GENERAL ASSEMBLY

DATA SCIENCE
WEB SCRAPING

I. WHAT IS WEB SCRAPING?
II. ROBOTS.TXT
III. HTML & HTML TAGS
IV. HOW TO LOOK AT HTML CODE
V. BEAUTIFUL SOUP

I. WHAT IS WEB SCRAPING?

- A way of systematically pulling information from a website
- Allows you to simulate a human viewing the page and copying information
- Hacking OK Cupid
- Pull data based upon finding patterns in the structured data
- This is one way of “getting the data”.
- But be careful… you can get blocked

II. ROBOTS.TXT

- The robots exclusion standard allows website owners to specify whether they allow web “robots” or not.
- This tells you whether you can scrape a website or not.
- Located in the root directory of a website and called “robots.txt”.
- “www.google.com/robots.txt” or “http://www.dataschool.io/robots.txt”
- Read more: http://www.robotstxt.org/robotstxt.html

- Things to look for
  - User-agent: what type of robots do the following rules apply to
  - Disallow: what parts of the website are you not allowed to scrape
- Notice that you may be able to scrape parts of the website but not others

III. HTML & HTML TAGS

- HTML is the structured data underneath webpages
- Your web browser takes this code and interprets its meaning
- Basic format of an HTML tag
- <tag class="class_name" id="id_name"> ... </tag>
- Open and close tags: <tag> and </tag>
- Attributes of the tag: class="class_name", id="id_name"

IV. HOW TO LOOK AT HTML CODE

- View source code
  - This shows you the entire HTML code that makes up the webpage.
  - Good to have it all, but hard to find specific things
- Inspect Element
  - Allows you to bring up highlighted HTML for that specific item on the page
  - This is the preferred method

V. BEAUTIFUL SOUP

- We'll be using two libraries to help us create web scraping robots.
  - Requests
  - Beautiful Soup
- Requests "gets" the webpage's HTML from the web.
- Beautiful Soup changes the HTML into a searchable, structured object.