DATA2001 Assignment 2 (Weight: 25%)

The aim of this assignment is to gain practical experience in analysing unstructured data. You must complete this in Python using a Jupyter notebook. You will need to submit a single Jupyter notebook (.ipynb file) via Blackboard.

Dataset:

The dataset for this assignment consists of reviews of fine foods from amazon. It includes product, user information, ratings, and plaintext review. Reviews are from period Aug 31st - Oct 26th, 2012.

Example review is shown below,

Data format

```
product/productId: B001E4KFG0
review/userId: A3SGXH7AUHU8GW
review/profileName: delmartian
review/helpfulness: 1/1
review/score: 5.0
review/time: 1303862400
review/summary: Good Quality Dog Food
review/text: I have bought several of the Vitality canned dog food products and have
found them all to be of good quality. The product looks more like a stew than a
processed meat and it smells better. My Labrador is finicky and she appreciates this
product better than most.
```

This dataset has 28054 observations and 10 columns. The description of columns is given below,

Id – review id

ProductId – id of the product

UserId – id of the user

ProfileName – name of the user

HelpfulnessNumerator – fraction of users who found the review helpful (numerator part)

HelpfulnessDenominator – fraction of users who found the review helpful (denominator part)

Score – rating of the product

Time – time of the review (in unix time)

Summary – review summary

Text – text of the review

Original source for the data, can be accessed here: SNAP: Web data: Amazon Fine Foods reviews (stanford.edu)

The submitted notebook should address 7 tasks (see marking grid for mark allocation):

1. Data Preparation: Read the dataset using the “pandas” library. Columns ‘Id’, ‘Score’ and ‘Text’ are the only columns that should be used, and other columns must be ignored. To perform sentiment analysis, annotate the review dataset using already provided score column. Create new column ‘Label’. Assign ‘pos’ – positive for the product ratings 4 and 5. Assign ‘neg’ – negative for the product ratings 1 and 2. Product rating with score 3 should be ignored and dropped from the dataset. After annotating, produce the summary of the dataset. How many positive and negative reviews are present in the data. Print your summary.

2. Data Cleaning: Write the necessary scripts to clean the text in the review dataset and explain the steps along with the justification in less than 4 lines.

3. Build a logistic regression text classifier to categorise whether review has positive or negative sentiment. 70% of the reviews should be used for training and the remaining 30% for the testing. List the steps taken in your own words to build the model in less than 4 lines.

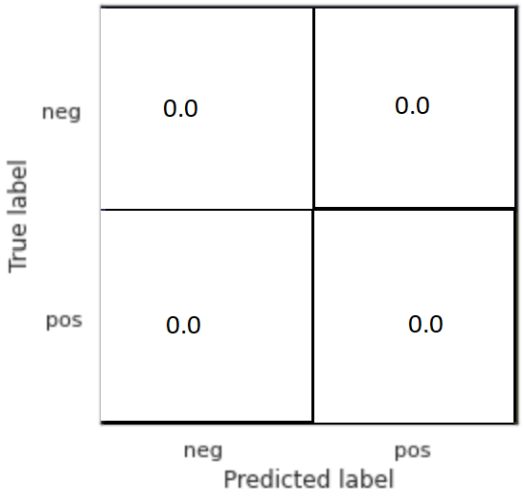
4. Evaluate the model built in the previous step and compare it with a baseline model that assigns positive label to all test samples. A) Report accuracy and only for the negative class report precision, recall, f1 score; and B) show a confusion matrix for both the baseline and the logistic regression models. Explain the cause of difference between the accuracy and the f1 score of the negative class for the baseline model.

   Result should be written in below format:

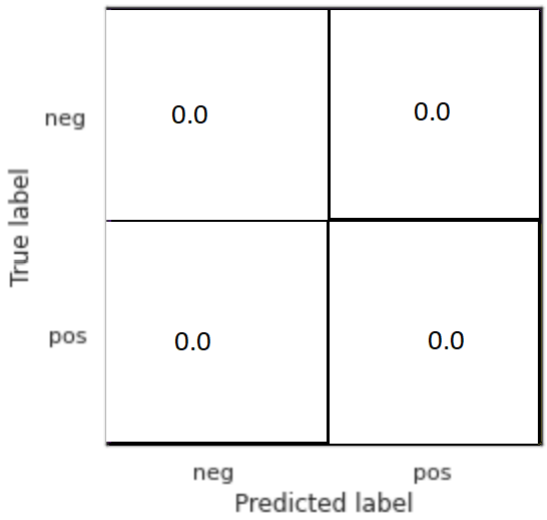
   A) Baseline: {‘accuracy’:0.0, ‘precision’:0.0, ’recall’:0.0, ‘f1-score’:0.0}

      Logisticregression: {‘accuracy’:0.0, ‘precision’:0.0, ’recall’:0.0, ‘f1-score’:0.0}

   B) Confusion matrix for baseline model:



Confusion matrix for logisticregression model:



5. Use the better performing model (hopefully your logistic regression model) to predict the sentiment for the reviews provided in the predict dataset (download 'predictdata.csv' from the blackboard). Report the results in a dataframe format with columns 'Id', 'Text' and 'Model Prediction'.

   Model Prediction Results

Id	Text	Model Prediction
1		
2		
3		
4		
5		

6. What are the most frequent words in the review dataset (use a world cloud to show this, remove stop words) and show which words play a significant role in classify whether the review is positive or negative.

7. Write in three lines with **your own words** about classification with unbalanced data. What is the issue with unbalanced data and how you could handle it in a better way (you don't need to implement your solution)? Provide the issue with imbalanced classification and your suggestion in less than 4 sentences that reader can understand clearly using research from the internet resources.

Note: Jupyter notebook should be commented properly and written in a way easier to understand for the reader. For marking purpose, code will be rerun to verify the results.