
Spis treści

Słowo wstępne	7
1. Analiza wielkich zbiorów danych	11
Praca z wielkimi zbiorami danych	12
Przedstawiamy Apache Spark i PySpark	14
Komponenty	14
PySpark	15
Ekosystem	16
Spark 3.0	17
PySpark i wyzwania w nauce o danych	17
O czym jest ta książka	18
2. Wprowadzenie do analizy danych za pomocą PySpark	19
Architektura systemu Spark	21
Instalacja interfejsu PySpark	22
Przygotowanie danych	24
Analiza danych za pomocą struktury DataFrame	28
Szybkie statystyki zbiorcze w strukturze DataFrame	32
Przetworzenie i przekształcenie struktury DataFrame	34
Złączenie struktur DataFrame i wybór cech	36
Ocena modelu	37
Dalsze kroki	39
3. Rekomendowanie muzyki i dane Audioscrobbler	40
Zbiór danych	41
Wymagania dla systemu rekomendacyjnego	43
Algorytm naprzemiennych najmniejszych kwadratów	45
Przygotowanie danych	46
Utworzenie pierwszego modelu	48
Wrywkowe sprawdzanie rekomendacji	52

Ocena jakości rekomendacji	53
Obliczenie wskaźnika AUC	55
Dobór wartości hiperparametrów	56
Przygotowanie rekomendacji	58
Dalsze kroki	60
4. Prognozowanie zalesienia za pomocą drzewa decyzyjnego	61
Drzewa i lasy decyzyjne	62
Przygotowanie danych	64
Pierwsze drzewo decyzyjne	68
Hiperparametry drzewa decyzyjnego	74
Regulacja drzewa decyzyjnego	75
Weryfikacja cech kategoryalnych	79
Losowy las decyzyjny	81
Prognozowanie	84
Dalsze kroki	84
5. Wykrywanie anomalii w ruchu sieciowym metodą grupowania według k-średnich	86
Grupowanie według k-średnich	87
Wykrywanie anomalii w ruchu sieciowym	88
Dane KDD Cup 1999	88
Pierwsza próba grupowania	89
Dobór wartości k	92
Wizualizacja w środowisku R	94
Normalizacja cech	97
Zmienne kategoryjne	99
Wykorzystanie etykiet i wskaźnika entropii	101
Grupowanie w akcji	102
Dalsze kroki	104
6. Wikipedia, algorytmy LDA i Spark NLP	105
Algorytm LDA	106
Algorytm LDA w interfejsie PySpark	106
Pobranie danych	107
Spark NLP	108
Przygotowanie środowiska	109
Przekształcenie danych	109
Przygotowanie danych za pomocą biblioteki Spark NLP	111
Metoda TF-IDF	114

Wyliczenie wskaźników TF-IDF	115
Utworzenie modelu LDA	116
Dalsze kroki	118
7. Geoprzestrzenna i temporalna analiza tras nowojorskich taksówek	119
Przygotowanie danych	120
Konwersja ciągów znaków na znaczniki czasu	122
Obsługa błędnych rekordów danych	124
Analiza danych geoprzestrzennych	125
Wprowadzenie do formatu GeoJSON	125
Biblioteka GeoPandas	126
Sesjonowanie w interfejsie PySpark	129
Budowanie sesji — dodatkowe sortowanie danych w systemie Spark	130
Dalsze kroki	132
8. Szacowanie ryzyka finansowego	133
Terminologia	134
Metody obliczania wskaźnika VaR	134
Wariancja-kowariancja	135
Symulacja historyczna	135
Symulacja Monte Carlo	135
Nasz model	136
Pobranie danych	137
Przygotowanie danych	137
Określenie wag czynników	140
Losowanie prób	142
Wielowymiarowy rozkład normalny	144
Wykonanie testów	145
Wizualizacja rozkładu zwrotów	148
Dalsze kroki	148
9. Analiza danych genomicznych i projekt BDG	150
Rozdzielenie sposobów zapisu i modelowania danych	151
Przygotowanie pakietu ADAM	153
Przetwarzanie danych genomicznych za pomocą pakietu ADAM	154
Konwersja formatów plików za pomocą poleceń pakietu ADAM	155
Pozyskanie danych genomicznych przy użyciu interfejsu PySpark i pakietu ADAM	155
Prognozowanie miejsc wiązania czynnika transkrypcyjnego na podstawie danych	
ENCODE	160
Dalsze kroki	164

10. Określanie podobieństwa obrazów	
za pomocą głębokiego uczenia i algorytmu PySpark LSH	166
PyTorch	167
Instalacja	167
Przygotowanie danych	168
Skalowanie obrazów za pomocą PyTorch	168
Wektoryzacja obrazów za pomocą modelu głębokiego uczenia	169
Osadzenie obrazów	169
Import osadzeń obrazów do pakietu PySpark	171
Określanie podobieństwa obrazów za pomocą algorytmu PySpark LSH	172
Wyszukiwanie najbliższych sąsiadów	173
Dalsze kroki	176
11. Zarządzanie cyklem uczenia maszynowego za pomocą platformy MLflow	177
Cykl uczenia maszynowego	177
Platforma MLflow	179
Śledzenie eksperymentów	180
Zarządzanie modelami uczenia maszynowego i udostępnianie ich	182
Tworzenie i stosowanie projektów za pomocą modułu MLflow Projects	185
Dalsze kroki	187