

Spis treści

Przedmowa 11

CZĘŚĆ I. PRZEWODNIK PO SIECIACH SPOŁECZNOŚCIOWYCH

Wstęp 25

1. Eksploracja Twittera: odkrywanie trendów, dowiadywanie się, o czym się rozmawia, i trochę więcej 27

1.1. Przegląd 27

1.2. Dlaczego Twitter to jest "to"? 28

1.3. Odkrywanie API Twittera 30

1.3.1. Podstawowa terminologia związana z Twitterem 30

1.3.2. Tworzenie połączenia z API Twittera 33

1.3.3. Odkrywanie trendów 36

1.3.4. Wyszukiwanie tweetów 40

1.4. Analiza 140 (lub więcej) znaków 46

1.4.1. Wyodrębnianie podmiotów z tweetów 47

1.4.2. Analizowanie tweetów i występujących w nich podmiotów z wykorzystaniem analizy częstości 49

1.4.3. Obliczanie różnorodności leksykalnej tweetów 51

1.4.4. Badanie wzorców w retweetach 53

1.4.5. Wizualizacja danych częstości za pomocą histogramów 55

1.5. Uwagi końcowe 59

1.6. Zalecane ćwiczenia 60

1.7. Zasoby online 61

2. Eksploracja Facebooka: analizowanie fanpage'y, znajomości i więcej 63

2.1. Przegląd 64

2.2. Interfejs API Graph Facebooka 64

2.2.1. Wprowadzenie do API Graph 66

2.2.2. Protokół Open Graph 70

2.3. Analiza połączeń grafu społecznościowego 75

2.3.1. Analizowanie stron Facebooka 78

2.3.2. Manipulowanie danymi z wykorzystaniem pakietu pandas 88

2.4. Uwagi końcowe 95

2.5. Zalecane ćwiczenia 96

2.6. Zasoby online 96

3. Eksploracja Instagrama: komputerowy wzrok, sieci neuronowe, rozpoznawanie obiektów i wykrywanie twarzy 99

3.1. Przegląd 100

3.2. Poznawanie API Instagrama 101

3.2.1. Tworzenie żądań do API Instagrama 101

3.2.2. Odczytywanie własnego kanału na Instagramie 103

3.2.3. Pobieranie medium według hashtagu 105

3.3. Anatomia posta na Instagramie 105

3.4. Szybki kurs na temat sztucznych sieci neuronowych 108

3.4.1. Trening sieci neuronowej pod kątem "oglądania" zdjęć 109

3.4.2. Rozpoznawanie cyfr pisanych odręcznie 111

3.4.3. Rozpoznawanie obiektów na zdjęciach przy użyciu wstępnie przeszkolonych sieci neuronowych 116

3.5. Wykorzystanie sieci neuronowych do postów na Instagramie 119

3.5.1. Oznaczanie zawartości obrazu 119

3.5.2. Wykrywanie twarzy na zdjęciach	121
3.6. Uwagi końcowe	122
3.7. Zalecane ćwiczenia	123
3.8. Zasoby online	124
4. Eksploracja sieci LinkedIn: stanowiska, współpracownicy i nie tylko	127
4.1. Przegląd	128
4.2. Poznawanie API LinkedIna	128
4.2.1. Tworzenie żądań do API LinkedIn	129
4.2.2. Pobieranie połączeń LinkedIn w pliku CSV	132
4.3. Krótki kurs grupowania danych	132
4.3.1. Normalizacja danych w celu umożliwienia analizy	135
4.3.2. Mierzenie podobieństwa	145
4.3.3. Algorytmy klasteryzacji	147
4.4. Uwagi końcowe	161
4.5. Zalecane ćwiczenia	161
4.6. Zasoby online	162
5. Eksploracja danych z plików tekstowych: obliczanie podobieństwa dokumentów, wyodrębnianie kolokacji i inne	163
5.1. Przegląd	164
5.2. Pliki tekstowe	164
5.3. Wprowadzenie do TF-IDF	166
5.3.1. Częstość terminu	166
5.3.2. Odwrotna częstość dokumentu	168
5.3.3. TF-IDF	169
5.4. Odpytywanie danych w języku naturalnym za pomocą TF-IDF	172
5.4.1. Natural Language Toolkit - wprowadzenie	172

5.4.2. Zastosowanie współczynnika TF-IDF do języka naturalnego	176
5.4.3. Wyszukiwanie podobnych dokumentów	177
5.4.4. Analiza bigramów w języku naturalnym	184
5.4.5. Refleksje na temat analizy danych języka naturalnego	193
5.5. Uwagi końcowe	194
5.6. Zalecane ćwiczenia	195
5.7. Zasoby online	195
6. Eksploracja stron internetowych: przetwarzanie języka naturalnego w celu zrozumienia języka ludzkiego, tworzenie podsumowań postów na blogu i inne	197
6.1. Przegląd	198
6.2. Scraping, parsowanie i crawling stron internetowych	199
6.2.1. Przeszukiwanie wszcz w crawlingu stron internetowych	202
6.3. Odkrywanie semantyki przez dekodowanie składni	205
6.3.1. Przetwarzanie języka naturalnego krok po kroku	207
6.3.2. Wykrywanie zdań w danych w języku naturalnym	210
6.3.3. Tworzenie streszczeń dokumentów	214
6.4. Zmiana paradygmatu. Analiza obiektów	222
6.4.1. Podsumowania danych w języku naturalnym	226
6.5. Jakość analiz do przetwarzania danych w języku naturalnym	230
6.6. Uwagi końcowe	234
6.7. Zalecane ćwiczenia	234
6.8. Zasoby online	235
7. Eksploracja skrzynek pocztowych: analiza, kto rozmawia z kim, o czym, jak często i nie tylko	237
7.1. Przegląd	238
7.2. Uzyskiwanie i przetwarzanie korpusu danych pocztowych	239
7.2.1. Uniksowe skrzynki pocztowe	239

7.2.2. Pobieranie danych Enron	243
7.2.3. Konwersja korpusu poczty na uniksowy format mbox	245
7.2.4. Konwertowanie uniksowych skrzynek pocztowych na obiekty DataFrame biblioteki pandas	247
7.3. Analiza korpusu Enron	249
7.3.1. Zapytania według zakresu dat (godzin)	250
7.3.2. Analiza wzorców w komunikacji nadawca-odbiorca	253
7.3.3. Wyszukiwanie wiadomości e-mail według słów kluczowych	257
7.4. Analiza własnych danych pocztowych	258
7.4.1. Dostęp do Twojej skrzynki Gmail za pomocą OAuth	260
7.4.2. Pobieranie i parsowanie wiadomości e-mail	262
7.4.3. Wizualizacja wzorców w e-mailu za pomocą frameworka Immersion	264
7.5. Uwagi końcowe	265
7.6. Zalecane ćwiczenia	265
7.7. Zasoby online	266
8. Eksploracja serwisu GitHub: badanie nawyków podczas współtworzenia oprogramowania, tworzenie grafów zainteresowań i nie tylko	269
8.1. Przegląd	270
8.2. Odkrywanie API GitHuba	270
8.2.1. Tworzenie połączenia do API serwisu GitHub	272
8.2.2. Tworzenie żądań do API GitHuba	275
8.3. Modelowanie danych za pomocą grafów właściwości	277
8.4. Analiza grafów zainteresowań serwisu GitHub	280
8.4.1. "Wysiewanie" grafu zainteresowań	281
8.4.2. Obliczanie miar centralności grafu	284
8.4.3. Rozszerzanie grafu zainteresowań z wykorzystaniem krawędzi "śledzi" dla użytkowników	287

8.4.4. Używanie węzłów jako punktów przestawnych w celu tworzenia bardziej wydajnych zapytań 296

8.4.5. Wizualizacja grafów zainteresowań 301

8.5. Uwagi końcowe 303

8.6. Zalecane ćwiczenia 304

8.7. Zasoby online 305

CZĘŚĆ II. TWITTER. RECEPTURY

9. Twitter. Receptury 309

9.1. Dostęp do interfejsu API Twittera dla celów programistycznych 310

9.2. Wykorzystanie OAuth w celu uzyskania dostępu do interfejsu API Twittera dla aplikacji produkcyjnych 311

9.3. Odkrywanie trendów 315

9.4. Wyszukiwanie tweetów 316

9.5. Konstruowanie wygodnych wywołań funkcji 318

9.6. Zapisywanie i przywracanie danych JSON z wykorzystaniem plików tekstowych 319

9.7. Zapisywanie danych JSON i uzyskiwanie dostępu do nich za pomocą MongoDB 320

9.8. Pobieranie próbek z mechanizmu firehose Twittera za pomocą API Streaming 323

9.9. Pobieranie danych szeregów czasowych 324

9.10. Wyodrębnianie podmiotów z tweetów 326

9.11. Znajdowanie najpopularniejszych tweetów w kolekcji 327

9.12. Znajdowanie najpopularniejszych obiektów w kolekcji tweetów 329

9.13. Tabularyzacja analizy częstości 330

9.14. Znajdowanie użytkowników, którzy retweetowali status 331

9.15. Wyodrębnianie przypisania retweeta 333

9.16. Wykonywanie odpornych na błędy żądań do Twittera 334

9.17. Pobieranie informacji o profilu użytkownika 337

- 9.18. Wyodrębnianie podmiotów tweeta z dowolnego tekstu 338
- 9.19. Pobieranie wszystkich znajomych lub obserwatorów użytkownika 339
- 9.20. Analiza znajomych i obserwatorów użytkownika 341
- 9.21. Zbieranie tweetów użytkownika 342
- 9.22. Crawling grafu znajomości 344
- 9.23. Analiza treści tweetów 346
- 9.24. Tworzenie streszczeń celów łączy 347
- 9.25. Analizowanie ulubionych tweetów użytkownika 350
- 9.26. Uwagi końcowe 352
- 9.27. Zalecane ćwiczenia 352
- 9.28. Zasoby online 353

CZĘŚĆ III. ZAŁĄCZNIKI

- A. Informacje o maszynie wirtualnej przeznaczonej dla tej książki 357
 - B. Elementarz OAuth 359
 - C. Porady i wskazówki na temat Pythona i środowiska Jupyter Notebook 363
- Skorowidz 365