Senior Design Bi-Weekly Status Report

# SE/CprE 492 Weekly Report #3

**Team**: sdmay21-35
**Team Email**: sdmay21-35@iastate.edu
**Advisors**

Ali Jannesari
Hung Phan

**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**: March 1st - March 14th

# Summary of Progress (Weekly Summary)

This period, the finishing touches were put on the front half of the development pipeline. In addition to tokenization, work was done in integrating the automatic scraping process with the tokenizer (with novel use of java code inside of python).

The tokenizer was completed this period, and according to the testing done so far, it is roughly 96% accurate with a sample size of 30. The tokenizer was initially created with transparency in mind, but additional hooks were added in order to seamlessly connect scraped HTML data with the parser and tokenizer using direct inputs and outputs.

The autotagger was tested and connected to the tokenization software this period, and appears to be working well when the tokenizer behaves. So far, 26 files have been passed through the entire pipeline successfully, with plenty more to come this next period.

# Meeting Notes (3-1-2021)

- Demo of tokenizer / data gathering pipeline
    - All is working great
    - Small bug in parsing negative numbers, will be fixed ASAP
- Deliverables for next week:
    - Fix Bug in Tokenizer
    - Start building a repertoire of auto-tagged training data

# Pending Issues

- Error Fixing in the Tokenizer
    - There are minor tokenization issues when it comes to negative numbers
        - How do we decide if -1 is “-”, “1”, or “-1”?
    - Fixing this issue will solve our major pipeline error
- Begin Autotagging in mass
    - Distribute work between the team to do some manual tagging
    - Build up a good repertoire (~100 tagged documents) for the initial round of training

# Past Week Accomplishments - Current Period - Individual Accomplishments

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

| Who | What | Individual Hours | Cumulative Hours (Starting 2/22/2020) |
|---|---|---|---|
| Ethan | Completing the tokenizer, debugging the tokenizer, integration debugging | 3.5 | 9.5 |
| Austin | Tags and planning future work | 0.5 | 3.5 |
| James | Sentence Splitting completion and debugging | 5 | 12 |
| Ahmad | Presentation slides, planning of training | 1.5 | 2 |
| Joseph | Scraper & Training | 4 | 6 |
| Ekene | Running the Tokenizer with data from leetcode | 2 | 3 |

# Plans for upcoming Reporting Period

Our plans for this upcoming reporting period **(March 1st - March 14th)** are broken down as follows:

| Who | What | Due When |
|---|---|---|
| Austin, Joseph | Finalize Tag Set, Implement Rules File | ASAP (March 12th) |
| Ethan | Debug the tokenization error (negative numbers) | ASAP (March 3rd) |
| All | Autotagging | Set 1 - March 8th<br>Set 2 - ? |
| Ethan, Joseph, James | Integration testing / bugfixing on the training pipeline | March 8th |
| Ahmad | Implementing training model on pipeline | Four weeks from 23rd of feb |
| Ekene | Find additional sources for documentation scraping | March 15th |
|  |  |  |
|  |  |  |