Software Engineering 491 - sddec19-01
Web Crawling for Data Breach Reports
Bi-Weekly Report 3
9/28 - 10/11
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 were pretty busy with midterms so not as much got done as we had hoped but we got a few things done. The ability to print out the confusion matrix to see false positives and false negatives will help us tune and gauge how good our model is on real world data.

Past 2 Weeks Accomplishments:
- Testing parameters
- Creating confusion matrix
- Starting on save/load functionality

Pending Issues:
- Might need Beautiful soup replacement for efficiency
- Need to figure out why certain links are getting denied even with following robots.txt
- Setup mongo on VM

Individual Contributions:

Team Member	Contribution	Bi- weekly Hours	Total Hours
Mark Schwartz	● Helped optimize/improve machine learning model ● Worked on Presentation ● Helped Jeremiah work on confusion matrix print out	~12	~36
Alec Lones	● Continued configuration on the VM	~12	~36

	● Determined VPN is NOT going to work, something about ISU and their vpn connection software just doesn't like linux as far as I can tell ● Finished database design with Jeremiah ● Discussed more ML implementations with Jeremiah ● Worked on presentaiton		
Nolan Kim	● Helped Jeremiah with decoupling code ● Worked on presentation	~12	~36
Jeremiah Brusegaard	● Worked on presentation ● Started working on save/load functions ● Working on showing confusion matrix	~12	~36

Plans for upcoming 2 weeks:

- Mark Schwartz:
  - Continue to test different ML parameters
  - Help Alec with UI
  - Help Jeremiah figure out how to use confusion matrix info to create better model
- Alec Lones:
  - MongoDB is stood up, but needs an interface created
  - Start work on Django UI
- Nolan Kim:
  - Try to configure scrapy in a way where it doesn't get blocked
  - Implement xpath to replace BeautifulSoup
- Jeremiah Brusegaard:
  - Finish save functions for Models for finding the "best" model
  - Figure out better parameters for Random Forest classifier
  - Research Random Forest classifier best practices
  - Get info from confusion matrix

Summary of weekly meeting:

We touched base with Ben about bug fixes. He still thinks we are doing well and on track. Another thing came up in our PIRM meeting where Daniels pointed out to us that we should be

looking for false positives and negatives and we are taking immediate action on this since it didn’t occur to us to look at this stat. We will try to use this to tune the model and also make sure it has “realistic” data to use.