CSE 465- COMPUTER ASSIGNMENT 2
NAIVE BAYES SPAM DETECTION
(DUE DATE: APRIL 17'17)

1. The dataset used in this assignment is a subset of 2005 TREC Public Spam Corpus. The data is found in the `data.zip`.
   Each line in the train/test files represents a single email with the following space-delimited properties: the first is the email ID (in the form /xxx/yyy), the second is whether it is 'spam' or 'ham' (non-spam), and the rest are words followed by their occurrence numbers. (Note that numbers may be words, so don't worry if a line contains multiple numbers in a row). The data has been pre-processed to remove non-word characters.

2. Using the training data, compute the prior probabilities $P(spam)$ and $P(ham)$. What is $P(spam)$?

3. Determine the vocabulary and compute the conditional probabilities $P(w_i|spam)$ and $P(w_i|ham)$ using the Laplace smoothing discussed in class. In this context we consider each word as a training example, so n is the total number of words (in either ham or spam documents) and n_c is the number of times w_i appeared in those documents (including multiple occurences in the same email).
   What are the 5 most likely words given that a document is spam? What are the 5 most likely words given that a document is ham?

4. Use these probabilities to classify the test data and report the accuracy (i.e. the percentage of correct classifications).

5. If you were a spammer, how would you modify your emails to beat the classifiers we have learned above?

6. Submit:

   (a) A high-level description on how your code works.

   (b) The accuracies you obtain.

   (c) If all your accuracies are low, what have you tried to improve and what do you suspect is failing.

   (d) Your code