CSC570AG Machine Learning
Homework 3

Spam filter based on a Naïve Bayes classifier
(due by the end of the day on Sunday, March 26th)

In this homework you will implement and evaluate a Naive Bayes classifier for spam filtering.

Download the dataset files hw3_train.zip and hw3_test.zip. The dataset is divided into two sets: a training set and a test set. Each set has two directories: spam and ham. All files in the spam folders are spam messages and all files in the ham folders are ham (non spam) messages. Use the training datasets to train a Naive Bayes classifier for text classification. Test the classifier's performance on the test set. Build word clouds for the training ham set and for the training spam set and compare them. Write a short report providing a) the steps of your work; b) a comparison between the ham and the spam word clouds; c) an evaluation of the performance of the classifier.

Commands you might need, which are not on the lecture slides:

VCorpus(DirSource("directory name"))	to build a volatile corpus from a directory
c()	to combine corpora.

All other commands you need to solve the homework are on the lecture slides. Your submission must consist of two text files:
- a text file, description.txt, no longer than a page
- a text file, script.txt, with the clean history (without any drafts, errors, or debugging) of your session. Provide short comments for your commands.

Let me know if you have questions.