Readme:

Follow the steps one-by-one

PageRankMR:

1. Before you run the PageRankMR, you must create input and output locations in HDFS. Use the following commands to create the input directory /user/cloudera/pagerank/input in HDFS:

**$ sudo su hdfs**

**$ hadoop fs -mkdir /user/cloudera**

**$ hadoop fs -chown cloudera /user/cloudera**

**$ exit**

**$ sudo su cloudera**

**$ hadoop fs -mkdir /user/cloudera/pagerank /user/cloudera/pagerank/input**

1. Move the files graph1.txt, graph2.txt, wiki-micro.txt and simplewiki-20150901-pages-articles-processed.xml(Put 1 file at a time and perform PageRank on it) to **/user/cloudera/pagerank/input.**

**Eg. $hadoop fs -put *filenames.ext* /user/cloudera/wordcount/input**

1. Compile the PageRankMR class.

To compile in a package installation of CDH:

**$ mkdir -p build**

**$ javac -cp /usr/lib/hadoop/\*:/usr/lib/hadoop-mapreduce/\* PageRankMR.java -d build -Xlint**

1. Create a JAR file for the PageRankMR application.

**$ jar -cvf pagerank.jar -C build/ .**

1. Run the DocWordCount application from the JAR file, passing the paths to the input and output directories in HDFS.

**$ hadoop jar pagerank.jar org.myorg.PageRankMR /user/cloudera/pagerank/input /user/cloudera/pagerank/output**

1. Output file display:

**$ hadoop fs -cat /user/cloudera/pagerank/output/\***

1. Rename the output file and save it in some folder.

**$ hadoop fs -get /user/cloudera/wordcount/output/part-r-00000 Downloads/Assignment3/graph1.out**

1. Delete the input file, output file and intermediate files(iter0-iter10, inputtask.txt) from output path

**$ hadoop fs -rm -r /user/cloudera/pagerank**

**Now Repeat the step a to h for next input file i.e. graph2.txt, wiki-micro.txt and simplewiki-20150901-pages-articles-processed.xml.**

1. For displaying Top100.out

**$ (head -100 Downloads/Assignment3/simplewiki.out) > Downloads/Assignment3/Top100.out**

**Extra Credit: Convergence of PageRank:**

Convergence is achieved when the error rate for any vertex in the graph falls below a given threshold.

To check the convergence of the PageRank, we need to maintain the output of previous iteration as a HashMap and compare it with the output of the next iteration. We should select a small threshold value which can be neglected during comparison of the outputs. The threshold can be in the order of 10-3 (0.001). The error rate is given as difference between i+1th and ith iteration of PageRank.

**Algorithm:**

IF (PageRank(i+1) – PageRank(i) < threshold )

THEN the PageRank is converged. No need to perform subsequent jobs.

**Note:**

**I have executed simplewiki-20150901-pages-articles-processed.xml on single node(not on Cluster)**

Files attached:

PageRankMR.java, graph1.out, graph2.out, wiki-micro.out, simplewiki.out and Top100.out.