Alexander Scharf
CS410 (Fall 2020)
Course Project

1.) Progress made thus far
- The data set has been extracted for both English and Japanese text
- The labeling of the data set is being done manually and took longer than expected just for English
- Decided to use spAcy as framework and researched its usage as part of tech review
- Ran preliminary analysis on data set (number of words, parts of speech, common words) for both “useful” and “not useful” reflections
- Coding for training model by referencing spAcy sample code
- Preliminary web app created and hosted on personal server (http://alexscharf.com/)

2.) Remaining tasks
- Complete labeling the English data set
- Begin labeling the Japanese data set
- Make adjustments to the code as needed for the Japanese data set
- Make the web app slightly more user friendly
- Conduct user tests
- Update document and clean up code to remove debugging

3.) Any challenges/issues being faced
- I thought I could outsource labeling the training data with Amazon Mechanical Turk, but labeling it required too much domain knowledge - taking more time than I thought.
- Unrelated to the direct goal of this project, but fiddling with the public facing web server took longer than expected. Ended up switching from Apache to Nginx.
- Still a bit unknown how much rewriting the original application will be necessary for Japanese or any other strange bugs like character encoding, especially with web app.