Austin James
CSCE 578
Project Proposal
3/31/2019

For my final project I propose to do a sort of combination of my second and third projects. I will take the GNU dictionary are use a vector space model and cosine similarities to compute a thesaurus using tf-idf values of dictionary entries. I think that using a vector space model with these values could solve some of the problems I came across with term variance when I was just comparing word-to-word.

Using a cosine similarity function will also allow me to tune my results in order to achieve the most accurate thesaurus. I can track average synonyms per words and also how many synonyms are valid when compared with the Moby thesaurus to tune my results. Similar to third project, I am planning on plotting these both on a scatter plot in order to find the best value.

I will then compare my generated thesaurus entries to those from the Moby Thesaurus in order to determine their accuracy. I can then compare accuracies with my other thesaurus that was made with word-to-word comparisons. I am hypothesizing that the tf-idf thesaurus will generate both more and more meaningful synonyms for each word.