CPRE 491

Weekly Report #1

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 – Continuing to develop my skills in Python, and combining that with learning the exercises involving NLP and word embedding. Research the different types of vectorization techniques, while noting their strengths and weaknesses.

-Kelly Jacobson- Researched different word embedding techniques and different vectorization techniques, including which might be best for software documentation. Made a tiny python program to apply vectorization to the words in a text file. I did this just to learn a little more about how the techniques we've been researching work with actual code.

-Zach Witte - Researched the best evaluation methods used in unsupervised learning and which ones might be best for our purposes. Looked into different clustering methods and some examples of how these methods can be implemented in Python. Also continued to research more about natural language processing and the different techniques revolving around it in order to get a better understanding of how it works.

-Sam Moore - Researched different algorithms when it comes to Natural Language Processing such as Gradient Boosting and Random Forests. Studied how the advantages and disadvantages of these techniques might play a part in NLP with software documentation. Continued to dive deeper into NLP through analyzing examples, reading research papers, and trying it out myself with online toolkits.

-Dan Vasudevan - In order to get a deeper understanding of the technicalities of how Natural Language Processing algorithms work, I researched different types of tokenization strategies that are used for the algorithms. I compared the positives and negatives of word, subword, and character tokenization and how they can be specifically applied to Software Documentation. I also looked at a previous NLP project that used character tokenization on C code.

-Austin Buller - I researched more types of algorithms for NLP and different types of work tokenization. I compared and evaluated the different types of tokenization to find out word tokenization method would be the most effective on software documentation. I also did the same for word vectorization.

-Jacob Kinser - Continued research on NLP techniques involving different tokenizations, vectorization models, and their strengths and weaknesses. Also discussing which techniques would work best for software documentation.

What we're planning to do in the coming week

-William Sengstock – To take a set of data and implement it in Jupyter Notebook. Using different word embedding techniques which we have been learning about and use them in a real example.

-Kelly Jacobson - Everyone has a similar task but I will be writing a program that performs some NLP tasks on a dataset. So I have to find a dataset to use, learn how to write the program, then write the program in Python with Jupyter Notebook. This needs to be done by Thursday's team meeting.

-Zach Witte- I plan to find a set of data online and create a small program in Python that utilizes the different techniques we have been researching in order to create a basic NLP language model based on the data set.

-Sam Moore - I will find a set of data to process from leetcode or GitHub or another online source, and actually implement a program in Jupyter Notebook to perform NLP on the dataset. This will help me get familiar with what we will be doing throughout the next two semesters and allow me to see how it works with a real-world example.

-Dan Vasudevan - This week I will find a data set I can use to implement a NLP algorithm. This will be my first actual exposure to using python to build an NLP algorithm. I have not decided what type of data set to use but I think it will most likely be related to software documentation.

-Austin Buller - This week I am planning on finding a dataset to practice implementing NLP. The data needs to be cleaned and vectorized using different methods to find out which is more effective. This will be done in Python using Jupyter Notebook and will be due next Thursday during the team meeting.

-Jacob Kinser- At our last client meeting, we were tasked with analyzing a set of data through different techniques we've researched so far. I plan on doing this throughout the next week in jupyter notebook while continuing to learn python.

Issues we had in the previous week

-William Sengstock – Because of differing schedules, sometimes getting everyone to meet at the same time can be difficult. That being said, I believe our group has figured out a meeting time where all members can be present.

-Kelly Jacobson - Our team meeting with the Client last week got canceled so we instead had to record ourselves talking about what we have done and send those to him. Our team meeting this week went well, talked about project requirements/constraints and what to do next.

-Zach Witte- There were some technical issues that resulted in last week's meeting being canceled. We could not meet at a later date due to schedule conflicts, and our TA meeting time got shifted. We adapted effectively, and now everything is running smoothly.

-Sam Moore - Some scheduling issues arose this last week and there was some miscommunication. We kept in touch and did what had to be done to continue without being set back.

-Dan Vasudevan - The only issue that came up was that there was a scheduling conflict and it was difficult to reschedule due to everyone's busy schedules. However, due to good communication from everyone we were able to work out the situation and move forward with the project.

Weekly Report #1

-Austin Buller- We only had some minor problems last week. Our TA meeting had to be moved because of a schedule issue/change. The second issue was caused by some technical problems our client was having and because of that, they missed the meeting.

-Jacob Kinser- We had a technology difficulty at a previous meeting with our client. We discussed the matter and resolved it by all creating a presentation recording on our research from the week. We also slightly changed our TA meeting time to accommodate all members.