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: Sept 18 - Oct 2

Summary of Progress in this Period

The work done during this reporting period begins to lay out the foundations of the work for the project.

- James found four different types of software documentation to consider when building our tagger, and decided that developer provided documentation and interactive programming challenges are the most important
- Austin began listing out new PoS token tags
- Austin and James began summarizing the cyclic dependency network research
- Ahmad, James, Joseph, and Ethan began analyzing the Stanford NLP implementations and their differences (including output consistency/inconsistency)
- Joseph began exploring how to train the Stanford CoreNLP

Pending Issues

1. Consistency errors between Python’s Stanza Library and Java’s CoreNLP Library
    a. During consistency testing, multiple differences were found between the output of the Java NLP and Python NLP
    b. This issue was exacerbated by the fact that python runs a jvm instance of CoreNLP, meaning that python is using a different version of CoreNLP than Java’s CoreNLP is
    c. More testing is to be done with consistency errors, however the group is moving towards Java as the primary language for future development due to its lower level library.
2. Choosing between Python’s Stanza Library and Java’s CoreNLP Library
    a. The group needs to come to a consensus related to the usage of Python or Java for future development
    b. The group of moving in the direction of Java’s CoreNLP due to it’s low level control and speed compared to the Python library
3. Development of the corpus of software documentation (and manual tagging)
    a. The team needs to develop the infrastructure to scrape software documentation and get useful data from it
    b. The current thought is to use web scraping to generate text which is capable of being passed into a Treebank generator

Plans for upcoming Reporting Period

Our plans for this upcoming reporting period **(Oct 2 - Oct 16)** are broken down as follows:

Who	What	Due When
James	Scrape Leetcode problems	Oct. 10
Austin, Ethan	Scrape JavaDocs	Oct. 10
James, Austin, Ethan	Clean scrapes into both an HTML version and plaintext version	Oct. 10
All	Run plaintext versions of software documentation through base POS tagger	Oct. 10
All	Create more code PoS tags or modify existing ones	Ongoing
Austin, James	Finish cyclic dependency network research	Oct. 10
All	Clean up tagged software documentation and manually tag code in documentation	Ongoing
Joseph	Explore automatic tagging for code inside software documentation	Oct. 10