Cmpe 493 Introduction to Information Retrieval, Spring 2018
Assignment 2 - Text Classification using Naive Bayes, Due: 30/03/2018 (Friday), 23:55

In this assignment you will implement a *Multinomial Naive Bayes* (NB) classifier with two different feature selection criteria for text document classification. In the first part, you will use all the words in the documents as a feature set. In the second part, you will select features via Mutual Information. You will choose the k (where $k = 50$) most discriminating words from each class and use **only** these words as features, not all words in the documents. Note that the features (words) should be selected from the training set.

You will use the Reuters-21578 data set, which you used in your first homework assignment. You should pre-process the SGML files to extract the training and test sets to be used in this assignment. You will build a Multinomial NB text classifier that assign an input test document into one of the following classes, which are the five most common topics in the Reuters-21578 corpus:

- earn
- acq
- money-fx
- grain
- crude

Therefore, your training and test sets will consist of the news articles that belong to one of the above five topics. Note that you should eliminate the news stories that belong to more than one of these topics. However, you should include the news stories that belong to only one of the above five topics, even if they belong to more than one topic. For example, the two documents, whose topic contents in SGML format are given below, are valid documents that you should include in your training and test sets:

- <TOPICS><D>money-fx</D></TOPICS>
- <TOPICS><D>money-fx</D><D>interest</D></TOPICS>

Although the second document belongs to two topics, only the first topic is included in our list of five topics. On the other hand, the following document should not be included in your training and test sets, since it belongs to two of the five topics that we are interested in.

- <TOPICS><D>money-fx</D><D>earn</D></TOPICS>

The training set will include the documents that are marked up by LEWISSPLIT="TRAIN":

- <REUTERS ... LEWISSPLIT="TRAIN" ...>

Your test set will include the documents that are marked up by LEWISSPLIT="TEST":

- <REUTERS ... LEWISSPLIT="TEST" ...>

You can use the preprocessing steps from your first assignment. You **must** remove stopwords. You can apply case-folding and stemming, though you are not required to do so. The text of a news story is enclosed under the <TEXT> tag. You should use the <TITLE> and the <BODY> fields to extract the text of a news story.

After creating the training and test sets, you should learn the parameters of your Multinomial Naive Bayes model using the training set with (i) using all words in the lexicon and (ii) with using the k most discriminating words based on Mutual Information.

Then, you should test your classifiers by using the test set and you should report the results of the two different classifiers.

Note that you are NOT allowed to use any external libraries in this homework.

Submission: You should submit a *".zip"* file named as YourNameSurname.zip containing the following files using the Moodle system:

1. Report:
   (a) Report the number of documents in each class in the training and test sets.
   (b) Report the k most discriminating words (where $k = 50$) for each class based on Mutual Information.
   (c) Report the *macro-averaged* and *micro-averaged* precision, recall, and F-measure values obtained by your two classifiers on the test set, as well as the performance values obtained for *each class separately* by using *Laplace smoothing* with $\alpha = 1$.
   (d) Include screenshots showing sample runs of your two programs.
2. Commented source code and readme: You may use any programming language of your choice. However, we need to be able to test your code. Submit a readme file containing the instructions for how to run your code.

Contact: For questions/comments you can contact Abdullatif Köksal (abdullatifkoksal@gmail.com).

Late Submission: You are allowed a total of 5 late days on homeworks with no late penalties applied. You can use these 5 days as you wish. For example, you can submit the first homework 2 days late, then the second homework 3 days late. In that case you will have to submit the remaining homeworks on time. After using these 5 extra days, 10 points will be deducted for each late day.