CSCI 620/Section 01-02/Rivero, Introduction to Big Data, Spring 2215

Assignment 5 – Data cleaning

Description

In this assignment, we are going to study different data cleaning concepts using MySQL and MongoDB over IMDB data.

Your tasks

1.- Provide a Grade project that takes the MovieInfo.json file available in myCourses as input and our previous MongoDB database. You must add new fields to the movie documents using two de-referencing approaches: title-based and id-based. These new fields are:

1. bechdel-test-title: Result of the Bechdel test (fails/passes) using original title for de-referencing.
2. bechdel-test-id: Result of the Bechdel test using movie id for de-referencing.
3. id-conflicts: Conflicts found when using de-referencing by id (the database document to be update already contains a bechdel-test-id field).
4. title-conflicts: Conflicts found when using de-referencing by original title (the database document to be update already contains a bechdel-test-title field).

Check the template and grading software for more info. (20 points)

2.- Provide a single SQL and a single aggregation queries to compute the median of a given attribute or field. Provide a single SQL and a single aggregation queries to compute the mode of a given attribute or field. These queries must be parameterized. For the median, you must use the fourth method described in Wikipedia (https://en.wikipedia.org/wiki/Quartile#Method_4). For the mode, recall that there can be more than one modes; your queries must retrieve all of them. Your queries must work using the grading software (5 points per query).

3.- Provide each of the following descriptions as a single SQL and aggregation queries. Check the query templates. Your queries must work using the grading software (10 points per query)

3.1.- Histogram of the ratings of movies of a given genre between a given pair of years.

3.2.- Scatter plot (and count) of genre vs. rating of movies between a given pair of years such that, for each combination, there are more than a given number of movies.

3.3.- Times series of average movie rating for a given genre with more than certain votes.

3.4.- Five-number summary of the rating of movies of a given genre between a given pair of years. (Use Method 4 (https://en.wikipedia.org/wiki/Quartile#Method_4) to compute quartiles and median. Note that, in the case of SQL, there is a view you must use.)

Submission instructions

- Use the software template provided in myCourses.
- Submit a single ZIP file to myCourses that must be named as your RIT user, e.g., crrvcs.zip. Do not include '@rit.edu.' The file must contain a folder named

'DeRefMongo' containing your Gradle project, and two folders named 'MedianMode' and 'AnalysisQueries' containing your SQL and MongoDB queries.

- Everything will be graded on a Linux machine, so you must always use the exact names provided in this document, software template and grading software.

Grading rubric

- Check the grading software.