Movie Recommendation System

Vaishali Bharatsinh jadeja(0781141)

DAB 402

A capstone project submitted

in partial fulfilment of the requirements for the degree of

DATA ANALYTICS FOR BUSINESS

in

Data Analysis

SAINT CLAIR COLLEGE FOR APPLIED ARTS AND TECHNOLOGY

Mississauga, ONTARIO, CANADA

SUBMITTED TO

Professor Savita Seharawat

ABSTRACT

A recommendation engine filters the data using different algorithms and recommends the most relevant items to users. It first captures the past behaviour of a customer and based on that, recommends products which the users might be likely to buy.A movie recommendation is important in our social life due to its strength in providing enhanced entertainment. Such a system can suggest a set of movies to users based on their interest, or the popularities of the movies.we propose a movie recommendation system that has the ability to recommend movies to a new user as well as the others. It mines movie databases to collect all the important information, such as, popularity and attractiveness, required for recommendation. Computation and processing service from service providers and product distributors. To recommend movies, first collects the ratings for users and then recommend the top list of items to the target user . In addition to this, users can check reviews of other users before watching movie. A different recommendation schemes have been presented includes collaborative filtering, content-based recommender system, and hybrid recommender system.We extract aspect-based specific ratings from reviews and also recommend reviews to users depends on user similarity and their rating patterns. Finally, validating the proposed movie recommendation system for various evaluation criteria, and also the proposed system shows better result than conventional systems.

INTRODUCTION

A recommender system is a simple algorithm whose aim is to provide the most relevant information to a user by discovering patterns in a dataset. The algorithm rates the items and shows the user the items that they would rate highly. The project focuses on a recommendation engine that filters the data using different algorithms and recommends the most relevant items to users. It first captures the past behavior of a customer and based on that, recommends products which the users might be likely to buy.An example of recommendation in action is when you visit Amazon and you notice that some items are being recommended to you or when Netflix recommends certain movies to you.

LITERATURE SURVEY

Different types of recommendation engines:

The most common types of recommendation systems are content-based and collaborative filtering recommender systems. In collaborative filtering, the behavior of a group of users is used to make recommendations to other users. The recommendation is based on the preference of other users.

A simple example would be recommending a movie to a user based on the fact that their friend liked the movie. There are two types of collaborative models Memory-based methods and Model-based methods. The advantage of memory-based techniques is that they are simple to implement and the resulting recommendations are often easy to explain. They are divided into two:

- **User-based collaborative filtering**: In this model, products are recommended to a user based on the fact that the products have been liked by users similar to the user. For example, if Derrick and Dennis like the same movies and a new movie come out that Derick like, then we can recommend that movie to Dennis because Derrick and Dennis seem to like the same movies.

- **Item-based collaborative filtering**: These systems identify similar items based on users' previous ratings. For example, if users A, B, and C gave a 5-star rating to books X and Y then when a user D buys book Y they also get a recommendation to purchase book X because the system identifies book X and Y as similar based on the ratings of users A, B, and C.

Model-based methods are based on Matrix Factorization and are better at dealing with sparsity. They are developed using data mining, machine learning algorithms to predict users' rating of unrated items. In this approach techniques such as dimensionality reduction are used to improve accuracy. Examples of such model-based methods include Decision trees, Rule-based Model, Bayesian Model, and latent factor models.

Content-based systems use metadata such as genre, producer, actor, musician to recommend items say movies or music. Such a recommendation would be for instance recommending Infinity War that featured Vin Diesel because someone watched and liked The Fate of the Furious. Similarly, you can get music recommendations from certain artists because you liked their music. Content-based systems are based on the idea that if you liked a certain item you are most likely to like something that is similar to it.

Architectural Diagram:-

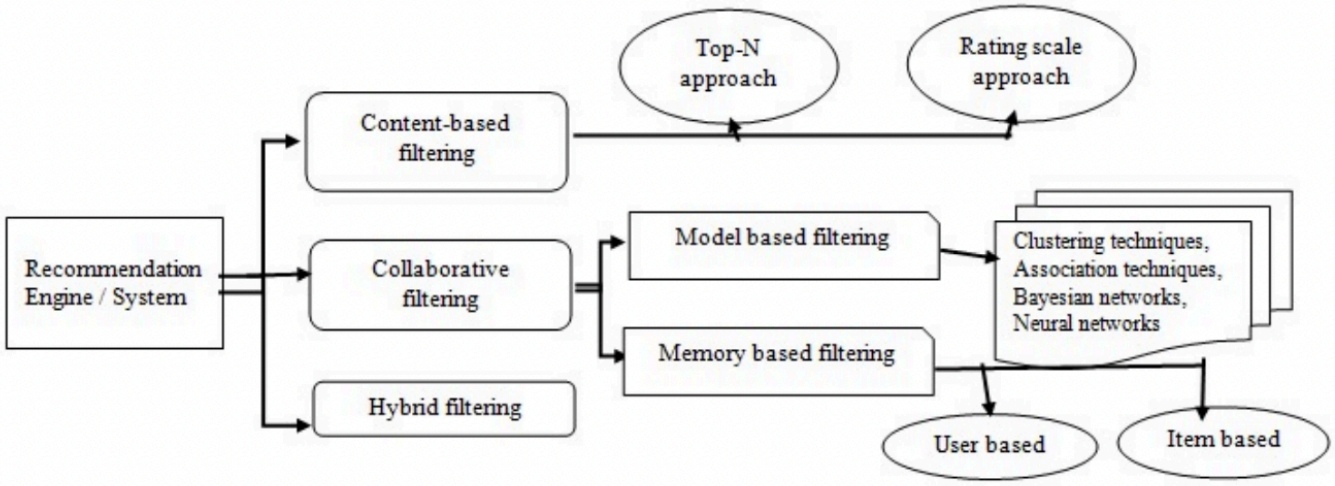


Figure.4: Architectural diagram

TECHNOLOGY USED:

1. Python

Python is an interpreted , high-level , general-purpose programming language . Created by Guido van Rossum and first released in 1991, Python's design philosophy emphasizes code readability with its notable use of significant whitespace . Its language constructs and object-oriented approach aim to help programmers write clear, logical code for small and large-scale projects. Python is dynamically typed and garbage - collected . It supports multiple programming paradigms , including procedural , object-oriented, and functional programming . Python is often described as a "batteries included" language due to its comprehensive standard library .

2. Jupyter Notebook

The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more.

METHODOLOGY

Recommendation System using collaborative filtering method will read the data entered by the user, like user's profile, rating and interest. Then the system will determine the parameters that will be used to compare it with other member's data in the database. The system will look for similarities with other members such as common interests. Based on these similarities, the system can provide appropriate community recommendations to the user.

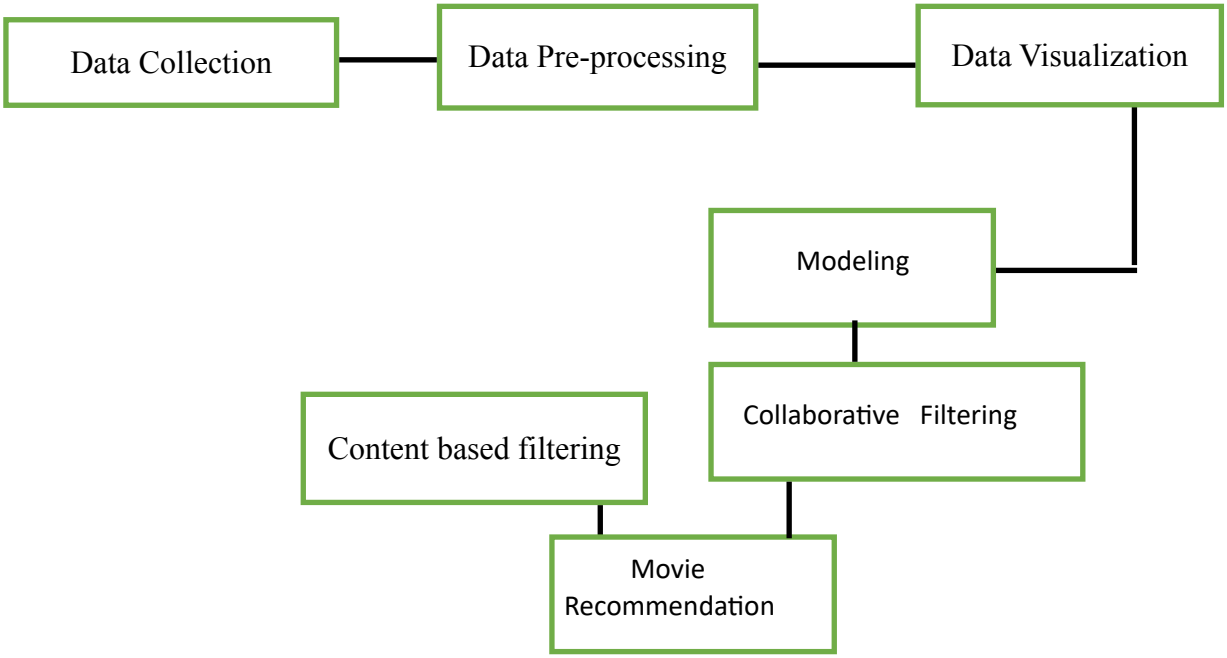


Figure.2 data model for movie recommendation system

DATA COLLECTION

DATASET

The Movie Database - The data which will be used in building our recommendation engine was obtained from Movie Lens, an online movie recommendation service. It contains over 1048575 ratings and 45463 movies. The dataset was developed by collecting movie preference information from 10656 users. In this project i have two data files – movies and ratings.

1. **Movies** – This file contains all the information regarding movies.it contains approximately 45463 and 19 columns.

2. **Ratings** – This file contains data about user ratings.it contains approximately 148000 records and 4 columns

Table 1. Movie file datatypes

Features(Movies)	**Data Types**
adult	bool
budget	int64
genres	object
movieid	int64
imdb_id	object
original_language	object
overview	object
popularity	float64
production_companies	object
production_countries	object

Features(Movies)	Data Types
release_date	object
revenue	float64
runtime	float64
spoken_languages	object
status	object
title	object
video	object
vote_average	float64
vote_count	float64

FeaturesrRatings(Ratings)	Types
Userid	Int64
Movieid	int64
Rating	Float64
Timestamp	int64

Table 2. Rating file datatypes

DATA PRE-PROCESSING

Summary of Categorical attributes

In data processing data type of some attributes is changed. These attributes are status, original languages and spoken languages. Summary of these attributes is given below-

Categorical attributes	count	Nunique	Top	Freq
Status	44349	6	Released	44010
Original_languages	44349	49	English	32062
Spoken_languages	44349	1921	English	22203

Table 3. Categorical attributes

Data Cleaning

Description of movies content data consists approximately 24 columns. Most of those columns were not necessary for our research paper and were therefore removed. The dataset contains plenty of empty values and duplicate values that need to be resolved. Additionally, there have been some entries for movies within the user rating files from movies, which did not correspond to any movies within description of movie content data. These entries were erased for easy processing.

Deleted attributes – The number of attributes that were deleted from original database are given below:

Attribues	Data Types
Belongs_to_collection	Object
Homepage	Object
Poster_path	Object
Tagline	Object

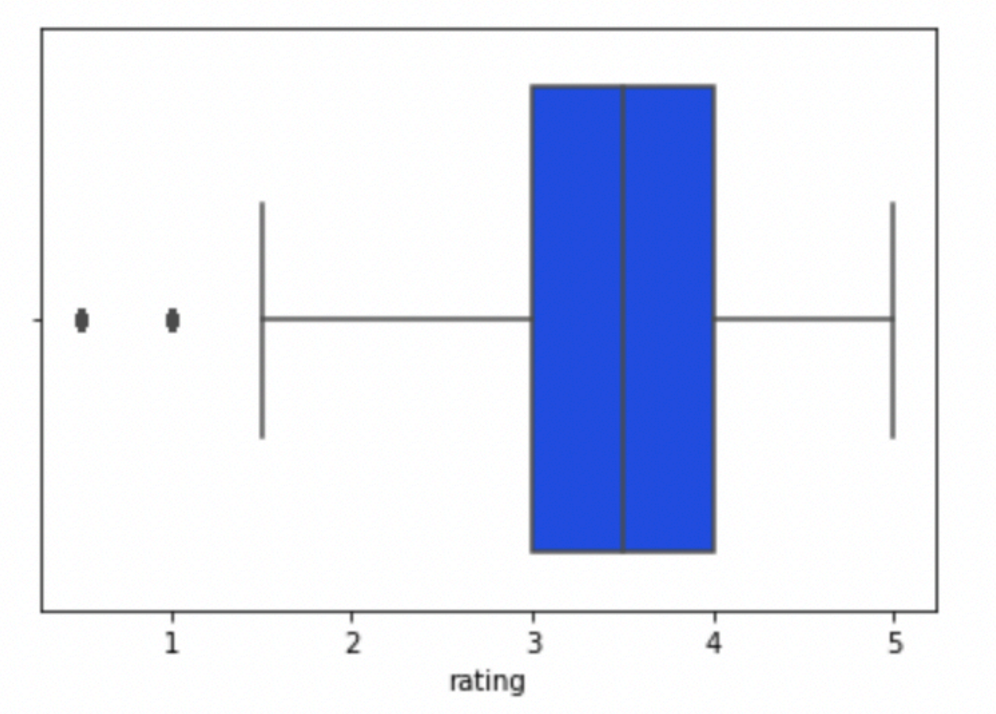
Table 4. Deleted attributes

Summary of null values:

Attributes	Count of null values
imdb_id	17
original_language	11
Overview	954
Popularity	3
production_companies	3
production_countries	3
release_date	87
Revenue	3
Runtime	260
spoken_languages	3
Status	84
Title	3

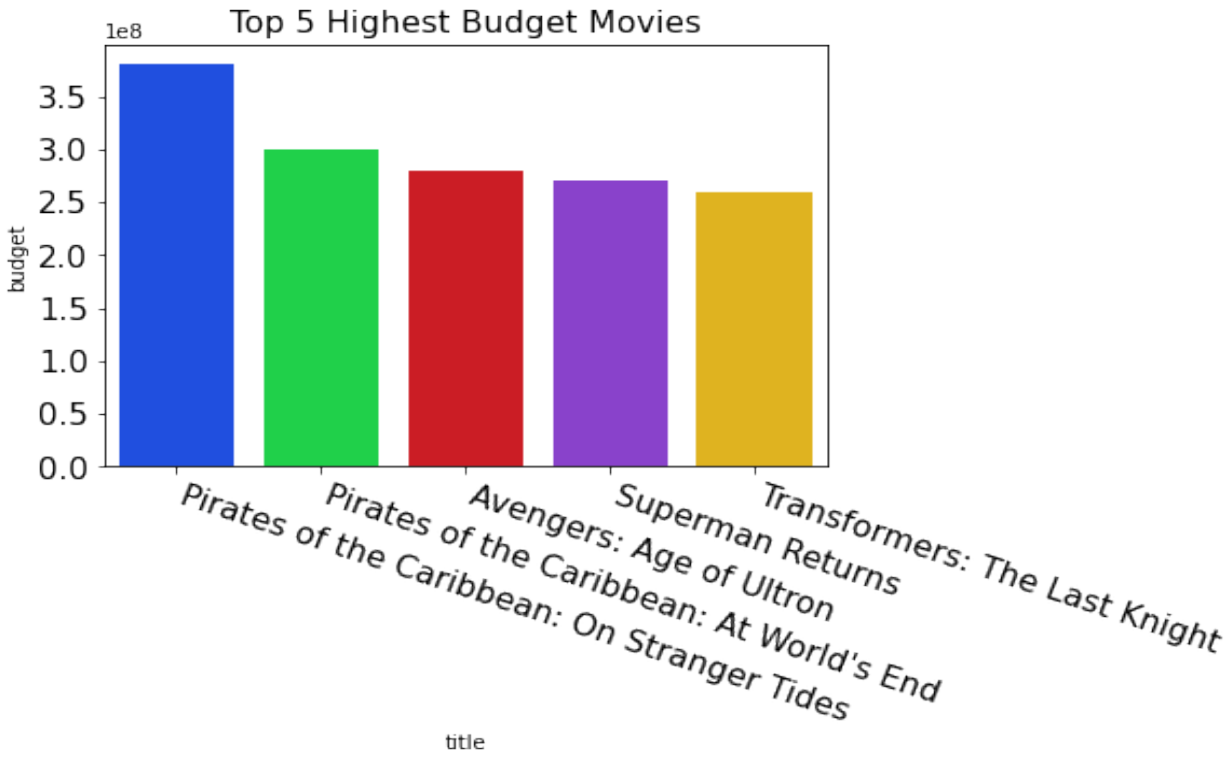
Outlier In rating file:

Outlier: An outlier is a data point in a data set that is distant from all other observations. A data point that lies outside the overall distribution of the dataset.



DATA visualization:

1. **Bar-plot of top 5 highest budget movie**



In this bar-plot name of top 5 highest budgeted movies are given on x-axis. On y-axis lowest to highest budget amount is given in millions. From bar-plot its visible that pirates of Caribbean is the movie with the highest budgeted movie with budget of more than 3.5 million.

2.Total number of rating:

In ratings file users give ratings 0 to5 . So on x axis is ratings and y axis it shows number of count.

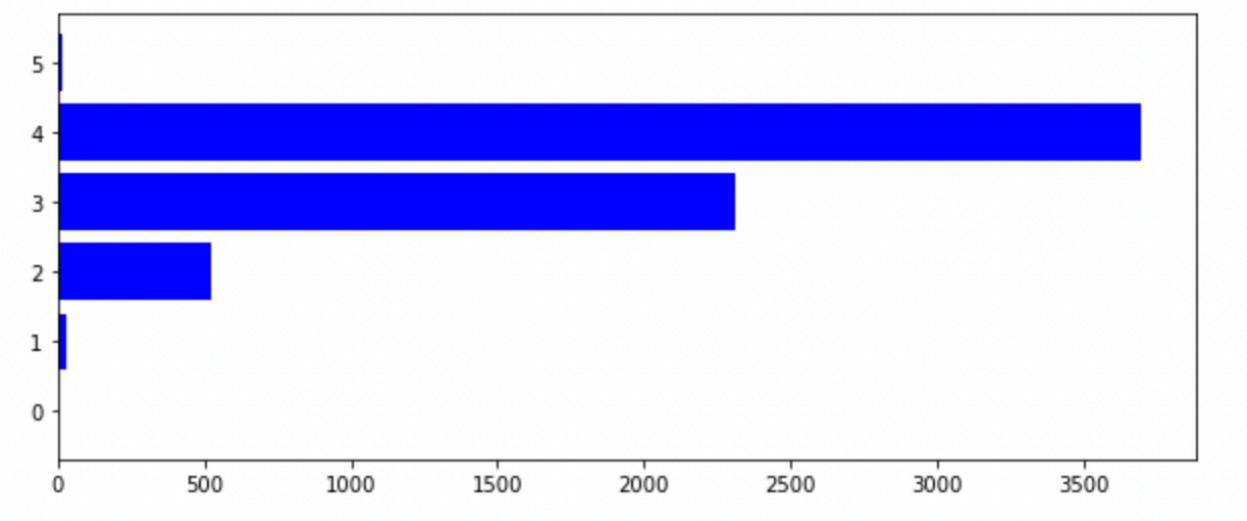


Figure.3 total no of rating

3. TOP 10 GENRE combination:

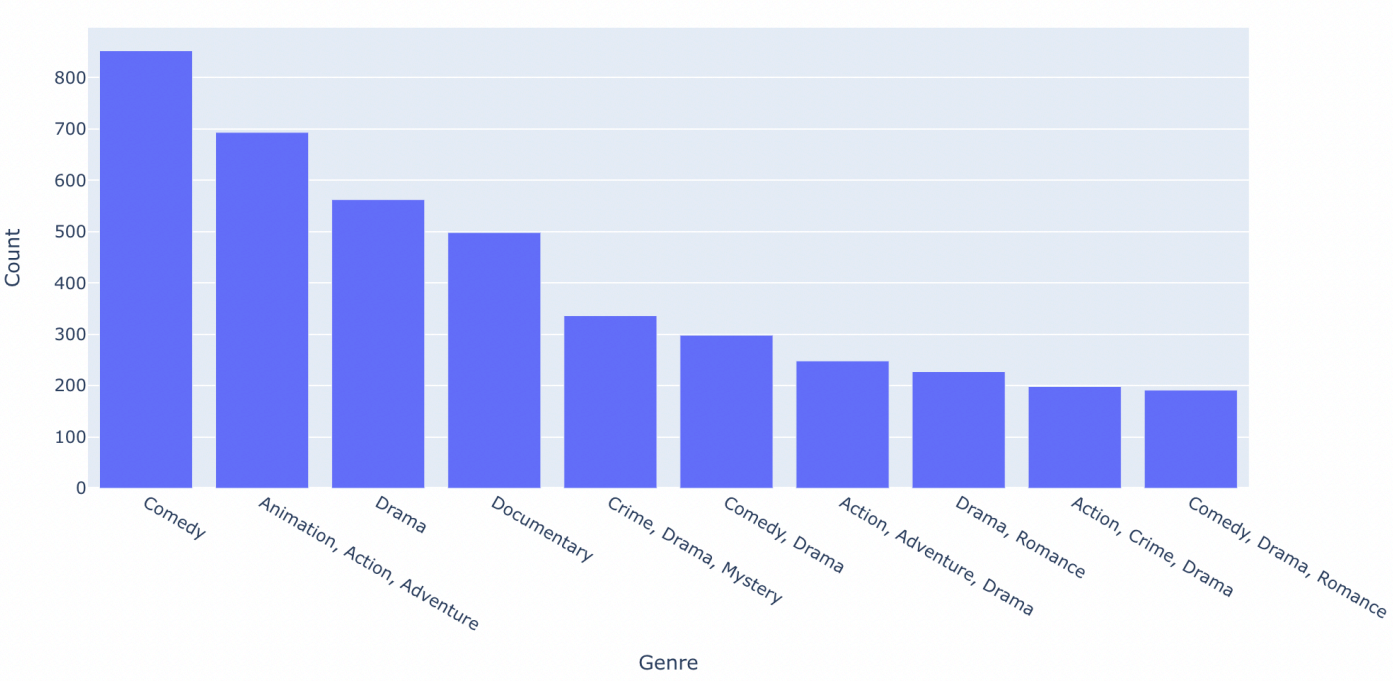


Figure.4 top 10 genre

4. Top 10 budget movie:

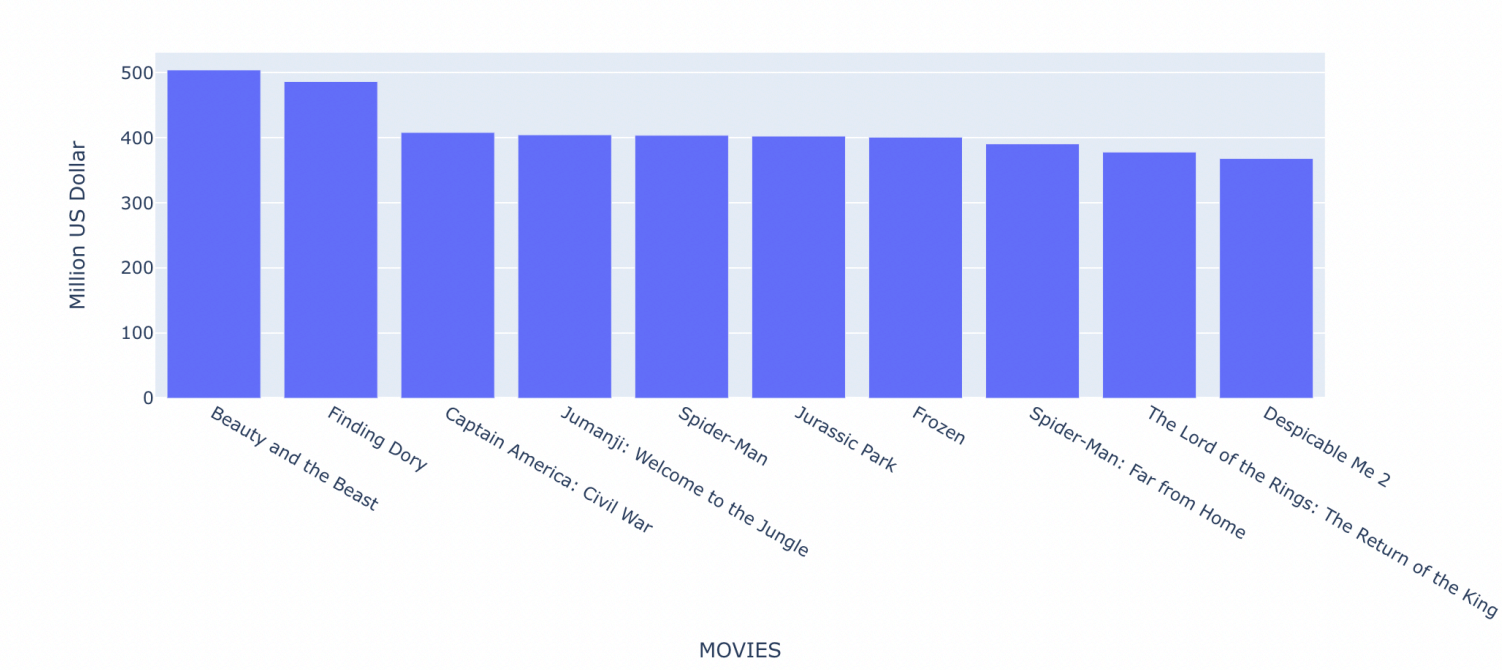
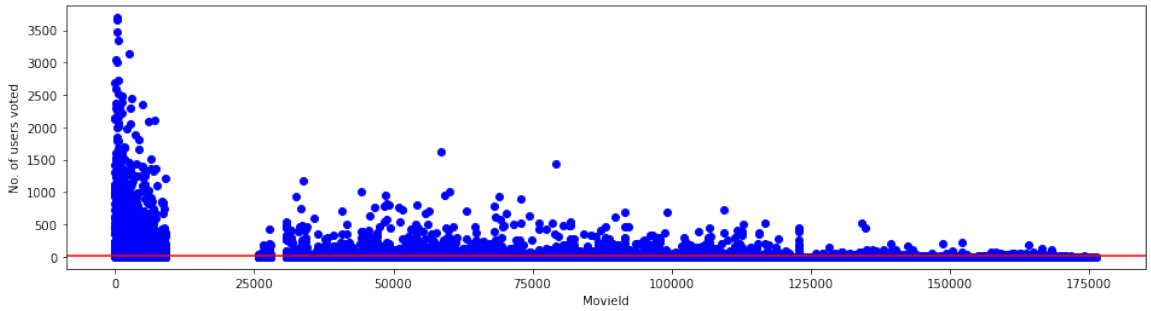


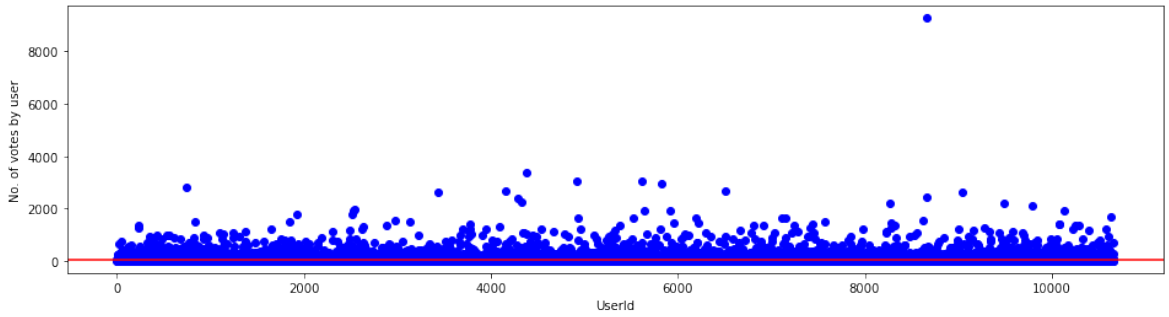
Figure.5 top 10 budget movie

5. Scatter plot of movies and ratings



In this scatter plot total number of movies id is given on x-axis. On y-axis number of user voted is given. From scatter plot it's the movies that are highly voted lies between 0 to 25000.

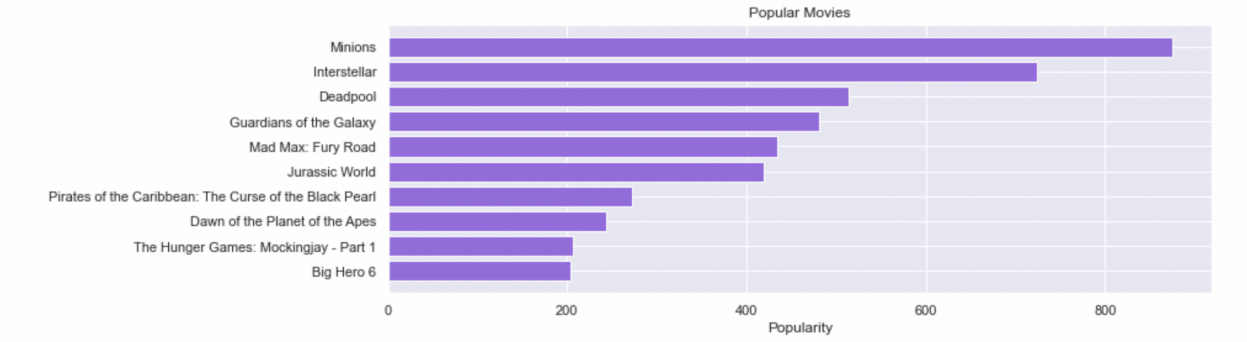
6. Scatter plot of users and votes



In this scatter plot total number of user id is given on x-axis. On y-axis number of votes by users is given. From scatter plot it's the movies that are highly voted users lies between 8000 to 10000.

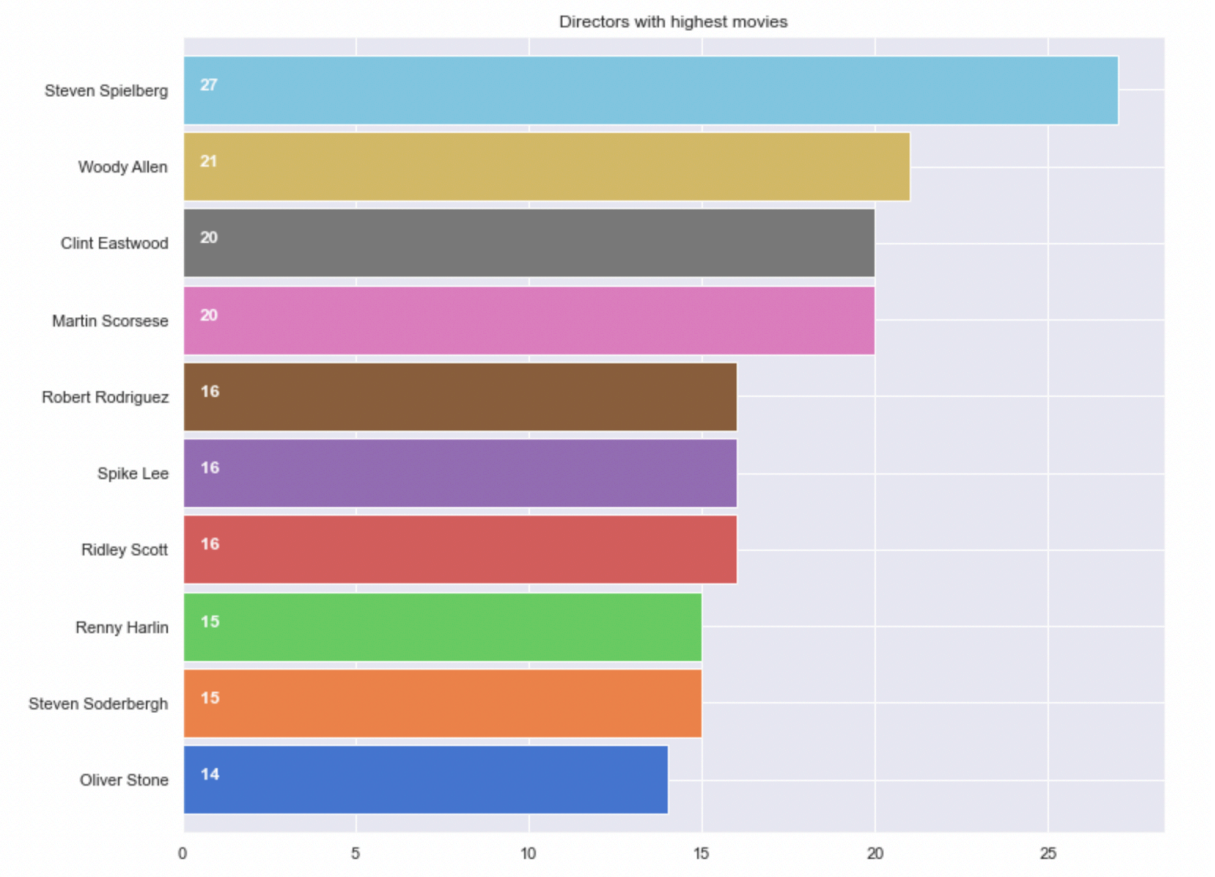
7. Top 10 Popular movie

Here, top 10 movies by popularity column.



8. Directors with highest movies

In this graph, top 10 directors with highest movies. So Steven Spielberg has the highest movie director.



Correlation:

A correlation matrix is a table containing correlation coefficients between variables. Each cell in the table represents the correlation between two variables. The value lies between -1 and 1. A correlation matrix is used to summarize data, as a diagnostic for advanced analyses and as an input into a more advanced analysis.

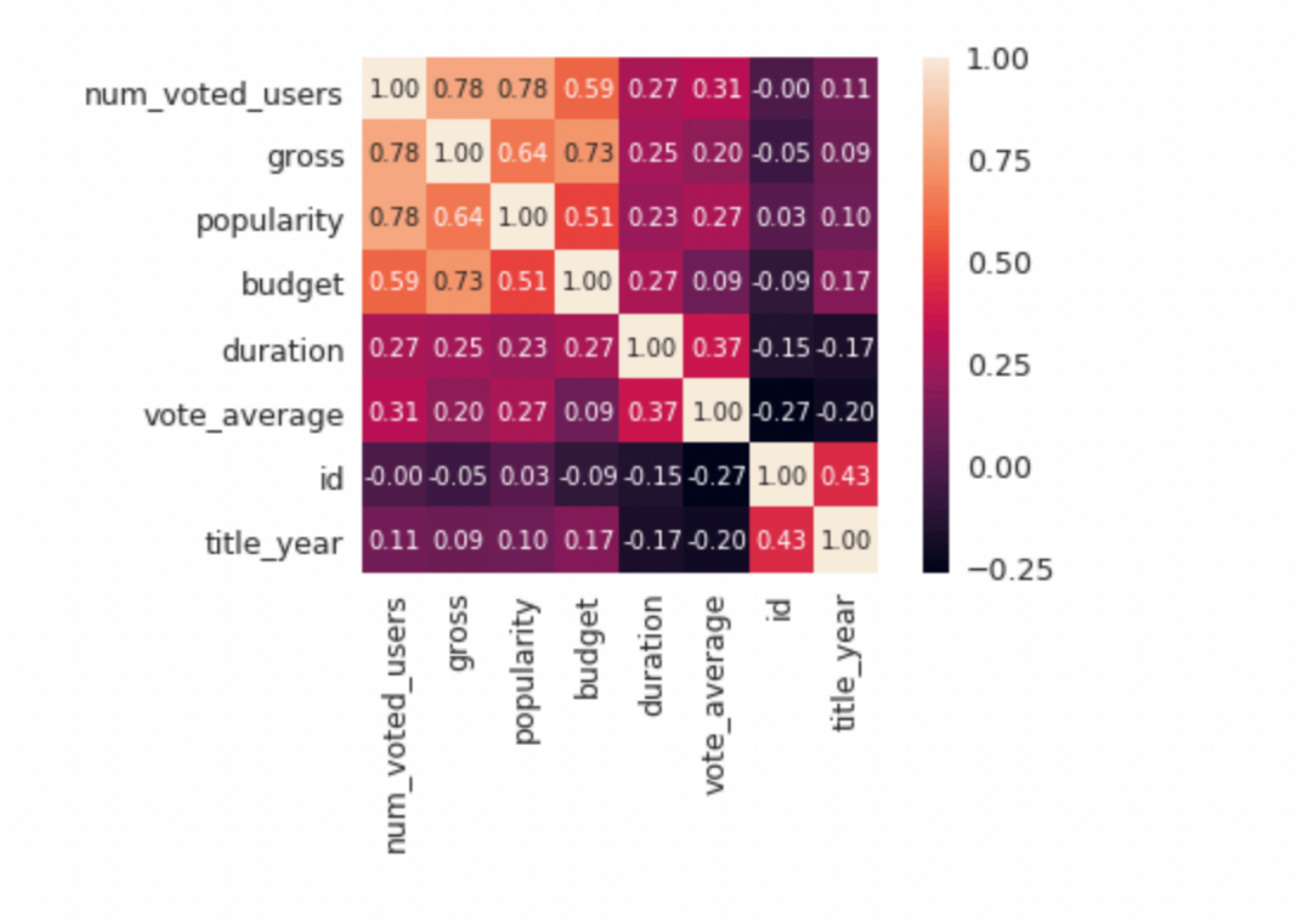


Fig. Heat map of correlation matrix

Data Modeling

Now that we have a decent dataset (or more), it's time to start studying it by creating graphs. Visualization is the finest approach to examine and convey your discoveries when working with massive amounts of data, and it is the next phase of any data analytics project.

Model implementation:

Collaborative filtering- in our project in collaborative filtering technique nearest neighbours algorithm is implemented.its functioning is explained below-

Calculate the K -Nearest Neighbours:

The **NearestNeighbors()** in the **sklearn.neighbors** library can be used to calculate the distance between movies using the **cosine similarity** and find the nearest neighbors for each movie. The number of the nearest neighbors (**n_neighbors**)is set to be 6. Since this includes the movie itself, generally it finds two nearest movies except the movie itself.

indices shows the indices of the nearest neighbors for each movie. Each row corresponds to the row in the dataframe. The first element in a row is the most similar (nearest) movie. Generally, it is the movie itself. The second element is the second nearest, and the third is the third nearest.

distances shows the distance between movies. A smaller number means the movie is closer.

```
Recommendations for Battle Royale II: Requiem:

1: Walk the Line, with distance of 0.7011097583023178:
2: The Dress, with distance of 0.7895589387239221:
3: Five Times Two, with distance of 0.7898581659753098:
4: A Tale of Two Cities, with distance of 0.805921764053694:
5: I'll Sleep When I'm Dead, with distance of 0.8103002835473896:
```

Collaborative filtering - The idea behind **Content-based (cognitive filtering)** recommendation system is to recommend an item based on a comparison between the content of the items and a user profile.In simple words,I may get recommendation for a movie

based on the description of other movies.in our project we have used **Term Frequency Inverse Document Frequency** (TF-IDF).

Once gathering all data as per our need we have chosen **TF-IDF** to create the vectorizer of our words. The reason behind choosing this algorithm is to give less weight to the words that are occurring frequently example (the, is, and etc.).

We have used the cosine similarity matrix to understand the overview column and identify similar patterns and structure from it. So we first calculate the cosine similarity matrix by importing linear kernel.

Then we have done reverse mapping of the movie titles and the indices of the dataframe which helped in identifying the index of a movie in our metadata dataframe and we also droped the duplicate values in the process.

Finally we defined a function which will get recommendations based on the movie title entered by the user.

If we select movie name(content_based('Blue in the Face’))

So the output is,

```
18867                                    The Raid
2885                                Licence to Kill
14018                                 The Collector
761                                   Trainspotting
36055                                  Ride Along 2
11610                                Cocaine Cowboys
20765                Don : The Chase Begins Again
18330          Sherlock Holmes: A Game of Shadows
21235                                Despicable Me 2
11047    Pirates of the Caribbean: Dead Man's Chest
```

Conclusion-

Recommender system has become more and more important because of the information overload. For content-based recommender system specifically, we attempt to find a new way to improve the accuracy of the representative of the movie. For the problems we mentioned at beginning, firstly, we use content-based recommender algorithm which means there is no cold start problem. We introduced the cosine similarity which is commonly used in industry. For the weight of features, we introduced TF-IDF which improve the representative of the movie. We introduce a new approach for setting weight for these features, the movie can be represented more accurately by TF-IDFC which is the key point of our research. In the end of the project, we use k-NN and various metrics to evaluate the improvement of the new approach. It is illustrated that the new approach contributes positively according to the evaluation.

Reference:

- [1] Roettgers, 2014. “Netflix spends $150 million on content recommendations every year”. URL:https://gigaom.com/2014/10/09/netflix-spends-150-million-on-content-recommendations-every-year/. Date last accessed: 02/03/2017
- [2] H. Li, F. Cai, and Z. Liao, ”Content-based filtering recommendation algorithm using HMM”, IEEE Fourth International Conference on Computational and Information Sciences, 275-277, 2012.
- [3] C. A. Gomez-Uribe and N. Hunt, ”The Netflix Recommender System: Algorithms, Business Value, and Innovation”, ACM Transactions on Management Information Systems 6(4):1-19, 2015. DOI: http://dx.doi.org/10.1145/2843948

- [4] O. Celma, ` Music Recommendation, Springer, Berlin, Heidelberg, 43-85, 2010. DOI: http://dx.doi.org/10.1007/978-3-642-13287-2 3
- [5] F. Zhang, Z. Zheng, Y. Xiang, N. J. Yuan, X. Xie, and Z. Li ”DRN: A Deep Reinforcement Learning Framework for News Recommendation”, Proceedings of the World Wide Web Conference, 167-176, 2018. DOI: http://dx.doi.org/10.1145/3178876.3185994
- [6] W.-C. Kang, C. Fang, Z. Wang, and J. McAuley, ”Visually-aware fashion recommendation and design with generative image models”, IEEE International Conference on Data Mining, 207-216, 2017.

Github:-