

# Spis treści

<b>Przedmowa</b> . . . . .	XI
0.1. Definicja eksploracji zasobów sieciowych . . . . .	XI
0.2. Cykl książek o eksploracji danych . . . . .	XI
0.3. Jak zbudowana jest ta książka? . . . . .	XII
0.4. Dlaczego ta książka jest potrzebna? . . . . .	XII
0.5. Podejście typu „biała skrzynka” – zrozumienie podstawowych struktur algorytmów i modeli . . . . .	XIII
0.5.1. Omówienie działania algorytmów . . . . .	XIII
0.5.2. Zastosowanie algorytmów do dużego zbioru danych . . . . .	XIII
0.5.3. Ćwiczenia do rozdziałów – upewnij się, że rozumiesz . . . . .	XIII
0.5.4. Ćwiczenia praktyczne – nauka eksploracji danych przez wykonywanie eksploracji danych . . . . .	XIV
0.6. Eksploracja danych jako proces . . . . .	XIV
0.7. Oprogramowanie . . . . .	XIV
0.8. Strona internetowa: <a href="http://www.dataminingconsultant.com">www.dataminingconsultant.com</a> . . . . .	XV
0.9. <i>Eksploracja zasobów internetowych</i> jako podręcznik . . . . .	XV
0.10. Podziękowania . . . . .	XVI
<b>I Analiza struktury sieci WWW</b> . . . . .	<b>1</b>
<b>1. Wyszukiwanie informacji tekstowych i wyszukiwanie w Internecie</b> . . . . .	<b>3</b>
1.1. Wyzwania sieci . . . . .	3
1.1.1. Wyszukiwarki internetowe . . . . .	4
1.1.2. Katalogi stron WWW . . . . .	5
1.1.3. Semantic Web . . . . .	5
1.2. Ściąganie stron internetowych . . . . .	6
1.2.1. Podstawy WWW . . . . .	6
1.2.2. Roboty internetowe . . . . .	8
1.3. Indeksowanie i wyszukiwanie według słów kluczowych . . . . .	13
1.3.1. Reprezentacja dokumentów . . . . .	16
1.3.2. Rozważania na temat implementacji . . . . .	21
1.3.3. Ranking ważności . . . . .	22
1.3.4. Zaawansowane wyszukiwanie tekstów . . . . .	31
1.3.5. Używanie struktury HTML do wyszukiwania słów kluczowych . . . . .	32
1.4. Ocena jakości wyszukiwania . . . . .	35

1.5.	Wyszukiwanie według podobieństwa . . . . .	39
1.5.1.	Miara (odległość) kosinusowa . . . . .	39
1.5.2.	Współczynnik Jaccarda . . . . .	42
1.5.3.	Podobieństwo dokumentów . . . . .	45
1.6.	Literatura . . . . .	47
1.7.	Ćwiczenia . . . . .	47
<b>2.</b>	<b>Ranking oparty na strukturze połączeń . . . . .</b>	<b>51</b>
2.1.	Wprowadzenie . . . . .	51
2.2.	Analiza sieci społecznych . . . . .	52
2.3.	Algorytm PageRank . . . . .	54
2.4.	Autorytety i koncentratory . . . . .	57
2.5.	Wyszukiwanie oparte na podobieństwie strukturalnym . . . . .	60
2.6.	Zaawansowane techniki tworzenia rankingów stron . . . . .	60
2.7.	Literatura . . . . .	62
2.8.	Ćwiczenia . . . . .	62
<b>II</b>	<b>Analiza zawartości sieci WWW . . . . .</b>	<b>65</b>
<b>3.</b>	<b>Grupowanie . . . . .</b>	<b>67</b>
3.1.	Wprowadzenie . . . . .	67
3.1.1.	Aglomeracyjne grupowanie hierarchiczne . . . . .	69
3.2.	Algorytm $k$ -średnich . . . . .	75
3.3.	Grupowanie oparte na prawdopodobieństwie . . . . .	79
3.3.1.	Problem skończonej mieszaniny . . . . .	81
3.3.2.	Problem klasyfikacji . . . . .	82
3.3.3.	Problem grupowania . . . . .	85
3.4.	Techniki wspólnego filtrowania (systemy rekomendacyjne) . . . . .	91
3.5.	Literatura . . . . .	92
3.6.	Ćwiczenia . . . . .	93
<b>4.</b>	<b>Ocena grupowania . . . . .</b>	<b>96</b>
4.1.	Podejścia do oceny grupowania . . . . .	96
4.2.	Funkcje kryterialne oparte na podobieństwie . . . . .	97
4.3.	Probabilistyczne funkcje kryterialne . . . . .	101
4.4.	Model oparty na zasadzie minimalnej długości opisu i ocena cech . . . . .	105
4.4.1.	Zasada minimalnej długości opisu . . . . .	107
4.4.2.	Ocena modelu opartego na zasadzie minimalnej długości opisu . . . . .	108
4.4.3.	Wybór cech . . . . .	111
4.5.	Ocena za pomocą odwzorowania klas do grup . . . . .	112
4.6.	Dokładność, kompletność i miara $F$ . . . . .	114
4.7.	Entropia . . . . .	117
4.8.	Literatura . . . . .	119
4.9.	Ćwiczenia . . . . .	119
<b>5.</b>	<b>Klasyfikacja . . . . .</b>	<b>122</b>
5.1.	Ogólne otoczenie i techniki oceny . . . . .	122
5.2.	Algorytm najbliższego sąsiada . . . . .	125
5.3.	Wybór cech . . . . .	128

5.4. Naiwny algorytm Bayesa . . . . .	132
5.5. Podejścia numeryczne . . . . .	139
5.6. Relacyjne uczenie się – <i>relational learning</i> . . . . .	141
5.7. Literatura . . . . .	146
5.8. Ćwiczenia . . . . .	146

### III Analiza użytkowania sieci WWW 149

<b>6. Wprowadzenie do analizy użytkowania sieci WWW . . . . .</b>	<b>151</b>
6.1. Definicja analizy użytkowania sieci WWW . . . . .	151
6.2. Metodologia <i>Cross-Industry Standard Process for Data Mining</i> . . . . .	152
6.3. Analiza kliknięć . . . . .	154
6.4. Pliki log serwera . . . . .	155
6.4.1. Pole adresu IP hosta . . . . .	156
6.4.2. Pole Data/Czas . . . . .	157
6.4.3. Pole żądania HTTP . . . . .	157
6.4.4. Pole kodu odpowiedzi HTTP . . . . .	157
6.4.5. Pole wielkości transferu (bajty) . . . . .	158
6.5. Format CLF . . . . .	158
6.5.1. Pole nazwy użytkownika . . . . .	159
6.5.2. Pole authuser . . . . .	159
6.6. Format ECLF . . . . .	159
6.6.1. Pole adresu strony odsyłającej . . . . .	159
6.6.2. Pole przeglądarki klienta . . . . .	159
6.6.3. Przykład rekordu pliku log . . . . .	160
6.7. Format Microsoft IIS . . . . .	161
6.8. Dodatkowe informacje . . . . .	162
6.9. Literatura . . . . .	162
6.10. Ćwiczenia . . . . .	162
<b>7. Wstępne przetwarzanie danych do analizy użytkowania sieci WWW . . . . .</b>	<b>164</b>
7.1. Konieczność wstępnego przetwarzania danych . . . . .	164
7.2. Czyszczenie i filtrowanie danych . . . . .	165
7.2.1. Badanie rozszerzeń stron i filtrowanie . . . . .	168
7.3. Usuwanie z pliku log wpisów robotów internetowych . . . . .	170
7.4. Identyfikacja użytkownika . . . . .	172
7.5. Identyfikacja sesji . . . . .	175
7.6. Uzupełnianie ścieżek . . . . .	177
7.7. Katalogi i przypisanie kategorii . . . . .	178
7.8. Dalsze kroki wstępnego przetwarzania danych . . . . .	181
7.9. Literatura . . . . .	181
7.10. Ćwiczenia . . . . .	182
<b>8. Eksploracyjna analiza użytkowania sieci WWW . . . . .</b>	<b>184</b>
8.1. Wprowadzenie . . . . .	184
8.2. Liczba żądań w sesji . . . . .	184
8.3. Długość sesji . . . . .	186
8.3.1. Procedura obliczania długości sesji . . . . .	186
8.4. Zależność między długością sesji a liczbą żądań użytkownika . . . . .	188

8.5. Średni czas na stronę . . . . .	190
8.6. Czas dla pojedynczych stron . . . . .	192
8.7. Literatura . . . . .	195
8.8. Ćwiczenia . . . . .	195
<b>9. Modelowanie użytkowania sieci WWW: grupowanie, reguły asocjacyjne i klasyfikacja . . . .</b>	<b>198</b>
9.1. Wprowadzenie . . . . .	198
9.2. Metodologia modelowania . . . . .	199
9.3. Definicja grupowania . . . . .	200
9.4. Algorytm grupowania BIRCH . . . . .	201
9.5. Analiza podobieństw i algorytm Apriori . . . . .	205
9.6. Dyskretyzacja zmiennych numerycznych . . . . .	206
9.7. Zastosowanie algorytmu Apriori do danych pliku log serwera CCSU . . . . .	208
9.8. Drzewa klasyfikacyjne i regresyjne . . . . .	211
9.9. Algorytm C4.5 . . . . .	215
9.10. Literatura . . . . .	218
9.11. Ćwiczenia . . . . .	218
<b>Indeks . . . . .</b>	<b>221</b>