Data set Link:

https://drive.google.com/file/d/1fQYedyo5KGYT6g6v5dJKnO9qOuJ8q4M8/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 comparison using N-gram:

Let a search engine powered by language model recommend which of the below sentences are most relevant w.r.t to given training corpus. Design a probabilistic language model to compare below test sentences for recommendation using Trigram. Use all the instances in the dataset as a training corpus.

Test Sentence 1: “The first thing that struck me about Oz was its brutality and unflinching scenes of violence, which set in right from the word GO.”

Test Sentence 2: “This was the most I'd laughed at one of Woody's comedies in years”

Part II

Perform the below sequential tasks on the given dataset.

i) Text Preprocessing:

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

ii) Feature Extraction:

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

Word Embedding using Skip Gram

iii) Similarity Analysis:

Use the vectorized representation from previous step and implement a method to identify and print the names of top two similar words 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)**