Sources for datasets

SVI datasets: https://urban-sustain.org/services/dataDownload.php
Covid datasets: https://urban-sustain.org/services/dataDownload.php
Crime Dataset: https://crime-data-explorer.fr.cloud.gov/pages/downloads

/analysis

We have used **spark framework** for our application part which does the data analysis for the actual proposed problem scope. The **source codes** and **build file** for the spark application are under '***analysis***' directory.

src/main/Q1.scala.

- This is our only class which contains all the codes for analysis
- Comments over each snippet describes the analysis it does

/data_processor

For processing each dataset and merging all the three datasets, we have used python or notebook script which are under '***data_processor***' directory.

Notebook_data_processing.ipynb

- normalize data,
- numerical value conversion,
- rearrange columns
- and save to csv

parseSVI.py

– reads the json file for svi datasets and converts to csv

parseCovid.py

- reads the json file for covid datasets and sends it to parse together

Parsetogether.py

– converts the parseCovid into csv

mergeAll.ipynb

– inner join all the three csvs and make a merged csv

Spark Configuration:

Spark-defaults.conf

spark.master spark://madison:41278

spark.eventLog.enabled true

spark.eventLog.dir hdfs://madison:41251/spark_log

spark.serializer org.apache.spark.serializer.KryoSerializer

spark.driver.memory 2g

spark.executor.extraJavaOptions -XX:+PrintGCDetails -Dkey=value -Dnumbers="one two three"

spark.kryoserializer.buffer.max 128m

spark-env.sh

export SPARK_MASTER_IP=madison

export SPARK_MASTER_PORT=41278

export SPARK_MASTER_WEBUI_PORT=41277

export SPARK_WORKER_CORES=3

export SPARK_WORKER_MEMORY=1g

export SPARK_WORKER_INSTANCES=4

slaves

montpelier
nashville
santa-fe
hartford
helena
montgomery
honolulu
frankfurt
boston
charleston
columbia
boise
indianaolis

Command for job submission

$SPARK_HOME/bin/spark-submit <JAR Path> --master <spark://MASTERNODE:PORT> --class <class-name> <input-file-path> <output-directory-name-by-state> <input-file-path>

Example:

$SPARK_HOME/bin/spark-submit target/scala-2.12/cs455_term_project_2.12-1.0.jar --master spark://madison:41278 --class Q1 /test_spark/co_combined_norm.csv CO