

---

# 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	63
Thinking About Web Crawler Models	65
<b>5. Scrapy.....</b>	<b>67</b>
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
<b>6. Storing Data.....</b>	<b>83</b>
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

<b>7. Reading Documents.....</b>	<b>107</b>
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
<b>8. Cleaning Your Dirty Data.....</b>	<b>121</b>
Cleaning in Code	121

Data Normalization	124
Cleaning After the Fact	126
OpenRefine	126
<b>9. Reading and Writing Natural Languages.....</b>	<b>133</b>
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
<b>10. Crawling Through Forms and Logins.....</b>	<b>153</b>
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
<b>11. Scraping JavaScript.....</b>	<b>163</b>
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	174
Handling Redirects	174
A Final Note on JavaScript	176
<b>12. Crawling Through APIs.....</b>	<b>177</b>
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
<b>13. Image Processing and Text Recognition.....</b>	<b>195</b>
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
<b>14. Avoiding Scraping Traps.....</b>	<b>217</b>
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 Honey pots	225
The Human Checklist	227
<b>15. Testing Your Website with Scrapers.....</b>	<b>229</b>
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?	239
<b>16. Web Crawling in Parallel.....</b>	<b>241</b>
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
<b>17. Scraping Remotely</b> .....	<b>257</b>
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
<b>18. The Legalities and Ethics of Web Scraping</b> .....	<b>265</b>
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
<b>Index</b> .....	<b>281</b>