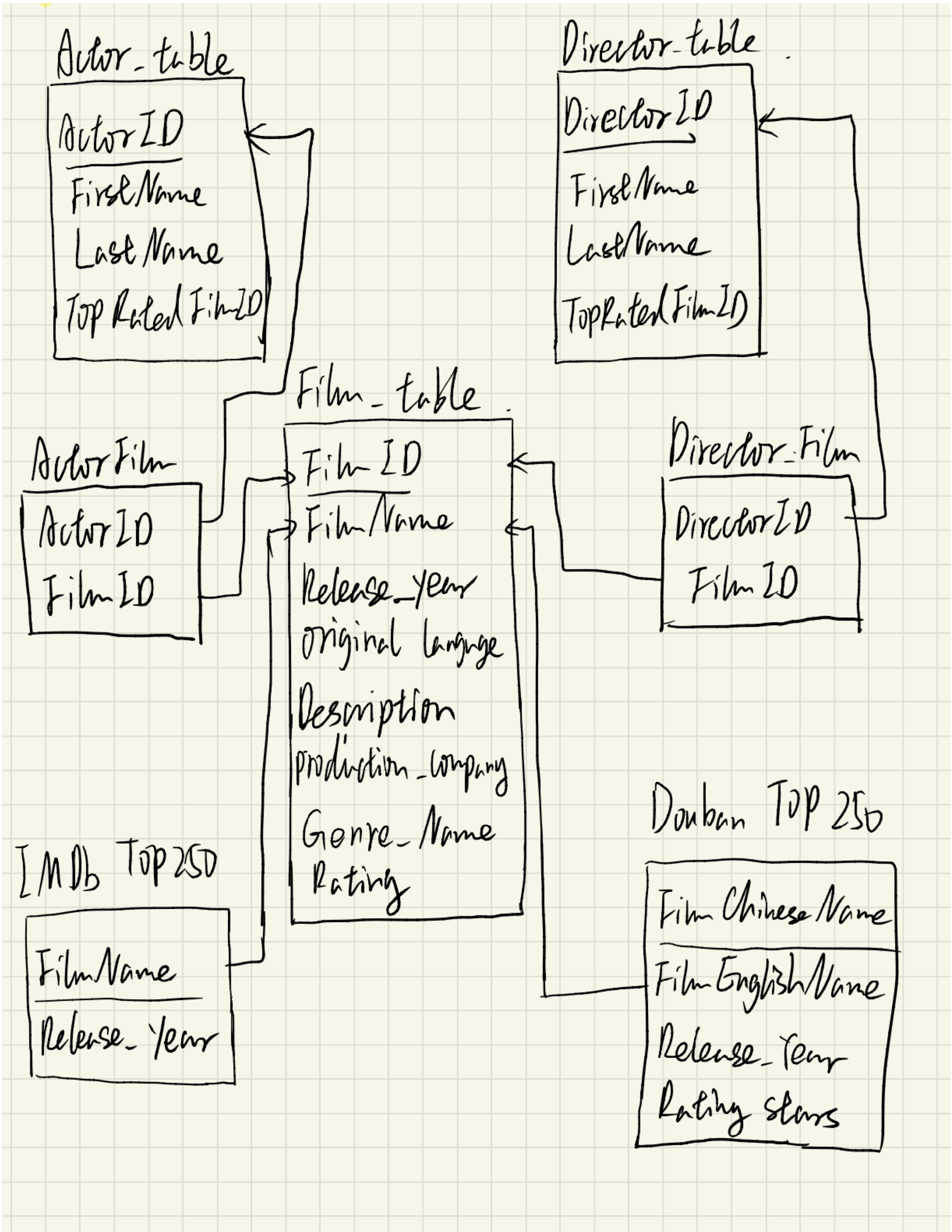
Author: Yifan Le
Data: 5.22.2019

Film database diagram:



Author: Yifan Le
Data: 5.22.2019

Some necessary explanation:

1. ActorFilm table and DirectorFilm table are relationships while others are objects.
2. In this diagram, data type of columns is clear. All "ID" columns' data type is int. Columns related with name, language, description, genre have data type of text. All "rating" columns are float. Only the Release_Year comes with epoch time.
3. File_table will be established by data from IMDb. I guess the scale will be like 20K rows of films.
4. IMDb Top 250 table does not contain "Rating" because FIle_table already covers it.
5. Douban Top 250 table works in Chinese so I create both FilmChineseName column and FilmEnglishName. Note that part of film in Douban Top 250 might not be covered in IMDb but we could still use it to compare taste of Users of IMDb and Douban, which somehow indicates the similarity or difference between audiences in US and those in China.
6. As far as I know, Douban uses different rating system with range from zero star to five stars. Thus I could not directly compare it with 10-point system. Ranking still works.

Data Source:

1. https://www.imdb.com/
2. https://www.imdb.com/chart/top
3. https://movie.douban.com/top250

Data Source Explanation: I believe I don't need to explain how IMDB and IMDB Top 250 works and it should be noted that Kaggle has highly related dataset and IMDB itself also provides with python library (https://imdbpy.sourceforge.io/) to make life easier. I preferred to use such API to fetch what I need rather than downloading from Kaggle. In addition, I would like to scrape data from douban.com, which is a similar website for Chinese users to rate and review films, books, etc. It seems that anti-web crawlers is applied but I think I can scrape some useful information from a static webpage, because what I actually need is only the movie name and rating in TOP 250.

Questions:

1. List five directors who has more works in IMDB Top 250, Douban Top 250.
2. List most popular films and their production companies after 2010 (ranked in Top 50 by both box office and amount of review/rating).
3. List Top 3 genres in different decades (based on rating of films).
4. List full name of Top 10 actors or actresses with their works (based on rating of films).