Software Engineering 492 - sddec19-01

Web Crawling for Data Breach Reports

Bi-Weekly Report 4

10/12 - 10/25

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 just working on adding functionality to the project soon we will allow the crawler to free roam outside of the initial urls and monitor what it does. We are working on a possible UI this was a stretch goal but we feel as though we could get something working before the end of the semester.

Past 2 Weeks Accomplishments:

- Got saving and loading working.
- Can print out confusion matrix.

Pending Issues:

- Need to figure out why certain links are getting denied even with following robots.txt
- Store Links and reports on database

Individual Contributions:

Team Member	Contribution	Bi- weekly Hours	Total Hours
Mark Schwartz	● Created a rough UI for testing ● Started making a graphical representation of the confusion matrix	~12	~48
Alec Lones	● Collected and generated lists of non-breach report sites to improve model with	~12	~48

	● Created python interface to interact with mongo db		
Nolan Kim	● Worked on determining whether our dataset is representative of the wild ● Experimented with scrapy configuration to try and find a config that isn’t blocked	~12	~48
Jeremiah Brusegaard	● Save and load done ● Confusion matrix prints out when model is created	~12	~48

Plans for upcoming 2 weeks:

- Mark Schwartz:
  - Finish confusion matrix graph/plot
  - Finish/polish UI
- Alec Lones:
  - Evaluate impact/need for further non-breach report data
  - Work with Jerry to evaluate DB interface design and speed
- Nolan Kim:
  - Figure out how to get VPN working from the command line
- Jeremiah Brusegaard:
  - Find model evaluation library to find best model
  - Turn model loose on open internet beyond seed urls

Summary of bi-weekly meeting:

We talked to Ben and just gave him status updates. He recommended features to prioritise and said we are still on track to finish this project. We also talked about how we want to approach the final presentation and what our plans for the rest of the semester are.