Lab 02	Map Reduce Programming
IT494 Big Data Processing, Autumn ’2023, DAIICT, Gandhinagar; pm_jat	

1. Find the most popular browser (made most requests) from the “web access log” file of lab01. You can know about the position of the agent in weblog from https://httpd.apache.org/docs/2.4/logs.html#accesslog

2. Perform the following task, a simple classification approach.

   Suppose you are given a set of data vectors in the file `iris.csv`. Read each row in data file as a data vector. You are also given a file containing class vectors in `iris_classes`. Read each row in this file as a class vector.

   Both data files are available under

   iris subfolder of the **dataset** folder shared with you. [save this link for future reference]
   https://drive.google.com/drive/folders/1Q0sy0NlD2nkjmzxuYURQoFt5XRZpcScs

   Your computation task goes as following:

   Iterate through all data vector and determine nearest class vector for each data vector. Let nearness be computed by `euclidean distance` between data vector and class vector.

   Your program should output ID of class vector for each data vector. Let ID of class vector be 1,2,3 in the order of their occurrence in the class file.

   Hint: It can be implemented as map only task. Load class vectors in “INIT” method of mapper. Store them in class level field that can be used in mapper method.

3. Compute top 10 earners from the “employee.csv” of lab01. You can only list employee numbers.

   For computing top-10, you can maintain “sorted map” with size of 10 at mappers. Can refer to “lecture slides”.

4. Compute the standard deviation of salary for each department from “employee.csv”

Submission Required: A single pdf file that contains the following:

1. All source code is pasted. Try having the code that is formatted and colored.
2. Link to “colab notebook” that contains source code and outputs of programs for all problems.

Using of markdown or latex for creating PDF would be better.