#6 Extra Credit

Several changes were made to improve the accuracy of the classifier:

1. Stop words (such as the, he, or she) refers to the most common words in a language. They are mostly neutral, so should be filtered out in computing. Therefore, a stop word list is added to the algorithm. Before the training phase, the stop words list is loaded from the disk (stop\_words.txt) and store into a separate hash set, and in both training phase and test phase, the algorithm refers to this set and ignore all the words in this set. By doing so, the accuracy is improved by approximately 1% before any other optimization is made.
2. Laplace smoothing replaces the original smoothing technique. The final value is set to 0.001. The chart below illustrates the error rate (in percent) regards to the Laplace k value.

|  |  |
| --- | --- |
| Laplace K | Error rate (%) |
| 0.0001 | 3.0 |
| 0.0005 | 3.0 |
| 0.001 | 3.0 |
| 0.005 | 3.4 |
| 0.01 | 3.4 |
| 0.05 | 4.2 |
| 0.1 | 4.6 |
| 0.5 | 5.6 |
| 1 | 6.4 |
| 5 | 8.8 |
| 10 | 11.2 |