SI 650/EECS 549 - Information Retrieval
Assignment 3

Due Thursday, April 6th, 23:59 EST.
Please submit as pdf attachment via Canvas.
General discussion encouraged, but everyone should come up with the solution independently.
If you received help from anyone, you should list the name(s) on the top of your submission.

This is an exercise of text classification, through the platform of an online data science competition: `https://inclass.kaggle.com/c/umich-si650-Identify-Personal-Attacks`.

Every document (i.e., a line in the data file) is a discussion comment from English Wikipedia. Your goal is to classify the topic of each comment into ONE of the two categories, based on whether it contains a personal attack.

0. the comment does NOT contain a personal attack

1. the comment DOES contain a personal attack

The training data contains 69,518 comments, already labeled with one of the above categories. The test data contains 46,346 comments that are unlabeled. The submission should be a .csv (comma separated free text) file with a header line "ID,Category" followed by exactly 46,346 lines. In each line, there should be exactly two integers, separated by a comma. The first integer is the line ID of a test question (0 - 46,345), and the second integer is the category your classifier predicts one of $\{0, 1\}$.

You can make 10 submissions per day. Once you submit your results, you will get an accuracy score computed based on 50% of the test data. This score will position you somewhere on the leaderboard. Once the competition ends, you will see the final accuracy computed based on 100% of the test data. The evaluation metric is the accuracy of your classifier - so the higher the better.

You can use any classifiers, any combination of features, and either supervised or semi-supervised methods. You can choose to use feature selection, or not. You can also be creative and make use of external data sources that do not contain the exact text comments in the data. More description and competition rules can be found on the competition web page.

Grading: We will evaluate your classification result using F1 score. You will receive at least 70 points if the F1 score of your best classifier beats a correctly implemented Naive Bayes classifier. The other 30 points will be given according to your position on the leader board.

The formula to compute your grade:

$$grade = 70 + 30 * 2/log_2(2 + rank)$$

(**Yes! The winner gets 108 points!**)

To secure a grade higher than 90, you have to beat the SVM benchmark.

Have fun! And don't waste your quota of submissions!

What to hand in: a one page memo describing the algorithms/features/tools you explored and the corresponding results. Please write down your name and the display name you used in the competition. We may request source code of some of your submissions.