Course ID: CSE-557
Total Points 100

Assignment 3

Deadline: May 4, 2022

Instructions

1. The assignment is to be attempted in pairs.
2. Programming Language: Python or in some cases Java (for analysis), preferable Javascript[d3.js] (for visualizaion)
3. For Plagiarism, institute policy will be followed
4. You need to submit the readme.pdf, Code files and Model files.
5. Report, models and code in .py format should be submitted in the classroom in a zip folder with the name A3_RollNumber1_RollNumber2.zip.
6. Attach a Report.pdf including all your hypothesis, procedure, steps and outcomes in the zip file.
7. You **can use any library** for pre-processing, training, doing experiments and post-processing in all questions.
8. One member should submit on google classroom while other member can mark turn in without the attachment.
9. In case of doubts, please comment on the classroom.

Dataset: Attached in the assignment post.

1 Patents and Publications dataset

Q1: (80 points)

In first assignment you worked on data from (https://drive.google.com/drive/folders/1H__WxZMqG0M0BEmf-1c5azJP-X0t7R4N) and in the second assingment you worked on real time/streaming data extracted from sources like google scholar APIs(https://pypi.org/project/scholarly/) and/or Semantic scholar APIs (https://pypi.org/project/semanticscholar/). Now it is time to utilize the understanding and knowledge built upon both the past problems you have worked with.

1. For this assignment you need to combine the information from both the historical/static and streaming/real time data to run queries and analysis.
2. You need to come up with 6 SQL queries which require information from both the type of datasets.
   (a) The queries should be insightful in the sense some meaningful information should be expressed through them.
   (b) The queries should be executable over windows of streaming data.
   (c) Do not copy the existing queries given in the past assignments as it is.
3. You should use some plotting/graphing library like d3.js to plot interactive graphs in a web app. You must do this for at least 2 queries.
4. Try to use the techniques discussed in the past two assignments to leverage the best resources possible
5. You will be graded upon how well you have understood the data till now, the quality of queries and the method/technology used to attain the final results

2 GhTorrent (Github) dataset

Q1: (80 points)

In first assignment you worked on pullrequests, repos and events data from https://drive.google.com/drive/folders/1XZJkuUu2IFluqxeadBc37B_PjNMcfZy0?usp=sharing and in the second assingment you worked with multiple end points of the github api (https://api.github.com/). Now it is time to utilize the understanding and knowledge built upon both the past problems you have worked with.

1. For this assignment you need to combine the information from both the historical/static and streaming/real time data to run queries and analysis.
2. You need to come up with 6 SQL queries which require information from both the type of datasets.
   (a) The queries should be insightful in the sense some meaningful information should be expressed through them.
   (b) The queries should be executable over windows of streaming data.
   (c) Do not copy the existing queries given in the past assignments as it is.
3. You should use some plotting/graphing library like d3.js to plot interactive graphs in a web app. You must do this for at least 2 queries.
4. Try to use the techniques discussed in the past two assignments to leverage the best resources possible
5. You will be graded upon how well you have understood the data till now, the quality of queries and the method/technology used to attain the final results

3 Bus Movements dataset

Q1: (80 points)

In the first assignment you worked on Delhi Open Transit data (https://opendata.iiitd.edu.in/data/static/) and in the second assingment you worked with a dump of real time/streaming data requested from (https://opendata.iiitd.edu.in/data/realtime/). Now it is time to utilize the understanding and knowledge built upon both the past problems you have worked with. (You can use the data provided to you via the google drive links before)

1. For this assignment you need to combine the information from both the historical/static and streaming/real time data to run queries and analysis.
2. You need to come up with 6 SQL queries which require information from both the type of datasets.
   (a) The queries should be insightful in the sense some meaningful information should be expressed through them.
   (b) The queries should be executable over windows of streaming data.
   (c) Do not copy the existing queries given in the past assignments as it is.
3. You should use some plotting/graphing library like d3.js to plot interactive graphs in a web app. You must do this for at least 2 queries.
4. Try to use the techniques discussed in the past two assignments to leverage the best resources possible
5. You will be graded upon how well you have understood the data till now, the quality of queries and the method/technology used to attain the final results

Course ID: CSE-557
Total Points 100

Assignment 3

Deadline: May 4, 2022

Milestones progress

(20 Points) for completing milestones and setup.

Milestones include

- Reading related papers in first week.
- Reporting progress and asking help to the TA every week at a time convenient for both.

PS: If the code doesn't run or gives any error, 0 points will be awarded.

Best of luck for the assignment.