Assignment 1

Data set Link:
https://drive.google.com/file/d/1e90oQiPhb5UcVdE8EjzXc0Fr1pIxEx_S/view?usp=sharing

Problem Statement:
The goal of Part I of the task is to use raw textual data in language models for recommendation based application.
The goal of Part II of task is to implement comprehensive preprocessing steps for a given dataset, enhancing the quality and relevance of the textual information. The preprocessed text is then transformed into a feature-rich representation using a chosen vectorization method for further use in the application to perform similarity analysis.

Part I

Sentence completion using N-gram: (3 Marks)

Recommend the top 3 words to complete the given sentence using N-gram language model. The goal is to demonstrate the relevance of recommended words based on the occurrence of Bigram within the corpus. Use all the instances in the dataset as a training corpus.

Test Sentence: " I like _____ "

Part II

Perform the below sequential tasks on the given dataset.

i) Text Preprocessing: (2 Marks)

a. Tokenization
b. Lowercasing
c. Stop Words Removal
d. Stemming
e. Lemmatization

ii) Feature Extraction: (2 Marks)

Use the pre-processed data from previous step and implement the below vectorization methods to extract features.

Word Embedding using TD-IDF

iii) Similarity Analysis: (3 Marks)

Use the vectorized representation from previous step and implement a method to identify and print the names of top two similar documents that exhibit significant similarity. Justify your choice of similarity metric and feature design. Visualize a subset of vector embedding in 2D semantic space suitable for this use case. **HINT: (Use PCA for Dimensionality reduction)**