EE/CprE/SE 492 GROUP PROGRESS REPORT

Group number: 17
Project title: Mining and Evaluating Verb tags and Other Important POS tags inside Software Documentation
Client: Hung Phan, Hiep Vo, Arushi Sharma
Advisor: Ali Jannesari
Team Members: William Sengstock, Austin Buller, Kelly Jacobson, Jacob Kinser, Sam Moore, Zach Witte, Dan Vasudevan

---

Project Summary:

The goal of our project is to research and experiment with different natural language processing (NLP) techniques to try and find the best way to tag (part of speech) software documentation. We want to use existing NLP models and techniques to create our own model that is best suited for processing software documentation and source code. The direction of the project has not greatly changed, but we think taking a neural network/deep learning approach might be the best way to address the issue of processing software documentation and code.

Accomplishments

Since the last report, we have been experimenting with tokenizers for source code and different types of neural networks.

Kelly - I have spent the past few weeks looking at the CuBERT tokenizer for source code, which is a Google research project. We wanted to experiment with its utilities for tokenizing python source code and compare its performance with the Python Tokenizer that comes from a standard Python library. First, I was just looking at how CuBERT works, including looking at what categories of token labels it has. Most are pretty similar to the Python Tokenizer. Then I was trying to implement a functioning model of CuBERT, but this is where I met failure. Despite my best efforts and consultation with Arushi, I could not get a CuBERT model. Some errors were caused by CuBERT requiring outdated installations of python and tensorflow, and some errors I lacked the understanding to solve. We decided CuBERT was not worth the hassle to get a working model and we will proceed with the Python Tokenizer for now. CuBERT does have some datasets that may still be useful for training whatever tokenizer model we end up with.

Jacob - Over the past few weeks I have looked deeper into different tokenizers. After researching different tokenizers, our group focused in on two different tokenizers - CuBERT tokenizer and python tokenizer. I spent around a week and a half researching CuBERT and attempted to create models to see if it would be beneficial to our research. I learned a lot about CuBERT but was unable to get a working model going. After discussions with Arushi, we decided that CuBERT had a difficult learning curve and that our efforts would be better spent elsewhere for the next couple of weeks. I then switched back to Python tokenizer where I have been trying to get more accurate tags for our tokens.

Zach - My main focus over the past few weeks has been neural networks and transformers. I have been researching what they are, how they apply to natural language processing, and how we might be able to implement them into our project. While I have looked at both the codeBERT model and the codeGPT model, I have mainly been looking at the codeBERT model and similar BERT implementations. I have found a pretrained codeBERT model available on github, and I have been looking into ways to train a BERT model from scratch using source code and software documentation.

Austin - My job over the past few weeks has been to research and understand transformers for neural networks. I have specifically been looking at the CodeBERT model and how it can apply to software documentation. CodeBert can be used as a pretrained model or we can train a neural network from scratch and use the CodeBert framework. After implementing a pre-trained CodeBert model the next step is to train a CodeBert model from a dataset of the group's choice and then to fine tune that model.

William - My responsibilities for the project in the past few weeks have been researching more into neural networks and how transformers operate with natural language processing. Experimentation involving neural networks and code has also been a task of mine, working with other group mates with the CodeBERT mode. I am looking into implementing a model from scratch to utilize source code tokenization, along with software documentation.

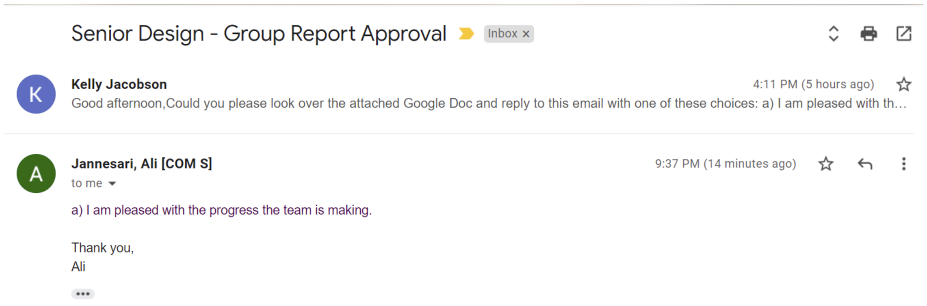
Samuel - My responsibilities for the project these last few weeks have been to look into the python tokenizer, understand what type of datasets we will be using to train this tokenizer, and figuring out the different functions of this tokenizer along with ways we can improve it. These tasks required research and collaboration with my team in order to understand the tokenizer's full functionality and all of its features.

Dhanush - My main focus over the past week was to experiment and improve the python tokenizer for python source code. This first included running the tokenizer on source code and checking its accuracy. Then we identified certain parts of the results that we thought need improvement so we wrote code to fix those parts.

Pending issues

With midterms happening, we have been struggling a little bit to keep up with all of our tasks for this project. This last week, Arushi let us have an extra week to work on our current objectives. We should have enough time by our next meeting to get everything done that we need to.
We have not yet met with Hiep Vo to talk about our project, but Arushi has been in contact with him and we now have some datasets he recommends we use.

Advisor Input/Signature:
Ali Jannesari



Senior Design - Group Report Approval

Kelly Jacobson 4:11 PM (5 hours ago)

Good afternoon,Could you please look over the attached Google Doc and reply to this email with one of these choices: a) I am pleased with th…

Jannesari, Ali [COM S] 9:37 PM (14 minutes ago)

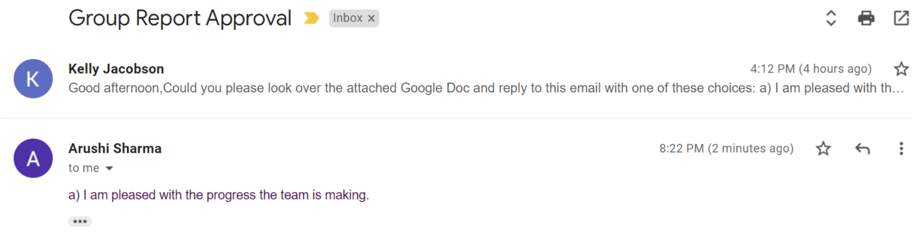
to me

a) I am pleased with the progress the team is making.

Thank you,
Ali

Client Input/Signature:

Arushi Sharma



Group Report Approval

Kelly Jacobson 4:12 PM (4 hours ago)

Good afternoon,Could you please look over the attached Google Doc and reply to this email with one of these choices: a) I am pleased with th…

Arushi Sharma 8:22 PM (2 minutes ago)

to me

a) I am pleased with the progress the team is making.