SI 601 Winter 2016 Homework 4 (100 points)

Due at 5:30pm on Wednesday, Feb. 3, 2016

Part 1 (75 points)

The provided 'movie_actors_data.txt' file contains a JSON string on each line. For example, the first line is:

{"rating": 9.3, "genres": ["Crime", "Drama"], "rated": "R", "filming_locations": "Ashland, Ohio, USA", "language": ["English"], "title": "The Shawshank Redemption", "runtime": ["142 min"], "poster": "http://img3.douban.com/lpic/s1311361.jpg", "imdb_url": "http://www.imdb.com/title/tt0111161/", "writers": ["Stephen King", "Frank Darabont"], "imdb_id": "tt0111161", "directors": ["Frank Darabont"], "rating_count": 894012, "actors": ["Tim Robbins", "Morgan Freeman", "Bob Gunton", "William Sadler", "Clancy Brown", "Gil Bellows", "Mark Rolston", "James Whitmore", "Jeffrey DeMunn", "Larry Brandenburg", "Neil Giuntoli", "Brian Libby", "David Proval", "Joseph Ragno", "Jude Ciccolella"], "plot_simple": "Two imprisoned men bond over a number of years, finding solace and eventual redemption through acts of common decency.", "year": 1994, "country": ["USA"], "type": "M", "release_date": 19941014, "also_known_as": ["Die Verurteilten"]}

The fields we are interested in are imdb_id , title , rating, genres, actors, and year. You will parse the JSON strings, and load the data into three tables in SQLite, and then write SQL queries to retrieve the data specified.

You will create three tables:

- The "movie_genre" table, which has two columns: imdb_id and genre. A movie typically has multiple genres, and in this case, there should be one row for each genre. If some movie does not have any genre, ignore that movie.
- The "movies" table, which has four columns: imdb_id, title, year, rating
- The "movie_actor" table, which has two columns imdb_id and actor. A movie typically has multiple actors, and in this case, there should be one row for each actor.

1. (10 points) Parse input file to get needed data for the three tables and load them into appropriate Python data structure.

2. (5 points) Create the movie_genre table and load data into it

3. (5 points) Create the movies table and load data into it

4. (5 points) Create the movie_actor table and load data into it

5. (10 points) Write an SQL query to find top 10 genres with most movies and print out the results

6. (10 points) Write an SQL query to find number of movies broken down by year in chronological order

7. (10 points) Write an SQL query to find all Sci-Fi movies order by decreasing rating, then by decreasing year if ratings are the same.

8. (10 points) Write an SQL query to find the top 10 actors who played in most movies in and after year 2000. In case of ties, sort the rows by actor name.

9. (10 points) Write an SQL query for finding pairs of actors who co-stared in 3 or more movies. The pairs of names must be unique. This means that ‘actor A, actor B’ and ‘actor B, actor A’ are the same pair, so only one of them should appear.

In each pair of actors you print out, the two actors must be ordered alphabetically. The pairs are ordered in decreasing number of movies they co-stared in. In case of ties, the rows are ordered by actors’ names.

You will need to join the movie_actor table with itself to get this data. It is a bit tricky. If you cannot do it with SQL statement, you can also write some Python code that works on the Python data structure that you used to create the movie_actor table. That’ll mean much more lines of code, and if you do it that way, you’ll get 5 points instead of 10 points. You will only get 10 points if you solve it with pure SQL.

When you run your Python code, it should print out EXACTLY such output in your terminal:

```
Top 10 genres:
Genre, Movies
Drama,166
Thriller,66
Crime,60
Adventure,56
Mystery,42
Comedy,41
Action,40
Romance,35
Fantasy,30
War,29

Movies broken down by year:
Year, Movies
1921, 1
1922, 1
1925, 1
1926, 1
1927, 1
1930, 1
1931, 2
1934, 1
1936, 1
1939, 3
1940, 3
1941, 2
1942, 1
1943, 1
1944, 2
1946, 3
1948, 3
1949, 1
1950, 4
1951, 2
1952, 2
1953, 2
1954, 5
1955, 1
1956, 1
1957, 6
1958, 2
1959, 5
1960, 2
1961, 2
1962, 4
1963, 2
1964, 2
1965, 1
1966, 3
```

1967, 2
1968, 3
1969, 2
1971, 1
1972, 2
1973, 3
1974, 2
1975, 5
1976, 3
1977, 2
1978, 1
1979, 5
1980, 3
1981, 2
1982, 3
1983, 1
1984, 3
1985, 2
1986, 3
1987, 3
1988, 5
1989, 1
1990, 1
1991, 3
1992, 2
1993, 4
1994, 5
1995, 8
1996, 2
1997, 3
1998, 5
1999, 5
2000, 5
2001, 5
2002, 5
2003, 9
2004, 5
2005, 3
2006, 4
2007, 7
2008, 7
2009, 7
2010, 7
2011, 6
2012, 8

Sci-Fi movies:
Title, Year, Rating
Inception, 2010, 8.8
Star Wars: Episode V - The Empire Strikes Back, 2008, 8.8
Star Wars, 1977, 8.8
The Matrix, 1999, 8.7
Terminator 2: Judgment Day, 1991, 8.6
WALL·E, 2008, 8.5
Aliens, 1986, 8.5
Back to the Future, 1985, 8.5
Alien, 1979, 8.5
A Clockwork Orange, 1971, 8.5
Eternal Sunshine of the Spotless Mind, 2004, 8.4
Star Wars: Episode VI - Return of the Jedi, 2002, 8.4
2001: A Space Odyssey, 1968, 8.4
Metropolis, 1927, 8.4
Blade Runner, 1982, 8.3
Donnie Darko, 2001, 8.2
The Thing, 1982, 8.2
Twelve Monkeys, 1995, 8.1
The Terminator, 1984, 8.1
Stalker, 1979, 8.1
District 9, 2009, 8.0
Star Trek, 2009, 8.0
The Truman Show, 1998, 8.0
Jurassic Park, 1993, 8.0

In and after year 2000, top 10 actors who played in most movies:
Actor, Movies
Christian Bale, 5
John Ratzenberger, 5
Michael Caine, 5
Leonardo DiCaprio, 4
Morgan Freeman, 4
Orlando Bloom, 4
Billy Boyd, 3
Bob Peterson, 3
Cate Blanchett, 3
Cillian Murphy, 3

Pairs of actors who co-stared in 3 or more movies:
Actor A, Actor B, Co-stared Movies
Christian Bale, Michael Caine, 4
Joe Pesci, Robert De Niro, 4
Al Pacino, John Cazale, 3
Alec Guinness, Anthony Daniels, 3
Alec Guinness, Carrie Fisher, 3
Alec Guinness, David Prowse, 3
Alec Guinness, Harrison Ford, 3
Alec Guinness, Kenny Baker, 3
Alec Guinness, Mark Hamill, 3
Alec Guinness, Peter Mayhew, 3
Anthony Daniels, Carrie Fisher, 3
Anthony Daniels, David Prowse, 3
Anthony Daniels, Harrison Ford, 3
Anthony Daniels, Kenny Baker, 3
Anthony Daniels, Mark Hamill, 3
Anthony Daniels, Peter Mayhew, 3
Benito Stefanelli, Clint Eastwood, 3
Bibi Andersson, Gunnar Björnstrand, 3
Billy Boyd, Cate Blanchett, 3
Billy Boyd, Orlando Bloom, 3
Billy Boyd, Sean Astin, 3
Carrie Fisher, David Prowse, 3
Carrie Fisher, Harrison Ford, 3
Carrie Fisher, Kenny Baker, 3
Carrie Fisher, Mark Hamill, 3
Carrie Fisher, Peter Mayhew, 3
Cate Blanchett, Orlando Bloom, 3
Cate Blanchett, Sean Astin, 3
Charles Chaplin, Hank Mann, 3
Christian Bale, Gary Oldman, 3
Christian Bale, Morgan Freeman, 3
Cillian Murphy, Michael Caine, 3
Daisuke Katô, Takashi Shimura, 3
Daisuke Katô, Toshirô Mifune, 3
David Prowse, Harrison Ford, 3
David Prowse, Kenny Baker, 3
David Prowse, Mark Hamill, 3
David Prowse, Peter Mayhew, 3
Frank Vincent, Joe Pesci, 3
Frank Vincent, Robert De Niro, 3
Gary Oldman, Michael Caine, 3
Gary Oldman, Morgan Freeman, 3
Harrison Ford, Kenny Baker, 3
Harrison Ford, Mark Hamill, 3
Harrison Ford, Peter Mayhew, 3
Kenny Baker, Mark Hamill, 3
Kenny Baker, Peter Mayhew, 3
Mark Hamill, Peter Mayhew, 3
Marlon Brando, Rudy Bond, 3
Michael Caine, Morgan Freeman, 3
Orlando Bloom, Sean Astin, 3
Takashi Shimura, Toshirô Mifune, 3

Part 2 (25 points)

Write a python program named si601_w16_hw4_part2_youruniquename.py that takes two command line arguments: genre and k

The program should print out the top k actors who played roles in most movies in the provided genre.

You should use the sqlite3 database file you created in Part 1.

Some example runs of my program are shown below. Your program should produce the same output when provided with the same command line arguments.

```
wyhs-MacBook-Air:si601_w16_hw4 wyh$ python si601_w16_hw4_part2_solution.py Drama 5
Top 5 actors who played in most Drama movies:
Actor, Drama Movies Played in
Robert De Niro, 10
James Stewart, 6
Al Pacino, 5
Robert Duvall, 5
Charles Chaplin, 4
wyhs-MacBook-Air:si601_w16_hw4 wyh$ python si601_w16_hw4_part2_solution.py Drama 10
Top 10 actors who played in most Drama movies:
Actor, Drama Movies Played in
Robert De Niro, 10
James Stewart, 6
Al Pacino, 5
Robert Duvall, 5
Charles Chaplin, 4
Claude Rains, 4
Diane Keaton, 4
Helena Bonham Carter, 4
Joe Pesci, 4
John Cazale, 4
wyhs-MacBook-Air:si601_w16_hw4 wyh$ python si601_w16_hw4_part2_solution.py Comedy 10
Top 10 actors who played in most Comedy movies:
Actor, Comedy Movies Played in
Charles Chaplin, 5
John Ratzenberger, 5
Bob Peterson, 3
Hank Mann, 3
John Goodman, 3
Wallace Shawn, 3
Al Ernest Garcia, 2
Billy Crystal, 2
Brad Garrett, 2
Carol Cleveland, 2
wyhs-MacBook-Air:si601_w16_hw4 wyh$ python si601_w16_hw4_part2_solution.py Sci-Fi 10
Top 10 actors who played in most Sci-Fi movies:
Actor, Sci-Fi Movies Played in
Harrison Ford, 4
Alec Guinness, 3
Anthony Daniels, 3
Carrie Fisher, 3
David Prowse, 3
Kenny Baker, 3
Mark Hamill, 3
Peter Mayhew, 3
Sigourney Weaver, 3
Arnold Schwarzenegger, 2
```

What to submit:

A zip file named si601_w16_hw4_youruniquename.zip that contains:

1. Your python program files for part 1 and part 2
2. Your output files for part 1 and part 2 (.db, .dot and .pdf files).