

БНТУ

Научная библиотека



\* 8 0 1 2 3 4 1 3 7 \*

Райан Митчелл

# Современный скрапинг веб-сайтов с помощью Python

---

АВТОМАТИЗАЦИЯ СБОРА ДАННЫХ С ИНТЕРНЕТ-САЙТОВ  
ПРИ ПОМОЩИ ЯЗЫКА PYTHON

2-Е МЕЖДУНАРОДНОЕ ИЗДАНИЕ



Санкт-Петербург · Москва · Екатеринбург · Воронеж  
Нижний Новгород · Ростов-на-Дону  
Самара · Минск

2021

НАВУКОВАЯ БІБЛІЯТЭКА

Беларускага нацыянальнага  
тэхнічнага ўніверсітэта

Інв. № 1883834

# Краткое содержание

Введение.....	13
<b>Часть I. Разработка веб-скраперов</b>	
Глава 1. Ваш первый веб-скрапер.....	23
Глава 2. Углубленный синтаксический анализ HTML-кода.....	35
Глава 3. Разработка веб-краулеров.....	54
Глава 4. Модели веб-краулинга.....	70
Глава 5. Scrapy.....	90
Глава 6. Хранение данных.....	107
<b>Часть II. Углубленный веб-скрапинг</b>	
Глава 7. Чтение документов.....	135
Глава 8. Очистка «грязных» данных.....	151
Глава 9. Чтение и запись текстов на естественных языках.....	163
Глава 10. Сбор данных из форм и проверка авторизации.....	185
Глава 11. Веб-скрапинг данных JavaScript.....	195
Глава 12. Веб-краулинг с помощью API.....	212
Глава 13. Обработка изображений и распознавание текста.....	231
Глава 14. Как избежать ловушек веб-скрапинга.....	256

**Глава 15.** Тестирование сайтов с помощью веб-скраперов.....270

**Глава 16.** Параллельный веб-краулинг.....284

**Глава 17.** Удаленный веб-скрапинг.....303

**Глава 18.** Законность и этичность веб-скрапинга.....313

Движемся дальше.....331

Об авторе.....333

Об обложке.....334

# Оглавление

Введение.....	13
Что такое веб-скрапинг .....	14
Почему это называется веб-скрапингом .....	14
Об этой книге.....	16
Условные обозначения.....	17
Использование примеров кода .....	18
Благодарности .....	19
От издательства.....	20
 <b>Часть I. Разработка веб-скраперов</b>	
<b>Глава 1.</b> Ваш первый веб-скрапер.....	23
Установка соединения.....	23
Знакомство с BeautifulSoup .....	26
Установка BeautifulSoup .....	27
Работа с BeautifulSoup .....	29
Надежное соединение и обработка исключений .....	31
<b>Глава 2.</b> Углубленный синтаксический анализ HTML-кода.....	35
Иногда молоток не требуется.....	35
Еще одна тарелка BeautifulSoup.....	37
Функции find() и find_all() .....	38

Другие объекты BeautifulSoup .....	41
Навигация по деревьям.....	41
Регулярные выражения.....	46
Регулярные выражения и BeautifulSoup.....	50
Доступ к атрибутам.....	51
Лямбда-выражения.....	52
<b>Глава 3. Разработка веб-краулеров .....</b>	<b>54</b>
Проход отдельного домена.....	54
Сбор информации со всего сайта.....	59
Сбор информации с нескольких сайтов .....	65
<b>Глава 4. Модели веб-краулинга.....</b>	<b>70</b>
Планирование и определение объектов .....	71
Работа с различными макетами сайтов .....	75
Структурирование веб-краулеров.....	80
Веб-краулинг с помощью поиска .....	80
Сбор данных с сайтов по ссылкам .....	84
Сбор данных со страниц нескольких типов.....	87
Размышления о моделях веб-краулеров .....	88
<b>Глава 5. Scrapy .....</b>	<b>90</b>
Установка Scrapy .....	90
Пишем простой веб-скрапер.....	92
«Паук» с правилами .....	94
Создание объектов Item .....	98
Вывод объектов Item .....	100
Динамический конвейер.....	101
Ведение журнала Scrapy .....	105
Дополнительные ресурсы.....	106
<b>Глава 6. Хранение данных .....</b>	<b>107</b>
Медиафайлы .....	107
Хранение данных в формате CSV .....	111
MySQL .....	113

Установка MySQL.....	114
Несколько основных команд .....	117
Интеграция с Python .....	120
Приемы и рекомендуемые методики работы с базами данных .....	124
Шесть шагов в MySQL .....	126
Электронная почта.....	130
 <b>Часть II. Углубленный веб-скрапинг</b>	
<b>Глава 7. Чтение документов.....</b>	<b>135</b>
Кодировка документов .....	135
Текст.....	136
Текстовые кодировки и глобальный Интернет.....	137
CSV .....	142
PDF.....	144
Microsoft Word и файлы .docx .....	147
<b>Глава 8. Очистка «грязных» данных.....</b>	<b>151</b>
Очистка данных в коде .....	151
Очистка задним числом .....	157
OpenRefine.....	158
<b>Глава 9. Чтение и запись текстов на естественных языках.....</b>	<b>163</b>
Обобщение данных .....	164
Модели Маркова.....	168
Natural Language Toolkit .....	175
Установка и настройка.....	176
Статистический анализ с помощью NLTK.....	177
Лексикографический анализ с помощью NLTK.....	179
Дополнительные ресурсы.....	183
<b>Глава 10. Сбор данных из форм и проверка авторизации.....</b>	<b>185</b>
Библиотека Requests .....	185
Отправка простейшей формы .....	186
Переключатели, флажки и другие поля ввода .....	188

Передача файлов и изображений.....	190
Обработка данных авторизации и параметров cookie.....	191
Другие проблемы с формами .....	194
<b>Глава 11. Веб-скрапинг данных JavaScript.....</b>	<b>195</b>
Краткое введение в JavaScript.....	196
Популярные библиотеки JavaScript.....	197
Ajax и динамический HTML.....	200
Выполнение JavaScript в Python с помощью Selenium .....	201
Дополнительные веб-драйверы Selenium.....	207
Обработка перенаправлений.....	208
Последнее замечание о JavaScript .....	210
<b>Глава 12. Веб-краулинг с помощью API.....</b>	<b>212</b>
Краткое введение в API.....	212
API и HTTP-методы.....	214
Подробнее об ответах на API-запросы.....	216
Синтаксический анализ JSON.....	217
Недокументированные API.....	219
Поиск недокументированных API.....	221
Документирование недокументированных API.....	222
Автоматический поиск и документирование API.....	223
Объединение API с другими источниками данных .....	225
Дополнительные сведения об API .....	229
<b>Глава 13. Обработка изображений и распознавание текста.....</b>	<b>231</b>
Обзор библиотек.....	232
Pillow.....	232
Tesseract.....	233
Обработка хорошо отформатированного текста.....	237
Автоматическая коррекция изображений.....	240
Веб-скрапинг текста, представленного в виде изображений на сайтах.....	243
Чтение капчи и обучение Tesseract.....	247
Получение капчи и отправка решений.....	253

---

<b>Глава 14. Как избежать ловушек веб-скрапинга .....</b>	<b>256</b>
Этический момент.....	257
Выдать скрипт за человека.....	258
Настройте заголовки .....	258
Обработка данных cookie с помощью JavaScript .....	260
Своевременность — наше все.....	263
Основные средства защиты форм.....	263
Значения скрытых полей ввода.....	264
Как справляться с полями-приманками.....	265
Контрольный список: как выдать программу за человека.....	268
<b>Глава 15. Тестирование сайтов с помощью веб-скраперов.....</b>	<b>270</b>
Основы тестирования .....	271
Что такое юнит-тесты .....	271
Python-модуль unittest .....	272
Тестирование с помощью Selenium.....	277
Взаимодействие с сайтом .....	277
unittest или Selenium.....	282
<b>Глава 16. Параллельный веб-краулинг .....</b>	<b>284</b>
Процессы или потоки.....	285
Многопоточный веб-краулинг.....	285
Состояния гонки и очереди.....	288
Модуль threading .....	291
Многопроцессный веб-краулинг .....	294
Пример многопроцессного веб-краулинга.....	297
Обмен данными между процессами.....	298
Многопроцессный веб-краулинг: еще один подход .....	301
<b>Глава 17. Удаленный веб-скрапинг.....</b>	<b>303</b>
Зачем использовать удаленные серверы.....	303
Как избежать блокировки IP-адреса.....	304
Портируемость и расширяемость .....	305
Tog .....	306



Удаленный хостинг.....	308
Запуск веб-скрапера из учетной записи веб-хостинга.....	308
Запуск из облака .....	310
Дополнительные ресурсы.....	312
<b>Глава 18. Законность и этичность веб-скрапинга .....</b>	<b>313</b>
Торговые марки, авторские права, патенты... спасите-помогите!.....	314
Посягательство на движимое имущество .....	317
Акт о компьютерном мошенничестве и злоупотреблении.....	319
Файл robots.txt и условия использования.....	320
Три веб-скрапера .....	324
eBay против Bidder's Edge и посягательство на движимое имущество .....	325
Соединенные Штаты Америки против Ауэрнхаймера и Акт о компьютерном мошенничестве и злоупотреблении .....	327
Филд против Google: авторские права и robots.txt .....	329
Двигаемся дальше.....	331
Об авторе.....	333
Об обложке.....	334