Assignment 2

Problem Statement

You are given a Twitter dataset (same dataset used in Assignment 1). You have to create a Neo4j database and write a Python program to insert the data into the database. The choice of data model(s) is at your discretion. Nodes and relationships should be associated with the required properties and constraints.

Once your database is loaded with the given data, you have to write Python/Php program(s) to perform the following operations. You have to build a web interface that will take input for each of the following operations and will display the final result.

1. Given a user, retrieve all tweets posted by that user
2. Given a user, retrieve all the users mentioned by the given user.
3. Retrieve the top 20 most frequently co-occurring hashtags. Two hashtags are said to be co-occurring if they co-appear in the same tweet
4. Given a hashtag, retrieve the top 20 most frequently co-occurring user-mentions with the hashtag. A user-mention and hashtag are said to be co-occurring if they co-appear in the same tweet.
5. Retrieve all the hashtags appearing from a given location.
6. Retrieve the top 5 user-user pairs where a user-user pair is ranked by the number of retweets. For example if user1 have retweeted tweets by user2 10 times, then the number of retweets for user1-user2 pair will be 10.
7. Retrieve the top 5 user-user pairs where a user-user pair is ranked by the number of replies. For example if user1 have replied to tweets by user 2 10 times, then the number of replies for user1-user2 pair will be 10.
8. Given a user, delete all tweets by that user

****Results of operations 8 may not be displayed in the web interface**

Evaluation Criteria:

1. On the day of evaluation, you should come prepared with the database and codes for the above problem. You will be given another set of similar queries that you will have to implement on the evaluation day. Your evaluation will be based on this new set of queries.
2. You will have to complete your implementation within the first 90 minutes. The last 90 minutes will be for evaluation.
3. Evaluation will be divided into two parts:
    a. Correctness of output (5 marks). Correct output for a query will fetch you full marks for that query and incorrect output will fetch 0 marks. No partial marks will be awarded.
    b. Honesty and authenticity in coding (5 marks). Plagiarism will be heavily penalized and may fetch you negative marks. Authenticity of your code will be checked offline. For this, you will have to submit your codes. Refer to the submission instructions given below.

Instructions :

1. Codes should be well documented.
2. Files to be submitted:
    a. Submit all the program files. The code for all the queries should be clubbed in one program file and name it query.php/query.py. Your submitted code should include only the queries implemented on the evaluation day and not the queries in this set of problems.
    b. Readme file – Include a readme file stating the necessary details and name it readme.txt.
    c. A file describing your data-model(schema) design in pdf format. Name the file datamodel.pdf.
3. Create a single compressed file (.zip or .tar) that will include all the program files. Do not include any database files.
4. The email-id for submission will be communicated soon.