ASSIGNMENT 4

Text Classification

Text classification is one of the fundamental tasks in data mining and machine learning. In this assignment, you will get an opportunity to train classifiers to recognize the topics represented by documents.

Dataset

You will use the 20 Newsgroups dataset. It contains newsgroup documents relating to 20 different topics. It can be downloaded from:
http://qwone.com/~jason/20Newsgroups/

There are 3 different versions available for download. It is recommended that you use the "bydate" version. It has documents split up into training and testing sets.

You can limit yourself to any 5 topics that are most interesting to you. The topics can be inferred from the names of the directories.

Classifiers

- You should choose at least 5 different classifiers and train them using the "train" part of the dataset.
- As before, you should test the accuracy of the algorithms using the "test" part of the data.
- In the model evaluation class, you learned various measures such as precision, recall, F-score. **You have to compare the classifiers using at least these 3 measures besides accuracy.**

Languages/Tools

You are free to use any language, package or tool that you want. Just be sure to mention this in the README file.

To help you get started, I am attaching a smaller project that identifies emails as spam or ham(not spam). It is in the R language using the package RTextTools. You can use it as a template to get started, if you want. There is also a paper that explains the package in detail.

********** How to run example: **********
1. Open the folder example.
2. Install RTextTools library
3. When running the R file, make sure you set your working directory to where the file is.
This can be done by:
setwd("PATH to where the R file is located")
5. Source the file.
****

You are free to choose any other language or library.

What to turn in

- Code
- README file (it should include the languages/tools that you used and how to compile your code)
- A brief report that should include the comparison table of the classifiers using the model evaluation metrics indicated earlier.

Please do not submit the data files. Doing so will take up excessive space on eLearning.