DSCI 551 – HW1 (Spring 2021)

Chat Data Analysis using Python & Firebase

100 points, Due 2/14

In this homework, we will analyze your Zoom chat logs and also roster data. We provide you with an example chat log, 551-0119.txt, (which contains all chat messages at our first meeting of the semester). It is a tab-delimited file, where each line records a chat: time, person, and message. Note that person has an extra colon at the end which you will need to remove in your code. The messages may also contain a '\r' character at the end which should be removed in your code too. The roster data, 551-tue-roster.csv, is a CSV file which contains student names and the participation location.

Note:

- Student names are in different format than that in chat log. All student names need to be converted into the format: John Smith, with first and last names separated by a space.
- You should use Pandas DataFrame whenever you can (e.g., in stats.py). Other libraries permitted in this homework are: sys, re, json, and requests.
- 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.

    Execution format: python stats.py <chat-log-file> <output-file>
    For example, python stats.py 551-0119.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.

    Execution format: python nochats.py <chat-log-file> <roster-file> <output-file>
    For example, python nochats.py 551-0119.txt 551-tue-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 in the format specified below.

       Execution format: python convert_chats.py <chat-log-file> <output-file>
           For example, python convert_chats.py 551-0119.txt chats.json
       Format of your output file:
           [{"Time":"00:01:44","Person":"David Chen","Message":"1"},…,]

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

       Execution format: python convert_roster.py <roster-file> <output-file>
           For example, python convert_roster.py 551-tue-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.

    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.
       For example,
           00:21:15 list
           00:22:20 variety,…
           …

Submission: a zip file that contains all the above scripts with specified names. Name your zip file: John_Smith_hw1.zip.