Assignment 3
COL362/632

Due date: March 18, 2016, 11:55pm IST

General Instructions

1. Do this assignment in groups of 2 people.
2. Use Postgres 9.3 for this assignment. These tutorials will help you get started
   - `http://www.w3resource.com/PostgreSQL/tutorial.php`
   - `http://www.tutorialspoint.com/postgresql/index.htm`
3. You will submit 1 file: Fname1Lname1-Fname2Lname2.sql. ONLY ONE MEMBER OF THE GROUP SHOULD SUBMIT ON MOODLE.
4. The .sql files are run automatically, so please ensure that there are no syntax errors in the file. *If we are unable to run your file, you get an automatic reduction to 0 marks.*
5. The format of the file should be as follows. All queries should be in a *single* line. Leave a blank line after each query. An example solution file can be found here (`http://www.cse.iitd.ac.in/~prajna/datasets/example.sql`).
6. Many of the queries below require an 'ORDER BY' clause. Using these clauses would by default return results in *increasing* order. If you omitted this clause, your answer will be evaluated as incorrect and zero marks will be awarded.
7. Errors with respect to equality and inequality conditions will also be evaluated as incorrect and zero marks will be awarded.
8. **No changes are allowed in the i) data, ii) attribute names, iii) table names**
9. **You are NOT allowed to use views.**

1 Dataset1

1.1 Instructions

1. This assignment will make use of IMDB Dataset. It consists of the following tables which can be downloaded from the following URLs.

    1. `http://www.cse.iitd.ac.in/~prajna/datasets/imdb2010/actor.txt`
       This file lists the information of all the actors.
    2. `http://www.cse.iitd.ac.in/~prajna/datasets/imdb2010/movie.txt`
       This file lists the information of all the movies.
    3. `http://www.cse.iitd.ac.in/~prajna/datasets/imdb2010/directors.txt`
       It contains the information about all directors.
    4. `http://www.cse.iitd.ac.in/~prajna/datasets/imdb2010/casts.txt`
       Lists which actor has acted in which movie.

5. `http://www.cse.iitd.ac.in/~prajna/datasets/imdb2010/movie_directors.txt`
   Lists which director has directed which movie.
6. `http://www.cse.iitd.ac.in/~prajna/datasets/imdb2010/genre.txt`
   Lists the genre of the movies.
7. `http://www.cse.iitd.ac.in/~prajna/datasets/imdb2010/users.txt`
   A list of all the users who rate movies.
8. `http://www.cse.iitd.ac.in/~prajna/datasets/imdb2010/ratings.txt`
   Ratings for movies by the users.

2. Your database should include following eight tables and you should use only these tables while writing solution of the queries.

   1. actor

Column Name	Data Type
id	int
fname	varchar
lname	varchar
gender	char

   Note: gender can take a single character value, "M" or "F"

   2. movie

Column Name	Data Type
id	int
name	varchar
year	date

   Note: year describes the date when the movie was released

   3. directors

Column Name	Data Type
id	int
fname	varchar
lname	varchar

   4. casts

Column Name	Data Type
pid	int
mid	int
role	varchar

   5. movieDirectors

Column Name	Data Type
did	int
mid	int

Column Name	Data Type
mid	int
genre	varchar

6. genre

7. userProfile
   Note : registered is the date when the user registered to the website. Country describes the country

Column Name	Data Type
userid	varchar
gender	char
age	int
country	varchar
registered	date

   the user belongs to.

8. ratings
   Note: Here rating is a number between 1-10.

Column Name	Data Type
mid	int
userid	varchar
rating	int

1.2 Queries

1. List the average number of movies per director. (1 column)

$$Average = \frac{\text{Total number movies}}{\text{Total number of directors}}.$$

   Note: Report result up to 6 digits of precision.

2. List the number of users rating movies from India. (1 column)

3. List the average age of people who rate movies. Report result upto 3 digits of precision. (1 column)

4. List the first and last names of all the actors who played in the movie 'Officer 444' order alphabetically. (2 columns)

5. List all the directors (both first and last names) who directed a 'Film-Noir' movie in a leap year order alphabetically. (2 columns)

6. List all the actors (both first and last names) who acted in a film between 1990 - 2000 (inclusive) order alphabetically. (2 columns)

7. List the number of users who have an account but have never rated any movie.
   (1 column)

8. List the number of users who have registered after 1st January, 2006. (1 column)

9. List all directors (only first names) who directed 500 movies or more, in descending order of the number of movies they directed. Return the directors' names and the number of movies each of them directed. (2 columns)

10. Find out the movies in the decreasing order of the number of actors acted in them. List the movie name and the number of actors. Resolve ties alphabetically by the name of the movies.(2 columns)

11. For each director, calculate the ratio of the number of female actors to the number of male actors in the movies directed by the director. List the name (only first name) of the director and the ratio (6 digits of precision) order by first name. (2 columns)

$$ActorGenderRatio = \frac{\text{Total number female actors}}{\text{Total number of male actors}}.$$

12. For each actor, return the genre of the movie that he/she has worked on maximum number of times. If an actor has two or more such genres, return all of them. Order the result alphabetically on actor's name. (2 columns)

13. List the movie names in the increasing order of ActorGenderRatio of the actors. ActorGenderRatio should be calculated as defined in query number 11. (1 column)

14. List the top 10 movies rated by users in the year 1990-2000 (both inclusive) order alphabetically.(1 column)

    Note: Result should contain exactly 10 tuples. Resolve ties alphabetically.

15. List the favorite movies of users in the age range from 13 to 21 (inclusive) order by movie name.

16. List the 5 most *active countries* rating the movies ordered by rank. A country "A" is more active than country "B" if people from country "A" have rated movies greater number of times than people from country "B". Resolve ties alphabetically. (1 column) *Note: Result should contain exactly 5 tuples.*

17. Find out the value of *GenderRatingRatio*. (1 column)

$$GenderRatingRatio = \frac{\text{Total number of times a movie has been rated by female users}}{\text{Total number of times a movie has been rated by male users}}$$

    (1 column) *Note: Report result up to 6 digits of precision.*

18. List the user and his *favourite director* ordered by userid. A director "ABC" is a favorite director of a user "u" if "u" has rated the movie directed by "ABC" the highest among all other movies rated by "u". There can be more than one favorite director of a particular user. In such a case, list all the favourite directors of the user. List both the first and last names of the director. (3 columns)

19. List the user and his *statistically favourite director* ordered by userid. A director "ABC" is a statistically favorite director of a user "u" if the average rating that "u" has given to movies directed by "ABC" is the highest among all other average ratings given to other directors. There can be more than one statistically favorite director of a particular user. In such a case, list all of them. (2 columns)

20. List all the genres, their average ratings, number of times a movie of that genre has been rated and the percentage of movies rated in this genre order by average rating. Resolve ties alphabetically. Report average and percentage upto 6 digits of precision. (4 columns)

21. List the actor who is most popular in India. Actor "A" is more popular than actor "B" in a country "C" if people from country "C" have given higher average rating to actor "A" than actor "B". The average rating for an actor is the average rating of all the movies he/she has acted in. Consider only those actors whose movies have been rated at least 5 times. In case of tie, return all the actors order alphabetically. (1 column)

22. List the users whose favorite movie is the same as the favorite movie of the user with with userid = "user_000085" ordered by userid. The favorite movie of a user is the movie that he/she has given the maximum rating to. Order the result in order of the user ids. (1 column)

23. List the year and the average rating of the movies released in that year order by average ratings. (2 columns)

24. List the director who has directed the maximum number of movies where the actor plays the role of "Himself" or "Themselves". Return the first name of the director and the number of movies. (2 columns)

25. List the top 10 active users on this website ordered by rank. A user "A" is more active than user "B" if he/she has rated more number of movies than user "B". (1 column).

    Note: Result should contain exactly 10 tuples. Resolve ties based on alphabetical ordering of userids

26. List the most popular movie in each country. The most popular movie/movies is the one that has got the highest average rating across all the users of that country. In case of a tie, return all movies order alphabetically. (2 columns)

27. Compute the top-5 movies for each country. List the movie or movies which appear the largest number of times in the top-5 lists. If there is more than one movie, list them in alphabetical ordering. (1 column)

28. Compute the top-5 movies for each user. List the "overall top-10" by listing the movies which appear in the maximum number of top-5 lists along with the number of times they appear in decreasing order of rank. (2 columns)

    Note: Result may contain more than 10 tuples. Resolve ties alphabetically.

29. Find the actor who has acted in most number of movies. For the year 1990 to 1995 (both inclusive) , list the average rating of his/her movies. Note that there can be more than one actor. In such a case, list all of them. (2 columns)

30. List the favorite actor for each director in alphabetial order of the director's name. The favorite actor of a director is the actor with which the director has worked the maximum number of times. There can be more than one favorite actors of a director. In such a case, list all of them alphabetically. (2 columns)