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

Kelly - Since the last report, I have still been working with tokenizers. I was able to modify the python tokenizer so that we can create our own specific token labels. With Jack, we tried training a hugging face language model from scratch, but ran into issues with the dataset and using the python tokenizer instead of the hugging face tokenizer. Arushi instructed us on what we needed to do to process our data and format it. This last week, I created a program for formatting our data along with creating tokens. It creates three files. One has all of the tokens. Another has all the comments and strings, which we should be able to apply a natural-language tokenizer to. And the last has the original python source code, with all the comments and strings removed. We should soon be able to move on from figuring out the tokenizer to actually processing and analyzing the tokens.

Jacob - Over the last month I have worked on modifying the python tokenizer to better fit our project. With Kelly's help, we were able to break down and add existing tags to make the tokenizer more accurate. Kelly and I also attempted to train a hugging face language model from scratch but ran into issues doing that. We have been working with Arushi on steps for moving forward. More recently I have been helping other members train the codeBERT model. After debugging the model I got it running, then I noticed the estimated completion time was too long to complete on my computer (as I need my computer's resources for other classes) Our advisors are working on getting us VM for our team to use. Our team is waiting to hear back so we can train the model.

Zach - I have been working on training a codeBERT model from scratch using software documentation and python source code to train the model itself. I have been working with Austin, Will, and recently Jack to accomplish this. After following multiple tutorials, I believe we have a model that we can successfully train. We were running into a few problems with training the model on our local machines, so we are getting access to a virtual machine in order to train the model.

William - Throughout these last couple of weeks I have mainly been focusing on attempting to train a codeBERT model from scratch to take in data. When it comes to the dataset that we will be using, I have been collaborating with the other group members on their modified python tokenizer. Along with Zach and Austin, we have been in contact with Arushi on how to go about implementing these models from scratch. That being said, there have been memory issues when it comes to the time it takes the model to run on our computers. We requested access to a virtual machine, and from there will continue our work on building the model and passing in data to it.

Samuel - Throughout these last couple of weeks I have been focusing on finding the most ideal dataset and formatting it in such a way that our model can use it most efficiently. We have run into a few problems regarding finding datasets that are large enough to provide us with ample training data along with validation data. Our model also focuses on python software documentation so that was another limitation we needed to take into consideration. After research with the group, we found a large enough dataset to use and are almost finished with formatting the data so our model can take it as input.

Austin - Since the last report, I have been attempting to train a CodeBERT model from scratch with software documentation data with William and Zach. We have been following a few online tutorials on how to implement the model. This new model will have to be implemented on a HPC cluster in the future as the computer resource demands are now beyond any of our groups computers.

Dan - Over the past 2 weeks I have primarily been focused on finding the best dataset to train our model on. Overall, there are a lot of sources where we could find this data but the difficult part is formatting it in a way our machine learning model can understand. Additionally, a constraint for our data is that it has to be python source code so all other coding languages do not work. After looking through the dataset options I did end up finding some options that fit our requirements.

Pending issues

Memory space is an issue when it comes to running the model we are trying to train from scratch. On our codeBERT model, the completion time well exceeded the limits for our personal computers. For that, we have contacted Arushi about receiving access to a virtual machine so we can run the model. We should receive access fairly soon and will be able to continue on with our work.

Advisor Input/Signature:

Jannesari, Ali [COM S]

to me, William

Mon, Mar 28, 8:58 AM

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

Thank you,
Ali