Cmpe 493 Introduction to Information Retrieval, Spring 2017
Assignment 2 - Movie Review Sentiment Classification, Due: 04/05/2017 (Thursday), 17:00

In this assignment you will implement a Multinomial Naive Bayes classifier 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/pabo/movie-review-data/)[1] to train and test your system.

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 from them. Assume that a word consists of letters from the English alphabet. Discard all tokens that contain different characters such as digits, punctuation marks, or other special symbols (e.g. $). Learn the parameters of your model using the training set, and test your classifier 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 system on the test set, as well as the performance values obtained for *each class separately without using smoothing*
   (b) Report the *macro-averaged* and *micro-averaged* precision, recall, and F-measure values obtained by your system on the test set, as well as the performance values obtained for *each class separately* by using *Laplace smoothing* with $\alpha = 1$.
   (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 a total of 3 late days (including weekends) on homeworks with no late penalties applied. You can use these 3 days as you wish. For example, you can submit the first homework 2 days late, then the second homework 1 day late. In that case you will have to submit the remaining homeworks on time. After using these 3 extra days, 10 points will be deducted for each late day.

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