While preparing to use the previous labs for this assignment, I came across some issues. I noticed that some of the feature vector words had a TF-IDF value of 0, which should only occur for words that are found in every document. These noise words should've been filtered out. After extensive investigation, it was discovered this was due to the method in which TF-IDF values were selected in lab1. Essentially there was a list of TF-IDF values that were being sorted in the wrong order (ascending vs descending), and the lowest values were being selected instead of the highest. Lab1 was modified to behave as expected and additional tests were ran. This modification resulted in a more sparse, but richer dataset. TF-IDF values near the overall mean were selected to be placed in the output feature vectors to combat sparsity while maintaining meaningfulness. Lab2 was used to determine impact on Classifier accuracy (as a case study into the effectiveness of the new classification). As the tables show, The new methodology is slightly more accurate, though feature vectors take about 2.5x longer to make (Classifier on-line, off-line cost is equivalent).

Bugged Implementation

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | KNN (300 words) | Decision Tree (300 words) | KNN (800 words) | Decision Tree (800 words) |
| Avg # Words | 300 | 300 | 795 | 795 |
| Accuracy(%) | 40.38 | 32.33 | 39.76 | 33.63 |

Avg Num Words: Number words used in each Cross Validation Trial

Accuracy: (#correct classes / #correct+#incorrect)

Fixed Implementation

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | KNN (300 words) | Decision Tree (300 words) | KNN (800 words) | Decision Tree (800 words) |
| Avg # Words | 300 | 300 | 792 | 792 |
| Accuracy(%) | 41.20 | 37.73 | 53.21 | 46.02 |

Avg Num Words: Number words used in each Cross Validation Trial

Accuracy: (#correct classes / #correct+#incorrect)

Representative Sampling was used in this lab for the similarity calculations. Running the baseline/True Jaccard similarity on the entire corpus of documents required about 22,000^2 calculations. This was allowed to run for 5.5 hours and resulted in 50% of the document pair comparisons being completed. If allowed to continue, it would thus yield an estimated time of 11 hours for the entire corpus. The Representative Sampling methodology from lab4 was employeed to work with a smaller amount of data: 5,000 points. This smaller data sample requires 5,000^2 comparisons, about 1/20 as many. The real time values are even better though, it takes the small dataset 4.5 minutes to run all Jaccard similarity pairs compared to the 11 hours of the full dataset, a savings of 99.3%. This additional savings could be due to additional code changes that took place in between trials, namely treating word vectors as bit vectors (allowing bitwise comparisons) instead of lists of booleans (required iterating over entire list).