

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: Oct 16 - Oct 23

Summary of Progress in this Period

This period saw the project move into a documentation gathering segment and concluded with movement into training and scraped data parsing. The next period will see both of these topics covered in much more detail.

- Performing a test run of training a Stanford CRClassifier model
- Automatic code tagging
- Tagged leetcode parts of speech
- Javadoc HTML parsing into useful information.
 - Added functionality for different types of included tags
 - Can maintain `<code>`, `<p>`, ``, etc
 - Added detail categories for constructors and fields
- We will be using xpos instead of upos in terms of parts of speech

Pending Issues

There are several issues currently pending to be covered in the next period of the project.

1. Extend javadoc HTML parser
 - a. Finish parsing detail segments
 - b. Link parsed data to actual HTML document it comes from
2. Generate a general HTML parser
 - a. Be able to parse leetcode HTML and javadoc HTML
 - b. At minimum be able to choose what tags to maintain and remove.
3. Find inconsistencies in tagging between java documentation and real tags
 - a. Tag by hand to make this happen.
4. Create training data
 - a. Automatically tag what we can
 - b. Make tool for filling in missing data
5. Work on training the CRFClassifier
 - a. Make tool that we can feed tagged data to and train the model

Plans for upcoming Reporting Period

Our plans for this upcoming reporting period (**Oct 23 - Nov 6**) are broken down as follows:

Who	What	Due When
All	Create more code PoS tags or modify existing ones	Ongoing
All	Clean up tagged software documentation and manually tag code in documentation	Ongoing
All	Start building a set of tagged software	Ongoing
Austin, James	Cyclic dependency network research	Nov 6
Ethan, James	Parse HTML into more beneficial data.	Late November
Joseph, Ahmad	Work on building tagged code training CRF model.	Ongoing
James, Ethan	Summarize the types of errors between Leetcode and Javadoc	Oct 30
Joseph, Austin	Finalize how to extract the labels/tags	Ongoing
Ethan	Summarize a dictionary between the text and code in the output	Oct 30