1.Result

question a).

```
yunke_zhu@cluster-dade-m:~/quiz3/test/3_1$ ./run.sh shingles01
Deleted shingles01
packageJobJar: [] [/usr/lib/hadoop-mapreduce/hadoop-streaming-2.9.2.jar] /tmp/streamjob1297935483830645454.jar tmpDir=null
19/10/19 00:05:01 INFO client.RMProxy: Connecting to ResourceManager at cluster-dade-m/10.128.0.10:8032
19/10/19 00:05:01 INFO client.AHSProxy: Connecting to Application History server at cluster-dade-m/10.128.0.10:10200
19/10/19 00:05:02 INFO client.RMProxy: Connecting to ResourceManager at cluster-dade-m/10.128.0.10:8032
19/10/19 00:05:02 INFO client.AHSProxy: Connecting to Application History server at cluster-dade-m/10.128.0.10:10200
19/10/19 00:05:02 INFO mapred.FileInputFormat: Total input files to process : 5
19/10/19 00:05:02 INFO mapreduce.JobSubmitter: number of splits:21
19/10/19 00:05:02 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enabled
19/10/19 00:05:02 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1571440208858_0008
19/10/19 00:05:03 INFO impl.YarnClientImpl: Submitted application application_1571440208858_0008
19/10/19 00:05:03 INFO mapreduce.Job: The url to track the job: http://cluster-dade-m:8088/proxy/application_1571440208858_0008/
19/10/19 00:05:03 INFO mapreduce.Job: Running job: job_1571440208858_0008
19/10/19 00:05:10 INFO mapreduce.Job: Job job_1571440208858_0008 running in uber mode : false
19/10/19 00:05:10 INFO mapreduce.Job:  map 0% reduce 0%
19/10/19 00:05:21 INFO mapreduce.Job:  map 19% reduce 0%
19/10/19 00:05:22 INFO mapreduce.Job:  map 29% reduce 0%
19/10/19 00:05:24 INFO mapreduce.Job:  map 33% reduce 0%
19/10/19 00:05:29 INFO mapreduce.Job:  map 43% reduce 0%
19/10/19 00:05:32 INFO mapreduce.Job:  map 52% reduce 0%
19/10/19 00:05:33 INFO mapreduce.Job:  map 67% reduce 0%
19/10/19 00:05:37 INFO mapreduce.Job:  map 71% reduce 0%
19/10/19 00:05:38 INFO mapreduce.Job:  map 76% reduce 0%
19/10/19 00:05:39 INFO mapreduce.Job:  map 81% reduce 0%
19/10/19 00:05:41 INFO mapreduce.Job:  map 86% reduce 0%
19/10/19 00:05:42 INFO mapreduce.Job:  map 90% reduce 0%
19/10/19 00:05:43 INFO mapreduce.Job:  map 100% reduce 0%
19/10/19 00:05:51 INFO mapreduce.Job:  map 100% reduce 14%
19/10/19 00:05:52 INFO mapreduce.Job:  map 100% reduce 29%
19/10/19 00:05:53 INFO mapreduce.Job:  map 100% reduce 43%
19/10/19 00:05:54 INFO mapreduce.Job:  map 100% reduce 71%
19/10/19 00:05:55 INFO mapreduce.Job:  map 100% reduce 86%
19/10/19 00:05:56 INFO mapreduce.Job:  map 100% reduce 100%
19/10/19 00:05:56 INFO mapreduce.Job: Job job_1571440208858_0008 completed successfully
19/10/19 00:05:56 INFO mapreduce.Job: Counters: 51
        File System Counters
                FILE: Number of bytes read=65654221
                FILE: Number of bytes written=137226736
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=3970954
                HDFS: Number of bytes written=5985
                HDFS: Number of read operations=98
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=21
        Job Counters
                Killed map tasks=1
                Killed reduce tasks=1
                Launched map tasks=22
                Launched reduce tasks=7
                Data-local map tasks=22
                Total time spent by all maps in occupied slots (ms)=555618
                Total time spent by all reduces in occupied slots (ms)=160443
                Total time spent by all map tasks (ms)=185206
                Total time spent by all reduce tasks (ms)=53481
                Total vcore-milliseconds taken by all map tasks=185206
                Total vcore-milliseconds taken by all reduce tasks=53481
                Total megabyte-milliseconds taken by all map tasks=568952832
```

```
                FILE: Number of bytes read=65654221
                FILE: Number of bytes written=137226736
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=3970954
                HDFS: Number of bytes written=5985
                HDFS: Number of read operations=98
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=21
        Job Counters
                Killed map tasks=1
                Killed reduce tasks=1
                Launched map tasks=22
                Launched reduce tasks=7
                Data-local map tasks=22
                Total time spent by all maps in occupied slots (ms)=555618
                Total time spent by all reduces in occupied slots (ms)=160443
                Total time spent by all map tasks (ms)=185206
                Total time spent by all reduce tasks (ms)=53481
                Total vcore-milliseconds taken by all map tasks=185206
                Total vcore-milliseconds taken by all reduce tasks=53481
                Total megabyte-milliseconds taken by all map tasks=568952832
                Total megabyte-milliseconds taken by all reduce tasks=164293632
        Map-Reduce Framework
                Map input records=10998
                Map output records=3559964
                Map output bytes=58534251
                Map output materialized bytes=65655061
                Input split bytes=2099
                Combine input records=0
                Combine output records=0
                Reduce input groups=72194
                Reduce shuffle bytes=65655061
                Reduce input records=3559964
                Reduce output records=105
                Spilled Records=7119928
                Shuffled Maps =147
                Failed Shuffles=0
                Merged Map outputs=147
                GC time elapsed (ms)=5910
                CPU time spent (ms)=94000
                Physical memory (bytes) snapshot=13225725952
                Virtual memory (bytes) snapshot=123421634560
                Total committed heap usage (bytes)=12215910400
        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=3968855
        File Output Format Counters
                Bytes Written=5985
19/10/19 00:05:56 INFO streaming.StreamJob: Output directory: shingles01
yunke_zhu@cluster-dade-m:~/quiz3/test/3_1$ ./total_pair_statistics.sh shingles01
Most similarity is clinton and obama, which their similarity is:   0.5993
Least similarity is reagan and gwbush, which their similarity is:   0.5485
```

question b).

```
yunke_zhu@cluster-dade-m:~/quiz3/test/3_2$ ./run.sh shingles02
Deleted shingles02
packageJobJar: [] [/usr/lib/hadoop-mapreduce/hadoop-streaming-2.9.2.jar] /tmp/streamjob1101453972537817848.jar tmpDir=null
19/10/19 00:03:54 INFO client.RMProxy: Connecting to ResourceManager at cluster-dade-m/10.128.0.10:8032
19/10/19 00:03:54 INFO client.AHSProxy: Connecting to Application History server at cluster-dade-m/10.128.0.10:10200
19/10/19 00:03:54 INFO client.RMProxy: Connecting to ResourceManager at cluster-dade-m/10.128.0.10:8032
19/10/19 00:03:54 INFO client.AHSProxy: Connecting to Application History server at cluster-dade-m/10.128.0.10:10200
19/10/19 00:03:55 INFO mapred.FileInputFormat: Total input files to process : 5
19/10/19 00:03:55 INFO mapreduce.JobSubmitter: number of splits:21
19/10/19 00:03:55 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enabled
19/10/19 00:03:55 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1571440208858_0007
19/10/19 00:03:56 INFO impl.YarnClientImpl: Submitted application application_1571440208858_0007
19/10/19 00:03:56 INFO mapreduce.Job: The url to track the job: http://cluster-dade-m:8088/proxy/application_1571440208858_0007/
19/10/19 00:03:56 INFO mapreduce.Job: Running job: job_1571440208858_0007
19/10/19 00:04:02 INFO mapreduce.Job: Job job_1571440208858_0007 running in uber mode : false
19/10/19 00:04:02 INFO mapreduce.Job:  map 0% reduce 0%
19/10/19 00:04:12 INFO mapreduce.Job:  map 14% reduce 0%
19/10/19 00:04:14 INFO mapreduce.Job:  map 24% reduce 0%
19/10/19 00:04:15 INFO mapreduce.Job:  map 29% reduce 0%
19/10/19 00:04:19 INFO mapreduce.Job:  map 38% reduce 0%
19/10/19 00:04:20 INFO mapreduce.Job:  map 48% reduce 0%
19/10/19 00:04:26 INFO mapreduce.Job:  map 57% reduce 0%
19/10/19 00:04:27 INFO mapreduce.Job:  map 62% reduce 0%
19/10/19 00:04:28 INFO mapreduce.Job:  map 76% reduce 0%
19/10/19 00:04:30 INFO mapreduce.Job:  map 81% reduce 0%
19/10/19 00:04:33 INFO mapreduce.Job:  map 90% reduce 0%
19/10/19 00:04:34 INFO mapreduce.Job:  map 100% reduce 0%
19/10/19 00:04:43 INFO mapreduce.Job:  map 100% reduce 14%
19/10/19 00:04:44 INFO mapreduce.Job:  map 100% reduce 29%
19/10/19 00:04:45 INFO mapreduce.Job:  map 100% reduce 43%
19/10/19 00:04:46 INFO mapreduce.Job:  map 100% reduce 57%
19/10/19 00:04:47 INFO mapreduce.Job:  map 100% reduce 71%
19/10/19 00:04:49 INFO mapreduce.Job:  map 100% reduce 100%
19/10/19 00:04:49 INFO mapreduce.Job: Job job_1571440208858_0007 completed successfully
19/10/19 00:04:50 INFO mapreduce.Job: Counters: 51
        File System Counters
                FILE: Number of bytes read=76216036
                FILE: Number of bytes written=158350366
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=3970954
                HDFS: Number of bytes written=6089
                HDFS: Number of read operations=98
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=21
        Job Counters
                Killed map tasks=1
                Killed reduce tasks=1
                Launched map tasks=22
                Launched reduce tasks=8
                Data-local map tasks=22
                Total time spent by all maps in occupied slots (ms)=536088
                Total time spent by all reduces in occupied slots (ms)=174039
                Total time spent by all map tasks (ms)=178696
                Total time spent by all reduce tasks (ms)=58013
                Total vcore-milliseconds taken by all map tasks=178696
                Total vcore-milliseconds taken by all reduce tasks=58013
                Total megabyte-milliseconds taken by all map tasks=548954112
                Total megabyte-milliseconds taken by all reduce tasks=178215936
```

```
                HDFS: Number of bytes read=3970954
                HDFS: Number of bytes written=6089
                HDFS: Number of read operations=98
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=21
        Job Counters
                Killed map tasks=1
                Killed reduce tasks=1
                Launched map tasks=21
                Launched reduce tasks=7
                Data-local map tasks=21
                Total time spent by all maps in occupied slots (ms)=569556
                Total time spent by all reduces in occupied slots (ms)=179475
                Total time spent by all map tasks (ms)=189852
                Total time spent by all reduce tasks (ms)=59825
                Total vcore-milliseconds taken by all map tasks=189852
                Total vcore-milliseconds taken by all reduce tasks=59825
                Total megabyte-milliseconds taken by all map tasks=583225344
                Total megabyte-milliseconds taken by all reduce tasks=183782400
        Map-Reduce Framework
                Map input records=10998
                Map output records=3559964
                Map output bytes=69096066
                Map output materialized bytes=76216876
                Input split bytes=2099
                Combine input records=0
                Combine output records=0
                Reduce input groups=1023808
                Reduce shuffle bytes=76216876
                Reduce input records=3559964
                Reduce output records=105
                Spilled Records=7119928
                Shuffled Maps =147
                Failed Shuffles=0
                Merged Map outputs=147
                GC time elapsed (ms)=5834
                CPU time spent (ms)=105340
                Physical memory (bytes) snapshot=13461430272
                Virtual memory (bytes) snapshot=123465150464
                Total committed heap usage (bytes)=12408848384
        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=3968855
        File Output Format Counters
                Bytes Written=6089
19/10/19 01:15:48 INFO streaming.StreamJob: Output directory: shingles02
yunke_zhu@cluster-dade-m:~/quiz3/test/3_2$ ./total_pair_statistics.sh shingles02
Most similarity is clinton and obama, which their similarity is:   0.2065
Least similarity is bush and obama, which their similarity is:   0.1734
yunke_zhu@cluster-dade-m:~/quiz3/test/3_2$
```

question c).

```
yunke_zhu@cluster-dade-m:~/quiz3/test/3_3$ ./run.sh shingles03
Deleted shingles03
packageJobJar: [] [/usr/lib/hadoop-mapreduce/hadoop-streaming-2.9.2.jar] /tmp/streamjob3617927430257033953.jar tmpDir=null
19/10/19 01:00:00 INFO client.RMProxy: Connecting to ResourceManager at cluster-dade-m/10.128.0.10:8032
19/10/19 01:00:00 INFO client.AHSProxy: Connecting to Application History server at cluster-dade-m/10.128.0.10:10200
19/10/19 01:00:00 INFO client.RMProxy: Connecting to ResourceManager at cluster-dade-m/10.128.0.10:8032
19/10/19 01:00:00 INFO client.AHSProxy: Connecting to Application History server at cluster-dade-m/10.128.0.10:10200
19/10/19 01:00:01 INFO mapred.FileInputFormat: Total input files to process : 5
19/10/19 01:00:01 INFO mapreduce.JobSubmitter: number of splits:21
19/10/19 01:00:01 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enabled
19/10/19 01:00:01 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1571440208858_0020
19/10/19 01:00:01 INFO impl.YarnClientImpl: Submitted application application_1571440208858_0020
19/10/19 01:00:01 INFO mapreduce.Job: The url to track the job: http://cluster-dade-m:8088/proxy/application_1571440208858_0020/
19/10/19 01:00:01 INFO mapreduce.Job: Running job: job_1571440208858_0020
19/10/19 01:00:07 INFO mapreduce.Job: Job job_1571440208858_0020 running in uber mode : false
19/10/19 01:00:07 INFO mapreduce.Job:  map 0% reduce 0%
19/10/19 01:00:15 INFO mapreduce.Job:  map 5% reduce 0%
19/10/19 01:00:16 INFO mapreduce.Job:  map 14% reduce 0%
19/10/19 01:00:19 INFO mapreduce.Job:  map 24% reduce 0%
19/10/19 01:00:20 INFO mapreduce.Job:  map 33% reduce 0%
19/10/19 01:00:21 INFO mapreduce.Job:  map 38% reduce 0%
19/10/19 01:00:23 INFO mapreduce.Job:  map 48% reduce 0%
19/10/19 01:00:28 INFO mapreduce.Job:  map 52% reduce 0%
19/10/19 01:00:30 INFO mapreduce.Job:  map 71% reduce 0%
19/10/19 01:00:31 INFO mapreduce.Job:  map 76% reduce 0%
19/10/19 01:00:32 INFO mapreduce.Job:  map 81% reduce 0%
19/10/19 01:00:35 INFO mapreduce.Job:  map 86% reduce 0%
19/10/19 01:00:36 INFO mapreduce.Job:  map 90% reduce 0%
19/10/19 01:00:37 INFO mapreduce.Job:  map 95% reduce 0%
19/10/19 01:00:38 INFO mapreduce.Job:  map 100% reduce 0%
19/10/19 01:00:45 INFO mapreduce.Job:  map 100% reduce 14%
19/10/19 01:00:46 INFO mapreduce.Job:  map 100% reduce 43%
19/10/19 01:00:47 INFO mapreduce.Job:  map 100% reduce 57%
19/10/19 01:00:49 INFO mapreduce.Job:  map 100% reduce 71%
19/10/19 01:00:50 INFO mapreduce.Job:  map 100% reduce 86%
19/10/19 01:00:51 INFO mapreduce.Job:  map 100% reduce 100%
19/10/19 01:00:51 INFO mapreduce.Job: Job job_1571440208858_0020 completed successfully
19/10/19 01:00:51 INFO mapreduce.Job: Counters: 51
        File System Counters
                FILE: Number of bytes read=17903242
                FILE: Number of bytes written=41724778
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=3970954
                HDFS: Number of bytes written=5999
                HDFS: Number of read operations=98
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=21
        Job Counters
                Killed map tasks=1
                Killed reduce tasks=1
                Launched map tasks=21
                Launched reduce tasks=7
                Data-local map tasks=21
                Total time spent by all maps in occupied slots (ms)=470898
                Total time spent by all reduces in occupied slots (ms)=140877
                Total time spent by all map tasks (ms)=156966
                Total time spent by all reduce tasks (ms)=46959
                Total vcore-milliseconds taken by all map tasks=156966
```

```
                Total vcore-milliseconds taken by all map tasks=156966
                Total vcore-milliseconds taken by all reduce tasks=46959
                Total megabyte-milliseconds taken by all map tasks=482199552
                Total megabyte-milliseconds taken by all reduce tasks=144258048
        Map-Reduce Framework
                Map input records=10998
                Map output records=759810
                Map output bytes=16383580
                Map output materialized bytes=17904082
                Input split bytes=2099
                Combine input records=0
                Combine output records=0
                Reduce input groups=232431
                Reduce shuffle bytes=17904082
                Reduce input records=759810
                Reduce output records=105
                Spilled Records=1519620
                Shuffled Maps =147
                Failed Shuffles=0
                Merged Map outputs=147
                GC time elapsed (ms)=5861
                CPU time spent (ms)=66730
                Physical memory (bytes) snapshot=13179662336
                Virtual memory (bytes) snapshot=123379810304
                Total committed heap usage (bytes)=12258902016
        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=3968855
        File Output Format Counters
                Bytes Written=5999
19/10/19 01:00:51 INFO streaming.StreamJob: Output directory: shingles03
yunke_zhu@cluster-dade-m:~/quiz3/test/3_3$
yunke_zhu@cluster-dade-m:~/quiz3/test/3_3$
yunke_zhu@cluster-dade-m:~/quiz3/test/3_3$ ls
cal_similarity.py  each_statistics.sh  mapper.py  original  reducer.py  run.sh  total_pair_statistics.sh
yunke_zhu@cluster-dade-m:~/quiz3/test/3_3$ ./total_pair_statistics.sh shingles03
rm: cannot remove 'final_statistics.log': No such file or directory
Most similarity is clinton and obama, which their similarity is:   0.1700
Least similarity is reagan and gwbush, which their similarity is:   0.1386
yunke_zhu@cluster-dade-m:~/quiz3/test/3_3$
```

question d).

```
yunke_zhu@cluster-dade-m:~/quiz3/test/3_4$ ./run.sh shingles04
Deleted shingles04
packageJobJar: [] [/usr/lib/hadoop-mapreduce/hadoop-streaming-2.9.2.jar] /tmp/streamjob5302851374959615372.jar tmpDir=null
19/10/19 01:06:24 INFO client.RMProxy: Connecting to ResourceManager at cluster-dade-m/10.128.0.10:8032
19/10/19 01:06:24 INFO client.AHSProxy: Connecting to Application History server at cluster-dade-m/10.128.0.10:10200
19/10/19 01:06:24 INFO client.RMProxy: Connecting to ResourceManager at cluster-dade-m/10.128.0.10:8032
19/10/19 01:06:24 INFO client.AHSProxy: Connecting to Application History server at cluster-dade-m/10.128.0.10:10200
19/10/19 01:06:25 INFO mapred.FileInputFormat: Total input files to process : 5
19/10/19 01:06:25 INFO mapreduce.JobSubmitter: number of splits:21
19/10/19 01:06:25 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system
19/10/19 01:06:25 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1571440208858_0021
19/10/19 01:06:25 INFO impl.YarnClientImpl: Submitted application application_1571440208858_0021
19/10/19 01:06:25 INFO mapreduce.Job: The url to track the job: http://cluster-dade-m:8088/proxy/application_1571440208858_0021/
19/10/19 01:06:25 INFO mapreduce.Job: Running job: job_1571440208858_0021
19/10/19 01:06:33 INFO mapreduce.Job: Job job_1571440208858_0021 running in uber mode : false
19/10/19 01:06:33 INFO mapreduce.Job:  map 0% reduce 0%
19/10/19 01:06:43 INFO mapreduce.Job:  map 14% reduce 0%
19/10/19 01:06:44 INFO mapreduce.Job:  map 29% reduce 0%
19/10/19 01:06:45 INFO mapreduce.Job:  map 33% reduce 0%
19/10/19 01:06:52 INFO mapreduce.Job:  map 43% reduce 0%
19/10/19 01:06:53 INFO mapreduce.Job:  map 62% reduce 0%
19/10/19 01:06:54 INFO mapreduce.Job:  map 67% reduce 0%
19/10/19 01:06:59 INFO mapreduce.Job:  map 76% reduce 0%
19/10/19 01:07:01 INFO mapreduce.Job:  map 81% reduce 0%
19/10/19 01:07:02 INFO mapreduce.Job:  map 95% reduce 0%
19/10/19 01:07:03 INFO mapreduce.Job:  map 100% reduce 0%
19/10/19 01:07:10 INFO mapreduce.Job:  map 100% reduce 29%
19/10/19 01:07:13 INFO mapreduce.Job:  map 100% reduce 57%
19/10/19 01:07:14 INFO mapreduce.Job:  map 100% reduce 86%
19/10/19 01:07:15 INFO mapreduce.Job:  map 100% reduce 100%
19/10/19 01:07:15 INFO mapreduce.Job: Job job_1571440208858_0021 completed successfully
19/10/19 01:07:15 INFO mapreduce.Job: Counters: 51
        File System Counters
                FILE: Number of bytes read=21376633
                FILE: Number of bytes written=48671560
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=3970954
                HDFS: Number of bytes written=6014
                HDFS: Number of read operations=98
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=21
        Job Counters
                Killed map tasks=1
                Killed reduce tasks=1
                Launched map tasks=21
                Launched reduce tasks=7
                Data-local map tasks=21
                Total time spent by all maps in occupied slots (ms)=504522
                Total time spent by all reduces in occupied slots (ms)=140583
                Total time spent by all map tasks (ms)=168174
                Total time spent by all reduce tasks (ms)=46861
                Total vcore-milliseconds taken by all map tasks=168174
                Total vcore-milliseconds taken by all reduce tasks=46861
                Total megabyte-milliseconds taken by all map tasks=516630528
                Total megabyte-milliseconds taken by all reduce tasks=143956992
        Map-Reduce Framework
                Map input records=10998
                Map output records=759810
```

```
                FILE: Number of bytes written=48671560
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=3970954
                HDFS: Number of bytes written=6014
                HDFS: Number of read operations=98
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=21
        Job Counters
                Killed map tasks=1
                Killed reduce tasks=1
                Launched map tasks=21
                Launched reduce tasks=7
                Data-local map tasks=21
                Total time spent by all maps in occupied slots (ms)=504522
                Total time spent by all reduces in occupied slots (ms)=140583
                Total time spent by all map tasks (ms)=168174
                Total time spent by all reduce tasks (ms)=46861
                Total vcore-milliseconds taken by all map tasks=168174
                Total vcore-milliseconds taken by all reduce tasks=46861
                Total megabyte-milliseconds taken by all map tasks=516630528
                Total megabyte-milliseconds taken by all reduce tasks=143956992
        Map-Reduce Framework
                Map input records=10998
                Map output records=759810
                Map output bytes=19856966
                Map output materialized bytes=21377473
                Input split bytes=2099
                Combine input records=0
                Combine output records=0
                Reduce input groups=532857
                Reduce shuffle bytes=21377473
                Reduce input records=759810
                Reduce output records=105
                Spilled Records=1519620
                Shuffled Maps =147
                Failed Shuffles=0
                Merged Map outputs=147
                GC time elapsed (ms)=5553
                CPU time spent (ms)=73560
                Physical memory (bytes) snapshot=13211164672
                Virtual memory (bytes) snapshot=123366350848
                Total committed heap usage (bytes)=12180783104
        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=3968855
        File Output Format Counters
                Bytes Written=6014
19/10/19 01:07:15 INFO streaming.StreamJob: Output directory: shingles04
yunke_zhu@cluster-dade-m:~/quiz3/test/3_4$ ./total_pair_statistics.sh shingles04
Most similarity is clinton and obama, which their similarity is:   0.0565
Least similarity is reagan and gwbush, which their similarity is:   0.0400
yunke_zhu@cluster-dade-m:~/quiz3/test/3_4$
```

The result is that all methods show that **Clinton** and **Obama**'s speech are most similar. And **Reagan** and **Gwbush** are least similar through 3 partition(a,c,d). And question b shows **Bush** and **Obama** are least similar.

2.Code

mapper3_1.py

```
#!/usr/bin/env python
import sys
import os
import json
import re
import subprocess
class Mapper:


    def MAP(self):
#--- get all lines from stdin ---


        #n = 3

        #filepath = "aa/bb/cc/real.rar"
        filepath = os.environ["mapreduce_map_input_file"]


        filepath = filepath.split("/")[-1]
        #pattern = re.compile("[a-zA-Z][a-zA-Z0-9]*")
        #line = sys.stdin.readline()
        for line in sys.stdin:
               #--- remove leading and trailing whitespace---
            line = line.strip()
                                #filepath = "123"
                                    #--- split the line into words ---
            words = line.split()
                                                    #--- output tuples [word, 1] in tab-
            line = re.sub(r'[^\w]','',line)
            for i in range(len(line)-1):

                word = line[i:i+4]
                #if 'tar' not in filepath:
                #    print "word: %s\t filepath: %s" %(word,filepath)
                print '%s\t%s' % (word.lower(), filepath)

            #print "111_111"
            #print "www_temp.rar"
```

```
            #print "qqq_tt.rar"
'''
        for lines in fl:
            words = lines.split()


            for i in range(len(words)-1):

                #words[i] = pattern.findall(words[i])
                #words[i+1] = pattern.findall(words[i+1])
                words[i] = re.sub(r'[^\w]','', words[i])
                words[i+1] = re.sub(r'[^\w]','', words[i+1])

                print (words[i].lower() + "|" + words[i+1].lower() + " " + "1")
'''
```

mapper3_2.py

```
#!/usr/bin/env python
import sys
import os
import json
import re
import subprocess
class Mapper:


    def MAP(self):
#--- get all lines from stdin ---


        #n = 3

        #filepath = "aa/bb/cc/real.rar"
        filepath = os.environ["mapreduce_map_input_file"]


        filepath = filepath.split("/")[-1]
        #pattern = re.compile("[a-zA-Z][a-zA-Z0-9]*")
        #line = sys.stdin.readline()
        for line in sys.stdin:
               #--- remove leading and trailing whitespace---
            line = line.strip()
                            #filepath = "123"
```

```
                                #--- split the line into words ---
            words = line.split()
                                                    #--- output tuples [word, 1] in tab-
            line = re.sub(r'[^\w]','',line)
            for i in range(len(line)-1):

                word = line[i:i+7]
                #if 'tar' not in filepath:
                #    print "word: %s\t filepath: %s" %(word,filepath)
                print '%s\t%s' % (word.lower(), filepath)

            #print "111_111"
            #print "www_temp.rar"
            #print "qqq_tt.rar"
'''
        for lines in fl:
            words = lines.split()


            for i in range(len(words)-1):

                #words[i] = pattern.findall(words[i])
                #words[i+1] = pattern.findall(words[i+1])
                words[i] = re.sub(r'[^\w]','', words[i])
                words[i+1] = re.sub(r'[^\w]','', words[i+1])

                print (words[i].lower() + "|" + words[i+1].lower() + " " + "1")
'''
```

mapper3_3.py

```
#!/usr/bin/env python
import sys
import os
import json
import re
import subprocess
class Mapper:


    def MAP(self):
#--- get all lines from stdin ---
```

```
        #n = 3


        filepath = os.environ["mapreduce_map_input_file"]
        #filepath = "aa/bb/cc/ss/real.rar"
        filepath = filepath.split("/")[-1]
        #pattern = re.compile("[a-zA-Z][a-zA-Z0-9]*")
        #line = sys.stdin.readline()
        #print "%s" % filepath

        #print "%s\t%s" % ("beor","ss.rar")

        for line in sys.stdin:
               #--- remove leading and trailing whitespace---
            line = line.strip()
                            #filepath = "123"
                                #--- split the line into words ---
            words = line.split()

            #--- output tuples [word, 1] in tab-delimited format---
            for i in range(len(words)-2):
                word1 = re.sub(r'[^\w]','', words[i]).lower()
                word2 = re.sub(r'[^\w]','', words[i+1]).lower()

                print '%s\t%s' % (word1+word2, filepath)
'''
        for lines in fl:
            words = lines.split()


            for i in range(len(words)-1):

                #words[i] = pattern.findall(words[i])
                #words[i+1] = pattern.findall(words[i+1])
                words[i] = re.sub(r'[^\w]','', words[i])
                words[i+1] = re.sub(r'[^\w]','', words[i+1])

                print (words[i].lower() + "|" + words[i+1].lower() + " " + "1")
'''



exp = Mapper()
exp.MAP()
```

mapper3_4.py

```
#!/usr/bin/env python
import sys
import os
import json
import re
import subprocess
class Mapper:


    def MAP(self):
#--- get all lines from stdin ---


        #n = 3


        filepath = os.environ["mapreduce_map_input_file"]
        #filepath = "aa/bb/cc/ss/real.rar"
        filepath = filepath.split("/")[-1]
        #pattern = re.compile("[a-zA-Z][a-zA-Z0-9]*")
        #line = sys.stdin.readline()
        #print "%s" % filepath

        #print "%s\t%s" % ("beor","ss.rar")

        for line in sys.stdin:
               #--- remove leading and trailing whitespace---
            line = line.strip()
                                #filepath = "123"
                                    #--- split the line into words ---
            words = line.split()

            #--- output tuples [word, 1] in tab-delimited format---
            for i in range(len(words)-2):
                word1 = re.sub(r'[^\w]','', words[i]).lower()
                word2 = re.sub(r'[^\w]','', words[i+1]).lower()
                word3 = re.sub(r'[^\w]','', words[i+2]).lower()
                print '%s\t%s' % (word1+word2+word3, filepath)
'''
        for lines in fl:
            words = lines.split()
```

```
            for i in range(len(words)-1):

                #words[i] = pattern.findall(words[i])
                #words[i+1] = pattern.findall(words[i+1])
                words[i] = re.sub(r'[^\w]','', words[i])
                words[i+1] = re.sub(r'[^\w]','', words[i+1])

                print (words[i].lower() + "|" + words[i+1].lower() + " " + "1")
'''



exp = Mapper()
exp.MAP()
```

reducer3_1.py

```
#!/usr/bin/env python
import sys

d = dict()
for line in sys.stdin:
    words = line.strip().split('\t')
    pair = words[0]
    filename = words[1]


    if filename in d.keys():
        d[filename].add(pair)
        #d[filename]="replaced"
    else:
        s = set()
        d[filename] = s
        d[filename].add(pair)

        #d[filename] = "new"
    #print '%s\t%s' %(pair,filename)
if 'speeches' in d:
    del d['speeches']




cp = d.copy()

for x in d:
    del cp[x]
    for y in cp:
        if y == x:
            continue
        else:
            intersect = len((d[x]) & d[y])
            print 'file1 : %s and file2 : %s\t intersect number is %d' % (x,y , interse
    print "size of file %s is %d " % (x,len(d[x]))
```

reducer3_2.py

```
#!/usr/bin/env python
import sys

d = dict()
for line in sys.stdin:
    words = line.strip().split('\t')
    pair = words[0]
    filename = words[1]



    if filename in d.keys():
        d[filename].add(pair)
        #d[filename]="replaced"
    else:
        s = set()
        d[filename] = s
        d[filename].add(pair)

        #d[filename] = "new"
    #print '%s\t%s' %(pair,filename)
if 'speeches' in d:
    del d['speeches']







cp = d.copy()

for x in d:
    del cp[x]
    for y in cp:
        if y == x:
            continue
        else:
            intersect = len((d[x]) & d[y])
            print 'file1 : %s and file2 : %s\t intersect number is %d' % (x,y , interse
    print "size of file %s is %d " % (x,len(d[x]))
```

reducer3_3.py

```
#!/usr/bin/env python
import sys
```

```
d = dict()
for line in sys.stdin:
    words = line.strip().split('\t')

    if len(words) < 2:
        continue


    pair = words[0]


    filename = words[1]


    if filename in d.keys():
        d[filename].add(pair)
        #d[filename]="replaced"
    else:
        s = set()
        d[filename] = s
        d[filename].add(pair)

        #d[filename] = "new"
    #print '%s\t%s' %(pair,filename)
if 'speeches' in d:
    del d['speeches']




cp = d.copy()

for x in d:
    del cp[x]
    for y in cp:
        if y == x:
            continue
        else:
            intersect = len((d[x]) & d[y])
            print 'file1 : %s and file2 : %s\t intersect number is %d' % (x,y , interse
    print "size of file %s is %d " % (x,len(d[x]))
```

reducer3_4.py

```
#!/usr/bin/env python
import sys

d = dict()
for line in sys.stdin:
    words = line.strip().split('\t')

    if len(words) < 2:
        continue


    pair = words[0]


    filename = words[1]


    if filename in d.keys():
        d[filename].add(pair)
        #d[filename]="replaced"
    else:
        s = set()
        d[filename] = s
        d[filename].add(pair)

        #d[filename] = "new"
    #print '%s\t%s' %(pair,filename)
if 'speeches' in d:
    del d['speeches']




cp = d.copy()

for x in d:
    del cp[x]
    for y in cp:
        if y == x:
            continue
        else:
            intersect = len((d[x]) & d[y])
            print 'file1 : %s and file2 : %s\t intersect number is %d' % (x,y , interse
    print "size of file %s is %d " % (x,len(d[x]))
```

run.sh

```
outputfile=$1
hadoop fs -rm -r $outputfile
hadoop jar /usr/lib/hadoop-mapreduce/hadoop-streaming.jar -files ./mapper.py,./reduce
r.py -mapper ./mapper.py -reducer ./reducer.py  -input /user/five-books -output $outp
utfile
```

each_statistics.sh

```
filename=$3
s1="hadoop fs -cat $3\/\* | grep 'size of file $1.tar' | awk '{s+=\$6} END {print s}'
"
s2="hadoop fs -cat $3\/\* | grep 'size of file $2.tar' | awk '{s+=\$6} END {print s}'
"
s3="hadoop fs -cat $3\/\* | grep 'intersect' | grep -w $1 | grep -w $2 | awk '{s+=\$1
1} END {print s}'"
x1=$(eval $s1)
x2=$(eval $s2)
x3=$(eval $s3)
bottom=$(($x1 + $x2 - $x3))
#echo "$(($x3) // ($x1 + $x2 - $x3))"
echo "$1 $2 $x3 $bottom"
```

total_pair_statistics

```
filename=$1
declare -a arr=("reagan" "bush" "clinton" "gwbush" "obama")
rm final_statistics.log
## now loop through the above array
for i in "${arr[@]}"
do
        for j in "${arr[@]}"
        do
                if [ "$i" != "$j" ]
                then
                        str="bash each_statistics.sh $i $j $filename >> final_statistic
                        eval $str
                fi
              # or do whatever with individual element of the array
        done
done

cat final_statistics.log | ./cal_similarity.py
```

cal_similarity.y

```
#!/usr/bin/env python
import sys
import os
import json
import re
import subprocess
class Mapper:


    def MAP(self):
#--- get all lines from stdin ---

        maxVal = 0.0
        minVal = 1.0
        for line in sys.stdin:
            line = line.strip()
            words = line.split()

            val = float(words[2]) / float(words[3])

            if val < minVal:
                minVal = val
                outMin = words[0] + " and " + words[1]
            if val > maxVal:
                maxVal = val
                outMax = words[0] + " and " + words[1]


        print "Most similarity is %s, which their similarity is: %8.4f" %(outMax, maxVa
        print "Least similarity is %s, which their similarity is: %8.4f" %(outMin,minVa







exp = Mapper()
exp.MAP()
```

3. Conclusion

1. Mappers are taking charge of the input and transfer the input as "key = shingles; value = filename" to the reducer, where filename represents the president's name.

2. Reducers will store the key pair into the set. Since the **same keys will go to the same reducer**, thus we can calculate the intersection between each two presidents. Meanwhile we can get the total shingles for each president.
3. After each reducer finishes their job, all the pata in part-0000's are like below:

```
yunke_zhu@cluster-dade-m:~$ hadoop fs -cat shingles01/part-00000
file1 : bush.tar and file2 : obama.tar.gz        intersect number is 4309
file1 : bush.tar and file2 : reagan.tar  intersect number is 4413
file1 : bush.tar and file2 : gwbush.tar  intersect number is 3940
file1 : bush.tar and file2 : clinton.tar         intersect number is 4116
size of file bush.tar is 5256
file1 : obama.tar.gz and file2 : reagan.tar      intersect number is 5108
file1 : obama.tar.gz and file2 : gwbush.tar      intersect number is 4478
file1 : obama.tar.gz and file2 : clinton.tar     intersect number is 4735
size of file obama.tar.gz is 6562
file1 : clinton.tar and file2 : reagan.tar       intersect number is 4862
file1 : clinton.tar and file2 : gwbush.tar       intersect number is 4246
size of file clinton.tar is 6030
file1 : gwbush.tar and file2 : reagan.tar        intersect number is 4516
size of file gwbush.tar is 5526
size of file reagan.tar is 7110
yunke_zhu@cluster-dade-m:~$
```

4. Then the formula of union between two presidents(A and B) is:|A| + |B| - |A & B|(intersection) So we only need to collect all the output and do the calculations.
5. I write several scripts in order to make my assembly line more effectively, where:

   run.sh: used for running the hadoop hdfs and restore the result to each shingles **each_statistics.sh**: used for combining two presidents' intersection and union, the sample result is shown above:

```
yunke_zhu@cluster-dade-m:~/quiz3/test/3_1$ ./each_statistics.sh reagan gwbush shingles01
reagan gwbush 31244 56962
yunke_zhu@cluster-dade-m:~/quiz3/test/3_1$ ./each_statistics.sh clinton obama shingles01
clinton obama 33162 55337
yunke_zhu@cluster-dade-m:~/quiz3/test/3_1$
```

   total_pair_statistics.sh: This is a loop script to output all the two pair of presidents' intersection and union and then record the results in **final_statistics.log** and call **cal_similarity** to calculate the maximum similarity and minimum similarity. The sample **cal_similarity** output is shown above:

```
yunke_zhu@cluster-dade-m:~/quiz3/test/3_1$ cat final_statistics.log
reagan bush 30845 55831
reagan clinton 33843 58351
reagan gwbush 31244 56962
reagan obama 35673 60240
bush reagan 30845 55831
bush clinton 28959 50303
bush gwbush 27285 47989
bush obama 30008 52973
clinton reagan 33843 58351
clinton bush 28959 50303
clinton gwbush 29450 51342
clinton obama 33162 55337
gwbush reagan 31244 56962
gwbush bush 27285 47989
gwbush clinton 29450 51342
gwbush obama 31166 53345
obama reagan 35673 60240
obama bush 30008 52973
obama clinton 33162 55337
obama gwbush 31166 53345
yunke_zhu@cluster-dade-m:~/quiz3/test/3_1$
```

4. Command

```
./run.sh -outputfile
```

```
./total_pair_statistics.sh -outputfile
```