Cmpe 493 Introduction to Information Retrieval, Fall 2020
Assignment 4 - Spam Email Filtering, Due: 05/02/2021, 23:59

In this assignment you will implement a spam/non-spam filter using the Multinomial Naive Bayes (NB) Algorithm. You will use a subset of the Ling-Spam corpus[1] to train and test your system. The provided training and the test sets (included in the *dataset.zip file*) each contain 240 spam and 240 legitimate email messages. Each email message is provided as a separate file. All files start with a "subject:" heading. Stopword removal and lemmatization have already been performed.

Preprocess the files by extracting the individual tokens from them. Develop two versions of the NB algorithm. (i) In the first version, you should use all the words in the documents in the training set as features.

(ii) In the second version, you should select features via Mutual Information. You should choose the k (where k = 100) most discriminating words from each class and use only these words as features. Note that the features (words) should be selected from the training 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) What is the size of your vocabulary when you use all words as features?
   (b) Report the k most discriminating words (where k = 100) for each class based on Mutual Information.
   (c) Report the macro-averaged precision, recall, and F-measure values obtained by the **two versions** of your classifier on the test set, as well as the performance values obtained for each class separately by using Laplace smoothing with $\alpha = 1$.
   (d) Perform randomization test to measure the significance of the difference between the macro-averaged F-scores of the two versions of your classifier (i.e., without feature selection and with feature selection).
   (e) Include a screenshot showing a sample run of your program.
2. Commented source code, executable, 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.

Honor Code:

You should work individually on this assignment and all the source code should be written by you. You are NOT allowed to use any available libraries or any code written by other people. Violation

[1] I. Androutsopoulos, J. Koutsias, K.V. Chandrinos, George Paliouras, and C.D. Spyropoulos, "An Evaluation of Naive Bayesian Anti-Spam Filtering". In Potamias, G., Moustakis, V. and van Someren, M. (Eds.), Proceedings of the Workshop on Machine Learning in the New Information Age, 11th European Conference on Machine Learning (ECML 2000), Barcelona, Spain, pp. 9-17, 2000.

of the Honor Code will be strictly penalised, not only by a zero grade from the homework, but also by filing a petition to the Disciplinary Committee.

Late Submission:

You are allowed 7 late days (one week) for this assignment with no late penalty. After 7 days, 10 points will be deducted for each late day.