## Spam Filter

The dataset we will be using is a subset of 2005 TREC Public Spam Corpus. You can find it as ***data.zip*** in the same folder as this description on Blackboard (under Assignments/Final Projects/Project2). It contains a training set and a test set. Both files use the same format: each line represents the space-delimited properties of an email, with the first one being the email ID, the second one being whether it is a spam or ham (non-spam), and the rest are words and their occurrence numbers in this email. In preprocessing, non-word characters have been removed, and features selected similar to what Mehran Sahami did in his original paper (PDF provided in the same folder under Assignments/Final Project/Project2) using Naive Bayes to classify spams.

Implement the naive Bayes algorithm classify spam. Use your algorithm to learn from the training set and report accuracy on the test set.

Try various smoothing parameters for the Naive Bayes learner. Which parameters work best?

**Extra Credit:** Features selected makes learning much easier, but it also throws out useful information. For example, exclamation mark (!) often occurs in spams. Even the format of email sender matters: in the case when an email address appears in the address book, a typical email client will replace it with the contact name, which means that the email is unlikely to be a spam (unless, of course, you are a friend of the spammer!). Sahami's paper talked about a few such features he had used in his classifier. For extra credit, you can play with the original files and come up with useful features that improve your classifier. ***Index.zip*** (in this folder) contains the list of the files used in train/test.