Software Engineering 491 - sddec19-01
Web Crawling for Data Breach Reports
Bi-Weekly Report 1
8/30 - 9/13
Client: Benjamin Blakely
Faculty Advisor: Benjamin Blakely

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

Bi-weekly Summary:
We are getting back into the swing of school so mainly focused on refamiliarizing ourselves with the prototype and project plan. Started creating action items for the group to make sure the project is delivered on time. Worked out issues with lemmatizer so that it works correctly now.

Past 2 Weeks Accomplishments:
- Improved the speed of the scraper.
- Started research on making the project multithreaded in order to decouple the code and devote more power to the machine learning module.
- Worked on the PIRM review presentation.
- Fixed lemmatizer module since it was not actually lemmatizing the majority of words encountered.

Pending Issues:
- Machine Learning model is overfitting
- Might need Beautiful soup replacement for efficiency
- Need to figure out why certain links are getting denied even with following robots.txt

Individual Contributions:

Team Member	Contribution	Bi- weekly Hours	Total Hours
Mark Schwartz	● Researched xpathing to improve our scraper ● Worked on PRIM review presentation	~12	~12
Alec Lones	● Continuing to investigate improvements to the scraper	~12	~12

	● Worked on PRIM review presentation		
Nolan Kim	● Researched multithreading in Python ● Worked on PRIM review presentation	~12	~12
Jeremiah Brusegaard	● Fixed lemmatizer because it was not tagging parts of speech ● Thoroughly tested lemmatizer to ensure proper lemmatization across multiple websites ● Worked on PRIM review presentation	~12	12

Plans for upcoming week:

- Mark Schwartz:
  - Continue to work on improving the scraper
  - Help create the XPath querier to speed up the scraper.
- Alec Lones:
  - Continue working to improve scraper and ML
  - Create a new scraper prototype that eliminates beautiful soup
- Nolan Kim:
  - Create prototype crawler+machine learning module that is multithreaded.
  - Create an XPATH query to isolate the body tag of an HTML document without any other tags inside the body tag. This method would be more efficient than using BeautifulSoup.
- Jeremiah Brusegaard:
  - Work on creating work tasks for group so we can get this done by the end of the semester
  - Train Machine learning model and fix overfitting issue
  - Gather information about the data provided from the client and probably find more data to train the model

Summary of weekly meeting:

We did not meet with the client as of this time due to scheduling conflicts. So our first meeting will be on Monday the 16th. Plans to ask about where the data came from and if we need to seek out large amounts of our own data. Ask client for recommendation on fixing the overfitting problem with the ML model.