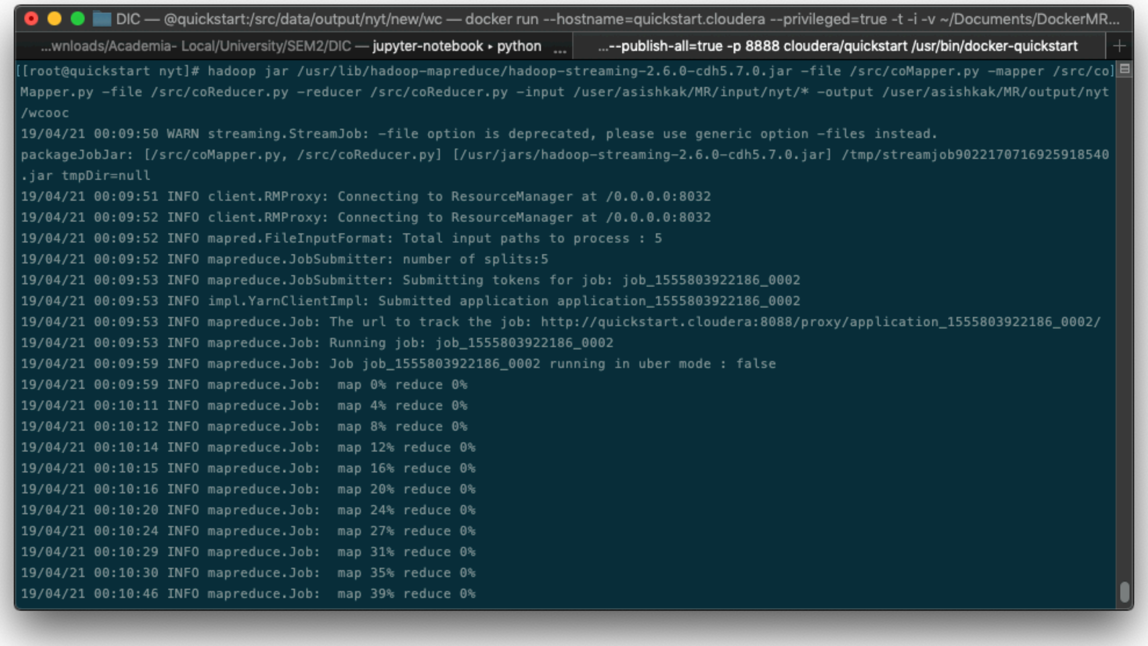
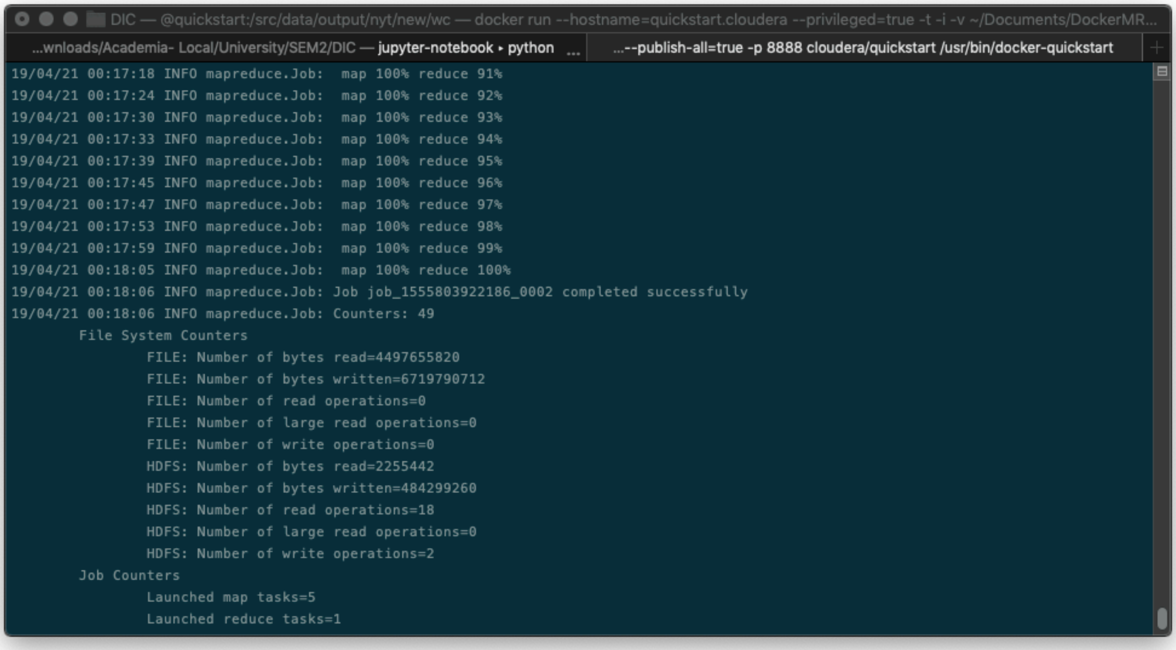
NYT

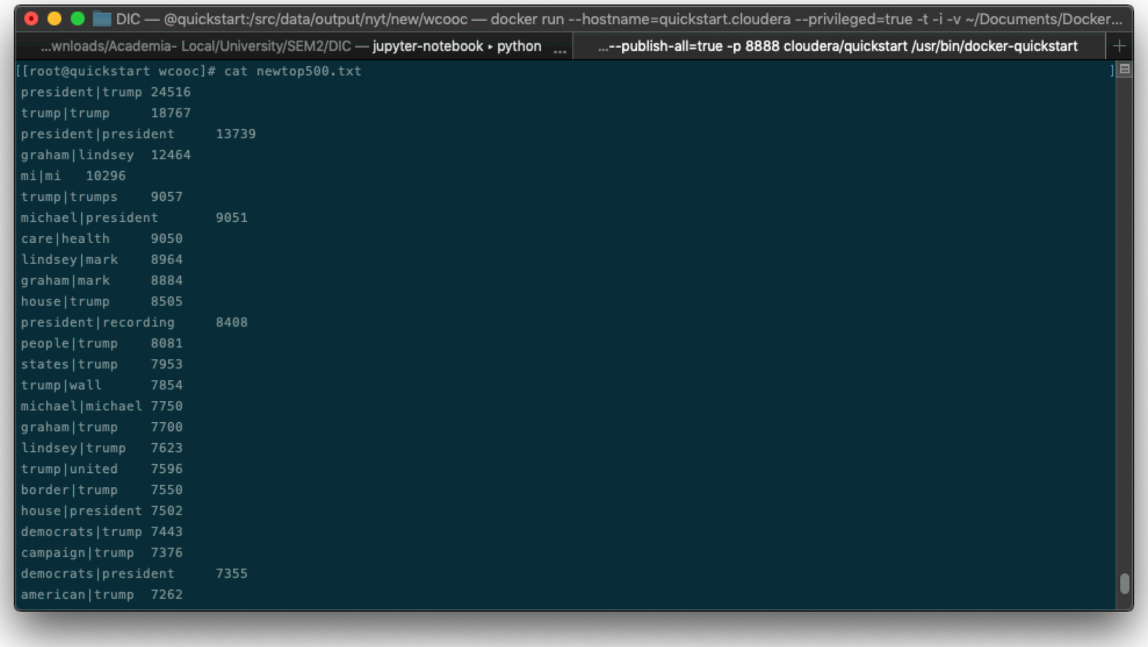
Co-occurrence


DIC — @quickstart:/src/data/output/nyt/new/wc — docker run --hostname=quickstart.cloudera --privileged=true -t -i -v ~/Documents/DockerMR...
...wnloads/Academia- Local/University/SEM2/DIC — jupyter-notebook ▸ python ...
...--publish-all=true -p 8888 cloudera/quickstart /usr/bin/docker-quickstart

```
[[root@quickstart nyt]# hadoop jar /usr/lib/hadoop-mapreduce/hadoop-streaming-2.6.0-cdh5.7.0.jar -file /src/coMapper.py -mapper /src/co]
Mapper.py -file /src/coReducer.py -reducer /src/coReducer.py -input /user/asishkak/MR/input/nyt/* -output /user/asishkak/MR/output/nyt
/wcooc
19/04/21 00:09:50 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [/src/coMapper.py, /src/coReducer.py] [/usr/jars/hadoop-streaming-2.6.0-cdh5.7.0.jar] /tmp/streamjob9022170716925918540
.jar tmpDir=null
19/04/21 00:09:51 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
19/04/21 00:09:52 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
19/04/21 00:09:52 INFO mapred.FileInputFormat: Total input paths to process : 5
19/04/21 00:09:52 INFO mapreduce.JobSubmitter: number of splits:5
19/04/21 00:09:53 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1555803922186_0002
19/04/21 00:09:53 INFO impl.YarnClientImpl: Submitted application application_1555803922186_0002
19/04/21 00:09:53 INFO mapreduce.Job: The url to track the job: http://quickstart.cloudera:8088/proxy/application_1555803922186_0002/
19/04/21 00:09:53 INFO mapreduce.Job: Running job: job_1555803922186_0002
19/04/21 00:09:59 INFO mapreduce.Job: Job job_1555803922186_0002 running in uber mode : false
19/04/21 00:09:59 INFO mapreduce.Job:  map 0% reduce 0%
19/04/21 00:10:11 INFO mapreduce.Job:  map 4% reduce 0%
19/04/21 00:10:12 INFO mapreduce.Job:  map 8% reduce 0%
19/04/21 00:10:14 INFO mapreduce.Job:  map 12% reduce 0%
19/04/21 00:10:15 INFO mapreduce.Job:  map 16% reduce 0%
19/04/21 00:10:16 INFO mapreduce.Job:  map 20% reduce 0%
19/04/21 00:10:20 INFO mapreduce.Job:  map 24% reduce 0%
19/04/21 00:10:24 INFO mapreduce.Job:  map 27% reduce 0%
19/04/21 00:10:29 INFO mapreduce.Job:  map 31% reduce 0%
19/04/21 00:10:30 INFO mapreduce.Job:  map 35% reduce 0%
19/04/21 00:10:46 INFO mapreduce.Job:  map 39% reduce 0%
```

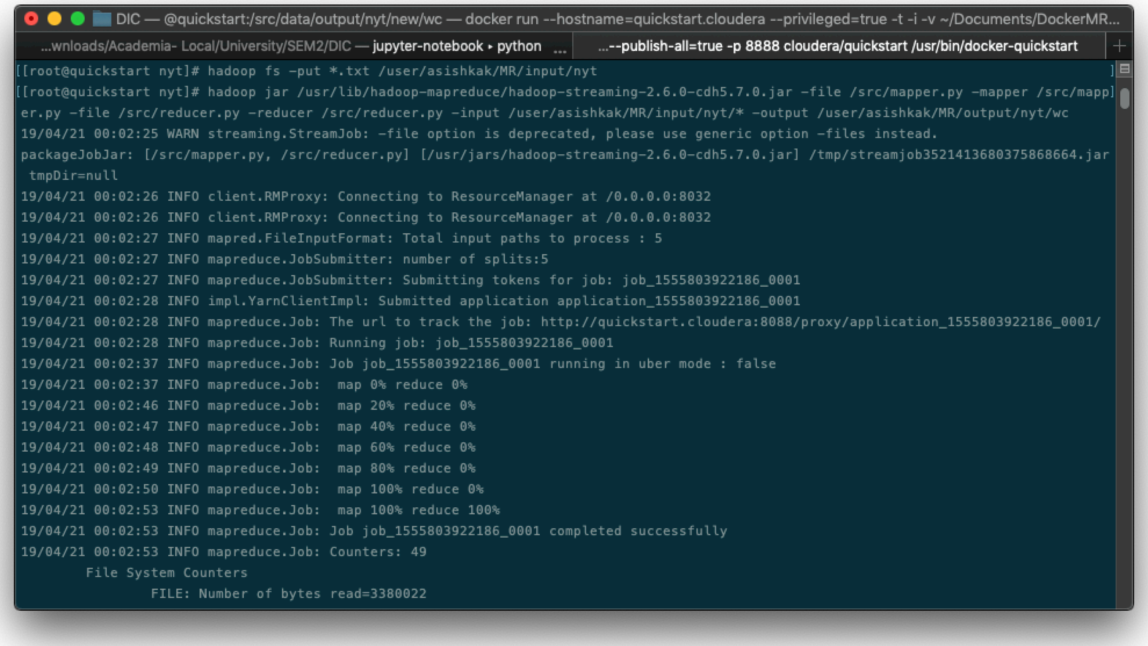

DIC — @quickstart:/src/data/output/nyt/new/wc — docker run --hostname=quickstart.cloudera --privileged=true -t -i -v ~/Documents/DockerMR...
...wnloads/Academia- Local/University/SEM2/DIC — jupyter-notebook ▸ python ...
...--publish-all=true -p 8888 cloudera/quickstart /usr/bin/docker-quickstart

```
19/04/21 00:17:18 INFO mapreduce.Job:  map 100% reduce 91%
19/04/21 00:17:24 INFO mapreduce.Job:  map 100% reduce 92%
19/04/21 00:17:30 INFO mapreduce.Job:  map 100% reduce 93%
19/04/21 00:17:33 INFO mapreduce.Job:  map 100% reduce 94%
19/04/21 00:17:39 INFO mapreduce.Job:  map 100% reduce 95%
19/04/21 00:17:45 INFO mapreduce.Job:  map 100% reduce 96%
19/04/21 00:17:47 INFO mapreduce.Job:  map 100% reduce 97%
19/04/21 00:17:53 INFO mapreduce.Job:  map 100% reduce 98%
19/04/21 00:17:59 INFO mapreduce.Job:  map 100% reduce 99%
19/04/21 00:18:05 INFO mapreduce.Job:  map 100% reduce 100%
19/04/21 00:18:06 INFO mapreduce.Job: Job job_1555803922186_0002 completed successfully
19/04/21 00:18:06 INFO mapreduce.Job: Counters: 49
	File System Counters
		FILE: Number of bytes read=4497655820
		FILE: Number of bytes written=6719790712
		FILE: Number of read operations=0
		FILE: Number of large read operations=0
		FILE: Number of write operations=0
		HDFS: Number of bytes read=2255442
		HDFS: Number of bytes written=484299260
		HDFS: Number of read operations=18
		HDFS: Number of large read operations=0
		HDFS: Number of write operations=2
	Job Counters
		Launched map tasks=5
		Launched reduce tasks=1
```

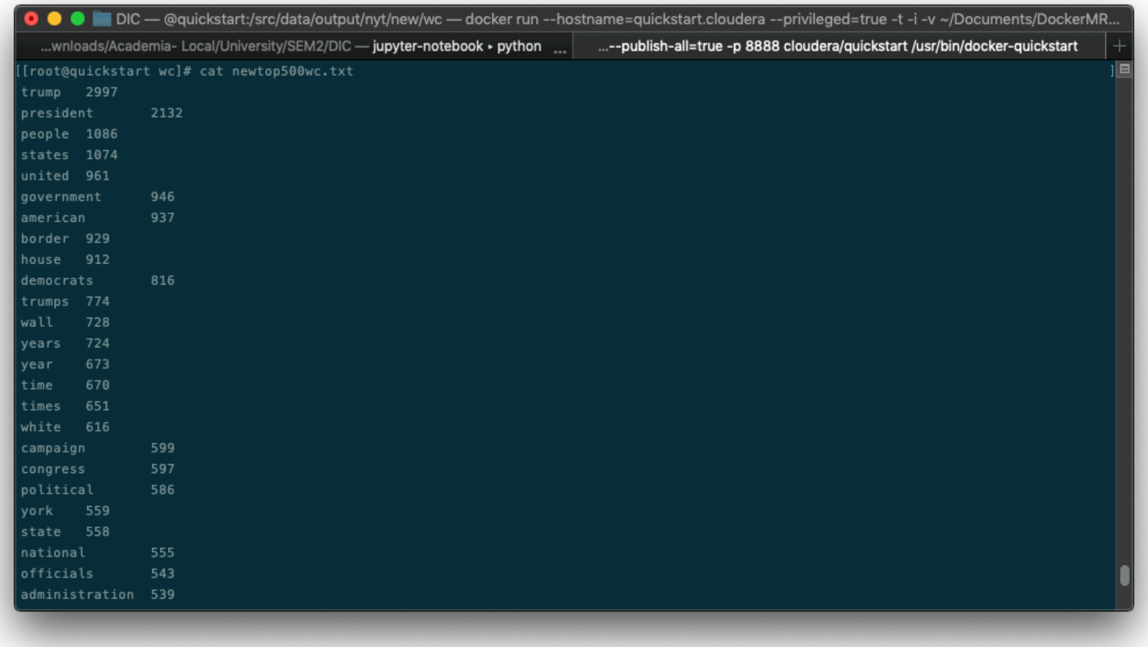

DIC — @quickstart:/src/data/output/nyt/new/wcooc — docker run --hostname=quickstart.cloudera --privileged=true -t -i -v ~/Documents/Docker...
...wnloads/Academia- Local/University/SEM2/DIC — jupyter-notebook ▸ python ...
...--publish-all=true -p 8888 cloudera/quickstart /usr/bin/docker-quickstart

```
[[root@quickstart wcooc]# cat newtop500.txt
 president|trump 24516
 trump|trump     18767
 president|president     13739
 graham|lindsey  12464
 mi|mi   10296
 trump|trumps    9057
 michael|president       9051
 care|health     9050
 lindsey|mark    8964
 graham|mark     8884
 house|trump     8505
 president|recording     8408
 people|trump    8081
 states|trump    7953
 trump|wall      7854
 michael|michael 7750
 graham|trump    7700
 lindsey|trump   7623
 trump|united    7596
 border|trump    7550
 house|president 7502
 democrats|trump 7443
 campaign|trump  7376
 democrats|president     7355
 american|trump  7262
```

Word Count


DIC — @quickstart:/src/data/output/nyt/new/wc — docker run --hostname=quickstart.cloudera --privileged=true -t -i -v ~/Documents/DockerMR...
...wnloads/Academia- Local/University/SEM2/DIC — jupyter-notebook ▸ python ...
...--publish-all=true -p 8888 cloudera/quickstart /usr/bin/docker-quickstart

```
[[root@quickstart nyt]# hadoop fs -put *.txt /user/asishkak/MR/input/nyt
[[root@quickstart nyt]# hadoop jar /usr/lib/hadoop-mapreduce/hadoop-streaming-2.6.0-cdh5.7.0.jar -file /src/mapper.py -mapper /src/mapp
 er.py -file /src/reducer.py -reducer /src/reducer.py -input /user/asishkak/MR/input/nyt/* -output /user/asishkak/MR/output/nyt/wc
 19/04/21 00:02:25 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
 packageJobJar: [/src/mapper.py, /src/reducer.py] [/usr/jars/hadoop-streaming-2.6.0-cdh5.7.0.jar] /tmp/streamjob3521413680375868664.jar
  tmpDir=null
 19/04/21 00:02:26 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
 19/04/21 00:02:26 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
 19/04/21 00:02:27 INFO mapred.FileInputFormat: Total input paths to process : 5
 19/04/21 00:02:27 INFO mapreduce.JobSubmitter: number of splits:5
 19/04/21 00:02:27 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1555803922186_0001
 19/04/21 00:02:28 INFO impl.YarnClientImpl: Submitted application application_1555803922186_0001
 19/04/21 00:02:28 INFO mapreduce.Job: The url to track the job: http://quickstart.cloudera:8088/proxy/application_1555803922186_0001/
 19/04/21 00:02:28 INFO mapreduce.Job: Running job: job_1555803922186_0001
 19/04/21 00:02:37 INFO mapreduce.Job: Job job_1555803922186_0001 running in uber mode : false
 19/04/21 00:02:37 INFO mapreduce.Job:  map 0% reduce 0%
 19/04/21 00:02:46 INFO mapreduce.Job:  map 20% reduce 0%
 19/04/21 00:02:47 INFO mapreduce.Job:  map 40% reduce 0%
 19/04/21 00:02:48 INFO mapreduce.Job:  map 60% reduce 0%
 19/04/21 00:02:49 INFO mapreduce.Job:  map 80% reduce 0%
 19/04/21 00:02:50 INFO mapreduce.Job:  map 100% reduce 0%
 19/04/21 00:02:53 INFO mapreduce.Job:  map 100% reduce 100%
 19/04/21 00:02:53 INFO mapreduce.Job: Job job_1555803922186_0001 completed successfully
 19/04/21 00:02:53 INFO mapreduce.Job: Counters: 49
         File System Counters
                 FILE: Number of bytes read=3380022
```


DIC — @quickstart:/src/data/output/nyt/new/wc — docker run --hostname=quickstart.cloudera --privileged=true -t -i -v ~/Documents/DockerMR...
...wnloads/Academia- Local/University/SEM2/DIC — jupyter-notebook ▸ python ...
...--publish-all=true -p 8888 cloudera/quickstart /usr/bin/docker-quickstart

```
[[root@quickstart wc]# cat newtop500wc.txt
 trump   2997
 president       2132
 people  1086
 states  1074
 united  961
 government      946
 american        937
 border  929
 house   912
 democrats       816
 trumps  774
 wall    728
 years   724
 year    673
 time    670
 times   651
 white   616
 campaign        599
 congress        597
 political       586
 york    559
 state   558
 national        555
 officials       543
 administration  539
```