GENERAL ASSEMBLY

INTRO TO DATA SCIENCE

LECTURE 14: SCREEN SCRAPING

I. Screen scraping vs. APIs

II. The basics of screen scraping

III. Python libraries

IV. Lab: Three Screen scraping examples

V. Classwork: Scraping your own data

I. Screen scraping vs. APIs

Screen scraping vs. APIs

- Which option is better?!

Screen scraping vs. APIs

- Which option is better?!
  - Advantages of APIs

Screen scraping vs. APIs

- Which option is better?!
  - Advantages of APIs
    - Data should be clean
    - Managed system should rarely change
    - More control over data selection
    - Tutorials and data definitions

Screen scraping vs. APIs

- Which option is better?!
    - Advantages of Screen scraping

Screen scraping vs. APIs

- Which option is better?!
  - Advantages of Screen scraping
    - Can always be used on any site
    - Websites are generally better maintained than APIs
    - No or fewer fees!
    - Can be done anonymously
    - Often the fastest method

Screen scraping vs. APIs

- Which option is better?!

- Depends on your goals and options

Discuss in groups of 4: Is it easier to scrape data from your sites or use an API? Why?

II. The basics of screen scraping

The basics of screen scraping

- Finding a site with the data you need —> Done

The basics of screen scraping

- Finding a site with the data you need —> Done
- Considerations for Fetching the Data

The basics of screen scraping

- Finding a site with the data you need —> Done
- Considerations for Fetching the Data
  - Do you need to login?

The basics of screen scraping

- Finding a site with the data you need —> Done
- Considerations for Fetching the Data
  - Do you need to login?
  - Are there drill-downs or selection options?

The basics of screen scraping

- Finding a site with the data you need —> Done
- Considerations for Fetching the Data
    - Do you need to login?
    - Are there drill-downs or selection options?
    - Is the data paginated (what happens when your out of pages?)

The basics of screen scraping

- Finding a site with the data you need —> Done
- Considerations for Fetching the Data
  - Do you need to login?
  - Are there drill-downs or selection options?
  - Is the data paginated (what happens when your out of pages?)
  - Is this a onetime load or repeated use?

The basics of screen scraping

- Finding a site with the data you need —> Done
- Considerations for Fetching the Data
  - Do you need to login?
  - Are there drill-downs or selection options?
  - Is the data paginated (what happens when your out of pages?)
  - Is this a onetime load or repeated use?
  - Is the data you need in the source html?

The basics of screen scraping

- Finding a site with the data you need —> Done
- Considerations for Fetching the Data
- How to peer through the web page?

The basics of screen scraping

- Finding a site with the data you need —> Done
- Considerations for Fetching the Data
- How to peer through the web page?
  - View < Developer < View Source

The basics of screen scraping

- Finding a site with the data you need —> Done
- Considerations for Fetching the Data
- How to peer through the web page?
  - View < Developer < View Source
  - View < Developer < Java Script (CMD + Option + J)
    - Look at Elements and Network to see whats going on

The basics of screen scraping

- Finding a site with the data you need —> Done
- Considerations for Fetching the Data
- How to peer through the web page?
  - View < Developer < View Source
  - View < Developer < Java Script (CMD + Option + J)
    - Look at Elements and Network to see whats going on
    - Use Network to see Headers, Cookies, Params

Using your Javascript Console, choose a url and explore how the url, headers, and parameters change as you navigate

III. Python libraries

Python Libraries

- Requests
- Beautiful Soup
- Re (regex can be helpful, but manual)
- Pandas

Python Libraries

- Requests
- Beautiful Soup
- Re (regex can be helpful, but manual)
- Pandas

- Scrapy
- Urllib
- mechanize
- many more …
- https://classic.scraperwiki.com/docs/python/python_libraries/

Python Libraries

- Requests
  - We’ll use to create and maintain web sessions
  - Very important for logins or semi-protected sites
- Beautiful Soup
- Re (regex can be helpful, but manual)
- Pandas

Python Libraries

- Requests
- Beautiful Soup
  - Read through HTML and extract tagged data
- Re (regex can be helpful, but manual)
  - Can also be used to extract data, but this is less reliable
- Pandas

Python Libraries

- Requests
- Beautiful Soup
- Re (regex can be helpful, but manual)
- Pandas
    - read html tables straight into dataframes

IV. Lab

V. Classwork:

Scrape your own websites and store the data as a csv, but NOT IN THE DROPBOX!

Resources

- http://blog.hartleybrody.com/web-scraping/
- http://rhodesmill.org/brandon/chapters/screen-scraping/
- https://classic.scraperwiki.com/docs/python/python_libraries/