Instructions

As mentioned in the assignment the code generates 4 files – 1 for output of each task and a time comparison file for Task2. The entire assignment consists of 3 scala files that perform the tasks as mentioned.

The output files generated by the code go in the format of Utkarsh_Gera_<task_id>.csv in the output folder path provided in the command line to the program. It is assumed that your system is setup with scala and spark environments and have necessary things in the PATH variable of the system.

In order to run the jar file provided in the Solution folder please follow the following command

spark-submit –class Task1 Utkarsh_Gera.jar <input file path> <output folder path>

spark-submit –class Task2 Utkarsh_Gera.jar <input file path> <output folder path>

spark-submit –class Task3 Utkarsh_Gera.jar <input file path> <output folder path>

You will find the output files generated in the output folder path you gave. You may give relative or absolute folder/file paths as per need. However if the program finds any issue in the folder paths or you don't provide these 2 path arguments/provide more than 2 arguments the program will exit with a proper message like:

Need at two arguments - Input file and output folder path

Input file doesn't exist

Output folder doesn't exist

To avoid such issues give the file path in the quotes in case it contains space or any other character. ex:

spark-submit --class Task3 assignment1.jar "/home/user/Documents/projects/stack-overflow-2018-developer-survey/survey_results_public.csv" "/home/user/Documents/projects/stack-overflow-2018-developer-survey"