DSCI 551 – HW1 (Spring 2021)

Chat Data Analysis using Python & Firebase

100 points, Due 2/14

In this homework, we will analyze your D2L chat logs and also roster data. We provide you with an example chat log, 551-0125.txt, (which contains all chat messages at our first meeting of the semester). The chat was recorded in the format of: date, time, from whom, message. Note that there is a whitespace between digital time and 'PM', and you need to remove it in your code. The messages may also contains a '\r' character at the end which should be removed in your code too. Each line of record is followed by a blank line. The roster data, 551-mw-roster.csv, is a CSV file which contains student names and the participation location. Note that student names are in different format than that in chat log. Note that your codes may be tested with additional chat logs and roster data with the same format.

1. [Analysis, 20 points]
    a. Write a Python script "stats.py" that computes the total number of chats for each person who participated in the chat. Output the statistics in a JSON file named stats.json.

       Execution format: python stats.py <chat-log-file> <output-file>
       For example, python stats.py 551-0125.txt stats.json

       Format of your output file:

       [{"Person":"John Smith","Message":8},…]

    b. Write a Python script "nochats.py" that finds the students who did not have chat messages and their participation locations. Write output also to a JSON file named nochats.json.

       Execution format: python nochats.py <chat-log-file> <roster-file> <output-file>
       For example, python nochats.py 551-0125.txt 551-mw-roster.csv nochats.json
       Format of your output file:
       [{"Name":"David Chen","Participating from":"United States of America"},…]

2. [Data Conversion, 30 points]
    a. Write a Python script "convert_chats.py" to convert a given log file into JSON file named chats.json in the format specified below.

       Execution format: python convert_chats.py <chat-log-file> <output-file>
       For example, python convert_chats.py 551-0125.txt chats.json
       Format of your output file:
       [{"Time":"4:37PM","Person":"David Chen","Message":"1"},…,]

b. Write a Python script "convert_roster.py" to convert a given log file into JSON file named roster.json in the format specified below.

Execution format: python convert_chats.py <roster-file> <output-file>
For example, python convert_chats.py 551-mw-roster.csv roster.json
Format of your output file:
[{"Name":"John Smith","Participating from":"United States of America"},…]

3. [Searching with Firebase, 50 points]
   a. Write a script "load.py" to load the JSON data for the chat and roster you generated in Part 2 into a Firebase database and create any index structure that you need to answer the following questions.

      Execution format: python load.py chats.json roster.json

   b. Write a Python script "search-person.py" that finds all students whose name contains at least one of the specified keywords (case insensitive).
   For example, python search-person.py 'john smith' will find all students whose name contains either 'john' or 'smith' or both.
   Return the student names one line per student.

      Execution format: python search-person.py 'John SMITH'

      Example output: (please print the output directly)
      John Smith
      John Allen
      Mary Smith

      If no student has the same names as your inputs in this class, please print 'Student Not Found'.

   c. Write a Python script "search-message.py" that finds all messages made by a given student.
   For example, python search-message.py 'john smith' will find all chat messages made by a student whose name is 'john smith' (case insensitive).
   Output the messages tab separated and one line per message. Please do not download the whole json file from firebase. Hint: You can filter names before requests records from firebase.

      Execution format: python search-message.py 'John SMITH'

      Example output: (please print the output directly)
      4:06pm list
      5:12pm variety

If no student has the same name as your input in this class, please print 'Student Not Found'; if this student is in this class but has no chat messages, please print 'This student is quiet.'

You should use Pandas DataFrame whenever you can. Other libraries permitted in this homework are: sys, re, json, csv, and requests. No firebase libraries like firebase_admin or python-firebase are allowed, requests would be convenient enough for this homework.

Scripts that take long run time (e.g. more than 2 minutes) will lead to deductions of points.

Please use Python 3.8 and make sure your scripts can work on ec2 for all the homework.

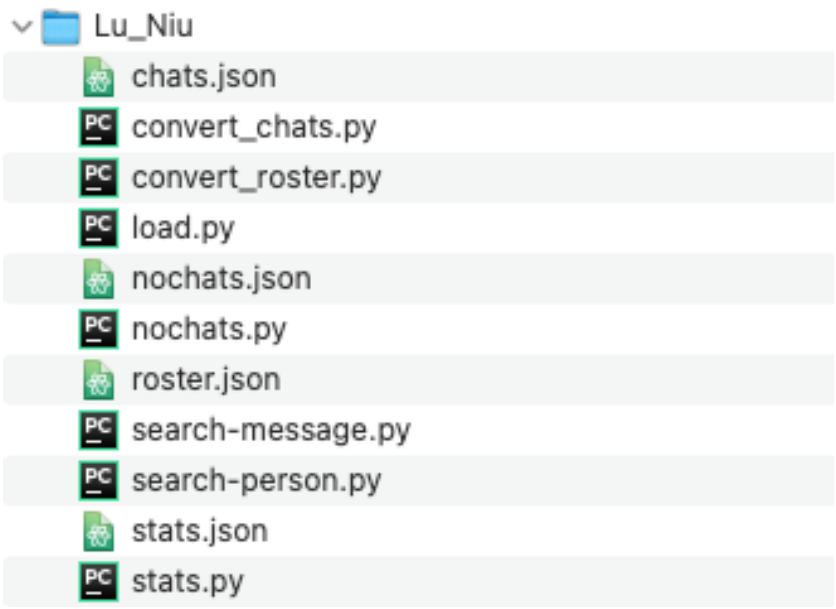
Submission:

Please include all the 7 python scripts and 4 output files below in a folder:

stats.py, nochats.py, convert_chats.py, convert_roster.py, load.py, search-person.py, search-message.py, stats.json, nochats.json, chats.json, roster.json

Please make sure all your scripts' names are correct and do not add directories Q1, Q2…

For example:



Name your zip file: **Firstname_Lastname_hw1.zip**.