Homework 4

This week we are performing sentiment analysis on Rotten Tomatoes movie reviews (published since 2023) using a sentiment lexicon (SentiWordNet and VADER), and different classification models. I am not providing code to read in the CSV file. **You will have to read the file and select the movie reviews as the feature data and the associated sentiment as the labels.** There will be two different approaches in this homework:

* Unsupervised learning, where we pretend we do not have any labels and apply a sentiment lexicon to auto-label the data (but we will compare the lexicon output to the actual labels)
* Supervised learning using different machine learning models

As in the last homework, we will be applying different stemming/lemmatization and vectorization techniques, and calculating accuracy and F1-scores.  
NOTE: the sentiment labels in this data are “POSITIVE” and “NEGATIVE”, meaning there is no “NEUTRAL” category, as a result, we will not be classifying anything as neutral either.

# Part 1 – Unsupervised learning

Write code that accomplishes the following for ALL reviews in the dataset (i.e., NO train-test split):

1. Apply VADER polarity analysis on each review
2. Apply SentiWordNet analysis on each review
   1. Apply POS-tagging and label each word in the data using the sentiwordnet lexicon
   2. Assign an overall sentiment based on the word labels
3. Decide whether to classify ‘0’ scores as positive or negative
4. Calculate the accuracy and f1-scores by comparing the two lexicon labels to the original labels

# Part 2 – Supervised learning

Write code that accomplishes the following (feel free to create a new python file):

1. Apply a stemming technique of your choice, as well as the wordnet lemmatizer (= two different techniques)
2. Apply different vectorization techniques
   1. Filter out duplicate words and send the resulting document through sklearn’s CountVectorizer
   2. Tf-idf vectorizer
3. Train three different machine learning models using 10-fold CV, must include Naïve Bayes, and choose two more from the following options:
   1. Logistic Regression
   2. Decision Tree
   3. Support Vector Machine
   4. Random Forest

Be careful which ones you choose, you will be expected to think critically about the performance of these models.

1. Evaluate the models using accuracy and F1-score (micro and macro).

# Part 3 – Questions

* Unsupervised learning
  1. How did you classify a score of 0?
  2. Which of the lexicons performed better?
* Supervised learning
  1. Which models did you choose and why?
  2. Look at the final metric scores and provide some insights into your results based on what we’ve learned in class.
* General
  1. Can we compare our unsupervised results to our supervised results, why or why not?
  2. Did you notice anything about the accuracy score and micro F1 score? Why do you think this is the case?
  3. Did you notice a significant difference in runtime between the supervised and unsupervised approaches? If so, which one was slower, and why do you think that is?