Assignment 2
Clustering

Mining of Massive Datasets Spring 2024

Due Date: 27th March 2024
Submission: Google Classroom

In this assignment, you have to cluster the datasets provided to you **using Apache Pyspark. You have to submit your Python code and a Word document explaining and analyzing your results and findings.**

1. *Perform Kmeans Clustering using your own Pyspark code on **dataset DS1** (you can use the code provided in class and modify it according to your requirements).*
   a. Run K-means for different values of K.
      i. For each value of K, run K-means multiple times.
      ii. Report your findings (error in each clustering, the time required, K that gives the best result, and the number of iterations to convergence for different runs.)
   b. Examine the **quality of clusters** and also of **clusterings**.
      i. Report the errors: within-cluster sum of squared error (WSSE), between-cluster sum of the square error (BSSE), and silhouette coefficient (SC) for each run of K-mean. Write your PySPARK code to calculate BSSE, WSSE, and SC.

2. *Perform **BISECTING Kmeans** Clustering using your own Pyspark code on **dataset DS1**.*
   a. Run *BISECTING Kmeans* for different values of K.
      i. For each value of K, run K-means multiple times.
      ii. Report your findings (error in each clustering, the time required, K that gives the best result)
   b. Examine the **quality of clusters** and also of **clusterings**.
      i. Report the errors: within-cluster sum of squared error (WSSE), between-cluster sum of the square error (BSSE), and silhouette coefficient (SC) for each run of K-mean. Write your PySPARK code to calculate BSSE, WSSE, and SC.

3. Perform **K-MEANS** clustering using **PYSPARK MLLIB Kmeans function** on the given **dataset DS2, DS3.**
   a. Use the ***Silhouette method*** to find the optimal value of K.
      i. Run K-means multiple times for optimal K. Report your findings (error in each clustering, the time required, the number of iterations to convergence for different runs.)
      ii. Report the errors: within-cluster sum of squared error (WSSE), between-cluster sum of the square error (BSSE), and silhouette coefficient (SC) for each run of K-mean. Use **PYSPARK MLLIB library for calculating** BSSE, WSSE, and SC.
   b. RUN Kmeans with K greater than the optimal K and post-process to improve the clustering results. Post-processing can help when clusters are of different sizes, densities, or shapes.

4. Repeat Part 3 above using the ***Bisecting Kmeans*** clustering function provided in **PYSPARK MLLIB.**

5. ***Compare the clustering results of the K-means and Bisecting K-means for all the datasets.***

   NOTE: Draw different plots to visualize the clustering results and include plots in your report.