
Spis treści

Wykaz oznaczeń	IX
Wykaz skrótów	XI
Wprowadzenie	1
Część I. Wprowadzenie do eksploracji danych tekstowych	5
1. Trendy w rozwoju systemów informatycznych eksploracji danych	7
2. Metody eksploracji danych tekstowych	11
2.1. Przebieg analizy dokumentu tekstowego i charakterystyka stosowanych metod	12
2.2. Określenie celu, zakresu i kosztów analizy	13
2.3. Przekształcenie zbioru dokumentów źródłowych	13
2.3.1. Informacja o częstości występowania poszczególnych terminów ...	13
2.3.2. Postać ustrukturyzowana	15
2.4. Wybór metody obliczeniowej	15
3. Architektura oprogramowania do eksploracji danych tekstowych na przykładzie pakietu SAS Text Analytics firmy SAS Institute	17
3.1. Rozpoczęcie pracy z programem Enterprise Miner (Text Miner)	19
3.1.1. Tworzenie nowego projektu i biblioteki	19
3.1.2. Tworzenie diagramów analizy danych	21
3.1.3. Określanie źródła danych projektu	22
3.2. Metodyka SEMMA	23
3.2.1. Etap Próbkowanie	24
3.2.2. Etap Eksploracja	24
3.2.3. Etap Modyfikacja	25
3.2.4. Etap Modelowanie	26
3.2.5. Etap Ocena	27

3.3.	Text Miner – etapy przetwarzania	28
3.4.	Text Miner – komponenty	30
3.4.1.	Właściwości węzła Klastrowanie tekstu	30
3.4.2.	Właściwości węzła Filtrowanie tekstu	31
3.4.3.	Właściwości węzła Import tekstu	32
3.4.4.	Właściwości węzła Parsowanie tekstu	33
3.4.5.	Właściwości węzła Profil tekstu	35
3.4.6.	Właściwości węzła Generator reguł tekstu	35
3.4.7.	Właściwości węzła Temat tekstu	36
3.5.	Przykład: Klasteryzacja zbioru zdań	37
3.5.1.	Konfiguracja diagramu przepływu danych	37
3.5.2.	Konfiguracja poszczególnych węzłów i interpretacja wyników ...	38
3.5.3.	Podsumowanie	48
Część II. Przetwarzanie informacji zawartej w dokumencie tekstowym		49
4.	Wybór funkcji wagującej macierzy częstości występowania terminów	51
4.1.	Wagi częstości	51
4.2.	Wagi wyrażenia	52
4.3.	Przykład obliczeniowy	53
4.4.	Podsumowanie	54
5.	Redukcja wymiarowości macierzy częstości występowania terminów	57
5.1.	Analiza semantyczna zmiennych ukrytych	57
5.1.1.	Rozkład SVD	58
5.1.2.	Przykład obliczeniowy rozkładu SVD	58
5.2.	Podsumowanie	62
6.	Wybór algorytmu klastrowania dokumentów tekstowych	63
6.1.	Określenie miary podobieństwa grupy dokumentów	63
6.2.	Algorytmy klastrowania	63
6.3.	Grupowanie za pomocą węzła Klastrowanie tekstów	66
6.3.1.	Węzeł Klastrowanie tekstu – algorytm Hierarchiczny	66
6.3.2.	Węzeł Klastrowanie tekstu – algorytm Maksymalizacja oczekiwania	66
6.3.3.	Węzeł Klastrowanie tekstu – właściwość Terminy opisowe	66
6.4.	Grupowanie za pomocą węzła Temat tekstu	69
6.4.1.	Tematy definiowane przez użytkownika	72
6.5.	Posumowanie	73

7. Zarys metodyki tworzenia modeli predykcyjnych oraz porównywania zdolności predykcyjnych modeli	75
7.1. Tworzenie modelu predykcyjnego	75
7.2. Ocena błędu klasyfikacji	76
7.2.1. Krzywe ROC	77
7.2.2. Wykresy wzrostu	77
7.3. Przykład: Użycie węzła Importowanie tekstu oraz porównywanie modeli predykcyjnych	78
7.3.1. Konfiguracja diagramu przepływu danych oraz poszczególnych węzłów	79
7.4. Podsumowanie	83
8. Klastrowanie dokumentów nadzorowane przez użytkownika	85
8.1. Charakterystyka węzła Generator reguł tekstu	85
8.2. Podsumowanie	88
Część III. Wydobywanie i organizacja wiedzy z dokumentów tekstowych w instytucji	89
9. Zarys zagadnień związanych z wydobywaniem i organizacją wiedzy w instytucji	91
9.1. Wprowadzenie	91
9.1.1. SAS Crawler	92
9.1.2. SAS Search and Indexing	93
9.1.3. SAS Information Retrieval Studio	94
9.2. Podsumowanie	95
10. Klasyfikacja dokumentów	97
10.1. SAS Content Categorization Studio	97
10.1.1. Metody klasyfikacji dokumentów dostępne w SAS CCS	99
10.1.2. Wydobywanie konceptów dostępne w SAS CCS	101
10.1.3. Wydobywanie kontekstu dostępne w SAS CCS	106
10.1.4. Zakładanie nowego projektu	108
10.1.5. Metodyka planowania projektu	110
10.1.6. Tworzenie nowej kategorii	113
10.1.7. Zasady używania kategoryzatora statystycznego	114
10.1.8. Zasady używania kategoryzatora generującego reguły automatycznie	117
10.1.9. Zasady używania kategoryzatora bazującego na regułach	121
10.1.10. Praca z konceptami	125
10.2. Przykład: Zastosowania klasyfikacji dokumentów w celu wspomaganie diagnostyki w departamencie radiodiagnostyki	135
10.3. Podsumowanie	142

11. Analiza sentymentu	143
11.1. SAS Sentiment Analysis Studio	144
11.1.1. Metoda oceny sentymentu dla dokumentu	145
11.1.2. Zakładanie nowego projektu	147
11.1.3. Testowanie istniejących modeli	157
11.1.4. Tworzenie modeli hybrydowych	158
11.1.5. SAS Sentiment Analysis Server	158
11.2. Przykład analizy sentymentu użytkowników telefonów komórkowych	158
11.3. Podsumowanie	164
Część IV. Inne zagadnienia przetwarzania dokumentów tekstowych	165
12. Inne elementy przetwarzania danych tekstowych	167
12.1. Porównywanie dokumentów za pomocą metryk	167
12.1.1. Odległość kosinusowa	167
12.1.2. Metryka Jaccarda	168
12.2. Wydobywanie jednostek specjalnych z dokumentów	171
Słownik pojęć związanych z eksploracją danych tekstowych	173
Dodatek A: Podstawy obsługi środowiska SAS i język 4GL	177
A.1. Wprowadzenie do obsługi systemu SAS	177
A.1.1. Struktura zbioru danych SAS	180
A.1.2. Formaty i informaty	182
A.2. Język 4GL	182
A.2.1. Blok typu DATA STEP	183
A.2.2. Blok typu PROC STEP	183
Dodatek B: Podstawy języka makr	187
B.1. Makrozmiennne	187
B.2. Makroprogramy	187
Dodatek C: Wizualna interpretacja danych	189
C.1. Przegląd typów wykresów stosowanych dla danych tekstowych	190
Bibliografia	193
Indeks pojęć	195
Spis rysunków	197
Spis tabel	203