Software Engineering 491 - sddec19-01
Web Crawling for Data Breach Reports
Bi-Weekly Report 2
9/14 - 9/27
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 figured out how to make the crawler even faster than before to the point where we can crawl 5000 links in 7 minutes. Also there was bugs in the code that we didn't know where bugs and when fixed significantly enhanced our machine learning performance and accuracy. Set up a vm so we have all dependencies stored on it making our project more portable.

Past 2 Weeks Accomplishments:
- Setup VM for portability
- Fixed bugs related to Machine Learning
- Greatly increased speed of crawler
- Fixed pathing issues related to scrapy import statements in our code

Pending Issues:
- 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	● Helped optimize/improve machine learning model ● Tested other possible ml algorithms/parameters	~12	~24
Alec Lones	● Setup ubuntu VM for development	~12	~24

	● Attempted to install vpn on the vm (not liking it yet) ● Worked on setting up MongoDB (discussed with Jeremiah on a database structure) ● Discussed ML Models with Jeremiah		
Nolan Kim	● Fixed an issue where scrapy wouldn't recognize the correct paths for modules in python. ● Helped Jeremiah decouple the code.	~12	~24
Jeremiah Brusegaard	● Found and fixed bug for machine learning ● Helped tweak scraper, enabled caching for quickly re scraping sites ● Helped Alec install and setup project on VM	~12	~24

Plans for upcoming 2 weeks:

- Mark Schwartz:
  - Continue to test different ML parameters
  - Figure out how to save ML models to save the better ones
- Alec Lones:
  - Continue poking vm to see if I can get the VPN working
  - Finish constructing the MongoDB database
  - Poke Scrapy some more
  - Work with Jeremiah on ML model
- Nolan Kim:
  - Try to configure scrapy in a way where it doesn't get blocked
  - Implement xpath to replace BeautifulSoup
- Jeremiah Brusegaard:
  - Create a way to save Models for finding the "best" model
  - Figure out better parameters for Random Forest classifier
  - Research Random Forest classifier best practices

Summary of weekly meeting:

We talked to our client Ben about what we wanted to accomplish to get a finished product by the end of the semester. Reviewed feedback from the presentations we have had to give this semester and last semester. Ben thinks that we are in a good place to finish this project on time.