Software Engineering 491 - sddec19-01

Web Crawling for Data Breach Reports

Week 6 Report

3/30 - 4/5

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:

Team finished initial draft of the design document. We have also pushed the prototype further and are running into some issues that need to be resolved with multithreading and needing a database. We currently have a homebrewed server setup for testing, but will need to explore getting a VM at school.

Past Week Accomplishments:

- Projects are split into modules
- Got vectorization "correctly working"
- Design Doc was turned in
- Websites are sanitized and stored
- Machine Learning has a module that should be able to start training

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	Worked on design document, helped Jeremiah with optimization of the crawler by storing the vectorized web pages in a text document rather than in memory.	~6	~48

Alec Lones	Worked on the design document Assisted Jeremiah in setting up a test server for the prototype Learned a bit about beautiful soup and goose3	~6	~48
Nolan Kim	● Design document work ● Researched vectorization to assist Jeremiah in building the machine learning module	~6	~48
Jeremiah Brusegaard	● Worked on design document ● Worked on fixing vectorization so the machine learning module correctly works ● Split the original file into modules so that they can be run independently of one another ● Currently training model on different sets	~6	~48

Plans for upcoming week:

- Mark Schwartz:
    - Help Jeremiah break the data set into a test set and training set.
    - Set up a VM and Database
    - Send request for VM and database
- Alec Lones:
    - Continue learning about Goose3 and Beautiful soup
    - See if there is a way to multithread the scraper, or figure out removing the bottleneck
    - Continue assisting Jeremiah in scraping and storing data
- Nolan Kim:
    - Figure out how to run multiple scrapers in multiple threads
    - Get a VM set up to run the database
    - Make scraper functions more robust
- Jeremiah Brusegaard:
    - Hopefully have a trained model that can make somewhat accurate predictions probably more than 80% accuracy

Summary of weekly meeting:

Talked about what needs to get done for coming week and what we want to have accomplished by summer. Assigned tasks to members on things to do over the summer. Also talked about getting virtual machine from ETG to run a scraper.