Project 3 – IMDB MOVIE REVIEW
Due on Sat 9th Apr 11:59pm EST

Context: IMDB dataset having 25K movie reviews for natural language processing or Text analytics.
This is a dataset for binary sentiment classification containing substantially more data than previous benchmark datasets. We provide a set of 12,500 highly polar movie reviews for training and 12,500 for testing. Please use less data eg 6K reviews if you are facing memory issues but make sure to use equal number of positive and negative sentiment reviews. Mention clearly in the notebook, if you have used a reduced dataset.
For more dataset information, please go through the following link,
http://ai.stanford.edu/~amaas/data/sentiment/

Dataset Source: Click here

Task: Goal of this project is to predict the number of positive and negative reviews using classification

Implementation:

- Preprocess Text Data(Remove punctuation, Perform Tokenization, Remove stopwords and Lemmatize/Stem)
- Perform TFIDF Vectorization
- Exploring parameter settings using GridSearchCV on Random Forest & Gradient Boosting Classifier. Use Xgboost instead of Gradient Boosting if it's taking a very long time in GridSearchCV
- Perform Final evaluation of models on the best parameter settings using the evaluation metrics
- Report the best performing model

Submission Instructions: Please just submit one jupyter notebook containing all the code and make use of markdown cells to include the comments, answers, reasoning, analysis, etc.

Note: Name of your file should be your "Project3-id_Firstname_Lastname.ipynb"