

Web Data Mining z użyciem języka Python

*Odkrywaj i wyodrębniaj informacje ze stron
internetowych za pomocą języka Python*

Dr Ranjana Rajnish
Dr Meenakshi Srivastava

APN Promise 2023

Spis treści

<i>O autorkach</i>	xiii
<i>O recenzencie</i>	xiv
<i>Podziękowania</i>	xv
<i>Przedmowa</i>	xvi
1 Eksploracja sieci Web – Wprowadzenie	1
Wstęp	1
Struktura	1
Cele	2
Wprowadzenie do eksploracji sieci Web	2
Sieć World Wide Web	3
Ewolucja sieci World Wide Web	6
Internet i Web 2.0	8
Eksplorowanie, modelowanie i analizowanie danych	9
<i>Podstawy eksploracji sieci Web</i>	12
<i>Kategorie eksploracji sieci Web</i>	13
Różnica między eksploracją danych i eksploracją sieci Web	16
<i>Zastosowania eksploracji sieci Web</i>	17
Eksploracja sieci Web i język Python	20
<i>Podstawowe biblioteki Pythona do eksploracji sieci Web</i>	21
Jak Python pomaga w eksploracji sieci Web?	23
<i>Wyrażenia regularne</i>	24
<i>Programy z obsługą sieci</i>	27
<i>Usługi internetowe</i>	31
<i>Rzut okiem na to, jak sposób Python ułatwia to wszystko</i>	32
Podsumowanie	34
Punkty do zapamiętania	35
Test zdobytej wiedzy	36
<i>Odpowiedzi</i>	37
Pytania	38
Kluczowe pojęcia	38

2 Taksonomia eksploracji sieci Web	39
Wstęp	39
Struktura	39
Cel	40
Wprowadzenie do eksploracji sieci Web	40
Eksploracja zawartości sieci Web	42
<i>Podstawowe zastosowania eksploracji zawartości sieci Web</i>	<i>44</i>
<i>Zawartość strony internetowej</i>	<i>45</i>
<i>Wstępne przetwarzanie zawartości</i>	<i>46</i>
<i>Analiza zawartości strony internetowej</i>	<i>48</i>
Eksploracja struktury sieci Web	49
Eksploracja korzystania z sieci Web	50
Kluczowe pojęcia	51
<i>Wskaźniki rankingowe</i>	<i>52</i>
<i>PageRank</i>	<i>52</i>
<i>Koncentratory i autorytety</i>	<i>54</i>
<i>Roboty internetowe</i>	<i>54</i>
<i>Zapach informacji</i>	<i>55</i>
<i>Profil użytkownika</i>	<i>55</i>
<i>Bibliometryki online</i>	<i>56</i>
<i>Rodzaje wskaźników bibliometrycznych</i>	<i>56</i>
Podsumowanie	57
Do zapamiętania	57
Test zdobytej wiedzy	58
<i>Odpowiedzi</i>	<i>60</i>
Pytania	60
Kluczowe terminy	60
 3 Główne zastosowania eksploracji sieci Web	 61
Wstęp	61
Struktura	61
Cele	62
Spersonalizowane aplikacje klienckie – handel elektroniczny	62
Wyszukiwanie w sieci	63
<i>Najczęściej stosowane metody śledzenia w witrynie</i>	<i>69</i>

Spersonalizowane portale i sieci Web	70
Optymalizacja wydajności usług internetowych	71
<i>Współczynnik odrzuceń</i>	72
<i>Średni czas na stronie</i>	72
<i>Unikalni użytkownicy</i>	72
Eksploracja procesów	74
Reguły asocjacyjne	75
Eksploracja reguł asocjacyjnych	77
Komponenty algorytmu Apriori	78
<i>Wsparcie i częste zbiory elementów</i>	79
<i>Wiarygodność</i>	79
<i>Podniesienie</i>	79
<i>Kroki w algorytmie Apriori</i>	80
Wzorce sekwencji	81
<i>Baza danych sekwencji</i>	82
<i>Podsekwencja kontra nadsekwencja</i>	82
<i>Minimalne wsparcie</i>	83
<i>Prefiks i sufiks</i>	84
<i>Projekcja</i>	84
Eksploracja reguł asocjacyjnych i biblioteki Pythona	85
<i>Pandas</i>	85
<i>Mlxtend</i>	85
Podsumowanie	88
Do zapamiętania	89
Test zdobytej wiedzy	90
<i>Odpowiedzi</i>	92
Pytania	92
Kluczowe pojęcia	92
4 Podstawy języka Python	93
Wstęp	93
Struktura	93
Cele	94
Wprowadzenie do języka Python	94
Podstawy Pythona	95

<i>Programowanie w Pythonie</i>	96
<i>„Hello World” – pierwszy skrypt w Pythonie</i>	97
<i>Instrukcje warunkowe/selekcji</i>	99
<i>Pętle/instrukcje iteracji</i>	102
<i>Funkcje</i>	106
<i>Listy</i>	110
Podstawy HTML: badanie strony internetowej	112
Podstawowe biblioteki Pythona	114
Instalacja Pythona	115
<i>Platforma uniksowa i linuksowa</i>	116
<i>Platforma Windows</i>	117
<i>Macintosh</i>	120
Wprowadzenie do popularnych IDE i PDE	124
<i>IDLE</i>	124
<i>Atom</i>	125
<i>Sublime Text</i>	125
<i>PyDev</i>	126
<i>Spyder</i>	126
<i>PyCharm</i>	126
<i>Google Colab</i>	127
Instalacja dystrybucji Anaconda	127
Podsumowanie	131
Do zapamiętania	131
Test zdobytej wiedzy	132
<i>Odpowiedzi</i>	134
5 Ekstrakcja danych z sieci Web	135
Wstęp	135
Struktura	136
Cele	136
Wprowadzenie do ekstrakcji danych z sieci Web	137
Ekstrakcja danych z sieci Web	137
<i>Zastosowania ekstrakcji danych z sieci Web</i>	139
<i>Działanie ekstraktora danych z sieci Web</i>	141
<i>Wyzwania związane z ekstrakcją danych z sieci Web</i>	145

<i>Moduły Pythona używane do ekstrakcji danych</i>	146
<i>Legalność ekstrakcji danych z sieci Web</i>	147
Wyodrębnianie i wstępne przetwarzanie danych.	151
Obsługa tekstu, obrazów i filmów.	152
<i>Obsługa tekstu</i>	154
<i>Obsługa obrazów</i>	155
<i>Wyodrębnianie filmów ze strony internetowej</i>	160
Ekstrakcja danych z dynamicznych witryn internetowych.	167
Zabezpieczenie CAPTCHA.	170
<i>Studium przypadku: Implementacja ekstrakcji danych w celu</i> <i>opracowania ekstraktora wyszukującego najnowsze wiadomości</i>	173
Podsumowanie	185
Do zapamiętania.	186
Test zdobytej wiedzy	187
<i>Odpowiedzi</i>	189
Pytania	190
Kluczowe pojęcia	190
6 Eksploracja opinii.	191
Wstęp	191
Struktura	192
Cele	192
Pojęcia związane z eksploracją opinii	192
<i>Biblioteka NLTK do analizy nastrojów</i>	195
<i>Eksploracja opinii/analiza nastrojów na różnych poziomach</i>	196
Zbieranie recenzji	198
<i>Źródła danych używane do eksplorowania opinii</i>	198
Praca z danymi	199
Wstępne przetwarzanie danych.	200
<i>Tokenizacja</i>	200
Oznaczanie części mowy.	202
Ekstrakcja cech	205
<i>Worek słów</i>	206
<i>TF-IDF</i>	206
Studium przypadku dotyczące analizy nastrojów	207

Podsumowanie	209
Do zapamiętania	210
Test zdobytej wiedzy	210
<i>Odpowiedzi</i>	212
Pytania	212
Kluczowe pojęcia	213
7 Eksploracja struktury sieci Web	215
Wstęp	215
Struktura	215
Cele	216
Wprowadzenie do eksploracji struktury sieci Web	216
Pojęcia związane z eksploracją struktury sieci Web	218
Rodzaje eksploracji struktury sieci Web	219
Eksploracja grafów sieci Web	221
Wyodrębnianie informacji z Internetu	224
Eksploracja sieci Deep Web	228
Wyszukiwanie w sieci i hiperłącza	231
Analiza hiperłączy w sieci Web	232
Algorytm Hyperlink Induced Topic Search (HITS)	234
Algorytm oparty na podziale	240
Implementacja w Pythonie	248
Podsumowanie	250
Do zapamiętania	251
Test zdobytej wiedzy	254
<i>Odpowiedzi</i>	256
Pytania	256
Kluczowe pojęcia	257
8 Analiza sieci społecznych w języku Python	259
Wstęp	259
Struktura	260
Cele	260
Wprowadzenie do analizy sieci społecznych	260
Tworzenie sieci	264

<i>Rodzaje grafów</i>	267
Analizowanie sieci	275
Wskaźniki odległości w połączeniach sieci	278
<i>Odległość</i>	278
<i>Średnia odległość</i>	281
<i>Ekscentryczność</i>	281
<i>Średnica</i>	282
<i>Promień</i>	282
<i>Obwód</i>	283
<i>Centrum</i>	284
Influencerzy w sieci	284
Studium przypadku dotyczące zbioru danych Facebooka	286
Podsumowanie	293
Do zapamiętania	294
Test zdobytej wiedzy	296
<i>Odpowiedzi</i>	297
Pytania	297
Kluczowe pojęcia	298
9 Eksploracja korzystania z sieci Web	299
Wstęp	299
Struktura	299
Cele	300
Proces eksploracji korzystania z sieci Web	300
Źródła danych	302
Rodzaje danych	303
<i>Dane dotyczące korzystania</i>	303
<i>Dane dotyczące treści</i>	306
<i>Dane dotyczące struktury</i>	306
<i>Dane dotyczące użytkownika</i>	306
Kluczowe elementy wstępnego przetwarzania danych	
korzystania z sieci Web	307
<i>Czyszczenie danych</i>	307
<i>Identyfikacja użytkownika</i>	308
<i>Identyfikacja sesji</i>	309

<i>Identyfikacja ścieżki</i>	309
Modelowanie danych	310
<i>Eksploracja reguł asocjacyjnych</i>	310
<i>Wzorzec sekwencji</i>	311
<i>Grupowanie</i>	311
<i>Eksploracja klasyfikacji</i>	311
Odkrywanie i analiza wzorców	312
<i>Reguła asocjacyjna do odkrywania wiedzy</i>	313
<i>Odkrywanie wzorców poprzez grupowanie</i>	313
<i>Eksploracja wzorców sekwencji w celu odkrywania wiedzy</i>	314
<i>Nauka poprzez klasyfikację</i>	314
<i>Analiza wzorców</i>	315
Prognozy dotyczące wzorca transakcji	315
<i>Budowanie systemu rekomendacyjnego opartego na treści</i>	317
<i>Profil produktu</i>	317
<i>Profil użytkownika</i>	317
Podsumowanie	318
Do zapamiętania	318
Test zdobytej wiedzy	319
<i>Odpowiedzi</i>	321
Pytania	322
Kluczowe pojęcia	322
Indeks	323