Course ID: CSE-557
Total Points 100

Assignment 2

Deadline: April 11, 2022

Instructions

1. The assignment is to be attempted in pairs.
2. Programming Language: Python or in some cases Java.
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 A2_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 already worked on data from https://drive.google.com/drive/folders/1H_WxZMqGOMOBEmf-1c5azJP-X0t7R4N?usp=sharing. Now it is time to stream some data and figure out how to handle it in real-time using Kafka:

1. Write a publisher to read data from google scholar APIs (https://pypi.org/project/scholarly/) and/or Semantic scholar APIs (https://pypi.org/project/semanticscholar/) and publish them to the Kafka topic. (Note: Alternate approaches like other pip packages and APIs are also allowed to get this data). (**Points:30**). Read the data in the consumer.
   For the rest of the queries below,
   - You can choose to filter the data at producer and put it in the topic and just read and display the data at consumer
   - Or you can choose to add filters in topic level (You can use partitions. please look it up.)
   - or just post all messages to topic from publisher as you read from API. Then read it at consumer and then apply your query to filter and retain only those records that satisfy the condition.
2. Find all authors who have published in ICML and NeurIPS in 2017. . (**Points:10**)
3. Count average number of citations to papers published in PAKDD in 2018. (**Points: 10**)
4. Find the total number of publications from any of the authors in IIIT Delhi in ML in the year 2019. (**Points: 10**)
5. Get the i10 index of the author with most number of publications for every 2 minute window of streaming. Do this for several windows. (**Points: 10**)
6. Report the performance difference for any one of the above queries when you do filtering at topic level, consumer and producer level (**Points: 10**)

2 GhTorrent (Github) dataset

Q1: (80 points) :

1. Report the number of create events in pytorch in the last week. You have to refer github API to get this (https://api.github.com/) (Hint: For this case it is (https://api.github.com/repos/pytorch/pytorch/events). So you have to read from appropriate API and publish it to topic. Filtering can be done anywhere as mentioned above. (**Points:30**)
2. Report the number of watch events and deleteEvents and corresponding repos for your user ID for the last two weeks or earlier if you have no events in that period. (if you have no events for a long time use the userId VenkteshV. You can sue this API to get the data : curlhttps://api.github.com/users/VenkteshV/events. (**Points:10**)
3. Find the number of pushEvents in total for the numpy repository. (**Points: 10**)
4. How many repositories does your userID have in total? (**Points: 10**)
5. Count the number of repositories that allow forking (with an UserID based query to API). (**Points: 10**)
6. Find the number of issues raised and also report corresponding repositories for the organization "chennaiTricolor" (Hint: https://api.github.com/orgs/chennaitricolor/events)(**Points: 10**).

3 Bus Movements Dataset

Q1: (80 points) Please download the Bus Movements Dataset from https://drive.google.com/drive/u/1/folders/1RQSj5OFaUTag837x8YPTw2T-WEx3oNYj. Stream this data and handle it in real-time using Kafka. Report the results for the following queries:

1. Write a publisher to read data from the above mentioned source and publish them to the Kafka topic. (**Points:30**). Read the data in the consumer.

   For the rest of the queries below,

   - You can choose to filter the data at producer and put it in the topic and just read and display the data at consumer
   - Or you can choose to add filters in topic level (You can use partitions. please look it up.)
   - or just post all messages to topic from publisher as you read from API. Then read it at consumer and then apply your query to filter and retain only those records that satisfy the condition.
2. Report which vehicle is travelling the fastest for all the routes. **(Points:10)**
3. Report the latitude and longitude of all buses that are travelling on the route the fastest and slowest buses are on at the timestamp - 1613984417. **(Points:10)**
4. Report the average speed of all buses on every route between 1200hrs and 1300hrs. **(Points:10)**
5. Report the vehicle IDs of all buses closest to IIITD between 1500hrs and 1600hrs. **(Points:10)**
6. Report instances where a bus goes beyond a threshold speed of 40. **(Points:5)**
7. Report the number of buses travelling at a time on each route. **(Points:5)**

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.