NLP Assignment 1

Set 3

Link to the Dataset:

https://drive.google.com/file/d/1x0oiWyLUns9002jTDj2CzIE6yqbglN__/view?usp=sharing

Note: Use 50% of dataset from the original dataset given

Description of Data:

This is the Amazon Fine food review dataset to predict the sentiment of food review each record consists of the following attributes:

The column or features in the dataset:

- Id
- ProductId — unique identifier for the product
- UserId — unqiue identifier for the user
- ProfileName
- HelpfulnessNumerator — number of users who found the review helpful
- HelpfulnessDenominator — number of users who indicated whether they found the review helpful or not
- Score — rating between 1 and 5
- Time — timestamp for the review
- Summary — brief summary of the review
- Text — text of the review

1. Download the dataset and Create a dataframe named as **food** then check the head, info, and describe methods on created dataframe **food**. **(2 Marks)**
2. Create another dataframe name called **Review** with Score and Text column. Perform pre-processing steps like Removing Punctuations, Numbers, and Special Characters, Stop Words in dataset. **(2 Marks)**
3. Normalize review by using Stemming or Lemmatization. **(2 Marks)**
4. Preprocessed text review should be included in the Review data frame as cleaned_text. Plot word cloud for the tweets. **(1 Marks)**
5. Create two objects **X** and **y**. **X** will be the 'cleaned_text ' column of Review data frame and **y** will be the ' Score' column. **(6 Marks)**
    a. Create a TF-IDF object and split the data into training and testing sets. Train a Decision tree model and Display the confusion Matrix.
    b. Create a BoW object and split the data into training and testing sets. Train a decision tree model and Display the confusion Matrix.
    c. Compare TF-IDF and BoW. ***Answer without justification will not be awarded marks.***
6. Parse the last 4 rows of 'text' using Viterbi Parser [Use toy_pcfg1 and toy_pcfg2 to get the probabilistic context free grammars; use the PCFG suitable for each sentence] **(3 marks)**
7. Display the HMM POS tagging on the first 4 rows of 'cleaned_text'. **(4 Marks)**