm wyodrębniania odczytuje ten plik konfiguracji w trakcie wyodrębniania, aby określić, które typy obiektów nielingwistycznych powinny być wyodrębniane.

W pliku tym obowiązuje następująca składnia:

#nazwa<TAB>*Język*<TAB>*Kod*

Tabela 41. Składnia pliku konfiguracji.

Etykieta kolumny	Opis
#name	Nazwa, za pośrednictwem której obiekty nielingwistyczne będą identyfikowane w dwóch pozostałych plikach wymaganych do wyodrębniania obiektów nielingwistycznych. W używanych tutaj nazwach rozróżniana jest wielkość liter.
Language	Język dokumentów . Najlepiej jest wybrać konkretny język. Istnieje jednak również opcja **Any**. Możliwe opcje to: 0 = dowolny język używany, gdy wyrażenie regexp nie jest charakterystyczne dla języka i może być używane w kilku szablonach z różnymi językami; przykład: IP/URL/adres e-mail; 1 = francuski; 2 = angielski; 4 = niemiecki; 5 = hiszpański; 6 = holenderski; 8 = portugalski; 10 = włoski.
Code	Kod części mowy. Większość obiektów, z nielicznymi wyjątkami, ma kod „s”. Możliwe wartości to: s = słowo zatrzymujące; a = przymiotnik; n = rzeczownik. Aktywne obiekty nielingwistyczne są najpierw wyodrębniane, a następnie stosowane są do nich wzorce wyodrębniania w celu określenia ich ról w szerszym kontekście. Na przykład wartościom procentowym przypisywana jest rola „a”. Załóżmy, że jako obiekt nielingwistyczny zostanie wyodrębniony wyraz 30%. Zostałby on zidentyfikowany jako przyrostek. Teraz, gdyby tekst zawierał wyrazy „30% salary increase” (30-procentowy wzrost wynagrodzenia), to obiekt „30%” zostałby dopasowany do wzorca części mowy „ann” (przymiotnik rzeczownik rzeczownik).

Kolejność definiowania obiektów

Kolejność, w jakiej obiekty są zadeklarowane w tym pliku, jest istotna i wpływa na sposób ich wyodrębniania. Są one stosowane w kolejności, w jakiej są wymienione. Zmiana kolejności wpłynie na uzyskiwane wyniki. Najbardziej jednoznaczne obiekty nielingwistyczne powinny być zdefiniowane przed bardziej ogólnymi.

Na przykład obiekt nielingwistyczny „Aminoacid" jest zdefiniowany w następujący sposób:

```
regexp1=($(AA)-?$(NUM))
```

gdzie $(AA) odpowiada „(ala|arg|asn|asp|cys|gln|glu|gly|his|ile|leu|lys|met|phe|pro|ser)", czyli konkretnym 3-literowym sekwencjom oznaczającym konkretne aminokwasy.

Z drugiej strony obiekt nielingwistyczny „Gene" jest bardziej ogólny i zdefiniowany w następujący sposób:

```
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?
```

Jeśli w sekcji konfiguracji obiekt „Gene" zostanie zdefiniowany przed obiektem „Aminoacid", to „Aminoacid" nigdy nie zostanie wykryty, ponieważ regexp3 z definicji obiektu „Gene" zawsze zostanie dopasowane jako pierwsze.

Reguły formatowania konfiguracji

- Wpisy w kolumnach rozdzielaj znakiem TAB.
- Nie usuwaj żadnych wierszy.
- Przestrzegaj składni przedstawionej w powyższej tabeli.
- Aby dezaktywować wpis, umieść symbol # na początku odpowiedniego wiersza. Aby aktywować wpis, usuń symbol # z początku wiersza.

Obsługa języków

W każdym ze współczesnych języków istnieją charakterystyczne sposoby wyrażania pojęć, budowania zdań i stosowania skrótów. W sekcji Language Handling można edytować wzorce wyodrębniania, wymuszać definicje dla takich wzorców i deklarować skróty dla języka wybranego z listy rozwijanej Language.

- Wzorce wyodrębniania
- Definicje wymuszone
- Skróty

Wzorce wyodrębniania

Podczas wyodrębniania informacji z dokumentów mechanizm wyodrębniania stosuje zestaw wzorców wyodrębniania części mowy do budowania „stosów" wyrazów w tekście w celu wykrycia terminów kandydackich (wyrazów i fraz) do wyodrębnienia. Można dodawać lub modyfikować wzorce wyodrębniania.

Do części mowy należą elementy gramatyczne, takie jak rzeczowniki, przymiotniki, przysłówki, imiona, inicjały, partykuły itd. Szereg takich elementów składa się na wzorzec wyodrębniania części mowy. W produktach IBM Corp. do eksploracji tekstu każdą część mowy reprezentuje jeden znak, dzięki czemu łatwiej jest definiować własne wzorce. Na przykład przymiotnik reprezentuje mała litera *a*. Zestaw obsługiwanych kodów wyświetlany jest domyślnie w górnej części sekcji każdego z domyślnych wzorców wyodrębniania wraz z zestawem zapisów wzorców i przykładów, które ułatwiają zrozumienie zasad posługiwania się kodami.

Reguły formatowania wzorców wyodrębniania

- Jeden wzorzec na wiersz.
- Aby dezaktywować wzorzec, umieść znak # na początku wiersza.

Kolejność wzorców wyodrębniania na liście jest bardzo ważna, ponieważ mechanizm wyodrębniania odczytuje daną sekwencję wyrazów tylko raz i przypisuje ją do pierwszego pasującego wzorca wyodrębniania.

Obsługiwane kody części mowy

Poniżej znajduje się spis wszystkich obsługiwanych kodów części mowy zdefiniowanych w języku angielskim skompilowanego słownika.

Wszystkie części mowy używane do konkretnego szablonu są wyświetlane u góry **Advanced Resources > Extraction patterns**.

Główna różnica między podstawowym szablonem zasobów a szablonem opinii polega na tym, że w szablonie podstawowym używane są podstawowe określniki („d") i przyimki („c"), a w szablonie opinii ich rozbudowane odpowiedniki („e" i „r"). Ponadto w szablonie opinii wszystkie wyrazy zawierające części mowy „a" i „Q" są przetwarzane wyłącznie jako „Q". We wszystkich szablonach opinii użycie „0", „1" i „2" jest ograniczone. Patrz**Advanced Resources > Language Handling (English) > Forced Definitions and Extraction patterns**.

Inne szablony w języku angielskim mogą używać części mowy, których nie wymieniono w słowniku (na przykład „w" i „W" w szablonie analizy rynku). Jednak w takim wypadku te części mowy są przypisane do konkretnych wyrazów w obszarze **Advanced Resources > Forced Definitions**.

Tabela 42. Obsługiwane kody części mowy

Code	Znaczenie	Przykład
a	przymiotnik	abdominal, blue...
W	nieużywany	nieużywany
b	przysłówek	frequently, often, very, ...
P	nieużywany	nieużywany
c	przyimek	„of"
Z	kod wewnętrzny dla wyrazów o nieprawidłowej pisowni	
d	określnik	„the"
D	nieużywany	nieużywany
e	rozbudowany określnik	the, an, my, your...
E	nieużywany	nieużywany
f	imię	John, Mary...
F	nieużywany	nieużywany
g	nieużywany	nieużywany
G	przymiotniki narodowości	french, american...
h	nieużywany	nieużywany
H	nieużywany	nieużywany
i	nieużywany	nieużywany
I	nieużywany	nieużywany
j	nieużywany	nieużywany
J	nieużywany	nieużywany
k	nieużywany	nieużywany
K	nieużywany	nieużywany
l	nieużywany	nieużywany
L	nieużywany	nieużywany

Tabela 42. Obsługiwane kody części mowy (kontynuacja)

Code	Znaczenie	Przykład
m	przymiotnik lub nieznana część mowy	dog, ibm
M	nieużywany	nieużywany
n	rzeczownik	pies
N	nieużywany	nieużywany
o	spójnik	„and”, „&”
O	nieużywany	nieużywany
p	imiesłów bierny	abandoned, accessorized...
P	nieużywany	nieużywany
q	nieużywany	nieużywany
Q	kwalifikator	expensive, small, good, ...
r	rozbudowany przyimek	of, among, against, from...
R	nieużywany	nieużywany
s	stop lista	dowolne słowo, które nie jest potrzebne do wyodrębnienia
S	nieużywany	nieużywany
t	tytuł	mrs., mrs, captain, brig., ...
T	przymiotniki techniczne	tumor-restricted... (wszystkie „T” są również „a”)
u	nieznane według definicji, nie występują w słowniku	
U	nieużywany	nieużywany
v	czasownik	eat, eats, ate, eating, ...
V	bezokolicznik	eat, ...
w	nieużywany	nieużywany
W	nieużywany	nieużywany
x	czasownik posiłkowy	be
X	nieużywany	nieużywany
y	przyimki w nazwiskach	von, di, de, ... (służy do wyodrębniania nazw osób: John von Doe)
Y	nieużywany	nieużywany
z	nieużywany	nieużywany
Z	nieużywany	nieużywany
0	przysłówek wyrażający opinię	Tylko w szablonach opinii. Patrz **Advanced resources > Language Handling (English) > Forced Definitions**.
1	„to” w opiniach	Patrz **Advanced resources > Language Handling (English) > Forced Definitions**
2	określony kwalifikator	Tylko w szablonach opinii. Patrz **Advanced resources > Language Handling (English) > Forced Definitions**.
3	nieużywany	nieużywany
4	nieużywany	nieużywany
5	nieużywany	nieużywany
6	nieużywany	nieużywany
7	nieużywany	nieużywany

Tabela 42. Obsługiwane kody części mowy (kontynuacja)

Code	Znaczenie	Przykład
8	nieużywany	nieużywany
9	nieużywany	nieużywany

Definicje wymuszone

Podczas wyodrębniania informacji z dokumentów mechanizm wyodrębniania przeszukuje tekst i przypisuje każdy napotkany wyraz do części mowy. W niektórych przypadkach ten sam wyraz mógłby pełnić więcej niż jedną rolę, zależnie od kontekstu. Jeśli chcesz wymusić traktowanie wyrazu zawsze jako określonej części mowy lub całkowicie wykluczyć wyraz z przetwarzania, możesz zrobić to w sekcji **Forced Definition** na karcie Advanced Resources. Więcej informacji zawiera Rozdział 17, “Informacje o zasobach zaawansowanych”, na stronie 189.

Aby wymusić traktowanie danego wyrazu jako określonej części mowy, musisz dodać w tej sekcji wiersz o formacie:

termin:*kod*

Tabela 43. Opis składni.

Pozycja	Opis
termin	Nazwa terminu.
code	Jednoznakowy kod reprezentujący część mowy. Można wymienić maksymalnie sześć różnych części mowy dla jednego terminu pojedynczego. Ponadto można zapobiec wyodrębnianiu wyrazu w wyrazy złożone/frazy, korzystając z kodu s (mała litera), na przykład dodatkowe:s.

Reguły formatowania definicji wymuszonych

- Jeden wiersz na wyraz.
- Terminy nie mogą zawierać dwukropków.
- Aby całkowicie zablokować wyodrębnianie wyrazu, użyj kodu części mowy s.
- Maksymalnie sześć kodów części mowy na wiersz. Obsługiwane kody części mowy są wyświetlane w sekcji Extraction Patterns. Więcej informacji zawiera temat “Wzorce wyodrębniania” na stronie 196.
- Aby uzyskać dopasowania częściowe, użyj gwiazdki (*) jako symbolu wieloznacznego na końcu łańcucha. Na przykład wprowadzenie add*:s spowoduje, że takie wyrazy, jak add, additional, additionally, addendum i additive nigdy nie będą wyodrębniane jako termin ani jako część terminu złożonego. Jeśli jednak dopasowany wyraz jest jawnie zadeklarowany jako termin w skompilowanym słowniku lub na liście wymuszonych definicji, to nadal będzie wyodrębniany. Na przykład wprowadzenie zarówno add*:s, jak i addendum:n, spowoduje, że znaleziony w tekście wyraz addendum zostanie wyodrębniony.

Skróty

Gdy mechanizm wyodrębniania przetwarza tekst, z zasady traktuje każdą napotkaną kropkę jako koniec zdania. Zwykle takie założenie jest poprawne, jednak ten sposób traktowania kropek nie ma zastosowania, gdy tekst zawiera skróty.

Jeśli podczas wyodrębniania terminów z tekstu okazuje się, że niektóre skróty są traktowane nieprawidłowo, należy je jawnie zadeklarować w tej sekcji.

Uwaga: Jeśli skrót występuje już w definicji synonimu lub jest zdefiniowany jako termin w słowniku typu, to nie trzeba go tutaj dodawać.

Reguły formatowania skrótów

- Definiuj jeden skrót w jednym wierszu.

Rozdział 18. Informacje o regułach powiązań w tekście

Analiza powiązań w tekście (TLA — Text link analysis) to technika dopasowywania wzorców służąca do wyodrębniania relacji istniejących w tekście w oparciu o zestaw reguł. Gdy przy wyodrębnianiu włączona jest analiza powiązań w tekście, dane tekstowe są porównywane z tymi regułami. Po znalezieniu dopasowania wzorzec analizy powiązań jest wyodrębniany i prezentowany. Reguły te definiuje się na karcie Text Link Rules.

Na przykład samo wyodrębnienie pojęć odzwierciedlających proste cechy organizacji może nie dostarczyć wartościowej wiedzy. Jednak stosując technikę TLA, można ujawnić powiązania między różnymi organizacjami lub osoby powiązane z organizacją. Technikę TLA można też zastosować do wyodrębniania opinii o takich zagadnieniach, jak odczucia osób o danym produkcie lub usłudze.

Aby skorzystać z techniki TLA, musisz mieć do dyspozycji zasoby zawierające reguły analizy powiązań w tekście. Wybierając szablon, można zorientować się, które szablony zawierają reguły TLA, po obecności lub braku ikony w kolumnie TLA.

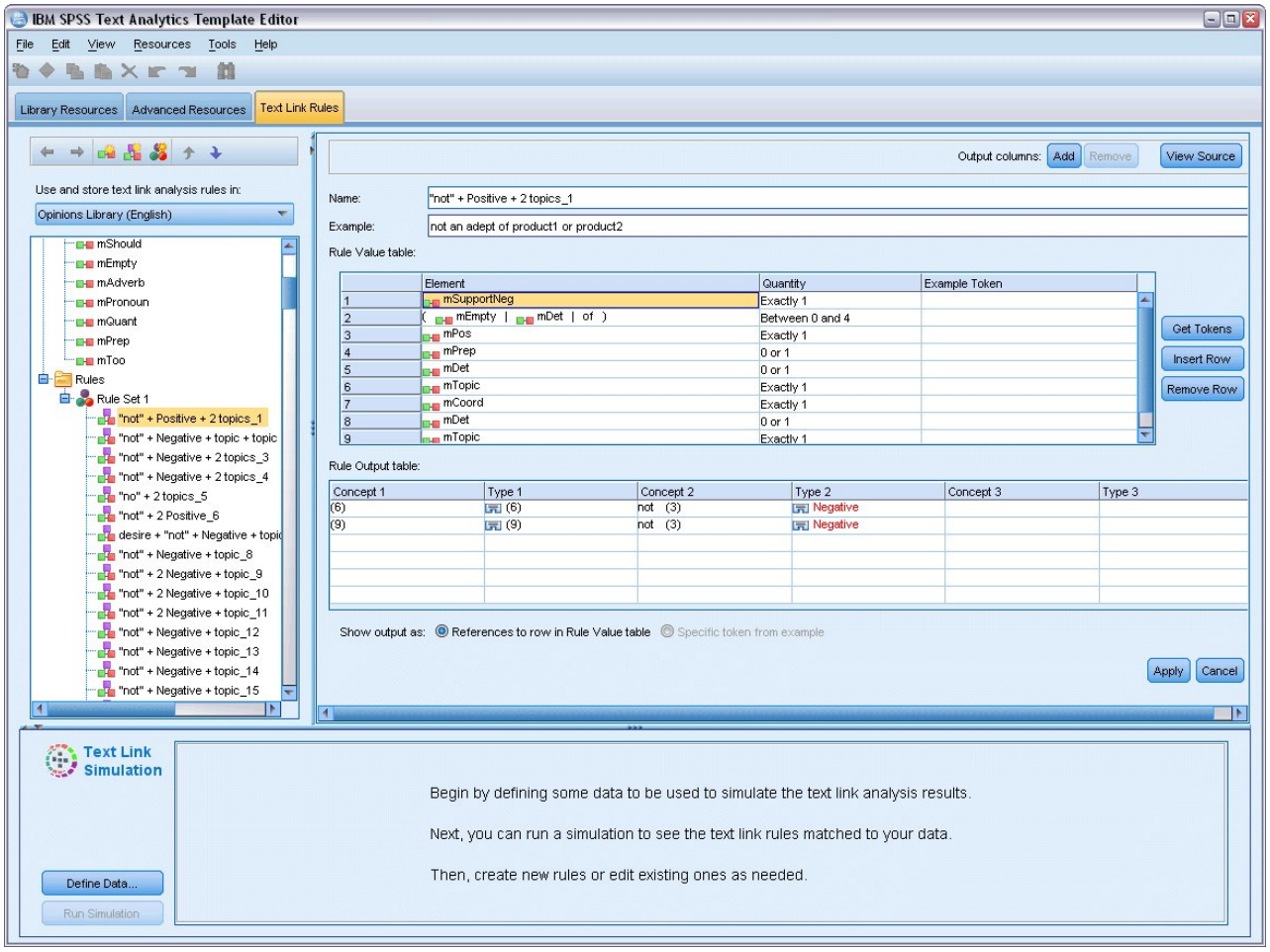
Wzorce analizy powiązań w tekście są znajdowane w danych tekstowych na etapie dopasowywania wzorców podczas wyodrębniania. Na tym etapie reguły są porównywane z danymi tekstowymi, a gdy zostanie znalezione dopasowanie, informacja ta jest wyodrębniana jako wzorzec. Niekiedy zachodzi potrzeba przeprowadzenia pogłębionej analizy powiązań w tekście lub zmiany sposobu dopasowania. Można w takich przypadkach dostroić reguły, dostosowując je do specyficznych potrzeb. Służy do tego karta Text Link Rules.

Uwaga: Obsługę zmiennych wycofano w wersji 13. Zamiast nich należy korzystać z makr. Więcej informacji zawiera temat “Praca z makrami” na stronie 206.

Gdzie opracowywać reguły powiązań w tekście

Użytkownik może edytować i tworzyć reguły bezpośrednio na karcie Text Link Rules w widoku Template Editor lub Resource Editor. Aby zorientować się, w jaki sposób reguły mogą wyszukiwać tekst, można uruchomić symulację na tej karcie. W trakcie symulacji procedura jest uruchamiana tylko na próbie danych symulacji i reguły powiązań w tekście są stosowane w celu sprawdzenia, czy jakiekolwiek wzorce pasują do tej próby. Wszystkie reguły, które pasują do tekstu, są następnie wyświetlane w panelu symulacji. Na podstawie dopasowań można w razie potrzeby zmodyfikować reguły i makra, aby zmienić sposób, w jaki tekst jest dopasowywany.

W odróżnieniu od innych zasobów zaawansowanych, reguły TLA są charakterystyczne dla biblioteki, dlatego można używać tylko reguł TLA z jednej biblioteki na raz. Z poziomu Template Editor lub Resource Editor przejdź do karty **Text Link Rules**. Na tej karcie można określić w szablonie bibliotekę zawierającą reguły TLA, których chcesz użyć lub które chcesz zmodyfikować. Z tego powodu zaleca się przechowywanie wszystkich reguł w jednej bibliotece, chyba że istnieje bardzo konkretny powód, by tego nie robić.



Rysunek 42. Karta Text Link Rules

Ważne: Karta Text Link Rules nie jest dostępna w przypadku zasobów języka japońskiego.

Od czego zacząć

Istnieje kilka sposobów rozpoczęcia pracy z edytorem na karcie Text Link Rules:

- Rozpocznij od symulacji wyników na próbie tekstu i zmodyfikuj lub utwórz reguły dopasowania na podstawie wyników wyodrębniania wzorców uzyskanych za pomocą bieżącego zestawu reguł.
- Utwórz nową regułę od początku lub zmodyfikuj istniejącą regułę.
- Pracuj bezpośrednio w widoku źródłowym.

Kiedy modyfikować lub tworzyć reguły

Wprawdzie reguły analizy powiązań w tekście dostarczane z każdym szablonem są często wystarczające do wyodrębnienia wielu prostych lub złożonych relacji z tekstu, niekiedy jednak konieczne może być wprowadzenie pewnych zmian do tych reguł lub utworzenie własnych reguł. Na przykład:

- Aby ujawnić pojęcie lub relację, która nie jest wyodrębniana w oparciu o istniejące reguły, poprzez utworzenie nowej reguły lub makra.
- Aby zmienić domyślne zachowanie typu dodanego do zasobów. Zwykle wymaga to edycji makra, takiego jak mTopic lub mNonLingEntities. Więcej informacji zawiera temat “Makra specjalne: mTopic, mNonLingEntities, SEP” na stronie 208.

- Aby dodawać nowe typy do istniejących reguł analizy powiązań w tekście i do makr. Na przykład, jeśli uważasz, że typ <Organization> jest zbyt szeroki, możesz utworzyć nowe typy dla organizacji z różnych sektorów gospodarki, takich jak <Pharmaceuticals>, <Car Manufacturing>, <Finance> i tak dalej. W takim przypadku należy zmodyfikować reguły analizy odsyłaczy tekstowych i/lub utworzyć makro, aby uwzględniać te nowe typy i odpowiednio je przetwarzać.
- Aby dodać typy do istniejącej reguły analizy powiązań w tekście. Załóżmy na przykład, że użytkownik ma regułę, która wychwytuje tekst john doe called jane doe, ale chce, aby ta reguła, która wychwytuje komunikację telefoniczną (called), wychwytywała także wymianę wiadomości e-mail. Można dodać do reguły typ obiektu nielingwistycznego oznaczający adres e-mail, aby wychwytywała ona również takie teksty, jak: johndoe @ ibm.com emailed janedoe @ ibm.com.
- Aby nieznacznie zmodyfikować istniejącą regułę, zamiast tworzyć nową. Załóżmy na przykład, że użytkownik ma regułę, która znajduje tekst xyz is very good, ale chce, aby ta reguła wychwytywała również tekst xyz is very, very good.

Symulowanie wyników analizy powiązań w tekście

Aby ułatwić sobie definiowanie reguł analizy powiązań w tekście i lepiej zrozumieć, w jaki sposób określone zdania są dopasowywane podczas analizy, często warto wybrać przykładowy fragment tekstu i przeprowadzić symulację. W trakcie symulacji, procedura jest uruchamiana tylko na próbie danych symulacji i reguły powiązań w tekście są stosowane w celu sprawdzenia, czy jakiekolwiek wzorce pasują do tej próby. Celem jest uzyskanie wyników symulacji, a następnie użycie tych wyników do udoskonalenia reguł, utworzenia nowych i uzyskania lepszej orientacji w działaniu mechanizmu dopasowywania. Dla każdego fragmentu tekstu (zdania, wyrazu lub klauzuli, w zależności od kontekstu) wynik symulacji zawiera zbiór leksemów i wszelkie reguły TLA, które ujawniły wzorzec w tym tekście. **Leksem** jest to dowolny wyraz lub wielowyrazowa fraza zidentyfikowana w procesie wyodrębniania.

W odróżnieniu od innych zasobów zaawansowanych, reguły TLA są charakterystyczne dla biblioteki, dlatego można używać tylko reguł TLA z jednej biblioteki na raz. Z poziomu Template Editor lub Resource Editor przejdź do karty **Text Link Rules**. Na tej karcie można określić w szablonie bibliotekę zawierającą reguły TLA, których chcesz użyć lub które chcesz zmodyfikować. Z tego powodu zaleca się przechowywanie wszystkich reguł w jednej bibliotece, chyba że istnieje bardzo konkretny powód, by tego nie robić.

Ważne! Jeśli używany jest plik danych, to powinien zawierać krótki tekst, aby czas przetwarzania był jak najkrótszy. Celem symulacji jest sprawdzenie, w jaki sposób tekst jest interpretowany, i zrozumienie działania reguł w odniesieniu do tego tekstu. Te informacje ułatwią użytkownikowi pisanie i edytowanie własnych reguł. Aby uzyskać bardziej kompletny zestaw danych, użyj węzła analizy powiązań w tekście lub uruchom strumień z sesją interaktywną i włączonym wyodrębnianiem TLA. Ta symulacja służy tylko do testowania i tworzenia reguł.

Definiowanie danych dla symulacji

Aby zorientować się, w jaki sposób reguły mogą wyszukiwać tekst, można uruchomić symulację na próbie danych. Pierwszym krokiem jest zdefiniowanie danych.

Definiowanie danych

1. Kliknij przycisk **Define data** w panelu symulacji w dolnej części karty **Text Link Rules**. Jeśli dane nie zostały wcześniej zdefiniowane, zamiast tego wybierz kolejno opcje **Tools > Run Simulation** z menu. Zostanie otwarty kreator danych symulacji.
2. Określ typ danych, wybierając jedną z następujących opcji:
   - **Paste or enter text directly** Dostępne jest pole tekstowe, w którym można wkleić tekst ze schowka lub ręcznie wprowadzić tekst, który ma zostać przetworzony. Można wprowadzić jedno zdanie na każdy wiersz lub użyć znaków interpunkcyjnych, takich jak kropki lub przecinki, do podziału zdania. Po wprowadzeniu tekstu można rozpocząć symulację, klikając opcję **Run Simulation**.
   - **Specify a file data source** Ta opcja wskazuje, że ma zostać przetworzony plik, który zawiera tekst. Kliknij przycisk **Next**, aby przejść do kroku kreatora, w którym można zdefiniować plik do przetworzenia. Po wybraniu pliku można rozpocząć symulację, klikając opcję **Run Simulation**. Obsługiwane są następujące typy plików:

.txt i .text. Wybrany plik jest odczytywany w trakcie symulacji bez wstępnego przetwarzania. Cały plik jest traktowany w taki sam sposób, jak w przypadku połączenia węzła File List lub Text Mining.

Ważne: Jeśli używany jest plik danych, to powinien zawierać krótki tekst, aby czas przetwarzania był jak najkrótszy. Celem symulacji jest sprawdzenie, w jaki sposób tekst jest interpretowany, i zrozumienie działania reguł w odniesieniu do tego tekstu. Te informacje ułatwią użytkownikowi pisanie i edytowanie własnych reguł. Aby uzyskać bardziej kompletny zestaw danych, użyj węzła analizy powiązań w tekście lub uruchom strumień z sesją interaktywną i włączonym wyodrębnianiem TLA. Ta symulacja służy tylko do testowania i tworzenia reguł.

3. Aby rozpocząć proces symulacji, kliknij przycisk **Run Simulation**. Zostanie wyświetlone okno dialogowe postępu. Jeśli użytkownik jest w sesji interaktywnej, podczas symulacji używane są ustawienia wyodrębniania aktualnie wybrane w sesji interaktywnej (sekcja **Tools > Extraction Settings** w widoku Concepts and Categories). Jeśli używane jest okno Template Editor, podczas symulacji używane są domyślne ustawienia wyodrębniania, które są takie same jak te pokazane na karcie Expert węzła Text Link Analysis. Aby uzyskać więcej informacji, patrz "Zrozumienie wyników symulacji".

Zrozumienie wyników symulacji

Aby zorientować się, w jaki sposób reguły mogą wyszukiwać tekst, można uruchomić symulację na próbie danych i przejrzeć jej wyniki. Następnie można zmienić zestaw reguł, aby lepiej pasował do danych. Po zakończeniu wyodrębniania i symulacji zostaną wyświetlone wyniki symulacji.

Dla każdego „zdania" zidentyfikowanego podczas wyodrębniania przedstawianych jest kilka informacji, w tym dokładna treść zdania, leksemy znalezione w tym zdaniu, a także wszystkie reguły, które pasowały do tekstu w zdaniu. Przez „**zdanie**" rozumiemy wyraz, zdanie lub klauzulę, w zależności od tego, w jaki sposób mechanizm wyodrębniania podzielił tekst na porcje.

Leksem jest to dowolny wyraz lub wielowyrazowa fraza zidentyfikowana w procesie wyodrębniania. Na przykład w zdaniu *My uncle lives in New York* mechanizm wyodrębniania może znaleźć następujące leksemy: *my*, *uncle*, *lives*, *in* i *new york*. Ponadto wyraz *uncle* (wujek) mógłby zostać wyodrębniony jako pojęcie typu <Unknown> (nieznanego), a nazwa *new york* mogłaby zostać wyodrębniona jako pojęcie typu <Location> (miejsce). Wszystkie pojęcia są leksemami, ale nie wszystkie leksemy są pojęciami. Pojęcia mogą być także innymi makrami, literałami łańcuchowymi i odstępami między wyrazami. Pojęciami mogą być tylko te wyrazy lub frazy, które mają przypisany typ.

Gdy użytkownik pracuje w sesji interaktywnej lub edytorze zasobów, operuje na poziomie pojęć. Reguły TLA są bardziej szczegółowe, a w definicji reguły mogą być używane pojedyncze leksemy ze zdania, nawet jeśli nie są nigdy wyodrębniane i przypisywane do typów. Możliwość wykorzystania leksemów, które nie są pojęciami, zapewnia większą elastyczność działania reguł, pozwalając na wychwytywanie złożonych relacji w tekście.

Jeśli w danych symulacji istnieje więcej niż jedno zdanie, można przechodzić do przodu i do tyłu między wynikami, klikając przyciski **Next** i **Previous**.

W przypadkach gdy zdanie nie pasuje do żadnej reguły TLA w wybranej bibliotece (nazwa biblioteki podana jest powyżej drzewa na tej karcie), wyniki są traktowane jako niezgodne, a przyciski **Next Unmatched** i **Previous Unmatched** są włączone, aby użytkownik wiedział, że istnieje tekst, dla którego żadna reguła nie znalazła dopasowania, i mógł szybko przechodzić do tych przypadków.

Po utworzeniu nowych reguł, edycji reguł lub zmianie zasobów bądź ustawieniu wyodrębniania można ponownie uruchomić symulację. Aby ponownie uruchomić symulację, kliknij przycisk **Run Simulation** w panelu symulacji. Ponownie zostaną użyte te same dane wejściowe.

W wynikach symulacji są przedstawione następujące pola i tabele:

Input text. Rzeczywiste „zdanie" zidentyfikowane przez proces wyodrębniania w danych symulacyjnych zdefiniowanych w kreatorze. Przez „zdanie" rozumiemy wyraz, zdanie lub klauzulę, w zależności od tego, w jaki sposób mechanizm wyodrębniania podzielił tekst na porcje.

System View. Zbiór leksemów rozpoznanych przez proces wyodrębniania.

- **Input Text Token.** Każdy leksem znaleziony w tekście wejściowym. Pojęcie leksemu zostało zdefiniowane wcześniej w tym temacie.
- **Typed As.** Jeśli leksem został zidentyfikowany jako pojęcie i został mu przypisany typ, to powiązana nazwa typu (np. <Unknown>, <Person>, <Location>) jest wyświetlana w tej kolumnie.
- **Matching Macro.** Jeśli leksem został dopasowany do istniejącego makra, to nazwa tego makra jest wyświetlana w tej kolumnie.

Rules Matched to Input Text. W tej tabeli wyświetlane są wszystkie reguły TLA, które zostały dopasowane na podstawie tekstu wejściowego. Dla każdej dopasowanej reguły zostanie wyświetlona nazwa reguły w kolumnie **Rule Output** i powiązane wartości wyjściowe dla tej reguły (pary pojęcie + typ). Można kliknąć dwukrotnie nazwę dopasowanej reguły, aby otworzyć regułę w panelu edytora powyżej panelu symulacji.

Przycisk **Generate Rule**. Po kliknięciu tego przycisku w panelu symulacji nowa reguła zostanie otwarta w oknie edytora reguł powyżej panelu symulacji. Tekst wejściowy zostanie wykorzystany jako przykład. Podobnie, każdy leksem z przypisanym typem lub dopasowany do makra w trakcie symulacji jest automatycznie wstawiany w kolumnie Elements w tabeli **Rule Values**. Jeśli leksem ma określony typ *i* pasuje do makra, to dla uproszczenia w regule będzie używane makro. Na przykład, jeśli używane są podstawowe zasoby języka angielskiego, zdanie „*I like pizza*" może podczas symulacji otrzymać typ <Unknown> i zostać dopasowane do makra mTopic. W takim przypadku wartość mTopic zostanie użyta jako element w wygenerowanej regule. Więcej informacji zawiera temat "Praca z regułami powiązań w tekście" na stronie 209.

Nawigacja wśród reguł i makr w drzewie

Jeśli podczas wyodrębniania prowadzona jest analiza powiązań w tekście, stosowane są reguły TLA zapisane w bibliotece wybranej na karcie **Text Link Rules**.

W odróżnieniu od innych zasobów zaawansowanych, reguły TLA są charakterystyczne dla biblioteki, dlatego można używać tylko reguł TLA z jednej biblioteki na raz. Z poziomu Template Editor lub Resource Editor przejdź do karty **Text Link Rules**. Na tej karcie można określić w szablonie bibliotekę zawierającą reguły TLA, których chcesz użyć lub które chcesz zmodyfikować. Z tego powodu zaleca się przechowywanie wszystkich reguł w jednej bibliotece, chyba że istnieje konkretny powód, by tego nie robić.

Możesz określić bibliotekę, w której chcesz pracować na karcie Text Link Rules, wybierając tę bibliotekę do listy rozwijanej **Use and store text link analysis rules in:** na tej karcie. Jeśli podczas wyodrębniania prowadzona jest analiza powiązań w tekście, stosowane są reguły TLA zapisane w bibliotece wybranej na karcie **Text Link Rules**. Z tego powodu, jeśli zdefiniowano reguły powiązań w tekście (reguły TLA) w więcej niż jednej bibliotece, tylko pierwsza biblioteka, w której znaleziono reguły TLA, zostanie użyta do analizy powiązań w tekście. Z tego powodu zaleca się przechowywanie wszystkich reguł w jednej bibliotece, chyba że istnieje bardzo konkretny powód, by tego nie robić.

Po wybraniu makra lub reguły w drzewie zawartość tego elementu jest wyświetlana w panelu edycji po prawej stronie. Jeśli klikniesz prawym przyciskiem myszy dowolny element w drzewie, zostanie otwarte menu kontekstowe czynności, takich jak:

- Utworzenie nowego makra w drzewie, a następnie otwarcie go w edytorze do prawej stronie.
- Utworzenie nowej reguły w drzewie, a następnie otwarcie jej w edytorze do prawej stronie.
- Utworzenie nowego zestawu reguł w drzewie.
- Wycinanie, kopiowanie i wklejanie elementów dla uproszczenia edycji.
- Usuwanie makr, reguł i zestawów reguł z zasobów.
- Wyłączanie makr, reguł i zestawów reguł, aby były ignorowane podczas przetwarzania.
- Przenoszenie reguł w górę lub w dół w celu zmiany kolejność przetwarzania.

Ostrzeżenia w drzewie

Ostrzeżenia w drzewie są wyświetlane z żółtym trójkątem i poinformują, że może wystąpić problem. Zatrzymaj wskaźnik myszy nad niepoprawnym makrem lub regułą, aby wyświetlić objaśnienie kontekstowe. W większości przypadków zostanie wyświetlone objaśnienie takie, jak: **Warning: No example provided; Enter an example**, które oznacza, że należy wprowadzić odpowiedni przykład.

Jeśli nie masz przykładu lub jeśli przykład nie jest zgodny z regułą, nie będzie można użyć funkcji Get Tokens, dlatego zaleca się, aby wprowadzić tylko jeden przykład na każdą regułę.

Gdy reguła jest podświetlona na żółto, oznacza to, że edytor TLA nie zna danego typu lub makra. Komunikat będzie podobny do następującego: **Warning: Unknown type or macro**. Informuje on, że element zdefiniowany w widoku źródłowym w postaci $nazwam na przykład $myType, nie jest typem istniejącym w bibliotece, ani nie jest makrem.

Aby zaktualizować dane w mechanizmie sprawdzania składni, należy przełączyć się do innej reguły lub makra; nie trzeba niczego rekompilować. Na przykład, jeśli reguła A wyświetla ostrzeżenie, ponieważ brakuje przykładu, kliknij górną lub dolną regułę, a następnie wróć do reguły A, aby sprawdzić, czy teraz jest poprawna.

Praca z makrami

Makra mogą uprościć strukturę reguł analizy powiązań w tekście, ponieważ umożliwiają grupowanie typów, innych makr i literałów (wyrazów) za pomocą operatora LUB (|). Korzyść ze stosowania makr polega nie tylko na możliwości wielokrotnego użycia tego samego makra w wielu regułach analizy powiązań w tekście, lecz także na propagowaniu zmiany w makrze do wszystkich reguł, w których jest stosowane. Większość gotowych reguł TLA dostarczanych z produktem zawiera predefiniowane makra. Makra są widoczne u góry drzewa w lewym panelu karty Text Link Rules.

W wynikach symulacji są przedstawione następujące pola i tabele:

Name. Unikalna nazwa identyfikująca makro. Zaleca się poprzedzanie nazw makr małą literą m, co pozwoli na szybkie identyfikowanie makr w regułach. Odwołując się do makr w regułach (ręcznie, tj. poprzez edycję bezpośrednią lub w widoku źródłowym), należy stosować przedrostek $, aby mechanizm wyodrębniania wiedział, że ma wyszukać nazwę makra. Jeśli jednak przeciągniesz i upuścisz nazwę makra lub dodasz makro z menu kontekstowego, to program automatycznie rozpozna makro i nie doda znaku $.

Tabela **Macro Value**.

- Szereg wierszy odzwierciedlających wszystkie możliwe wartości, jakie to makro może reprezentować. W wartościach tych rozróżniana jest wielkość liter.
- Wartości mogą być typami, literałami łańcuchowymi, odstępami między wyrazami, makrami lub dowolnymi kombinacjami tych elementów. Więcej informacji zawiera “Elementy obsługiwane w regułach i makrach” na stronie 214.
- Aby wprowadzić wartość elementu w makrze, kliknij dwukrotnie wiersz, w którym chcesz pracować. Pojawi się edytowalne pole, w którym można wprowadzić odwołanie do typu, odwołanie do makra, literał łańcuchowy albo odstęp między wyrazami. Zamiast tego można kliknąć w komórce prawym przyciskiem myszy, aby wyświetlić menu kontekstowe z listą często używanych makr, nazw typów i nazw typów nielingwistycznych. Aby odwołać się do typu lub makra, należy poprzedzić nazwę makra lub typu znakiem „$”. Na przykład $mTopic jest odwołaniem do makra mTopic. Wprowadzając więcej niż jeden argument, należy używać nawiasów ( ) do zgrupowania argumentów lub znaku | oznaczającego logiczne LUB.
- Za pomocą przycisków po prawej stronie tabeli Macro Value można dodawać i usuwać z niej wiersze.
- Każdy element należy wprowadzać w osobnym wierszu. Na przykład, aby utworzyć makro reprezentujące 3 literały łańcuchowe, takie jak am LUB was LUB is, należałoby wprowadzić każdy literał w osobnym wierszu, a tabela zawierałaby wówczas 3 wiersze.

Tworzenie i edytowanie makr

Użytkownik może tworzyć nowe lub edytować istniejące makra. Należy przestrzegać wskazówek i opisów właściwych dla edytora makr. Więcej informacji zawiera temat “Praca z makrami”.

Tworzenie nowych makr

1. Z menu wybierz opcje **Tools > New Macro**. Możesz też kliknąć ikonę New Macro na pasku narzędzi drzewa, aby otworzyć nowe makro w edytorze.
2. Wprowadź unikalną nazwę i zdefiniuj elementy wartości makra.
3. Po zakończeniu kliknij przycisk **Apply**, aby sprawdzić, czy nie wystąpiły błędy.

Edytowanie makr

1. Kliknij nazwę makra w drzewie. Makro zostanie otwarte w panelu edytora po prawej stronie.
2. Wprowadź zmiany.
3. Po zakończeniu kliknij przycisk **Apply**, aby sprawdzić, czy nie wystąpiły błędy.

Wyłączanie i usuwanie makr

Wyłączanie makr

Jeśli chcesz, aby makro było ignorowane podczas przetwarzania, możesz je wyłączyć. Może to powodować ostrzeżenia lub błędy w regułach, które wciąż odwołują się do wyłączonego makra. Należy zachować ostrożność podczas usuwania i wyłączania makr.

1. Kliknij nazwę makra w drzewie. Makro zostanie otwarte w panelu edytora po prawej stronie.
2. Kliknij prawym przyciskiem myszy na nazwie.
3. Z menu kontekstowego wybierz opcję **Disable**. Ikona makra staje się szara, a samo makro staje się nieedytowalne.

Usuwanie makr

Jeśli chcesz się pozbyć się makra, możesz je usunąć. Może to powodować lub błędy w regułach, które wciąż odwołują się do usuniętego makra. Należy zachować ostrożność podczas usuwania i wyłączania makr.

1. Kliknij nazwę makra w drzewie. Makro zostanie otwarte w panelu edytora po prawej stronie.
2. Kliknij prawym przyciskiem myszy na nazwie.
3. Z menu kontekstowego wybierz opcję **Delete**. Makro zniknie z listy.

Sprawdzanie, zapisywanie i anulowanie

Stosowanie zmian w makrach

Po kliknięciu poza edytorem makra lub wybraniu opcji **Apply** makro jest automatycznie przeglądane w poszukiwaniu błędów. Jeśli zostanie znaleziony błąd, należy wyeliminować go przed przejściem do innej części aplikacji.

Jednak mniej poważne błędy powodują wyświetlenie jedynie ostrzeżenia. Na przykład, jeśli makro zawiera niekompletne lub nieprzywoływane definicje typów lub innych makr, zostanie wyświetlony komunikat ostrzegawczy. Po kliknięciu przycisku **Apply** jakiekolwiek nieskorygowane ostrzeżenia spowodują, że po lewej stronie nazwy makra w drzewie reguł i makr w lewym panelu pojawi się ikona ostrzeżenia.

Zastosowanie makra nie oznacza, że makro zostanie trwale zapisane. Zastosowanie spowoduje sprawdzanie poprawności reguły i zgłoszenie ewentualnych błędów i ostrzeżeń.

Zapisywanie zasobów w sesji pracy z interaktywnym pulpitem roboczym

1. Aby zapisać zmiany wprowadzone w zasobach podczas sesji w interaktywnym panelu roboczym, dzięki czemu można je będzie wykorzystać przy następnym uruchomieniu strumienia, należy:
   - Zaktualizować węzeł modelowania, aby mieć pewność, że będzie można uzyskać te same zasoby przy następnym uruchomieniu strumienia. Więcej informacji zawiera temat “Aktualizowanie węzłów modelowania i zapisywanie” na stronie 77. Następnie należy zapisać strumień. Chcąc zapisać strumień, należy zrobić to w oknie głównym IBM SPSS Modeler po zmianie węzła modelowania.

2. Aby zapisać zmiany wprowadzone w zasobach podczas sesji w interaktywnym panelu roboczym, dzięki czemu można je będzie wykorzystać w innych strumieniach, należy:

- Zaktualizować używany szablon lub utworzyć nowy. Więcej informacji zawiera temat “Tworzenie i modyfikowanie szablonów” na stronie 155. Nie spowoduje to zapisania zmian dla bieżącego węzła (patrz poprzedni krok).
- Lub zaktualizować używany pakiet TAP. Więcej informacji zawiera temat “Aktualizowanie pakietów analizy tekstu” na stronie 131.

Zapisywanie zasobów w oknie Template Editor

1. Należy najpierw opublikować bibliotekę. Więcej informacji zawiera temat “Publikowanie bibliotek” na stronie 173.
2. Następnie należy zapisać szablon za pomocą opcji **File > Save Resource Template** w menu.

Anulowanie zmian makra

1. Aby odrzucić zmiany, kliknij przycisk **Cancel**.

Makra specjalne: mTopic, mNonLingEntities, SEP

Szablon Opinions (i jemu podobne) oraz szablony Basic Resources zawierają dwa marka specjalne o nazwach mTopic i mNonLingEntities.

mTopic

Domyślnie makro mTopic grupuje wszystkie zawarte w szablonie typy, które prawdopodobnie mają związek z opinią, takie jak następujące typy z biblioteki *Core* : <Person> (osoba), <Organization> (organizacja), <Location> (miejsce) i tak dalej, pod warunkiem że typ nie jest typem opinii (na przykład <Negative> (negatywna) lub <Positive> (pozytywna)) bądź typem zdefiniowanym jako obiekt nielingwistyczny na karcie Advanced Resources.

Za każdym razem, gdy użytkownik tworzy nowy typ w szablonie Opinions (lub podobnym), program zakłada, że o ile ten typ nie jest określony w innym makrze lub sekcji obiektów nielingwistycznych karty Advanced Resource, będzie traktowany tak samo, jak inne typy zdefiniowane w makrze mTopic.

Załóżmy, że utworzyliśmy w zasobach nowe typy na podstawie szablonu Opinions: <Vegetables> (warzywa) i <Fruit> (owoce). Nie musimy wprowadzać żadnych zmian, by nowe typy były traktowane tak samo, jak typy mTopic, możemy więc automatycznie ujawniać opinie pozytywne, negatywne, neutralne i kontekstowe dotyczące nowych typów. W trakcie wyodrębniania zdanie „*I enjoy broccoli, but I hate grapefruit*” (Lubię brokuły, ale nie cierpię grejpfrutów) wygeneruje 2 wzorce wynikowe:

broccoli <Vegetables> + like <Positive>

grapefruit <Fruit> + dislike <Negative>

Jeśli jednak chcemy traktować te typy inaczej niż pozostałe typy z makra mTopic, możemy dodać nazwę typu do istniejącego makra, na przykład mPos, by zgrupować wszystkie typy opinii pozytywnych, lub utworzyć nowe makro, do którego będzie się można później odwoływać w jednej lub wielu regułach.

Ważne! Jeśli utworzony zostanie nowy typ, taki jak <Vegetables>, to zostanie on uwzględniony jako typ w makrze mTopic, ale jego nazwa nie będzie jawnie widoczna w definicji makra.

mNonLingEntities

Podobnie, jeśli dodamy nowe obiekty nielingwistyczne w sekcji **Nonlinguistic Entities** karty Advanced Resources, to będą one automatycznie przetwarzane jako typy z makra mNonLingEntities, chyba że nakażemy inaczej. Więcej informacji zawiera temat “Obiekty nielingwistyczne” na stronie 192.

SEP

Dostępne jest także predefiniowane makro SEP, które odpowiada globalnemu separatorowi zdefiniowanemu na komputerze lokalnym, z reguły przecinkowi (,).

Praca z regułami powiązań w tekście

Reguła analizy powiązań w tekście to zapytanie boolowskie, które jest używane do przeprowadzenia dopasowania w zdaniu. Reguły analizy powiązań w tekście zawiera jeden lub więcej z następujących argumentów: typy, makra, literały łańcuchowe lub odstępy między wyrazami. Aby można było wyodrębnić wyniki TLA, potrzebna jest co najmniej jedna reguła analizy powiązań w tekście.

W edytorze reguł na karcie Text Link Rules wyświetlane są następujące obszary i pola:

Pole **Name**. Unikalna nazwa reguły powiązań w tekście.

Pole **Example**. Opcjonalnie można uwzględnić przykładowe zdanie lub sekwencję słów, która zostałaby wychwycona przez tę regułę. Zaleca się podawanie przykładów. W edytorze można wygenerować leksemy na podstawie przykładu, aby zobaczyć, w jaki sposób tekst będzie dopasowywany do reguły i jakie wyniki zostałyby wygenerowane. **Leksem** jest to dowolny wyraz lub wielowyrazowa fraza zidentyfikowana w procesie wyodrębniania. Na przykład w zdaniu *My uncle lives in New York* mechanizm wyodrębniania może znaleźć następujące leksemy: *my*, *uncle*, *lives*, *in* i *new york*. Ponadto wyraz *uncle* (wujek) mógłby zostać wyodrębniony jako pojęcie typu <Unknown> (nieznanego), a nazwa *new york* mogłaby zostać wyodrębniona jako pojęcie typu <Location> (miejsce). Wszystkie pojęcia są leksemami, ale nie wszystkie leksemy są pojęciami. Pojęcia mogą być także innymi makrami, literałami łańcuchowymi i odstępami między wyrazami. Pojęciami mogą być tylko te wyrazy lub frazy, które mają przypisany typ.

Tabela **Rule Value**. Ta tabela zawiera elementy reguły, które są używane w celu dopasowania reguły do zdania. Można dodawać lub usuwać wiersze tabeli za pomocą przycisków po jej prawej stronie. Tabela zawiera 3 kolumny:

- Kolumna **Element**. Wprowadź wartości lub kombinacje wartości typów, literałów łańcuchowych, odstępów między wyrazami (<Any Token>) lub makr. Więcej informacji zawiera temat “Elementy obsługiwane w regułach i makrach” na stronie 214. Kliknij dwukrotnie komórkę elementu, aby wprowadzić informacje bezpośrednio. Zamiast tego można kliknąć w komórce prawym przyciskiem myszy, aby wyświetlić menu kontekstowe z listą często używanych makr, nazw typów i nazw typów nielingwistycznych. Należy pamiętać, że wpisując informacje bezpośrednio w komórce, należy poprzedzić nazwę typu z makra lub typu znakiem $, np. $mTopic dla makra mTopic. Kolejność tworzenia wierszy elementów wpływa na sposób, w jaki reguła zostanie dopasowana do tekstu. Podając więcej niż jeden argument, należy używać nawiasów ( ) do grupowania argumentów i znaku | jako logicznego operatora LUB. Należy pamiętać, że w wartościach rozróżniana jest wielkość liter.
- Kolumna **Quantity**. Wskazuje minimalną i maksymalną liczbę wystąpień elementu wymaganą, by został uznany za dopasowany. Na przykład, jeśli chcesz zdefiniować odstęp (lub szereg wyrazów) o długości od 0 do 3 wyrazów, możesz wybrać opcję **Between 0 and 3** z listy lub wprowadzić liczby bezpośrednio w oknie dialogowym. Wartością domyślną jest „**Exactly 1**”, czyli dokładnie 1 wyraz. W niektórych przypadkach element powinien być opcjonalny. Wówczas będzie miał minimum równe 0, a maksimum większe od 0 (tj. 0 lub 1, pomiędzy 0 i 2). Należy zwrócić uwagę, że pierwszy element w regule nie może być opcjonalny, co oznacza, że nie może mieć minimum równego 0.
- Kolumna **Example Token**. Po kliknięciu przycisku **Get Tokens** program dzieli tekst z pola **Example** na leksemy i wypełnia tę kolumnę leksemami, które pasują do elementów zdefiniowanych przez użytkownika. Można również wyświetlić te elementy w tabeli wynikowej.

Tabela **Rule Output** Każdy wiersz w tej tabeli definiuje sposób, w jaki wynikowy wzorzec TLA zostanie przedstawiony w wynikach. Wynik reguły może zawierać wzorce obejmujące maksymalnie sześć par kolumn pojęcie/para, z których każda reprezentuje jeden *parametr*. Na przykład wzorzec typów <Location> + <Positive> składa się z dwóch parametrów, czyli 2 par pojęcie/typ.

Uwaga: Terminy w kolumnie **Element** tabeli **Rule Value** lub w dowolnej z kolumn **Concept** tabeli **Rule Output** nie mogą zaczynać się od żadnego z następujących znaków: `, #, %, ^, *, _, -, :, <, >, /, \, ".

Język zapewnia swobodę wyrażania tych samych ogólnych pojęć na wiele różnych sposobów, dlatego możesz zdefiniować wiele różnych reguł w celu wychwytywania tego samego ogólnego pojęcia. Na przykład tekst *"Paris is a*

place I love" i tekst *"I really, really like Paris and Florence"* odzwierciedlają to samo ogólne pojęcie — że Paryż jest lubiany — ale wyrażone w różny sposób. By je wychwycić, potrzeba dwóch różnych reguł. Łatwiej jest jednak pracować z wynikowymi wzorcami, jeśli podobne pojęcia są pogrupowane. Z tego powodu, choć możesz mieć 2 różne reguły wychwytujące te 2 frazy, możesz zdefiniować ten sam wynik dla obu reguł, na przykład jako wzorzec typów <Location> + <Positive>. Taki wynik będzie odzwierciedlał znaczenie obu tekstów. Jak widać, wyniki nie zawsze odzwierciedlają strukturę i kolejności wyrazów w tekście oryginalnym. Ponadto taki wzorzec typów może pasować do innych fraz i generować takie wzorce pojęć, jak: paris + like i tokyo + like.

Aby ułatwić sobie szybkie i bezbłędne definiowanie wyników, można użyć menu kontekstowego, aby wybierać elementy, które mają być ujęte w wynikach. Można także przeciągać i upuszczać elementy z tabeli Rule Value do wyników. Na przykład, jeśli masz regułę, która zawiera odwołanie do makra mTopic w wierszu 2 tabeli Rule Value, i chcesz, by ta wartość znalazła się w wynikach, możesz po prostu przeciągnąć i upuścić element mTopic do pierwszej pary kolumn w tabeli Rule Output. Spowoduje to automatyczne wpisanie pojęcia i typu w parze. A jeśli chcesz, aby wyniki rozpoczynały się od typu zdefiniowanego przez trzeci element (wiersz 3) z tabeli Rule Value, przeciągnij typ z tej tabeli do komórki **Type 1** w tabeli wynikowej. Tabela zostanie zaktualizowana i pojawi się w niej odwołanie do wiersza w nawiasie (3).

Możesz także ręcznie wprowadzać takie odwołania do tabeli, klikając dwukrotnie komórkę w kolumnie **Concept**, a następnie wprowadzając symbol $, po którym następuje numer wiersza, na przykład $2, aby odwołać się do elementu zdefiniowanego w wierszu 2 tabeli Rule Value. Wprowadzając informacje ręcznie, należy również zdefiniować kolumnę **Type**; wprowadź symbol #, po którym następuje numer wiersza, na przykład #2, aby odwołać się do elementu zdefiniowanego w wierszu 2 tabeli Rule Value.

Można stosować obie metody łącznie. Załóżmy, że masz typ <Positive> w wierszu 4 wartość tabeli Rule Value. Możesz przeciągnąć go do kolumny Type 2, a następnie kliknąć dwukrotnie komórkę w kolumnie Concept 2, po czym ręcznie wprowadzić słowo *not* przed typem. Kolumna wynikowa zawierałaby wartość not (4) lub, jeśli jesteś trybie edycji lub trybie kodu źródłowego, wartość not $4. Teraz można kliknąć prawym przyciskiem myszy w kolumnie Type 1 i wybrać, na przykład, makro o nazwie mTopic. Taka definicja wyniku mogłaby wygenerować taki wzorzec pojęcia, jak: car + bad.

Większość reguł ma tylko jeden wiersz wyników, ale istnieją przypadki, gdy możliwy i pożądany jest więcej niż jeden wynik. W takim przypadku należy zdefiniować jeden wynik na każdy wiersz tabeli Rule Output.

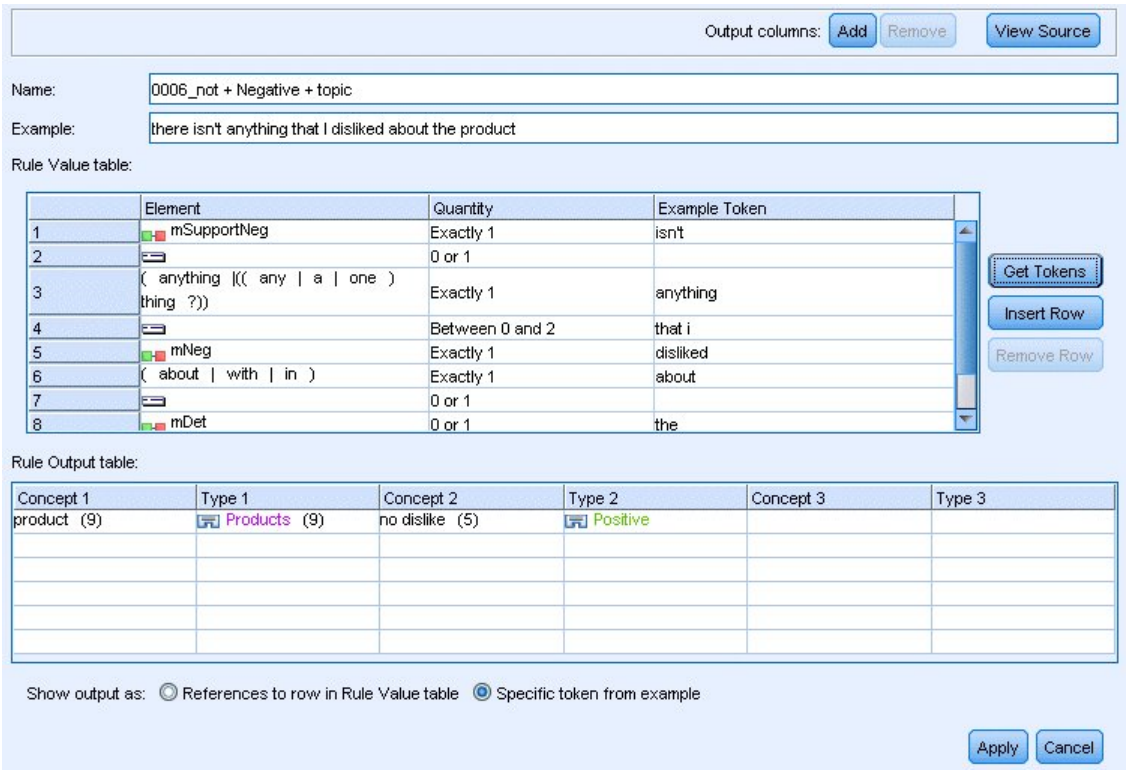
Ważne: Należy pamiętać, że podczas wyodrębniania wzorców TLA wykonywane są także inne operacje lingwistyczne. Jeśli więc definicja wyniku ma wartość t$3\t#3, oznacza to, że wzorzec będzie zawierał ostateczne pojęcie trzeciego elementu i ostateczny typ trzeciego elementu po wykonaniu kompletnego przetwarzania lingwistycznego (wyszukiwania synonimów i innych operacji grupowania).

- **Show output as**. Domyślnie opcja **References to row in Rule Value table** jest zaznaczona, a wynik jest wyświetlany przy użyciu odwołań liczbowych do wierszy, zgodnie z definicją na karcie Rule Value. Jeśli wcześniej kliknięto opcję Get Tokens i masz leksemy w kolumnie Example Tokens tabeli Rule Value, można wyświetlić wyniki dla tych konkretnych leksemów, wybierając odpowiednią opcję.

Uwaga: Jeśli w tabeli wynikowej nie ma wystarczającej liczby wynikowych par pojęcie/typ, można dodać kolejną parę, klikając przycisk Add na pasku narzędzi edytora. Jeśli obecnie wyświetlane są 3 pary i klikniesz przycisk Add, to do tabeli dodane zostaną 2 kolumny (Concept 4 i Type 4). Oznacza to, że tabela wynikowa dla wszystkich reguł będzie zawierała 4 pary. Istnieje również możliwość usunięcia nieużywanych par, o ile żadna z reguł w zestawie reguł w tej bibliotece nie korzysta z usuwanych par.

Przykładowa reguła

Załóżmy, że zasoby zawierają następujące reguły analizy powiązań w tekście oraz że włączono wyodrębnianie wyników TLA:



Rysunek 43. Karta Text Link Rules: edytor reguł

Przy każdym wyodrębnianiu mechanizm wyodrębniania odczyta każde zdanie i spróbuje przeprowadzić dopasowanie zgodnie z następującą sekwencją:

Tabela 44. Przykład sekwencji wyodrębniania

Element (wiersz)	**Opis argumentów**
1	Pojęcie z jednego z typów reprezentowanych przez makra mPos lub mNeg albo z typu <Uncertain>.
2	Pojęcie, któremu przypisano jeden z typów reprezentowanych przez makro mTopic.
3	Jeden z wyrazów reprezentowanych przez makro mBe.
4	Opcjonalny element, 0 lub 1 wyraz, tzw. odstęp między wyrazami lub <Any Token>
5	Pojęcie, któremu przypisano jeden z typów reprezentowanych przez makro mTopic.

W tabeli wynikowej widzimy, że od reguły oczekuje się wzorca, który obejmuje pojęcia lub typy odpowiadające makru mTopic zdefiniowane w wierszu 5 tabeli **Rule Value** + pojęcia lub typy odpowiadające makru mPos, mNeg bądź typowi <Uncertain> zdefiniowane w wierszu 1 tabeli **Rule Value**. Wzorcem tym może być: sausage + like lub <Unknown> + <Positive>.

Tworzenie i edytowanie reguł

Można tworzyć nowe reguły lub edytować istniejące. Należy przestrzegać wskazówek i opisów właściwych dla edytora reguł. Więcej informacji zawiera temat “Praca z regułami powiązań w tekście” na stronie 209.

Tworzenie nowych reguł

1. Z menu wybierz opcje **Tools > New Rule**. Można też kliknąć ikonę New role na pasku narzędzi drzewa, aby otworzyć nową regułę w edytorze.
2. Wprowadź unikalną nazwę i zdefiniuj elementy wartości reguły.

3. Po zakończeniu kliknij przycisk **Apply**, aby sprawdzić, czy nie wystąpiły błędy.

Edytowanie reguł

1. Kliknij nazwę reguły w drzewie. Reguła zostanie otwarta w panelu edycji po prawej stronie.
2. Wprowadź zmiany.
3. Po zakończeniu kliknij przycisk **Apply**, aby sprawdzić, czy nie wystąpiły błędy.

Wyłączanie i usuwanie reguł

Wyłączanie reguł

Jeśli chcesz, aby reguła była ignorowana podczas przetwarzania, możesz ją wyłączyć. Należy zachować ostrożność podczas usuwania i wyłączania reguł.

1. Kliknij nazwę reguły w drzewie. Reguła zostanie otwarta w panelu edycji po prawej stronie.
2. Kliknij prawym przyciskiem myszy na nazwie.
3. Z menu kontekstowego wybierz opcję **Disable**. Ikona reguły staje się szara, a sama reguła staje się nieedytowalna.

Usuwanie reguł

Jeśli chcesz się pozbyć reguły, możesz ją usunąć. Należy zachować ostrożność podczas usuwania i wyłączania reguł.

1. Kliknij nazwę reguły w drzewie. Reguła zostanie otwarta w panelu edycji po prawej stronie.
2. Kliknij prawym przyciskiem myszy na nazwie.
3. Z menu kontekstowego wybierz opcję **Delete**. Reguła zniknie z listy.

Sprawdzanie, zapisywanie i anulowanie

Stosowanie zmian w regułach

Po kliknięciu poza edytorem reguły lub wybraniu opcji **Apply** reguła jest automatycznie przeglądana w poszukiwaniu błędów. Jeśli zostanie znaleziony błąd, należy wyeliminować go przed przejściem do innej części aplikacji.

Jednak mniej poważne błędy powodują wyświetlenie jedynie ostrzeżenia. Na przykład, jeśli reguła zawiera niekompletne lub nieprzywoływane definicje typów lub makr, zostanie wyświetlony komunikat ostrzegawczy. Po kliknięciu przycisku **Apply** jakiekolwiek nieskorygowane ostrzeżenia spowodują, że po lewej stronie nazwy makra w drzewie reguł i makr w lewym panelu pojawi się ikona ostrzeżenia.

Zastosowanie reguły nie oznacza, że reguła zostanie trwale zapisana. Zastosowanie spowoduje sprawdzanie poprawności reguły i zgłoszenie ewentualnych błędów i ostrzeżeń.

Zapisywanie zasobów w sesji pracy z interaktywnym pulpitem roboczym

1. Aby zapisać zmiany wprowadzone w zasobach podczas sesji w interaktywnym panelu roboczym, dzięki czemu można je będzie wykorzystać przy następnym uruchomieniu strumienia, należy:
   - Zaktualizować węzeł modelowania, aby mieć pewność, że będzie można uzyskać te same zasoby przy następnym uruchomieniu strumienia. Więcej informacji zawiera temat “Aktualizowanie węzłów modelowania i zapisywanie” na stronie 77. Następnie należy zapisać strumień. Chcąc zapisać strumień, należy zrobić to w oknie głównym IBM SPSS Modeler po zmianie węzła modelowania.
2. Aby zapisać zmiany wprowadzone w zasobach podczas sesji w interaktywnym panelu roboczym, dzięki czemu można je będzie wykorzystać w innych strumieniach, należy:
   - Zaktualizować używany szablon lub utworzyć nowy. Więcej informacji zawiera temat “Tworzenie i modyfikowanie szablonów” na stronie 155. Nie spowoduje to zapisania zmian dla bieżącego węzła (patrz poprzedni krok).
   - Lub zaktualizować używany pakiet TAP. Więcej informacji zawiera temat “Aktualizowanie pakietów analizy tekstu” na stronie 131.

Zapisywanie zasobów w oknie Template Editor

1. Należy najpierw opublikować bibliotekę. Więcej informacji zawiera temat “Publikowanie bibliotek” na stronie 173.
2. Następnie należy zapisać szablon za pomocą opcji **File > Save Resource Template** w menu.

Anulowanie zmian w regułach

1. Aby odrzucić zmiany, kliknij przycisk **Cancel** w panelu edytora.

Kolejność przetwarzania reguł

Jeśli podczas wyodrębniania prowadzona jest analiza powiązań w tekście, „zdanie” (klauzula, wyraz, fraza) będzie dopasowywane do kolejnych reguł, dopóki nie zostanie stwierdzona zgodność lub nie zostanie zbadania ostatnia reguła. Położenie w drzewie określa kolejność, w jakiej reguły są badane. Zgodnie ze sprawdzoną procedurą reguły należy uporządkować od najbardziej zawężających do najbardziej ogólnych. Najbardziej zawężające reguły powinny być umieszczone w górnej części drzewa. Aby zmienić położenie konkretnej reguły lub zestawu reguł, wybierz opcję **Move up** lub **Move down** z menu kontekstowego drzewa reguł i makr albo użyj przycisków strzałek w górę i w dół na pasku narzędzi.

W widoku źródłowym nie można zmienić kolejności reguł, przenosząc je w edytorze. Im wyżej reguła jest wyświetlana w widoku źródłowym, tym wcześniej będzie przetwarzana. Zaleca się zmienianie kolejności reguł wyłącznie w drzewie, co pozwoli uniknąć problemów z kopiowaniem/wklejaniem.

Ważne! W poprzednich wersjach produktu IBM SPSS Modeler Text Analytics wymagane były unikalne, liczbowe identyfikatory reguł. Począwszy od wersji 18.1 można jedynie wskazać kolejność przetwarzania poprzez przenoszenie reguły w górę lub w dół w drzewie lub przez określenie ich pozycji w widoku źródła.

Załóżmy na przykład, że tekst zawiera następujące dwa zdania:

I love anchovies

I love anchovies and green peppers

Dodatkowo przyjmijmy, że istnieją dwie reguły analizy powiązań w tekście z następującymi wartościami:

A

	Element	Quantity	Example Token
1	Positive	Exactly 1	
2	mDet	0 or 1	
3	mTopic	Exactly 1	
4			
5			
6			
7			

B

	Element	Quantity	Example Token
1	Positive	Exactly 1	
2	mDet	0 or 1	
3	mTopic	Exactly 1	
4	(SEP | and | or)	1 or 2	
5	mDet	0 or 1	
6	mTopic	Exactly 1	
7			

Rysunek 44. 2 reguły przykładowe

W widoku źródłowym wartości reguły mogą wyglądać następująco:

A: value = $Positive $mDet? $mTopic

B: value = $Positive $mDet? $mTopic ($SEP|and|or){1,2} $mDet? $mTopic

Jeśli reguła **A** jest wyżej w drzewie niż reguła **B**, to reguła **A** będzie przetwarzana jako pierwsza, a pierwsze zdanie *I love anchovies and green peppers* zostanie dopasowane do $Positive $mDet? $mTopic i wygeneruje niekompletny wzorzec wynikowy (anchovies + like), ponieważ reguła nie szukała 2 dopasowań do $mTopic.

Dlatego, aby wychwycić prawdziwe znaczenie tekstu, najbardziej zawężająca reguła, w tym przypadku **B**, musi zostać umieszczona wyżej w drzewie niż reguła bardziej ogólna, w tym przypadku **A**.

Praca z zestawami reguł (wiele przejść)

Zestawy reguł są przydatnym rozwiązaniem do grupowania pokrewnych reguł w drzewie reguł i makr, pozwalającym na przetwarzanie danych w wielu przejściach. Definicja zestawu reguł składa się tylko z jego nazwy i służy do porządkowania reguł w użytecznych grupach. W niektórych kontekstach tekst jest zbyt skomplikowany i zróżnicowany, by można było go przetworzyć w jednym przejściu. Na przykład tekstowe dane wywiadowcze mogą zawierać informacje o powiązaniach między osobami, które dadzą się ujawnić na podstawie metody kontaktu (*x zadzwonił do y*), relacji rodzinnych (*y jest szwagrem x*), transakcji pieniężnych (*x przelał 1000 zł na konto y*) i tak dalej. W takim przypadku przydatna jest możliwość tworzenia specjalnych zestawów reguł analizy powiązań w tekście, z których każda dotyczyć będzie określonego rodzaju relacji, np. ujawniania kontaktów, ujawniania powiązań rodzinnych itd.

Aby utworzyć zestaw reguł, wybierz opcję „Create Rule Set" z menu kontekstowego drzewa reguł i makr lub z paska narzędzi. Można teraz tworzyć nowe reguły bezpośrednio pod węzłem zestawu reguł w drzewie i przenosić istniejące reguły do zestawu reguł.

Podczas wyodrębniania z użyciem zasobów, w których reguły są pogrupowane w zestawy, mechanizm wyodrębniania musi wielokrotnie przechodzić przez tekst, aby w każdym przejściu dopasowywać inny rodzaj wzorców. W ten sposób jedno „zdanie" może być dopasowane do każdej reguły w zestawie reguł, a bez zestawu reguł byłoby dopasowane tylko do jednej reguły.

Uwaga: Jeden zestaw może zawierać maksymalnie 512 reguł.

Tworzenie nowych zestawów reguł

1. Z menu wybierz kolejno opcje **Tools > New Rule Set**. Zamiast tego możesz kliknąć ikonę New Rule Set na pasku narzędzi drzewa. Zestaw reguł pojawi się w drzewie reguł.
2. Dodaj nowe reguły do tego zestawu lub przenieś do niego istniejące reguły.

Wyłączanie zestawów reguł

1. Kliknij prawym przyciskiem myszy nazwę zestawu reguł w drzewie.
2. Z menu kontekstowego wybierz opcję **Disable**. Zestaw reguł stanie się szary, a wszelkie zawarte w nim reguły także będą wyłączone i ignorowane podczas przetwarzania.

Usuwanie zestawów reguł

1. Kliknij prawym przyciskiem myszy nazwę zestawu reguł w drzewie.
2. Z menu kontekstowego wybierz opcję **Delete**. Zestaw reguł i zawarte w nim reguły zostaną usunięte z zasobów.

Elementy obsługiwane w regułach i makrach

Jako wartości parametrów w regułach analizy powiązań w tekście oraz makrach dopuszczalne są następujące argumenty:

Makra

Makra można użyć bezpośrednio w regule analizy powiązań w tekście lub w innym makrze. Jeśli wprowadzasz nazwę makra ręcznie lub w widoku źródłowym (a nie wybierasz jej z menu kontekstowego), koniecznie poprzedź ją znakiem dolara ($), na przykład $mTopic. W nazwach makr rozróżniana jest wielkość liter. W menu kontekstowych można wybierać dowolne makra zdefiniowane w danej chwili na karcie Text Link Rules.

Typy

Można pisać bezpośrednio w treści reguły analizy powiązań w tekście lub w makrze. Jeśli wprowadzasz nazwę typu ręcznie lub w widoku źródłowym (a nie wybierasz jej z menu kontekstowego), koniecznie poprzedź ją znakiem dolara ($), na przykład $Person. W nazwach typów rozróżniana jest wielkość liter. Korzystając z menu kontekstowego, możesz wybrać dowolny typ z obecnie używanego zestawu zasobów.

Próba odwołania się do nierozpoznanego typu spowoduje wyświetlanie komunikatu ostrzegawczego, a do czasu wyeliminowania problemu obok reguły w drzewie reguł i makr widoczna będzie ikona ostrzeżenia.

Literały łańcuchowe

Aby uwzględnić informacje, które nigdy nie zostały wyodrębnione z analizowanego tekstu, można zdefiniować literał łańcuchowy, a mechanizm wyodrębniania będzie poszukiwał tego literału. Wszystkim wyodrębnionym wyrazom lub frazom przepisane zostały typy i dlatego nie można ich używać w literałach łańcuchowych. Jeśli użyjesz wyrazu, który został wyodrębniony, to zostanie on zignorowany, nawet jeśli ma przypisany typ <Unknown>.

Literał łańcuchowy może składać się z jednego lub większej liczby słów. Przy definiowaniu listy literałów łańcuchowych obowiązują następujące reguły:

- Listę łańcuchów należy ująć w nawiasy, np. (his). Chcąc wskazać kilka alternatywnych literałów łańcuchowych, należy rozdzielić je separatorem LUB, na przykład (a|an|the) lub (his|hers|its).
- Można używać wyrazów prostych lub złożonych.
- Wyrazy na liście należy rozdzielać znakiem |, który pełni funkcję logicznego operatora LUB.
- Jeśli chcesz wyszukiwać zarówno formę pojedynczą, jak i mnogą, wprowadź obie. Program nie generuje automatycznie różnych form gramatycznych wyrazu.
- Używaj tylko małych liter.
- Aby wielokrotnie używać tych samych literałów łańcuchowych, zdefiniuj je jako makro i używaj tego makra w innych makrach i regułach analizy powiązań w tekście.
- Jeśli łańcuch zawiera kropki lub łączniki, to należy uwzględnić je w literale. Na przykład, aby wyszukiwać w tekście łańcuch a.k.a, wprowadź literalnie litery i kropki: a.k.a.

Operator wykluczenia

Znak ! służy jako operator wykluczenia, który powoduje, że zanegowane w ten sposób wyrażenie nie zajmie określonego miejsca. Operator wykluczenia możesz dodać wyłącznie ręcznie, bezpośrednio edytując komórkę (kliknij komórkę dwukrotnie w tabeli Rule Value lub Macro Value table) lub w widoku źródłowym. Na przykład, jeśli do reguły analizy powiązań w tekście dodasz zapis $mTopic @{0,2} !($Positive) $Budget, to program będzie wyszukiwał teksty zawierające (1) termin przypisany do dowolnego z typów w makrze mTopic, (2) odstęp między wyrazami o długości od zera do dwóch wyrazów, (3) nie będzie wyszukiwał instancji terminu przypisanego do typu <Positive>, (4) termin przypisany do typu <Budget>. Taka reguła może wychwycić "*cars have an inflated price tag*" (samochody o zawyżonych cenach), ale zignorowałaby frazę "*store offers amazing discounts*" (sklep oferuje fantastyczne zniżki).

Aby skorzystać z tego operatora, należy ręcznie wpisać wykrzyknik i nawiasy do komórki elementu, po uprzednim dwukrotnym kliknięciu tej komórki.

Odstępy między wyrazami (<Any Token>)

Odstęp między wyrazami, określany też symbolem <Any Token> definiuje przedział liczby leksemów, które mogą być obecne między dwoma elementami. Odstępy między wyrazami są przydatne przy dopasowywaniu bardzo podobnych fraz, które mogą różnić się tylko nieznacznie ze względu na obecność dookreślników, przedimków, przymiotników lub innych podobnych wyrazów.

Tabela 45. Przykład elementów w tabeli Rule Value bez odstępu między wyrazami

#	Element
1	 Nieznane
2	 mBeHave
3	 Dodatnie

Uwaga: W widoku źródłowym ta wartość jest zdefiniowana jako: $Unknown $mBeHave $Positive

Pasuje ona zdania takie jak "*the hotel staff was nice*", gdzie *hotel staff* należy do typu <Unknown>, *was* należy do makra mBeHave, a *nice* ma typ <Positive>. Ale nie znajdzie zdania "*the hotel staff was very nice*".

Tabela 46. Przykład elementów w tabeli Rule Value z odstępem między wyrazami (<Any Token>)

#	Element
1	 Nieznane
2	 mBeHave
3	
4	 Dodatnie

Uwaga: W widoku źródłowym ta wartość jest zdefiniowana jako: $Unknown $mBeHave @{0,1} $Positive

Jeśli do reguły dodasz odstęp między wyrazami, reguła znajdzie zarówno zdanie "*the hotel staff was nice*", jak i zdanie "*the hotel staff was very nice*".

W widoku źródłowym lub podczas bezpośredniej edycji odstęp między wyrazami oznaczony jest jako @{#,#}, gdzie @ oznacza odstęp między wyrazami, a {#,#} określa minimalną i maksymalna liczbę wyrazów między poprzedzającym a następującym elementem. Na przykład @{1,3} powoduje, że dwa określone elementy zostaną znalezione, jeśli między nimi znajduje się co najmniej jeden wyraz, ale nie więcej niż trzy wyrazy. @{0,3} powoduje, że dwa określone elementy zostaną znalezione, jeśli między nimi znajduje się 0, 1, 2 lub 3 wyrazy, ale nie więcej niż trzy.

Przeglądanie i praca w trybie źródłowym

Dla każdej reguły i makra edytor TLA generuje bazowy kod źródłowy, który jest używany przez mechanizm wyodrębniający do dopasowywania i generowania wyników analizy TLA. Jeśli preferujesz pracę bezpośrednio z kodem, możesz wyświetlić ten kod źródłowy i edytować go bezpośrednio. W tym celu kliknij przycisk „View Source" w górnej części edytora. Nastąpi przejście do aktualnie wybranej reguły lub makra i podświetlenie tego elementu. Jednak zaleca się korzystanie z paneli edytora, co ograniczy ryzyko popełnienia błędów.

Po zakończeniu przeglądania lub edycji źródła kliknij opcję **Exit Source**. Jeśli wprowadzisz niepoprawną składnię reguły, trzeba będzie ją skorygować przed wyjściem z widoku źródłowego.

Ważne: W przypadku edycji w widoku źródłowym zdecydowanie zaleca się edytowanie reguł i makr pojedynczo. Po zakończeniu edycji makra należy sprawdzić poprawność wyników, przeprowadzając wyodrębnianie. Jeśli wynik jest zadowalający, zaleca się zapisanie szablonu przed wprowadzeniem następnej zmiany. Jeśli wynik jest niezadowalający lub wystąpił błąd, należy przywrócić zapisane zasoby.

Makra w widoku źródłowym

```
[macro]
name = nazwa_makra
value = ([type_name|macro_name|literal_string|word_gap])
```

Tabela 47. Wpisy makr

[macro]	Każde makro musi zaczynać się od wiersza ze słowem [macro], które wyznacza początek makra.
name	Nazwa definicji makra. Każda nazwa musi być unikalna.
value	Kombinacja jednego lub większej liczby typów, literałów łańcuchowych, odstępów między wyrazami lub makr. Więcej informacji zawiera "Elementy obsługiwane w regułach i makrach" na stronie 214. Wprowadzając więcej niż jeden argument, należy używać nawiasów ( ) do zgrupowania argumentów lub znaku | oznaczającego logiczne LUB.

Oprócz wskazówek i zasad składniowych przedstawionych w sekcji dotyczącej makr z pracą w widoku źródłowym wiąże się konieczność przestrzegania kilku dodatkowych wytycznych, które nie są wymagane podczas pracy w widoku edytora. Podczas pracy z makrami w trybie kodu źródłowego należy też przestrzegać następujących zasad:

- Każde makro musi zaczynać się od wiersza ze słowem [macro], które wyznacza początek makra.
- Aby wyłączyć element, należy umieścić wskaźnik komentarza (#) na początku wiersza.

Przykład. W tym przykładzie zdefiniowano makro o nazwie mTopic. mTopic oznacza obecność terminu pasującego do *jednego* z następujących typów: <Product>, <Person>, <Location>, <Organization>, <Budget> lub <Unknown>.

```
[macro]
name=mTopic
value=($Unknown|$Product|$Person|$Location|$Organization|$Budget|$Currency)
```

Reguły w widoku źródłowym

```
[pattern(ID)]
name = nazwa_wzorca
value = [$nazwa_typu|nazwa_makra|odstępy_między_wyrazami|literały_łańcuchowe]
output = $digit[\t]#digit[\t]$digit[\t]#digit[\t]$digit[\t]#digit[\t]
```

Tabela 48. Wpisy reguł

[pattern (<ID>)]	Wskazuje początek tej reguły analizy powiązań w tekście i udostępnia unikalny identyfikator liczbowy określający kolejność przetwarzania.
name	Określa unikalną nazwę tej reguły analizy powiązań w tekście.

Tabela 48. Wpisy reguł (kontynuacja)

value	Określa składnię i argumenty, które mają być dopasowywane do tekstu. Więcej informacji zawiera temat “Elementy obsługiwane w regułach i makrach” na stronie 214.
wyniki	Format wyjściowy dla wynikowych wzorców wykrytych w tekście. Wyniki nie zawsze odzwierciedlają dokładne pozycje pierwotnych elementów w tekście źródłowym. Ponadto można zdefiniować wiersze wyników dla jednej reguły analizy powiązań w tekście, umieszczając każdy wynik w osobnym wierszu. Składnia wyników: • Wyniki należy oddzielać kodem tabulacji \t, na przykład $1\t#1\t$3\t#3 • $ i numer oznacza termin pasujący do argumentu zdefiniowanego w parametrze o tym numerze. Zatem $1 oznacza termin pasujący do pierwszego argumentu zdefiniowanego dla wartości. • # i numer oznacza nazwę typu elementu na tej pozycji. Jeśli element jest listą literałów łańcuchowych, przypisany zostanie typ <Unknown>. • Wartość Null\tNull nie spowoduje utworzenia żadnych wyników.

Oprócz wskazówek i zasad składniowych przedstawionych w sekcji dotyczącej reguł z pracą w widoku źródłowym wiąże się konieczność przestrzegania kilku dodatkowych wytycznych, które nie są wymagane podczas pracy w widoku edytora. Podczas pracy z regułami w trybie kodu źródłowego należy też przestrzegać następujących zasad:

- Definiując dwa lub większą liczbę elementów, należy ująć je w nawiasy, niezależnie od tego, czy są opcjonalne (na przykład ($Negative|$Positive) *lub* ($mCoord|$SEP)?). $SEP oznacza przecinek.
- Pierwszy element w regule analizy powiązań w tekście nie może być opcjonalny. Na przykład nie można rozpocząć od value = $mTopic? ani value = @{0,1}.
- Możliwe jest powiązanie ilości (lub liczby wystąpień) z leksemem. Jest to przydatne, jeśli chcemy zdefiniować jedną regułę, która obejmuje wszystkie przypadki, zamiast pisać osobne reguły dla każdego przypadku. Na przykład, można użyć literału łańcuchowego ($SEP|and), próbując znaleźć , (przecinek) albo wyraz and. Jeśli uzupełnisz definicję o ilość, literal przyjmie postać ($SEP|and){1,2} i pasuje do każdego z następujących wystąpień: "," "and" ", and".
- Między nazwą makra a znakami $ i ? nie są dozwolone spacje w wartości reguły analizy powiązań w tekście.
- Spacje nie są dozwolone w definicji wyników reguły analizy powiązań w tekście.
- Aby wyłączyć element, należy umieścić wskaźnik komentarza (#) na początku wiersza.

Przykład. Załóżmy, że zasoby zawierają następującą reguły analizy TLA oraz że włączono wyodrębnianie wyników TLA:

```
## 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
```

Przy każdym wyodrębnianiu mechanizm wyodrębniania odczyta każde zdanie i spróbuje przeprowadzić dopasowanie zgodnie z następującą sekwencją:

Tabela 49. Przykład sekwencji wyodrębniania

Pozycji w zbiorze	Opis argumentów
1	Nazwisko osoby ($Person),
2	Jeden albo dwa spośród następujących elementów: przecinek ($SEP), określnik ($mDet), czasownik pomocniczy ($mSupport), łańcuchy „then” lub „as”,
3	0 lub 1 wyraz (@{0,1})
4	Funkcja ($Function)

Tabela 49. Przykład sekwencji wyodrębniania (kontynuacja)

Pozycji w zbiorze	Opis argumentów
5	Jeden z następujących łańcuchów: „of”, „with”, „for”, „in”, „to” lub „at”,
6	0 lub 1 wyraz (@{0,1})
7	Nazwa organizacji ($Organization)
8	0, 1 lub 2 wyrazy (@{0,2})
9	Nazwa miejsca ($Location)

Ta reguła analizy powiązań w tekście znalazłaby takie zdania lub frazy, jak:

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

Ta przykładowa reguła analizy powiązań w tekście wygenerowałaby następujące wyniki:

```
jean doe <Person> hr director <Function> ibm <Organization> france <Location>
```

Gdzie:

- jean doe odpowiada elementowi $1 (pierwszemu elementowi reguły TLA), a <Person> to typ terminu jean doe (#1),
- hr director to termin odpowiadający elementowi $4 (4. elementowi w regule TLA), a <Function> to typ terminu hr director (#4),
- ibm to termin odpowiadający $7 (7. element 7 w regule TLA), a <Organization> to typ terminu ibm. (#7),
- france to termin odpowiadający elementowi $9 (9. elementowi w regule TLA), a <Location> to typ terminu france (# 9).

Zestawy reguł w widoku źródłowym

```
[set(<ID>)]
```

Gdzie [set (<ID>)] oznacza początek zestawu reguł i zawiera unikalny identyfikator liczbowy określający kolejność przetwarzania zestawów.

Przykład. Następujące zdanie zawiera informacje na temat osób, ich funkcji w ramach firmy, a także aktywności firmy w obszarze fuzji/przejęć.

Org1 Inc has entered into a definitive merger agreement with Org2 Ltd, said John Doe, CEO of Org2 Ltd.

Można zapisać jedną regułę z kilkoma wynikami, by uwzględnić wszystkie możliwe wyniki, na przykład:

```
## Org1 Inc entered into a definitive merger agreement with Org2 Ltd, said
John Doe, CEO of Org2 Ltd.

[pattern(020)]
name=020
value = $Organization @{0,4} $ActionNouns @{0,6}  $mOrg @{1,2}
 $Person @{0,2} $Function @{0,1} $Organization
output = $1\t#1\t$3\t#3\t$5\t#5
output = $7\t#7\t$9\t#9\t$11\t#11
```

co spowoduje wygenerowanie następujących 2 wzorców wynikowych:

- org1 inc<Organization> + merges with <ActiveVerb> + org2 ltd<Organization>
- john doe <Person> + ceo <Function> + org2 ltd<Organization>

Ważne! Należy pamiętać, że podczas wyodrębniania wzorców TLA wykonywane są także inne operacje lingwistyczne. W tym przypadku na etapie grupowania synonimów termin merger zostanie zgrupowany z terminem merges with. A ponieważ merges with należy do typu <ActiveVerb>, to ta nazwa typu pojawi się w ostatecznym wynikowym wzorcu TLA. Jeśli więc definicja wyniku ma wartość t$3\t#3, oznacza to, że wzorzec będzie zawierał ostateczne pojęcie trzeciego elementu i ostateczny typ trzeciego elementu po wykonaniu kompletnego przetwarzania lingwistycznego (wyszukiwania synonimów i innych operacji grupowania).

Zamiast pisać skomplikowane reguły, takich jak przedstawiona powyżej, często łatwiej jest operować na dwóch regułach. Pierwsza będzie wyspecjalizowana wyszukiwaniu wzmianek o fuzjach i przejęciach:

```
[set(1)]
## Org1 Inc has entered into a definitive merger agreement with Org2 Ltd
[pattern(44)]
name=firm + action + firm_0044
value=$mOrg @{0,20} $ActionNouns @{0,6} $mOrg
output(1)=$1\t#1\t$3\t#3\t$5\t#5
```

i zwróci org1 inc<Organization> + merges with <ActiveVerb> + org2 ltd <Organization>

Druga będzie wyspecjalizowana w wyszukiwaniu osób/funkcji/firm:

```
[set(2)]
## said John Doe, CEO of Org2 Ltd
[pattern(52)]
name=individual + role + firm_0007
value=$Person @{0,3} $mFunction (at|of)? ($mOrg|$Media|$Unknown)
output(1)=$1\t#1\t$3\tFunction\t$5\t#5
```

i zwróci john doe <Person> + ceo <Function> + org2 ltd <Organization>

Uwagi

Informacje te zostały opracowane dla produktów i usług oferowanych na całym świecie.

Produktów, usług lub opcji opisywanych w tym dokumencie IBM nie musi oferować we wszystkich krajach. Informacje o produktach i usługach dostępnych w danym kraju można uzyskać od lokalnego przedstawiciela IBM. Odwołanie do produktu, programu lub usługi IBM nie oznacza, że można użyć wyłącznie tego produktu, programu lub usługi IBM. Zamiast nich można zastosować ich odpowiednik funkcjonalny pod warunkiem, że nie narusza to praw własności intelektualnej IBM. Jednakże cała odpowiedzialność za ocenę przydatności i sprawdzenie działania produktu, programu lub usługi pochodzących od producenta innego niż IBM spoczywa na użytkowniku.

IBM może posiadać patenty lub złożone wnioski patentowe na towary i usługi, o których mowa w niniejszej publikacji. Przedstawienie niniejszej publikacji nie daje żadnych uprawnień licencyjnych do tychże patentów. Pisemne zapytania w sprawie licencji można przesyłać na adres:

IBM Director of Licensing
IBM Corporation
North Castle Drive, MD-NC119
Armonk, NY 10504-1785
U.S.A.

Zapytania dotyczące zestawów znaków dwubajtowych (DBCS) należy kierować do lokalnych działów własności intelektualnej IBM (IBM Intellectual Property Department) lub wysłać je na piśmie na adres:

Intellectual Property Licensing
Legal and Intellectual Property Law
IBM Japan, Ltd.
19-21, Nihonbashi-Hakozakicho, Chuo-ku
Tokio 103-8510, Japonia

INTERNATIONAL BUSINESS MACHINES CORPORATION DOSTARCZA TĘ PUBLIKACJĘ W STANIE, W JAKIM SIĘ ZNAJDUJE ("AS IS") BEZ UDZIELANIA JAKICHKOLWIEK GWARANCJI (RĘKOJMIĘ RÓWNIEŻ WYŁĄCZA SIĘ), WYRAŹNYCH LUB DOMNIEMANYCH, A W SZCZEGÓLNOŚCI DOMNIEMANYCH GWARANCJI PRZYDATNOŚCI HANDLOWEJ, PRZYDATNOŚCI DO OKREŚLONEGO CELU ORAZ GWARANCJI, ŻE PUBLIKACJA TA NIE NARUSZA PRAW OSÓB TRZECICH. Ustawodawstwa niektórych krajów nie dopuszczają zastrzeżeń dotyczących gwarancji wyraźnych lub domniemanych w odniesieniu do pewnych transakcji; w takiej sytuacji powyższe zdanie nie ma zastosowania.

Informacje zawarte w niniejszej publikacji mogą zawierać nieścisłości techniczne lub błędy drukarskie. Informacje te są okresowo aktualizowane, a zmiany te zostaną uwzględnione w kolejnych wydaniach tej publikacji. IBM zastrzega sobie prawo do wprowadzania ulepszeń i/lub zmian w produktach i/lub programach opisanych w tej publikacji w dowolnym czasie, bez wcześniejszego powiadomienia.

Wszelkie wzmianki w tej publikacji na temat stron internetowych innych podmiotów zostały wprowadzone wyłącznie dla wygody użytkownika i w żadnym wypadku nie stanowią zachęty do ich odwiedzania. Materiały dostępne na tych stronach nie są częścią materiałów opracowanych dla tego produktu IBM, a użytkownik korzysta z nich na własną odpowiedzialność.

IBM ma prawo do używania i rozpowszechniania informacji przysłanych przez użytkownika w dowolny sposób, jaki uzna za właściwy, bez żadnych zobowiązań wobec ich autora.

Licencjobiorcy tego programu, którzy chcieliby uzyskać informacje na temat programu w celu: (i) wdrożenia wymiany informacji między niezależnie utworzonymi programami i innymi programami (łącznie z tym opisywanym) oraz (ii) wspólnego wykorzystywania wymienianych informacji, powinni skontaktować się z:

IBM Director of Licensing
IBM Corporation
North Castle Drive, MD-NC119
Armonk, NY 10504-1785
U.S.A.

Informacje takie mogą być udostępnione, o ile spełnione zostaną odpowiednie warunki, w tym, w niektórych przypadkach, zostanie uiszczona stosowna opłata.

Licencjonowany program opisany w niniejszej publikacji oraz wszystkie inne licencjonowane materiały dostępne dla tego programu są dostarczane przez IBM na warunkach określonych w Umowie IBM z Klientem, Międzynarodowej Umowie Licencyjnej IBM na Program lub w innych podobnych umowach zawartych między IBM i użytkownikami.

Dane dotyczące wydajności i cytowane przykłady zostały przedstawione jedynie w celu zobrazowania sytuacji. Faktyczne wyniki dotyczące wydajności mogą się różnić w zależności do konkretnych warunków konfiguracyjnych i operacyjnych.

Informacje dotyczące produktów innych podmiotów niż IBM zostały uzyskane od dostawców tych produktów, z ich publicznych ogłoszeń lub innych dostępnych publicznie źródeł. IBM nie testował tych produktów i nie może potwierdzić dokładności pomiarów wydajności, kompatybilności ani żadnych innych danych związanych z tymi produktami. Pytania dotyczące możliwości produktów innych podmiotów należy kierować do dostawców tych produktów.

Wszelkie stwierdzenia dotyczące przyszłych kierunków rozwoju i zamierzeń IBM mogą zostać zmienione lub wycofane bez powiadomienia.

Publikacja ta zawiera przykładowe dane i raporty używane w codziennej pracy. W celu kompleksowego ich zilustrowania, podane przykłady zawierają nazwiska osób prywatnych, nazwy przedsiębiorstw oraz nazwy produktów. Wszystkie te nazwy/nazwiska są fikcyjne i jakiekolwiek podobieństwo do istniejących nazw/nazwisk jest całkowicie przypadkowe.

Znaki towarowe

IBM, logo IBM i ibm.com są znakami towarowymi lub zastrzeżonymi znakami towarowymi International Business Machines Corp. zarejestrowanymi w wielu systemach prawnych na całym świecie. Pozostałe nazwy produktów i usług mogą być znakami towarowymi IBM lub innych przedsiębiorstw. Aktualna lista znaków towarowych IBM dostępna jest w serwisie WWW IBM, w sekcji "Copyright and trademark information" Informacje o prawach autorskich i znakach towarowych), pod adresem www.ibm.com/legal/copytrade.shtml.

Intel, logo Intel, Intel Inside, logo Intel Inside, Intel Centrino, logo Intel Centrino, Celeron, Intel Xeon, Intel SpeedStep, Itanium i Pentium są znakami towarowymi lub zastrzeżonymi znakami towarowymi Intel Corporation lub przedsiębiorstw podporządkowanych Intel Corporation w Stanach Zjednoczonych i w innych krajach.

Linux jest zastrzeżonym znakiem towarowym Linusa Torvaldsa w Stanach Zjednoczonych i/lub w innych krajach.

Microsoft, Windows, Windows NT oraz logo Windows są znakami towarowymi Microsoft Corporation w Stanach Zjednoczonych i/lub w innych krajach.

UNIX jest zastrzeżonym znakiem towarowym The Open Group w Stanach Zjednoczonych i/lub w innych krajach.

Java oraz wszystkie znaki towarowe i logo dotyczące języka Java są znakami towarowymi lub zastrzeżonymi znakami towarowymi Oracle i/lub przedsiębiorstw afiliowanych.

Inne nazwy produktów i usług mogą być znakami towarowymi IBM lub innych podmiotów.

Indeks

Znaki specjalne

A

B

C

D

E

F

G

H

I

J

K

L

Ł

M

N

O

P

R

S

T

U

W

Z

Drukowane w USA