October 8 - October 22 Summary

I did two things in these two weeks:

- I was able to create a pipeline in Python. This pipeline performs all the (modified) steps in the final report document shared last week, with two caveats:
    - It is currently a (very) stupid pipeline, and will form pairs out of pretty much every noun it finds.
    - It currently assumes we're using the FLAIR model (which needs to be trained), rather than BioBERT. This assumption was made for two reasons:
        - At the time of programming the pipeline, I still hadn't gotten a reply from the authors of BioBERT on the issue, and it was getting too late.
        - FLAIR offers a Python API along with their command-line program, which easily fits into a pipeline (BioBERT does not do this).

    The pipeline has been pushed to the GitHub repository, shared with Saad.
- Another person commented on the issue, and I now understand the format of the `test.tsv` file. Each line should contain a word, followed by a tab, followed by the class label. I found this odd, but on asking, the owner closed the issue as out of scope of that particular issue. I updated the `generate-tsv.py` file to conform to this standard. For class labels, it simply sets every word to the `B` class (beginning of a term). The idea is that at this stage we merely want the predictions, and we're less interested in the performance itself.
    - I am currently running this model on the assigned VM. If that does not work, **this repository** seems to have code for BioFLAIR. We would simply need to replace the `test.txt` file they use, discarding the middle (part of speech) column, and modify their fine-tuning code to only work with two columns (this is configured in line 4).