Introduction

Everyone loves movies irrespective of age, gender, race, colour, or geographical location. The most interesting fact which I found is how we all are connected via this medium and yet how unique our choices and combinations are. It is difficult to to say that everyone would like a particular movie because some people like thriller, romance or some people like sci-fi genres movie, while few people like lead actors and directors.

So, here's where a role of data analyst come, we extract all the patterns based on user's behaviour and from the data(movie) itself. So, without any further delay let's jump and discuss our recommendation system.

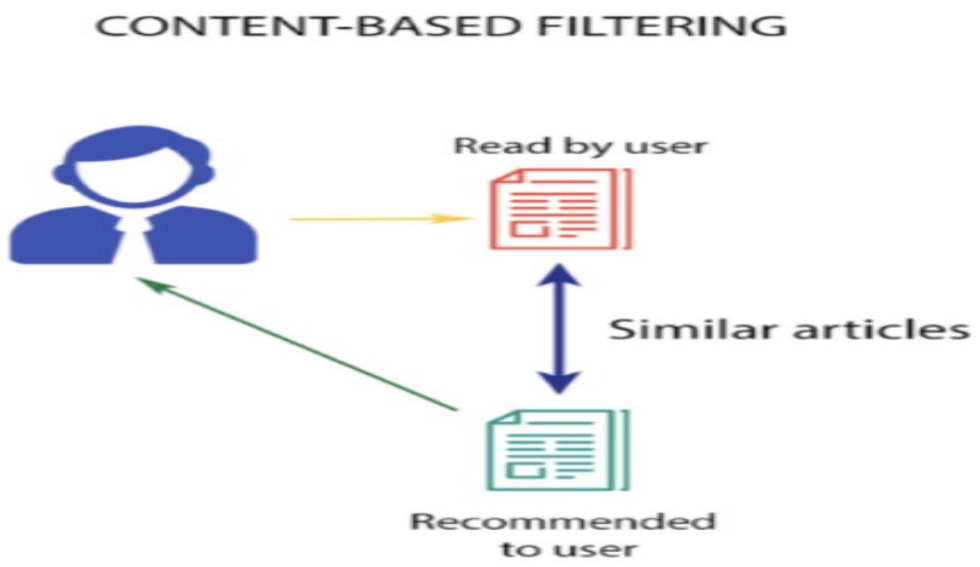
What is a Recommendation System?

In a simple words a Recommendation System is a filtration program whose primary goal is to predict the "preference" or "rating" of a user towards a specific domain. In our case, this is domain-specific recommendation system, where we will predict only those movie which user would prefer based on data of his/her choice.

What are the different filtration strategies?

- Content-based Filtering

  In this filtration strategy the algorithm recommends products which user has liked in the past. Here we are going to use cosine similarity which will be computed from the data we have. For example, if the user likes sci-fi movies like spider-man then we can recommend other marvels movie like Dr. Strange.



- Collaborative-based Filtering

  This filtration strategy is based on the combination of user's interest and compare it with other user's interest. The history of all the users plays an Important role here. For example, if user 'A' likes 'Thor' and 'Batman' and the user 'B' likes 'Batman' and 'Justice League' then they have similar interests because these movies belong to super-hero genre. So, there are high probability that the user 'A' would like 'Justice League' and the user 'B' would like's 'Thor'.

- Hybrid-based Filtering

This is a special type of recommendation system which is a combination of content and collaborative filtering method. Hybrid recommender system approaches can be implemented in different ways like by using content and collaborative-based methods to generate predictions separately and then combining the prediction or we can just add the capabilities of collaborative-based methods to a content-based approach (and vice versa).

About our Project

This project is based on Content-based Filtering methodology. We will create tags based on different columns like ‘genres’, ‘keyword’, ‘cast’, ‘crew’.

Project Flow (High Level Overview)

Data ----- > Preprocessing ------ > ML Model -----> To Website ----- > Deploy

About Dataset

This is a TMDB 5000 Movie Dataset. There are two csv files.

1. Movies: The movie dataset contains information of movies in which we have different columns like genres, homepage, id, keywords etc.
2. Credits: The credits dataset contains information of cast and crew with the name of movie in which they have worked.

Approach of Problem

We will first combine both dataset into single dataset. Then we will extract only those columns which will be helpful in creating tags. Once that done then we proceed for data pre-processing, in which we will do data cleaning, and converting it to useful insight like extracting names of top actors from crew column and converting it to list. After that we will make Machine Learning model in which we will convert lists to array and use “Cosine Similarity” and “Vectorization” to calculate distance between movies. After completing this our Machine Learning model is ready. Now the final part is to make a website and deploy the project and for that we will use Streamlit library for frontend and Heroku cloud for deployment.