Software Engineering 491 - sddec19-01
Web Crawling for Data Breach Reports
Week 6 Report
4/6 - 4/12
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:
Got VM from ETG it now has the project on it with dependencies installed to run the project. We are working on getting a database setup on it to move from storing lemmatized data in text files to a database. Also we met with the client and discussed goals to have by the end of the semester and the presentation we need to give.

Past Week Accomplishments:
- Vm is setup
- VM has project setup to be able to crawl
- Got machine learning algorithms to test, currently some are above 80% accuracy

Pending Issues:
Need database
Better way of storing lemmatized sites
Scraping seems to be currently single threaded and is bottlenecking the system. Might consider multi threading this?

Individual Contributions:

Team Member	Contribution	Weekly Hours	Total Hours
Mark Schwartz	● Set up VM ● Created a tester function to simplify testing of the model	~6	~54
Alec Lones	● Continued to assist Jeremiah with scraping and storing data	~6	~54

	● Explored multithreading scrapy spiders		
Nolan Kim	● Researched and played around with multithreaded Scrapy spiders	~6	~54
Jeremiah Brusegaard	● Tested different Machine learning models ● Found some that give above 80% accuracy, still need to run tests to make sure data isn't being overfitted	~6	~54

Plans for upcoming week:

- Mark Schwartz:
  - Help set up database/learn MongoDB
  - Set up a demo for presentation
- Alec Lones:
  - Continue to investigate multithreading scrapy
  - Continue to assist in scraping and storing data
  - Continue to investigate beautiful soup and goose3 (mostly as it relates multithreading now)
- Nolan Kim:
  - Learn MongoDB
  - Make breachCrawler multithreaded
- Jeremiah Brusegaard:
  - run tests for machine learning
  - comment and document code
  - Help Mark get a demo code runner going to be able to give a url and a classification

Summary of weekly meeting:

Talked to Ben about progress and upcoming presentation. We also asked him some questions about machine learning and issues we were having with it in our project. Overall Ben is very happy about our progress this semester and is looking forward to our implementation next semester. Current plans are to wrap up this semester with the presentation and distribute code/server access to continue experimenting over the summer.