**HW 3 HBase Homework**

**Due date 20th Oct**

|  |  |
| --- | --- |
| Instruction | Points |
| * Download twitter data from the link “https://www.kaggle.com/datasets/gpreda/covid19-tweets” * Insert this data to the HBase: * Keep three column families: (i) Users (ii) Tweets (iii) Extra * Add different columns data under these 3 column families. * Keep a unique row key for each entry. | 15%  \*required  Group |
| * Create versions for the column family that includes “User Bio” column. * Put multiple data in “User Bio” for some specific entry. (Put command) * Get one or more versions of “User Bio” for that entry to see if versioning works. (Get command) | 12.5%  Group |
| * Find how many tweets are from different locations (“user\_location”) | 15%  Student 1 |
| * Find how many users are created in each year. Find the number of users created in each month of 2020. (“user\_created”). Do the same task for the users who are verified. | 12.5%  Student 1 |
| * Find users whose “Screen Name” includes special characters (other than alphabet and number). Do the same task for the users who are verified. | 15%  Student 2 and 3 |
| * Find “Name” of the users who are popular! For example, Show the “Name” of the users who have “User Followers” count more than 6 digits. Show the number of verified popular users and unverified popular users. | 15%  Student 2 and 3 |
| * Find “Tweet Content” that starts with #covid19 (not case sensitive). Do the same task for the users who are verified and popular. | 15%  Student 2 and 3 |

Submit your queries with the screenshots of their outputs  
(either from terminal or programming IDE)

If the result of some query is too long, you can just take one screenshot of the first few result data.