CSCI-620 Data Management with the IMDb Dataset

[Developing tools to interact with the IMDb dataset]

Aishwarya Rao
ar2711@rit.edu

Apurav Khare
ak2816@rit.edu

Martin Qian
jq3513@rit.edu

Prateek Kalasannavar
pk6685@rit.edu

ABSTRACT

This project aims to explore a dataset by understanding it, modeling it to a normalized relational schema so that it can be stored and retrieved from a relational database management system. The project also focuses on developing an interface that allows fast and easy access to the dataset by abstracting complex query scenarios, like search by specific parameters within and across tables, and aggregate queries.

1. PROJECT STATUS

As established in Phase 0 and Phase 1, the deliverable of Phase 2 in Data Management includes three things. One, a document specifying different query scenarios on the IMDb dataset, their equivalent queries and sample outputs. Two, a script that handles requests from the user interface and maps these requests to the relevant query and responds with appropriate results. Finally, a user interface that allows the user to pick scenarios, specify various filters and view the results in a readable format. The following sections cover the technology and methodology used to implement these tasks, query scenarios, and screenshots of the output.

1.1 Technology Used

We used a web interface that allows users to interact with our database through a browser. The front end is built with ReactJS and BootStrap. It consists of tabs for every query scenario and allows entry of appropriate filters as seen in 1.1. On submitting a customized query, the request is handled by a Python script which uses the the user entered values to build a query and send it to the database. It collects the response and sends it back to ReactJS for the user to view. The database is stored using MySQL Server and the queries are built accordingly.

1.2 Requirements and running the code

Our development phases uses the following technologies,

- Python 3.6 with Flask
- ReactJS
- MySQL Server version -

`frontend.png`

Due to memory and storage constraints, the database size was cut down to only those entries after the year 2005.

1.3 Query Scenarios

- List the names of alive actors whose name starts with a given keyword (such as “Phi”)and did not participate in any movie in a given year (such as 2014)
- List the names of alive producers who have produced more than a given number (such as 50) of talk shows in a given year (such as 2017) and whose name contains a given keyword (such as “Gill”)
- List the average runtime for movies whose original title contain a given keyword such as (“star”)and were written by somebody who is still alive
- List the names of alive producers with the greatest number of long-run movies produced (runtime greater than 120min)
- List the unique name pairs of actors who have acted together in more than a given number (such as 2) movies

and sort them by average movie rating (of those they acted together

- List the actors that have worked in x movies (say 10) from one genre (say horror)
- List the actors and directors that have worked together atleast X times.
- List the highest rated episodes in year x sorted in high to low manner.
- Writer, director that have worked together in atleast x different TV Shows
- List the most popular TV shows between the years x and y that are still running.