INLS 690-270 Data Mining: Methods and Applications
Assignment 2

Please submit as a zip file attachment via Canvas.

This is a text classification task through the online data competition platform Kaggle. You may use the following invitation link to join the competition but please refrain from posting the invitation link elsewhere: https://www.kaggle.com/t/52aa6b92acd943bb9a4f27329c0c08a5.

In this task, every document (a line in the data file) is a product review. Your goal is to classify each review into ONE of the two categories:

- 1: the review expresses positive sentiment.
- 0: the review expresses negative sentiment.

The training data contains 29,996 examples, already labeled with one of the above categories (1 or 0). The test data contains 6,000 examples that are unlabeled. The submission should be a .csv (comma-separated free text) file with a beginning header line `id, label` followed by exactly 6,000 lines. In each line, there should be exactly two integers, separated by a comma. The first integer is the line ID of a test example (`1, 2, ..., 6000`), and the second integer is the category your classifier predicts one of {`0,1`}. Refer to the sample submission file for more details.

You can make 10 submissions per day (reset at midnight UTC, i.e. 7 pm EST). 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 the other 50% of the test data.

The evaluation metric is the *accuracy* of your classifier:

$$\text{accuracy} = \frac{\text{\# of correctly classified test examples}}{\text{total \# of test examples}}$$

so the higher the better.

You can use any classifier or combination of classifiers, any combination or selection of features, and either supervised, semi-supervised, or even transfer learning approaches. You can be creative and make use of external data sources. **please do not attempt to link back the original dataset** in Kaggle, such as https://www.kaggle.com/kritanjalijain/amazon-reviews or https://www.kaggle.com/bittlingmayer/amazonreviews. Instead, you can make use of a much larger product review dataset (almost 4 million reviews) that can be downloaded at https://drive.google.com/drive/folders/1ZDa2qHMPxQSXxg-Bn-LnoXwi41ZoZI-r.

Please try to implement a classification method that can outperform the baseline method: a logistic regression classifier using unigram count features derived only from the review text with regularization hyperparameter $C = 1$; its accuracy ≈ 0.88. Its implementation is released in Sakai → Resources → Assignments → Homework Assignment 3. The formula to compute your grade (rounded to the closest integer):

$$grade = 80 + 20 * 2/\log_2(2 + rank)$$

Yes! The winner will get 105 points (105% grade of this homework)! The 25^{th} position will get 88 points.

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. In addition, please attach source code to implement the algorithm in your submissions.