README(Shortest Path)

We have created a map-reduce program for shortest path for the Game of Thrones dataset. This entire project has been implemented in Hadoop Virtual Machine. We ran the MR initially on a small graph dataset and then on large graph dataset.
The small dataset was synthetically created by us and the large was downloaded from the online available resources. These datasets were downloaded from the network of Game of Thrones books, which includes book 1 and book 2.

Link to Large data: https://www.kaggle.com/mmmarchetti/game-of-thrones-dataset
Attributes of the dataset used:

Column Name	Description
Source	Starting point of an edge
Target	End point of an edge from its source node
weight	Distance from source to target

The main file does all the pre-processing of the raw data which includes stripping of unnecessary details and splitting of data in readable lines. This data is then converted into a format suitable to the mapper i.e., in key-value pairs. The data is stored in a new input file in the format of nodeid, distance from the source (0 for the source node), and corresponding neighbours.

The mapper takes input in above format and separates the nodes and values with respect to the node id and calculates the distance of the node from the source node and outputs a path. These are then fed into the reducer, which in turns checks the minimum distance of the particular node and updates the result with it.

Since, this has to be iterated multiple times to obtain the optimal shortest path, we created a shell script to do all the mapper reducer operation on multiple iterations.

How to run

- Synthetic small data text file for shortest path algorithm is provided in the zip file. The name of the file is small_data.txt

- The main.py file is required to run to process the input data to the key-value format to feed in the mapper.
  1. Run the command: python main.py < small_data.txt
  2. The output of the above command will be written into “mapper_input.txt” file.
  3. Copy "mapper_input.txt” into shortestPath/data folder

- Start Hadoop by running:
  start-hadoop.sh

- Make a new directory on HDFS for the input files:
  hdfs dfs -mkdir shortestPath

- Copy the local directory /data to the HDFS directory :
  hdfs dfs -copyFromLocal $HOME/shortestPath/data shortestPath/data

- Make sure you copied them to the right place:
  hdfs dfs -ls shortestPath/data

- Run bash launch.sh to run the MR job for shortest path algorithm.
  Commands: bash launch.sh OR bash launch.sh mapper_input.txt 5
  Note: ‘5' denotes number of iterations

- You can now view the results by running:
  hdfs dfs -cat shortestPath5/output/*
  Note: ‘5' denotes number of iterations

- Or if you need the results locally, copy from hdfs then cat:
  hdfs dfs -copyToLocal shortestPath5/output/ $HOME/shortestPath_output

- Make sure you always stop the HDFS file system before you leave. You can save the current exploration.
  stop-hadoop.sh