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 proposed to develop a light-weight web-based application. Our project will mainly use Python programming language, and Flask micro-framework as the scaffolding for developing the web-based application. We will also use MongoDB as the DBMS for managing and storing the curated IMDb dataset.

1. *Supported Queries*

The following are supported queries and functionalities that we are planning to implement:

* + List of movies yearly (basic query)
  + List of movies based on genre (basic query)
  + List of actors and the number of movies starred by the actor (basic query)
  + 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)