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

Jacob - Over the last month I have worked on different models surrounding CodeBERT. Originally we were attempting to train a codeBERT model from scratch, but we decided it would be better to fine tune an existing mode instead. We have created a few different models with no success so far, but we are close and hope to have this model done the week of 4/18.

William - Throughout these last couple of weeks, I have dealt with codeBERT; trying to make a model from scratch and fine-tuning one. When my team members and I were trying to make a model from scratch, we ran into memory issues and had to obtain access to a VM for it to run. It is in the queue now to run in the near future, so we plan to have a model done soon. Continuing, I have been attempting to fine-tune a model to support the dataset we have been working on. The goal is to fine-tune the model to include POS tagging for software documentation (source code, per example).

Kelly - Since the last report, I have continued working on data-preprocessing and tokenization. Data-preprocessing is a huge chunk of the pipeline. The model we create requires a prepared dataset of tokens and their labels so that the model can learn how to properly apply labels to new tokens. I created a program that can process a raw file containing python source code and plain English text. It processes the file, applies tokenization using modified versions of the python tokenizer and SpaCy tokenizer, and compiles the tokens and labels into a vocab file that we can easily load into our model. The next steps rely on getting the CodeBERT model to use our own data.

Austin - Over the past month I have been working on training CodeBERT models either from scratch or using a pre-trained model and trying to fine tune the model to work more efficiently with software documentation. In the past week I have started to look into improving the tokenizer we are using which could lead to better performance in our CodeBERT model.

Dan - Over the past few weeks I have worked primarily with fine tuning the CodeBERT machine learning model with POS tagging data. This has included different aspects to it including preprocessing the datasets to match the model's requirements, writing code to read in the data using the pre-existing CodeBERT model and writing code to evaluate the accuracy of the results. In the preprocessing stage, we also had to work with Python tokenizer to label the data in the appropriate manner.

Samuel - Over the past month I have been working with a pre-trained codeBERT model because we thought that would suit our datasets the best. We decided as a team to take the pre-trained model and fine tune it with data that we want the model to be able to handle in the future (POS tagged software documentation). The pre-trained model needed to be adjusted in many ways including the format of the data we train and test it with, the arguments we give the model, etc. Over the past month, I have been working with my team to create a functional model that can be trained on the data that we chose.

Zach - Over the past few weeks, I have been working with codeBERT, and how to utilize it to better process and tag software documentation and code. My main focus was creating a BERT model from scratch, training it using strictly software documentation and code. We ended up switching directions, and the last 2 weeks I started helping other team members with fine tuning an existing codeBERT model to fit our needs for the project.

Pending issues

Our group currently has no pending issues.

Advisor Input/Signature: