

# **Text Mining: metody, narzędzia i zastosowania**

## **Wykorzystanie SAS Text Analytics**

### **Spis treści**

#### **Wykaz oznaczeń**

#### **Wykaz skrótów**

#### **Wprowadzenie**

### **Część I. Wprowadzenie do eksploracji danych tekstowych**

#### **1. Trendy w rozwoju systemów informatycznych eksploracji danych**

#### **2. Metody eksploracji danych tekstowych**

- 2.1. Przebieg analizy dokumentu tekstowego i charakterystyka stosowanych metod
- 2.2. Określenie celu, zakresu i kosztów analizy
- 2.3. Przekształcenie zbioru dokumentów źródłowych
  - 2.3.1. Informacja o częstotliwości występowania poszczególnych terminów
  - 2.3.2. Postać ustrukturyzowana
- 2.4. Wybór metody obliczeniowej

#### **3. Architektura oprogramowania do eksploracji danych tekstowych na przykładzie pakietu SAS Text Analytics firmy SAS Institute**

- 3.1. Rozpoczęcie pracy z programem Enterprise Miner (Text Miner)
  - 3.1.1. Tworzenie nowego projektu i biblioteki
  - 3.1.2. Tworzenie diagramów analizy danych
  - 3.1.3. Określanie źródła danych projektu
- 3.2. Metodyka SEMMA
  - 3.2.1. Etap Próbkowanie
  - 3.2.2. Etap Eksploracja
  - 3.2.3. Etap Modyfikacja
  - 3.2.4. Etap Modelowanie
  - 3.2.5. Etap Ocena
- 3.3. Text Miner - etapy przetwarzania
- 3.4. Text Miner - komponenty
  - 3.4.1. Właściwości węzła Klastrowanie tekstu
  - 3.4.2. Właściwości węzła Filtrowanie tekstu
  - 3.4.3. Właściwości węzła Import tekstu
  - 3.4.4. Właściwości węzła Parsowanie tekstu
  - 3.4.5. Właściwości węzła Profil tekstu
  - 3.4.6. Właściwości węzła Generator reguł tekstu
  - 3.4.7. Właściwości węzła Temat tekstu
- 3.5. Przykład: Klasteryzacja zbioru zdań
  - 3.5.1. Konfiguracja diagramu przepływu danych
  - 3.5.2. Konfiguracja poszczególnych węzłów i interpretacja wyników
  - 3.5.3. Podsumowanie

## **Część II. Przetwarzanie informacji zawartej w dokumencie tekstowym**

### **4. Wybór funkcji wagującej macierzy częstości występowania terminów**

- 4.1. Wagi częstości
- 4.2. Wagi wyrażenia
- 4.3. Przykład obliczeniowy
- 4.4. Podsumowanie

### **5. Redukcja wymiarowości macierzy częstości występowania terminów**

- 5.1. Analiza semantyczna zmiennych ukrytych
  - 5.1.1. Rozkład SVD
  - 5.1.2. Przykład obliczeniowy rozkładu SVD
- 5.2. Podsumowanie

### **6. Wybór algorytmu klastrowania dokumentów tekstowych**

- 6.1. Określenie miary podobieństwa grupy dokumentów
- 6.2. Algorytmy klastrowania
- 6.3. Grupowanie za pomocą węzła Klastrowanie tekstów
  - 6.3.1. Węzeł Klastrowanie tekstu - algorytm Hierarchiczny
  - 6.3.2. Węzeł Klastrowanie tekstu - algorytm Maksymalizacja oczekiwań
  - 6.3.3. Węzeł Klastrowanie tekstu - właściwość Terminy opisowe
- 6.4. Grupowanie za pomocą węzła Temat tekstu
  - 6.4.1. Tematy definiowane przez użytkownika
- 6.5. Podsumowanie

### **7. Zarys metodyki tworzenia modeli predykcyjnych oraz porównywania zdolności predykcyjnych modeli**

- 7.1. Tworzenie modelu predykcyjnego
- 7.2. Ocena błędu klasyfikacji
  - 7.2.1. Krzywe ROC
  - 7.2.2. Wykresy wzrostu
- 7.3. Przykład: Użycie węzła Importowanie tekstu oraz porównywanie modeli predykcyjnych
  - 7.3.1. Konfiguracja diagramu przepływu danych oraz poszczególnych węzłów
- 7.4. Podsumowanie

### **8. Klastrowanie dokumentów nadzorowane przez użytkownika**

- 8.1. Charakterystyka węzła Generator reguł tekstu
- 8.2. Podsumowanie

## **Część III. Wydobywanie i organizacja wiedzy z dokumentów tekstowych w instytucji**

### **9. Zarys zagadnień związanych z wydobywaniem i organizacją wiedzy w instytucji**

- 9.1. Wprowadzenie
  - 9.1.1. SAS Crawler
  - 9.1.2. SAS Search and Indexing
  - 9.1.3. SAS Information Retrieval Studio
- 9.2. Podsumowanie

## **10. Klasyfikacja dokumentów**

### 10.1. SAS Content Categorization Studio

10.1.1. Metody klasyfikacji dokumentów dostępne w SAS CCS

10.1.2. Wydobywanie konceptów dostępne w SAS CCS

10.1.3. Wydobywanie kontekstu dostępne w SAS CCS

10.1.4. Zakładanie nowego projektu

10.1.5. Metodyka planowania projektu

10.1.6. Tworzenie nowej kategorii

10.1.7. Zasady używania kategoryzatora statystycznego

10.1.8. Zasady używania kategoryzatora generującego reguły automatycznie

10.1.9. Zasady używania kategoryzatora bazującego na regułach

10.1.10. Praca z konceptami

10.2. Przykład: Zastosowania klasyfikacji dokumentów w celu wspomaganie diagnostyki w departamencie radiodiagnostyki

10.3. Podsumowanie

## **11. Analiza sentymentu**

### 11.1. SAS Sentiment Analysis Studio

11.1.1. Metoda oceny sentymentu dla dokumentu

11.1.2. Zakładanie nowego projektu

11.1.3. Testowanie istniejących modeli

11.1.4. Tworzenie modeli hybrydowych

11.1.5. SAS Sentiment Analysis Server

11.2. Przykład analizy sentymentu użytkowników telefonów komórkowych

11.3. Podsumowanie

## **Część IV. Inne zagadnienia przetwarzania dokumentów tekstowych**

### **12. Inne elementy przetwarzania danych tekstowych**

12.1. Porównywanie dokumentów za pomocą metryk

12.1.1. Odległość kosinusowa

12.1.2. Metryka Jaccarda

12.2. Wydobywanie jednostek specjalnych z dokumentów

## **Słownik pojęć związanych z eksploracją danych tekstowych**

### **Dodatek A: Podstawy obsługi środowiska SAS i język 4GL**

A.1. Wprowadzenie do obsługi systemu SAS

A.1.1. Struktura zbioru danych SAS

A.1.2. Formaty i informaty

A.2. Język 4GL

A.2.1. Blok typu DATA STEP

A.2.2. Blok typu PROC STEP

### **Dodatek B: Podstawy języka makr**

B.1. Makrozmiennicze

B.2. Makroprogramy

## **Dodatek C: Wizualna interpretacja danych**

C.1. Przegląd typów wykresów stosowanych dla danych tekstowych

**Bibliografia**

**Indeks pojęć**

**Spis rysunków**

**Spis tabel**