Austin James

CSCE 578

Assignment 3 Analysis

4/23/2019

For this project I attempted to generate a thesaurus from a dictionary using cosine similarities of definition vectors in a vector space model of definitions. I used tf-idf values to form normalized vectors, creating a space that was n-dimensional, where n was the number of different words used in the entire set of definitions. I then compared the thesaurus entries I generated to the corresponding Moby thesaurus entry, to see what percentage of my generated synonyms appeared in an edited thesaurus.

In order to decrease run-time, I did not create the full n x m matrix, where n was the number of terms and m was the set of all words used in definitions. Instead I used a dictionary to store the vectors and only added values that were greater than 0. This allowed me to only compare like term values in different definitions, and I avoided comparing all term values against all term values.

After some calibration, I decided that cosine similarity of 0.4 or better was enough to prove a synonym. Calibrating this value for this project was tougher than for my assignment 3 program because there is not one metric, such as average cluster size, that can give an indication of where the threshold for similarity should be. In order to find that threshold in this assignment I had to use the eye test on some output and see what created the most logical thesaurus entries. I was also able to use the comparison output to help find which value produced the highest percentage of synonyms that appeared in the Moby thesaurus.

This approach worked much better than my approach to the same problem for assignment 2. Tf-idf comparison of definitions produced a much more logical set of thesaurus entries than simply matching terms who had a given number of similar terms in their definitions. Some specific examples I noticed were that entries for atomic elements and animal species were much more logical. This program was able to pair element abbreviations with name, and also was able to cluster words pertaining to a specific animal species or family, with little other synonyms being added. I thought that this provided a certain level of proof that using tf-idf values certainly improved accuracy, because dictionary entries for those kinds of terms tend to use rare words that would only pertain to closely related terms.