CS432/532: NoSQL Project 1-Page Proposal

Project Title: IMDB Data Analysis

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

I. 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, we explore the distribution of the number of votes of movie rating. Fourth, we explore the relationship between actors.

II. Software Design and Implementation

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. Python libraries such as NumPy, Pandas, and NetworkX maybe use to calculate some of the analysis we need. Python visualization libraries such as Matplotlib and Seaborn are used. The features in this project are going to be developed by us with the help of several libraries that we have mentioned.

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.

A. Software Design and NoSQL-Databse 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.

B. Supported Queries/Functionalities that we Plan to Implement

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.

[3] Yueming. "IMDB 5000 Movie Dataset." *Kaggle*, 16 Dec. 2017, www.kaggle.com/carolzhangdc/imdb-5000-movie-dataset.