



## SAS® Text Miner 4.2

Wykorzystaj wartość ukrytą w informacjach tekstowych.

### Do czego służy SAS® Text Miner?

SAS Text Miner wykorzystuje zaawansowane algorytmy lingwistyczne, oparte na silniku Teragram, dostępne w ramach podstawowego rozwiązania do eksploracji danych (ang. data mining) SAS® Enterprise Miner™. Połączenie strukturalnej, ilościowej analizy danych z analizą niestrukturalną zwykłego tekstu daje kompletny i zrozumiały wgląd w zintegrowane środowisko modelowania predykcyjnego. Automatyzacja ręcznego przetwarzania źródeł danych tekstowych, wykorzystanie interaktywnych raportów szczegółowych oraz dostarczenie algorytmów do prowadzenia zaawansowanych analiz umożliwiają wydajniejsze przewidywanie przyszłych trendów i wykorzystywanie nowych możliwości przy jednoczesnym zmniejszeniu ryzyka.

### Dlaczego SAS® Text Miner?

SAS Text Miner oszczędza pieniądze i zasoby, dzięki automatyzacji czasochłonnych procesów czytania i rozumienia tekstu. Łącząc utrudnione i nieustrukturalizowane źródła danych dostarcza dokładniejszy i bardziej kompletny wgląd w organizację. Na podstawie e analiz przeprowadzonych przy użyciu obu typów dokumentów tworzone są modele opisowe i predykcyjne, które dostarczają więcej obszarów do wnioskowania.

### Dla kogo jest przeznaczony SAS® Text Miner?

Program SAS Text Miner zaprojektowano głównie z myślą o analitykach i statystykach biznesowych, którzy muszą przekopywać się przez duże ilości tekstu w celu wydobycia informacji, pomysłów i trendów. Program znajdzie zastosowanie w każdej branży i w sektorze publicznym i jest szczególnie przydatny dla organizacji, które aktywnie budują modele predykcyjne.

Organizacje kumulują ogromne ilości tekstu. Informacje od klientów, wiadomości e-mail, dokumenty sieci Web, blogi, wpisy w portalu Twitter, okólniki, roszczenia gwarancyjne, sondaże, artykuły prasowe, wyniki badań, CV, informacje o klientach, wywiad konkurencyjny ... lista może ciągnąć się w nieskończoność. Nikt nie ma czasu, by to wszystko przeczytać, nie mówiąc już o organizacji, klasyfikacji lub doszukiwaniu się głębszego znaczenia w tych wszystkich fragmentach informacji.

Aby z zebranych danych wydobyć jak największą wartość, trzeba je przeanalizować, i podjąć działania. Ze względu na niejednoznaczność języka mówionego, najważniejsze wiadomości ukryte w danych tekstowych są trudne do wykrycia, a ich przetwarzanie często niemożliwe. Większość organizacji nie dysponuje mechanizmem połączenia informacji strukturalizowanych i niestrukturalizowanych i nie ma możliwości wykorzystania ich w procesie podejmowania decyzji.

Dzięki programowi SAS Text Miner można dokonać klasyfikacji dokumentów w ramach zdefiniowanych kategorii danych i odnaleźć wyraźne związki lub skojarzenia między tematami. Interaktywna eksploracja umożliwia wykrycie wzorów w zbiorach dokumentów i zastosowanie wyciągniętych wniosków bezpośrednio w przygotowywaniu modeli predykcyjnych, co zapewni wydobyć maksymalnej wartości ze wszystkich źródeł informacji.

### Kluczowe korzyści

- **Skrócenie czasu potrzebnego na podjęcie decyzji dzięki zautomatyzowanym procesom.** Wdrożenie inteligentnych algorytmów i technik przetwarzania słów sprawia, że czasochłonne zadania wykonywane na ogół ręcznie, tj. kategoryzacja, oznaczanie lub tworzenie bibliotek tematycznych i indeksów dokumentów, realizowane są automatycznie i wydajnie.

- **Wykrycie niedostrzeżonych wcześniej związków i powiązań.** Po co ograniczać analityków tekstu do wyszukiwania określonych terminów i znanych wyrażen? Dzięki rozbudowanemu, interaktywnemu interfejsowi użytkownika, który zaznacza ścieżki i odnośniki umożliwiające dokładniejszą analizę dokumentów, program SAS Text Miner oferuje unikalną, opartą na danych metodę identyfikacji nowych koncepcji.
- **Wizualna prezentacja zaawansowanego podglądu danych z możliwością przejścia do konkretnej frazy w dokumencie.** SAS Text Miner oferuje wizualną prezentację całego procesu eksploracji danych z możliwością podglądu żądanych szczegółów, ilustrujących powiązania między danymi oraz z możliwością eksploracji powiązań między pozycjami w zbiorach dokumentów.
- **Program umożliwia rozpoznawanie trendów i wyszukiwanie możliwości biznesowych dzięki pełnemu pakietowi narzędzi do modelowania predykcyjnego.** Analiza informacji, tj. pisma od klientów i notatki centrum obsługi klienta, oferują wartościowe informacje o braku zadowolenia wśród klientów oraz wgląd w potrzeby związane z produktami lub usługami.

### Opis rozwiązania

SAS Text Miner oferuje szeroki pakiet narzędzi lingwistycznych i narzędzi do modelowania analitycznego służących do odnajdywania i wydobywania wiedzy z wielu dokumentów tekstowych. Po konwersji tekstu na format rozpoznawany przez silnik eksploracji danych tematy i zagadnienia identyfikowane są jako wyraźne związki, by dokumenty można było łączyć w powiązane grupy dla celów oznaczenia przed modelowaniem predykcyjnym. Teraz dostępne są zaawansowane algorytmy wyszukiwania, rozbudowane sprawdzanie pisowni oraz analizowanie dokumentów pod kątem szeregu tematów. Wyniki uzyskane

w programach SAS Enterprise Content Categorization lub SAS Concept Creation do modułu SAS Text Miner mogą zostać bezpośrednio zintegrowane z eksploracją danych w celu uzupełnienia parametrów określonych przez użytkownika.

### Dostęp do szeregu formatów dokumentów i języków

SAS Text Miner potrafi odczytywać tekst zapisany w wielu formatach dokumentów; a kreator przetwarzania wstępnego pomaga użytkownikowi w konwersji plików na zbiory danych SAS celem wprowadzenia ich do programu SAS Text Miner. Dzięki temu można dokonać analizy informacji z poziomu pojedynczego, zintegrowanego systemu wykorzystującego dane pochodzące z wielu różnych źródeł, w tym z Internetu i sieci społecznościowych, dzięki opcji przeszukiwania sieci Web. Spersonalizowane procedury i słowniki są dostępne dla języków arabskiego, chińskiego, holenderskiego, angielskiego, francuskiego, niemieckiego, włoskiego, japońskiego, koreańskiego, polskiego, portugalskiego, hiszpańskiego i szwedzkiego. Wydobycie obiektów możliwe jest w każdym z obsługiwanych języków. Języki, które nie są obecnie obsługiwane, można zakodować i analizować przy pomocy kodowania Unicode UTF-8.

### Elastyczny, przyjazny dla użytkownika interfejs

Architektura klienta Java/serwera SAS oferuje zawierające dużo informacji podsumowania graficzne, ułatwiające

przejście do dokumentów tekstowych w celu uzyskania większej ilości informacji. Dzięki wielopoziomym połączeniom z serwerem procesy obliczeniowe można oddzielić od interfejsu użytkownika. Zaawansowane serwery UNIX i Windows mogą zostać oddelegowane do intensywnej eksploracji danych, podczas gdy użytkownicy pracują na komputerach stacjonarnych. Dzięki temu osiągnięto niespotykaną dotychczas elastyczność konfiguracji w zakresie od platform dla użytkowników indywidualnych po rozwiązania dla przedsiębiorstwa. Dodatkowo interfejs automatycznie generuje kody wyników punktowych w trakcie budowania modeli. Taki kod można wyeksportować i wykorzystać w innym programie do analizy wywiadu biznesowego, w tym Microsoft Excel, SAS Enterprise Content Categorization, SAS® Enterprise Guide® lub JMP®.

### Kompleksowa analiza tekstu

Analiza tekstu dokonuje rozkładu danych tekstowych i generuje tworzy prezentację ilościową nadającą się do celów eksploracji danych. Moduł do analizy tekstu (nowa wersja w SAS Text Miner 4.2 ) dokonuje rozkładu tekstu na części mowy, adresy, numery telefonów i nazwy firm, w tym tematy i rdzenie wyrazów. Ten zaawansowany analizator składni umożliwia określenie, jakie słowa zostaną zignorowane lub wskazanie, jakie słowa traktowane będą jako synonimy. Zaprzeczenia, zwroty wielowyrzowe i obiekty zdefiniowane przez użytkownika są dodatkowym uzupełnieniem poprzedniej wersji tej funkcji.

### Redukcja wymiarów

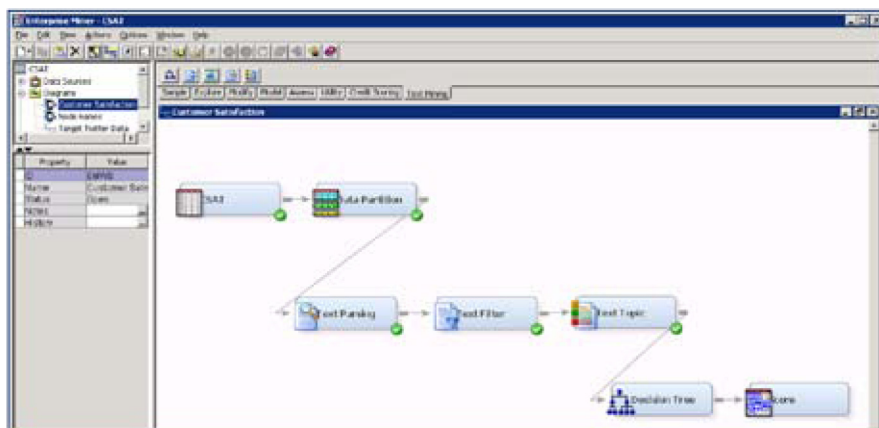
Zaawansowane filtrowanie przy pomocy ważenia, zintegrowanego sprawdzania pisowni oraz przekształcania danych jakościowych na kompaktowe formaty jest możliwe dzięki złożonym technikom redukcji wymiarów wykorzystywanym przez program SAS Text Miner. Przeanalizowane dokumenty można przekształcić na odwzorowanie liczbowe przy pomocy rozkładu według wartości osobliwych (SVD), zwijania wyrażen lub połączenia obu tych technik.

### Identyfikacja tematu tekstu i grupowanie

Zaawansowane algorytmy automatycznie grupują dokumenty pod kątem podobnych motywów i tematów w oparciu o ich treść. Zamiast stawiania wymagania, by dokumenty dotyczyły tego samego tematu („grupowanie twarde”), nowy moduł Text Topic (temat tekstu) w programie SAS Text Miner 4.2 działa w oparciu o założenie, że każdy dokument może dotyczyć kilku lub ani jednego z określonych tematów. Użytkownik może zdefiniować te tematy lub mogą one zostać wybrane automatycznie przez narzędzie. Interaktywny interfejs modułu Text Topic umożliwi na bieżąco podgląd grup dokumentów związanych z różnymi tematami oraz modyfikację definicji tematów. Alternatywnie, jeśli potrzebne jest grupowanie twarde, moduł Text Mining (eksploracja tekstu) wykorzystywany jest do umieszczenia tematów w hierarchii grup lub liście grup. Procesy grupowania oparte na algorytmie maksymalizacji wartości oczekiwanej (EM) stosują techniki grupowania przestrzennego w celu zorganizowania dokumentów w grupy znaczeniowe. Podsumowania grup można wyświetlić w sposób łatwy do zinterpretowania w kontekście oryginalnych dokumentów tekstowych. Środowisko interaktywnej wizualizacji umożliwia analitykom zbadanie koncepcji i powiązań między dokumentami oraz dynamiczne wprowadzanie modyfikacji celem dalszego dostosowania analiz do potrzeb.

### Filtrowanie tekstu

Moduł Text Filter (nowa wersja w SAS Text Miner 4.2) oferuje zintegrowaną funkcję przeszukiwania całego tekstu i automatyzuje sprawdzanie pisowni



SAS Text Miner 4.2 zawiera trzy nowe moduły (Text Parsing, Text Filter oraz Text Topic). Zwroty wielowyrzowe, opcje przeszukiwania całego tekstu oraz możliwość dodania obiektów zdefiniowanych przez użytkownika, to tylko niektóre z nowych opcji.

ni, łączenie koncepcji oraz tworzenie podzbiorów zwrotów i dokumentów. Interaktywna kwerenda odnajdzie poszczególne dokumenty pasujące do wprowadzonych kryteriów wyszukiwania. Filtry mogą bazować na dowolnym wyznaczniku, w tym na obecności lub braku obecności konkretnych wyrażen, a interaktywna wizualizacja umożliwia sondowanie do momentu odnalezienia dokumentów i wyrażen odpowiadających danej specyfikacji. Mapy koncepcji łączą wyrażenia i obiekty w wizualny, interaktywny sposób, co umożliwia wykrycie niedostrzeżonych wcześniej wzorów.

### **Bezpośrednia implementacja wyników z innych programów SAS® Analytics**

Płynna integracja z najlepszym oprogramowaniem SAS do modelowania predykcyjnego lub innymi narzędziami z nowej oferty SAS Text Analytics oferuje szeroką gamę narzędzi do eksploracji danych tekstowych i strukturalnych, a także narzędzi do wstępnego przetwarzania, oznaczania i wykorzystywania otrzymanych danych. Organizacje wykorzystujące nagradzane oprogramowanie analityczne SAS mogą wdrożyć analitykę do swojego środowiska operacyjnego w celu skutecznego zidentyfikowania i rozwiązania krytycznych kwestii biznesowych.

## **Główne cechy**

### **Uniwersalny dostęp do danych**

- Dostęp do wielu postaci danych tekstowych, w tym plików PDF, rozszerzonej tablicy ASCII, HTML, formatów Microsoft Office, arkuszy kalkulacyjnych, prezentacji, wiadomości e-mail i formatów baz danych.
- Funkcje przeszukiwania sieci Web, w tym portali społecznościowych, tj. Twitter, oraz kanałów wiadomości.
- Możliwość wydobywania, przekształcania oraz wgrywania danych tekstowych do zestawu danych SAS celem eksploracji.

### **Obsługa wielu języków**

- Obsługa kodowania for Latin-1, znaków dwubitowych oraz UTF-8.
- Języki europejskie (kodowanie Latin-1): holenderski, angielski, francuski, niemiecki, włoski, polski\*\*\*, portugalski, hiszpański i szwedzki.
- Języki wschodnie (obsługa znaków dwubitowych): arabski, chiński, japoński, koreański.

### **Elastyczny, przyjazny dla użytkownika interfejs**

- Eksploracja tekstu realizowana jest przez cztery osobne moduły odpowiadające najpopularniejszemu zadaniu. Można je łączyć w dowolny sposób w zależności od zadaniapotrzeb. Te moduły tekstowe współpracują bezpośrednio z szeregiem modułów SAS Enterprise Miner i mogą zostać rozbudowane przy pomocy algorytmów zdefiniowanych przez użytkownika lub określenie nowej reguły biznesowej dotyczącej modelowania predykcyjnego, grupowania, wizualizacji oraz raportowania i dzięki temu można ich używać jako kodu punktowego SAS.
- Wykresy przebiegu procesów analizy eksploracji danych można modyfikować, zapisywać i udostępniać innym użytkownikom.
- Elastyczne narzędzia raportowania umożliwiają publikację wyników w zwięzłym formacie HTML.
- Wykres Concept Link (powiązania koncepcji) przedstawia wizualne powiązania między zwrotami.

### **Moduł Tex Parsing (analizy składniowej tekstu) \*\*\***

- Domyślne lub zmodyfikowane listy filtrowania usuwają z analizy wyrażenia o niskiej lub zerowej zawartości informacji.
- Automatyczne sprawdzanie pisowni.
- Obcinanie końcówek fleksyjnych w celu określenia rdzeni słów.
- Oznaczanie części mowy na podstawie kontekstu.
- Wyodrębnienie grupy rzeczownika w celu zidentyfikowania koncepcji na poziomie wyrażenia, tj. „wywiad konkurencyjny.”
- Obsługa wielu różnych typów obiektów, w tym nazwisk i nazw firm, lokalizacji, dat, adresów, wymiarów, a także adresów e-mail i www. Obiekty te można zmodyfikować dla każdego obsługiwanego języka.
- Wielowyrzowe zwroty zdefiniowane przez użytkownika, tj. „przeciągnij i upuść”.
- Domyślne i zdefiniowane przez użytkownika listy synonimów.
- Kompleksowe funkcje obejmują dzielenie złożeń na poszczególne składowe.

### **Techniki redukcji wymiarów**

- Zwijanie wyrażen automatycznie identyfikuje n wyrażen o największej wadze w dokumencie.
- Rozkład według wartości osobliwych (SVD) przekształca każdy dokument w n-wymiarową przestrzeń, w której im bliżej siebie dokumenty są położone, tym bardziej są do siebie podobne.

### **Moduł Text Topic (temat tekstu) \*\*\***

- Przeglądarka taksonomii wyświetla automatycznie wygenerowane tematy domyślne oraz tematy stworzone przez użytkownika.
- Dokumenty można skategoryzować jako przynależące do zera, jednego lub większej liczby różnych tematów.
- Tematy można modyfikować w sposób interaktywny w łatwym w obsłudze i intuicyjnym środowisku wizualnym.

### **Algorytmy grupowania tekstu**

- Grupowanie oparte na algorytmie maksymalizacji wartości oczekiwanej łączy dokumenty w dyskretne, niepokrywające się grupy (tzw. grupowanie twarde) przy pomocy technik grupowania przestrzennego.
- Grupowanie hierarchiczne ułatwia automatyczne łączenie dokumentów w taksonomie.

## Wymagania techniczne

### Obsługiwane platformy

- AIX: wersja 5.3 i 6.1 na architekturze POWER
- HP-UX Itanium: HP-UX 11iv2 (11.23), 11iv3 (11.31)
- Linux dla x86 (x86-32): RHEL 4 i 5, SuSE SLES 9 i 10
- Microsoft Windows (x86-32): Windows XP Professional, Windows Vista\*, rodzina Windows Server 2003
- Microsoft Windows dla x64 (EM64T/AMD64): Windows XP Professional dla x64, Windows Vista\* dla x64, Windows Server 2003 dla x64
- Solaris na SPARC: wersja 9, 10
- Solaris na x64: wersja 10

\* UWAGA: obsługiwane edycje systemu Windows Vista to Enterprise, Business i Ultimate

### Obsługiwane przeglądarki internetowe

- Internet Explorer 6 w systemie Windows XP Pro
- Internet Explorer 7 w systemach Windows XP Pro i Windows Vista\*
- Firefox 2.0 w systemach Windows XP Pro, Windows Vista\* oraz Linux x86 (SuSE i RHEL)

### Wymagane/opcjonalne oprogramowanie warstwy środkowej (middle tier)

- Klient SAS i warstwa pośrednia wymagają środowiska Sun JRE 1.5i.

### Wymagane oprogramowanie

- Wymagane jest Oprogramowanie SAS Enterprise Miner, które jest wymagane i musi być zainstalowany na tym samym komputerze co SAS Text Miner; lub wymagane jest oprogramowanie SAS Enterprise Miner for Desktop jest wymagany, które i musi być zainstalowany na tym samym komputerze co SAS Text Miner for Desktop

## Główne cechy c.d.

- Tworzenie profili grup i tematów przez dodawanie danych strukturalnych z oryginalnych dokumentów w celu zwiększenia dokładności analizy (np. wiek, skłonność do zakupu itp.).

### Moduł Text Filter (filtrowanie tekstu) \*\*\*

- Zawiera zwięzły podgląd dokumentów i słownictwa lub wszystkich wyrażen wykrytych podczas analizy tekstu.
- Automatycznie sprawdza pisownię, mapując słowa napisane błędnie do właściwych wyrażen.
- Stosuje wyszukiwania typu Google lub klauzule SQL WHERE do analizy podzbiorów (np. przeprowadza osobne analizy gwarancji dla każdej marki lub modelu samochodu).
- Może w sposób programowy i interaktywny rozróżniać i usuwać nieistotne wyrażenia, mapować skróty i wyświetlać inne równoważne zwroty.

### Wielostronny podgląd danych

- Połącz dane tekstowe z tradycyjną eksploracją danych strukturalnych w celu zautomatyzowania, wizualizacji, klasyfikacji i wdrożenia wyników modelowania predykcyjnego.
- Płynnie połącz dane ilościowe i jakościowe z analizą tekstu w celu poprawy trafności prognoz.
- Zaawansowane techniki, takie jak sieci neuronowe, wnioskowanie oparte na śladach pamięciowych, modele regresyjne oraz drzewa decyzji można rozszerzać przy pomocy modułu SAS Enterprise Miner Code, który umożliwi większą innowacyjność i szybsze wdrożenie przy zmniejszeniu ryzyka.
- Ocenę sprawności poszczególnych modeli można wyświetlić na jednym ekranie w celu wybrania najodpowiedniejszego do wdrożenia kodu punktowego, który zostanie wybrany do kategoryzacji nowych dokumentów.
- Dane otrzymane z programu SAS Enterprise Content Categorization mogą zostać bezpośrednio zintegrowane z analizą eksploracji danych. Tematy i motywy wykryte przez SAS Text Miner będą cennym wkładem do SAS Enterprise Content Categorization, szczególnie w sytuacjach, w których nie istniały wcześniej żadne taksonomie. \*\*\*

\*\*\* Nowa funkcja w SAS Text Miner 4.2 (grudzień 2009)

| TERM          | FREQ | # DOCS | KEEP | WEIGHT | ROLE    | ATTRIBUTE |
|---------------|------|--------|------|--------|---------|-----------|
| sas institute | 2391 | 525    | [X]  | 0.034  | COMPANY | Alpha     |
| be            | 1755 | 498    | [X]  | 0.044  | Verb    | Alpha     |
| software      | 857  | 485    | [X]  | 0.042  | Noun    | Alpha     |
| use           | 694  | 354    | [X]  | 0.051  | Verb    | Alpha     |
| data          | 1201 | 340    | [X]  | 0.116  | Noun    | Alpha     |
| system        | 533  | 262    | [X]  | 0.151  | Noun    | Alpha     |

Zaawansowana składnia wyszukiwania modułu Interactive Filter Viewer wyszukuje dokumenty w oparciu o znajdujące się w nich słowa lub zwroty i oferuje elastyczność w dzieleniu analizy na podgrupy.

## SAS Institute Polska

ul. Gdańska 27/31  
01-633 Warszawa  
tel. +48 22 560 46 00 do 02  
fax. +48 22 560 46 04

www.sas.com/poland

