All Your Data Are Belong To Us:

Web Scraping with Python

Christopher Byrd

What is web scraping?

Extracting data from websites using the Hypertext Transfer Protocol (HTTP)

Legality of Web Scraping

Note: I am not a lawyer, and this is neither legal advice nor encouragement of illegal behaviour

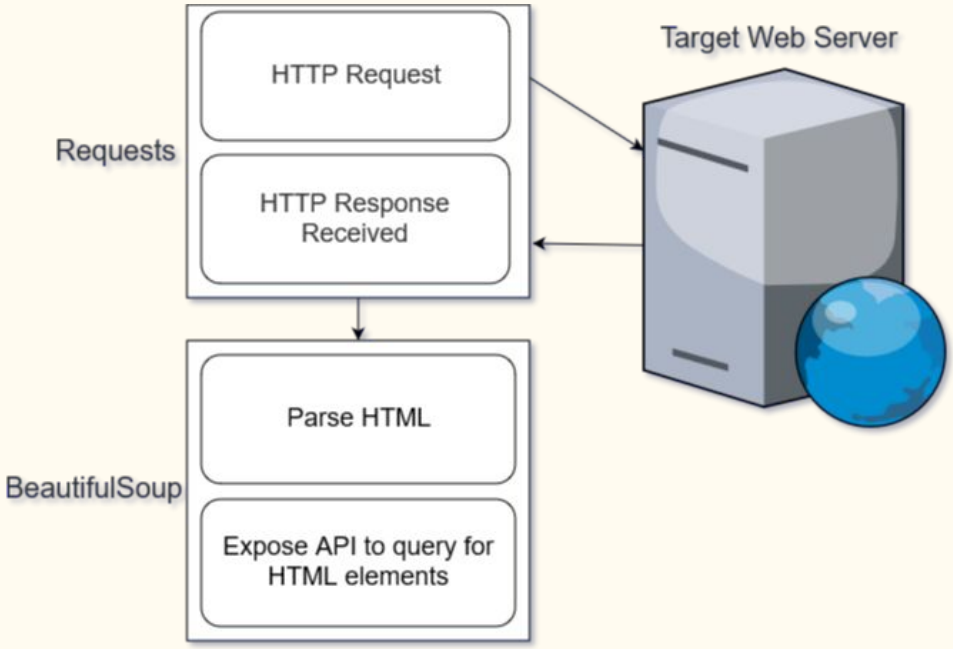
- In **Nguyen v. Barnes & Noble, Inc.**, it was determined that a Browsewrap Agreement (Terms of Use at the bottom of a page) are not enforceable without “sufficient notice”
- With that said, site owners have used the following three claims against scrapers:
  - Copyright infringement
  - Violation of Computer Fraud & Abuse Act
  - “Trespass to Chattel”
- **Scrape at your own risk**

What's needed?

- Python 3 (we'll be using 3.6)
- Requests module
- BeautifulSoup module
- PIP 3 (to install the above modules)
- Text editor of your choosing

Procedure

1. Call the appropriate requests function on the desired URL
    a. requests.get() or requests.post()
2. Verify the response status code is desired (typically 200 OK)
3. Pass response content to BeautifulSoup
4. Use BeautifulSoup to query page content as desired



Okay...
But what do those steps look like in code?

```
#!/usr/bin/env python3

import requests
import sys
from bs4 import BeautifulSoup


page = requests.get("http://www.website.com")

if page.status_code != 200:
    print("{} status code received. Exiting.".format(page.status_code))
    sys.exit(1)

soup = BeautifulSoup(page.content, 'html.parser')
```

About today's demo...

The code from today's demo will be available at:

https://github.com/ChrisByrd14/AllYourData

The demo will consist of three parts:

1. Python package installation, and basic web scraping
2. Saving scraped data to a CSV file
3. Storing scraped data in a SQLite 3 database

The site we'll be scraping today is http://books.toscrape.com/

Let’s begin...