Max Calehuff
Kyrill Rekun

Project: **NLP Approach to Multi-label Classification of MRI Reports**
Github: https://github.com/UrbanPancake/deep-learning-project

Week #1 Progress Report

The first thing we spent time doing was cleaning, parsing, and tokenizing the MRI text. This was a bit of a pain because of the weird formatting from when the text was pulled from the source. Ultimately It was just some regex work to get everything cleaned and splitting correctly. Next we tokenized the MRI text just like we have been doing in class.

Here we began to think about creating a dictionary of keywords that identify which body part/organ the MRI is being done on. For example, we want to flag MRI reports about lungs as having lung related words because that means this report has something to do with potential lung diagnoses. Since we have 16 non-exclusive classes in our multilabel problem, we want to ideally identify reports by subject matter and then subsequently use only lung related data for identifying lung diagnoses. We have a basic understanding of how to group and split the different diagnosis from our research, but Max will double check with his mentor who should have more domain knowledge if diagnosis segmentation is correct and viable.

For now, we focused on replicating the binary classification model from fastai in pytorch using GRU and LSTM. Kyrill wrote most of the code for this model but ran into some technical issues with running the model. His local NVidia 1080 GPU just wouldn't finish running the model and would crash midway. From the results from a small number of epochs we should be able to get within range of the fastai model after a bit more training. Max plans to rerun the notebooks on his practicum gpu to get the actual results.

Next week is fine tuning the binary model and then decided what model pipeline we want to use to shrink the multilabel problem into a more manageable task.

Responsibilities:

Max: Research work, running model on GPU
Kyrill: Text cleaning, Tokenization, and writing the binary classification model code.