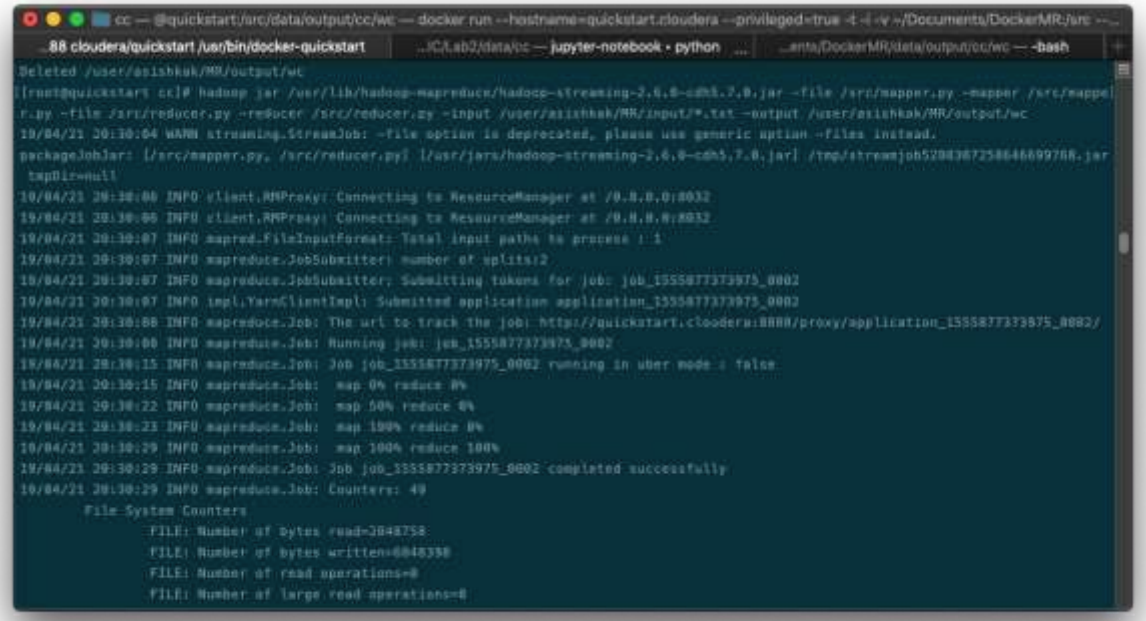
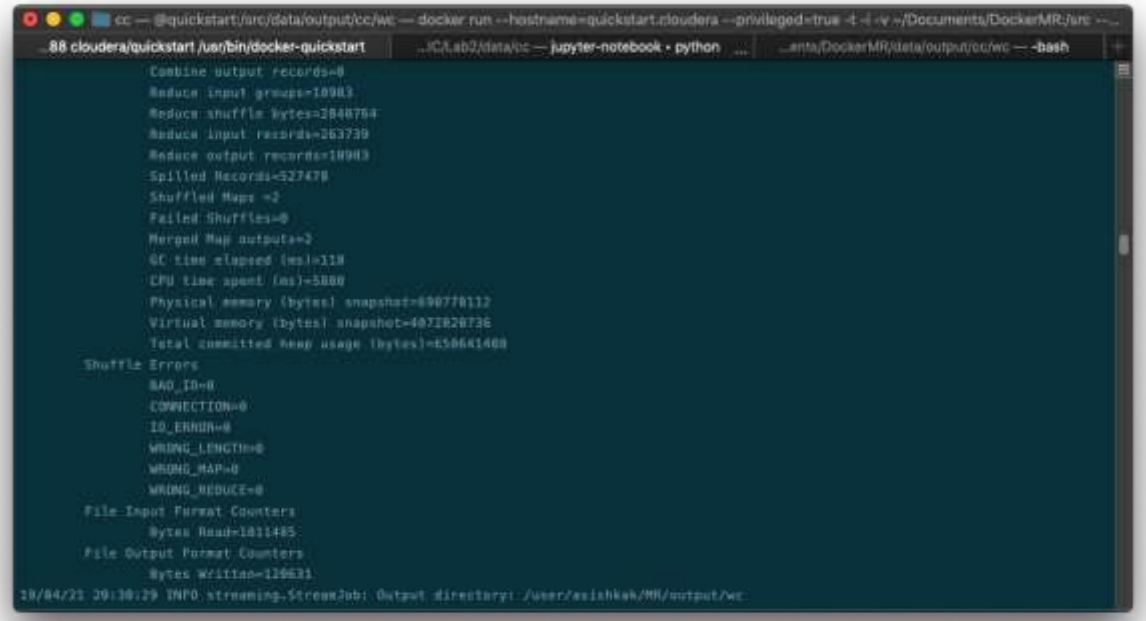
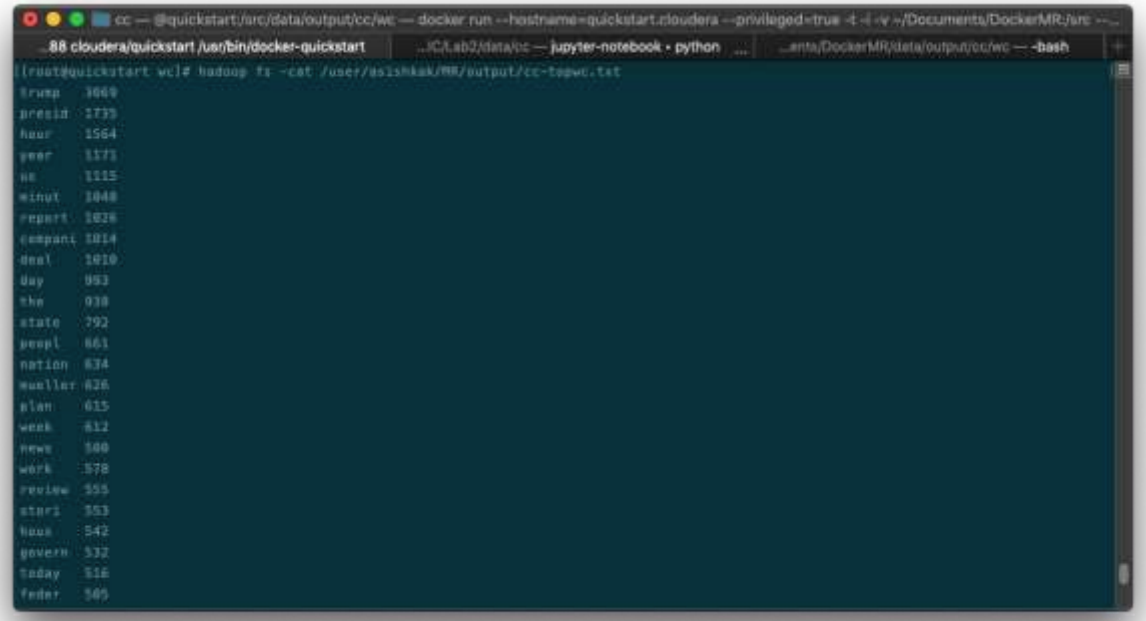
Common Crawl

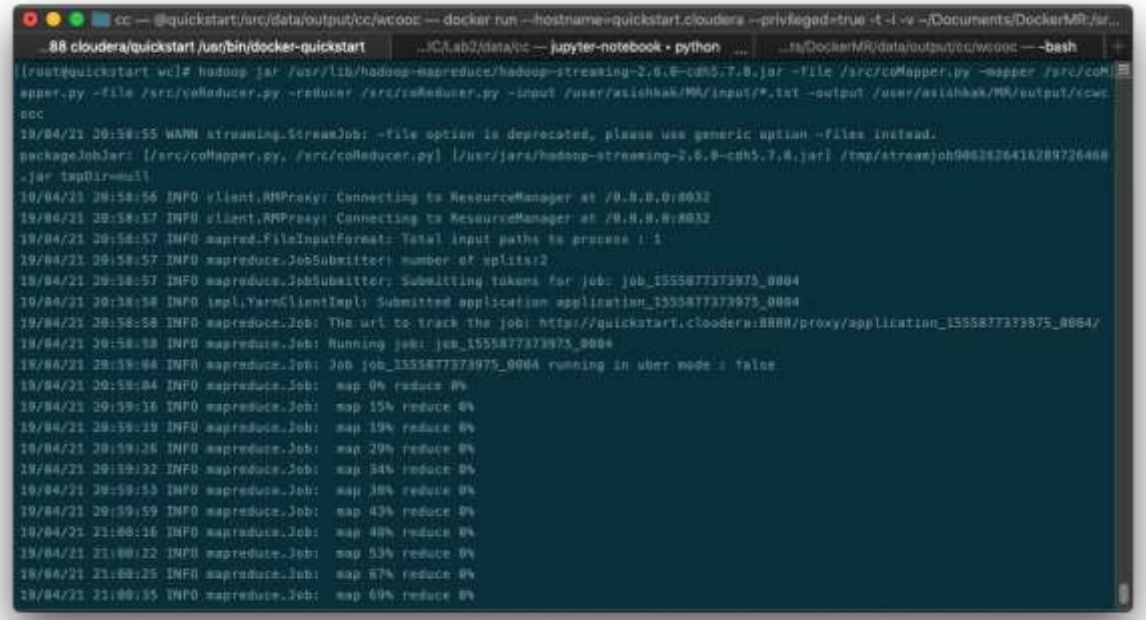
Word Count

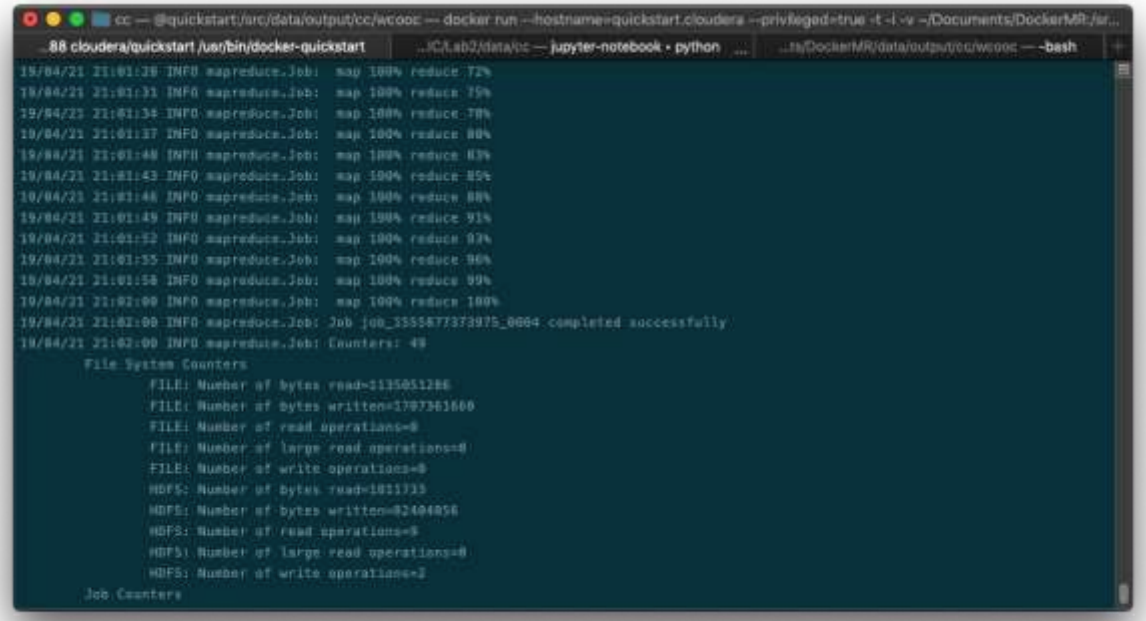


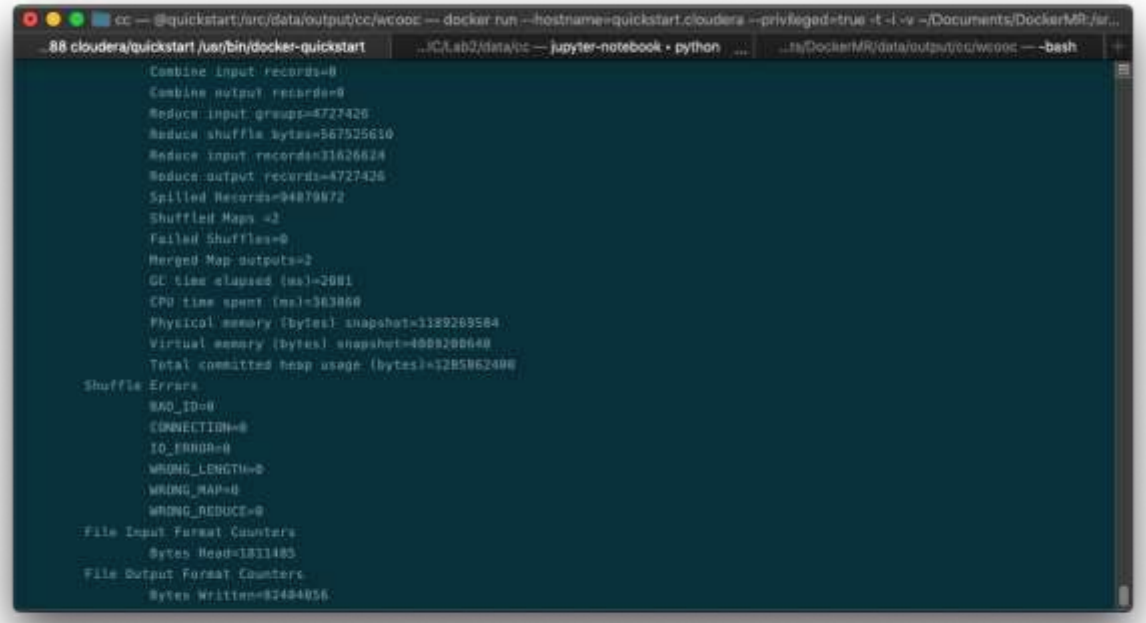


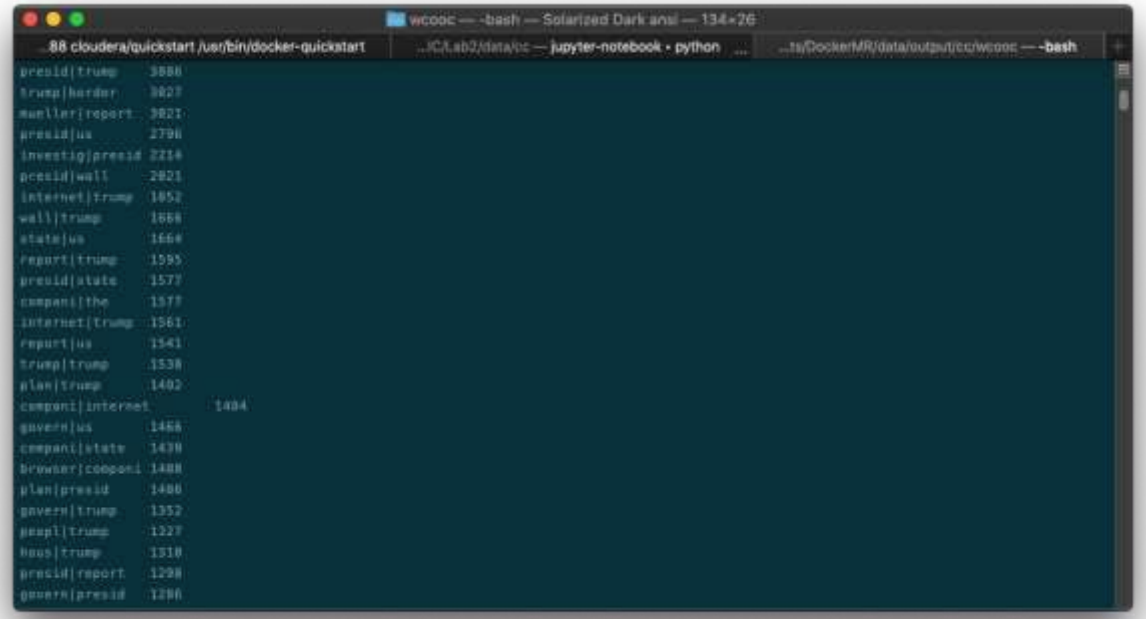


Word Co-occurrence




cc — @quickstart:/src/data/output/cc/wcooc — docker run --hostname=quickstart.cloudera --privileged=true -t -i -v ~/Documents/DockerMR:/sr...
_88 cloudera/quickstart /usr/bin/docker-quickstart
...IC/Lab2/data/cc — jupyter-notebook • python
...ts/DockerMR/data/output/cc/wcooc — -bash
19/04/21 21:01:28 INFO mapreduce.Job: map 100% reduce 72%
19/04/21 21:01:31 INFO mapreduce.Job: map 100% reduce 75%
19/04/21 21:01:34 INFO mapreduce.Job: map 100% reduce 78%
19/04/21 21:01:37 INFO mapreduce.Job: map 100% reduce 80%
19/04/21 21:01:40 INFO mapreduce.Job: map 100% reduce 83%
19/04/21 21:01:43 INFO mapreduce.Job: map 100% reduce 85%
19/04/21 21:01:46 INFO mapreduce.Job: map 100% reduce 88%
19/04/21 21:01:49 INFO mapreduce.Job: map 100% reduce 91%
19/04/21 21:01:52 INFO mapreduce.Job: map 100% reduce 93%
19/04/21 21:01:55 INFO mapreduce.Job: map 100% reduce 96%
19/04/21 21:01:58 INFO mapreduce.Job: map 100% reduce 99%
19/04/21 21:02:00 INFO mapreduce.Job: map 100% reduce 100%
19/04/21 21:02:00 INFO mapreduce.Job: Job job_1555677373975_0004 completed successfully
19/04/21 21:02:00 INFO mapreduce.Job: Counters: 49
File System Counters
FILE: Number of bytes read=1135051286
FILE: Number of bytes written=1707361660
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=1011335
HDFS: Number of bytes written=82404056
HDFS: Number of read operations=9
HDFS: Number of large read operations=0
HDFS: Number of write operations=2
Job Counters


cc — @quickstart:/src/data/output/cc/wcooc — docker run --hostname=quickstart.cloudera --privileged=true -t -i -v ~/Documents/DockerMR:/sr...
_88 cloudera/quickstart /usr/bin/docker-quickstart
...IC/Lab2/data/cc — jupyter-notebook • python
...ts/DockerMR/data/output/cc/wcooc — -bash
Combine input records=0
Combine output records=0
Reduce input groups=4727426
Reduce shuffle bytes=567525610
Reduce input records=31626624
Reduce output records=4727426
Spilled Records=94879872
Shuffled Maps =2
Failed Shuffles=0
Merged Map outputs=2
GC time elapsed (ms)=2081
CPU time spent (ms)=363860
Physical memory (bytes) snapshot=1189269504
Virtual memory (bytes) snapshot=4008280640
Total committed heap usage (bytes)=1285062400
Shuffle Errors
BAD_ID=0
CONNECTION=0
IO_ERROR=0
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_REDUCE=0
File Input Format Counters
Bytes Read=1011485
File Output Format Counters
Bytes Written=82404056


wcooc — -bash — Solarized Dark ansi — 134×26
...88 cloudera/quickstart /usr/bin/docker-quickstart
...IC/Lab2/data/cc — jupyter-notebook • python ...
...ts/DockerMR/data/output/cc/wcooc — -bash
presid|trump 3886
trump|border 3827
mueller|report 3821
presid|us 2790
investig|presid 2214
pcesid|wall 2021
internet|trump 1852
wall|trump 1666
state|us 1664
report|trump 1595
presid|state 1577
compani|the 1577
internet|trump 1561
report|us 1541
trump|trump 1538
plan|trump 1402
compani|internet 1404
govern|us 1468
compani|state 1430
browser|compani 1408
plan|presid 1406
govern|trump 1352
peopl|trump 1327
hous|trump 1310
presid|report 1298
govern|presid 1296