Joanna Wu
CPSC 499 Write Up and Notes

In doing some preliminary research on Python Scripting and web scraping, I've jotted down some of my notes here:

- BeautifulSoup seems to be the primary Python library used by most for web scraping
- Data can be exported to many different formats, including an excel spreadsheet/csv
- Requests is used to send and receive URLs
- First web scrapers were believed to be created in 1993, to serve in a search engine
- There are multiple ways to accomplish the goal of scraping uniform data from a web page:
    - Copy and paste
        - Slowest and most labor intensive
    - Text pattern matching
        - Can use regular expressions
    - HTTP programming
        - Works with static and dynamic web pages, using socket programming
    - HTML parsing
    - DOM parsing
    - Vertical aggregation
        - Domain specific harvesting, very scalable
    - Semantic annotation recognizing
    - Computer vision web-page analysis
        - Attempts to use machine learning and computer vision to interpret pages visually
- Cloud versus local
    - Local web scrapers can run on your computer, not very scalable
    - Cloud based allows for more simultaneous tasks
- Some common uses of web scraping
    - Price monitoring
    - Market research
    - Mundane / repetitive tasks
    - Gathering contact information
    - Real estate

    - Sentiment analysis
    - News and content monitoring
- Preventing web scraping- there are also multiple ways in which web scraping can be prevented by websites
    - Legal statement
    - Use a server that has a firewall built in, to protect your server from attacks
    - Use CSRF tokens
        - A hidden form field to make it harder to parse data on the website
    - Prevent hotlinking
        - This makes it so that if a link or image is taken off of the website, it won't be displayed
    - Blacklist specific IP addresses
        - Useful if you can identify IP addresses which have been used for scraping
    - Throttling requests
        - Limiting the number of requests from one IP address
    - Change website structure frequently