Problem 3

Results on the program (Manually formatted into matrix):

Original Case Results:

	True Spam	True Ham
Classified Spam	320	4
Classified Ham	80	396

Ignore Case Results:

	True Spam	True Ham
Classified Spam	312	3
Classified Ham	88	397

"Limited Fields" Case Results:

	True Spam	True Ham
Classified Spam	357	51
Classified Ham	43	349

Overall, the various Naive Bayes classifiers did pretty well across all cases, with all accuracies greater than 0.88 over the entire test sets.

It was suprising that we lost net accuracy when we ignored the case on the words. I expect this to help reduce "overfitting" since the difference between lower and upper case words is relatively arbitrary. However, on second look at the data, I realize that the spam commonly uses UPPER CASE LETTERS to emphasis a point, so there is some information encoded in the case on letters.

The "limited fields" case did surpising well. While only using about 4 fields per message, we still were able to achieve over 0.88 accuracy. We did not do as well classifying hams, but the increased accuracy in classifying spams makes up for this. And with fewer words to deal with, I can imagine that using only the fields is desirable for major organizations who need to deal with large number of emails.

As for a better Naive Bayes, I would definitely still consider the To, From, CC, and Subject fields. I would also consider the some of the tags that occur within the email body. These HTML tags (e.g <div align=3D"left">) make the email seem colorful and are usually employed in advertisements and marketting. Many of the spams have this intricate, fancy typesetting but the simple hams lack this, focusing on practical sharing of information and text. I just noticed this of many spams but very few hams in the test data.