README

We have created a map-reduce program for word co-occurrences for the authors. This entire project has been implemented in Hadoop Virtual Machine. We ran the MR initially on a small JSON dataset and then on large JSON dataset.
The small dataset was downloaded from the publicly available DBLP publications and the large was called through APIs. These datasets were downloaded from the publications which consist of authors, title, publisher, year etc.

In the mapper we started off by doing the preprocessing of data which consists of removing the unnecessary data and finding the attributes which has authors as field to do the further processing. We then searched for the authors key in the JSON data and then selected the text fields of the authors along with the co-authors. The associative array was used to store the multiple co-authors related to an author. The author is stored in string along with the co-author and their count in dictionary format. This is sent as an input to the reducer.

The reducer compares the key and value of the dictionary based on the author. If same author is encountered, it compares its co-authors and increments the counter if the data is found or else appends the co-author to the list.

How to run

- **For small data-** JSON file is provided in the zip file

- **For large data-** We have used requests library to call the API
  pip install requests
- The main.py file is required to run to call the API for large data(1000 entries) and the JSON file is downloaded in the root directory of the main.py file.

Common for both large and small data

- Start Hadoop by running:
  start-hadoop.sh

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

- Copy the small_data.json file from the provided zip folder to $HOME/data

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

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

- The mapper and reducer python file run by calling the below commands.
  hadoop jar $HOME/hadoop/share/hadoop/tools/lib/hadoop-streaming-2.6.4.jar \
  -file $HOME/mapper.py -mapper $HOME/mapper.py \
  -file $HOME/reducer.py -reducer $HOME/reducer.py \
  -input input/data/* -output output/occurrences

  Note- Make sure you specify the output file for every job or delete the old one.

- You can now view the results by running:
  hdfs dfs -cat output/occurrences/*

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

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