**Proposal for Final Project: Find Movie Similarity from Plot Summaries**

**BY Group 14: Dheeraj Rawat (8989521)**

**: Priya Bhardwaj (8940263)**

**Clear Problem Statement (Objective, Motivation, Method)**

**Objective:** The objective of this project is to predict the similarity between movies based on their plot summaries using Natural Language Processing (NLP) techniques and clustering algorithms.

**Motivation:** The ever-increasing content available on streaming platforms requires more efficient and accurate ways to recommend similar movies to users by analyzing aggregated content from sources such as IMDB and Kaggle, we can immediately find patterns and parallels that are not obvious, to improve the quality of movie recommendations.

**Methods:** We will collect movie data collected from IMDB and Wikipedia, pre-process text data using tokenization and stemming, transform text data using TF-IDF Vectorization, and use K Means clustering algorithm to generate similar movies clusters The resulting clusters will be visualized using graph and plot to display interrelationship movies.

**Interesting Choice of Problem**

The trouble of locating movie similarity from plot summaries is exciting because it combines diverse elements of statistics science, which include text processing, machine learning, and data visualization. It has realistic programs within the entertainment industry, particularly in improving advice systems for streaming services like Netflix, thereby enhancing user experience.

**Reasonable (Doable) Choice of Problem**

This problem is reasonable and doable within the given timeframe and assets. There is ample availability of film plot facts from public databases like IMDB and Kaggle. The methodologies involved, consisting of TF-IDF Vectorization and K Means clustering, are nicely set up and can be carried out the use of current Python libraries like NLTK, Scikit-study, and Matplotlib.

**Review of the Data Source**

The primary data sources for this assignment might be IMDB and Kaggle, which provide comprehensive and well-structured plot summaries for a wide variety of films. Additionally, the task might be extended to consist of the Netflix Movie dataset, allowing for a comparison and validation of results across different datasets.

**Indication of Reference Code**

1. **Data Collection**: Scraping or downloading plot summaries from IMDB and Kaggle.
2. **Preprocessing**: Tokenizing and stemming the text data.
3. **TF-IDF Vectorization**: Transforming the text data into a format suitable for clustering.
4. **Clustering**: Applying the K Means algorithm to create clusters of similar movies.
5. **Visualization**: Plotting the Graph to visualize the relationships between movie clusters.