PROBLEM 3 Bob Wagner

Program Output:

```
[#]              True Spam  True Ham
Classified Spam  TP         FP
 Classified Ham  FN         TN

[1]              True Spam  True Ham
Classified Spam  320        4
 Classified Ham  80         396

[2]              True Spam  True Ham
Classified Spam  312        3
 Classified Ham  88         397

[3]              True Spam  True Ham
Classified Spam  357        51
 Classified Ham  43         349
```

To juxtapose the performance of the Naive Bayes spam classifier, I will use the first confusion matrix as a base/control case, as all words were included and left as is. As you can see in this matrix, nearly all Ham messages were classified correctly, but a significant portion (20%) of the Spam messages were incorrectly classified as Ham. The performance of this classifier can be argued both favorably and unfavorably. Personally, I think that the incorrect classification of Ham messages should have a significantly higher weight than incorrectly classifying Spam messages; missing a Ham message because of this reason could have some unintended and serious consequences relative to the task of sifting through junk mail. In my opinion, this would be the "best" performing classifier out of the three confusion matrices.

In the second confusion matrix, all words were converted to lower case. We observe that the Ham classification marginally improved, whereas the the Spam classification became marginally worse. *This suggests that case does not play an important part in the efficient classification of messages into Ham or Spam.*

In the third confusion matrix, only words in the To, From, Cc, and Subject fields were used to train and classify messages. This essentially reduces the vocabulary to email addresses and typically short subject lines. We see that the correct Ham classification is *severely* impacted with a 12% increase in incorrect classification. However, the correct classification of Spam messages is somewhat improved. *This suggests that the information contained within the To, From, Cc, and Subject fields should be a larger factor in the classification of messages, but that the body of the message is critically important as well.*