Cmpe 493 Introduction to Information Retrieval, Spring 2020
Assignment 2 - Movie Review Sentiment Classification, Due: 04/05/2020 (Monday), 23:59

In this assignment you will implement three Naive Bayes (NB) classifiers, Multinomial NB, Bernoulli NB, and Binary NB, for identifying the sentiment (positive/negative) of a given movie review. You will use the Cornell Movie Review Data Set (polarity dataset v2.0)(https://www.cs.cornell.edu/people/pa review-data/)[1] to train and test your classifiers.

The training and test sets are provided in the *data.zip file*. The positive reviews are in the *pos* folder and the negative reviews are in the *neg* folder. The training set contains 700 positive and 700 negative movie reviews. The test set contains 300 positive and 300 negative movie reviews. Each review is provided as a separate file. The tokens have already been lower-cased.

Preprocess the files by extracting the individual words (tokens) from them. Learn the parameters of your models using the training set, and test your classifiers by using the provided test set.

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 *macro-averaged* and *micro-averaged* precision, recall, and F-measure values obtained by your classifiers on the test set, as well as the performance values obtained for *each class separately* by using *Laplace smoothing* with $\alpha = 1$.
   (b) Compare and discuss the performance of each NB model for this task. Perform randomization tests to measure the significance of the differences between the micro-averaged F-scores of the algorithms.
   (c) Include a screenshot showing a sample run of your program.
2. Commented source code and readme: You may use any programming language of your choice. However, I need to be able to test your code. Submit a readme file containing the instructions for how to run your code.

Late Submission: You are allowed 7 late days (one week) for this assignment with no late penalty. However, I suggest you to turn in the assignment in two weeks, since you will have to work on the other assignments afterwards. After 7 days, 10 points will be deducted for each late day (if you have difficulty in completing the assignment due to any reason, please contact me).

---

[1] Bo Pang and Lillian Lee, A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts, Proceedings of ACL 2004.