TV Show Search Engine Using IMDB Dataset

1. Introduction

This project presents a TV Show Search Engine designed to retrieve show information based on keyword queries. It leverages the IMDB dataset, which contains metadata about TV shows, including their title, summary, genres, actors, ratings, and more. The goal is to create an efficient and user-friendly system to explore media content.

2. Data Description and Exploration

The dataset includes the following columns:

- **Title**: Name of the show.
- **About**: Brief summary of the show.
- **Episode Duration**: Duration of each episode in minutes.
- **Genres**: Comma-separated list of genres.
- **Actors**: Comma-separated list of main actors.
- **Rating**: User rating out of 10.
- **Votes**: Number of user votes.
- **Years**: Start and end years of the show.

Key observations include:

1. Missing values were handled to ensure data quality.
2. Text columns such as "title" and "about" were preprocessed for consistency.

Visualizations:

- Distribution of shows by genres highlights the most frequent categories.
- Average ratings by genre offer insights into user preferences.

3. Data Cleaning and Normalization

Text columns underwent the following steps:

1. **Text Cleaning**: Removed special characters and converted text to lowercase.
2. **Stop-word Removal**: Removed common words like "the" and "and" that do not carry meaningful information.
3. **Combined Search Text**: Created a consolidated column combining the title, about, genres, and actors for better search relevance.

4. Feature Extraction

TF-IDF was used to convert text into numerical vectors. This method assigns weights to words based on their importance in the dataset. A matrix of 5000 features was generated to represent the textual data.

5. Model Building

The search engine uses **Cosine Similarity** to find relevant TV shows:

1. A query (e.g., "science fiction adventure") is cleaned and vectorized.
2. The vectorized query is compared to the dataset using cosine similarity.
3. Shows with the highest similarity scores are retrieved.

Example Query:
For the query "science fiction adventure," the system retrieves shows like *Stranger Things* and *Doctor Who*, processing results in approximately 0.03 seconds.

6. Results and Evaluation

The system was evaluated on:

1. **Relevance**: The search returned highly relevant shows based on user queries.
2. **Performance**: The average query response time was under a second.

Performance Metrics:

- **Precision and Recall**: Manually evaluated by comparing retrieved results with actual relevance.
- **Query Processing Time**: Measured for efficiency.

7. Conclusion and Future Work

This project successfully implemented a TV Show Search Engine using the IMDB dataset. The system provides accurate and fast results. Future enhancements could include:

1. Advanced ranking algorithms to improve relevance.
2. Incorporating user feedback for personalized recommendations.
3. Adding filters for attributes like year range, rating thresholds, and actor-specific searches.