# Python scraping

# Use cases

1. PADI: iterate over countries and save dive locations. Run once per year
2. Funda: iterate over cities, zip codes or provinces. Run once per month?

|  |  |  |  |
| --- | --- | --- | --- |
| **Module** | **Docstrings** | **Test** | **Jupyter** |
| Dicts | OK | OK | Later |
| Lists | OK | OK | Later |
| Soups | OK | Get\_table | OK |
| Drivers | OK | OK | OK |
| Times | OK | OK | OK |
| Strings | OK | OK | Later |
| Urls | OK | OK | OK |
| Utils | OK | OK (most functions) | Later |

|  |  |  |  |
| --- | --- | --- | --- |
| **Module** | **Docstrings** | **Test** | **Jupyter** |
| website | Later |  |  |
| config | OK | OK | OK |

<https://blog.devgenius.io/web-scraping-with-beautiful-soup-equality-copies-and-parsing-part-of-a-document-570a8531d1b8>

# Fingerprinting

Fingerprinting can be used to identify a user. Deviceinfo.me is recommended by protonmail.com

<https://www.deviceinfo.me/>

Screen size

Last clicked position in page: X:655  ,  Y:7176  (pixels)

Browser Window Size:

Outer: 1936 x 1056 (pixels)

Inner: 1920 x 910 (pixels)

# Drivers

6-12-2020

Issue: funda recognizes test software?

<https://www.jondelamotte.com/blog/firefox-stuck-in-remote-control-mode/>

<https://firefox-source-docs.mozilla.org/testing/marionette/Intro.html>

Tutorial

<https://stackoverflow.com/questions/58873022/how-to-make-selenium-script-undetectable-using-geckodriver-and-firefox-through-p>

<https://intoli.com/blog/making-chrome-headless-undetectable/>

# API Tutorials

## Intermediate API tutorial

<https://www.dataquest.io/blog/last-fm-api-python/>

## Javascript/API

<https://towardsdatascience.com/data-science-skills-web-scraping-javascript-using-python-97a29738353f>

## Public APIs

<https://github.com/public-apis/public-apis>

# Scraping tutorial

<https://realpython.com/beautiful-soup-web-scraper-python/>

## Part 1: Inspect Your Data Source

Decipher the Information in URLs

Inspect the Site Using Developer Tools

In Chrome, you can open up the developer tools through the menu View → Developer → Developer Tools. You can also access them by right-clicking on the page and selecting the Inspect option, or by using a keyboard shortcut.

## Part 2: Scrape HTML Content From a Page

**Static pages**

Use requests

import requests

URL = 'https://www.monster.com/jobs/search/?q=Software-Developer&where=Australia'

page = requests.get(URL)

**Dynamic pages**

Use requests-html (not covered) or selenium.

For example, requests-html is a project created by the author of the requests library that allows you to easily render JavaScript using syntax that’s similar to the syntax in requests. It also includes capabilities for parsing the data by using Beautiful Soup under the hood.

## Part 3: Parse HTML Code With Beautiful Soup

Beautiful Soup is a Python library for parsing structured data. It allows you to interact with HTML in a similar way to how you would interact with a web page using developer tools.

Beautiful Soup is a Python library for pulling data out of HTML and XML files.

## Webscraping references

**Basics**

<https://realpython.com/beautiful-soup-web-scraper-python/#part-3-parse-html-code-with-beautiful-soup>

<https://www.dataquest.io/blog/web-scraping-tutorial-python/>

<https://realpython.com/python-requests/>

**Advanced**

<https://www.digitalocean.com/community/tutorials/how-to-scrape-web-pages-with-beautiful-soup-and-python-3>

<https://blog.hartleybrody.com/web-scraping-cheat-sheet/>

<https://blog.hartleybrody.com/web-scraping/>

<https://www.datacamp.com/community/tutorials/web-scraping-using-python>

<http://www.compjour.org/warmups/govt-text-releases/intro-to-bs4-lxml-parsing-wh-press-briefings/>

**Funda**

<https://github.com/khpeek/funda-scraper>

<https://www.bartbarnard.nl/blog/?p=599>

**Package**

<https://www.scrapinghub.com/learn-scrapy/>

<https://www.scraping-bot.io/how-to-scrape-a-website-without-getting-blocked/>

# Network

## VPNS

**Nord VPN**

<https://gist.github.com/JamoCA/eedaf4f7cce1cb0aeb5c1039af35f0b7>

## Tor and Python

Less relevant: Tor is too slow.

<https://medium.com/@jasonrigden/using-tor-with-the-python-request-library-79015b2606cb>

<https://stackoverflow.com/questions/30286293/make-requests-using-python-over-tor>

## Proxies

Instead of connecting directly to a server that can fulfill a requested resource, such as a file or web page for example, the client directs the request to the proxy server, which evaluates the request and performs the required network transactions.

Oct 2021: use a Proxy via a Chrome Extension

<https://hide.me/nl/proxy>

How to work with extensions in Selenium

<https://stackoverflow.com/questions/34222412/load-chrome-extension-using-selenium>

Note: first check credibility of hide.me. The alternative is to Use e.g. ProtonVPN on the PC instead of the browser.

<https://www.scrapehero.com/how-to-rotate-proxies-and-ip-addresses-using-python-3/>

## SOCKS

SOCKS is an Internet protocol that exchanges network packets between a client and server through a proxy server. SOCKS5 optionally provides authentication so only authorized users may access a server. Practically, a SOCKS server proxies TCP connections to an arbitrary IP address, and provides a means for UDP packets to be forwarded.