CSCI 620/Section 02/Mior, Introduction to Big Data, Spring 2235

Assignment 6 – Data cleaning

Description

An additional data file is provided on myCourses which contains, for each line, a JSON document describing user accounts on other sites in the Stack Exchange Network.

Your tasks

1. Briefly describe the information stored in the file.
   (10 points)

2. Start by modifying your previous MongoDB Users collection to include the AccountId for each user from the original dataset.

   Then add the information provided by the data file to your MongoDB database of users. Specifically, you need to include the reputation and number of each class of badge received for each other site along with the name of the site. You can choose any value if there are duplicates. Provide a program to perform this operation and report how many successful updates you were able to perform with the given data.

   (Hint: The documents in the file contain the account IDs of each user.)
   (35 points)

3. Assuming that the documents in the file did not contain the account IDs, how would you perform the matching process? Provide a program that analyzes the issues you may find in such a scenario. Consider how many values in the new data set match existing documents and the number of cases where new values match more than one existing document. Include the results in your report.
   (25 points)

4. Provide programs that extract the following data to be plotted from MongoDB. Use your favorite visualization program or library to plot and report your results. Your plots should be included in your report.
   (10 points each)

   4.1. A five-number summary (minimum, first quartile, median, second quartile, maximum) of the reputation of all users.
   4.2. For the top 10 most used tags, the average number of questions for each tag as a bar chart.
   4.3. Number of questions asked each year as a time series plot.