Shaan Mistry
13 March 2022
Professor Long

Assignment 7 WRITEUP.pdf

k-Nearest Neighbors Algorithm

The distances between the feature vectors of each text were computed with three different formulas. Each formula had different pros and cons depending on the input to the program.

Manhattan Distance

The Manhattan distance was most affected the limit of a noise filter. Its accuracy increased greatly as the limit for the noise filter increased.

Euclidean Distance

The Euclidean distance performed similarly to the Manhattan distance. However, it stayed more consistent when the size of the inputted sample passage was changed.

Cosine Distance:

This biggest difference between cosine distance and the first two distance metrics is that cosine distance is only affected by the words that are the same. If the word is only in one of the texts, then the vector component for that word is 0.

Observations

Varying Passage Size

I noticed that the program identified larger passages much better than smaller passages. I tested this by inputting varying sized William Shakespeare passages. As I increased the size of the anonymous sample passages, the program was more likely to have William Shakespeare as the 1-nearest author.

Varying Noise Word Limit

I noticed that as I change the limit of the noise words, the top k-Nearest changed based on the size of the inputted passage.

There seemed to be a golden ratio of the noise limit based on the sample. The program would fail if the noise limit was too small or too large based on the relative to the size of the sample text. If the size of the anonymous sample text is smaller, than limit would be smaller, and if it was larger than the limit would be larger.

Conclusion

The success of the program depends on the size of the texts in relation to size of the noise limit. As the input sizes decreases, the top k-Nearest neighbors have very similar distances which makes it difficult to identify the true author. This is because smaller passages have less data so the algorithm will make smaller distances. A smaller size also makes passages more affected by the bloom filter as a couple words in the filter could completely change the outcome of the program. Therefore, we can conclude that using large samples with a relative-sized noise filter will make it easy for the program to identify the correct author. In addition, the metric choice also plays a large role on the ability of the program to identify the correct author. The Euclidean distance seemed the most accurate overall. However, if the anonymous sample is small, it may be better to use cosine distance because it isn't affected by words that aren't in both texts. This means it can compare a very large sample text and a very small text much better than its counterparts. Likewise, if the anonymous sample is large, the Manhattan distance will perform well if paired with a large noise filter.