Text Mining

Assignment: 1

(15 points)

Q.1 (6 Points)

a. Use Reuters-21578 train data set to create a vocabulary of size 500 using 'tm.py' package provided with this assignment. Start with creating an object of 'TextMiner' class by initializing it with a TMSK style properties file. Note 'tm.py' uses a properties file just like TMSK. Create a properties file with no special parameter settings and properly set all the tags (especially bodytags and doctag) in the properties file. Also ensure that 'tm.py' package is in the same directory from where you are running your python code. Use 'tokenize' and 'mkdict' functions implemented in the package.

Or if you wish to use TMSK

Create a properties file for mkdict for Reuters-21578 train data set with no special parameter settings and run using a size of 500. Don't forget to set all the tags (especially bodytags and doctag) in the properties file and ensure that it is in the current directory where the java command is being run.

b. Now create another dictionary of size 500 after removing stop words and stemming the tokens. Use 'stopwords' and 'stem' functions implemented in the package. Run and note the differences in the dictionary from the previous run.

Or if you wish to use TMSK

Modify the parameters in the properties file to use stop words and stemming (use the files provided in AuxFilesLesson1.zip linked to in Lesson 1); run and note the differences in the dictionary from the previous run.

c. Generate a local dictionary for the category "earn". This dictionary will be built from documents that have the topic "earn". Use the 'mkdict' function implemented in the package to understand how you can pass category label to generate a category specific dictionary file.

Or if you wish to use TMSK

Generate a local dictionary for the category "earn". This dictionary will be built from documents that have the topic "earn". Check the dictionary file. Lets say you don't want any of the numeric features or ones that include the characters #, & and -. You could edit the file and delete these manually. Or else you could edit the properties file and regenerate the local dictionary. (hint: add the characters you want to exclude to BOTH whitespace-chars and word-delimiters). Do it this way. Note that the dictionary is still not perfect, but we will let the prediction programs decide which features to use. Submit the properties file you used as the solution for this question.

Q.2 (5 Points)
Use the dictionary obtained in Q1(c) to generate vectors for the category "earn" for both the training and test data sets. Use the 'vectorize' function implemented in the package to perform this task.

Or if you wish to use TMSK

Use the dictionary obtained in Q1(c) to generate vectors for the category "earn" for both the training and test data sets. Submit the properties file for the test set.

Q.3 (1 Point)
What component in the software set would need to be replaced in order to run a classifier for periods as end-of-sentence?

Q.4 (3 Points)
Spreadsheet data for text-mining problems differs from spreadsheet data for general data-mining problems. List any 3 differences.

Optional part for students with knowledge of Python

If you know Python, you may find it useful to try and do questions 1 and 2 in Python instead of using TMSK. Note that we won't be teaching Python, so you are a bit on your own although we will provide additional hints via the discussion forum as he sees fit. If you are planning to follow this option please **do read the file** "General Guidance for Using Python in this Course" provided at the General Course Information and Instructions section of this course. This file gives you important information on the required software to use Python in this course.

We have provided a sample python module 'tm' which contains some basic building blocks useful for this assignment. You can freely import 'tm' module to write your answers.