MINISTRY OF EDUCATION AND SCIENCE OF UKRAINE

IVAN FRANKO NATIONAL UNIVERSITY OF LVIV

Department
of Foreign Languages
for Sciences

SUMMARY

Web Scraping with Python: Collecting More Data from the Modern Web

by Ryan Mitchell
O'Reilly Media, Sebastopol (USA), 2018, 1-66 pp.

Submitted by
the student of the Faculty of
Applied Mathematics and
Informatics
O. Kravets

Lviv-2022

One of the main things which are discussed in the passage I read is writing a simple automated program. You can query web servers, request data and parse it to extract the information you need.

In the first part the author focuses on web scraping mechanics: using Python to request information from a web server, performing basic handling of the server's response, and interacting with sites in an automated fashion. Part II explores a variety of more specific tools and applications to fit any web scraping scenario you're likely to encounter.

The first chapter starts with the basics of sending a GET request (a request to fetch, or "get," the content of a web page) to a web server for a specific page, reading the HTML output from that page, and doing some simple data extraction in order to isolate the content that you are looking for. In the second section, the author describes how to view complex HTML pages to get only the information you are looking for. In third chapter , readers can start look at real-world problems, with scrapers traversing multiple pages and even multiple sites. And in the fourth section Ryan Mitchell focuses primarily on web crawlers that collect a limited number of "types" of data (such as restaurant reviews, news articles or company profiles) from a variety of websites, and that store these data types as Python objects that read and write

To my mind, this book is designed to serve not only as an introduction to web scraping, but as a comprehensive guide to collecting, transforming, and using data from uncooperative sources. Although it uses the Python programming language and covers many Python basics, it should not be used as an introduction to the language. I think if you don't know any Python at all, this book might be a bit of a challenge.

I also want to note that technical books are often able to focus on a single language or technology, but web scraping is a relatively disparate subject, with practices that require the use of databases, web servers, HTTP, HTML, internet security, image processing, data science, and other tools. This book attempts to cover all of these, and other topics, from the perspective of 'data gathering.' It should not be used as a complete treatment of any of these subjects, but I believe they are covered in enough detail to get you started writing web scrapers.

I have been able to easily translate the lessons into practical projects. The author spells out the considerations for using each particular technique and the situations in which some might be more effective than others. The chapters are mutually reinforcing but you may find yourself skipping around if you already know what you want to get out of web scraping.

Therefore, to summarize, if you want to go deep into the web scraping domain, Web Scraping With Python by Ryan Mitchell is the best book I recommend. If you follow the techniques discussed in this book, you can scrap any website, the tools and examples included in this book allow you to easily automate repetitive tasks.