README(Page Rank)

We have created a map-reduce program for page rank for the Game of Thrones dataset. This entire project has been implemented in Hadoop Virtual Machine. We ran the MR initially on a small dataset and then on large dataset. The small dataset was synthetically created from the book example and the large was downloaded from the online available resources. These datasets were downloaded from the network of Game of Thrones series.

Link to Large data: https://networkofthrones.wordpress.com/data/
We used series data from the above link.
Attributes of the dataset used:

Column Name	Description
Source	Starting point of an edge
Target	End point of an edge from its source node

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, their corresponding outgoing nodes and Page rank.

The mapper takes input in above format and separates the nodes and neighbours with respect to Nodes and Values. In this handling of dangling nodes is also done, so as to not create empty lists. The division of the nodes and values helps in calculation of the page ranks of the nodes. This is then fed to the reducer, this sums all the incoming page ranks to a node and outputs the final ranking to the nodes. This ranking will help us determine the dominating and submissive nodes.

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

How to run

- Synthetic small data text file for Page Rank is provided in the pagerank zip file. The name of the file is nodes_small.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 pg_main.py
  2. The output of the above command will be written into “small_pg_input.txt” file.
  3. Copy "small_pg_input.txt” into pagerank/data folder

- Start Hadoop by running:
  start-hadoop.sh

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

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

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

- To run the MR job for shortest path algorithm.
  - Commands- bash launch.sh (‘input_file_name’)(‘number_of_iterations’)
  - Ex- bash pg_launch.sh small_pg_input.txt 5
  - **Note**- launch.sh can be run without specifying any parameters.
- You can now view the results by running:
  - hdfs dfs -cat pagerank_final5/output/*
  - where 5 refers to the number of the iterations

- Or if you need the results locally, copy from hdfs then:
  hdfs dfs -copyToLocal pagerank_final5/output/ $HOME/pagerank_output

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