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 |  |  |
| Accuracy(%) | 41.20 | 37.73 |  |  |

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

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