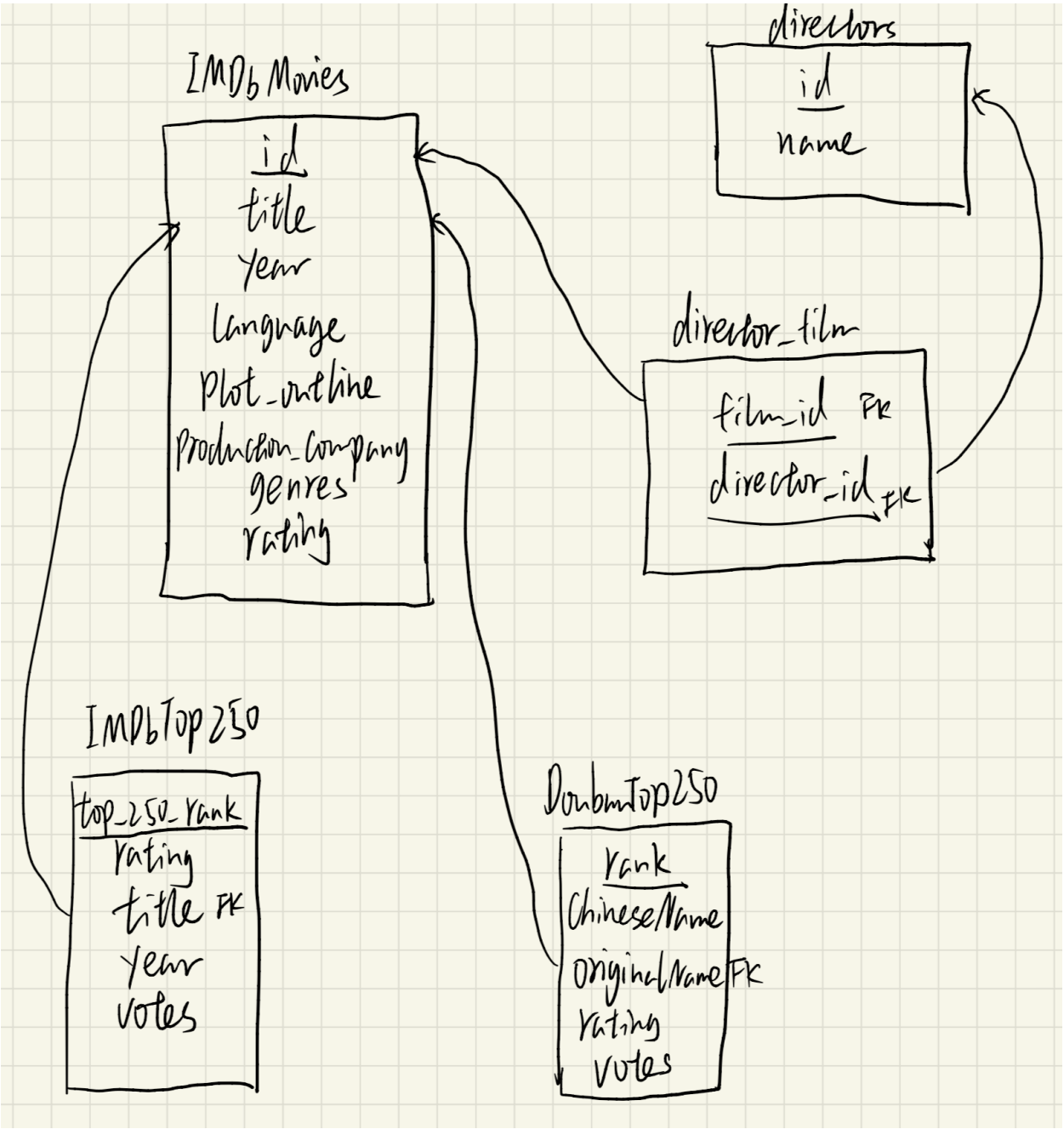
Film database diagram:



Some necessary explanation:

1. DirectorFilm table are relationships while others are objects.
2. In this diagram, data types of columns is clear. Columns of id, rating, votes, year contain data in int. Columns related with name, language, description, genre have data type of text.

3. Douban Top 250 website works in Chinese so I create both ChineseName column and OriginalName, which means if it is a Chinese film, the name will be the same in two columns. Note that part of film in Douban Top 250 might not be covered in IMDb but we could still use it to compare the taste of Users of IMDb and Douban, which somehow indicates the similarity or difference between audiences in the US and those in China.

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 used such API to fetch what I need rather than downloading from Kaggle. In addition, I scraped 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 scraped some useful information from the static webpage https://movie.douban.com/top250, because what I actually need is only the movie name and rating in TOP 250.

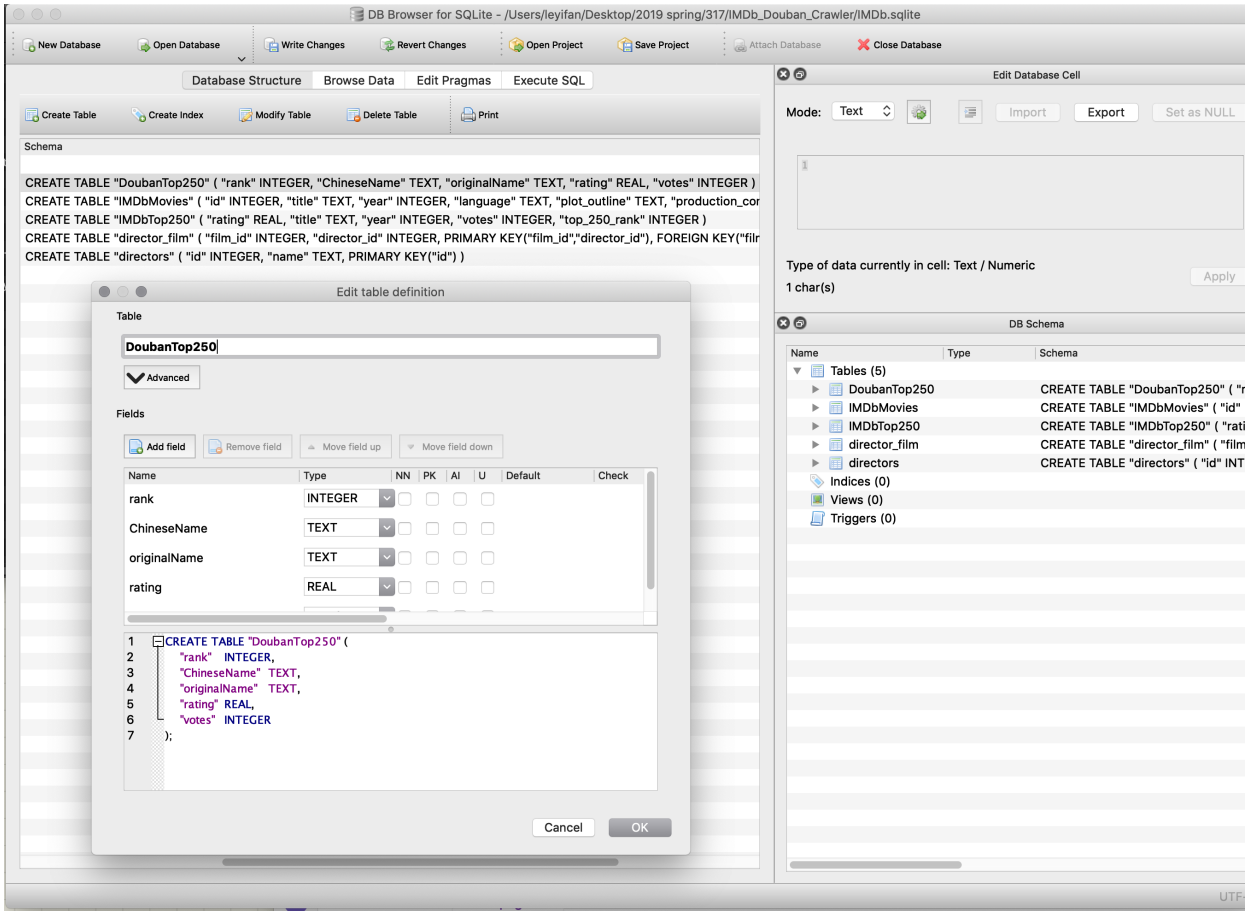
The part above line is the updated version for part1, and explanation for part2 as well as sql problems lists as follows.

First, I used JupyterNotebook in Anaconda environment (python3.6). All files in .ipynb is used to create corresponding .csv file except csv2sqlite.ipynb which write csv as table into IMDb.sqlite with help of pandas and sqlite3. "new_IMDb.csv" is different from "IMDb.csv" on columns they contain but they could both be created from "importIMDb.ipynb". So does "new_IMDbTop250".

How to create the IMDb.sqlite from scratch:

1. Use imdb.IMDb().get_top250_movies() to create IMDbTop250.csv. (See more details in importIMDbTop250.ipynb)
2. Use beautifulsoup to scrape data from https://movie.douban.com/top250 into top250.xlsx then convert it to DoubanTop250.csv. (See more details in importDoubanTop250.ipynb)
3. First, Get IMDbID for films in IMDb database from https://grouplens.org/datasets/movielens/. Again, use imdb.IMDb().get_movie(movieid) to fetch data of files. Note, as shown in links.csv, there is at least 50K+ films in IMDb database but it really takes time to fetch them one by one. Thus I downloaded about 30K films into "IMDb.csv". (See more details in importDoubanTop250.ipynb)

4. Extract directors in “IMDb.csv” to “directors.csv”. (See more details in importDirector.ipynb)
5. Create mappings of director and film in “director_film.csv”. (See more details in importDirector.ipynb)
6. Drop some columns in “IMDb.csv” and “IMDBTop250.csv” and get “new_” version.
7. Write those five .csv files into IMDb.sqlite. (See more details in csv2sqlite.ipynb)
8. Open IMDb.sqlite in DB browser for SQLite. Set primary key, foreign key of all tables via its “Modify Table” function. (Edit->Modify Table)
9. According to information in “Database Structure”, we can see whether it is all set.



Github link: https://linco1n3.github.io/IMDb_Douban_Crawler/

I will update README after final exams done :).

Q&A:

1.List filmname and corresponding year both in douban Top 250 and IMDb Top250

```
SELECT IMDbMovies.title, IMDbMovies.year FROM
(DoubanTop250 JOIN IMDbTop250 ON DoubanTop250.originalName =
IMDbTop250.title) f
JOIN IMDbMovies on IMDbMovies.title = f.title
ORDER By IMDbMovies.year
```

Psycho	1960
2001: A Space Odyssey	1968
A Clockwork Orange	1971
The Godfather	1972
One Flew Over the Cuckoo's Nest	1975
Witness for the Prosecution	1982
Once Upon a Time in America	1984
Dead Poets Society	1989
The Silence of the Lambs	1991
Terminator 2: Judgment Day	1991
Schindler's List	1993
The Shawshank Redemption	1994
Forrest Gump	1994
The Lion King	1994
Pulp Fiction	1994
Se7en	1995
Braveheart	1995
Before Sunrise	1995
The Usual Suspects	1995
Trainspotting	1996
12 Angry Men	1997
Good Will Hunting	1997
The Truman Show	1998

Lock, Stock and Two Smoking Barrels	1998
Saving Private Ryan	1998
Psycho	1998
Fight Club	1999
The Matrix	1999
The Sixth Sense	1999
The Green Mile	1999
American Beauty	1999
Memento	2000
Requiem for a Dream	2000
The Lord of the Rings: The Fellowship of the Ring	2001
A Beautiful Mind	2001
Monsters, Inc.	2001
The Lord of the Rings: The Two Towers	2002
The Pianist	2002
Catch Me If You Can	2002
The Lord of the Rings: The Return of the King	2003
Pirates of the Caribbean: The Curse of the Black Pearl	2003
Before Sunset	2004
Hotel Rwanda	2004
V for Vendetta	2005
The Prestige	2006
WALL·E	2008
The Dark Knight	2008
Hachi: A Dog's Tale	2009

3 Idiots	2009
Up	2009
Mary and Max	2009
Inglourious Basterds	2009
Inception	2010
Shutter Island	2010
How to Train Your Dragon	2010
Toy Story 3	2010
The Dark Knight	2011
Harry Potter and the Deathly Hallows: Part 2	2011
The Dark Knight Rises	2012
Django Unchained	2012
Interstellar	2014
The Grand Budapest Hotel	2014
Gone Girl	2014
Whiplash	2014
Inside Out	2015
Room	2015
Spotlight	2015
Mad Max: Fury Road	2015

2. List six directors who has more works in intersection of IMDB Top250 and Douban Top 250

```
SELECT directors.name, count(*) FROM
(DoubanTop250 JOIN IMDbTop250 ON DoubanTop250.originalName =
IMDbTop250.title) f
JOIN IMDbMovies on IMDbMovies.title = f.title
JOIN director_film on IMDbMovies.id = director_film.film_id
JOIN directors on directors.id = director_film.director_id
GROUP by directors.name
ORDER By count(*)  DESC
```

LIMIT 6

Christopher Nolan	6
David Fincher	3
Pete Docter	3
Peter Jackson	3
Quentin Tarantino	3
Steven Spielberg	3

3.List films amount in diffrent decades.
SELECT CAST(round(year/10-0.5) * 10 as INT) as decade, count(*) as film_amout
FROM IMDbMovies GROUP BY round(year/10-0.5)

1890	4
1900	7
1910	98
1920	265
1930	1211
1940	1378
1950	1814
1960	2003
1970	2464
1980	2876
1990	4493
2000	7443
2010	5568

4.List 50 most popular films and their production company after 2010
SELECT title, production_company FROM IMDbMovies
WHERE year > 2009
ORDER by rating DESC
LIMIT 50

Wheels	Loaded Dice Films
Human Planet	British Broadcasting Corporation
The Runner from Ravenshead	Little Crew Studios
Shining Night: A Portrait of Composer Morten Lauridsen	Song Without Borders
Inception	Warner Bros.
A Very Potter Sequel	StarKid Productions
Drishyam	Aashirvad Cinemas
Hladnokrvno	Elegy film
The Phantom of the Opera at the Royal Albert Hall	The Really Useful Theatre Company
Katyar Kaljat Ghusali	Essel Vision
The Jinx: The Life and Deaths of Robert Durst	HBO Documentary Films
Toy Masters	Urban Archipelago Filmed Entertainment
Senna	Universal Pictures
Interstellar	Paramount Pictures
Crystal Lake Memories: The Complete History of Friday the 13th	1428 Films
Butterfly Girl	Lytta Productions
The Men Who Built America	Stephen David Entertainment
A Christmas Carol	BBC Cymru Wales
The Two Escobars	All Rise Films
The Intouchables	Quad Productions
The War You Don't See	Dartmouth Films
Whiplash	Bold Films
Generation War	teamWorx Produktion für Kino und Fernsehen GmbH

Avengers: Infinity War	Marvel Studios
Bey Yaar	CineMan Productions
Once Brothers	ESPN Films
RangiTaranga	Sri Devi Entertainers
The Dark Knight Rises	Warner Bros.
Pilot	Mammoth Screen
Django Unchained	The Weinstein Company
Batman: The Dark Knight Returns, Part 2	Warner Premiere
Louis C.K. Oh My God	Pig Newton
José and Pilar	JumpCut
Lucia	Audience films
Symphony of the Soil	Lily Films
The Salt of the Earth	Decia Films
Most Likely to Succeed	One Potato Productions
The Story of Film: An Odyssey	Hopscotch Films
One Little Pill	Zard Productions
Kaakkaa Muttai	Fox STAR Studios
The Invisible Front	Aspectus Memoria
Elizabeth Ekadashi	Essel Vision Productions
Natarang	Zee Talkies
The Dream Team	National Basketball Association
Soodhu Kavvum	Thirukumaran Entertainment
Papanasam	Rajkumar Theaters Private Limited
Temple Grandin	HBO Films
Toy Story 3	Walt Disney Pictures

Inside Job	Sony Pictures Classics
Foo Fighters: Back and Forth	Allentown Productions