Big Data and Intelligent Analytics
Spring Semester 2022

INSTRUCTOR: Sri Krishnamurthy
analyticsneu@gmail.com

NOTE: This covers part 1 to get started. Part 2 will be discussed in class next week!

Assignment 2: Working with large datasets

In this assignment, you will work with large datasets to ingest, process, store it so you can access it through different means.

Preparation:

1. Review:
    - https://proceedings.neurips.cc/paper/2020/file/fa78a16157fed00d7a80515818432169-Paper.pdf
    - https://papers.nips.cc/paper/2020/file/fa78a16157fed00d7a80515818432169-Supplemental.pdf
2. https://sevir.mit.edu/sites/default/files/About_SEVIR.pdf
3. https://github.com/MIT-AI-Accelerator/neurips-2020-sevir

Case:

Now that you demonstrated your understanding of the SEVIR datasets in Assignment 1, your team wants to try out different models that can leverage these datasets. The Data science team has told them that they will be building models using machine learning and your team (data engineering) will be expected to help automate and productionize the analytics. They came across the papers[1,2] listed above and want to replicate the experiments. The team is particularly interested in temporal datasets and in synthetic data generation use cases and the paper provides them exactly those examples : Nowcasting

(temporal datasets) and SynRad generation (cross sectional data). Your team wants to familiarize yourselves with the models (3) and asks you to replicate the models.

Tasks:

Part 1:

1. Get familiar with the models by reading the papers[1]
2. You are not expected to train the models. But you will work with pre-trained models. Download the models (https://github.com/MIT-AI-Accelerator/neurips-2020-sevir/tree/master/models)
3. Start with the Synrad usecase in [3] and run the notebooks in https://github.com/MIT-AI-Accelerator/neurips-2020-sevir/tree/master/notebooks
4. Try out the sample datasets and complete the notebooks
5. Note that the sample datasets for the Nowcasting notebooks are missing. You will need to create test datasets to try out the models.
6. Generate test data using the generators in **https://github.com/MIT-AI-Accelerator/neurips-2020-sevir/tree/master/src/data**
7. Start with Synrad to get a feel for the generators and generate data for 2-3 eventids. Compare the formats you generated to the ones given in the sample datasets. You can use https://www.hdfgroup.org/downloads/hdfview/ to view the files.
8. Now create datasets for the Nowcasting usecase and try out the sample notebooks https://github.com/MIT-AI-Accelerator/neurips-2020-sevir/tree/master/notebooks

Part 2: Building a pipeline to automate the tasks

To be discussed next week

Part 2: Will be announced next week.

Deliverables (Due Feb 25th 11.59am):

1. A 2-5 page report in https://github.com/googlecodelabs/tools format to illustrate your understanding of various steps and outcomes.
2. Github with
    a. Links to the notebook and any other supporting files
    b. Links to dashboard
3. You will be given 10 minutes to present your analysis in class on Feb 25th