# DSCI 551 – HW1 (Spring 2021)

# Chat Data Analysis using Python & MySQL

## 100 points, Due 3/20

Similar to homework 1, we will analyze your Zoom chat logs and also roster data. The sample data are the same as ones provided in homework 1. However, in this homework, you will be using MySQL to store the data. Refer to handout on how to set up python connector for MySQL. We will test your code assuming data are stored in dsci551 database with user dsci551 and password Dsci-551. You need to use MySQL functions to perform searching and analysis whenever you can. Note that your codes may be tested using additional chat and roster data in the same format as the samples.

1. [Data import, 30 points] Write a data importer "import.py" that loads the data into MySQL as two tables, one for chat log, the other for roster. You can design the structure of tables yourself.

Execution format:

python import.py <chat-log-file> <roster-file>

1. [Analysis, 30 points]
   1. 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 <output-json-file>

Format of your output file:

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

* 1. 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 <output-file>

For example, python nochats.py nochats.json

Format of your output file:

[{"Name":"David Chen","Participating from":"United States of America"},…]

1. [Searching with MySQL, 40 points]
   1. [20 points] 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.

* 1. [20 points] 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,…

…

You should use Pandas DataFrame, [MySQL fulltext search function](https://dev.mysql.com/doc/refman/8.0/en/fulltext-natural-language.html) for this homework. Other libraries permitted in this homework are: sys, re, json, and requests.

Submission: a zip file that contains all the above scripts with specified names. Name your zip file: John\_Smith\_hw3.zip.