Software Engineering 492 - sddec19-01
Web Crawling for Data Breach Reports
Bi-Weekly Report 5
10/25-11/7
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 found a good model using cross validation. Working on testing what happens when we let it crawl. Finally we need to integrate the database now that the database is stood up and the API calls are done for it.

Past 2 Weeks Accomplishments:
- Got a high quality model with a Mathews correlation coefficient of .95
- Database API calls are finished and ready to crawl
- Crawler can crawl new websites with boolean variable flipped to true and classify new sites

Pending Issues:
- Need to figure out why certain links are getting denied even with following robots.txt
- Store Links and reports on database
- Need to limit the amount of scraping/machine learning the crawler does so that it doesn't eat up lots of RAM

Individual Contributions:

Team Member	Contribution	Bi- weekly Hours	Total Hours
Mark Schwartz	● Made a prototype GUI ● Helped Jeremiah find good ML models	~12	~60
Alec Lones	● Assisted Jeremiah with analyzing ML results ● Help Jeremiah and Nolan with beginnings of database integration	~12	~60

	● Assisted Jeremiah with making changes to our ML algorithm (dropping Hash vectorizer and using a count instead)		
Nolan Kim	● Made a prototype where the machine learning is integrated with the crawler	~12	~60
Jeremiah Brusegaard	● Found really high quality model ● worked with Alec to get DB working ● Helped nolan get crawler working ● Made ability for model to print out more stats about it when it gets saved	~12	~60

Plans for upcoming 2 weeks:

- Mark Schwartz:
  - Finish GUI
  - Assist in improving ML model
- Alec Lones:
  - Continue assisting Jeremiah with ML algorithm progress
  - Continue updating the Mongo interface as needed
- Nolan Kim:
  - Try to stop the crawler from eating up so much RAM
- Jeremiah Brusegaard:
  - Integrate DB with crawler
  - Store base link information for model

Summary of bi-weekly meeting:

Client said we are still making good progress. We used the grid search recommended by him to find parameters for getting the best possible model. We are just working on finishing stuff so the project is done by the end of the semester.