McGill **Artificial Intelligence** Society

Intro to Web Scraping

Fall 2019 Workshop 3

By Rick Wu and Cheng Lin

Agenda

1. What is web scraping?
2. Introduction to HTML, web dev terminology
3. Scraping with **requests** and **BeautifulSoup4**
4. Statically vs. dynamically loaded sites
5. Scraping with **selenium**
6. Recap

What is web scraping?

Web Scraping

There are two main ways to collect data from the internet:

1. Web APIs
    a. Interface for programmers to easily retrieve data from website's server
2. Web scraping
    a. Extracting data by parsing the HTML files of websites

Intro to HTML, etc.

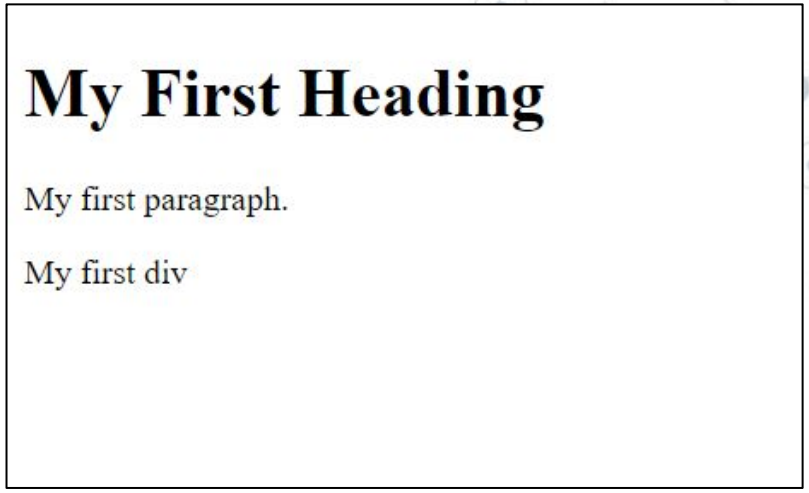
Intro to HTML

Hypertext Markup Language (HTML) is the standard markup language for web sites

- Dictates *what* data is displayed
- Can be viewed with **Inspect Element** in Google Chrome

index.html

```
1   <!DOCTYPE html>
2   <html>
3   <body>
4
5   <h1>My First Heading</h1>
6
7   <p>My first paragraph.</p>
8
9   <div id="my_div">My first div</div>
10
11  </body>
12  </html>
```



Web browsers interpret HTML documents as a **tree** of elements (HTML Document Object Model)

- Elements are separated in HTML by **tags**
- An element is a **parent** of the elements embedded within it
- An element is a **child** of the element it is embedded in

Scraping with requests and BS4

Requests Python Package

- Python wrapper for HTTP requests
- In the context of scraping, **requests** lets us retrieve the HTML of a URL

BeautifulSoup4 Python Package

- Python package for HTML or XML parsing
- In the context of scraping, **BeautifulSoup4** lets us iterate, search, and traverse the HTML DOM tree

Coding Example

(https://colab.research.google.com/drive/1zqHSAM_IIveQonDOT4M7Sgnl3GhbMTXT)

Statically vs. dynamically loaded sites

What happens if our site has Javascript?

- The **requests** package isn't a web browser, so it doesn't interpret (aka run) any of the Javascript embedded within the HTML
- This means that on dynamically updated sites, the data you see in a browser is not necessarily what you get when using **requests**
  - You get exactly what you see in "view page source"

Scraping with Selenium

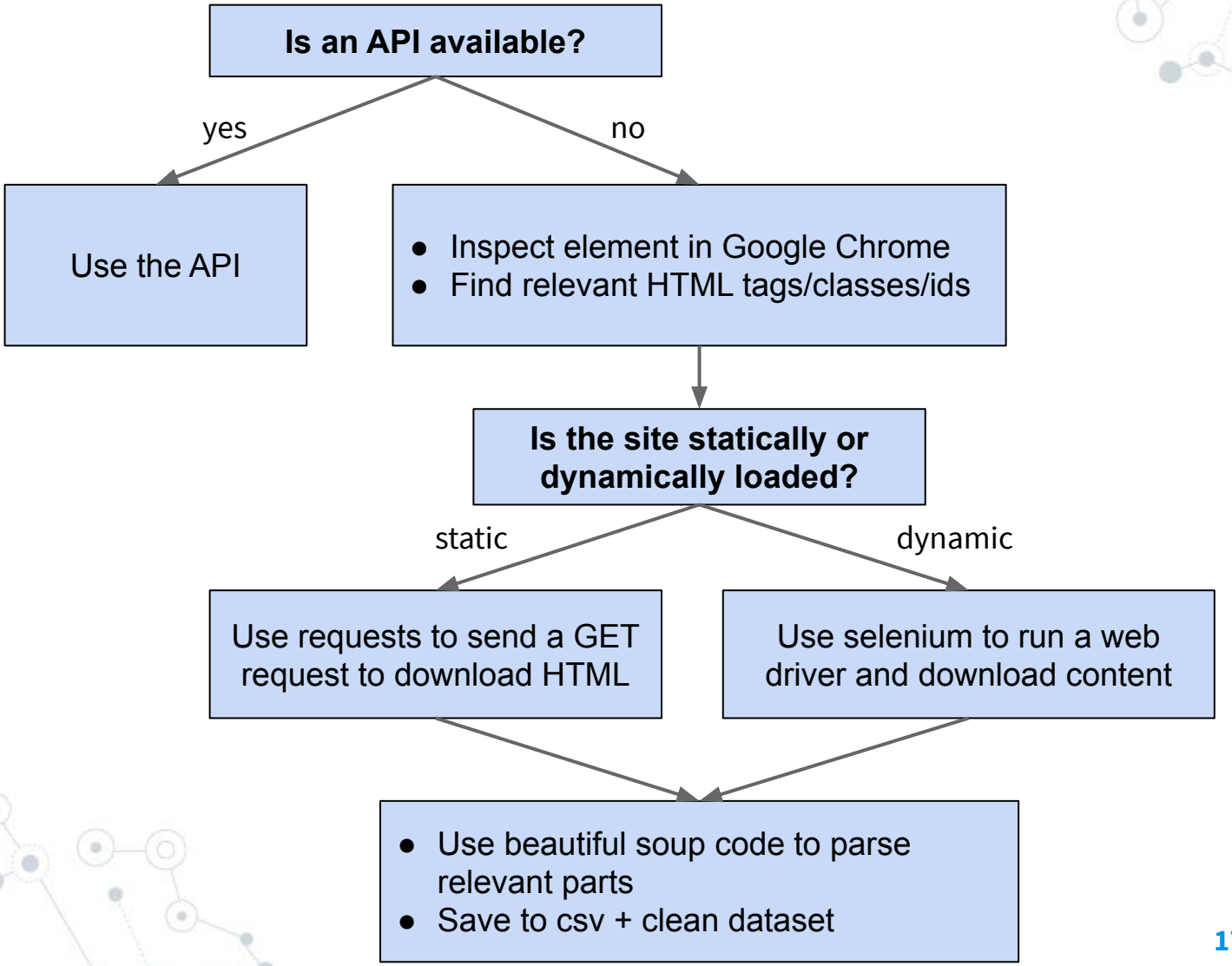
Selenium Python Package

- Selenium is a Web Browser Automation Tool
  - Mainly used to automate web applications for testing purposes
- Selenium allows you to interface with a web browser:
  - Load webpages as a client (web browser)
  - Click buttons
  - Enter information in forms
  - Search for specific information on the web pages

Coding Example

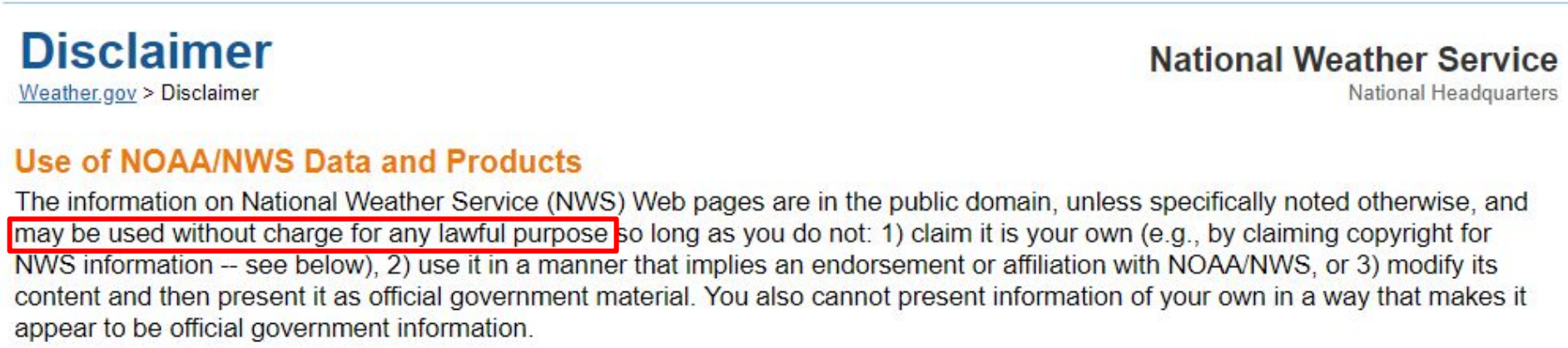
(https://colab.research.google.com/drive/1zqHSAM_IIveQonDOT4M7Sgnl3GhbMTXT)

In Summary


Is an API available?
yes
no
Use the API
Inspect element in Google Chrome
Find relevant HTML tags/classes/ids
Is the site statically or dynamically loaded?
static
dynamic
Use requests to send a GET request to download HTML
Use selenium to run a web driver and download content
Use beautiful soup code to parse relevant parts
Save to csv + clean dataset

Always abide by website policies!

- Web scraping may be against a website's Terms of Service—always respect site policies!
- Your IP address may be banned from a website if you scrape too frequently or maliciously
- Use an API if possible, don't bombard site with requests, abide by robots.txt files
- https://benbernardblog.com/web-scraping-and-crawling-are-perfectly-legal-right/



Disclaimer

Weather.gov > Disclaimer

National Weather Service
National Headquarters

Use of NOAA/NWS Data and Products

The information on National Weather Service (NWS) Web pages are in the public domain, unless specifically noted otherwise, and may be used without charge for any lawful purpose so long as you do not: 1) claim it is your own (e.g., by claiming copyright for NWS information -- see below), 2) use it in a manner that implies an endorsement or affiliation with NOAA/NWS, or 3) modify its content and then present it as official government material. You also cannot present information of your own in a way that makes it appear to be official government information.

Further Readings

- Tutorial: Python Web Scraping Using BeautifulSoup
- Scraping Dynamic Javascript Text
- Web Scraping Using Selenium — Python
- Introduction to Web Scraping using Selenium

Thanks!

Check out our next workshop on Nov 18th:

Workshop 4: Image Classification

https://www.facebook.com/events/1395381947292454/

Also give us feedback on this workshop:

tiny.cc/MAIS-F2019-W3-feedback

Thanks!

Any questions?

You can find us at:

https://mcgillai.com/