**Assignment 2: Sentiment Analysis**

In this assignment, you will practice sentiment analysis with textual data.

You are provided with a dataset “MovieReview-Sample.csv” which contains 2,000 movie review text, and a labeled sentiment. Label “0” is Negative and label “1” is Positive.

**Question 1**: Performance Comparisons

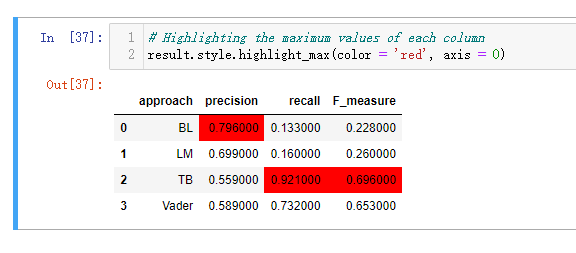
You are asked to use three approaches taught in Lab 2 to perform sentiment analysis on the dataset: 1) Using Bing Liu’s Lexicon; 2) Using LM dictionary; 2) Using TextBlob; and 3) Using Vader (either from NLTK or from Vader directly).

Please report the following:

1. Report Precision, Recall and F measure achieved by each tool. Notice that you will calculate them by comparing your prediction and the gold standard (label 0 and 1). Please present the result in a comparison table and highlight the highest performance.

(Hint, you should report precision not accuracy. This means you need to calculate positive precision, negative precision and then average precision)

1. Provide your analysis of the performances. If you are in charge of identifying the appropriate software to perform sentiment analysis for movie reviews, which one will you choose? Give 1-2 reasons.



I will choose TextBlob method to perform sentiment analysis for movie reviews,.

Firstly, for movie reviews, the consequence of wrongly predict a good review or bad review is not that significant when it compare to wrongly predict a cancer patient. Which means the high and low values about precision and recall is not that important, hence we should focus on F measure (combine precision and recall). Hence, we should choose TextBlob method to perform sentiment analysis for movie reviews.

Secondly, TextBlob method have two highest performances out of three, mathematically we should choose TextBlob method to perform sentiment analysis for movie reviews.

**Question 2**: Ensemble

You are going to using ensemble method to improve the performance of individual tool. Can you think of a way to ensemble the three methods/tools to improve the performance?

(Hint 1: you may choose the 3 best performing algorithms to ensemble. There is no need to include inferior algorithms from the previous step. Hint 2: the simplest form or ensemble is a majority vote, or a weighted majority vote based on the algorithm performances). Report your performance improvement (in percentage) over any single models.

Bonus: I also provide the original full dataset “Movie\_review\_Polarity\_CSV.zip”. Please notice that this file contains pos.csv and neg.csv. You may run your algorithm on the full dataset and see if the performance hold from the sample dataset.

Submission:

1. Word Report
2. Python program. Please make sure your python program can run successfully.

Other instructions:

1. DO NOT submit your dataset. Only submit Word and python program.
2. Do not use absolute path to read your input data (it won’t run on your TA’s computer)
3. Name all your files FirstName\_LastName.xxx. This will make our grading easier.
4. Do not zip your file. Submit two files directly.

Thank you!