COMP 3380 Final Project

Souvik Ray and Rahul Kumar

Summary

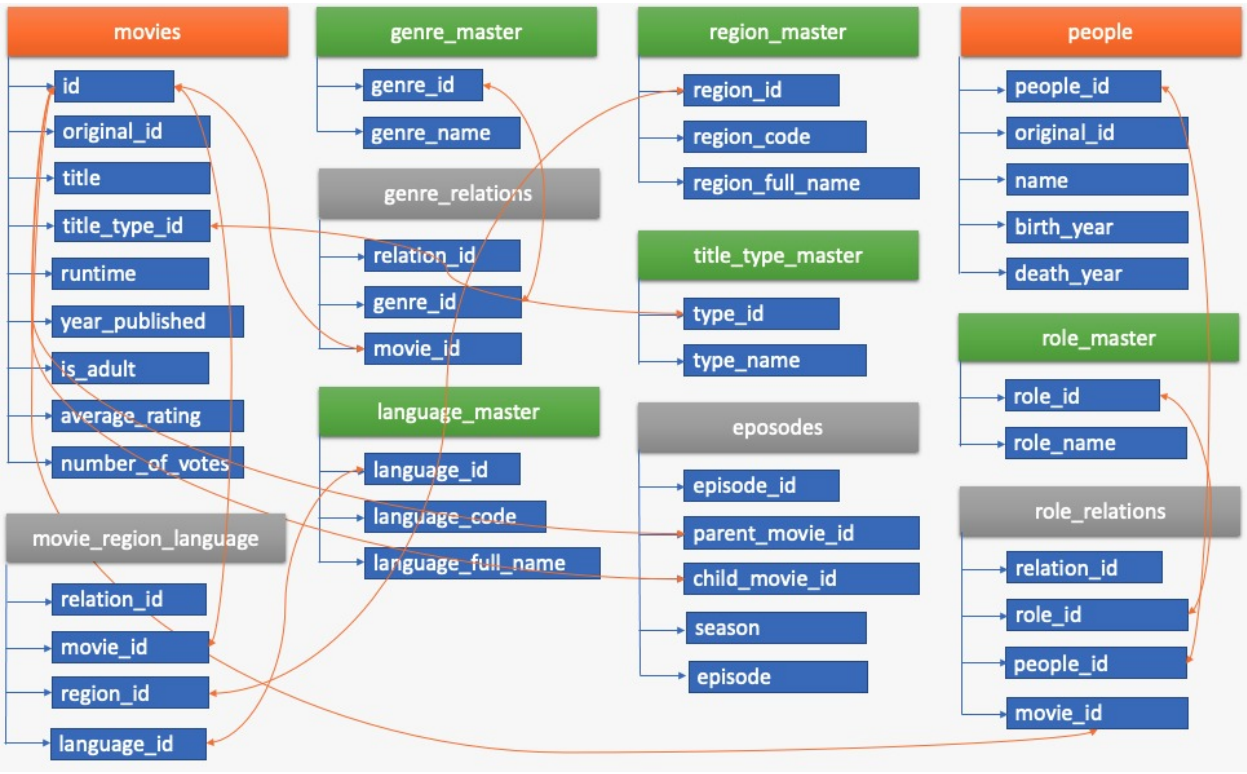
1. This IMDB database was chosen to allow us to access various information which were of different types and could be divided into different tables for quick access of data while having a significant amount of rows to work with.

This database contains 7 different tsv files. The total file size exceeds 5.5 GB and has approximately 12 million data entries in each file.

This link is different than the one previously chosen because the dataset available at the earlier link was removed before we had the opportunity to finish the project.

The link to the dataset - https://www.imdb.com/interfaces/

2. The image below gives an overview of how the database is structured.

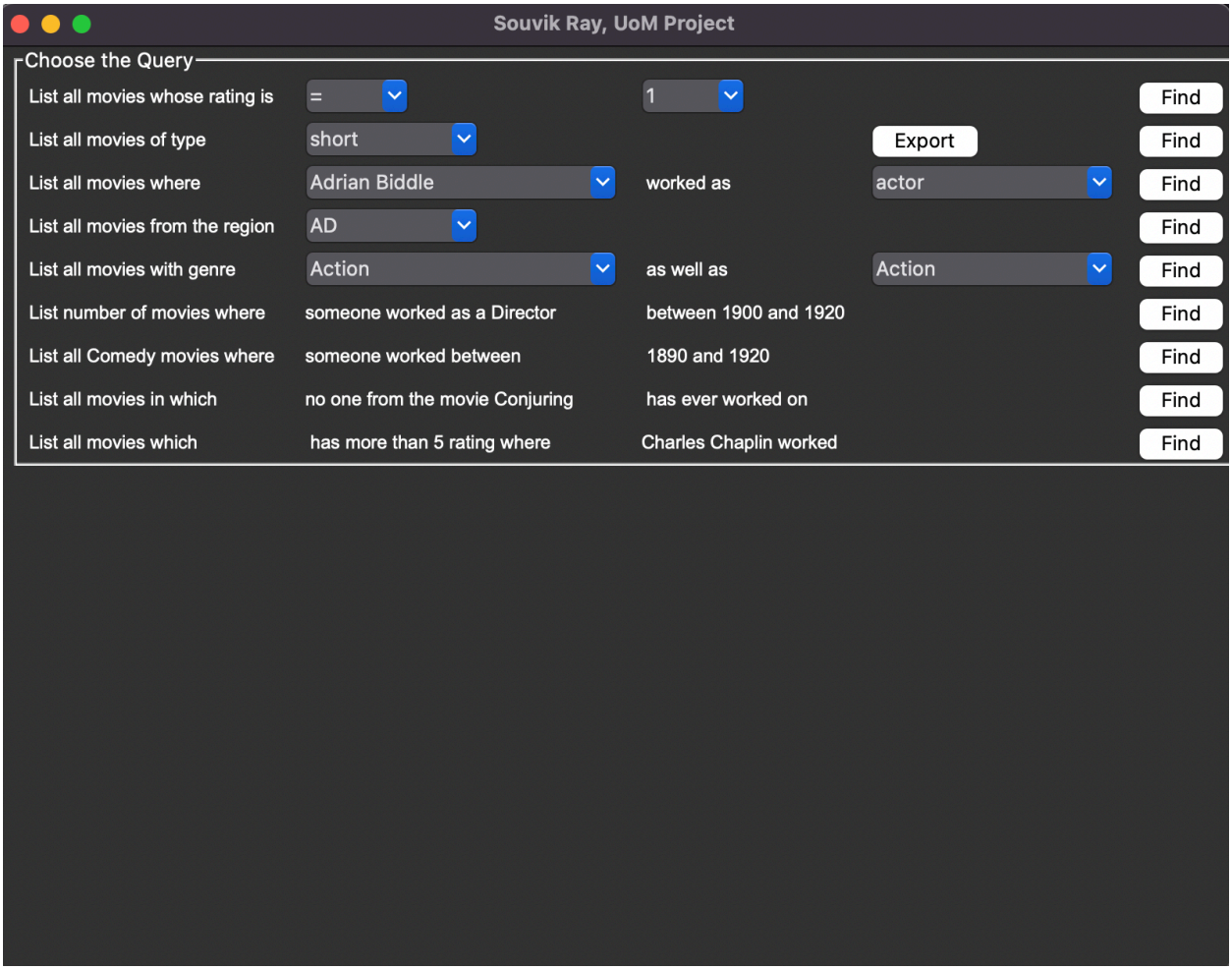


Movies and People contain most of the information and have been marked in orange. 4 master data tables have been created to give each language, genre, region, title and role a specific id which allows us to reference and parse the data later.

3. This database was created in DB Browser fro SQLite. It was populated using the above IMDB datasets and the process was automated using python. The front end UI has also been created using python. To populate the database please use the **upload_data.py** and uncomment the functions as required at the bottom of the file.

To start the front end UI please use the **data_analysis.py** file. It might require the user to install several modules like **PyQt5** and **tkinter**. Attached below is an image of how the UI looks like and how it displays results.

List all movie of type (x) can also be exported into a csv file format which automatically creates a new document in the current folder/directory.



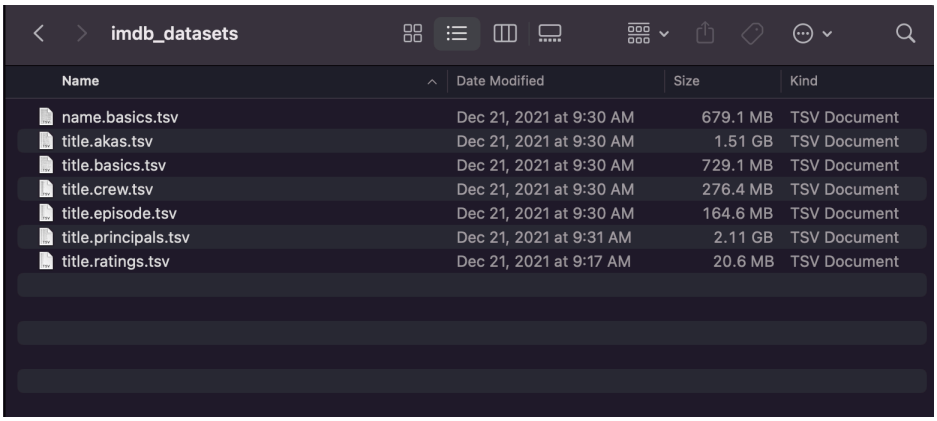
4. Queries —

    1. List all movies whose rating is greater than, less than or equal to the selected rating.
    2. List all movies of that type.
    3. List all movies where person x worked in y role.
    4. List all movies from the chosen region.
    5. List all movies in which is a part of both genres.
    6. List all movies someone worked as a director between certain years.
    7. List all comedy movies someone worked between certain years.
    8. List all movies where no one from the movie conjuring has ever worked on.
    9. List all the movies which has a 5 star rating where Charlie Chaplin has also worked.
    10. Count of all the movies of all genres in a given range of years.

5. Links -

Please find below some important links and information -

1. The entire project folder link - https://1drv.ms/u/s!AhVTjYyRL-ywpzJS51po3itZACxW?e=RnlBG3
2. Dataset - https://www.imdb.com/interfaces/
3. The datasets must have the same name as they were when they are downloaded from IMDB website. They must be kept in a single folder called ‘imdb_datasets’.



4. Due to file size issues we

decided to not include the datasets in the submitted folder. For every program to work - please download and keep all datasets in the supplied folder named - ‘imdb_datasets’.