CPRE 491

Weekly Report #4

Team 17

Project Title: Mining and Evaluating Verb tags and Other Important POS tags inside Software Documentation

Date: 10/18/21

Members:

-William Sengstock – Team Leader

-Kelly Jacobson -

-Zach Witte -

-Sam Moore -

-Dan Vasudevan -

-Austin Buller -

-Jacob Kinser-

What we've accomplished in the past week/what we've been researching

-William Sengstock – I continued to work with POS tagging involving NLTK and spaCy for the model using alice.txt. Also, I used the software documentation provided by our client and applied POS tagging and tokenization to that.

-Kelly Jacobson - Building on the model from last week, I worked on comparing spaCy and NLTK. I ran comparisons to see the differences in how each model performed POS tagging. I also compared how each model was different in Lemmatization.

-Zach Witte - I continued to build on the spacy model that my group has been working on. I looked into clustering methods that I could use on the software documentation provided to us and I applied K-means clustering to the data.

-Sam Moore - This past week I was not able to make the meetings due to previously scheduled interviews, but I worked with Dan to test the data that was given to us in terms of clustering.

-Dan Vasudevan - This last week I worked on doing some more analysis with the previous NLP models I created and tested it with another data set. I was able to compare the performance of the models on the software documentation data to the text data and a lot came from it.

-Austin Buller - This past week I continued to work on the model I built in the previous week using word2vec. I focused on using normal text data instead of software documentation and used the model to cluster words.

-Jacob Kinser- This past week I continued my research on word2vec while also testing out clustering. Using the software documentation provided by our client, I test how removing stop words and punctuations affected the documentation. I also started testing different clustering algorithms.

What we're planning to do in the coming week

-William Sengstock – For this next week we will continue to work on POS tagging for software documentation, and analyzing how it differs from normal text. I believe we will be working in larger groups, and comparing more word embeddings so we better understand the differences between them.

-Kelly Jacobson - We are supposed to be working as one big group this week on one big model. I will probably be focusing on making comparisons between models. We want to see what each model does differently, why that might be good or bad, and also report on what we find.

-Zach Witte - This week, we will all be working as one group to write about POS tagging and the best practices when it comes to analyzing software documentation. We will compare what we have worked on and learned so far to help us as well as build upon what we have been working on the past few weeks.

-Sam Moore - This coming week, we will continue to improve our understanding of how clustering works and how it is crucial for improving a model. We will also take notes regarding what exactly we do and how we do it in order to document our understanding and communicate our findings to our clients.

-Dan Vasudevan - We will continue to work on POS tagging software documentation but in a larger group. We will try to add more models and possibly use different data sets too. Additionally, we will write summaries about what we have learned from the assignment.

-Austin Buller - This next week we will focus more on using different datasets and libraries like Stanford NLP to evaluate the differences in outcomes. Knowing the performance differences between these libraries will be important for the final deliverable next semester.

-Jacob Kinser- This next week I plan to highlight more differences in word2vec while also continuing to improve on my clustering work. I will be taking notes of the results of my research to note to our client in our next meeting.

Issues we had in the previous week

-William Sengstock – Recovering from being sick gave me less time to do research on word embeddings with the software documentation. That being said, I feel better now and should be able to apply myself fully for the upcoming week.

-Kelly Jacobson - I didn't really have any issues this last week. Just that the comparison models I had taken a very long time to do analysis, and I was not able to just let them run. This meant I did not have all the important information that comes from these models. Next week, I will probably be working with different data that should speed up the process and I can revisit my design to see if I can speed it up.

-Zach Witte - I did not have any major issues this week. I had some coding errors, and I had some issues importing certain packages in Jupyter notebook.

-Sam Moore - I had a few issues just understanding how clustering worked in general and understanding how I was supposed to use it. I fixed those issues by communicating with my team members and doing more research on the topic.

-Dan Vasudevan - None of the issues I encountered were significant. I just had some small issues importing the new data set I used but I quickly learned how to import it properly.

-Austin Buller - I didn't have any major issues this past week, I only had small issues with getting my model to run correctly.

-Jacob Kinser- This past week I had a few issues getting my clustering algorithm to work, I eventually got a slightly working demo that I plan to continue going into next week.