Homework 3 - Due March 10th, 2021 at 11:59 Eastern Standard Time

How to do this assignment: Using Google Colab (https://colab.research.google.com/) please answer the following questions in a new notebook. You are to write the question on a Text field, and your programming answer on a Code field. Please write your name and student ID number on a Text field at the very beginning of this notebook. Note that when you are testing your code, the output needs to be shown

What to submit: Via iCollege you are to download a copy of the executed Colab notebook in PDF format, and submit this file ALONGSIDE a link (in the comments/description section) to your own GitHub repository where this Colab notebook is stored. In other words, you are expected to keep your own personal GitHub repository for this class and place ALL your assignments and project code in that location.

Questions (150 total points): ***NOTE: You can only use Python's built-in regular functions, scikit-learn as your ML library for these exercises, NLTK as your NLP library. Any code using any other ML or NLP libraries will not be graded. Note: You can use pandas and/or numpy as these are NOT machine learning libraries.***

Using the Cornell Movie Review data (http://www.cs.cornell.edu/people/pabo/movie-review-data/) use the polarity_dataset V2.0 and write the following pieces of code:

1. Using NLTK tokenize all documents, separated by polarity, remove stop words, and list the top 20 most frequent tokens (and their counts) for the positive reviews, and the top 20 most frequent tokens (and their counts). What kind of things do you notice are different between the two sets? (30 points)

2. Using the code from previous lectures, build 3 polarity classifiers using the following parameters (20 points). Note: just train the models.

a) **For training: use 50% of the positive dataset and 70% of the negative dataset. For your model use: NaiveBayes with the TF-IDF vectorizer.**
b) **For training: use 50% of the negative dataset and 70% of the positive dataset. For your model use: NaiveBayes with the TF-IDF vectorizer.**
c) **For training: use 25% of the negative dataset and 25% of the positive dataset. For your model use: SVM with the TF-IDF vectorizer.**

3. Using the models from question 2, evaluate them on their individual rest of the dataset. This is, for a) 50% positive and 30% negative, for b) 50% negative and 30% positive, and for c) 75% negative and 75% positive. Calculate and show ONLY the following metrics for each model: Accuracy, Precision, Recall, Macro F1-score. (15 points).

4) Using the model performance metrics from question 3, answer the following questions. Please provide logical and intuitive rationale for your answers, simple answers like: because it has the best score, will not be sufficient. (40 points):

- **a) What is the best performing model?**
- **b) Why do you think this is the best performing model?**
- **c) How does class imbalance play in determining polarity?**
- **d) Do you think either more data or a better model is a better approach for this kind of task?**

5) Using NLTK and VADER, calculate the sentiment score for all documents in the positive polarity. Calculate the polarity threshold needed (and reasonable) to have the majority of the document labels match. Do the same for the negative class. Provide the threshold needed, the reason why you think this threshold is reasonable, and the accuracy percentage (how many documents are correctly labeled using this threshold). (45 points):

Bonus (40 points): Repeat questions 2,3 and 4 removing all stopwords. Answer the following questions: Did this change the results in any way? Why do you think so?