Problem description: Spam classification

(Please Do Not Circulate the materials used in our homeworks outside of the class.)

Due date: October 27, 12:00pm (October 28, 00:00am) GMT+08:00

===

In this problem, we will ***use the naive Bayes algorithm to build a spam classifier with Laplace smoothing***. y =1 represents "Yes, it is a spam", y = 0 represents "No, this is not a spam".

As we have learnt in previous class, in order to get the text emails into a form usable by naive Bayes, we need to define a proper and effective "Dictionary" and preprocess the emails. This work could be time-consuming and requires some extra technique. To make this homework simpler, we've already done some preprocessing on the messages and the work to extract feature vectors out of the email documents has also been done for you, so you can just ***load the matrices (MATRIX.TRAIN, MATRIX.TRAIN.50 ~ 1400, MATRIX.TEST) with function "readMatrix" in Matlab (refer files "readMatrix.m" "nb_train.m" and "nb_test.m" for call format)***. The "dictionary" that we used to get the matrices is stored in file "***TOKENS_LIST***".

(a) Use the code outline provided in ***"nb_train.m"*** to train your parameters. Train your parameters using the document-word matrix in ***MATRIX.TRAIN***

(b) Use the parameters you get in (a) to classify the test set data by filling in the code in ***"nb test.m"***. Report the test set error on ***MATRIX.TEST***

(c) Using the parameters fit in part (a), ***find the 5 tokens that are most indicative of the SPAM class*** (i.e., have the highest positive value on the measure above). The numbered list of tokens in the file ***TOKENS_LIST*** should be useful for identifying the words/tokens.

(d) Repeat part (a), but with training sets of size ranging from 50, 100, 200, . . . , up to 1400, by using the files ***MATRIX.TRAIN.****. Plot the test error each time (***use MATRIX.TEST as the test data) to obtain a learning curve (test set error vs. training set size).*** You may need to change the call to readMatrix in nb train.m to read the correct file each time. Which training-set size gives the best test error?

DO NOT COPY