
Spis treści

Przedmowa	9
Słowo wstępne	11
1. Analiza wielkich zbiorów danych	13
Wyzwania w nauce o danych	15
Przedstawiamy Apache Spark	16
O czym jest ta książka	18
2. Wprowadzenie do analizy danych za pomocą Scala i Spark	21
Scala dla badaczy danych	22
Model programowania w Spark	23
Wiązanie rekordów danych	23
Pierwsze kroki — powłoka Spark i kontekst SparkContext	24
Przesyłanie danych z klastra do klienta	29
Wysyłanie kodu z klienta do klastra	32
Tworzenie list danych i klas wyboru	33
Agregowanie danych	36
Tworzenie histogramów	38
Statystyki sumaryzacyjne ciągłych wartości	39
Tworzenie współdzielonego kodu wyliczającego statystyki sumaryczne	40
Prosty wybór zmiennych i ocena zgodności rekordów	44
Następny krok	45
3. Rekomendowanie muzyki i dane Audioscrobbler	47
Zbiór danych	48
Algorytm rekomendacyjny wykorzystujący metodę naprzemiennych najmniejszych kwadratów	49
Przygotowanie danych	51

Utworzenie pierwszego modelu	54
Wrywkowe sprawdzanie rekomendacji	56
Ocena jakości rekomendacji	57
Obliczenie metryki AUC	59
Dobór wartości hiperparametrów	60
Przygotowanie rekomendacji	62
Dalsze kroki	63
4. Prognozowanie zalesienia za pomocą drzewa decyzyjnego	65
Szybkie przejście do regresji	65
Wektory i cechy	66
Przykłady treningowe	67
Drzewa i lasy decyzyjne	68
Dane Covtype	70
Przygotowanie danych	71
Pierwsze drzewo decyzyjne	72
Hiperparametry drzewa decyzyjnego	76
Regulacja drzewa decyzyjnego	77
Weryfikacja cech kategoryalnych	79
Losowy las decyzyjny	81
Prognozowanie	83
Dalsze kroki	83
5. Wykrywanie anomalii w ruchu sieciowym metodą grupowania według k-średnich	85
Wykrywanie anomalii	86
Grupowanie według k-średnich	86
Włamania sieciowe	87
Dane KDD Cup 1999	87
Pierwsza próba grupowania	88
Dobór wartości k	90
Wizualizacja w środowisku R	93
Normalizacja cech	94
Zmienne kategoryalne	96
Wykorzystanie etykiet i wskaźnika entropii	97
Grupowanie w akcji	98
Dalsze kroki	100

6. Wikipedia i ukryta analiza semantyczna	101
Macierz słowo – dokument	102
Pobranie danych	104
Analiza składni i przygotowanie danych	104
Lematyzacja	105
Wyliczenie metryk TF-IDF	106
Rozkład według wartości osobliwych	108
Wyszukiwanie ważnych pojęć	110
Wyszukiwanie i ocenianie informacji za pomocą niskowymiarowej reprezentacji danych	113
Związek dwóch słów	114
Związek dwóch dokumentów	115
Związek słowa i dokumentu	116
Wyszukiwanie wielu słów	117
Dalsze kroki	118
7. Analiza sieci współwystępowania za pomocą biblioteki GraphX	121
Katalog cytowań bazy MEDLINE — analiza sieci	122
Pobranie danych	123
Analiza dokumentów XML za pomocą biblioteki Scala	125
Analiza głównych znaczników i ich współwystępowania	126
Konstruowanie sieci współwystępowania za pomocą biblioteki GraphX	128
Struktura sieci	131
Połączone komponenty	131
Rozkład stopni wierzchołków	133
Filtrowanie krawędzi zakłócających dane	135
Przetwarzanie struktury EdgeTriplet	136
Analiza przefiltrowanego grafu	138
Sieci typu „mały świat”	139
Kliki i współczynniki klastrowania	139
Obliczenie średniej długości ścieżki za pomocą systemu Pregel	141
Dalsze kroki	145
8. Geoprzestrzenna i temporalna analiza tras nowojorskich taksówek	147
Pobranie danych	148
Przetwarzanie danych temporalnych i geoprzestrzennych w systemie Spark	148
Przetwarzanie danych temporalnych za pomocą bibliotek JodaTime i NScalaTime	149

Przetwarzanie danych geoprzestrzennych za pomocą Esri Geometry API i Spray	150
Użycie interfejsu API Esri Geometry	151
Wprowadzenie do formatu GeoJSON	152
Przygotowanie danych dotyczących kursów taksówek	154
Obsługa dużej liczby błędnych rekordów danych	155
Analiza danych geoprzestrzennych	158
Sesjonowanie w systemie Spark	161
Budowanie sesji — dodatkowe sortowanie danych w systemie Spark	162
Dalsze kroki	165
9. Szacowanie ryzyka finansowego metodą symulacji Monte Carlo	167
Terminologia	168
Metody obliczania wskaźnika VaR	169
Wariancja-kowariancja	169
Symulacja historyczna	169
Symulacja Monte Carlo	169
Nasz model	170
Pobranie danych	171
Wstępne przetworzenie danych	171
Określenie wag czynników	174
Losowanie prób	176
Wielowymiarowy rozkład normalny	178
Wykonanie testów	179
Wizualizacja rozkładu zwrotów	181
Ocena wyników	182
Dalsze kroki	184
10. Analiza danych genomicznych i projekt BDG	187
Rozdzielenie sposobów zapisu i modelowania danych	188
Przetwarzanie danych genomicznych za pomocą wiersza poleceń systemu ADAM	190
Format Parquet i format kolumnowy	195
Prognozowanie miejsc wiązania czynnika transkrypcyjnego na podstawie danych ENCODE	197
Odczytywanie informacji o genotypach z danych 1000 Genomes	203
Dalsze kroki	204

11. Analiza danych neuroobrazowych za pomocą pakietów PySpark i Thunder	205
Ogólne informacje o pakiecie PySpark	206
Budowa pakietu PySpark	207
Ogólne informacje i instalacja biblioteki pakietu Thunder	209
Ładowanie danych za pomocą pakietu Thunder	210
Podstawowe typy danych w pakiecie Thunder	214
Klasyfikowanie neuronów za pomocą pakietu Thunder	216
Dalsze kroki	221
A Więcej o systemie Spark	223
Serializacja	224
Akumulatory	225
System Spark i metody pracy badacza danych	226
Formaty plików	228
Podprojekty Spark	229
MLlib	229
Spark Streaming	230
Spark SQL	230
GraphX	230
B Nowy interfejs MLLib Pipelines API	231
Samo modelowanie to za mało	231
Interfejs API Pipelines	232
Przykład procesu klasyfikacji tekstu	233
Skorowidz	236