

Spark K-Means

Java implimentation of K-means for Apache Spark.

Prerequisites

To run the algorithms you will need Java version 8, Spark version 2.3.2, and the latest version of Eclipse. Spark 2.3.2 is incompatible with newer versions of Java, i.e. 9 >, so the use of Java 8 is a strict requirement.

Specific instructions are provided for both running the algorithm locally and on a cluster.

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 into the `src` directory.
3. Add the Spark jars to your project.
4. Since you will be running Spark in Eclipses (local mode), replace line 25 in the main method with

```
SparkConf conf = new SparkConf().setAppName("SparkKMeans").setMaster("local[*]");
```

5. Copy and paste the following values into your run configuration.

Parameters:

```
80  
200  
13500  
20  
0.0001  
0  
path/to/the/input  
path/to/the/output
```

To see what different parameters these values correspond to, refer to lines 39-46. Since Spark is running locally, specify the input path as the location where you choose to store the files on your system. The output path can be whatever you desire given that the directory does not already exist.

6. Hit run.

Cluster 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 into the `src` directory.
3. Add the Spark jars to your project.
4. Export the project to a jar and place the jar in your cluster.
5. Pass the following values to the job submission.

Parameters:

```
80
200
13500
20
0.0001
0
path/to/the/input
path/to/the/output
```

To see what different parameters these values correspond to, refer to lines 38-45.

Since Spark is running in distributed mode, place the data files into your cluster and specify the location as part of your input parameters. The output path can be whatever you specify given that the directory does not already exist.

6. Submit the job.