

УДК 004.738.1 :004.738.52Python

ББК 32.971.353

**М66**

**Митчелл, Райан.**

**М66**     Скрапинг веб-сайтов с помощью Python. Сбор данных из современного интернета / Р. Митчелл ; пер. с англ. А. В. Груздева. — 2-е изд., эл. — 1 файл pdf: 282 с. — Москва : ДМК Пресс, 2023. — Систем. требования: Adobe Reader XI либо Adobe Digital Editions 4.5 ; экран 10". — Текст : электронный.

ISBN 978-5-89818-305-9

Изучите методы скрапинга и краулинга веб-сайтов, чтобы получить доступ к неограниченному объему данных в любом уголке Интернета в любом формате. С помощью этого практического руководства вы узнаете, как использовать скрипты Python и веб-API, чтобы одновременно собрать и обработать данные с тысяч или даже миллионов веб-страниц.

Идеально подходящая для программистов, специалистов по безопасности и веб-администраторов, знакомых с языком Python, эта книга знакомит не только с основными принципами работы веб-скраперов, но и углубляется в более сложные темы, такие как анализ сырых данных или использование скраперов для тестирования интерфейса веб-сайта. Примеры программного кода, приведенные в книге, помогут разобраться в этих принципах на практике.

УДК 004.738.1 :004.738.52Python

ББК 32.971.353

**Электронное издание на основе печатного издания:** Скрапинг веб-сайтов с помощью Python. Сбор данных из современного интернета / Р. Митчелл ; пер. с англ. А. В. Груздева. — Москва : ДМК Пресс, 2016. — 280 с. — ISBN 978-5-97060-223-2. — Текст : непосредственный.

Все права защищены. Любая часть этой книги не может быть воспроизведена в какой бы то ни было форме и какими бы то ни было средствами без письменного разрешения владельцев авторских прав.

Материал, изложенный в данной книге, многократно проверен. Но поскольку вероятность технических ошибок все равно существует, издательство не может гарантировать абсолютную точность и правильность приводимых сведений. В связи с этим издательство не несет ответственности за возможные ошибки, связанные с использованием книги.

В соответствии со ст. 1299 и 1301 ГК РФ при устранении ограничений, установленных техническими средствами защиты авторских прав, правообладатель вправе требовать от нарушителя возмещения убытков или выплаты компенсации.

ISBN 978-5-89818-305-9

© 2015 Ryan Mitchell

© Оформление, перевод, ДМК Пресс, 2016

# Содержание

<b>Предисловие .....</b>	<b>10</b>
<b>Вступление .....</b>	<b>13</b>
<b>ЧАСТЬ I. ПОСТРОЕНИЕ СКРАПЕРОВ .....</b>	<b>20</b>
<b>Глава 1. Ваш первый скрапер .....</b>	<b>21</b>
Соединение с Интернетом .....	21
Введение в BeautifulSoup .....	24
Установка BeautifulSoup .....	24
Запуск BeautifulSoup .....	26
Как обеспечить надежный скрапинг .....	28
<b>Глава 2. Продвинутый парсинг HTML .....</b>	<b>31</b>
Вам не всегда нужен молоток .....	31
Еще одно применение BeautifulSoup .....	32
find() и findAll() .....	34
Другие объекты BeautifulSoup .....	36
Навигация по дереву синтаксического разбора .....	37
Работа с дочерними элементами и элементами-потомками .....	38
Работа с одноуровневыми элементами .....	39
Работа с родительскими элементами .....	40
Регулярные выражения .....	41
Регулярные выражения и BeautifulSoup .....	46
Работа с атрибутами .....	47
Лямбда-выражения .....	48
За рамками BeautifulSoup .....	48
<b>Глава 3. Запуск краулера .....</b>	<b>50</b>
Обход отдельного домена .....	50
Краулинг всего сайта .....	54
Сбор данных по всему сайту .....	57
Краулинг Интернета .....	59
Краулинг с помощью Scrapy .....	65
<b>Глава 4. Использование API .....</b>	<b>70</b>
Как работают API .....	71
Общепринятые соглашения .....	72
Методы .....	72

## 6 ❖ Содержание

Аутентификация.....	73
Ответы .....	74
Вызовы API.....	75
Echo Nest.....	76
Несколько примеров.....	76
Twitter.....	78
Приступаем к работе.....	78
Несколько примеров.....	79
Google API.....	83
Приступаем к работе.....	83
Несколько примеров.....	84
Парсинг JSON-данных.....	86
Возвращаем все это домой.....	88
Подробнее о применении API .....	92
<b>Глава 5. Хранение данных .....</b>	<b>94</b>
Медиафайлы.....	94
Сохранение данных в формате CSV.....	97
MySQL.....	99
Установка MySQL.....	100
Некоторые основные команды.....	102
Интеграция с Python .....	106
Методы работы с базами данных и эффективная практика .....	109
«Шесть шагов» в MySQL.....	112
Электронная почта .....	115
<b>Глава 6. Чтение документов .....</b>	<b>117</b>
Кодировка документа.....	117
Текст.....	118
Кодировка текста и глобальный Интернет.....	119
CSV .....	124
Чтение CSV-файлов.....	124
PDF .....	126
Microsoft Word и .docx.....	128
<b>ЧАСТЬ II. ПРОДВИНУТЫЙ СКРАПИНГ .....</b>	<b>132</b>
<b>Глава 7. Очистка данных.....</b>	<b>133</b>
Очистка данных на этапе создания кода.....	133
Нормализация данных .....	136

Очистка данных постфактум .....	138
OpenRefine.....	139
<b>Глава 8. Чтение и запись естественных языков .....</b>	<b>144</b>
Аннотирование данных.....	145
Марковские модели.....	148
Шесть шагов Википедии: заключительная часть .....	152
Natural Language Toolkit .....	156
Установка и настройка .....	156
Статистический анализ с помощью NLTK .....	156
Лексикографический анализ с помощью NLTK .....	160
Дополнительные ресурсы .....	163
<b>Глава 9. Краулинг сайтов, использующих веб-формы.....</b>	<b>165</b>
Библиотека requests .....	165
Отправка простой формы .....	166
Радиокнопки, флажки и другие элементы ввода данных .....	168
Отправка файлов и изображений.....	170
Работа с логинами и cookies .....	171
Базовая HTTP-аутентификация .....	173
Другие проблемы при работе с формами .....	174
<b>Глава 10. Скрапинг JavaScript-кода .....</b>	<b>175</b>
Краткое введение в JavaScript .....	176
Распространенные библиотеки JavaScript.....	177
Ajax и динамический HTML.....	180
Выполнение JavaScript в Python с помощью библиотеки Selenium.....	181
Обработка редиректов.....	186
<b>Глава 11. Обработка изображений и распознавание текста .....</b>	<b>189</b>
Обзор библиотек .....	190
Pillow .....	190
Tesseract .....	191
NumPy .....	192
Обработка хорошо отформатированного текста.....	193
Скрапинг текста с изображений, размещенных на веб-сайтах.....	196

Чтение CAPTCHA и обучение Tesseract .....	198
Обучение Tesseract.....	200
Извлечение CAPTCHA и отправка результатов распознавания .....	204
<b>Глава 12. Обход ловушек в ходе скрапинга .....</b>	<b>208</b>
Обратите внимание на этический аспект .....	209
Учимся выглядеть как человек.....	210
Настройте заголовки.....	210
Обработка cookies .....	212
Время решает все.....	214
Общие функции безопасности, используемые веб-формами .....	215
Значения полей скрытого ввода .....	215
Обходим «горшочки с медом».....	217
Проверяем скрапер на «человечность» .....	219
<b>Глава 13. Тестирование вашего сайта с помощью скраперов.....</b>	<b>221</b>
Введение в тестирование .....	222
Что такое модульные тесты?.....	222
Питоновский модуль unittest .....	223
Тестирование Википедии.....	224
Тестирование с помощью Selenium.....	227
Взаимодействие с сайтом .....	227
Unittest или Selenium? .....	231
<b>Глава 14. Скрапинг с помощью удаленных серверов .....</b>	<b>233</b>
Зачем использовать удаленные серверы? .....	233
Как избежать блокировки IP-адреса.....	234
Переносимость и расширяемость.....	235
Tor.....	236
PySocks .....	237
Удаленный хостинг.....	238
Запуск с аккаунта веб-хостинга .....	238
Запуск из облака.....	240
Дополнительные ресурсы .....	241
Заглянем в будущее.....	242
<b>Приложение А. Кратко о том, как работает Python .....</b>	<b>244</b>
Установка и «Hello, World!» .....	244

<b>Приложение В. Кратко о том, как работает Интернет.....</b>	<b>248</b>
---	------------

<b>Приложение С. Правовые и этические аспекты веб-скрапинга .....</b>	<b>252</b>
---	------------

Товарные знаки, авторские права, патенты, о боже! .....	252
---	-----

Авторское право .....	254
-----------------------	-----

Посягательство на движимое имущество.....	256
---	-----

Закон о компьютерном мошенничестве и злоупотреблении.....	258
---	-----

robots.txt и Пользовательское соглашение .....	259
--	-----

Три на шумевших случая в практике веб-скрапинга .....	263
---	-----

eBay против Bidder's Edge и посягательство на движимое имущество.....	263
---	-----

США против Орнхаймера и Закон о компьютерном мошенничестве и злоупотреблении.....	265
---	-----

Филд против Google: авторское право и robots.txt .....	268
--	-----

<b>Об авторе .....</b>	<b>269</b>
------------------------	------------

<b>Колофон .....</b>	<b>270</b>
----------------------	------------

<b>Предметный указатель.....</b>	<b>271</b>
----------------------------------	------------