Brief Report

**Things we try but didn’t work out:**

We tried using rbf kernels and tuning gamma but it performed much worse than linear kernel. We also tried counting words that appear most in the spam/ham and not the other, but it hurt the precision and recall so we went with the most used words in the 10,000 emails. We also tried using the -j, -i, and -b options but found no value that helped increase the accuracy.

**Things we implemented:**

We also tried 1000, 2000, 3000, 4000, and 5000 most used words and decided to go with 3000 most used words because anything more than that took a lot of time to produce the svm files and also did not improve the precision or the recall. We extracted the words from the subject and the body separately and found that it helped improve the accuracy and the precision/recall. We tried different values of C and decided that 0.085 gives the highest value of precision and recall. We use the first 6000 emails for training and the rest 2000 emails for cross-validation.

**Observation:**

We wrote a program to check which emails the svmlight misclassified. But we noticed that even us, human, have a hard time telling if those emails are spam or not.