CISC 4610 Fall 2020 Extra Credit Homework: SQLite, Neo4j, and MongoDB

Due 12/03/2020

Introduction

For this extra credit assignment, you will be writing several more queries for SQLite, Neo4j, and MongoDB for the same data you have been working with in the other assignments. Starter code is provided in Python to load the JSON into the three databases and query it. You will complete this code with several queries for the three systems. You will submit your code and the output of your queries.

Setup

You should have Neo4j and MongoDB as well as the Python drivers for them and for SQLite set up on your system from the previous assignments. If you removed any of them, please see the setup instructions for the respective assignment. Also remember that you must start the Neo4j database manually and that a mongod process needs to be running (you might have set this up to run automatically at system startup). You can test your setup by executing "python runAll.py" in the directory where you saved this assignment.

Introduction to the code

The Python code is contained in the file "runAll.py". If you have all dependencies installed, you should be able to run the script as it is to populate the databases. By default, the code removes and reinserts the content of all three databases *every time you run it*! To avoid this and be able to test your queries more quickly:

After the first successful execution of the starter code, you can change line 19 to

loadData = False

Change this back to "True" after any time you accidentally make a change to the data or just want to make sure again that you have the right content loaded in the databases.

Also, if you are unable to run all three databases simultaneously on your system, you can change lines 20 through 22 as needed to test your queries for only one system at a time, e.g., Mongo only:

doSQLite = False
doNeo4j = False
doMongo = True

Tasks

The databases are populated automatically, as discussed above. To complete the assignment, solve the queries marked with "TODO" in the code. Please review the instructions for the previous assignments for details on the data and how to go about creating the queries.

Write code to query the databases

Implement the missing queries (marked by "TODO" comments) in the "querySqlite()", "queryNeo4j()", and "queryMongo()" functions (for MongoDB, use aggregation pipelines for all queries). Note that they are the same 4 queries for all three systems, plus one additional one for Neo4j. They are:

1. List the 10 Images with the greatest number of Landmarks contained in them. List them in descending order of the number of Landmarks they contain, followed by their URL alphabetically. List only the Image URLs and the numbers of Landmarks.
2. List all Landmark descriptions associated with more than one geographic Location in the data. List them in descending order of the number of Locations, followed by the description alphabetically. List only the descriptions and the numbers of Locations.
3. List the 10 Images with the greatest number of Image matches of either type (partial or full). List them in descending order of the number of matches, followed by their URL alphabetically. List only the Image URLs and the numbers of matches.
4. List the 10 documents (Images for which there is a JSON file) with the largest number of relationships of any kind (with Labels, Pages, etc.). List them in descending order of the number of relationships, followed by their URL alphabetically. List only the Image URLs and the numbers of relationships.

The additional 5th query for Neo4j is this:

5. List all "Landmark" nodes associated with more than one geographic Location in the data. List them in descending order of the number of Locations, followed by their description alphabetically. List only the description and the number of locations.
   Note that this query is slightly different from query 2. In a comment in your code, briefly explain why this query returns fewer results than the one above. What about the data causes this?

Submission

Submit the following two files in a zip file named "<lastname>_extra.zip" by 12/03/2020, 2:15pm:

1. Your completed version of "runAll.py" with 13 queries.
2. A text file containing the output of your code.

On Collaboration and Academic Integrity

All code must be produced by yourself (which includes the MongoDB Compass procedure described in the instructions for assignment 3). You may, of course, discuss the assignment with other students but keep those discussions limited to concepts and ideas, do not write the actual code together and *do not configure the stages in MongoDB Compass together*. You may also look for help online but if you use code snippets etc. found online, mark the source. Violations of these policies are subject to penalty.