Data Science and Big Data Analysis

Bonus Assignment

Utilizing Lamar Red HPC For Big Data Analysis

Choose one large online dataset (size of dataset is larger than 100 M) that you are interested in (provide link in your read me file).

Use any programming language that you are familiar with to finish this project on Lamar RED HPC (Python Matlab, and JAVA for big data analysis are available on HPC):

Application of Lamar RED HPC account is available under Module 1.

1: preprocess the dataset if needed using the techniques that you have learned.

2: perform the explorative analysis techniques on the dataset.

3: perform the visualization techniques on the dataset.

4: Choose either two classification techniques or two clustering techniques that you are interested in, apply them to the dataset that you selected, and compare the performance of those two techniques. You can choose any clustering/classification techniques, not limited to those that you introduce for this class.

Note: many programming languages, such as Python, and Matlab, etc., have both clustering algorithms in their library, you do not need to write your own code to implement those techniques, you only need to know how to use the them in the programming language that you chose. You might need to do preprocessing of the dataset before you use any classification/clustering techniques.

Submit your program and a project report that includes:

1: link of the dataset

2: instructions of how to download your dataset and compile and run your program

3: explanation of what preprocessing techniques you have applied

4: explain what parameters values you chose for each algorithm and why

4: screenshot of the steps and results of running your program in each step on HPC

Submit your program and project report on blackboard.