Table of Contents

Preface ix

Part I. Building Scrapers

1. Your First Web Scraper 3
Connecting 3
An Introduction to BeautifulSoup 6
Installing BeautifulSoup 6
Running BeautifulSoup 8
Connecting Reliably and Handling Exceptions 10

2. Advanced HTML Parsing 15
You Don't Always Need a Hammer 15
Another Serving of BeautifulSoup 16
find() and find_all() with BeautifulSoup 18
Other BeautifulSoup Objects 20
Navigating Trees 21
Regular Expressions 25
Regular Expressions and BeautifulSoup 29
Accessing Attributes 30
Lambda Expressions 31

3. Writing Web Crawlers 33
Traversing a Single Domain 33
Crawling an Entire Site 37
Collecting Data Across an Entire Site 40
Crawling Across the Internet 42

4. Web Crawling Models 49
Planning and Defining Objects 50
Dealing with Different Website Layouts 53

Structuring Crawlers 58
Crawling Sites Through Search 58
Crawling Sites Through Links 61
Crawling Multiple Page Types 64
Thinking About Web Crawler Models 65

5. Scrapy **67**
Installing Scrapy 67
Initializing a New Spider 68
Writing a Simple Scraper 69
Spidering with Rules 70
Creating Items 74
Outputting Items 76
The Item Pipeline 77
Logging with Scrapy 80
More Resources 81

6. Storing Data **83**
Media Files 83
Storing Data to CSV 86
MySQL 88
Installing MySQL 89
Some Basic Commands 91
Integrating with Python 94
Database Techniques and Good Practice 97
"Six Degrees" in MySQL 100
Email 103

Part II. Advanced Scraping

7. Reading Documents **107**
Document Encoding 107
Text 108
Text Encoding and the Global Internet 109
CSV 113
Reading CSV Files 113
PDF 115
Microsoft Word and .docx 117

8. Cleaning Your Dirty Data **121**
Cleaning in Code 121

Data Normalization 124
Cleaning After the Fact 126
OpenRefine 126

9. Reading and Writing Natural Languages..... 133
Summarizing Data 134
Markov Models 137
Six Degrees of Wikipedia: Conclusion 141
Natural Language Toolkit 144
Installation and Setup 144
Statistical Analysis with NLTK 145
Lexicographical Analysis with NLTK 147
Additional Resources 151

10. Crawling Through Forms and Logins..... 153
Python Requests Library 153
Submitting a Basic Form 154
Radio Buttons, Checkboxes, and Other Inputs 156
Submitting Files and Images 157
Handling Logins and Cookies 158
HTTP Basic Access Authentication 159
Other Form Problems 160

11. Scraping JavaScript..... 163
A Brief Introduction to JavaScript 164
Common JavaScript Libraries 165
Ajax and Dynamic HTML 167
Executing JavaScript in Python with Selenium 168
Additional Selenium Webdrivers 173
Handling Redirects 174
A Final Note on JavaScript 175

12. Crawling Through APIs..... 177
A Brief Introduction to APIs 177
HTTP Methods and APIs 179
More About API Responses 180
Parsing JSON 182
Undocumented APIs 183
Finding Undocumented APIs 184
Documenting Undocumented APIs 186
Finding and Documenting APIs Automatically 186
Combining APIs with Other Data Sources 189

More About APIs 192

13. Image Processing and Text Recognition..... 195
Overview of Libraries 196
Pillow 196
Tesseract 197
NumPy 199
Processing Well-Formatted Text 199
Adjusting Images Automatically 202
Scraping Text from Images on Websites 205
Reading CAPTCHAs and Training Tesseract 208
Training Tesseract 210
Retrieving CAPTCHAs and Submitting Solutions 213

14. Avoiding Scraping Traps..... 217
A Note on Ethics 217
Looking Like a Human 218
Adjust Your Headers 219
Handling Cookies with JavaScript 220
Timing Is Everything 222
Common Form Security Features 223
Hidden Input Field Values 223
Avoiding Honeypots 225
The Human Checklist 226

15. Testing Your Website with Scrapers..... 229
An Introduction to Testing 229
What Are Unit Tests? 230
Python unittest 230
Testing Wikipedia 232
Testing with Selenium 235
Interacting with the Site 235
unittest or Selenium? 238

16. Web Crawling in Parallel..... 241
Processes versus Threads 241
Multithreaded Crawling 242
Race Conditions and Queues 244
The threading Module 247
Multiprocess Crawling 249
Multiprocess Crawling 251
Communicating Between Processes 253

Multiprocess Crawling—Another Approach 255

17. Scraping Remotely..... 257
Why Use Remote Servers? 257
Avoiding IP Address Blocking 258
Portability and Extensibility 259
Tor 259
PySocks 261
Remote Hosting 261
Running from a Website-Hosting Account 262
Running from the Cloud 263
Additional Resources 264

18. The Legalities and Ethics of Web Scraping..... 265
Trademarks, Copyrights, Patents, Oh My! 265
Copyright Law 266
Trespass to Chattels 268
The Computer Fraud and Abuse Act 270
robots.txt and Terms of Service 271
Three Web Scrapers 274
eBay versus Bidder's Edge and Trespass to Chattels 274
United States v. Auernheimer and The Computer Fraud and Abuse Act 276
Field v. Google: Copyright and robots.txt 277
Moving Forward 278

Index..... 281