Web Scraping

Annie, Nakshatra, Prashanth, Steve

June 13, 2023

Introduction

- Web scraping is the process of extracting data from websites
- It involves automatically navigating web pages, accessing and retrieving data, and saving it for further analysis
- Web scraping is commonly used for various purposes, such as data collection, market research, competitor analysis, and price monitoring.
- It enables automation and efficiency in data extraction from multiple sources.
- Web scraping can provide valuable insights and data for businesses, researchers, and developers
- However, it's important to be aware of legal and ethical considerations when performing web scraping to respect website policies and terms of service

Beautiful Soup

- Powerful Python library for web scraping
- Simplifies parsing HTML and XML documents
- Provides methods for navigating and searching parsed data
- Popular choice for web scraping due to its ease of use
- Handles malformed HTML and adapts to different parsing requirements

Installation

- pip install beautifulsoup4
- pip3 for mac

HTML 101

HTML Page Structure

```
<!DOCTYPE html>        ← Tells version of HTML
<html>                 ← HTML Root Element

<head>                 ← Used to contain page HTML metadata
  <title>Page Title</title>   ← Title of HTML page
</head>

<body>                 ← Hold content of HTML
  <h2>Heading Content</h2>    ← HTML headling tag
  <p>Paragraph Content</p>    ← HTML paragraph tag
</body>

</html>
```

```python
from bs4 import BeautifulSoup
soup = BeautifulSoup(html_doc, 'html.parser')

print(soup.prettify())
# <html>
#  <head>
#   <title>
#    The Dormouse's story
#   </title>
#  </head>
#  <body>
#   <p class="title">
#    <b>
#     The Dormouse's story
#    </b>
#   </p>
#   <p class="story">
#    Once upon a time there were three little sisters; and their names were
#    <a class="sister" href="http://example.com/elsie" id="link1">
#     Elsie
#    </a>
#    ,
#    <a class="sister" href="http://example.com/lacie" id="link2">
#     Lacie
#    </a>
#    and
#    <a class="sister" href="http://example.com/tillie" id="link3">
#     Tillie
#    </a>
#    ; and they lived at the bottom of a well.
#   </p>
#   <p class="story">
#    ...
#   </p>
#  </body>
# </html>
```

Playing around

```
soup.title
# <title>The Dormouse's story</title>

soup.title.name
# u'title'

soup.title.string
# u'The Dormouse's story'

soup.title.parent.name
# u'head'

soup.p
# <p class="title"><b>The Dormouse's story</b></p>

soup.p['class']
# u'title'

soup.a
# <a class="sister" href="http://example.com/elsie" id="link1">Elsie</a>

soup.find_all('a')
# [<a class="sister" href="http://example.com/elsie" id="link1">Elsie</a>,
#  <a class="sister" href="http://example.com/lacie" id="link2">Lacie</a>,
#  <a class="sister" href="http://example.com/tillie" id="link3">Tillie</a>]

soup.find(id="link3")
# <a class="sister" href="http://example.com/tillie" id="link3">Tillie</a>
```

Common Things

```
for link in soup.find_all('a'):
    print(link.get('href'))
# http://example.com/elsie
# http://example.com/lacie
# http://example.com/tillie
```

```
print(soup.get_text())
# The Dormouse's story
#
# The Dormouse's story
#
# Once upon a time there were three little sisters; and their names were
# Elsie,
# Lacie and
# Tillie;
# and they lived at the bottom of a well.
#
# ...
```

Scraping the Web

Virtual Environments

Creating Virtual Environment

```
virtualenv beautifulsoup
```

Activating Virtual Environment

```
source beautifulsoup/bin/activate
```

Deactivating Virtual Environment

```
deactivate
```

Deleting Virtual Environment

```
sudo rm -rf beautifulsoup
```

Scraping Technique

Requesting and Loading the webpage

Parsing the content

Store the data

Loading Web Page

Fake Python

Fake Jobs for Your Web Scraping Journey

Senior Python Developer
Payne, Roberts and Davis

Stewartbury, AA

2021-04-08

Learn | Apply

Energy engineer
Vasquez-Davidson

Christopherville, AA

2021-04-08

Learn | Apply

Legal executive
Jackson, Chambers and Levy

Port Ericaburgh, AA

2021-04-08

Learn | Apply

Fitness centre manager
Savage-Bradley

East Seanview, AP

2021-04-08

Learn | Apply

Parsing Content

```
<div class="card">
  <div class="card-content">
    <div class="media">
      <div class="media-left">
        <figure class="image is-48x48">
          <img
            src="https://files.realpython.com/media/real-python-logo-thumbnail.7f0db70c2ed2.jpg"
            alt="Real Python Logo"
          />
        </figure>
      </div>
    </div>

    <div class="content">
      <p class="location">Stewartbury, AA</p>
      <p class="is-small has-text-grey">
        <time datetime="2021-04-08">2021-04-08</time>
      </p>
    </div>
    <footer class="card-footer">
      <a
        href="https://www.realpython.com"
        target="_blank"
        class="card-footer-item"
        >Learn</a
      >
    </footer>
  </div>
</div>
```

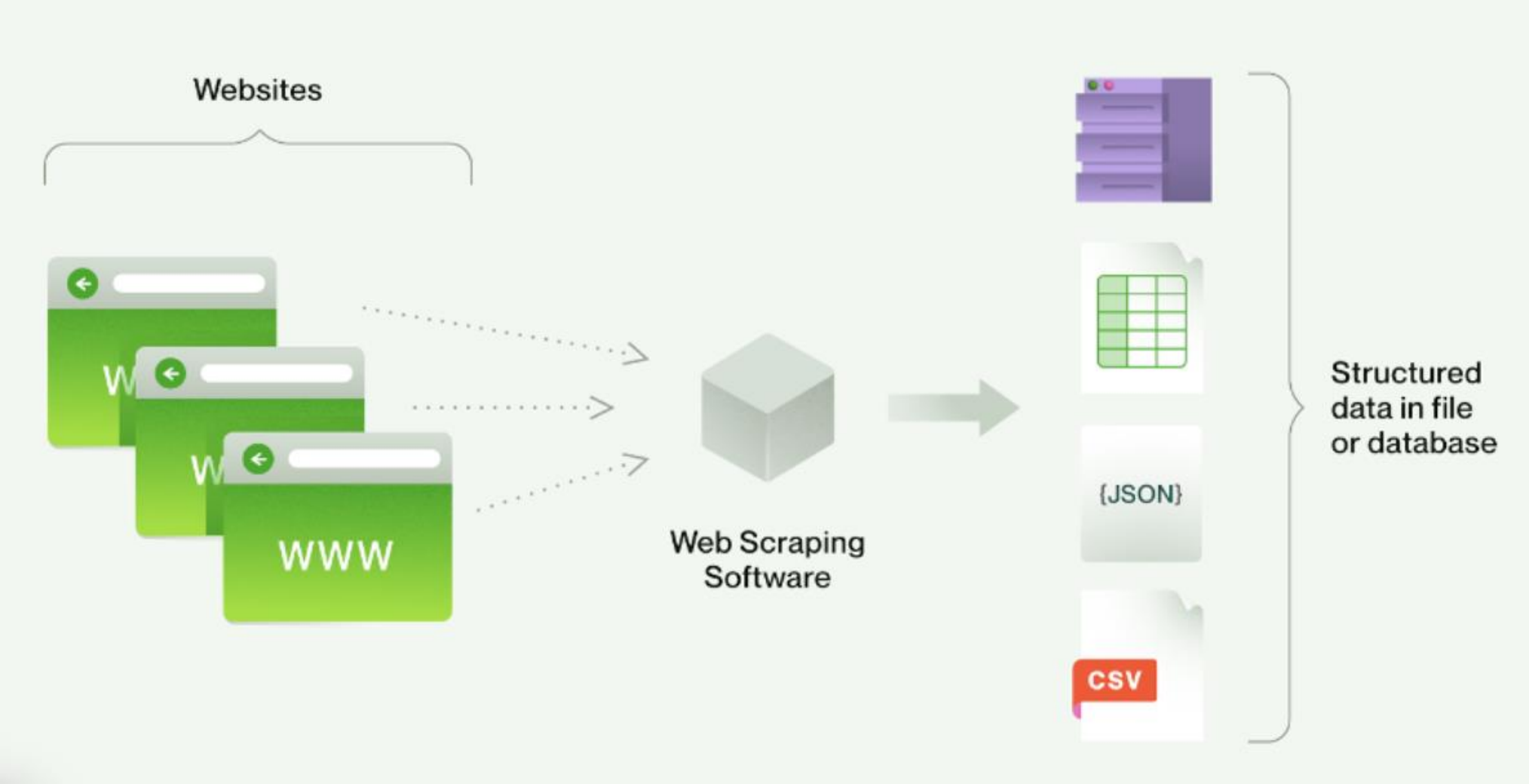
Processing Data

```
Senior Python Developer
Payne, Roberts and Davis
Stewartbury, AA

Energy engineer
Vasquez-Davidson
Christopherville, AA

Legal executive
Jackson, Chambers and Levy
Port Ericaburgh, AA
```

Storing Data



Libraries

Requesting and Loading the webpage – Requests, urllib, httplib

Parsing the content – Beautiful Soup, re, Scrapy

Store the data – SQLite, csv, JSON

Demo