1. Crawl the tweets regarding COVID19 and COVID VACCINE (use appropriate hashtags) from twitter (use tweepy with Apache storm)

2. Carry out the following experiment in an offline setting. You can for example crawl the tweets and store it in a database. Store at least 10k tweets.

● Use Jagadish et. al's algorithm to create the v-optimal histogram. Refer the course link shared by sir (http://www.mathcs.emory.edu/~cheung/Courses/584/Syllabus/06-Histograms/)

3. Now implement Guha's algorithm http://www.mathcs.emory.edu/~cheung/Courses/584/Syllabus/06-Histograms/ in a streaming scenario (stream tweets).

4. Analyze the created histogram by posing some queries and report the error.

Progress report

1. Date -24/04/2021
    a) Installed streaming platform -> Apache Kafka on the system and created a topic.
    b) Installed pykafka on conda
    c) To retrieve the topic from the python instead of terminal-- to do
    d) Writing the steps for installing the kafka on the windows.

Resources for Apache Kafka on windows-
https://towardsdatascience.com/running-zookeeper-kafka-on-windows-10-14fc70dcc771
https://www.youtube.com/watch?v=3XjfYH5Z0f0

Summary->

1. Install apache **kafka binary** package instead of source package.
2. Extract it and store it as C:/kafka for convenience.
3. We will notice that in bin files there will be windows folder and other sh files. The windows contain .bat files for all the utilities like zookeeper start , kafka start and so on..(sh files are for linux environment I guess)
4. Edit the properties file in the config folder->
   In server.properties
   log.dirs= C:/kafka/kafka-logs instead of tmp/kafka-logs
   In zookeeper properties
   dataDir=C:/kafka/zookeeper-data instead of tmp/zookeeper

   We are using only a single cluster now.
5. Now we will always start zookeeper first , only after that we will start the kafka server.
   Cmd-

.\bin\windows\zookeeper-se … .bat (press tab) .\config\zookeeper.properties
The default zookeeper port is 2181 , it will start there

6. Now in another cmd prompt, for starting kafka server
   C:\kafka\.bin\windows\kafka-se … .bat (press tab) .\config\server.properties

7. Command for starting producers and consumer in the terminal

https://towardsdatascience.com/how-to-do-a-sentiment-analysis-in-realtime-using-the-jupyter-notebook-kafka-and-nltk-4470aa8c3c30
https://dorianbg.wordpress.com/2017/11/11/ingesting-realtime-tweets-using-apache-kafka-tweepy-and-python/
https://towardsdatascience.com/track-real-time-gold-prices-using-apache-kafka-pandas-matplotlib-122a73728a88
plot
https://www.dataquest.io/blog/matplotlib-tutorial/

-pip install kafka-python
https://towardsdatascience.com/track-real-time-gold-prices-using-apache-kafka-pandas-matplotlib-122a73728a88

https://towardsdatascience.com/getting-started-with-apache-kafka-in-python-604b3250aa05

2. Date -25-04-2021
1.Done normal crawling from twitter based on certain hashtag
2.Used twitter crawling as producer in kafka and printed the message on the consumer side.
Used kafka-python library

3. To do- apply guha's algorithm on this streaming and make histogram;;
Report error

4. Save the data into the database. // not required with kafka - direct tweepy connection and save into the database or csv as per convenience.

Confluent one is very powerful kafka python library
https://towardsdatascience.com/using-kafka-to-optimize-data-flow-of-your-twitter-stream-90523d25f3e8
https://towardsdatascience.com/getting-started-with-apache-kafka-in-python-604b3250aa05
https://anaconda.org/conda-forge/python-confluent-kafka

Important link postgres and kafka connectivity

Check docker also!
https://hevodata.com/learn/connecting-kafka-to-postgresql-4-easy-steps/
https://hellokoding.com/kafka-connect-sinks-data-to-postgres-example-with-avro-schema-registry-and-python/
https://crate.io/docs/crate/howtos/en/latest/integrations/kafka-connect.html
https://medium.com/@wesleybos99/your-first-data-pipeline-with-kafka-8ed9728e37b0
https://www.confluent.io/hub/confluentinc/kafka-connect-jdbc
https://stackoverflow.com/questions/57158092/push-data-from-kafka-topic-to-postgresql-in-json
https://highalpha.com/data-stream-processing-for-newbies-with-kafka-ksql-and-postgres/
https://medium.com/@tilakpatidar/streaming-data-from-postgresql-to-kafka-using-debezium-a14a2644906d

Date -28-04-2021

To do
1.Implement Jagdish et al done
2.Implement Guha et al. done

Plot the histogram in case of Jagdish
Think of Guha as streaming
Twitter extract verify

Last day

1. **Dont close terminals of cluster server kafka**
2. **If you do … then killing need to be done manually by seeing ports on the terminal**
3. **Producer ne bhej diya , to consumer ek baar hi consume krega ...agar one time sent**