Parallel K-Means using Spark

Objective

This assignment aims to enhance your understanding of Spark MapReduce, .

Description

In this assignment, you are asked to implement a new parallel version of K-means clustering algorithm using the Spark map-reduce framework.

Specifications

You should implement a parallel version of the K-Means algorithm using the MapReduce framework. Then, evaluate your clustering algorithm using the IRIS dataset [1] as compared to the original one. In terms of run time and clustering accuracy.

Implement your algorithm in a generalized way (i.e can accept different sizes of feature vector).

Notes

- You do not have to implement the unparalleled K-Means.
- You should deliver a report that contains at least the following:
  - Your MapReduce K-Means algorithm.
  - The challenges you faced to implement it and how you solved it.
  - The evaluation results.

Grading Policies

- You should work in groups of 2 or 3 students.
- No late submission is allowed.
- Plagiarizing is not acceptable. Sharing code fragments between groups is prohibited and all the groups that are engaged in this action will be severely penalized. Not delivering the assignment will be much better than committing this offence.

References

[1] `http://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data`

Good Luck