CPRE 491

Weekly Report #2

Team 17

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

Date: 9/30/2021

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 - For this week I continued researching about NLTK, and then applying my knowledge to an actual model in Jupyter. Using a dataset that my group found online, we trained the model to tokenize, clean, and more to experiment with the functions.

-Kelly Jacobson - This week I followed a guide on creating an NLTK model for language processing. It worked with tweets and basically converted them to a format the program could read and cleaned up the text to get rid of special characters and filtered for only English tweets. Then I performed text analysis and created some models to visualize the data.

-Zach Witte - This week I worked with my group in jupyter notebook to experiment with making an NLTK model using our own data set that we found via kaggle. We cleaned, tokenized, and tried the K-means clustering method on our data. The goal was to get more familiar with the different tools available in NLTK and see how they affect a new data set.

-Sam Moore - This week I worked on researching exactly how POS in Natural Language Processing works, and what models can be used to achieve it. I studied the accuracy of certain models by feeding it data and checking accuracy based on the relationship between the result I expected and the actual result. I also experimented and researched word vectorization using the same toolkits and the same data, which I analyzed for any consistencies or inconsistencies.

-Dan Vasudevan - This week I worked on building an NLP model using data about wine reviews to predict the type of wine based on the review it was given. It worked well because I used pre-trained algorithms in the scikit-learn library to train the data. The two algorithms were the multinomial naive bayes algorithm and the support vector classification model.

-Austin Buller - This past week we began to start implementing what we have been researching the past few weeks. The data I worked with this week were articles titles and their categories. My original plan for the week was to create a model to predict the article category based on the title after I cleaned and vectorized the data. However, I ran into some issues with the raw data so I only got to vectorization.

-Jacob Kinser- The past week I have been experimenting in jupyter notebook with Zach and William to help familiarize myself more with what we have been researching over the previous weeks. Our group looked at analyzing Amazon reviews.

What we're planning to do in the coming week

-William Sengstock - Moving on to this week, I will continue to train NLP models, most likely with different datasets. Along with that, I will implement different word embeddings, such as StandfordNLP and spaCY, to compare and contrast them to NLTK.

-Kelly Jacobson - Similar to last week, I will build another NLP model this week. Instead of using NLTK I will look into also using Glove and SpaCy. I also want to find different data to work with and try to build it mostly on my own, rather than following a guide.

-Zach Witte - This week I will be working in another small group to experiment with different NLP libraries. We want to continue learning about NLP and making models while also trying to decide which libraries will be best suited for our main project.

-Sam Moore - For this week, my partner and I will experiment with a different model and different data, specifically software documentation data. We want to work to train the model and develop a strategy that leads to increased accuracy. We will also research

different ways to train the data and perhaps specific ways regarding software documentation data.

-Dan Vasudevan -For the upcoming week my partner and I will create a different model using a dataset with software documentation. We plan on using NLTK and tokenization/vectorization strategies that can help us train the model to be highly accurate.

-Austin Buller - For the next week I plan to work with a partner to create a new model based on a different data set. The goal is to train the model without giving it the answers directly. We will also use different NLP libraries, tokenization, and vectorization methods to find out what works best with software documentation.

-Jacob Kinser- For the upcoming week, I will be doing similar implementation and research as done over the past week. I will continue to familiarize myself with the different libraries while working in a group with Austin.

Issues we had in the previous week

-William Sengstock - There were not any major problems that arose this previous week, in my opinion. However, some of the functions in the NLTK toolkit were confusing at first, mainly because I am still learning about NLP and all the packages that come along with it.

-Kelly Jacobson - I simply did not have a lot of time last week to devote to this class so I did not like the way the model came out. A lot was just copy-pasted from the guide, and I don’t think I learned a ton when doing that. Hopefully this week I can devote more time to actually understanding what I’m doing.

-Zach Witte - There were no major issues this week. I struggled with fully understanding how to train our model properly, but this just requires more research and time experimenting with NLP libraries in Python.

-Sam Moore - There were not any major problems in this past week, but I did run into some trouble when it came to understanding exactly how the vectorization worked and what the output of it meant. This week, I will do more extensive research into understanding how the process works.

-Dan Vasudevan - Not many issues occurred last week but I think we could've been better at using explicit data cleansing techniques to verify that the tokenization strategies were accurate.

-Austin Buller - I ran into some issues with the dataset I was using. This was caused because python couldn't read or iterate through certain values in the data and it took a while to get everything up and running.

-Jacob Kinser- I ran into some problems trying to process the data. I was getting type errors that took a while to fix.