

Spis treści

Wstęp	13
Część I. Podstawy analizy danych	
Rozdział 1. Wprowadzenie do bibliotek Pythona	21
Wyjaśnienie pojęcia „analiza danych”	22
Standardowy proces analizy danych	23
Proces KDD	24
Proces SEMMA	25
Proces CRISP-DM	25
Analiza danych a danetyka	27
Role analityka danych i danetyka	27
Umiejętności analityka danych oraz danetyka	28
Instalacja środowiska Python 3	30
Instalacja i konfigurowanie Pythona w systemie Windows	30
Instalacja i konfigurowanie Pythona w Linuksie	31
Instalacja i konfigurowanie Pythona w systemie macOS	31
za pomocą instalatora graficznego	
Instalacja i konfigurowanie Pythona w systemie macOS	
za pomocą menedżera Homebrew	31
Oprogramowanie używane w tej książce	32
Używanie aplikacji IPython jako powłoki	33
Odczytywanie stron podręcznika	35
Źródła informacji na temat bibliotek analizy danych w Pythonie	35
Korzystanie z aplikacji JupyterLab	36
Stosowanie aplikacji Jupyter Notebook	37

Zaawansowane funkcje aplikacji Jupyter Notebook	38
Skróty klawiszowe	38
Instalowanie innych jąder	38
Realizowanie poleceń powłoki	39
Rozszerzenia	40
Podsumowanie	44
Rozdział 2. Biblioteki NumPy i pandas	45
Wymogi techniczne	46
Tablice NumPy	46
Własności tablic	48
Wybieranie elementów tablicy	49
Numeryczne typy danych tablic NumPy	50
Obiekty dtype	52
Kody znakowe typów danych	52
Konstruktory dtype	53
Atrybuty dtype	53
Manipulowanie wymiarami tablic	54
Łączenie tablic NumPy	55
Rozdzielanie tablic NumPy	58
Zmiana typu danych tablic NumPy	60
Tworzenie widoków i kopii NumPy	61
Fragmentowanie tablic NumPy	62
Indeksowanie logiczne i indeksowanie specjalne	64
Rozgłaszanie tablic	65
Tworzenie obiektów DataFrame biblioteki pandas	67
Obiekt Series biblioteki pandas	68
Odczytywanie i kwerendowanie danych Quandl	71
Opisywanie obiektów DataFrame	74
Grupowanie i łączenie obiektów DataFrame	76
Praca z brakującymi danymi	80
Tworzenie tabel przestawnych	81
Rozwiązywanie kwestii dat	82
Podsumowanie	84
Bibliografia	85
Rozdział 3. Statystyka	86
Wymogi techniczne	87
Atrybuty i ich typy	87
Typy atrybutów	87
Atrybuty dyskretne i ciągłe	88
Pomiar tendencji centralnej	89
Średnia arytmetyczna	89
Dominanta	90
Mediana	90
Pomiar dyspersji	90
Skośność i kurtosis	93

Określanie związków za pomocą współczynników kowariancji i korelacji	94
Współczynnik korelacji Pearsona	94
Współczynnik korelacji rang Spearmana	95
Współczynnik korelacji rang Kendalla	95
Centralne twierdzenie graniczne	96
Pozyskiwanie prób	96
Przeprowadzanie testów parametrycznych	98
Przeprowadzanie testów nieparametrycznych	102
Podsumowanie	107
Rozdział 4. Algebra liniowa	108
Wymogi techniczne	109
Dopasowywanie do wielomianów za pomocą biblioteki NumPy	109
Wyznacznik macierzy	111
Określanie rzędu macierzy	111
Macierz odwrotna w bibliotece NumPy	112
Rozwiązywanie równań liniowych za pomocą biblioteki NumPy	113
Rozkład macierzy za pomocą SVD	114
Wartości własne i wektory własne w bibliotece NumPy	115
Generowanie liczb losowych	116
Rozkład dwumianowy	117
Rozkład normalny	118
Testowanie normalności rozkładu danych za pomocą biblioteki SciPy	119
Tworzenie tablicy maskowanej za pomocą podpakietu numpy.ma	122
Podsumowanie	124
Część II.	
Eksploracyjna analiza danych i oczyszczanie danych	
Rozdział 5. Wizualizacja danych	127
Wymogi techniczne	127
Wizualizacja za pomocą pakietu Matplotlib	128
Akcesoria wykresu	129
Wykres punktowy	131
Wykres liniowy	132
Wykres kołowy	133
Wykres kolumnowy	134
Histogram	135
Wykres bąbelkowy	136
Tworzenie wykresów za pomocą biblioteki pandas	137
Zaawansowana wizualizacja za pomocą pakietu seaborn	139
Wykresy lm	140
Wykresy kolumnowe	142
Wykresy rozkładu	143
Wykresy pudełkowe	143
Wykresy KDE	144
Wykresy skrzypcowe	145

Wykresy zliczeń	146
Wykresy łączone	147
Mapy cieplne	148
Wykresy macierzowe	150
Wizualizacja interaktywna za pomocą biblioteki Bokeh	151
Tworzenie prostego wykresu	151
Glify	153
Szablony	154
Wykresy wielokrotne	157
Oddziaływania	159
Adnotacje	162
Najeżdżanie kursorem	163
Widżety	165
Podsumowanie	168
Rozdział 6. Pozyskiwanie, przetwarzanie i przechowywanie danych	169
Wymogi techniczne	170
Odczyt i zapis plików CSV za pomocą biblioteki NumPy	171
Odczyt i zapis plików CSV za pomocą biblioteki pandas	172
Odczyt i zapis plików arkusza kalkulacyjnego Excel	173
Odczyt i zapis plików JSON	174
Odczyt i zapis plików HDF5	175
Odczyt i zapis danych z tabel HTML-a	176
Odczyt i zapis plików Parquet	177
Odczyt i zapis danych z obiektu pickle	178
Łatwy dostęp do danych za pomocą modułu sqlalchemy	178
Odczyt i zapis danych w bazie danych MySQL	180
Wstawianie całego obiektu DataFrame do bazy danych	182
Odczyt i zapis danych w bazie danych MongoDB	183
Odczyt i zapis danych w bazie danych Cassandra	184
Odczyt i zapis danych w bazie danych Redis	185
PonyORM	186
Podsumowanie	187
Rozdział 7. Oczyszczanie nieuporządkowanych danych	188
Wymogi techniczne	189
Eksploracja danych	189
Filtrowanie danych w celu pozbycia się szumu	192
Filtrowanie po kolumnach	193
Filtrowanie po rzędach	195
Rozwiązywanie kwestii brakujących wartości	197
Usuwanie brakujących wartości	197
Uzupełnianie brakujących wartości	198
Rozwiązywanie kwestii elementów odstających	200
Techniki kodowania cech	202
Kodowanie „gorącojedynkowe”	202
Kodowanie etykietowe	204
Koder zmiennych porządkowych	205

Skalowanie cech	206
Metody skalowania cech	206
Przekształcanie cech	208
Rozdzielanie cech	210
Podsumowanie	211
Rozdział 8. Przetwarzanie sygnałów i szeregi czasowe	212
Wymogi techniczne	213
Moduł statsmodels	213
Średnie kroczące	213
Funkcje okna czasowego	216
Kointegracja	217
Rozkład STL	220
Autokorelacja	221
Modele autoregresyjne	223
Model ARMA	225
Generowanie sygnałów okresowych	227
Analiza Fouriera	230
Filtrowanie metodą analizy widmowej	231
Podsumowanie	233

Część III. Dokładna analiza uczenia maszynowego

Rozdział 9. Uczenie nadzorowane: analiza regresyjna	237
Wymogi techniczne	238
Regresja liniowa	238
Wieloraka regresja liniowa	239
Wielowspółliniowość	240
Usuwanie wielowspółliniowości	240
Zmienne fikcyjne	242
Projektowanie modelu regresji liniowej	243
Ocenianie skuteczności modelu regresyjnego	245
Współczynnik determinacji R-kwadrat	245
MSE	246
MAE	246
RMSE	246
Dopasowywanie regresji wielomianowej	247
Modele regresji używane w klasyfikacji	249
Regresja logistyczna	250
Charakterystyka modelu regresji logistycznej	251
Rodzaje algorytmów regresji logistycznej	252
Mocne i słabe strony regresji logistycznej	252
Implementacja regresji logistycznej za pomocą biblioteki scikit-learn	252
Podsumowanie	254

Rozdział 10. Uczenie nadzorowane: techniki klasyfikacji	255
Wymogi techniczne	256
Klasyfikacja	256
Naiwny klasyfikator Bayesa	258
Drzewa decyzyjne	261
Algorytm KNN	264
Maszyny wektorów nośnych	266
Terminologia	266
Podział danych na zestawy uczący i testowy	268
Wydzielanie	269
k-krotny sprawdzian krzyżowy	269
Metoda samowsporna	269
Ocena skuteczności modelu klasyfikacji	270
Macierz pomyłek	270
Dokładność	273
Precyzja	273
Czułość	273
Wskaźnik F1	273
Krzywa ROC i obszar AUC	274
Podsumowanie	276
Rozdział 11. Uczenie nienadzorowane: PCA i analiza skupień	277
Wymogi techniczne	278
Uczenie nienadzorowane	278
Redukowanie wymiarowości danych	279
Analiza głównych składowych	280
Przeprowadzanie PCA	280
Analiza skupień	283
Wyznaczanie liczby skupień	284
Grupowanie danych za pomocą algorytmu centroidów	288
Hierarchiczna analiza skupień	290
Algorytm DBSCAN	294
Widmowa analiza skupień	296
Ocenianie jakości analizy skupień	298
Wewnętrzna ocena jakości	298
Zewnętrzna ocena jakości	299
Podsumowanie	303
Część IV. Przetwarzanie języka naturalnego, analiza obrazów i obliczenia równoległe	
Rozdział 12. Analiza danych tekstowych	307
Wymogi techniczne	308
Instalacja bibliotek NLTK i spaCy	308
Normalizacja tekstu	309
Tokenizacja	310

Usuwanie słów nieinformatywnych	314
Rdzeniowanie słów i lematyzacja	315
Oznaczanie części mowy	317
Rozpoznawanie jednostek nazewniczych	318
Analiza zależności	319
Tworzenie chmury słów	320
„Worek słów”	321
Metoda TF-IDF	322
Analiza sentymentów za pomocą klasyfikacji tekstu	323
Klasyfikacja za pomocą „worka słów”	324
Klasyfikacja za pomocą metody TF-IDF	328
Podobieństwo tekstów	330
Indeks Jaccarda	331
Podobieństwo cosinusowe	332
Podsumowanie	333
Rozdział 13. Analiza obrazów	334
Wymogi techniczne	335
Instalacja biblioteki OpenCV	335
Omówienie danych obrazowych	336
Obrazy binarne	336
Obrazy w odcieniach szarości	337
Obrazy kolorowe	337
Modele barw	338
Rysowanie na obrazach	341
Pisanie na obrazach	345
Zmiana rozmiaru obrazu	346
Przekształcenie izometryczne obrazów	348
Zmiana jasności	350
Rozmywanie obrazu	352
Wykrywanie twarzy	355
Podsumowanie	358
Rozdział 14. Obliczenia równoległe za pomocą biblioteki Dask	359
Obliczenia równoległe za pomocą biblioteki Dask	360
Typy danych Dask	361
Tablice Dask	362
Ramki danych Dask	363
Worki Dask	368
Interfejs Dask Delayed	371
Skalowane wstępne przetwarzanie danych	373
Skalowanie cech w bibliotece Dask	373
Kodowanie cech w bibliotece Dask	374
Skalowane uczenie maszynowe	376
Obliczenia równoległe za pomocą biblioteki scikit-learn	377
Reimplementacja algorytmów uczenia maszynowego na potrzeby biblioteki Dask	378
Podsumowanie	382