BF768 Spring 2022 Homework 2

Due: Sunday February 27, 2022 at 11:59 pm EST

General Policy on Homework Collaboration: Except as otherwise noted, all problem sets/homeworks are to represent individual effort and are to be written up and turned in individually. This does not preclude talking about a problem set with other class members; in fact, working together is encouraged, since it is one of the skills of modern science. The only requirement is that if you work on a problem set with other people, please note on the write-up with whom you worked.

Description: You will be interfacing MySQL with Python and writing Python scripts that perform certain tasks in MySQL.

Part 1

Write a Python script "yourname_part1.py" that connects to your personal database in MySQL (using the pymysql module) and performs the following tasks:

a. Create a table called **Pathways** with fields **path_id** (integer) and **pathname** (length: 100 or fewer characters). The primary key is path_id. Do not make the primary key auto_increment. Include a "drop table" command for Pathways before your "create table" statement.
b. Load the data found in the file **pathways.tab** in the HW2 folder on Blackboard into the Pathways table. You will use the LOAD DATA LOCAL INFILE command discussed earlier.

Part 2

Write a Python script "yourname_read_execute.py" based on the file "read_execute.py" in the HW2 folder on Blackboard to do the following:

a. Create a function "read_query" that takes a filename as input (a string), opens/reads the file, and returns the contents as one string (i.e. removes the endline characters). You'll be using this function to parse SQL queries later.
b. Create a function "execute_query" that takes a SQL query, a database name, a username, and a password as inputs (each one a string), and then connects to the database, executes the query, and returns the results.
c. The main program should take the name of a text file containing a query (e.g.query1.sql from part 3, below), read and execute the query in it, and print the results (one row at a time). You will get the database, username, password, and filename from the system arguments.

The script “yourname_read_execute.py” will be called from the bioed command line as follows:

./yourname_read_execute.py <database> <username> <password> <queryfile.sql>

Since the homework graders have access to the miRNA database, they will be able to use your script to run queries on it. Since *you* have access to the miRNA database, you should be able to *test* your script on it.

Part 3

Write SELECT statements to do the following, and submit each as a separate file:

1. List all genes targeted by the let-7c miRNA, but not targeted by miR-16 (gene id, gene name). No duplicate rows should be displayed. Submit as “yourname_query1.sql”.
2. List the top 10 miRNAs with the most targeted genes (miRNA id, miRNA name, count). You can produce the top 10 with a LIMIT clause. Submit as “yourname_query2.sql”.
3. List the top 10 miRNA pairs with the most gene targets in common (miRNA id #1, miRNA name #1, miRNA id #2, miRNA name #2, count of genes in common). You can produce the top 10 with a LIMIT clause. Submit as “yourname_query3.sql”.
4. List the number of unique miRNAs with a score less than –0.6 in the targets table (count). Only one row should be reported. Submit as “yourname_query4.sql”.

Your homework submission should consist of six files:

- “yourname_part1.py”
- “yourname_read_execute.py”
- “yourname_query1.sql”
- “yourname_query2.sql”
- “yourname_query3.sql”
- “yourname_query4.sql”