**MOVIE RECOMMENDATION SYSTEM**

**Problem Statement:** These days there’s lot of digital data available on almost every platform which creates the issue of finding the appropriate data(whether it might be any data or item in marketing applications) in a vast set of data. In similar way, production companies/industries need to supply the proper data/products based on the user’s requirements and interest which is also required in the market analysis. If users are unable to get the data easily that they actually require, it might result in a big problem of sales of the companies. To resolve this issue and suggest users with their interested data(might be songs, movies, items etc) these recommendation systems act as powerful tools that aims to get the user’s rating and then recommend the data(here movies) on the basis of their matching interests.

**Recommendation Systems are basically built to suggest the relevant data based on the search (user’s), they might be used in recommending videos on YouTube, songs on spotify, movies on Netflix, items on amazon or flip kart, connections on LinkedIn and many more.**

There are mainly three types of recommendation systems.

Recommendation Systems

Collaborative Filtering

Hybrid (both)

Content Based

Involves both the approaches (Contend Based & Collaborative Filtering)

Recommends movies which are highly rated by the similar users (Mostly used approach)

Recommends movies similar to the searched movie

**I have built a Content-Based movie recommendation system that gives 10 most similar movies based on the movie we are searching.**

Datasets I have taken to build this system contains the details of around 5000 movies.

(Movies data of shape (4803, 20) & Credits data of shape (4803, 4))

**Steps I have performed on the dataset :**

* ***Data Cleaning and Analyzing******–*** In this step, Initially, I merged the two datasets on movie\_id column and then checked for null values and duplicates in data. After this, I have analyzed the importance of each column in our system building and removed the unnecessary columns which might create confusion.
* *Text Pre Processing –* In this step, I have transformed the data with few steps like extracting only the required words in each column by writing appropriate functions, splitting the text into separate strings, combining the three columns data into a new column as ‘Tags’ which is the main column that helps us search the most similar movie and then creating a new dataframe with three columns containing movie id, title and Tags. This will be the main dataset on which we will be performing our further steps to build a recommendation system.
* *Stemming the data –* Now performing stemming on Tags column which will reduce every string to its root word. This will avoid the problem of words repetition in different forms.
* *Vectorizing the textual data –* When the data is ready we need to convert it into numerical data i.e, vectors using CountVectorizer which makes the process easy of finding similarity among movies(Just by calculating distance between the movies we can recommend).And also stopwords(words with no meaning like – are, is, the, to etc) are ignored.
* *Finding Cosine Similarity –* I have calculated the cosine distance among the movies to find the similarity and the movies with minimum distance are displayed as the recommendations to the searched one (I did not calculate Euclidean distance because it is not an accurate measure to be applied here as it fetches the distance between two end points and we need the minimum angle between vectors).