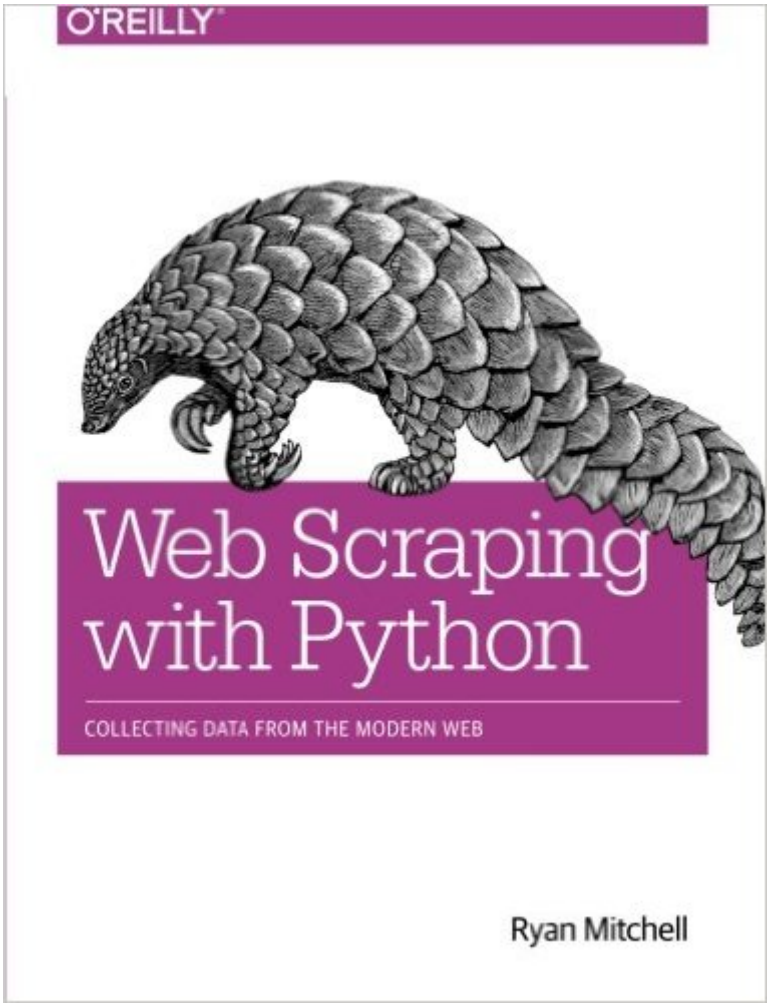
Web Scraping With Python: Collecting Data From The Modern Web



Synopsis

Learn web scraping and crawling techniques to access unlimited data from any web source in any format. With this practical guide, youâ ™ll learn how to use Python scripts and web APIs to gather and process data from thousandsâ ”or even millionsâ ”of web pages at once.Ideal for programmers, security professionals, and web administrators familiar with Python, this book not only teaches basic web scraping mechanics, but also delves into more advanced topics, such as analyzing raw data or using scrapers for frontend website testing. Code samples are available to help you understand the concepts in practice.Learn how to parse complicated HTML pagesTraverse multiple pages and sitesGet a general overview of APIs and how they workLearn several methods for storing the data you scrapeDownload, read, and extract data from documentsUse tools and techniques to clean badly formatted dataRead and write natural languagesCrawl through forms and loginsUnderstand how to scrape JavaScriptLearn image processing and text recognition

Book Information

Paperback: 256 pages

Publisher: O’Reilly Media; 1 edition (July 24, 2015)

Language: English

ISBN-10: 1491910291

ISBN-13: 978-1491910290

Product Dimensions: 7 x 0.6 x 9.2 inches

Shipping Weight: 1.2 pounds (View shipping rates and policies)

Average Customer Review: 4.6 out of 5 starsÂ Â See all reviewsÂ (33 customer reviews)

Best Sellers Rank: #25,476 in Books (See Top 100 in Books) #3 inÂ Books > Computers & Technology > Internet & Social Media > Web Browsers #5 inÂ Books > Computers & Technology > Web Development & Design > Web Services #10 inÂ Books > Computers & Technology > Internet & Social Media > Online Searching

Customer Reviews

Q&A with author Ryan Mitchell What got you interested in web scraping? In 2011, I started working for a company called Abine, that offered a service to remove customersÃ¢Â Â™ personal information from various sites on the Internet. In the early days of the company, the process of looking for someoneÃ¢Â Â™s personal information on all of these sites, filling out all these opt-out forms, faxing emailing, compiling reports to send back to the customers -- it all took a lot of time! I started looking into ways to streamline these processes, and add additional features. I built bots that

could search for profiles, store information in our database, fill out web forms, create documents, and send the emails and faxes automatically. Some of these sites were fairly bot-resistant, so I had to learn, and even invent, some interesting techniques to deal with them. I really fell in love with building bots and scraping the web, and continued to do it even after I left the company! Why is Python such a good fit for web scraping and building web crawlers? IÃ¢Â Â™ll be honest: As far as high performance programming languages go, Python does not win many speed contests. But with web scraping, youÃ¢Â Â™re not looking for speed -- sending and receiving data across the Internet will be thousands of times slower than any relatively tiny differences in language performance, so you can throw that metric out the window! What you need is something thatÃ¢Â Â™s lightweight, easy to deploy to remote machines, that can be installed and run anywhere, thatÃ¢Â Â™s easy to write and modify, and, perhaps most importantly: that has a plethora of well-document tools for just about any situation. Python has all of these in spades. WhatÃ¢Â Â™s the most interesting way youÃ¢Â Â™ve used web scraping, for professional or side projects? One of my favorite scraping projects, and something I introduce in Web Scraping with Python, is scraping Wikipedia for historical edits by IP address, time of the edit, and language. You can resolve the IP address to a geographic location, and explore when and where speakers of different languages are making edits. Lots of interesting sociological research potential there! A recent hobby of mine has also been automated CAPTCHA solving. I really enjoy analyzing new types of CAPTCHAs for vulnerabilities, writing scripts to pre-process the images, creating data sets for machine learning algorithms, and seeing how high I can get the success percentage of my bots! No real practical applications these days, but you never know when it will come in handy.

What information do you hope that readers of your book will walk away with? I try to stress a couple of things throughout the book: First, no website is bot-proof. Attempts to make websites more bot-proof generally also result in a loss of usability for human users. That loss of usability may be in the form of slower loading times, poor browser compatibility, lack of accessibility for users with mobility or visual impairments, or users on mobile devices. And many of these measures have no real deterring effect on web scrapers. If you can view the data in a browser, you can capture it with a scraper. Second, writing web scrapers that capture the data you want often involve combining multiple techniques, some creative thinking, and a dash of laziness. I canÃ¢Â Â™t count the number of times people have asked me to build a bot, or to help them build a bot, to collect data that could be easily obtained through an API! So sometimes your data collection problem can be solved using the information from only a single chapter in the book. On the other hand, I also provide an

example of a web scraper that uses JavaScript execution, HTML parsing, DOM interaction, and optical character recognition, all in one piece of code, in order to extract the text from book previews on ! (Sorry, !) When faced with a web scraping problem you should always 'work the steps' to try formulate a data extraction and processing plan -- itÃ¢Â Â™s not just about learning a single library or command! WhatÃ¢Â Â™s the most exciting or important thing happening in your space right now? Like many fields, especially computer science fields, thereÃ¢Â Â™s a lot being done with machine learning and big data. The percentage of page requests performed by humans and bots is about 50/50 right now, and as more humans are getting on the Internet, more bots are too -- and outpacing them! ThereÃ¢Â Â™s just so much data, and so many machines collecting that data, and so many connections we havenÃ¢Â Â™t been able to make before, waiting to be made. And these arenÃ¢Â Â™t just data scientists and server farm owners making them, either! The kind of research that once might have required months or years of surveys and data collection are now just a Python script, a database, and a weekend of coding away!

I really liked this book, for the following reasons:1. It is a great introduction to web scraping. The reader is given confidence to use well-known Python packages such as BeautifulSoup and get useful results from scraping webpages in a very short time.2. Where to go after learning the basics? - the author describes the tools, techniques and frameworks to use for scraping dynamic websites, including code examples. This is the most challenging part of the book because it frequently involves combining tools and the reader will have to get his/her hands dirty and learn by doing also. This is reasonable since different websites present different challenges.3. I liked the author’s writing style. She favors simple explanations, identifies potential pitfalls and makes clear, technical recommendations based on her experience.Highly recommended. I wish I had this book two years ago.

After learning the basics of Python I really struggled to dig into a project. Programming, just like a foreign language, leaves your memory quickly if you don’t use it. I had tried web scraping with several video tutorials, but couldn’t work through the tutorials to the point of understanding how to build one myself. Then...I got Ryan Mitchell’s book. This book sets you up with not only the basics, but also more advanced techniques that you’ll need to really build out your scraper. Ryan touches on other subjects such as using a database, working around data hidden in Javascript, cleaning up data, using NLPT, and more. You’ll get a solid foundation to launch into your own Web Scraping

project, and learn just enough about additional topics (like MySql) to integrate them into your scraper. I really appreciate how Ryan made this approachable for both a Python beginner, and for an intermediate user.

This is a truly excellent book. It is the closest I have seen in a book to the experience of sitting with a friendly, approachable expert who is ready to answer your questions intelligently and in a supportive way. You need the very basics of Python as can be learned from the Pycharm educational version but everything else is provided.

Good at describing how web scraping works but does not go into a lot more details on most areas. For example, it lacks completely as how to handle cookies. It’s true that the requests library handles the cookies automatically but if the cookies need to be manipulated, it doesn’t describe how such can be done.

Writing this sort of book is particularly challenging because readers will come in knowing different subsets of the material being addressed. However, I found that Ryan did a great job of separating out sections so that I could skip the parts that I already knew, and zoom in on the sections that I need for any specific project. In that way, it manages to serve as both a mostly-linear teaching book, but also a reference.I’d also really appreciated Ryan’s nuanced reasoning on the ethics of scraping, which got a layer deeper than simply throwing responsibility over to the reader with some version of "with great power comes great responsibility", and got into a thoughtful discussion of how to make the call about whether a potential application is legitimate.

This is mostly a beginners’ manual, so don’t expect extremely complicated programs or tips. However, if you are new to web scraping, this is a great introductory book to the tools available in Python and their uses. In my case, I had learned most of what was in the book using trial and error (and lots of time going through Stack Exchange questions!). If I had had this book before, I would have saved a lot of time learning the basics.

This book is excellent. I love the focus on Python 3 and all the techniques presented. I felt like it was Christmas day just reading the Table of Contents. THIS BOOK IS PACKED FULL OF INFORMATION. It is a joy to read and always has answers when I am looking. I have found it useful in my scraping at work and at home on multiple occasions. Easy read and a joy to have read. Thank

you Ryan for this book!

This book got me up and running pretty quickly. After reading the first few chapters along with a few Google searches I was able to build the web scraper that I wanted. The rest of the book as some nice little nuggets for advanced users. Great choice for the price.

Web Scraping with Python: Collecting Data from the Modern Web Python: Python Programming Course: Learn the Crash Course to Learning the Basics of Python (Python Programming, Python Programming Course, Python Beginners Course) Unsupervised Deep Learning in Python: Master Data Science and Machine Learning with Modern Neural Networks written in Python and Theano (Machine Learning in Python) Convolutional Neural Networks in Python: Master Data Science and Machine Learning with Modern Deep Learning in Python, Theano, and TensorFlow (Machine Learning in Python) Deep Learning in Python: Master Data Science and Machine Learning with Modern Neural Networks written in Python, Theano, and TensorFlow (Machine Learning in Python) Deep Learning in Python Prerequisites: Master Data Science and Machine Learning with Linear Regression and Logistic Regression in Python (Machine Learning in Python) Data Analytics: What Every Business Must Know About Big Data And Data Science (Data Analytics for Business, Predictive Analysis, Big Data) Data Analytics: Practical Data Analysis and Statistical Guide to Transform and Evolve Any Business. Leveraging the Power of Data Analytics, Data ... (Hacking Freedom and Data Driven) (Volume 2) Beginning Python Programming: Learn Python Programming in 7 Days: Treading on Python, Book 1 Python: Python Programming For Beginners - The Comprehensive Guide To Python Programming: Computer Programming, Computer Language, Computer Science Learn Python in One Day and Learn It Well: Python for Beginners with Hands-on Project. The only book you need to start coding in Python immediately Maya Python for Games and Film: A Complete Reference for Maya Python and the Maya Python API Python: Python Programming For Beginners - The Comprehensive Guide To Python Programming: Computer Programming, Computer Language, Computer Science (Machine Language) Deep Learning: Recurrent Neural Networks in Python: LSTM, GRU, and more RNN machine learning architectures in Python and Theano (Machine Learning in Python) Ruby: Programming, Master's Handbook: A TRUE Beginner's Guide! Problem Solving, Code, Data Science, Data Structures & Algorithms (Code like a PRO in ... web design, tech, perl, ajax, swift, python,) Java Programming: Master's Handbook: A TRUE Beginner's Guide! Problem Solving, Code, Data Science, Data Structures & Algorithms (Code like a PRO in ... web design, tech, perl, ajax, swift, python) Analytics: Data

Science, Data Analysis and Predictive Analytics for Business (Algorithms, Business Intelligence, Statistical Analysis, Decision Analysis, Business Analytics, Data Mining, Big Data) Python for Everybody: Exploring Data in Python 3 Scraping By: Wage Labor, Slavery, and Survival in Early Baltimore (Studies in Early American Economy and Society from the Library Company of Philadelphia) Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data (Data-Centric Systems and Applications)