Senior Design Bi-Weekly Status Report

Team: sdmay21-35
Team Email: sdmay21-35@iastate.edu
Team Members:

Ahmad Alramahi - Lead Developer
Austin Boling - Meeting Facilitator
Joseph Naberhaus - Project Lead
Ekene Okeke - Report Coordinator
Ethan Ruchotzke - Documentation Manager
James Taylor - Linguistics SME

Project: POS Tagger for Software Documentation
Report Period: Feb 22nd - Mar 8th

Summary of Progress in this Period

This period was the primary work period for the February milestone. The february milestone, as discussed in the semester 1 presentation, includes work in Tokenization and Sentence splitting. In addition, formatting and specific models need to be selected.

This period, primary work was done on the tokenizer. While we were unable to finish, we have a strong design and are well into implementing it, and expect to be done by March 1st.

We finalized the tag set we are planning on using moving forward, and the new tagset is defined inside of a google document in our shared drive. We now have a set of tokens we can train a model to recognize and tag for us.

In addition, more work needs to be done finding sources for training. We currently only have two general sites we can scrape HTML data from. We need to find more data in order to have a thoroughly trained model.

Moving forward, work needs to be done on finding training models which fit this product, and a discussion needs to be had on the formatting of the training pipeline's elements. This must be completed by March 1st.

Meeting Notes

- Extend the tokenization deadline by one week
- Find extra sources (50 / 50)

Pending Issues

- Refinement of the automatic tagging pipeline
    - Inaccuracies / room for extra automatic tagging
    - Minimization of manual tagging is a must
- Tokenizer
    - 2 week schedule
    - Ethan & James will be working on this
    - Implement the “middle” of the data pipeline (turning data into tokens)
    - Unfortunately, the 2 week schedule was too tight for this component. Work will continue in order to create a functional product. Planning and design are thorough, and the expectation is a working product by March 1st.
- Training
    - Preduce training that works with the data from tagging pipeline

Work Done - Current Period

Our hourly work done for the current reporting period is as follows.

Who	What	Individual Hours	Cumulative Hours (Starting 2/22/2020)
Ethan	Tokenization - Stage 1, Access to VM	6	6
Austin	Tags	3	3
James	Tokenization & Sentence Splitting	7	7
Ahmad	Planned future work	1/2	1/2
Joseph	Scraper & Training	2	2
	.		

Plans for upcoming Reporting Period

Our plans for this upcoming reporting period **(Feb 23rd - Mar 8th)** are broken down as follows:

Who	What	Due When
All (Led by Austin)	Finalize Tag Set	Prior to tokenizer completion
Ethan & James	Finish the tokenizer & sentence splitter	March 1st (End of February Milestone)
All	Manual Data Tagging	Four weeks from 8th February
All (Led by Joseph)	Setup Training Pipeline (Schedule out)	March 1st
Ahmad	Implementing training model on pipeline	Four weeks from 23rd of feb
		s