Parallel and Cloud Computing – Home Work 3

Report

The code has been written in Google Colab and the pyspark installation codes are also available in the .py code file. The code file name is *pyspark_kmeans.py*

1. Forming 0/1 Matrix from input data

Three files *movies.data.txt, ratings.dat.txt and users.dat.txt* has been downloaded from the following GitHub link

https://github.com/databricks/spark-training/tree/master/data/movielens/medium

The movies file has been converted into 0 and 1 matrix format based on the Genres in the column. The itertools has been used to split the values from movies file.

Pivot_table function has been used to convert the matrix.

First the table is stored without index and header, therefore the K-means can be easily implemented to find the best cluster.

One more file is stored with index and header to merge with the cluster value found from the clustering algorithm

2. Implementing K-Means Value

The spark is used to implement the K-Means machine learning algorithm

Spark.mllib has been used to predict the cluster using K-means algorithm and the cluster is tested for 0 to 20 cluster values.

Based on the square error the optimal cluster values has been taken from the clusters.

3. Implementing ALS

ALS has been found by implemented on input files.

Various sample ranks have been used in the ALS to predict the optimal vales.

Using root mean squared error, best model has been found based on rank, iteration and regularized parameter.

From overall performance we can say ALS performed better than K-Means algorithm.