SVM Implementation Ideas

For the implementation of spam emails using SVMs, we tried several ideas to improve our test set accuracy. In order to choose what words the SVM will use to consider the emails as spam/ham, we only added a few more words to our words text file. This gave us an accuracy of about 85% accuracy consistently. We then tried to change our C value; we changed C from 1 to 10, but our accuracy did not get any better. We then changed C to much smaller values (down to .01), but similarly, we did not get a better accuracy. We found that a good number was 1.75.

Since our accuracy was still below 90%, we changed other parameters. First, we modified our list of words. This is when we saw the greatest improvement in accuracy. We tried using all unique words in the emails, but this proved to be ineffective as the program took too long to run (upwards of 10 minutes). We modified the list of words to include the most frequent words found. More specifically, we chose all the words that appeared over 180 times over the 8000 emails. This resulted in a list of about 1300 words. This brought our accuracy up to about 96% very consistently and this improved our runtime.

Next, we played with the number of training examples. We started off using 6000 training examples, and 2000 cross validation examples. We modified this to 5000 training examples and 5500 training examples, and we found that 5500 training set is a good medium and 2500 emails for the CV test set. We figured that having more CV data would be beneficial. After using an extended list of words, we went back and modified C again, lowering it down from 1.75 to .05, and found that a consistent value was .01. This raised our accuracy to 97%, but we never achieved 98%.
Lastly, we wrote a function that would show us which files the program predicted incorrectly, so that we could examine those emails further. We were unable to find any useful patterns that would help us.

We ended with a consistent 97.6% test/cross-validation set accuracy. This is the model we submitted.