Washington State University Vancouver

Introduction to Databases - CS 351

Assignment 5 - Due: 11:59PM Mar. 21

Instructor:
Ben McCamish

February 28, 2019

Overall Assignment

This is a programming assignment that should be written in Python 3. You may use any external libraries for MySQL, but must cite them in your write up. Your write up should be written in LaTeX.

Interface

```
./program username password <query #>

Where 'program' is your python script. 'username' is the name of the user for the database.
'password' is the password. You can assume that the port and hostname of th edatabase is
localhost and 3306. The last argument is optional. If a number is specified, then print that
query. Otherwise, print all queries in the order they are shown here.
```

Part 1 - Data - 50%

Download the CSV file containing 5000 tuples from a movie database. Notice, that the data is not in Second Normal Form. This data should be inserted into a schema of your design in MySQL. The database must be in Second Normal Form. This means that the attributes containing multiple values, for instance genre, need to be in their own table. For example, the Genre attribute may be separated out to contain a list of genres with their IDs as the primary key. Then a table needs to be created that contains the ID of the genre, the ID of the movie, and the ID of the relationship. Once the genre information is separated, you should be left with three tables. This will continue until all of the divisible data is in its own table.

Your assignment needs include at least two methods. The first method should create all of your relations. The second method should insert and parse the data from the provided CSV file into your relations.

Write up

You should include a description of each relation in your schema, including what attributes are primary and foreign keys. Also include an analysis on what normal form this schema is in. You may put the database in a normal form higher than Second, but it may not be lower than Second.

Part 2 - Queries - 50%

For the following questions, write SQL queries that answer them. Your code must include a method for each query. As the third optional command line argument, pass the query that should run. If no value is passed as an argument, then print all in this order. All results should only shown 5 tuples.

1. **10%** What is the average budget of all movies? Your output should include just the average budget value.
2. **10%** Show only the movies that were produced in the United States. Your output must include the movie title and the production company name.
3. **10%** Show the top 5 movies that made the most revenue. Your output must include the movie title and how much revenue it brought in.
4. **10%** What movies have both the genre Science Fiction and Mystery. Your output must include the movie title and all genres associated with that genre.
5. **10%** Find the movies that have a popularity greater than the average popularity. Your output must include the movie title and their popularity.

Write up

Your write up must include the query used for each question. Show the first 5 tuples returned by each query in a table with the attribute names.

What to turn in (in a zip on Blackboard):

- Your write up as a PDF along with the source (LaTeX, Word, etc.)
- All code, well commented.
- README.txt on how to run your code (detailed if required).
- The database created by your program. You may also just dump your database.