Topic: Text Mining (NLP)

Instructions:
Please share your answers filled in-line in the word document. Submit code separately wherever applicable.

Please ensure you update all the details:
Name: Kavali Rakesh
Batch ID: DSWDMCON 18012022
Topic: Text Mining and NLP

Grading Guidelines:
1. An assignment submission is considered complete only when correct and executable code(s) are submitted along with the documentation explaining the method and results. Failing to submit either of those will be considered an invalid submission and will not be considered for evaluation.
2. Assignments submitted after the deadline will affect your grades.

Grading Criteria:

Submission	Date			Submission	Date
Correct	On time	A	100		
80% & above	On time	B	85	Correct	Late
50% & above	On time	C	75	80% & above	Late
50% & below	On time	D	65	50% & above	Late
		E	55	50% & below	
Copied/No Submission		F	45		

- **Grade A: (>= 90):** When all assignments are submitted on or before the given deadline.
- **Grade B: (>= 80 and < 90):**
  - When assignments are submitted on time but less than 80% of problems are completed.

    (OR)
  - All assignments are submitted after the deadline.

- **Grade C: (>= 70 and < 80):**
  - When assignments are submitted on time but less than 50% of the problems are completed.

    (OR)
  - Less than 80% of problems in the assignments are submitted after the deadline.

- **Grade D: (>= 60 and < 70):**
  - Assignments submitted after the deadline and with 50% or less problems.

- **Grade E: (>= 50 and < 60):**
  - Less than 30% of problems in the assignments are submitted after the deadline.

    (OR)
  - Less than 30% of problems in the assignments are submitted before the deadline.

- **Grade F: (< 50):** No submission (or) malpractice.

Hints:

1. Business Problem

1.1. What is the business objective?

1.2. Are there any constraints?

2. Work on each feature of the dataset to create a data dictionary as displayed in the below image:

Name of Feature	Description	Type	Relevance
ID	Customer ID	Quantitative, Nominal	Irrelevant, ID does not provide useful information

2.1 Make a table as shown above and provide information about the features such as its data type and its relevance to the model building. And if not relevant, provide reasons and a description of the feature.

3. Data Pre-processing

3.1 Data Cleaning, Feature Engineering, etc.

3.2 Outlier Treatment

4. Exploratory Data Analysis (EDA):

4.1. Summary

4.2. Univariate analysis

4.3. Bivariate analysis

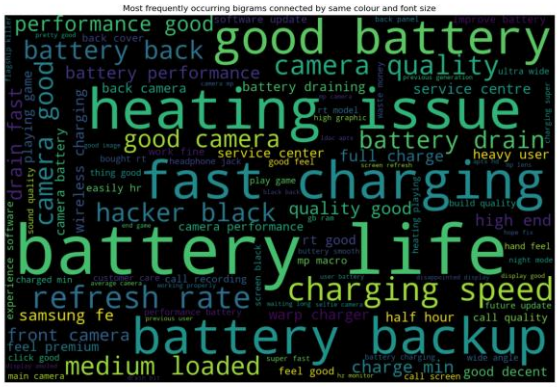
5. Model Building

5.1 Extract text data from websites such as Amazon, Snapdeal, IMDB, Twitter, etc.

5.2 Clean the data and build a word cloud for both positive and negative words. Perform Sentiment Analysis as well.

5.3 Briefly explain the model output in the documentation.

6. Write about the benefits/impact of the solution - in what way does the business (client) benefit from the solution provided?



Problem Statement: -

In the era of widespread internet use, it is necessary for businesses to understand what the consumers think of their products. If they can understand what the consumers like or dislike about their products, they can improve them and thereby increase their profits by keeping their customers happy. For this reason, they analyze the reviews of their products on websites such as Amazon or Snapdeal by using text mining and sentiment analysis techniques.

Task 1:

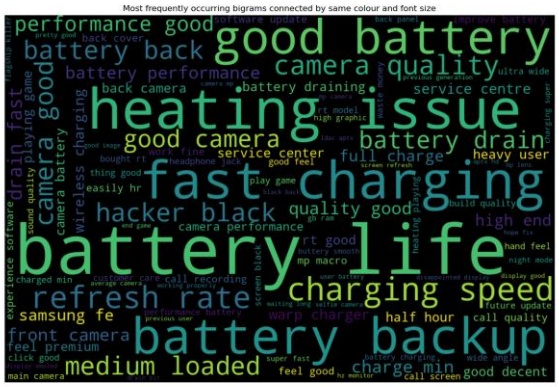
1. Extract reviews of any product from e-commerce website Amazon.
2. Perform sentiment analysis on this extracted data and build a unigram and bigram word cloud.

Corpus :

Positive words :

Negative words :

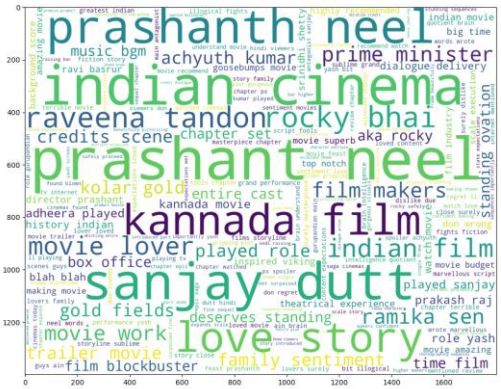
Bigram :



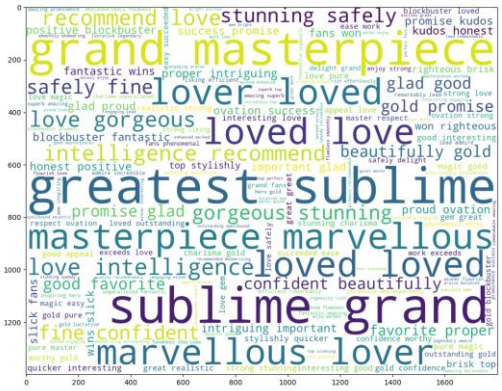
Task 2:

1. Extract reviews for any movie from IMDB and perform sentiment analysis.

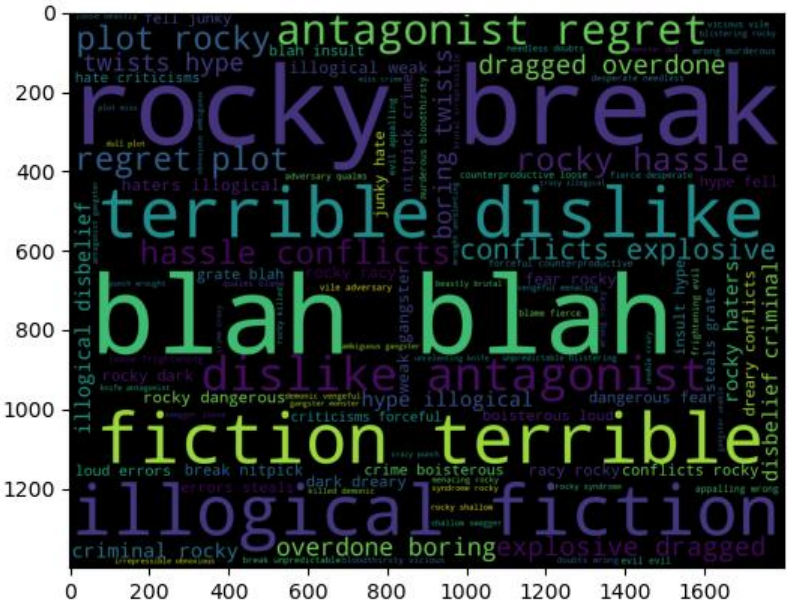
CORPUS :



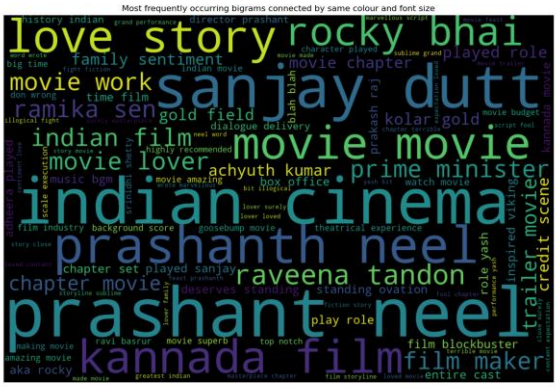
POSITIVE WORDS :



NEGATIVE WORDS:



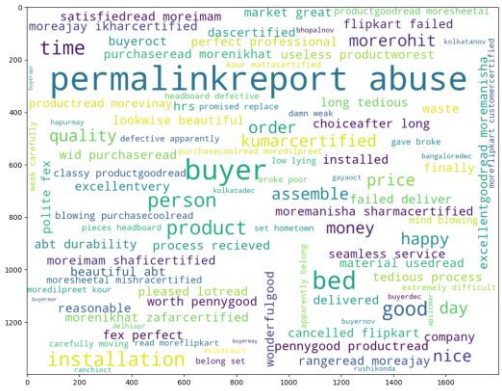
BIGRAM:



Task 3:

1. Choose any other website on the internet and do some research on how to extract text and perform sentiment analysis

CORPUS :



POSITIVE WORDS :

NEGATIVE WORDS: