ML Project Report

Title of Project: Movie Recommendation System

Student Name: Krishna Gupta

Enrollment Number : 01719011921

Email ID: krishg2k4@gmail.com

Contact Number: 8700490654

Google Drive Link:
https://drive.google.com/drive/folders/1emiUef3CSoopu5pPE5mStlI5JafPtHVC?usp=sharing

Google Website Link:
https://krishna1gupta-movie-recommender.streamlit.app/

Dataset Link:
https://www.kaggle.com/datasets/tmdb/tmdb-movie-metadata

MOVIE RECOMMENDATION SYSTEM

Machine Learning is the last invention that the humans need to ever make

~ NICK BOSTROM

Abstract

This project focuses on developing a movie recommendation system that suggests movies to users based on their selected movie. The system utilizes a provided dataset and recommends movies based on tag similarity. Tags used for similarity include the top 3 cast members, director, genre, specific keywords, and film overview. To process the data, a Bag of Words method is employed to convert the tags into vectors. The cosine similarity method is then utilized to identify movies with similar tags.

The project's objective is to enhance the movie-watching experience by providing personalized recommendations. By analyzing the tags associated with each movie, the system identifies related movies that share similar characteristics. This approach assists users in discovering movies aligned with their preferences and interests.

The implementation of the movie recommendation system involves Python programming and incorporates relevant machine learning and Natural Language Processing Techniques.

The utilization of cosine similarity and the Bag of Words method enhances the accuracy and relevance of the recommendations provided.

Keyword

Movie Recommendation System, cosine similarity, Bag of Words, Natural Language Processing,

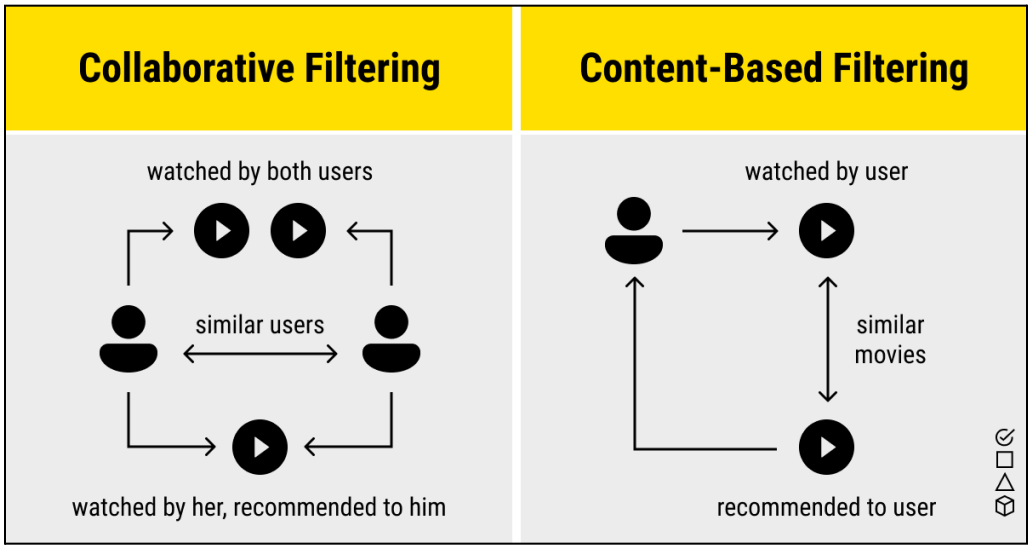
1. Introduction

Movies are becoming the main source of entertainment these days. In India only around 2000 movies are made in approximately 15-20 languages. There are thousands of movies created every year . People find it very difficult to choose among the large piles of movies based upon their interest.

Big streaming giants like Netflix and Amazon Prime have very robust recommendation systems so that the users spend maximum time on their platform.

The recommendation system is mainly of three types. They are as follows:-

1. Content Based
2. Collaborative Based
3. Hybrid Based



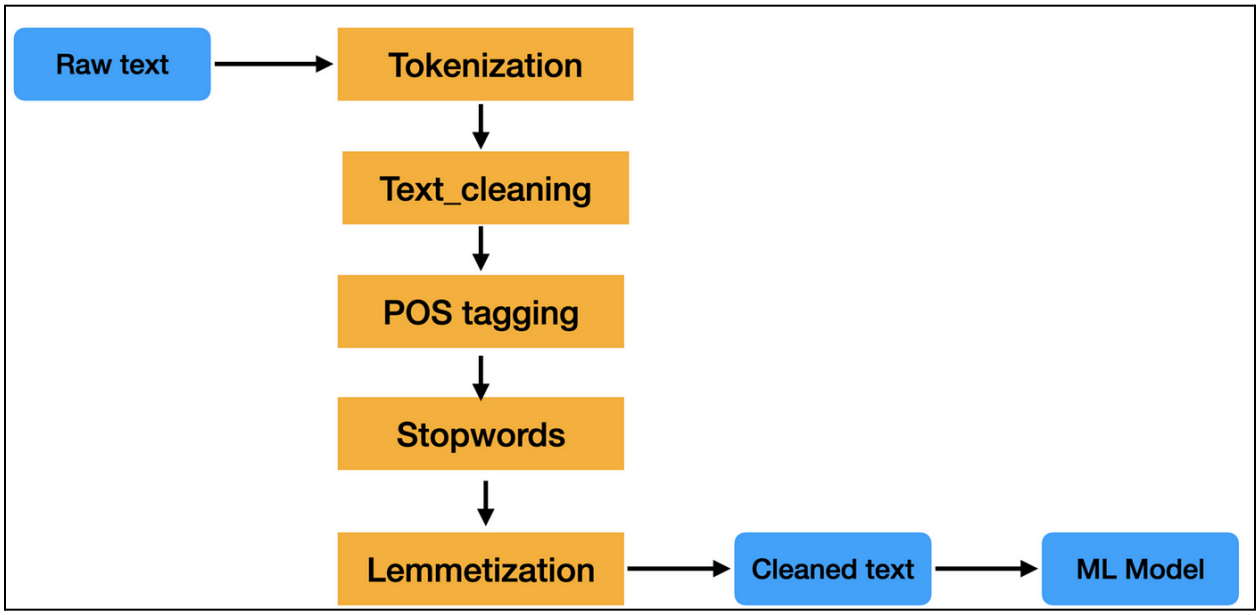
In my project, I have used the Content based recommendation system. A content-based recommendation system recommends items to users based on the

similarity between items' features and the user's preferences. It analyzes the content or attributes of items, such as text descriptions or metadata, and matches them to users' past preferences. For example, in a movie recommendation system, it suggests movies with similar genres, actors, or directors based on the user's previous movie choices. Content-based systems are effective in providing personalized recommendations but may have limitations in discovering new items outside the user's established preferences.

Basic preprocessing steps involve tokenization,stemming,removal of stop words, and making a specific tag column which include genre,Top 3 Cast,Director,Genre and Overview and Keywords associated.

Bag of words technique used to convert the tag column into the vectorized form. It is a popular approach used in natural language processing and text mining. It involves representing text documents as a collection of individual words, disregarding grammar and word order. The technique creates a "bag" or set of words from a corpus, where each word becomes a feature. The frequency or presence of these words is used to construct a numerical representation of the document. By treating each document as a unique combination of words, the Bag of Words technique enables text analysis, information retrieval, and various machine learning tasks such as sentiment analysis, document classification, and topic modeling.

2. Proposed Methodology



2.1 Data Collection

Dataset used in this work is collected from Kaggle. The name of the dataset is TMDB **4809 Movie Dataset**. This dataset was generated from The Movie Database API. This product uses the TMDb API but is not endorsed or certified by TMDb.This data set includes information related to 5000 movies. This dataset contains two csv files tmdb_5000_movies.csv and tmbd_5000_credits.csv. These both file contain info related to cast,crew,movie id, movie title ,budget,revenue, genre,overview etc

2.2 Text Preprocessing

This is one of the time consuming processes, most of the data is not in the required format and needs to be transformed in the desired format to make the recommendation. Text preprocessing steps involve :-

1. Removing the NA values :- Dropping the rows that contain the null value, since there are only three rows that contain the null value, hence that won't make any change which contains in total 4809 movies
2. Checking the duplicate values :- There are no duplicate values present in the dataset
3. Feature selection :- Feature selection done manually by considering the domain knowledge in the mind. Hence both the dataset combined based on the ID and in total 7 columns are selected to do further processing
4. Transforming the columns :- Columns are transformed to fetch the required information. Since the column contains the string of dictionary. The string converted to the list, then the dictionary is accessed and then useful information is extracted
5. Steeming :- Stemming done to bring the different tokenized words to their base word and reduce the unnecessary overhead

2.3 Clustering the movies

Once we get the clean data in the form of a column called 'tags' , text vectorization takes place with the help of CountVectorizer. Total words/ features are restricted to 5000. We get vectors data frame which is of dim 4806 * 5000 (4806 moves and 5000 features). Then with the help of cosine similarity clubbing of the similar movies takes place

3. Result and Discussion

Overall, the results demonstrate the effectiveness of our movie recommendation ML project, providing users with accurate and personalized movie suggestions based on content-based analysis. The project's findings contribute to the growing

field of recommendation systems and pave the way for further advancements in personalized movie recommendations.

4. Conclusion and Future Scope

The future scope involves :-

1. We have considered 5000 words to do analysis, we can reduce that amount to make more precise suggestions
2. Bag of words technique is incorporated in my project, one can replace it with more advanced techniques like Word2Vec and Deep Learning techniques like RNN,CNN and transformers
3. The project involves suggestion of only movies, one can include documentaries, and OTT shows
4. The hybrid approach, combining collaborative filtering and content-based techniques, can yield superior results compared to using either method alone. The system will be able to provide personalized recommendations while also suggesting movies with similar attributes to the ones users preferred.

5. Acknowledgement

I would like to express my sincere gratitude and heartfelt appreciation to Dr. Sanjay Singh, my esteemed ML professor, for providing me with the opportunity to undertake this meaningful ML project. Dr. Singh's guidance, expertise, and support throughout this journey have been invaluable. I would also like to extend my thanks

to my friends and family who have been a constant source of encouragement and support.

I am grateful for the knowledge, guidance, and inspiration I have received from Dr. Sanjay Singh and the unwavering support from my friends and family. Their contributions have played a significant role in shaping my academic journey and personal growth. I am truly fortunate to have such wonderful individuals in my life.

6. References

1. https://www.kaggle.com/datasets/tmdb/tmdb-movie-metadata
2. https://www.youtube.com/@campusx-official
3. https://chat.openai.com/
4. https://scholar.google.co.in/citations?user=7Uad2MUAAAAJ&hl=en