Assignment 1

Problem Statement:

You are given a Twitter dataset. You have to create a Cassandra database and write a Python program to insert the data into the database. The choice of data model(s) is at your discretion. The data model(s) choice should be done keeping in mind the Cassandra architecture and its goals.

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 an author name, display all tweets posted by that author sorted by decreasing order of date and time. The details of the tweet must include the tweet Id, tweet text, tweet author Id, tweet location and tweet language.
2. Given a keyword, retrieve the tweets containing the keyword and sort them by their popularity in decreasing order. Popularity of a tweet is based on it's like-count. Higher the like-count, more the popularity.
3. Given a hashtag, retrieve all tweets containing the hashtag and sort them in decreasing order of date and time.
4. Given an author name, retrieve all tweets that mentions the author. Sort them in decreasing order of date and time.
5. Retrieve all tweets of a particular date sorted in decreasing order of their popularity where popularity is based on like count of the tweet.
6. Retrieve all tweets from a given location
7. Given a date, retrieve top 20 popular hashtags over the last 7 days. The popularity of a hashtag is determined by its frequency of occurrence over the said period.
8. Given a date, delete all tweets posted on that day.

****Results of operation 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. 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.