Team Members:

1.Layla Reda Hassan

2.Maria Emad Ayad

3.Marina Samouil Hassab Allah

4.Mina Eid Ageeb

5.Abraam Atef Fekry

**Sentiment Analysis of movie reviews**

[Team ID: 21]

Sentiment Analysis of movie reviews

Introduction:

/\* Discuss in brief the project main idea and the objectives\*/

We have a dataset contains polarity of movie reviews (1000 positive reviews and 1000 negative reviews).

Our model classify the polarity of these reviews after applying preprocessing steps and lemmatization on the dataset

Methodology:

/\*Discuss the methodologies used in order to fulfil your objectives (i.e.The feature sets and the models implemented)\*/

Firstly, we read the text data in CSV file (.txt -> .csv)

Secondly, we apply preprocessing steps:

* We made all text lowercase
* We removed all stop words
* We removed punctuation {“ ”, / ,[,],….} by tokenization
* We lemmatize the text to get the source words

Thirdly, we shuffled the data and split it to train and validate, and then TF-IDF Vectorizer generate feature from them and pass it to the model.

Fourthly, we used *LogisticRegression()* model as it has higher accuracy compared to naïve Bayes .

Naïve Bayes: its accuracy rang [0.71:0.81]

Logistic Regression: its accuracy rang [0.80:0.86]

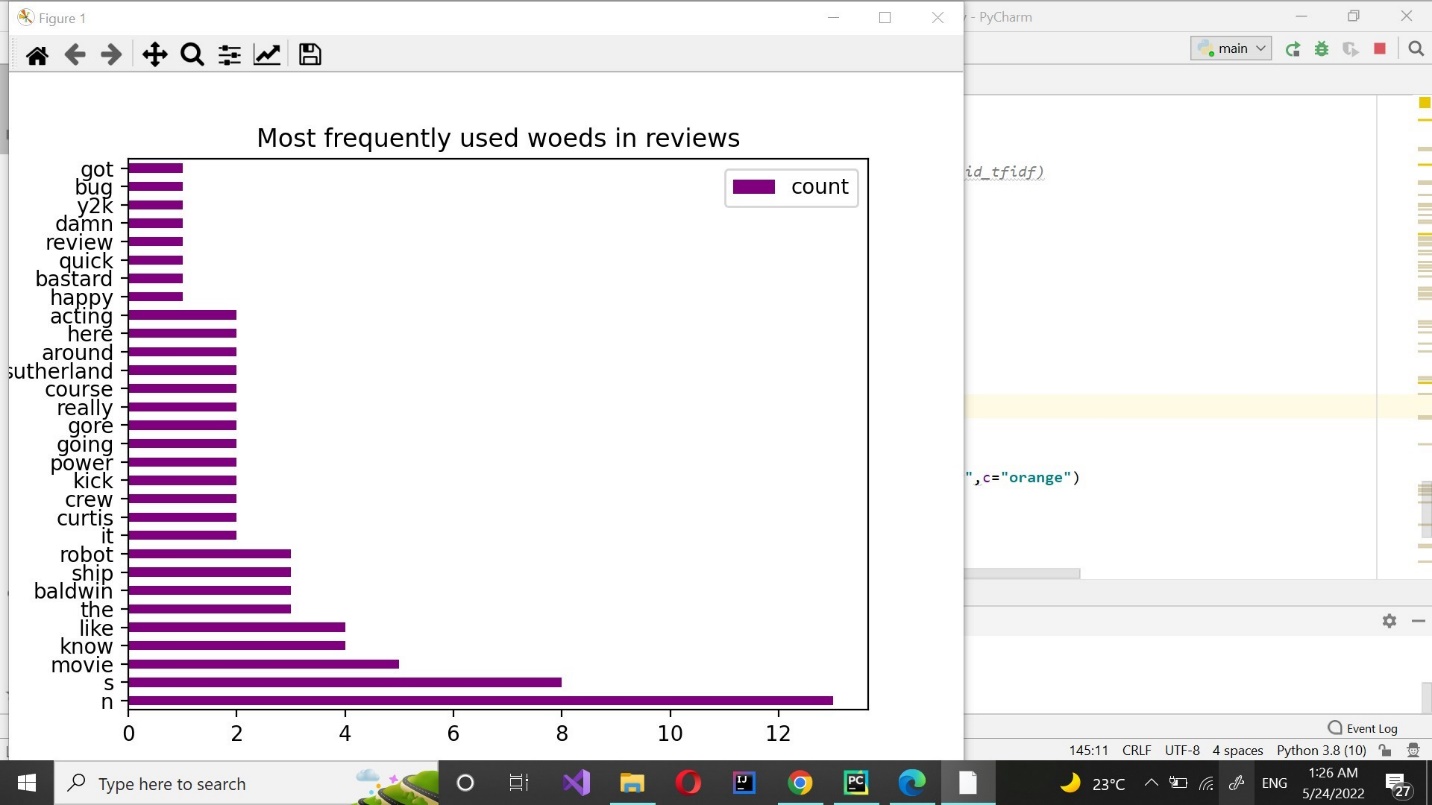
Data Set Summary:

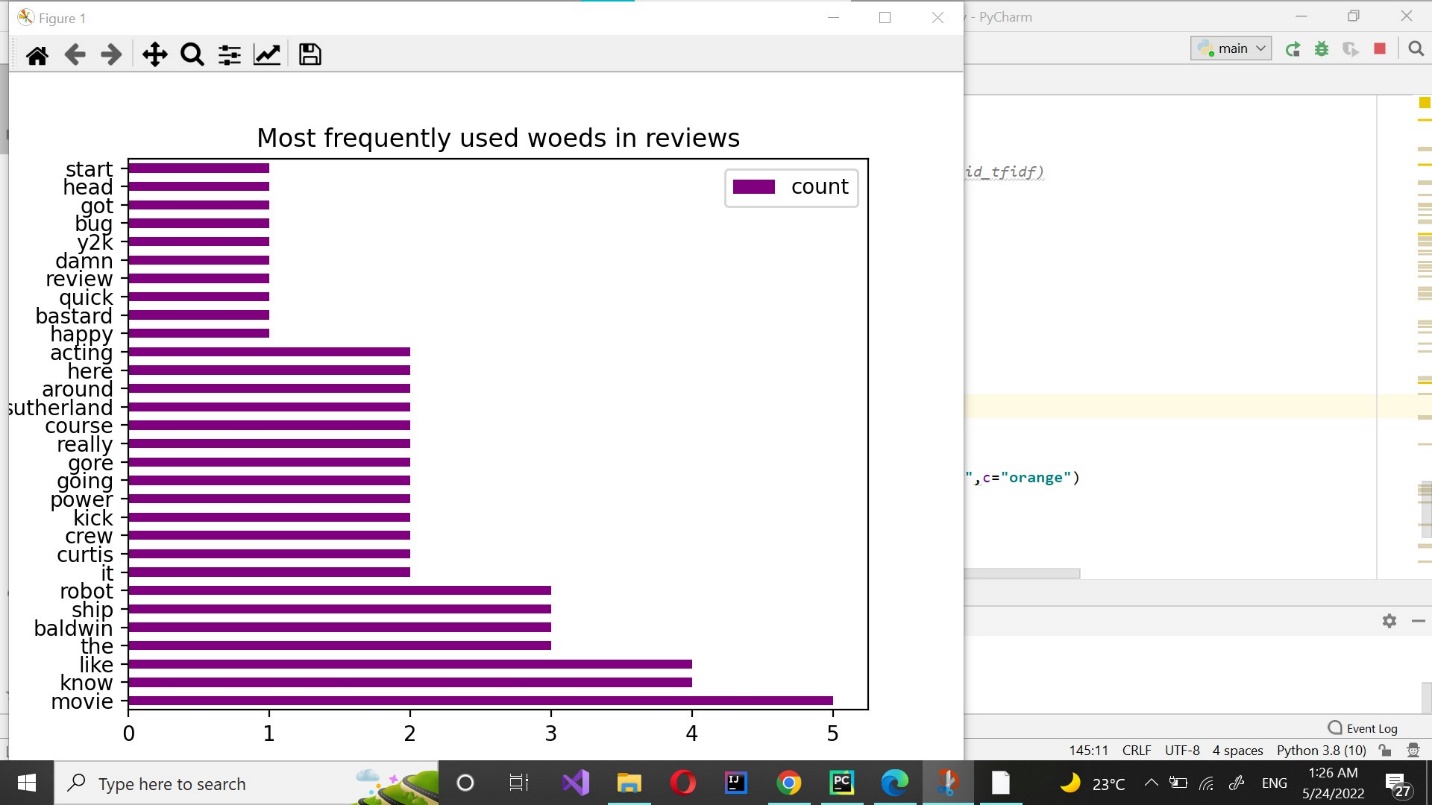
/\*Answer the following Questions:

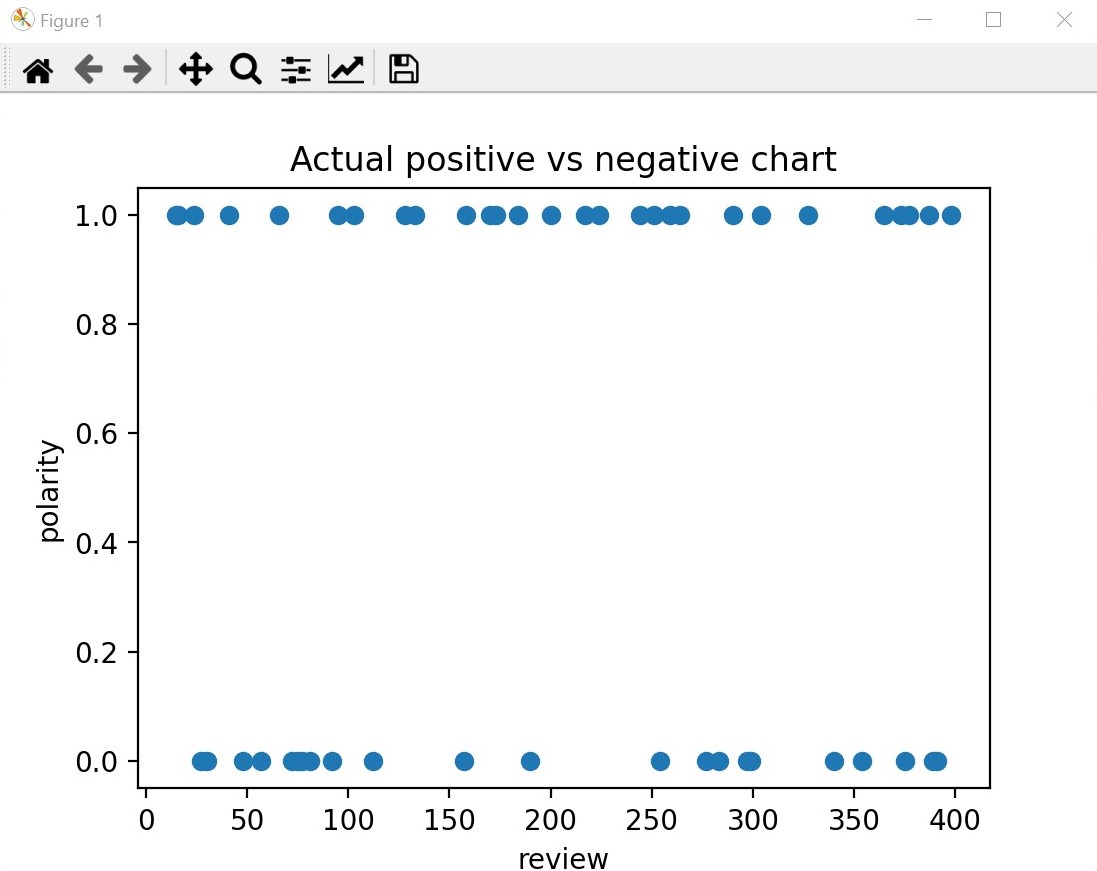
1-What is the data set used?

2- What is the summary of the dataset columns?

3- Visualize the dataset statistics\*/

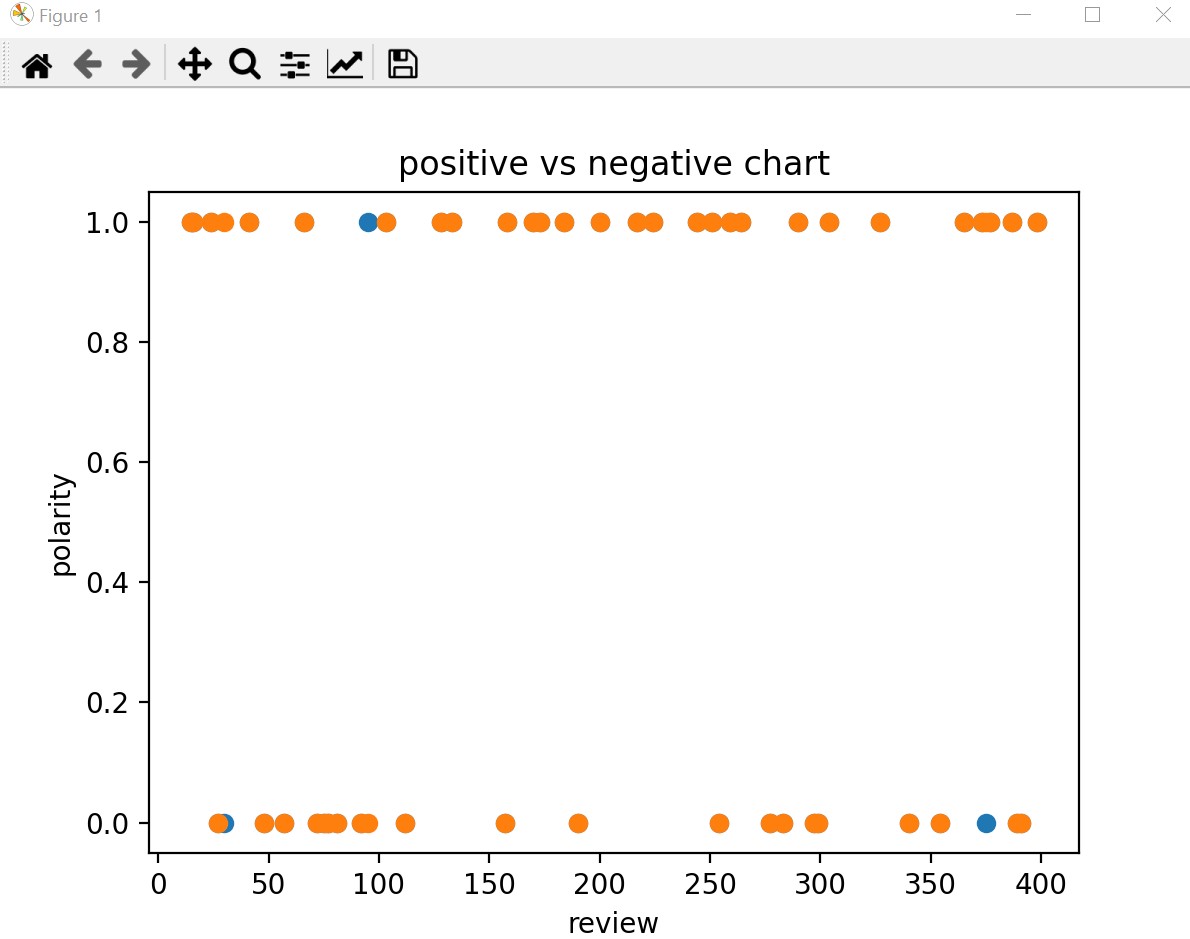
1. Our dataset contains 1000 negative reviews and 1000 positive reviews
2. After creating CSV file contains the reviews, its columns [polarity , reviews]
3. There is a graph showing the 30 most common repeated words of the first review as example

then we noticed that ‘n’ and ‘s’ are repeated many times although these aren’t important words, so we removed them to improve the classification accuracy 

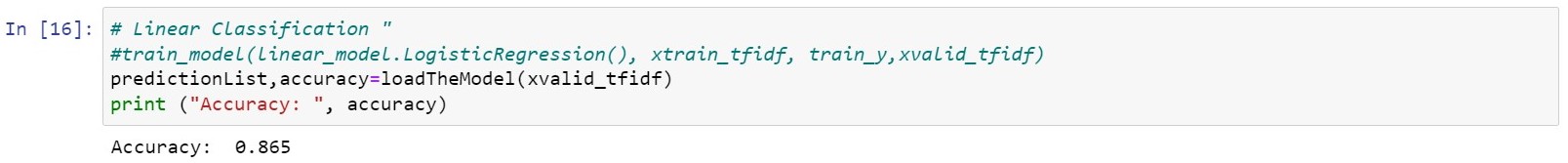
And there is another sample graph of 50 rows of positive and negative reviews

Results:

/\*Use suitable graphs to visualize your models results\*/

This is a graph to show the actual and predicted polarity of test texts

As the accuracy is high there is small wrong predictions

This is a screenshot of the run of our model to show the accuracy.