data science
student society

Introduction to Web Scraping

October 24, 2024

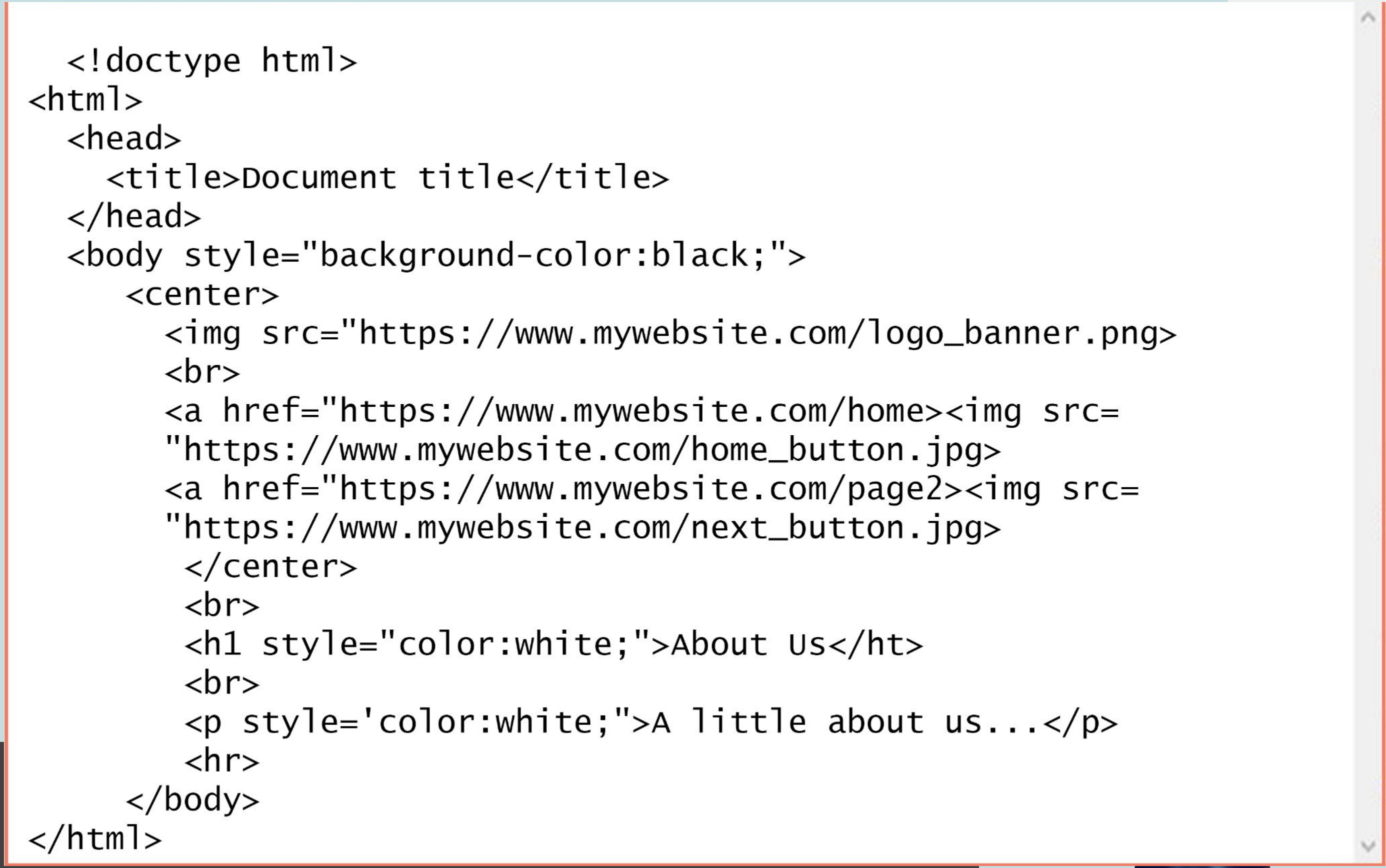
data
science
student
society

Please check in!

What is Web Scraping?

- Imagine you want to get a large amount of data from a website to use on a machine learning data science project
- Rather than copy + paste the information, you can directly obtain raw HTML from websites
    - This HTML contains data on formatting and also raw text
- Many websites have APIs to make this process easier

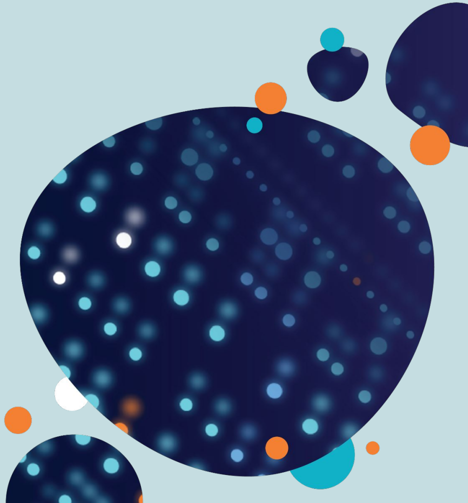
Aside: HTML Breakdown



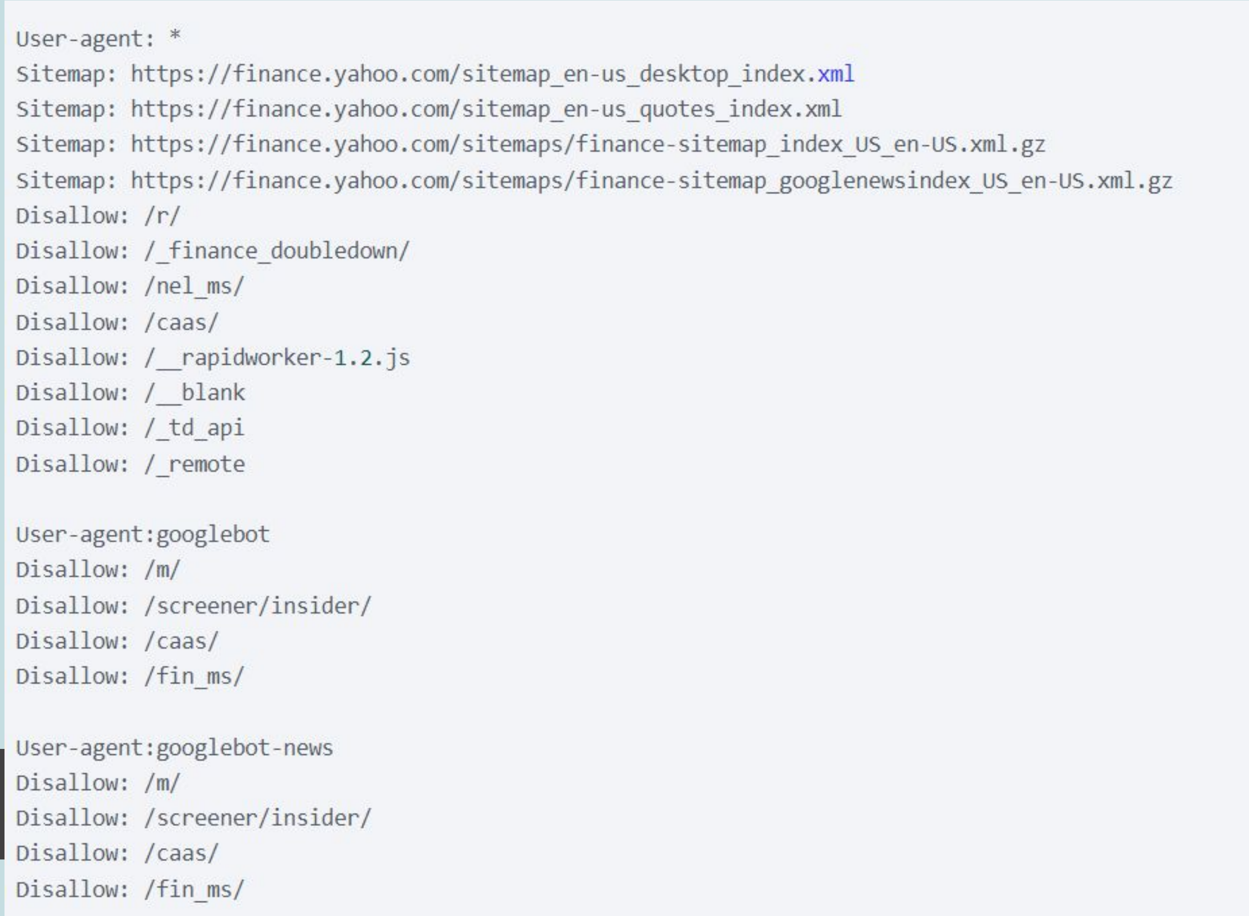
```
  <!doctype html>
<html>
  <head>
    <title>Document title</title>
  </head>
  <body style="background-color:black;">
     <center>
       <img src="https://www.mywebsite.com/logo_banner.png>
       <br>
       <a href="https://www.mywebsite.com/home><img src=
       "https://www.mywebsite.com/home_button.jpg>
       <a href="https://www.mywebsite.com/page2><img src=
       "https://www.mywebsite.com/next_button.jpg>
        </center>
        <br>
        <h1 style="color:white;">About Us</ht>
        <br>
        <p style='color:white;">A little about us...</p>
        <hr>
     </body>
</html>
```

CAUTION!!! 👀

- When you web scrape, you are acting like a web browser to access the source code
- Suspicious behavior will usually result in your IP address being blocked from accessing the site
- ALWAYS check the site’s ToS
- Then check for a *.com/robots.txt file
  - This tells you the rules for scraping



A robots.txt file

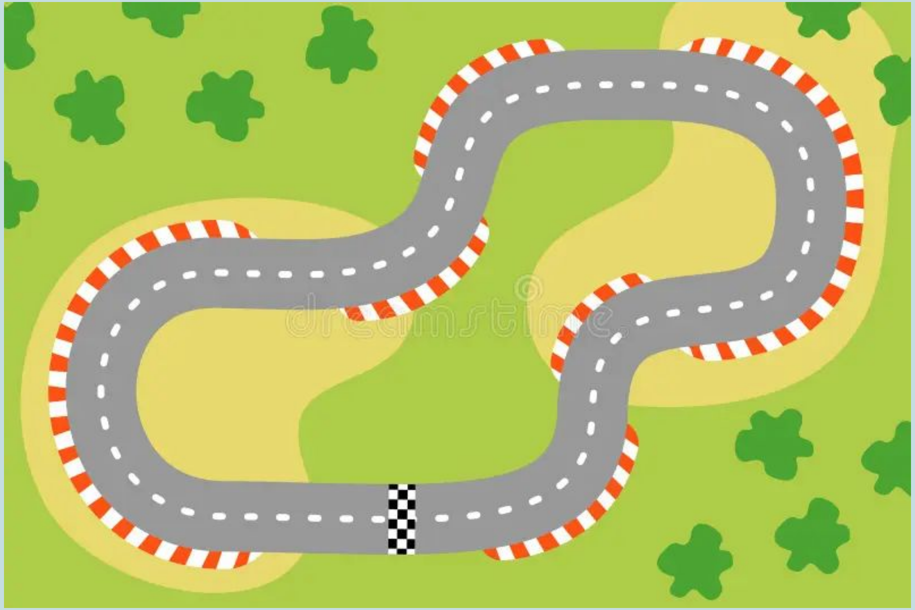


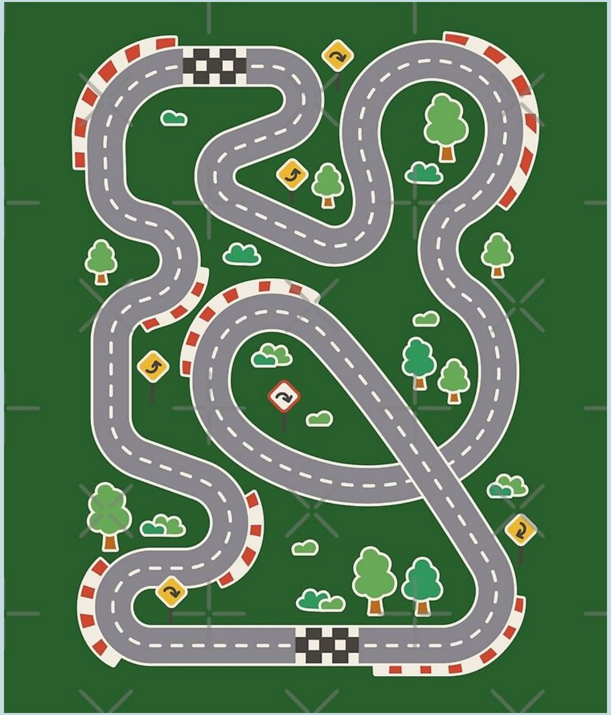
```
User-agent: *
Sitemap: https://finance.yahoo.com/sitemap_en-us_desktop_index.xml
Sitemap: https://finance.yahoo.com/sitemap_en-us_quotes_index.xml
Sitemap: https://finance.yahoo.com/sitemaps/finance-sitemap_index_US_en-US.xml.gz
Sitemap: https://finance.yahoo.com/sitemaps/finance-sitemap_googlenewsindex_US_en-US.xml.gz
Disallow: /r/
Disallow: /_finance_doubledown/
Disallow: /nel_ms/
Disallow: /caas/
Disallow: /__rapidworker-1.2.js
Disallow: /__blank
Disallow: /_td_api
Disallow: /_remote

User-agent:googlebot
Disallow: /m/
Disallow: /screener/insider/
Disallow: /caas/
Disallow: /fin_ms/

User-agent:googlebot-news
Disallow: /m/
Disallow: /screener/insider/
Disallow: /caas/
Disallow: /fin_ms/
```

Static vs. Dynamic Web Pages

- Static sites do not change, the content they present is static and is all premade assets
    - Each page on the website is represented by a matching html file
    - static site example
- Dynamic sites are pages that can, in real-time, update their content
    - Content on the page can be changed even without a separate html file and reference
    - Think of a newspaper vs. a live news ticker
    - Just requesting the html does not deliver because you’ll get the “default” html
    - dynamic site example

Web Scraping Example





Selenium & Chromedriver

- Selenium is a library that allows you to automate browsers
- Chromedriver allows you to control that automation
- Allows you to do things like:
  - Wait for dynamic to load
  - Interact with pages
    - Like clicking buttons
  - Scroll dynamically updating feeds
  - Fill out forms

Demo

- Clone the notebook here: https://github.com/Nolancchu/2024-Fall-Webscraping-Workshop/blob/main/workshop.ipynb

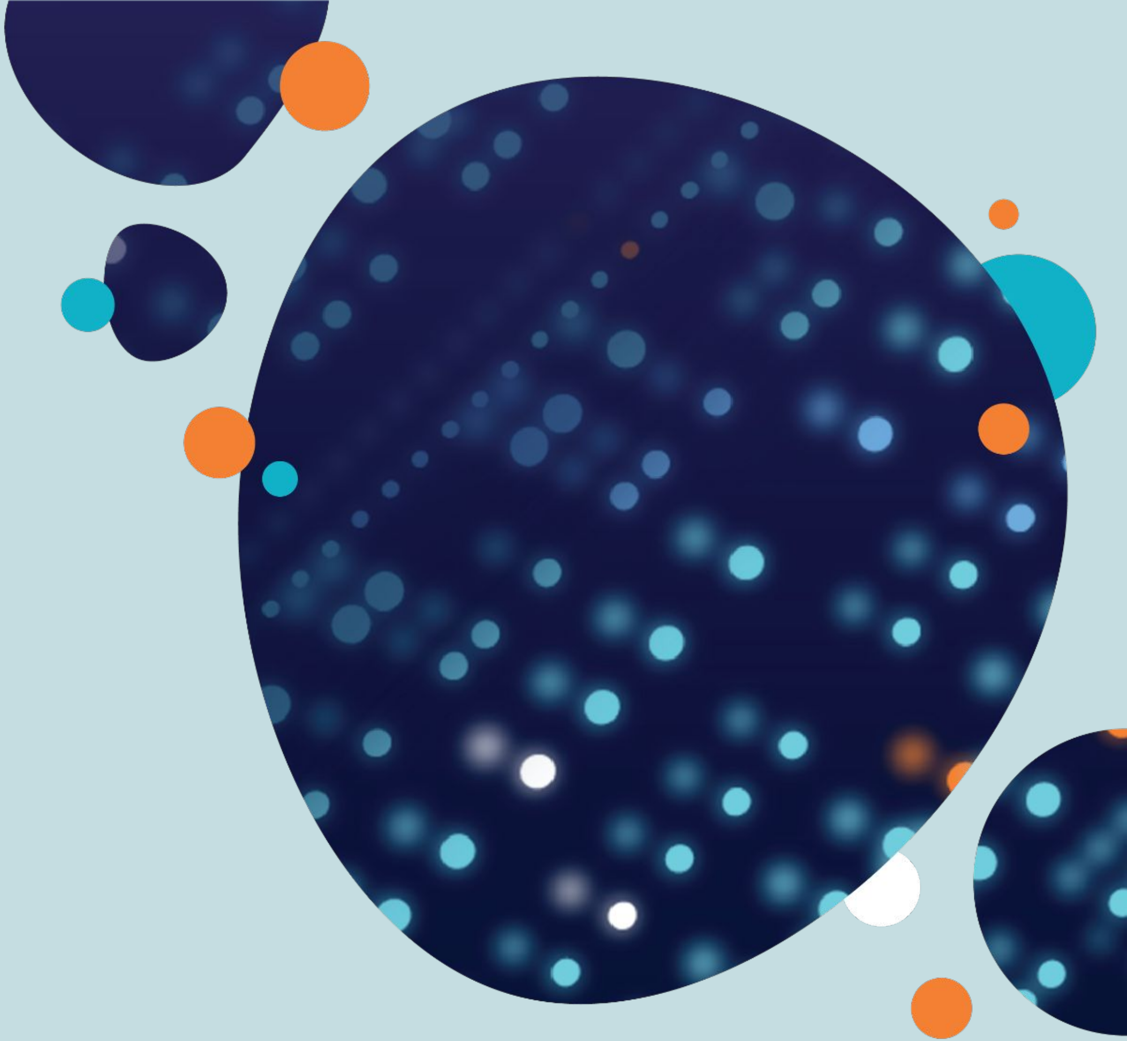
Further Resources

- Geeks4Geeks Web Scraping with Amazon Customer Reviews Tutorial: https://www.geeksforgeeks.org/web-scraping-amazon-customer-reviews/?ref=oin_asr10
- HTML Explained: https://www.w3schools.com/html/html_intro.asp
- Further explanation on robots.txt: https://www.zenrows.com/blog/robots-txt-web-scraping#web-scraping-using-robots-txt-steps

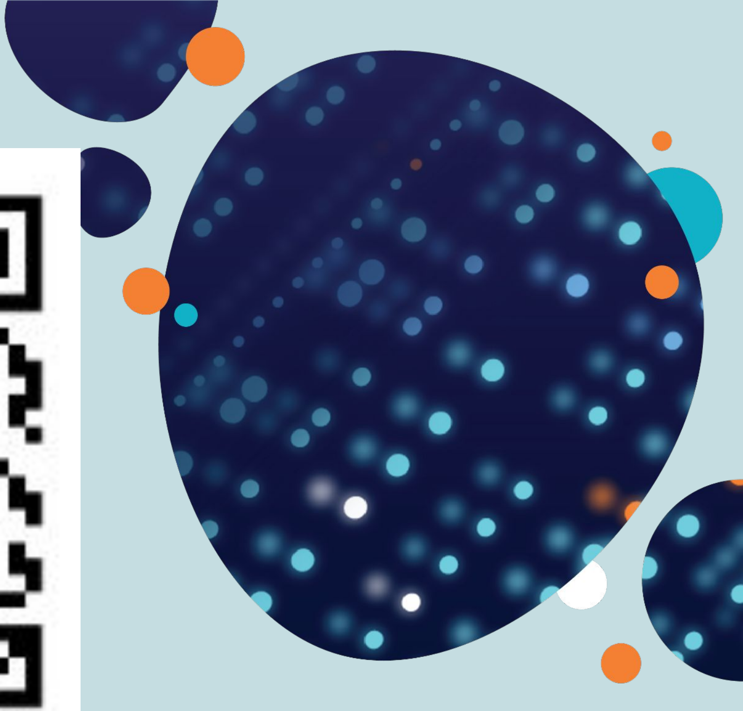
Note: Make sure to stay connected on the DS3 Discord for further workshops!!!!

data science
student society

Next Time:
Intro to EDA



data science student society



Leave your feedback here!