Software Engineering 491 - sddec19-01
Web Crawling for Data Breach Reports
Week 4 Report
3/1 - 3/8
Client: Benjamin Blakely
Faculty Advisor: Dr. Daniels

Team Members:

Mark Schwartz - Scraping Team
Alec Lones - Project Leader - -Machine Learning Team
Nolan Kim - Scraping Team - Git Master
Jeremiah Brusegaard - Machine Learning Team

Weekly Summary:

The team explored writing different scrapers with Scrapy in order to resolve any outstanding dependency issues, gain more experience with Scrapy, and practice scraping different websites. We also added multiple branches to the git repository in order to start pushing our experiments and collaborating together.

Past Week Accomplishments:

Accomplished creating multiple scrapers in order to scrape different websites
Gained experience working with Scrapy

Pending Issues:

Beautiful soup not correctly parsing just text. Instead it's getting some jquery.

Individual Contributions:

Team Member	Contribution	Weekly Hours	Total Hours
Mark Schwartz	Finished up a test run scraper to run through the data breach report pdfs given to us by Ben (our client). It saves all the text that is in quotes HTML tags.	~6	~30
Alec Lones	Wrote a test scraper to become more familiar with Scrapy	~6	~30

	Resolved a bunch of dependency issues between windows, pycharm, and scrapy Scraped a government food nutrition website as a test		
Nolan Kim	Continue to discover new features of scrapy, including how to download scraped files to your computer, how to run a spider from another script, and how to pipe output from a spider into another function.	~6	~30
Jeremiah Brusegaard	● Created a prototype for scraping the websites that we were given by Ben. ● Created a blacklisting feature for restricting it from only crawling twitter or other social media sites ● Started on lemmatizer	~6	~30

Plans for upcoming week:

- Mark Schwartz:
  - Improve my crawler to detect links and crawl to other websites
  - Play with lemmatizing/vectorization
  - Feed crawler data into said vectorization functions
- Alec Lones:
  - Continue to improve my test scraper
  - Log my govt food data (I think this might be useful in working with lemmatization, stemming, and vectorization)
  - Once my scraper is functioning where I want it, then I will move on to stemming, lemmatization, and vectorization
  - Jeremiah suggested trying out beautiful soup to help with website scraping so I will investigate that as well
- Nolan Kim:
  - Attempt to feed crawler output into a lemmatization agent
  - Work on the design document
- Jeremiah Brusegaard:
  - Finish lemmatizer so that I can start vectorization
  - Figure out when to stop the scraper
  - Mark the training data so that the future machine learning algorithm works

  - Figure out how to make the text parser better because it keeps getting

Summary of weekly meeting:

We didn’t meet with Ben because of scheduling conflicts. Talked about plans moving forward and doing the design document.Talked about Beautiful Soup for text parsing. Talked about what standards we want to have for scraping the websites since they are non standard in nature.