WASHINGTON STATE UNIVERSITY

DATA STRUCTURES FOR DATA ANALYTICS –

DATA219

FALL 2024

Final Project

Due: December 8

Instructor:
Madhumitha Sivakumaran

Overview

You have already done a lot of work towards this final project, through the course of the semester, and specifically as part of HW5. In HW5, you selected a dataset, proposed ways to clean it, and suggested data structures that could be used to answer questions of interest from the dataset. The final project is essentially an extension of your HW5. In this final project, you will:

1. Concretely define the question you are interested in answering from the dataset.
2. Implement your data analysis in pursuit of this question using the 2 different data structures you proposed for data storage.
3. Compare the performance of your analysis operations using the two different data structures.

Meeting these three requirements is essential for a successful project. Below are more guidelines before you attempt your final.

Cleaning Data

Your HW5 submission included the steps for data cleaning, which will be assessed in your final project. Your grading will depend on how well you address these data quality concerns specific to your dataset. Pay particular attention to (if these apply to your data):

1. **Completeness**: Have you identified missing data in some columns or rows? How do you plan to handle these missing values without losing a significant amount of data? Does it make sense to impute values?
2. **Accuracy**: Consider the accuracy of the data. For instance, if your data is related to air quality at a specific location over a certain period of time, is it possible to have two different air quality values for the same time instant? Having such data points could indicate inaccuracy in the dataset. Think about what accuracy means within the context of your dataset.
3. **Validity**: Continuing with the air quality dataset example, does your dataset have negative values of air quality? If so, then these values are "invalid," i.e., out of the expected range. Does your dataset have these issues?
4. **Consistency**: Are there any inconsistencies in the format of the data in a particular column? Are the units of measurement consistent across the values of a field?

Define the question to answer, and build your solution

You will provide a clearly defined question that you plan to answer using the dataset. You must identify a complex question that requires significant coding effort. For instance, if your question involves finding the maximum value in a series of data, you should implement your algorithm for obtaining the maximum value in the data structure you are using rather than simply calling the max()function.

Data structures and analysis

In preparation for your final project, you have already identified the data structures that will be used for data storage in your analysis. It is important to have well-justified reasons for the choice of data structures. For instance, if your data requires range queries, it is not advisable to use a hash. However, if you must use a hash, it is necessary to make it clear and demonstrate it. After justifying, you can use a hash and compare it against something that makes more sense, like a sorted list with binary search.

After selecting the data structures, the next step is to implement the analysis operations and compare the performance of the two data structures. The time it takes to run your analysis can be used to compare the performance. It is important to consider the Big O analysis learned in class and not just run the analysis for a single set of parameters. Instead, the performance of the data structures should be compared across varying input sizes to provide a comprehensive comparison.

Submission

Please **zip** the following and submit

1. The iPython notebook (.ipynb notebook, that is) with all your analysis
2. the dataset(s) you used
3. A 1-3 page PDF giving a summary of (checklist):
    - your data
    - chosen data structures
    - justifications why you chose the data structures
    - question you are trying to answer
    - cleaning steps
    - the operations you performed and your implementations of them
    - performance analysis for the two data structures

Notes:

If you are working as a pair, this pairing must be the same as that for HW5.

If working as a pair, ensure you both are putting in comparable effort. I will assess this when grading the pair. The partners may not get the same credit if it is ascertained that one of the members did much of the heavy lifting.

If working as a pair, only 1 member of the group should submit the zip file, clearly writing in the zip file name the names of the 2 members who worked together.

You will be required to present your work to me over a zoom call where you will walk me through your project. If you submitted as a pair, then you will both be required to present the work to me during a shared zoom call. I will send a schedule for signing up for this demo.

Grading rubric

10% Cleaning Data

10% Defining the question of interest

10% Data Structure 1 for storing data

10% Data Structure 2 for storing data

10% Implementation of the analysis performed on Data structure 1

10% Implementation of the analysis performed on Data structure 2

20% Comparison of operations with data structures

10% Writeup

10% Demo/Presentation.