Lab 2 Report

Yizhan Wu (UBIT: yizhanwu)

Zhenyu Yang (UBIT: zhenyuya)

Commands used to execute for wordcount:

start-dfs.sh

#Commands for word count in nyt news articles

hdfs dfs -put $HOME/Desktop/newsdata /input

hadoop jar $HADOOP_HOME/share/hadoop/mapreduce/hadoop-mapreduce-examples-3.1.2.jar wordcount /input /input/newsWords

hdfs dfs -get /input/newsWords /home/cse587/Desktop/newsWords

hdfs dfs -rm -r /input

#Commands for word count in tweets

hdfs dfs -put $HOME/Desktop/twitterdata /input

hadoop jar $HADOOP_HOME/share/hadoop/mapreduce/hadoop-mapreduce-examples-3.1.2.jar wordcount /input /input/twitterWords

hdfs dfs -get /input/twitterWords /home/cse587/Desktop/twitterWords

hdfs dfs -rm -r /input

#Commands for word count in common crawl

hdfs dfs -put $HOME/Desktop/crawldata /input

hadoop jar $HADOOP_HOME/share/hadoop/mapreduce/hadoop-mapreduce-examples-3.1.2.jar wordcount /input /input/crawlWords

hdfs dfs -get /input/crawlWords /home/cse587/Desktop/crawlWords

hdfs dfs -rm -r /input

#Stop hdfs

stop-all.sh

Commands used to execute for word cooccurence:

start-dfs.sh

#Commands for word co-occurrences in nyt news articles

hdfs dfs -put $HOME/Desktop/newsdata /input

hadoop jar $HADOOP_HOME/share/hadoop/tools/lib/hadoop-streaming-3.1.2.jar -files $HOME/Desktop/mappercoocc.py,$HOME/Desktop/reducercoocc.py,$HOME/Desktop/top10news.txt -mapper 'python3 mappercoocc.py top10news.txt' -reducer 'python3 reducercoocc.py' -input /input -output /input/newsWordsCoocc

hdfs dfs -get /input/newsWordsCoocc /home/cse587/Desktop/newsWordsCoocc

hdfs dfs -rm -r /input

#Commands for word co-occurrences in tweets

hdfs dfs -put $HOME/Desktop/twitterdata /input

hadoop jar $HADOOP_HOME/share/hadoop/tools/lib/hadoop-streaming-3.1.2.jar -files $HOME/Desktop/mappercoocc.py,$HOME/Desktop/reducercoocc.py,$HOME/Desktop/top10twitter.txt -mapper 'python3 mappercoocc.py top10twitter.txt' -reducer 'python3 reducercoocc.py' -input /input -output /input/twitterWordsCoocc

hdfs dfs -get /input/twitterWordsCoocc /home/cse587/Desktop/twitterWordsCoocc

hdfs dfs -rm -r /input

#Commands for word co-occurrences in commoncrawl

hdfs dfs -put $HOME/Desktop/crawldata /input

hadoop jar $HADOOP_HOME/share/hadoop/tools/lib/hadoop-streaming-3.1.2.jar -files $HOME/Desktop/mappercoocc.py,$HOME/Desktop/reducercoocc.py,$HOME/Desktop/top10cc.txt -mapper 'python3 mappercoocc.py top10cc.txt' -reducer 'python3 reducercoocc.py' -input /input -output /input/crawlWordsCoocc

hdfs dfs -get /input/crawlWordsCoocc /home/cse587/Desktop/crawlWordsCoocc

hdfs dfs -rm -r /input

#Stop hdfs

stop-all.sh

Visualization:

We used XAMPP to host local server in order to show visualization result. D3.js was used in the process.

Open XAMPP control panel, start Apache service, and then copy the “wordcloud” folder into “htdocs”. And then use the link below to see visualization result:

http://localhost/wordcloud/index.html