

Hadoop K-Means

Java implimentation of K-Means for Apache Hadoop/MapReduce 2.

Prerequisites

To run the algorithms you will need Java version 8, Hadoop version 2.9.1, and the latest version of Eclipse.

Map Reduce is not designed for the iterative nature of K-means. As such, deploying our Hadoop K-means implementation requires some abnormal setup from the user. As is, this algorithm is intended to be deployed on a cluster. If you desire to run it in local mode, then some modifications to the code are required.

Local Mode Walkthrough

1. Download both the java files and the dimension 200, SF1 through SF5 files.
2. Create a new project in Eclipse and place the java files in your `src` folder.
3. Add the Hadoop jars to your project.
4. Make the following modications to each respective line within Hadoop-KMeans.java:

- Uncomment line 62, comment out line 63
- Uncomment line 78, comment out line 79
- Uncomment line 97, comment out line 98

5. In your run configuration, copy and paste the following parameter values.

Parameters:

```
28
200
13000
20
0.013
0.0001
path/to/the/input
path/to/the/output
```

Refer to lines 52-57 for a description of the input parameters.

Since Hadoop is running locally, simply specify the input path as the location where you choose to store the files on your system. The output path can be whatever you specify given that the directory does not already exist.

6. Hit run in Eclipse.

Cluster Mode Walkthrough

1. Download both the java files and the dimension 200, SF1 through SF5 files.
2. At the root directory of your cluster's HDFS, create a directory and place the data files in this directory. At the time of execution, this will serve as the path to the input.
3. Create a new project in Eclipse and place the java files in your `src` folder.
4. Add the Hadoop jars to your project.
5. Export the project as a jar and add place the jar in your cluster.
6. As part of the job submission, copy and paste the following parameter values.

Parameters:

```
28
200
13000
20
0.013
0.0001
path/to/the/input
path/to/the/output
```

Refer to lines 52-57 for a description of the input parameters. The output path can be whatever you want given that the directory does not already exist.

7. Submit the job.