NLP Progress Report - Robin Cosbey and Josh Loehr

DATA PREPARATION

Given the .csv representations of the journal files in our corpuses, document files containing one sentence per line have been generated. The main body of each journal was kept and all other sections were removed (e.g. Acknowledgements, References, Citations, Appendices). We have also replaced headings, numbers, and references with single tokens. Additional processing has included the removal of stop words and stemming. Similar processing was completed with the abstracts corresponding to each journal. With this information, TF-IDF scores have been produced for each sentence as well as each sentence in the abstracts. Each label is the comparison of the journal sentences with each abstract sentence by way of cosine similarity. We plan to supply several features to our model including the sentence representations, the heading the sentence is within, and sentence length.

LEARNING/TEST PROCESS

Functionality for a data loader and RNN model have been implemented. The data loader takes in the label and feature files to produce numpy arrays ready to be fed into the model as well as representations of the document files for evaluation purposes and a file containing the number of sentences in each document for use by the model. The RNN model creates the graph with the provided file dimensions and trains the model. After training has been completed, the trained model is run with the test data set and extracted summaries are produced and stored.

FUTURE STEPS

1. Run model with binary label and feature files, produce summaries
2. Evaluate summaries produced
    a. By human comparison to the original abstracts
    b. ROUGE-N