Austin James
CSCE 578
Assignment 02 Analysis
02/27/2019

In this assignment I used the GNU dictionary to generate a thesaurus using the similarity of significant words in the definitions of terms. I then compared this generated thesaurus to the Moby Thesaurus to see how many synonyms appeared in the entries of a term in both thesauruses. I found that for some entries the generated thesaurus was quite accurate, with 66% or more of generated synonyms appearing in the Moby Thesaurus. However, other entries contained no synonyms that appeared in corresponding Moby Thesaurus entry.

I found that the terms that the thesaurus generator really had a difficult time with were the scientific words. I think this is because these terms don't really have synonyms. These terms have very specific definitions, and they can't usually be replaced by another term. The synonyms that were generated for those terms could be more accurately described as related terms.

I initially chose to define a synonym as a word whose definition contains 33% or more of the significant words in the definition of another term. I found that this produced far too many synonyms for each entry. I then decided to increase the percentage to 50%. This produced a reasonable amount of synonyms and a nice percentage of common synonyms for some words. I found that any increase beyond 50% began to produce some fairly odd results, and I did not think that these results were acceptable. Some words began to have no synonyms generated and some had thousands.

Some unexpected things that affected the results of the program came from the GNU dictionary. I was unable to get the GNU dictionary parser available online to integrate properly with my code, so I had to create my own. My parser removed all tags and tagged terms from

definitions which can affect the significant words in a definition. Also, if a term had multiple definitions, my parser combined them into one definition for the purpose of finding synonyms. This could probably reduced the number of synonyms generated for words with multiple definitions because other words had to have more significant words in common with the original word to be determined to be a synonym. My parser also removed stop words and number words from one to ten (“one”, “two”, etc.) from all definitions.

My conclusion after this assignment is that our language is far too complex for a program to create an accurate thesaurus from an average dictionary. While some entries the program generated were fairly correct, the vast majority were not. I think that the only way for a program like this to create an accurate thesaurus would be if it was generated from a dictionary that used a limited set of term in definitions in order to ensure that synonyms actually had an acceptable percentage of significant words in common.