Free Download: DS Career Guide

How to Learn Data Science & Machine Learning, Land a High-Paying Job, and Future-Proof Your Career

GET INSTANT ACCESS!

5 Tasty Python Web Scraping Libraries (https://elitedatascience.com/python-web-scraping-libraries)

Share (https://www.facebook.com/sharer.php?u=https%3A%2F%2Felitedatascience.com%2Fpython-web-scraping-libraries)

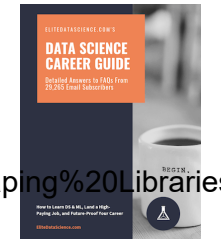
Google (https://plus.google.com/share?text=5+Tasty+Python+Web+Scraping+Libraries&url=https%3A%2F%2Felitedatascience.com%2Fpython-web-scraping-libraries)

out this site%20https://elitedatascience.com/python-web-scraping-libraries)
Linkedin (https://www.linkedin.com/shareArticle?trk=5+Tasty+Python+Web+Scraping+Libraries&url=https%3A%2F%2Felitedatascience.com%2Fpython-web-scraping-libraries)
Tweet (https://twitter.com/intent/tweet?text=5%20Tasty%20Python%20Web%20Scraping%20Libraries&url=https://elitedatascience.com/python-web-scraping-libraries)
(https://getpocket.com/save?title=5%20Tasty%20Python%20Web%20Scraping%20Libraries&url=https://elitedatascience.com/python-web-scraping-libraries)
(https://www.linkedin.com/shareArticle?trk=5%20Tasty%20Python%20Web%20Scraping%20Libraries&url=https://elitedatascience.com/python-web-scraping-libraries)
(http://www.facebook.com/sharer.php?u=https://elitedatascience.com/python-web-scraping-libraries)
(https://plus.google.com/share?text=5%20Tasty%20Python%20Web%20Scraping%20Libraries&url=https://elitedatascience.com/python-web-scraping-libraries)
(https://twitter.com/intent/tweet?text=5+Tasty+Python+Web+Scraping+Libraries&url=https://elitedatascience.com/python-web-scraping-libraries)
(http://service.weibo.com/share/share.php?url=https://elitedatascience.com/python-web-scraping-libraries&title=5%20Tasty%20Python%20Web%20Scraping%20Libraries)
(mailto:?subject=Check out this site%20&body=Check

Web scraping is a common and effective way of collecting data for projects and for work. In this guide, we'll be touring the essential stack of Python web scraping libraries.

Why only 5 libraries?

There are dozens of packages for web scraping out there… but you only need a handful to be able to scrape almost any site. This is an opinionated guide. We've decided to feature the 5 Python libraries for web scraping that we love most. Together, they cover all the important bases, and they are well-documented.



Free: Data Science Career Guide

Learn how to **land a high-paying job in data science** and **future-proof your career** with the most efficient **roadmap** to learning DS & ML for busy professionals.

First Name

Email

Send My Download

Do I need to learn every library below?

No, but everyone will need Requests, because it's how you communicate with websites. The rest depend on your use case. Here's a rule of thumb:

- You should learn at least one of BeautifulSoup or lxml. Pick depending on which is more intuitive for you (more on this below).
- Learn Selenium if you need to scrape sites with data tucked away by JavaScript.
- Learn Scrapy if you need to build a real spider or web-crawler, instead of just scraping a few pages here and there.

Why are they tasty?

Because they are yummy! So without further ado…

- The Farm: Requests
- The Stew: Beautiful Soup 4
- The Salad: lxml
- The Restaurant: Selenium
- The Chef: Scrapy

The Farm: Requests

The Requests library is vital to add to your data science toolkit. It's a simple yet powerful HTTP library, which means you can use it to access web pages.

We call it The Farm because you'll be using it to get the raw ingredients (i.e. raw HTML) for your dishes (i.e. usable data).

Its simplicity is definitely its greatest strength. It's so easy use that you could jump right in without reading documentation.

For example, if you want to pull down the contents of a page, it's as easy as:

```python
import requests
page = requests.get('http://examplesite.com')
contents = page.content
```

But that's not all that Requests can do. It can access API's, post to forms, and much more.

Plus, it's got character… It's the only library that calls itself Non-GMO, organic, and grass-fed. You gotta love that.

Resources

- Requests Quickstart Guide (http://docs.python-requests.org/en/master/user/quickstart/) – Official documentation. Covers practical topics like passing parameters, handling responses, and configuring headers.

The Stew: Beautiful Soup 4

After you have your ingredients, now what? Now you make them into a stew… a beautiful stew.

Beautiful Soup (BS4) is a parsing library that can use different **parsers**. A parser is simply a program that can extract data from HTML and XML documents.

Beautiful Soup's default parser comes from Python's standard library. It's flexible and forgiving, but a little slow. The good news is that you can swap out its parser with a faster one if you need the speed.

One advantage of BS4 is its ability to automatically detect encodings. This allows it to gracefully handle HTML documents with special characters.

In addition, BS4 can help you navigate a parsed document and find what you need. This makes it quick and painless to build common applications. For example, if you wanted to find all the links in the web page we pulled down earlier, it's only a few lines:

```
Python
from bs4 import BeautifulSoup
soup = BeautifulSoup(contents, 'html.parser')
soup.find_all('a')
```

site%20body=Check
out
this
site%20https://elitedatascience.com/python-
web-
scraping-
libraries)

This charming simplicity has made it one of the most beloved Python web scraping libraries!

Resources

- Beautiful Soup Documentation (https://www.crummy.com/software/BeautifulSoup/bs4/doc/) – Includes convenient quickstart guide.
- Really Short Example (https://gist.github.com/bradmontgomery/1872970) – Short example of using Beautiful Soup and Requests together.

(https://getpocket.com/save?
title=5%20Tasty%20Python%20Web%20Scraping%20Libraries&url=https://elitedatascience.com/python-
web-
scraping-
libraries)
(https://www.linkedin.com/shareArticle?
trk=5%20Tasty%20Python%20Web%20Scraping%20Libraries&url=https://elitedatascience.com/python-
web-
scraping-
libraries)
<<
(http://www.facebook.com/sharer.php?
u=https://elitedatascience.com/python-
web-
scraping-
libraries)
(https://plus.google.com/share?
text=5%20Tasty%20Python%20Web%20Scraping%20Libraries&url=https://elitedatascience.com/python-
web-
scraping-
libraries)
(https://twitter.com/intent/tweet?
text=5+Tasty+Python+Web+Scraping+Libraries&url=https://elitedatascience.com/python-
web-
scraping-
libraries)
(http://service.weibo.com/share/share.php?
url=https://elitedatascience.com/python-
web-
scraping-
libraries&title=5%20Tasty%20Python%20Web%20Scraping%20Libraries)
(mailto:?
subject=Check
out
this
site%20&body=Check

The Salad: lxml

lxml is a high-performance, production-quality HTML and XML parsing library. We call it The Salad because you can rely on it to be good for you, no matter which diet you're following.

Among all the Python web scraping libraries, we've enjoyed using lxml the most. It's straightforward, fast, and feature-rich.

Even so, it's quite easy to pick up if you have experience with either XPaths or CSS. Its raw speed and power has also helped it become widely adopted in the industry.

Beautiful Soup vs lxml

Historically, the rule of thumb was:

- If you need speed, go for lxml.
- If you need to handle messy documents, choose Beautiful Soup.

Yet, this distinction no longer holds. Beautiful Soup now supports using the lxml parser, and vice-versa. It's also pretty easy to learn the other once you've learned one.

So to start, we recommend trying both and picking the one that feels more intuitive for you. We prefer lxml, but many swear by Beautiful Soup.

site%20body=Check out this site%20https://elitedatascience.com/python-web-scraping-libraries)

Resources

- lxml Documentation (http://lxml.de/index.html#introduction) – Official documentation.
- HTML Scraping with lxml and Requests (http://docs.python-guide.org/en/latest/scenarios/scrape/) – Short and sweet tutorial on pulling a webpage with Requests and then using XPath selectors to mine the desired data. This is more beginner-friendly than the official documentation.

The Restaurant: Selenium

Sometimes, you do need to go to a restaurant to eat certain dishes. The farm is great, but you can't find everything there.

Likewise, sometimes the Requests library is not enough to scrape a website. Some sites out there use JavaScript to serve content. For example, they might wait until you scroll down on the page or click a button before loading certain content.

Other sites may require you to click through forms before seeing their content. Or select options from a dropdown. Or perform a tribal rain dance…

For these sites, you'll need something more powerful. You'll need Selenium (which can handle everything except tribal rain dancing).

Selenium is a tool that automates browsers, also known as a web-driver. With it, you can actually open a Google Chrome window, visit a site, and click on links. Pretty cool, right?

It also comes with Python bindings for controlling it right from your application. This makes it a breeze to integrate with your chosen parsing library.

Resources

- Selenium with Python (http://selenium-python.readthedocs.io/) – Documentation for Selenium's Python bindings.
- Webscraping with Selenium (http://thiagomarzagao.com/2013/11/12/webscraping-with-selenium-part-1/) –

site%20body=Check
out
this
site%20https://elitedatascience.com/python-
web-
scraping-
libraries)

- Excellent, thorough 3-part tutorial for scraping websites with Selenium.
- Scraping Hotel Prices (https://www.scrapehero.com/tutorial-web-scraping-hotel-prices-using-selenium-and-python/) – Code snippet for scraping hotel prices using Selenium and lxml.

The Chef: Scrapy

(https://getpocket.com/save?
title=5%20Tasty%20Python%20Web%20Scraping%20Libraries&url=https://elitedatascience.com/python-
web-
scraping-
libraries)
(https://www.linkedin.com/shareArticle?
trk=5%20Tasty%20Python%20Web%20Scraping%20Libraries&url=https://elitedatascience.com/python-
web-
scraping-
libraries)
<<
(http://www.facebook.com/sharer.php?
u=https://elitedatascience.com/python-
web-
scraping-
libraries)
(https://plus.google.com/share?
text=5%20Tasty%20Python%20Web%20Scraping%20Libraries&url=https://elitedatascience.com/python-
web-
scraping-
libraries)
(http://stackoverflow.com/questions/17975471/selenium-with-scrapy-for-
dynamic-page)).
(https://twitter.com/intent/tweet?
text=5+Tasty+Python+Web+Scraping+Libraries&url=https://elitedatascience.com/python-
web-
scraping-
libraries)
(http://service.weibo.com/share/share.php?
url=https://elitedatascience.com/python-
web-
scraping-
libraries&title=5%20Tasty%20Python%20Web%20Scraping%20Libraries)
(mailto:?
subject=Check
out
this
site%20&body=Check

Ok, we covered a lot just now. You've got Requests and Selenium for fetching HTML/XML from web pages. Then, you can use Beautiful Soup or lxml to parse it into useful data.

But what if you need more than that? What if you need a complete spider that can crawl through entire websites in a systematic way?

Introducing: Scrapy! Scrapy is technically not even a library… it's a complete web scraping framework. That means you can use it to manage requests, preserve user sessions, follow redirects, and handle output pipelines.

It also means you can swap out individual modules with other Python web scraping libraries. For instance, if you need to insert Selenium for scraping dynamic web pages, you can do that (see example

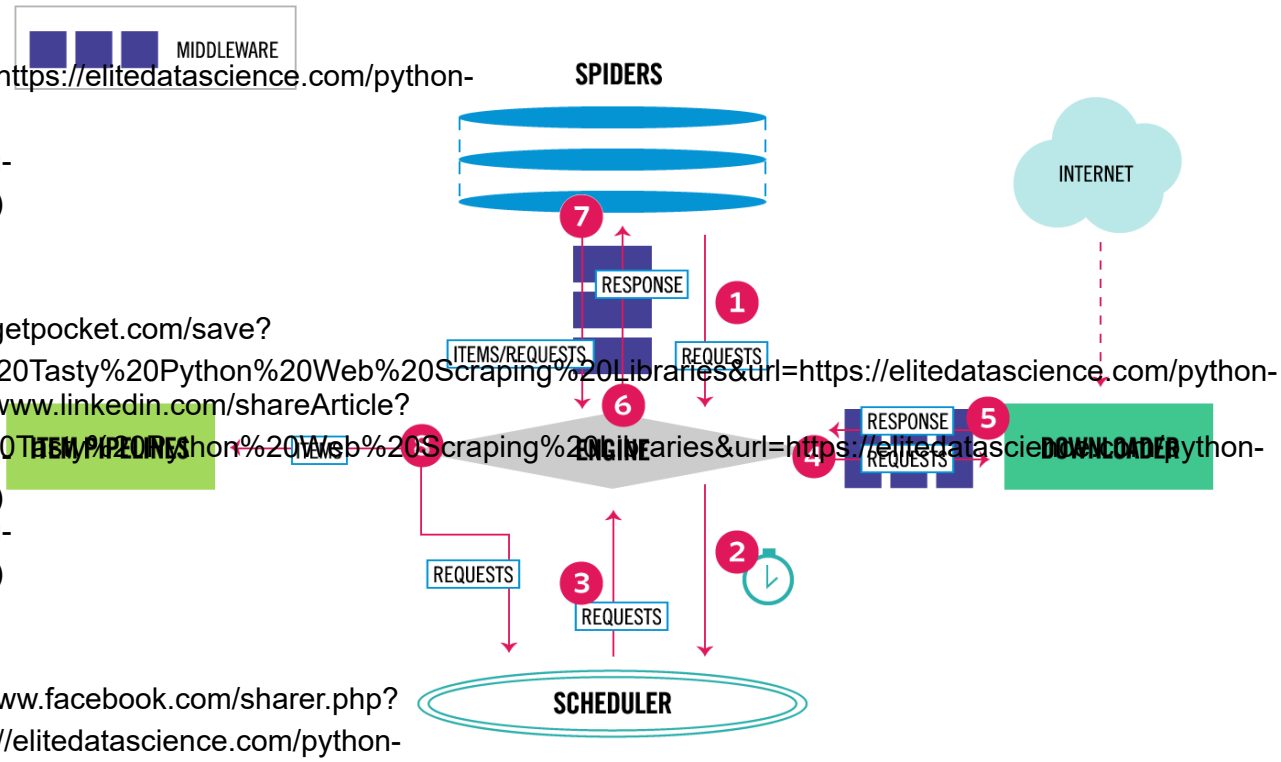
site%20body=Check
out
this
site%20https://elitedatascience.com/python-
web-
scraping-
libraries)

(https://getpocket.com/save?
title=5%20Tasty%20Python%20Web%20Scraping%20Libraries&url=https://elitedatascience.com/python-
web-
scraping-
libraries)

(https://www.linkedin.com/shareArticle?
trk=5%20Tasty%20Python%20Web%20Scraping%20Libraries&url=https://elitedatascience.com/python-
web-
scraping-
libraries)

<<

(http://www.facebook.com/sharer.php?
u=https://elitedatascience.com/python-
web-
scraping-
libraries)



Scrapy architecture, image borrowed from official documentation (https://doc.scrapy.org/en/latest/topics/architecture.html)

So if you need to reuse your crawler, scale it, manage complex data pipelines, or cook up some other sophisticated spider, then Scrapy was made for you.

(https://plus.google.com/share?
text=5%20Tasty%20Python%20Web%20Scraping%20Libraries&url=https://elitedatascience.com/python-
web-
scraping-
libraries)

Resources

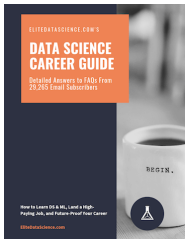
- Scrapy Documentation (https://scrapy.org/) - Official site with links to many other resources.
- Extracting data from websites with Scrapy (https://medium.com/@kaismh/extracting-data-from-websites-using-scrapy-e1e1e357651a#.sw7c9ycio) - Detailed tutorial for scraping an e-commerce site using Scrapy.
- Scrapinghub (https://scrapinghub.com/) - Cloud-based crawling service by the creators of Scrapy. The first cloud unit is free.

(https://twitter.com/intent/tweet?
text=5+Tasty+Python+Web+Scraping+Libraries&url=https://elitedatascience.com/python-
web-
scraping-
libraries)

(http://service.weibo.com/share/share.php?
url=https://elitedatascience.com/python-
web-
scraping-
libraries&title=5%20Tasty%20Python%20Web%20Scraping%20Libraries)

(mailto:?
subject=Check
out
this
site%20&body=Check

site%20&body=Check out this site%20https://elitedatascience.com/python-web-scraping-libraries)



Free: Data Science Career Guide

How to Learn Data Science & Machine Learning, Land a High-Paying Job, and Future-Proof Your Career

First Name

Email

Send My Download

(https://getpocket.com/save?title=5%20Tasty%20Python%20Web%20Scraping%20Libraries&url=https://elitedatascience.com/python-web-scraping-libraries)

(https://www.linkedin.com/shareArticle?trk=5%20Tasty%20Python%20Web%20Scraping%20Libraries&url=https://elitedatascience.com/python-web-scraping-libraries)

<<

(http://www.facebook.com/sharer.php?u=https://elitedatascience.com/python-web-scraping-libraries)

(https://plus.google.com/share?text=5%20Tasty%20Python%20Web%20Scraping%20Libraries&url=https://elitedatascience.com/python-web-scraping-libraries)

(https://twitter.com/intent/tweet?text=5+Tasty+Python+Web+Scraping+Libraries&url=https://elitedatascience.com/python-web-scraping-libraries)

(http://service.weibo.com/share/share.php?url=https://elitedatascience.com/python-web-scraping-libraries&title=5%20Tasty%20Python%20Web%20Scraping%20Libraries)

(mailto:?subject=Check out this site%20&body=Check

Share (https://www.facebook.com/sharer.php?u=https%3A%2F%2Felitedatascience.com%2Fpython-web-scraping-libraries)

Google (https://plus.google.com/share?text=5+Tasty+Python+Web+Scraping+Libraries&url=https%3A%2F%2Felitedatascience.com%2Fpython-web-scraping-libraries)

Linkedin (https://www.linkedin.com/shareArticle?trk=5+Tasty+Python+Web+Scraping+Libraries&url=https%3A%2F%2Felitedatascience.com%2Fpython-web-scraping-libraries)

Tweet (https://twitter.com/intent/tweet?text=5%20Tasty%20Python%20Web%20Scraping%20Libraries&url=https://elitedatascience.com/python-web-scraping-libraries)

« Previous Post

How to Learn Statistics for Data Science, The Self-Starter Way (https://elitedatascience.com/learn-statistics-for-data-science)

Next Post »

5 Heroic Python NLP Libraries (https://elitedatascience.com/python-nlp-libraries)

site%20&body=Check
out
this
site%20https://elitedatascience.com/python-
web-
scraping-
libraries)

Free Download: DS Career Guide

How to Learn Data Science & Machine Learning, Land a High-Paying Job, and Future-Proof Your Career

GET INSTANT ACCESS!

(https://getpocket.com/save?
title=5%20Tasty%20Python%20Web%20Scraping%20Libraries&url=https://elitedatascience.com/python-
web-
scraping-
libraries)

(https://www.linkedin.com/shareArticle?
trk=5%20Tasty%20Python%20Web%20Scraping%20Libraries&url=https://elitedatascience.com/python-
web-
scraping-
libraries)

LINKS

<<

Start Here (https://elitedatascience.com/?page_id=1584)

Homepage (/)

Login (https://elitedatascience.com/login)

(http://www.facebook.com/sharer.php?
u=https://elitedatascience.com/python-
web-
scraping-
libraries)

ARTICLES

Guides (https://elitedatascience.com/category/guides)

Concept Explainers (https://elitedatascience.com/category/explainers)

Code Tutorials (https://elitedatascience.com/category/tutorials)

(https://plus.google.com/share?
text=5%20Tasty%20Python%20Web%20Scraping%20Libraries&url=https://elitedatascience.com/python-
web-
scraping-
libraries)

Career Help (https://elitedatascience.com/category/career)

Tools & Resources (https://elitedatascience.com/category/resources)

SHARE

(https://twitter.com/intent/tweet?
text=5+Tasty+Python+Web+Scraping+Libraries&url=https://elitedatascience.com/python-
web-
scraping-
libraries)

Share (https://www.facebook.com/sharer.php?u=https%3A%2F%2Felitedatascience.com)

Google (https://plus.google.com/share?
text=5+Tasty+Python+Web+Scraping+Libraries&url=https%3A%2F%2Felitedatascience.com)

Linkedin (https://www.linkedin.com/shareArticle?
trk=5+Tasty+Python+Web+Scraping+Libraries&url=https%3A%2F%2Felitedatascience.com)

Tweet (https://twitter.com/intent/tweet?
text=5%20Tasty%20Python%20Web%20Scraping%20Libraries&url=https://elitedatascience.com)

(http://service.weibo.com/share/share.php?
url=https://elitedatascience.com/python-
web-
scraping-
libraries&title=5%20Tasty%20Python%20Web%20Scraping%20Libraries)

Copyright © 2016-2019 · EliteDataScience.com · All Rights Reserved · Terms
(https://elitedatascience.com/terms-of-service) · Privacy (https://elitedatascience.com/privacy-policy)

(mailto:?
subject=Check
out
this
site%20&body=Check