Mid-Term Assignment

Instructions

1. Solve the problem using map-reduce steps only.
2. Using KMeans library or APIs to solve the problem **will not be accepted** as a solution.
3. All steps should be explained in detail using comments or markdown text.
4. Each step should have a heading and detailed explanations in plain text. Show (print) the results of each step.
5. Databricks platform has to be used for this assignment.
6. Upload the data file to the location: ***/FileStore/tables/jio2022/customerdata.csv***
7. Print the final results
    a. First 10 customers for each segment
    b. Centroid values of each segment
8. The notebook should contain two sections:
    a. First section should demonstrate all the steps and their results
    b. Last section should use "chaining map reduce" to create clusters from the initial data.
9. All steps should be executed correctly
10. Use PEP8 standards for coding (Refer to link).
11. Submit the notebook as .dbc file.
12. The notebook should have the Names and IDs of your group at the beginning.
13. Please do not copy code from your friends. Plagiarism cases will be penalized. Direct copying will result in zero marks for both the groups involved.
14. Notebooks that are not properly documented or commented on will be penalized

Problem Statement

Given the customer's data, you have to segment the data into 3 categories which would be helpful to the company in designing customer campaigns. Data has to be segmented based on the income and age values of customers, using the K-means algorithm.

> 👉 Go through the following video till 4:26 to get a basic understanding of K-means: click Here

Parameters for K - Means :

- **Value of K=3**
- **No. of iterations =40**

Dataset

- Download the dataset from the following link:
  - Link to Data
- The data consists of 3 columns which are:
  - Customer Id
  - Income
  - Age

Calculations to be used for solving K-means problem

Data Normalization:

- min-max Normalization over column X:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

x = value in column X of a data point

min(x) = minimum value present in Column X

max(x) = maximum value present in Column X

x/ = normalized value of x

Centroid Calculation:

- Formula to Calculate the Centroid of each cluster

The centroid of a finite set of k points $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k$ in $\mathbb{R}^n$

$$\mathbf{C} = \frac{\mathbf{x}_1 + \mathbf{x}_2 + \cdots + \mathbf{x}_k}{k}.$$

Distance Calculation:

- Formula to Calculate Euclidean distance between data points.

In the Euclidean plane, let point p have Cartesian coordinates (p_1, p_2) and let point q have coordinates (q_1, q_2).

$$d(p, q) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2}.$$