Weekly Report 5

Jeremy Lachowicz

2/19/2022

1 Accomplished this week

I make a preprocessing library I can use to clean the text from the files. This includes number removers, word and sentence tokenizers, stop-word removers, punctuation removers, and various other symbol removers. I ran the texts through this and created a preliminary Word2Vec model for each of them. My biggest problem this week was finding a word tokenizer that worked with Ancient Greek. The CLTK word tokenizer that I had planned on using didn't work, so I ended up finding another tokenizer made for Greek by spaCy.

2 Accomplish next week

Extract the associated vectors from the models to calculate cosine similarity, and fine-tune accordingly. One of the main things I want to test this week is the impact preprocessing steps have on the Word2Vec accuracy. Since I have such small amounts of data, I think keeping stop-words and puncuation might lead to a better model, so that will be tested this week. I also want to find a way to create 2-grams and also see how this impacts the results (last week I used 1-grams).