Assignment 4 (due Nov 19) (6 points)

In this assignment you will be given a dataset (train and test) of documents containing 20 directories, where each document contains the text belonging to one newsgroup. You can download the dataset from this link:
http://qwone.com/~jason/20Newsgroups/20news-bydate.tar.gz
Perform the following tasks, and report the results accompanied by the R script.

Part 1

1- Load the dataset in R.
2- Preprocess the dataset using the following steps (hint: you can use the tm package, tmMap(Corpus,Function)):
   a. Remove XML from the document
   b. Remove the author of the message
   c. Remove stop words
   d. Remove extra spaces
   e. Transform all upper cases to lower case
   f. Remove punctuations
   g. Remove numbers
3- Create document-term matrix using TFIDF and stemming options in DocumentTermMatrix() function in the tm package.
4- Convert the DT matrix to the R data frame.
5- Train a SVM classifier for classifying documents on *comp,* that is, consider all documents in comp.* as in the positive class. The remaining documents belong to the negative class. Report the sensitivity and specificity on the test set.

Part 2

1- Combine the training and test datasets (the preprocessed version).
2- Implement a 5-fold Cross Validation (CV) and perform the classification task using a SVM classifier. There are two ways to implement CV. You can use the cvTools library, with documentation downloaded from;
   cran.r-project.org/web/packages/cvTools/cvTools.pdf
   In their example, they use regression, which you can change to SVM.

   Alternatively, you can implement the 5-fold CV on your own and build the confusion matrix on your own. The advantage of doing so is that you consolidate your understanding of CV. Whichever way you implement CV, report the sensitivity and specificity based on CV.

   In class, we emphasize that a good practice is to perform k-fold CV multiple times. For simplicity of the assignment, just conduct a 5-fold CV once. Use the seed 340.

Hand in: Apart from submitting your R script and reporting the sensitivities and specificities of Parts 1 and 2, comment on why there are differences between the values in Parts1 and 2.