Hadoop Distributed Similarity Group-by

Java implimentation of Distributed Similarity Groupby 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 Distributed Similarity Groupby. As such, deploying our Hadoop DSG implementation requires some abnormal setup from the user. As is, this algorithm is intended to be 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. Within your project create the following two directories `InputDir` and `InputDir/iter0`. Place the data files into `InputDir/iter0`.

- Note: You do not to provide an input directory as `InputDir/iter` is the hardcoded input directory.

4. Add the Hadoop jars to your project.
5. Make the following modications to each respective line in Main.java:

- Comment out lines 146-149
- Uncomment lines 151-154

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

Parameters:

```
100
28
200
200
0
0.00025
50000
```

You do not need to provde an output directory as this will be created for you at the time of execution.

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.

7. 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 directoy of your cluster's HDFS, create two directories, `InputDir` and within that directory `InputDir/iter0`, 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

```
100
28
200
200
0
0.00025
50000
```

You do not need to provde an output directory as this will be created for you at the time of execution.

7. Submit the job.