Big Data Analytics (CSE 3053)

Movie Watching Preferences

Submitted by

Aryaveer Sadal (20BDS0192)

Rashi Maheshwari (19BDS0006)

Under the kind guidance of

Anuradha G

School of Computing Science and Engineering

Vellore Institute of Technology, Vellore

May,2022

1. Abstract

Movies are one of the best sources of entertainment available for us. Our project is about gathering data on movie watching preferences and conclude what type of movie would a person chose if there were a list of movies in front of him. The criteria's we have chosen for this project are based up on genre, movie duration, ratings, language of the movie, place of watching movie, etc. So, we conducted a survey and collected the data on movie watching preferences. Our primary aim is to process the collected data by using big data techniques and to present a report. What we have done is that we have the given the most priority to one option and the rest are given less and less priority according to data, for example in genre, action is given the highest priority or number because it is has the most votes from the data collected through forms and like that all the others will get less and less priority. This is done for all the criteria. All the movies will be checked according to the priority system and the movie with the most priority will be selected. By the end of this project, we are expecting to show which movies are most likely preferred to watch from the available set of movies by the audience.

2. Literature Review

2.1. Netflix Bigdata Analytics- The Emergence of Data Driven Recommendation [1]

This paper has discussed about the diverse kinds of recommendation system like content based and collaborative filtering recommendation system and use of Hybrid recommendation technique by Netflix. In Netflix recommendation system collects user data such as the location of a user; content watched by the user, user interests, the data searched by the user, and the time at which user watched. They have also discussed how Netflix with the help of bigdata analytics can predict the viewing habits of subscriber and how it helped in producing the original content which would be an enormous success. The production or original content is decided by running Bigdata analytics and other data mining algorithms to determine the size of the audience who would watch these original contents. Netflix uses A/B test to personalize user experience. Whatever it shows on the platform (e.g.: images, video clips) is driven by data collected by A/B test.

Netflix keeps tracks of various things including, when a user stops a content, when a user fast forwards a content, time during which the user watch, location of the user, on what device does he watch, browsing behavior of user, reviews and ratings given by user.

2.2. Profiles and preferences of OTT users in Indian perspective [2]

This paper is in a novel approach to understand the user profiles and preferences from an Indian perspective. This paper also studies the major shift towards the consumption of old entertainment to new entertainment. The methodology used for this research is quite simple, first, a standardized questionnaire is prepared for collection of data. This questionnaire was conducted for age group 18 – 55 + who watch OTT for their entertainment. After the collection of data, various variables are selected which will be used for further analysis. Data is then visualized using SPSS Statistics software and visualization techniques. Apart from the visualization, this paper also focuses on the relation between various variables using Chi-Square Tests.

Not only, the collection of data was done from the same region, it also contained only 404 responses which limits the scope of study. This can result in inaccurate results. Thus, to make sense for any business decision there wants to be a bigger sample size and geography.

2.3. Movie Success Prediction using Historical and Current Data Mining [3]

In this paper, researchers proposed a model where they consider several factors, each factor is assigned by a weight and success/failure of the upcoming movies is predicted based on the factor's value. The dataset used is scraped from IMDb, Rotten Tomato and YouTube. This dataset contains around one thousand data entries. The entire system is divided into various modules: Data Collection, Data Cleaning, Data Analysis, Data Classification and Calculation. The success of a movie is based on revenue of the movie and so they have calculated how strongly some factors influences revenue, i.e., success of a movie. Finally, based on each factor, a movie is classified as poor, average, good and exceptionally good and according given a score. Then, the mean of these scores is calculated and then according to that mean a movie is predicted as disaster, flop, average, hit, super hit and blockbuster.

In this work, researchers have only used features related to a movie, but a movie's success does not depend only on those features, it also depends on what the audience wants or in other words what they prefer.

2.4. Predicting the Conceptual Appeal of Movies using Data Analytics [4]

This paper aims to explore what factors are necessary to predict the quality of a movie in terms of its concept, how to establish a relationship between distinct categories and what are various techniques that can used for predictive analysis. The proposed model used the concept of correlation between different parameters that would best suit the interest of the stakeholders. A correlation was found between genres, ratings and actors. Random Forest algorithm was used for prediction. This algorithm is a combination of many decision trees. This algorithm provides good accuracy, robust to noise and outliners, provides estimates for strong correlations and strengths, and solves the problem of overfitting the training data.

The studies conducted, aimed to determine those features of movies that would increase the box office success. However, it is quite clear that box office success is not the most crucial factor to ensure if a movie is conceptually appealing. And so, it is important to understand public opinions and preference.

2.5. Prediction of Movies popularity Using Machine Learning Techniques [5]

The proposed research aims to predict movies popularity. Researchers have used machine learning approach for the experimentation. Machine learning have powerful classification algorithms for classification. This research aims to improve previous research. The proposed methodology contains following steps: Data Extraction, Data Preprocessing, Data Integration and Transformation, Feature Selection and Classification. The dataset was collected from internet movie database (IMDb). It contains about 2000 data points. The data is inconsistent, missing and very noisy as well. To cater missing fields issue, they have used central tendency as a standard for filling missing values for attributes. They have also done some transformations that can easy the process of prediction and can provide consistency in data. Then, they have performed feature selection which will remove those attributes that do not add information to the analysis process.

Performing data mining on a large dataset is a difficult task because there are so many attributes that play an important part in analysis. Moreover, these attributes can have can be in different dimensions with lots of noisy data and missing fields. So, data preprocessing and transformation is necessary.

3. Architectural Representation

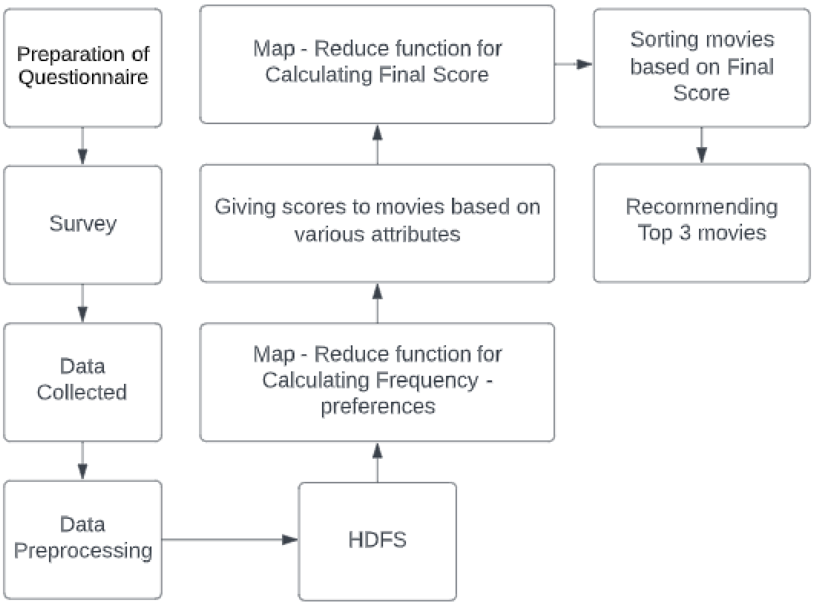

Preparation of Questionnaire
Survey
Data Collected
Data Preprocessing
HDFS
Map - Reduce function for Calculating Frequency - preferences
Giving scores to movies based on various attributes
Map - Reduce function for Calculating Final Score
Sorting movies based on Final Score
Recommending Top 3 movies

Fig: Our approach/architecture flow diagram

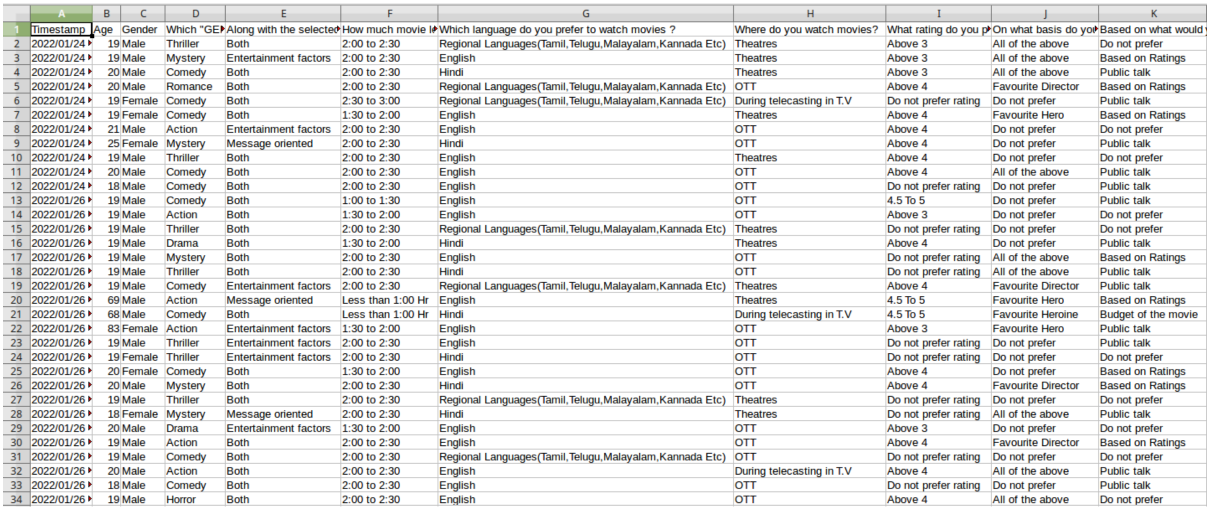
4. Module Explanation

4.1. Data Collection:

We prepared a questionnaire containing various questions that would describe someone’s movie preferences. We have taken various parameters under consideration while preparing questionnaire, such as their age, gender, genre they prefer, what do they enjoy in a movie, length of the movie they will go for, in which language do they prefer to watch movies in, on which platform they watch movies, what is basic rating a movie should have, why do they watch a movie, etc. Below is a table showing summary of the parameters.:

Parameter	**No. of Values**	**Values**
Timestamp	-	-
Age	-	-
Gender	3	Female, Male, Others
Which "GENRE" do you prefer?	8	Action, Comedy, Drama, Fantasy, Horror, Mystery, Romance, Thriller
Along with the selected genre in the above question, what do you enjoy more about a movie?	3	Message oriented, Entertainment factors, Both
How much movie length, do you prefer to watch? (in hours)	5	Less than 1:00 hr, 1:00 to 1:30, 1:30 to 2:00, 2:00 to 2:30, 2:30 to 3:00
Which language do you prefer to watch movies?	3	English, Hindi, Regional Languages
Where do you watch movies?	3	Theatres, OTT, During telecasting in TV
What rating do you prefer to go to a movie?	5	4.5 to 5, Above 4, Above 3, Above 2, Do not prefer
On what basis do you prefer to watch a movie?	5	Favourite Hero, Favourite Heroine, Favourite Director, All of the above, Do not prefer
Based on what would you decide to go to a movie?	4	Based on rating, Public talk, Budget of the movie, Do not prefer

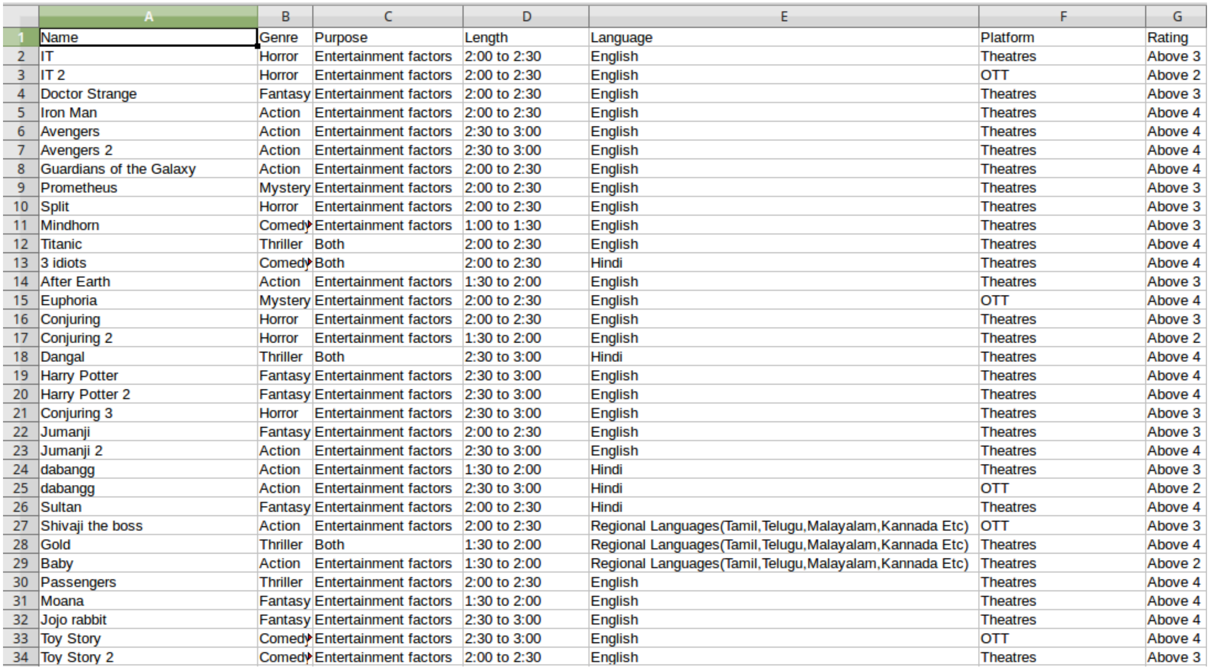
We made a google form and circulated for the survey. Data collected from the survey has about 2079 records.



	A	B	C	D	E	F	G	H	I	J	K
1	Timestamp	Age	Gender	Which "GE	Along with the selected	How much movie le	Which language do you prefer to watch movies ?	Where do you watch movies?	What rating do you p	On what basis do you	Based on what would y
2	2022/01/24	19	Male	Thriller	Both	2:00 to 2:30	Regional Languages(Tamil,Telugu,Malayalam,Kannada Etc)	Theatres	Above 3	All of the above	Do not prefer
3	2022/01/24	19	Male	Mystery	Entertainment factors	2:00 to 2:30	English	Theatres	Above 3	All of the above	Based on Ratings
4	2022/01/24	20	Male	Comedy	Both	2:00 to 2:30	Hindi	Theatres	Above 3	All of the above	Public talk
5	2022/01/24	20	Male	Romance	Both	2:00 to 2:30	Regional Languages(Tamil,Telugu,Malayalam,Kannada Etc)	OTT	Above 4	Favourite Director	Based on Ratings
6	2022/01/24	19	Female	Comedy	Both	2:30 to 3:00	Regional Languages(Tamil,Telugu,Malayalam,Kannada Etc)	During telecasting in T.V	Do not prefer rating	Do not prefer	Public talk
7	2022/01/24	19	Female	Comedy	Both	1:30 to 2:00	English	Theatres	Above 4	Favourite Hero	Based on Ratings
8	2022/01/24	21	Male	Action	Entertainment factors	2:00 to 2:30	English	OTT	Above 4	Do not prefer	Do not prefer
9	2022/01/24	25	Female	Mystery	Message oriented	2:00 to 2:30	Hindi	OTT	Above 4	Do not prefer	Public talk
10	2022/01/24	19	Male	Thriller	Both	2:00 to 2:30	English	Theatres	Above 4	Do not prefer	Do not prefer
11	2022/01/24	20	Male	Comedy	Both	2:00 to 2:30	English	OTT	Above 4	All of the above	Public talk
12	2022/01/24	18	Male	Comedy	Both	2:00 to 2:30	English	OTT	Do not prefer rating	Do not prefer	Public talk
13	2022/01/26	19	Male	Comedy	Both	1:00 to 1:30	English	OTT	4.5 To 5	Do not prefer	Public talk
14	2022/01/26	19	Male	Action	Both	1:30 to 2:00	English	OTT	Above 3	Do not prefer	Do not prefer
15	2022/01/26	19	Male	Thriller	Both	2:00 to 2:30	Regional Languages(Tamil,Telugu,Malayalam,Kannada Etc)	Theatres	Do not prefer rating	Do not prefer	Do not prefer
16	2022/01/26	19	Male	Drama	Both	1:30 to 2:00	Hindi	Theatres	Above 4	Do not prefer	Public talk
17	2022/01/26	19	Male	Mystery	Both	2:00 to 2:30	English	OTT	Do not prefer rating	All of the above	Based on Ratings
18	2022/01/26	19	Male	Thriller	Both	2:00 to 2:30	Hindi	OTT	Do not prefer rating	All of the above	Public talk
19	2022/01/26	19	Male	Comedy	Entertainment factors	2:00 to 2:30	Regional Languages(Tamil,Telugu,Malayalam,Kannada Etc)	Theatres	Above 4	Favourite Director	Public talk
20	2022/01/26	69	Male	Action	Message oriented	Less than 1:00 Hr	English	Theatres	4.5 To 5	Favourite Hero	Based on Ratings
21	2022/01/26	68	Male	Comedy	Both	Less than 1:00 Hr	Hindi	During telecasting in T.V	4.5 To 5	Favourite Heroine	Budget of the movie
22	2022/01/26	83	Female	Action	Entertainment factors	1:30 to 2:00	English	OTT	Above 3	Favourite Hero	Public talk
23	2022/01/26	19	Male	Thriller	Entertainment factors	2:00 to 2:30	English	OTT	Do not prefer rating	Do not prefer	Public talk
24	2022/01/26	19	Female	Thriller	Entertainment factors	2:00 to 2:30	Hindi	OTT	Do not prefer rating	Do not prefer	Do not prefer
25	2022/01/26	20	Female	Comedy	Both	1:30 to 2:00	English	OTT	Above 4	Do not prefer	Based on Ratings
26	2022/01/26	20	Male	Mystery	Both	2:00 to 2:30	Hindi	OTT	Above 4	Favourite Director	Based on Ratings
27	2022/01/26	19	Male	Thriller	Both	2:00 to 2:30	Regional Languages(Tamil,Telugu,Malayalam,Kannada Etc)	Theatres	Do not prefer rating	Do not prefer	Do not prefer
28	2022/01/26	18	Female	Mystery	Message oriented	2:00 to 2:30	Hindi	Theatres	Do not prefer rating	All of the above	Public talk
29	2022/01/26	20	Male	Drama	Entertainment factors	1:30 to 2:00	English	OTT	Above 3	Do not prefer	Do not prefer
30	2022/01/26	19	Male	Action	Both	2:00 to 2:30	English	OTT	Above 4	Favourite Director	Based on Ratings
31	2022/01/26	19	Male	Comedy	Both	2:00 to 2:30	Regional Languages(Tamil,Telugu,Malayalam,Kannada Etc)	OTT	Do not prefer rating	Do not prefer	Do not prefer
32	2022/01/26	20	Male	Action	Both	2:00 to 2:30	English	During telecasting in T.V	Above 4	All of the above	Public talk
33	2022/01/26	18	Male	Comedy	Both	2:00 to 2:30	English	OTT	Do not prefer rating	Do not prefer	Public talk
34	2022/01/26	19	Male	Horror	Both	2:00 to 2:30	English	OTT	Above 4	All of the above	Do not prefer

Fig: Dataset – Survey

Apart from the survey data, we are also considering another dataset that contains movie names and other basic information about them. We will award each movie in this dataset with a score and them according to the final score, we will recommend top 3 movies.



	A	B	C	D	E	F	G
1	Name	Genre	Purpose	Length	Language	Platform	Rating
2	IT	Horror	Entertainment factors	2:00 to 2:30	English	Theatres	Above 3
3	IT 2	Horror	Entertainment factors	2:00 to 2:30	English	OTT	Above 2
4	Doctor Strange	Fantasy	Entertainment factors	2:00 to 2:30	English	Theatres	Above 3
5	Iron Man	Action	Entertainment factors	2:00 to 2:30	English	Theatres	Above 4
6	Avengers	Action	Entertainment factors	2:30 to 3:00	English	Theatres	Above 4
7	Avengers 2	Action	Entertainment factors	2:30 to 3:00	English	Theatres	Above 4
8	Guardians of the Galaxy	Action	Entertainment factors	2:00 to 2:30	English	Theatres	Above 4
9	Prometheus	Mystery	Entertainment factors	2:00 to 2:30	English	Theatres	Above 3
10	Split	Horror	Entertainment factors	2:00 to 2:30	English	Theatres	Above 3
11	Mindhorn	Comedy	Entertainment factors	1:00 to 1:30	English	Theatres	Above 3
12	Titanic	Thriller	Both	2:00 to 2:30	English	Theatres	Above 4
13	3 idiots	Comedy	Both	2:00 to 2:30	Hindi	Theatres	Above 4
14	After Earth	Action	Entertainment factors	1:30 to 2:00	English	Theatres	Above 3
15	Euphoria	Mystery	Entertainment factors	2:00 to 2:30	English	OTT	Above 4
16	Conjuring	Horror	Entertainment factors	2:00 to 2:30	English	Theatres	Above 3
17	Conjuring 2	Horror	Entertainment factors	1:30 to 2:00	English	Theatres	Above 2
18	Dangal	Thriller	Both	2:30 to 3:00	Hindi	Theatres	Above 4
19	Harry Potter	Fantasy	Entertainment factors	2:30 to 3:00	English	Theatres	Above 4
20	Harry Potter 2	Fantasy	Entertainment factors	2:30 to 3:00	English	Theatres	Above 4
21	Conjuring 3	Horror	Entertainment factors	2:30 to 3:00	English	Theatres	Above 3
22	Jumanji	Fantasy	Entertainment factors	2:00 to 2:30	English	Theatres	Above 3
23	Jumanji 2	Action	Entertainment factors	2:30 to 3:00	English	Theatres	Above 4
24	dabangg	Action	Entertainment factors	1:30 to 2:00	Hindi	Theatres	Above 3
25	dabangg	Action	Entertainment factors	2:30 to 3:00	Hindi	OTT	Above 2
26	Sultan	Fantasy	Entertainment factors	2:00 to 2:30	Hindi	Theatres	Above 4
27	Shivaji the boss	Action	Entertainment factors	2:00 to 2:30	Regional Languages(Tamil,Telugu,Malayalam,Kannada Etc)	OTT	Above 3
28	Gold	Thriller	Both	1:30 to 2:00	Regional Languages(Tamil,Telugu,Malayalam,Kannada Etc)	Theatres	Above 4
29	Baby	Action	Entertainment factors	1:30 to 2:00	Regional Languages(Tamil,Telugu,Malayalam,Kannada Etc)	Theatres	Above 2
30	Passengers	Thriller	Entertainment factors	2:00 to 2:30	English	Theatres	Above 4
31	Moana	Fantasy	Entertainment factors	1:30 to 2:00	English	Theatres	Above 4
32	Jojo rabbit	Fantasy	Entertainment factors	2:30 to 3:00	English	Theatres	Above 4
33	Toy Story	Comedy	Entertainment factors	2:30 to 3:00	English	OTT	Above 4
34	Toy Story 2	Comedy	Entertainment factors	2:00 to 2:30	English	Theatres	Above 3

Fig: Dataset – Movies

4.2. Data Pre-processing:

The data obtained through survey include age, gender, genre they prefer, what do they enjoy in a movie, length of the movie they will go for, in which language do they prefer to watch movies in, on which platform they watch movies, what is basic rating a movie should have, why do they watch a movie, etc., a total of 2079 data. Usually, the amount of data on the network is huge, and there are too many data sources. It is easy to produce a lot of unreliable data. Low quality data will lead to the unreliability of data analysis, and there is a big gap with the actual data. Therefore, it is necessary to analyse the data and collect the data for pre-processing.

There were some unnecessary columns such as timestamp, that does not add any value to the data. So, we have removed such columns. We checked for inconsistent and noisy data, missing values, duplicate records.

4.3. Map-Reduce Concept:

We have used map-reduce concept at two places – calculating frequency of each class of a categorical variable and calculating final scores for movies.

1. Calculating frequency of each class of a categorical variable:
   From the dataset we can see that people have chosen various options that showcases their preference. The class that is more often chosen i.e., the one with more frequency is the most preferred. So, we decided to calculate the frequency of each class and then sort them to understand what people prefer. According to the ranks of each class we have awarded them with a number, such that the class with highest frequency, gets the highest number.

We have implemented this using map-reduce concept:

- Map: We have mapped the class as the key with 1 as the value. <Class$_i$, 1>
- Combine / Sort: We have combined the values with same key i.e., class.
- Reduce: For the reduce function, we have calculated the count for each class.

After this we have sorted the result.

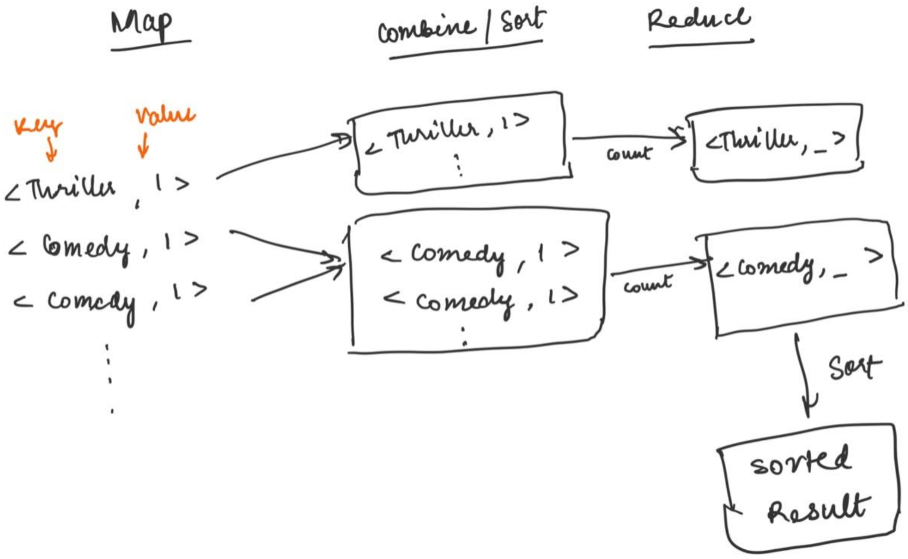


Fig: Map-reduce flowchart for calculating frequency

2. Calculating final scores for movies:
   From above we got the information about the preference and awarded each class with a number. We have then awarded each movie (present in movies dataset) with a score according to the class for each attribute. We have then calculated final score. We have implemented this using map-reduce concept:
   - Map: We have mapped the name of the movie as the key with the score of each attribute as the value. <Movie$_i$, Attribute$_j$>
   - Combine / Sort: We have combined the values with same key i.e., name of the movie.
   - Reduce: For the reduce function, we have calculated the sum of all the values for each movie.

   After this we have sorted the result and recommended the top 3 movies.

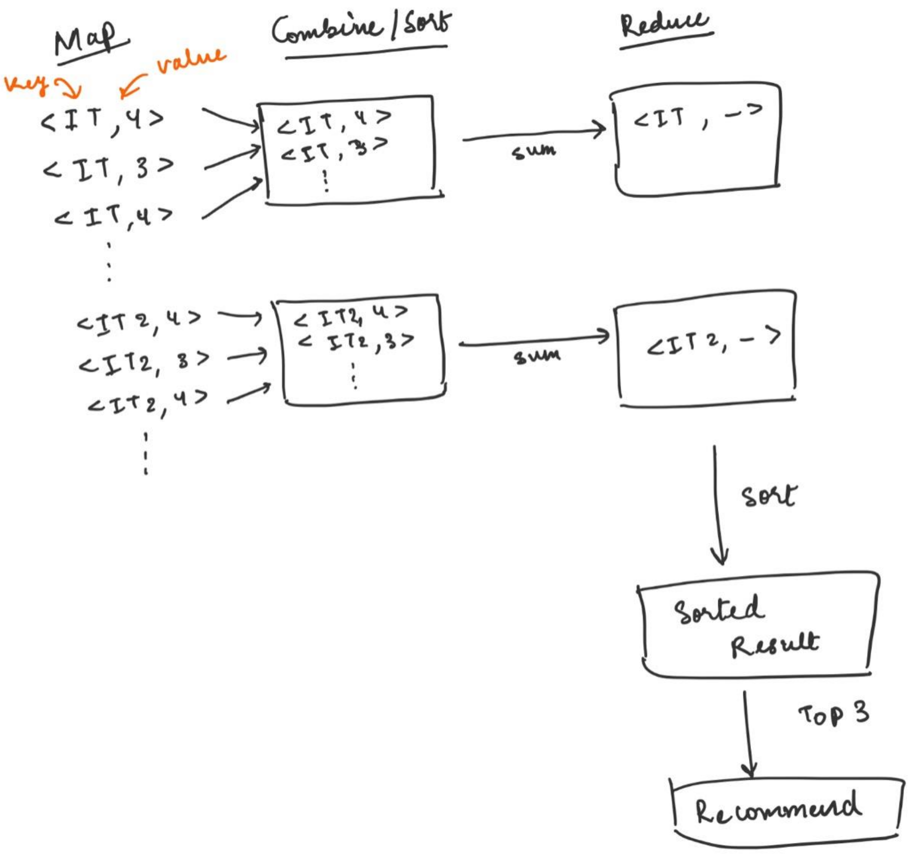


Fig: Map-reduce flowchart for calculating final score

Code link: https://drive.google.com/file/d/1VwCK7crfPHznQb9q1M8fFlOzNoxmL7Pd/view?usp=sharing

5. Conclusion and Future Enhancement

If we are given a set of films, using our method we can find out the movie which will be the most preferred out of the rest of them. From the data we have collected using google forms, each option in the question is given a number according to the preferences. For e.g., in genre Action is given the number 8 because it is the most chosen option out of the rest and other are also given numbers 7 and below according to the data. This is done for all the questions. So, when a movie is given along with its attributes, we can compare its attributes with the algorithm and get the number according to the match with data already processed through the forms. Now after all the attributes are compared with the data the sum of all the numbers is considered. This number tells how much the movie is preferred. Larger the number larger will be the movie preferred. So, when we are given set of movies, we calculate the number of all the movies and the movie with the highest number is selected and is the most preferred out of the rest of them. So, if u want to know about which movie is going to be preferred or which movie will be chosen by most of the people, using our method you can find that. Data collected had only 2079 records, which can be increased to get more information and improve our recommendation system. Relations between the variables can be analysed to find more insights and get a better result.

6. References

1. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3473148
2. https://ejmcm.com/pdf_7583_35d794dd02227a2ca49548f57c76a2c5.html
3. https://www.ijcaonline.org/archives/volume178/number47/chakraborty-2019-ijca-919415.pdf
4. https://www.ijert.org/predicting-the-conceptual-appeal-of-movies-using-data-analytics
5. http://paper.ijcsns.org/07_book/201608/20160820.pdf
6. https://www.hindawi.com/journals/complexity/2021/9947832/