PubMed

The implementation is all contained in the repo: https://github.com/kpetrakis/PubMed .

Due to the large size of the files, I didn't include any of them in the repo. To have basic reproducibility you must either place "PubMed Multi Label Text Classification Dataset Processed.csv" file in data/raw directory and run "python parse_data.py" or download the data tensors directly in the train/ dev/ and test/ directories from this Google Drive:

https://drive.google.com/drive/folders/16k8-PbThCXYkJygvtTpiFyC_OkS26cyu?usp=sharing

I recommend the second option as the encoding part might take a while (at least in my laptop it did!)

I only ran the experiment with sequence length 256 due to compute and time limitations.

I implemented the EDA part in a Jupyter Notebook (mainly due to plotting) and placed it in notebooks/ directory.

For the evaluation I used the basic metrics found in the literature for multi-label classification problems.

I do not consider this implementation to meet by any means production level code standards or results. I just tried to give you a basic outline of a project structure to tackle the given task.

Notes:

Due to the lack of a descent GPU on my part I had to run the experiment training and evaluation runs on Google Colab.