

Spis treści

Przedmowa	9
------------------------	----------

Część I. Tworzenie pełza czy

1. Twój pierwszy robot indeksujący	17
Połączenie	17
Wprowadzenie do biblioteki BeautifulSoup	19
Instalacja biblioteki BeautifulSoup	20
Korzystanie z biblioteki BeautifulSoup	22
Stabilne połączenia i obsługa wyjątków	23
2. Zaawansowana analiza składniowa HTML	27
Młotek nie zawsze jest potrzebny	27
Kolejna porcja BeautifulSoup	28
Funkcje find() i find_all()	30
Inne obiekty biblioteki BeautifulSoup	32
Poruszanie się po drzewach hierarchii	32
Wyrażenia regularne	36
Wyrażenia regularne w bibliotece BeautifulSoup	40
Uzyskiwanie dostępu do atrybutów	41
Wyrażenia lambda	41
3. Tworzenie robotów indeksujących	43
Poruszanie się po pojedynczej domenie	43
Pełzanie po całej witrynie	47
Gromadzenie danych z całej witryny	49
Pełzanie po internecie	51

4. Modele ekstrakcji danych.....	57
Planowanie i definiowanie obiektów	58
Obsługa różnych szat graficznych	61
Konstruowanie robotów indeksujących	65
Poruszanie się po witrynach za pomocą paska wyszukiwania	65
Poruszanie się po witrynach za pomocą odnośników	68
Poruszanie się pomiędzy różnymi typami stron	70
Właściwe podejście do procesu tworzenia modeli robotów indeksujących	72
5. Scrapy.....	73
Instalacja biblioteki Scrapy	73
Inicjowanie nowego pająka	74
Pisanie prostego robota indeksującego	75
Korzystanie z pająków przy użyciu reguł	76
Tworzenie elementów	80
Wyświetlanie elementów	81
Potoki elementów	82
Dzienniki zdarzeń w bibliotece Scrapy	85
Dodatkowe zasoby	86
6. Przechowywanie danych.....	87
Pliki multimedialne	87
Przechowywanie danych w plikach CSV	90
MySQL	92
Instalacja środowiska MySQL	92
Podstawowe polecenia	94
Integracja ze środowiskiem Python	97
Techniki bazodanowe i dobre rozwiązania	100
Sześć stopni oddalenia w środowisku MySQL	102
Alerty e-mail	105

Część II. Zaawansowana ekstrakcja danych

7. Odczytywanie dokumentów.....	109
Kodowanie dokumentu	109
Pliki tekstowe	110
Kodowanie tekstu a internet globalny	110
Format CSV	114
Odczyt plików CSV	115
Format PDF	116
Edytor Microsoft Word i pliki .docx	118

8. Oczyszczanie danych	123
Oczyszczanie na poziomie kodu	123
Normalizacja danych	126
Oczyszczanie pozyskanych danych	128
OpenRefine	128
9. Odczyt i zapis języków naturalnych	133
Podsumowywanie danych	134
Modele Markowa	137
Sześć stopni oddalenia od Wikipedii — podsumowanie	140
Natural Language Toolkit	142
Instalacja i konfiguracja	143
Analiza statystyczna za pomocą pakietu NLTK	143
Analiza leksykologiczna za pomocą pakietu NLTK	145
Dodatkowe zasoby	149
10. Kwestia formularzy i pól logowania	151
Biblioteka Requests	151
Przesyłanie podstawowego formularza	152
Przyciski opcji, pola zaznaczania i inne mechanizmy wprowadzania danych	153
Wysyłanie plików i obrazów	155
Pola logowania i ciasteczka	155
Podstawowe uwierzytelnianie protokołu HTTP	157
Inne problemy z formularzami	158
11. Ekstrakcja danych a język JavaScript	159
Krótkie wprowadzenie do języka JavaScript	160
Popularne biblioteki JavaScriptu	161
Ajax i dynamiczny HTML	163
Uruchamianie kodu JavaScriptu w środowisku Python	
za pomocą biblioteki Selenium	164
Dodatkowe obiekty WebDriver	169
Obsługa przekierowań	169
Końcowe uwagi na temat języka JavaScript	171
12. Ekstrakcja danych poprzez API	173
Krótkie wprowadzenie do API	173
Metody HTTP a API	175
Dodatkowe informacje na temat odpowiedzi API	176
Analizowanie składni formatu JSON	177

Nieudokumentowane API	178
Wyszukiwanie nieudokumentowanych API	180
Dokumentowanie nieudokumentowanych API	181
Automatyczne wyszukiwanie i dokumentowanie API	182
Łączenie API z innymi źródłami danych	184
Dodatkowe informacje na temat API	188
13. Przetwarzanie obrazów i rozpoznawanie tekstu	189
Przegląd bibliotek	190
Pillow	190
Tesseract	190
NumPy	193
Przetwarzanie prawidłowo sformatowanego tekstu	193
Automatyczne korygowanie obrazów	196
Ekstrakcja danych z obrazów umieszczonych w witrynach	198
Odczytywanie znaków CAPTCHA i uczenie aplikacji Tesseract	201
Uczenie aplikacji Tesseract	202
Ekstrakcja kodów CAPTCHA i przesyłanie odpowiedzi	205
14. Unikanie pułapek na boty	209
Kwestia etyki	209
Udawanie człowieka	210
Dostosuj nagłówki	210
Obsługa ciastek za pomocą języka JavaScript	212
Wyczucie czasu to podstawa	214
Popularne zabezpieczenia formularzy	214
Wartości ukrytych pól wejściowych	215
Unikanie wabików	216
Być człowiekiem	218
15. Testowanie witryn internetowych za pomocą robotów indeksujących	219
Wprowadzenie do testowania	219
Czym są testy jednostkowe?	220
Moduł unittest	220
Testowanie Wikipedii	222
Testowanie za pomocą biblioteki Selenium	224
Interakcje z witryną	225
Selenium czy unittest?	228

16. Zrównoleganie procesu ekstrakcji danych.....	229
Procesy i wątki.....	229
Wielowątkowa ekstrakcja danych.....	230
Wyścigi i kolejki.....	232
Moduł threading.....	235
Wieloprosesowa ekstrakcja danych.....	237
Przykład z Wikipedią.....	238
Komunikacja międzyprocesowa.....	240
Wieloprosesowa ekstrakcja danych — metoda alternatywna.....	242
17. Zdalna ekstrakcja danych z internetu.....	243
Powody korzystania z serwerów zdalnych.....	243
Unikanie blokowania adresu IP.....	243
Przenośność i rozszerzalność.....	245
Tor.....	245
PySocks.....	246
Hosting zdalny.....	247
Uruchamianie z poziomu serwisu hostingowego.....	247
Uruchamianie z poziomu chmury.....	248
Dodatkowe zasoby.....	250
18. Legalność i etyka ekstrakcji danych z internetu.....	251
Znaki towarowe, prawa autorskie, patenty, ojej!.....	251
Prawo autorskie.....	252
Naruszenie prawa własności rzeczy ruchomych.....	254
Ustawa o oszustwach i nadużyciach komputerowych.....	256
Plik robots.txt i warunki świadczenia usług.....	256
Trzy roboty indeksujące.....	259
Sprawa eBay przeciwko Bidder's Edge (prawo własności rzeczy ruchomych).....	260
Sprawa Stany Zjednoczone przeciwko Auernheimerowi (ustawa CFAA).....	261
Sprawa Field przeciwko Google (prawo autorskie i plik robots.txt).....	263
Co dalej?.....	263
Skorowidz.....	265