**HOMEWORK 1**

**K MEANS CLUSTERING OF AMAZON REVIEW**

**SUCHISMITHA VEDALA (1470929)**

**AIM:**

To cluster amazon review data using K means clustering.

**GIVEN:**

A set of product items and a list of items, with which product has which rating. We need to implement this using K-means and canopy algorithms using MapReduce concept.

**OVERVIEW:**

The K-means clustering implies dividing the large data into clusters , executing them separately and considering the result as whole later. The MAPREDUCE concept consists of two classes, mapper and reducer where the mapper class divides the data and performs logic whose output is given as input to the reducer.

**ARCHITECTURE:**

|  |  |
| --- | --- |
| FILENAME | STEPS INVOLVED |
| DataProcessing.Java | Step1 |
| Canopy\_KMeans.Java | Step 2,3,4,5 |
| SupporterClass.Java | This file is used for various supporting functions involved in the homework. |
| HomeWorkExecute.Java | This is the wrapper class to execute all the steps defined in the above classes. |

**DESCRIPTION:**

**DataProcessing.Java:**

This reads the reviews and splits the userid, productid and score and writes about the product id to an intermediate file.

**Canopy\_KMeans.Java:**

This has multiple classes namely CanopiesMapper, CanopiesReducer, ptocMapper, ptocReducer, K\_Means4Mapper, K\_MeansReducer and the FinalMapper and FinalRedducer.

STEPS:

* CanopiesMapper and CanopiesReducer classes are the selection of canopies from the data set.
* ptocMapper and ptocReducer classes are those used to map the product to canopies generated in the previous step.
* K\_Means4Mapper, K\_meansReducer are the ones used to apply the K Means algorithm to the processed product canopy file.
* FinalMapper and FinalReducer classes are used to map reduce the data that has been obtained from the K Means algorithm implemented classes.

**SupporterClass.Java:**

This class has functions to read, manipulate, process the data.

**HomeWorkExecute.Java:**

This is a wrapper for all the steps described above.

**OUTPUT**

The intermediate output files are in the cloudc52 subfolder of the homework1 folder.

**README**

Enter the following Commands in the terminal

1. hdfs dfs -rm -r -f /cloudc52/

2. ant clean

3. ant build

4. yarn jar build/lib/cloud\_Suchi.jar homework1.HomeWorkExecute /cloudc-data/amazon- tiny/100k\_lines.txt /cloudc52/out