CS432/532: Final Project Report

**Project Title: IMDB Data Analysis**

**Team Member(s): Rohan Ravindra Saraf and Satrio Baskoro Yudhoatmojo**

1. PROBLEM

IMDb is the world most popular sites for movies, TV and celebrity content which has been around since 1990 [1]. There are several datasets already curated by people that we can use for this project. We proposed to do an explorative analysis on the curated IMDb dataset. Based on our naïve search of the IMDb curated dataset, we found that two IMDb datasets [2, 3]. We decided to use the dataset from [3] for our project because it consists more attributes and the number data is larger than [2].

We propose to analyze the following problems. First, we analyze the top 10 movies genre on each year. Second, we explore the correlation between movie rating and movie revenue. Third, derive the conclusion about the profit-loss for the movie based on revenue. Also we worked to study the increasing and decreasing trends in the genres of the movies.

1. SOFTWARE DESIGN AND IMPLEMENTATION

Following are the details about the Software design, NoSQL – Database, and the tools that we have used in the developing the Project – 3.

1. *Software Design and NoSQL-Database and Tools Used*

We propose to store the curated dataset in MongoDB DBMS. We would preprocess the curated dataset in a way so that it is a good fit to store it in MongoDB.

In the application layer, we will use Python programming language for performing the analysis stated in Section I. Python programming language will be used to retrieve data from MongoDB and transform the data into the needed format for the analysis.

After retrieving the data using the python, we have used the HTML, CSS and JavaScript to represent the data in user friendly manner with the proper GUI. Also the top layer of the HTML, gives user the provision to give the input to the application by either selecting year, selecting genre, sliding the lower and upper bound of the IMDB Score etc.

To wrap the whole project, we will wrap the project in the form of web-based application. We may use micro-framework Flask as the scaffolding for developing this web-based application.

Also on top of all this we have used Microsoft Excel to clean the data from the Null and invalid values. We have done following things to clean and purify the data:

* For the Null values in the column with Number as Data type, we have filled it with Average of the column.
* For the Columns of Revenue, we have generated the random number between the specific ranges to fill out null values.
* Use of Excel formulas to trim and remove the unnecessary blanks spaces in between the words and at the end.
* For null values in the columns with the string datatype we have replace it with the dummy string.

1. *Data Import Process in the MongoDB Database*

As mentioned earlier, we have used the dataset [3], so we have imported it using a following command into the MongoDB Database.

**Command:**

*mongoimport --type csv -d (DB Name) -c (Collection Name) --headerline --drop (Filename/path to File)*

Here in our case we have given “**movies**” as the database name and “**movies**” as the collection name.

1. *Libraries for Animation and Chart/Graphs*

We have used the libraries of the animation and displaying the chart/graphs.

1. Animation: Wow.min.js and css 🡪 we need give the respective class to the HTML div and also give the time ins

‘ms’ for the animation and include the script in the page.

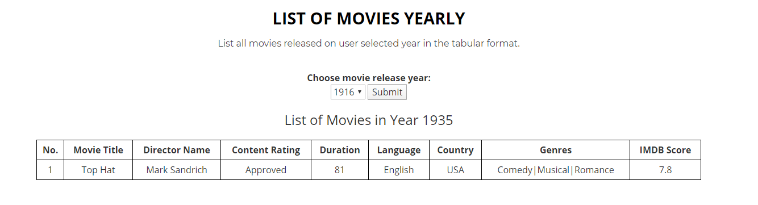
1. Graphs: Chart.min.js and css 🡪 we pass the finalized data to the library and library help us to generated the graphs.
2. *Supported Queries*

All the operations are the Selection Operations to get the details from the Project. We have some of the basic queries along with the sophisticated queries which mainly use to derive the conclusion using either a table or a pictorial graph.

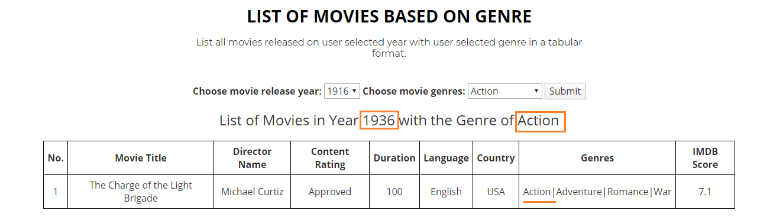
The following are supported queries and functionalities that we have implemented in the Project:

1. **List of movies yearly (basic query)**

This query help us to list the all movies for the selected year in the tabular format. We take the “**year**” as an input from the user via a drop down menu. After he clicks submit, we fire a select query in the MongoDB Database to fetch all the movies released in that particular year.

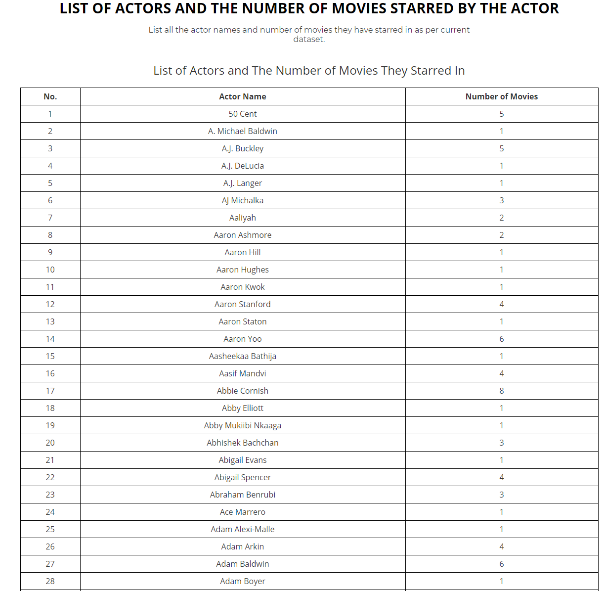


1. **List of movies based on genre (basic query)**

This query help us to list the all movies for the selected year and selected genre in the tabular format. We take the “**year**” and “**genre**” as an input from the user via a drop down menu. After he clicks submit, we fire a select query in the MongoDB database to fetch all the movies released in that particular year and search movies in that year based on genre.

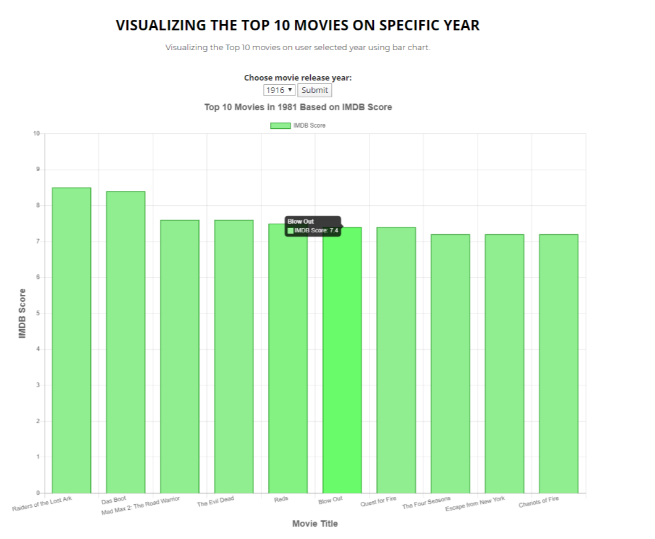
1. **List of actors and number of movies starred by the actor (basic query)**

This query help us to list the all actors and number of movies they starred in the tabular format. Here we fire a select query in the MongoDB database to fetch all the actors from the dataset, we have 3 columns for the name of actors as there are multiple actors in a single movie. Firstly we get all the actors and then we count the number of movies they starred in to get the actual count.



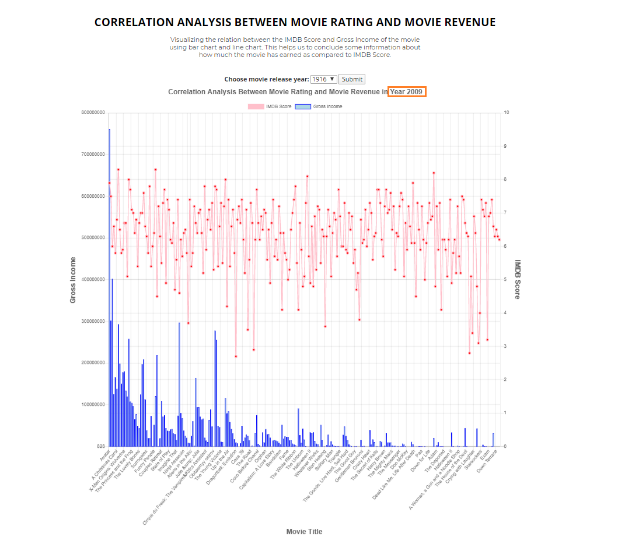
1. **Visualizing the top 10 movie genre on each year (sophisticated query)**

This query help us to visualize the top 10 movies for the selected year in the graphical format. We take the “year” as an input from the user via a drop down menu. After he clicks submit, we fire a select query in the MongoDB database to fetch all the movies released in that particular year and select the only top 10 of them and give a graph in a user friendly format. We select the top 10 movies based on the IMDB score that we have for the movie in the dataset.



1. **Correlation analysis between movie rating and movie revenue (sophisticated query).**

This query help us to conclude or derive a relation for movie about how much has earned as compared to the IMDB rating given by the people. It happens that the movie has not collected much revenue but based on the story and script makes the movie popular between the compelling them to give high IMDB Score. Vice Versa is also possible that movie collected a lot revenue to popularity of its actors but the actual story, script and other things are so-so, normal so it has low IMDB Score. Here we take the “year” as the input from the year so generate the scenario for that particular year.



* + Visualizing the top 10 movie genre on each year
  + Correlation analysis between movie rating and movie revenue
  + Number of profit-loss movie revenue in each year
  + Movies to watch based on ratings (basic query)
  + Actor’s popularity on each movie
  + Relationship network analysis between actors

REFERENCES

1. IMDB. “What Is IMDb?” *IMDb*, IMDb.com, help.imdb.com/article/imdb/general-information/what-is- imdb/G836CY29Z4SGNMK5?ref\_=helpart\_nav\_1#.
2. Promptcloud. “IMDB Data from 2006 to 2016 - Dataset by Promptcloud.” *Data.world*, 26 June 2017, [https://data.world/promptcloud/imdb-data- from-2006-to-2016.](https://data.world/promptcloud/imdb-data-from-2006-to-2016)
3. Yueming. “IMDB 5000 Movie Dataset.” *Kaggle*, 16 Dec. 2017, [www.kaggle.com/carolzhangdc/imdb-5000-movie-dataset.](http://www.kaggle.com/carolzhangdc/imdb-5000-movie-dataset)