October 11 - October 25 Activities Summary

- Computed percentage of medical terms in NPR dataset.
- Developed script to download PubMed data and parse resulting XML.
- Read BioBERT paper, went through source code.
    - Paper is a standard fine-tuning of the standard BERT model (trained on Wikipedia and Book corpus dataset) to PubMed data.
    - Downloaded pre-trained weights for Named Entity Recognition (NER) and pre-processed data from repository and tried running NER on our dataset.
    - Wrote script to generate `test.tsv`. This format is a speculation of the expected format, based on their documentation.
    - This paper presented at NeurIPS 2019 uses a knowledge distillation method to reduce the computational cost of the large BERT model. Their claim is that their method allows for knowledge distillation on the original BERT model, and the distilled model can be fine-tuned on task-specific datasets. This may be an avenue for future work. The authors of BioBERT do not seem to be interested in this at the moment.

Issues

- Their source code contains various tokenizers (all of which I understand). Their pre-processed data is tokenized. However, the code that runs predictions seems to do tokenization (although I'm unsure of this), and therefore I'm unsure whether they expect us to use their tokenizers, or simply let their prediction code take care of it.
- Although they provide pre-trained weights, it appears that you can't do predictions without a training set (used for training models) and/or a dev set (used for hyper-parameter tuning). Both of these require manual labeling, although prediction should not require this.
    - One of the issues in their repository asks this exact question: the response was to comment out some lines. However, on doing this, the code doesn't run, because while running evaluation, the code also creates some mapping file that is required by the prediction code.
    - As an attempt to fix this, I simply copied one of the `dev` sets from their pre-processed data and then ran the code (I'm not sure this is the right thing to do). I used a free GPU provided by Google Colab, since it took too long on a CPU. It seems to have failed on the last stage, stating that it requires some file. There does not seem to be any code in the repository that generates this file (as checked by `grep`). I raised a GitHub issue asking all of the above questions.