Lab 11

SUBMISSION REQUIREMENTS: Work out the R scripts and submit:

1. This document containing the R code you have worked out. Paste the R code under each of the questions.
2. R file for this lab.

Similar to the classification example given in this lecture, process and classify the newsgroup document data. You can download this data from:

http://qwone.com/~jason/20Newsgroups/

and save it on your computer, for example in your "…/news/text/" folder. Note that the data is separated into one test and one train folder, each containing 20 sub folders on different subjects. Choose 2 subjects to analyze (we will use sci.space and rec.autos in this lab) and 100 documents from each.

Hint: To select 100 documents from a folder, use the following code –

```
Doc1.Train.Source <- DirSource(paste(getwd(),"/20news-bydate-train/sci.space",sep=""))
Doc1.Train <- Corpus(URISource(Doc1.Train.Source$filelist[1:100]), readerControl=list(reader=readPlain))
```

a) For each subject select:
   - 100 documents for training from the train folder
   - 100 documents for testing from the test folder

b) Obtain the merged Corpus (of 400 documents), please keep the order as
   - Doc1.Train from the "sci.space" newsgroup train data
   - Doc1.Test from the "sci.space" newsgroup test data
   - Doc2.Train from the " rec.autos" newsgroup train data
   - Doc2.Test from the " rec.autos" newsgroup test data

c) Implement preprocessing (clearly indicate what you have used)

d) Create the Document-Term Matrix using the following arguments
   - minWordLength=2
   - minDocFreq=5

e) Split the Document-Term Matrix into
   - train dataset containing rows (1:100,201:300)
   - test dataset containing rows (101:200,301:400)

f) Use the abbreviations "Sci" and "Rec" as tag factors in your classification.

g) Classify text using the kNN() function (use k=3)

h) Display classification results as a R dataframe and name the columns as:

- "Doc"
- "Predict" - Tag factors of predicted subject ("Sci" or "Rec")
- "Prob" - The classification probability
- "Correct' - TRUE/FALSE

```
> result
  Doc Predict      Prob Correct
1   1     Sci 0.5000000    TRUE
2   2     Rec 0.5000000   FALSE
3   3     Sci 0.5000000    TRUE
4   4     Rec 0.5000000   FALSE
5   5     Sci 1.0000000    TRUE
6   6     Rec 0.5000000   FALSE
7   7     Sci 1.0000000    TRUE
8   8     Sci 0.5000000    TRUE
```