Assignment 1

Assignment on MapReduce

Due by Sun, Sept 25, 2024

Java programming language is recommended for this assignment, but you can use python as well. Submit a compressed archive (zip, tar, etc.) of your code, along with the input jar file and output file. Also, include a PDF document with answers and CLI screenshots (input/output commands with results) to the questions below. Note: Please provide concise answers.

Contact your TA for any questions related to this assignment or post clarification questions to the Piazza platform.

1. K-means

The k-means algorithm is the most well-known and commonly used clustering method.

- It takes the input parameter, k, and partitions a set of n objects into k clusters so that the resulting intra-cluster similarity is high whereas the inter-cluster similarity is low.

- Cluster similarity is measured according to the mean value of the objects in the cluster, which can be regarded as the cluster's 'center of gravity'.

- The algorithm proceeds as follows:

  - Firstly, randomly selects k objects from the whole objects which represent initial cluster centers.

  - Each remaining object is assigned to the cluster to which it is the most similar, based on the distance between the object and the cluster center.

  - The new mean for each cluster is then calculated. This process iterates until the criterion function converges.

2. Data Input

We are going to cluster data points dataset and this dataset is provided to you, download it from Quercus.

- *data_points.txt*

3. Questions

1) [Marks: 15] Implement a Map Reduce program for counting the number of lines in a document. Use the '*shakespeare.txt'* file and download it from Quercus. Please submit input/output files with code.

2) [Marks: 45] Apply K-means clustering on Map Reduce using k = 5 and k = 8 clusters on the given dataset, list the cluster labels or centroids, the number of iterations for convergence or use maximum iterations = 15 and time/duration.

3) [Marks: 10] Explain the advantages and disadvantages of using K-Means Clustering with MapReduce.

   Please read the paper which is provided with the assignment in the Quercus and answer the following questions.

4) [Marks: 10] Can we reduce the number of distance comparisons by applying the Canopy Selection? Which distance metric should we use for the canopy clustering and why?

5) [Marks: 10] Is it possible to apply Canopy Selection on MapReduce? If yes, then explain in words, how would you implement it.

6) [Marks: 10] Is it possible to combine the Canopy Selection with K-Means on MapReduce? If yes, then explain in words, how would you do that.