## Resources Used in the Project

Our Genre classifier used Google Colab with a TPU runtime and the free offered extension to the RAM to store all of the BERT embeddings

Our Authorship and NSP classifier used a Google Cloud N1 Ultramem machine with 40 CPUs, 961 GB of RAM, and no GPU running a Tensorflow 1.15 environment.

## Running the Code

The processing.ipynb file contains the code used to process the fantasy datasets. We used the same code base for the realistic datasets, so further realistic datasets could be generated by following the steps outlined below. We have included a zip file that contains all of the .csv files that we used in our experiments, as well as the .txt files that we used for processing.

Creating a new dataset: Given a text file, take one of the sections to process an existing dataset and change the names to reflect the .txt file to be used and the target .csv file to create. Run the notebook to create the new .csv file.

The Full Bert Type Classifier and NSPandAuthorship files can be run directly using our datasets to produce our results provided there is enough RAM to store the BERT embeddings. The comments in these files indicate what each part of the code does, and should allow for modifiable experiments.

In the NSPandAuthorship.ipynb file, the test can be switched over to doing the individual dataset test described in the writeup by switching the df_list input in getFeatures() with a custom list that only includes that particular dataframe. After the list is switched out, the used_labels list should be updated and the number of examples can be increased to use the full dataset. We used 10000 max examples to get the full datasets when we ran the code.

***** Our figures were generated by collecting data in individual runs and placing them manually in data structures. This is due to the large amount of RAM that BERT embeddings use preventing us from storing all of the tests in a single runtime. If you would like to recreate our figures, run each of the tests described in the writeup manually and then plot the data separately.*****