December 6 - December 20 Summary

With an established pipeline, my work these two weeks has been running the models. By "model", I mean the model used in the classification stage; all other pipeline steps remain the same. I am trying two models:

- The original model as discussed in the proposal: I am done running this.
- The twin classifier model, where we use the above model, but also run a disease-only model, then remove pairs where both the terms in the pair are (or are not) diseases. This model is currently running. I expect it will take a few more days.

Daniel showed me some code to make API queries to ROBOKOP, that could be used for validation. I will run validation on the pipeline with both models, and the expectation is that the latter, while more complex and slower, will do better. Currently, I expect a high recall but low precision (which I was told was a separate problem in itself).

Seeing the output, I see scope for improvement, as a "post-post-processing" step. For example, even with the twin classifier model, I see pairs such as `(neuropathic pain, pain)`, which are logically not possible, or `(diabetes, diabetic)`. I suspect lemmatization would help here, following which we could simply remove pairs where the result of lemmatization is a pair of the same terms. Occasionally, there are also reversed pairs: `(A, B)` and `(B, A)`. It should be easy to remove these.

With the negative news covered, I am seeing some evidence of a high recall. The model does occasionally give accurate results (which I verified by searching online), such as `(Fosamax, osteoporosis)`, `(marijuana, insomnia)`, and `(diazepam, anxiety)`.