BIG DATA ASSIGNMENT-2
UE18CS322

NAME : SAIPRAKASH L SHETTY

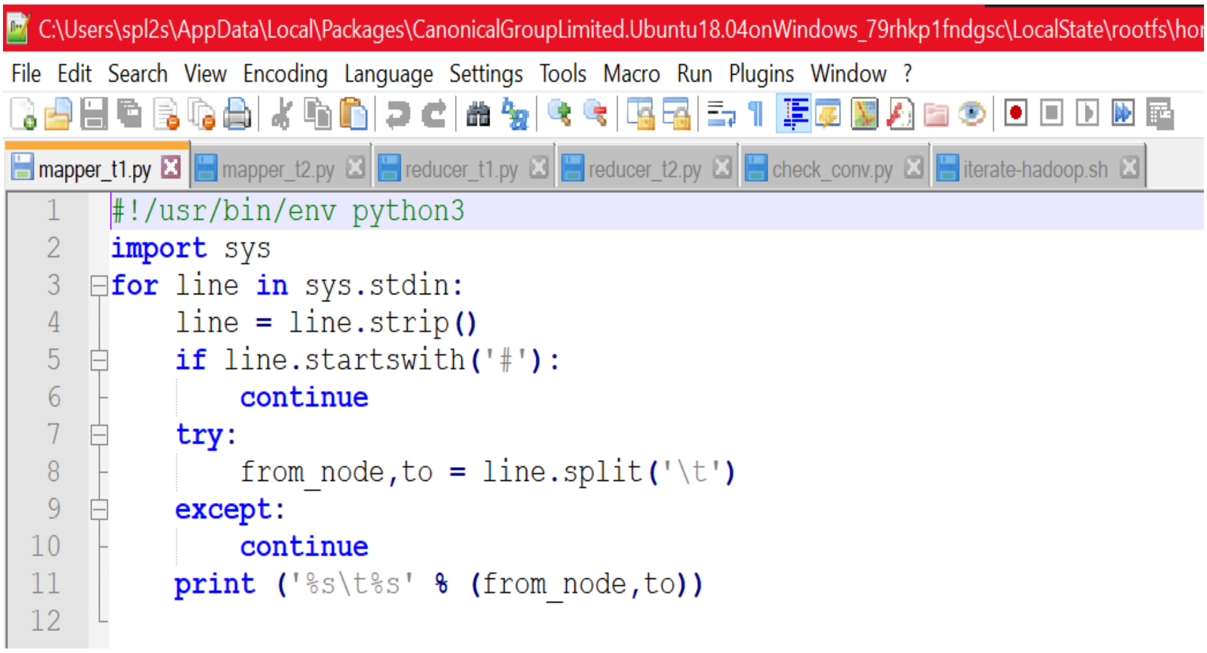
Title :

Implementation of PageRank Algorithm using Hadoop MapReduce

Task-1 : Creating an adjacency list from the given input file.

1. mapper_t1.py

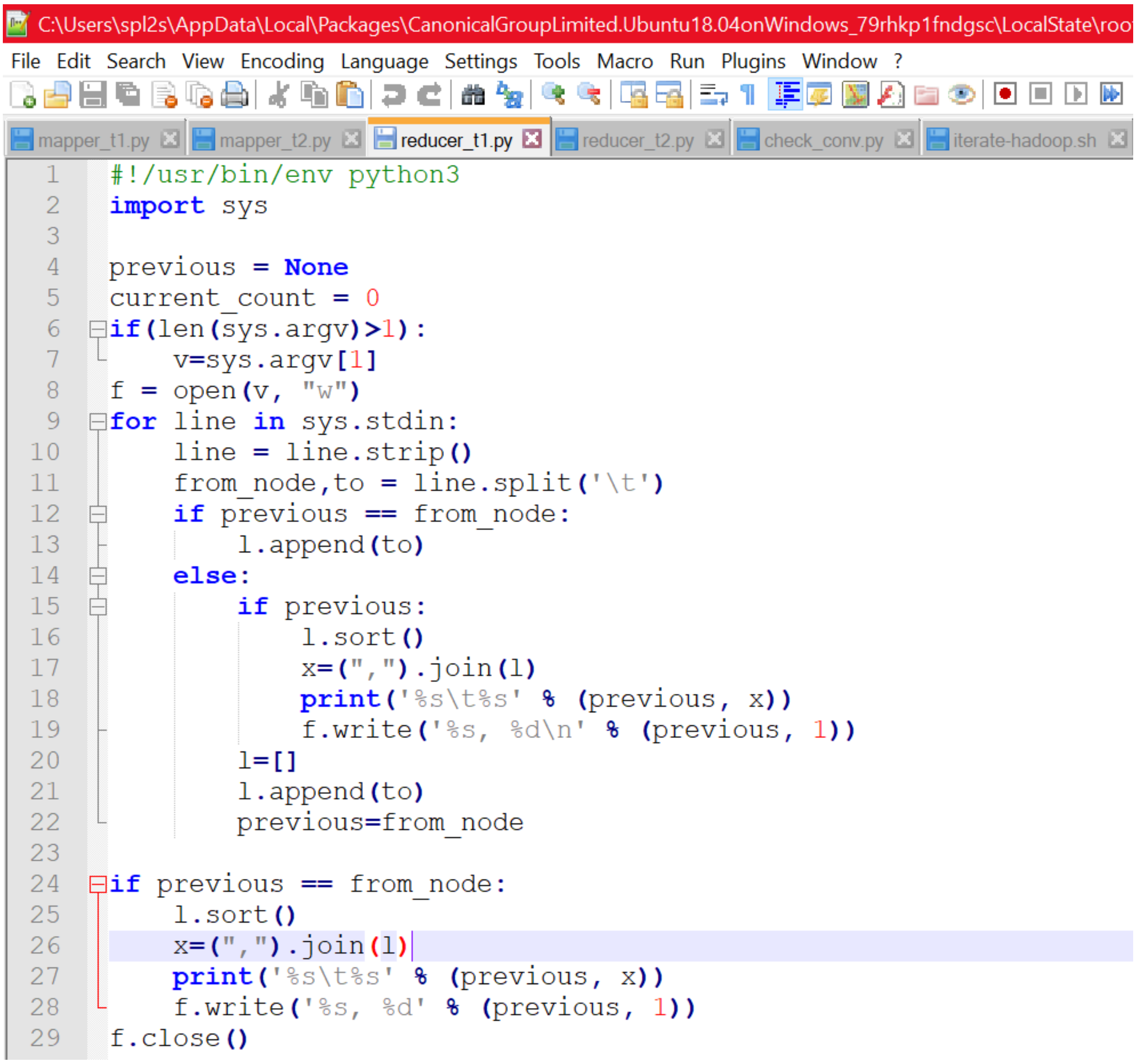
It receives the input file given to us which has the sorted graph with columns as from_node and to_node.


C:\Users\spl2s\AppData\Local\Packages\CanonicalGroupLimited.Ubuntu18.04onWindows_79rhkp1fndgsc\LocalState\rootfs\hor
File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?
mapper_t1.py
mapper_t2.py
reducer_t1.py
reducer_t2.py
check_conv.py
iterate-hadoop.sh

```
#!/usr/bin/env python3
import sys
for line in sys.stdin:
    line = line.strip()
    if line.startswith('#'):
        continue
    try:
        from_node,to = line.split('\t')
    except:
        continue
    print ('%s\t%s' % (from_node,to))

```

2. reducer_t1.py

It receives the output from the first mapper file and appends the to_nodes as a list with from_node. Opens “v” which stores 1 as the initial PageRank for all the nodes.

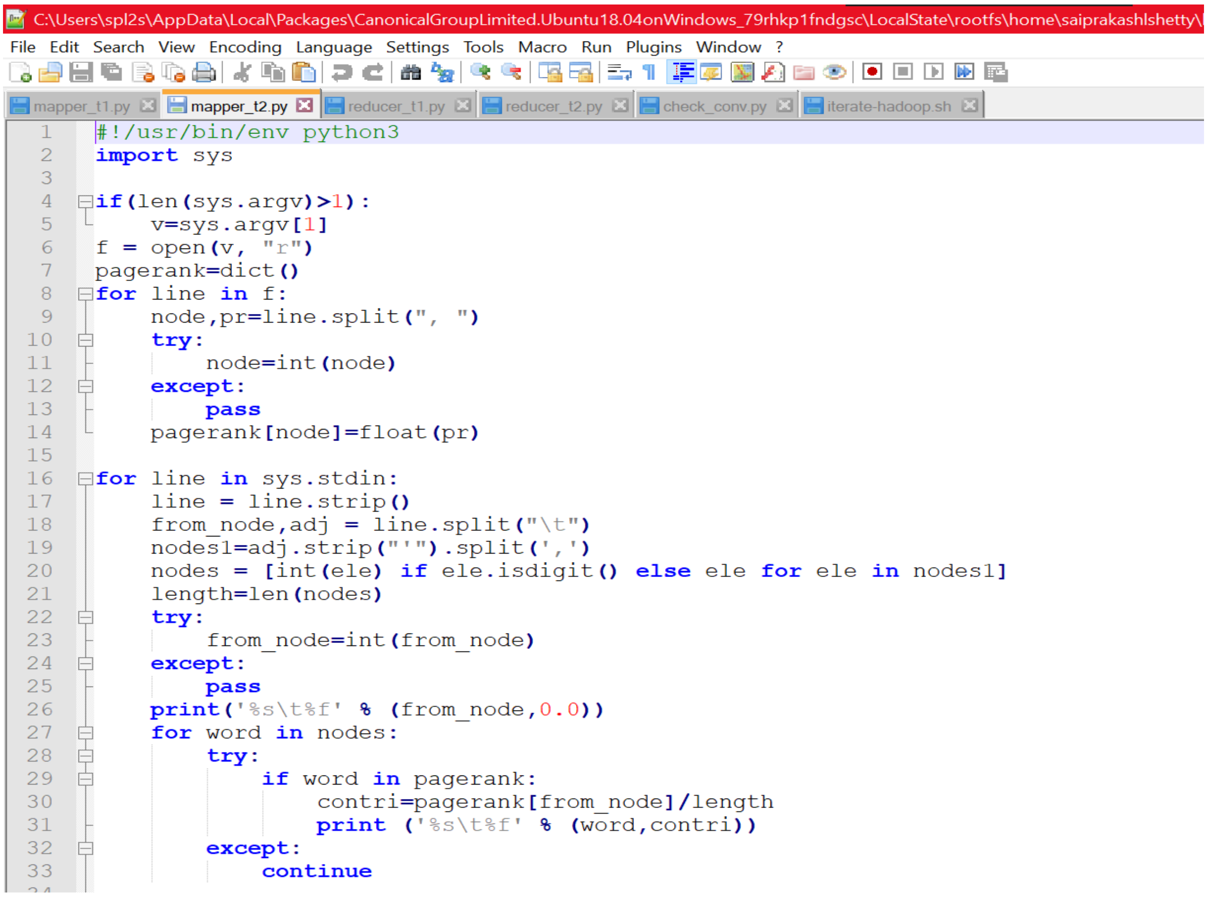

C:\Users\spl2s\AppData\Local\Packages\CanonicalGroupLimited.Ubuntu18.04onWindows_79rhkp1fndgsc\LocalState\roo
File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?
mapper_t1.py
mapper_t2.py
reducer_t1.py
reducer_t2.py
check_conv.py
iterate-hadoop.sh

```
#!/usr/bin/env python3
import sys

previous = None
current_count = 0
if(len(sys.argv)>1):
    v=sys.argv[1]
f = open(v, "w")
for line in sys.stdin:
    line = line.strip()
    from_node,to = line.split('\t')
    if previous == from_node:
        l.append(to)
    else:
        if previous:
            l.sort()
            x=(",").join(l)
            print('%s\t%s' % (previous, x))
            f.write('%s, %d\n' % (previous, 1))
        l=[]
        l.append(to)
        previous=from_node

if previous == from_node:
    l.sort()
    x=(",").join(l)
    print('%s\t%s' % (previous, x))
    f.write('%s, %d' % (previous, 1))
f.close()
```

Task-2 : Mapping and the reducing the initial adjacency list until it converges to the final PageRank

1. mapper_2t.py

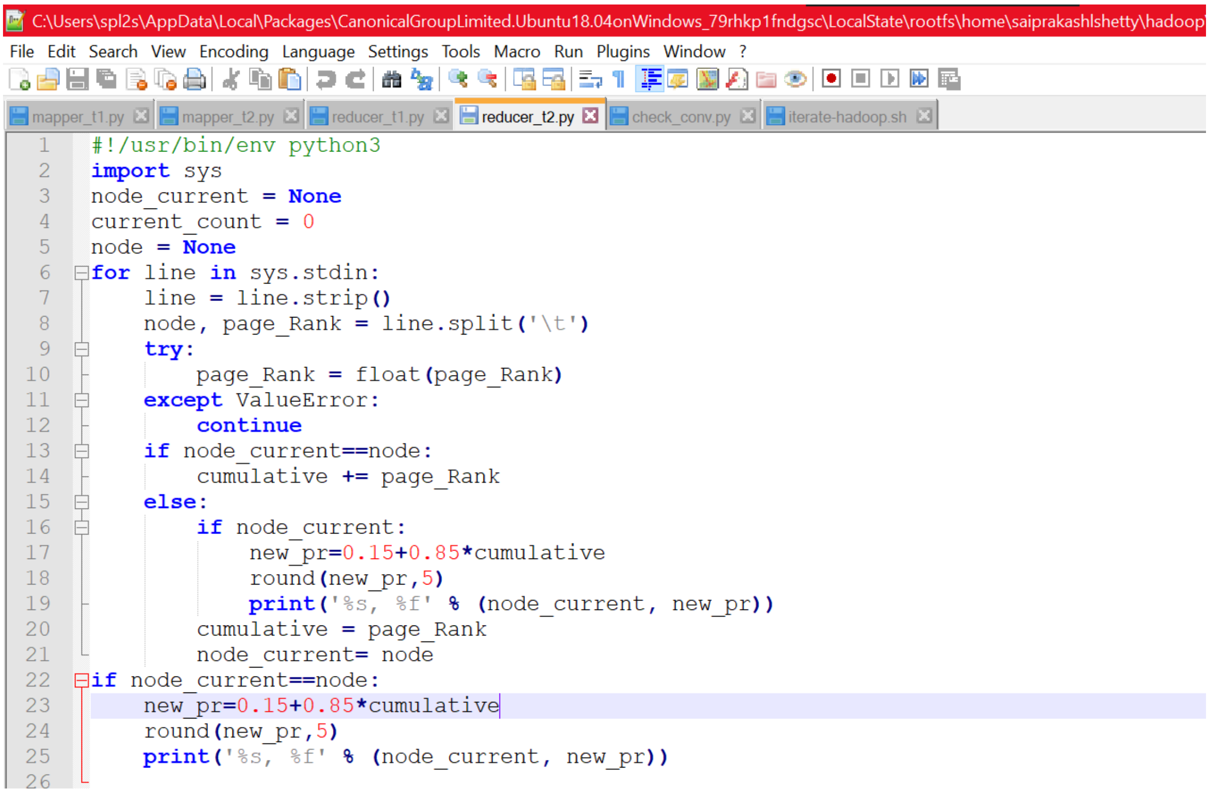
It reads the local file “v” and the adjacency list (output 1) from HDFS.
PageRank formula = 1/n.
where n= no of outgoing links from the from_node w.r.t the length of the adjacency list of each from_node.


C:\Users\spl2s\AppData\Local\Packages\CanonicalGroupLimited.Ubuntu18.04onWindows_79rhkp1fndgsc\LocalState\rootfs\home\saiprakashlshetty\
File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?
mapper_t1.py
mapper_t2.py
reducer_t1.py
reducer_t2.py
check_conv.py
iterate-hadoop.sh

```python
#!/usr/bin/env python3
import sys

if(len(sys.argv)>1):
    v=sys.argv[1]
f = open(v, "r")
pagerank=dict()
for line in f:
    node,pr=line.split(", ")
    try:
        node=int(node)
    except:
        pass
    pagerank[node]=float(pr)

for line in sys.stdin:
    line = line.strip()
    from_node,adj = line.split("\t")
    nodes1=adj.strip("'").split(',')
    nodes = [int(ele) if ele.isdigit() else ele for ele in nodes1]
    length=len(nodes)
    try:
        from_node=int(from_node)
    except:
        pass
    print('%s\t%f' % (from_node,0.0))
    for word in nodes:
        try:
            if word in pagerank:
                contri=pagerank[from_node]/length
                print ('%s\t%f' % (word,contri))
        except:
            continue
```

2. reducer_2t.py

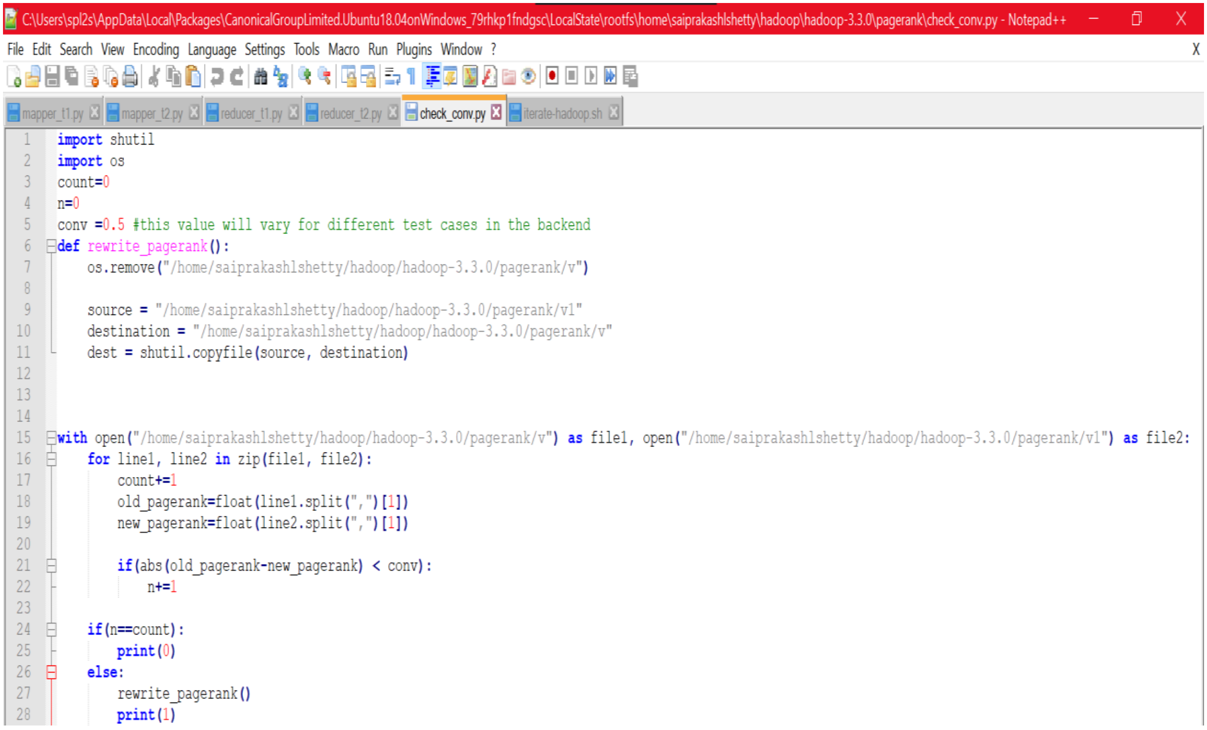
This reducer computes the PageRank by using the power method.


C:\Users\spl2s\AppData\Local\Packages\CanonicalGroupLimited.Ubuntu18.04onWindows_79rhkp1fndgsc\LocalState\rootfs\home\saiprakashlshetty\hadoop
File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?
mapper_t1.py
mapper_t2.py
reducer_t1.py
reducer_t2.py
check_conv.py
iterate-hadoop.sh

```python
#!/usr/bin/env python3
import sys
node_current = None
current_count = 0
node = None
for line in sys.stdin:
    line = line.strip()
    node, page_Rank = line.split('\t')
    try:
        page_Rank = float(page_Rank)
    except ValueError:
        continue
    if node_current==node:
        cumulative += page_Rank
    else:
        if node_current:
            new_pr=0.15+0.85*cumulative
            round(new_pr,5)
            print('%s, %f' % (node_current, new_pr))
        cumulative = page_Rank
        node_current= node
if node_current==node:
    new_pr=0.15+0.85*cumulative
    round(new_pr,5)
    print('%s, %f' % (node_current, new_pr))
```

- check_conv.py

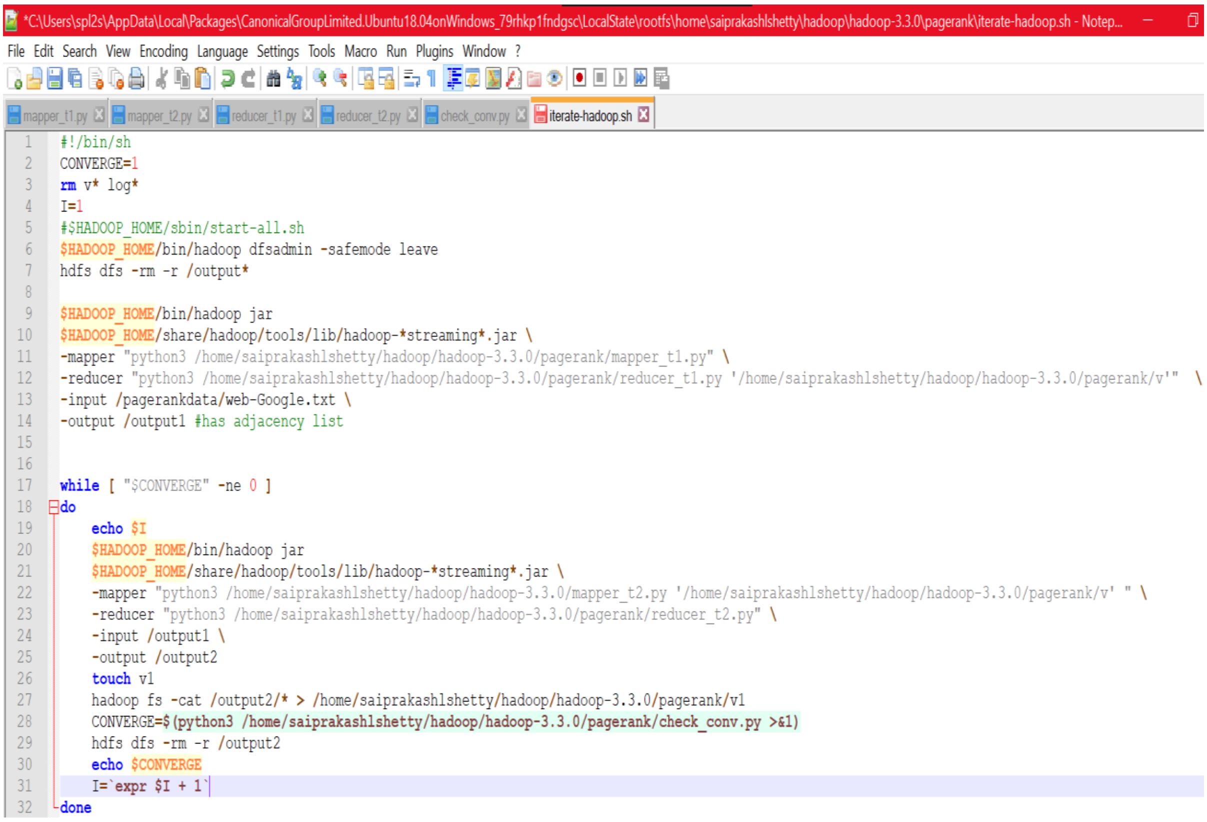
Difference between the old PageRank and new PageRank is calculated, it is also used to copy the output from hdfs into the “v”.


C:\Users\spl2s\AppData\Local\Packages\CanonicalGroupLimited.Ubuntu18.04onWindows_79rhkp1fndgsc\LocalState\rootfs\home\saiprakashlshetty\hadoop\hadoop-3.3.0\pagerank\check_conv.py - Notepad++
File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?
mapper_t1.py
mapper_t2.py
reducer_t1.py
reducer_t2.py
check_conv.py
iterate-hadoop.sh

```python
import shutil
import os
count=0
n=0
conv =0.5 #this value will vary for different test cases in the backend
def rewrite_pagerank():
    os.remove("/home/saiprakashlshetty/hadoop/hadoop-3.3.0/pagerank/v")

    source = "/home/saiprakashlshetty/hadoop/hadoop-3.3.0/pagerank/v1"
    destination = "/home/saiprakashlshetty/hadoop/hadoop-3.3.0/pagerank/v"
    dest = shutil.copyfile(source, destination)



with open("/home/saiprakashlshetty/hadoop/hadoop-3.3.0/pagerank/v") as file1, open("/home/saiprakashlshetty/hadoop/hadoop-3.3.0/pagerank/v1") as file2:
    for line1, line2 in zip(file1, file2):
        count+=1
        old_pagerank=float(line1.split(",")[1])
        new_pagerank=float(line2.split(",")[1])

        if(abs(old_pagerank-new_pagerank) < conv):
            n+=1

    if(n==count):
        print(0)
    else:
        rewrite_pagerank()
        print(1)
```

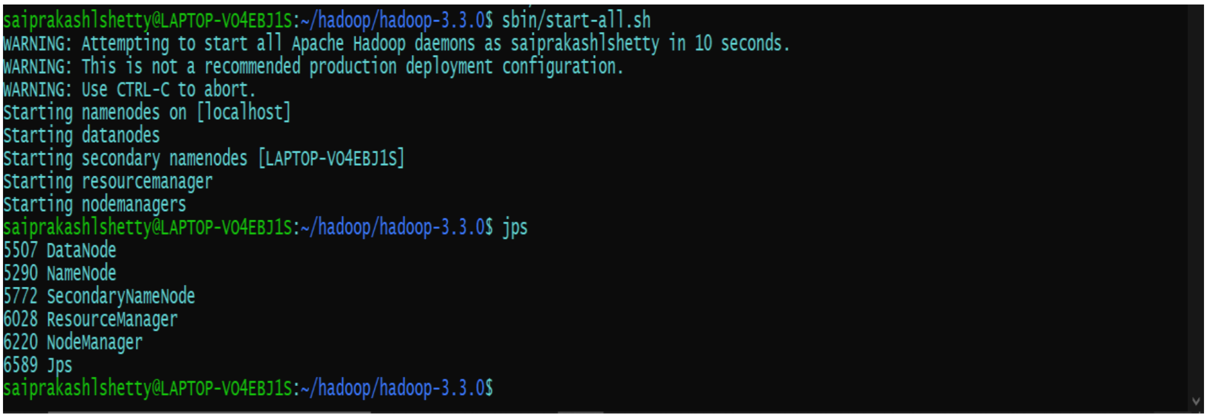
- iterate-hadoop.sh

Runs mapper and reducer files iteratively until the final PageRank converges to a negligible value as written.

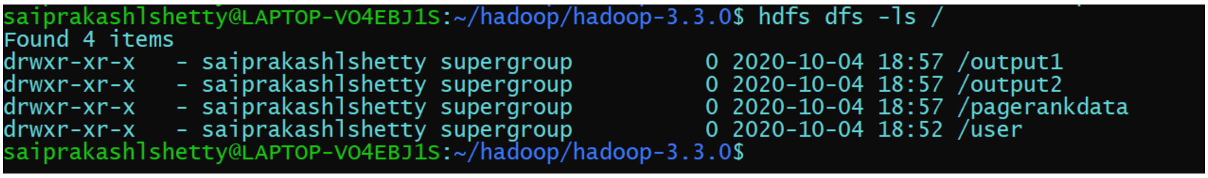

*C:\Users\spl2s\AppData\Local\Packages\CanonicalGroupLimited.Ubuntu18.04onWindows_79rhkp1fndgsc\LocalState\rootfs\home\saiprakashlshetty\hadoop\hadoop-3.3.0\pagerank\iterate-hadoop.sh - Notep...
File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?
mapper_t1.py
mapper_t2.py
reducer_t1.py
reducer_t2.py
check_conv.py
iterate-hadoop.sh

```sh
#!/bin/sh
CONVERGE=1
rm v* log*
I=1
#$HADOOP_HOME/sbin/start-all.sh
$HADOOP_HOME/bin/hadoop dfsadmin -safemode leave
hdfs dfs -rm -r /output*

$HADOOP_HOME/bin/hadoop jar
$HADOOP_HOME/share/hadoop/tools/lib/hadoop-*streaming*.jar \
-mapper "python3 /home/saiprakashlshetty/hadoop/hadoop-3.3.0/pagerank/mapper_t1.py" \
-reducer "python3 /home/saiprakashlshetty/hadoop/hadoop-3.3.0/pagerank/reducer_t1.py '/home/saiprakashlshetty/hadoop/hadoop-3.3.0/pagerank/v'"  \
-input /pagerankdata/web-Google.txt \
-output /output1 #has adjacency list


while [ "$CONVERGE" -ne 0 ]
do
    echo $I
    $HADOOP_HOME/bin/hadoop jar
    $HADOOP_HOME/share/hadoop/tools/lib/hadoop-*streaming*.jar \
    -mapper "python3 /home/saiprakashlshetty/hadoop/hadoop-3.3.0/mapper_t2.py '/home/saiprakashlshetty/hadoop/hadoop-3.3.0/pagerank/v' " \
    -reducer "python3 /home/saiprakashlshetty/hadoop/hadoop-3.3.0/pagerank/reducer_t2.py" \
    -input /output1 \
    -output /output2
    touch v1
    hadoop fs -cat /output2/* > /home/saiprakashlshetty/hadoop/hadoop-3.3.0/pagerank/v1
    CONVERGE=$(python3 /home/saiprakashlshetty/hadoop/hadoop-3.3.0/pagerank/check_conv.py >&1)
    hdfs dfs -rm -r /output2
    echo $CONVERGE
    I=`expr $I + 1`
done
```

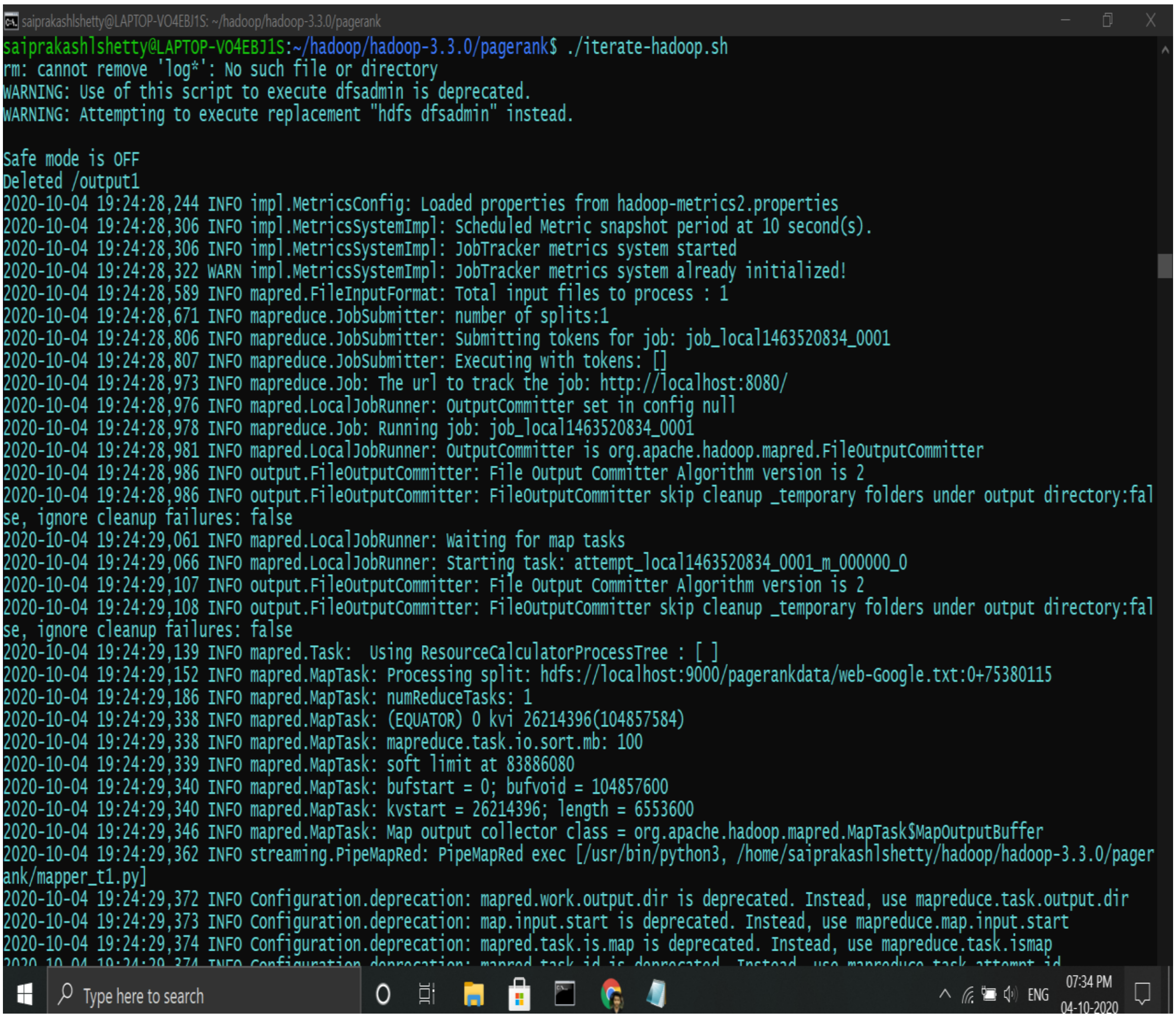
Execution screenshots



```
saiprakashlshetty@LAPTOP-VO4EBJ1S:~/hadoop/hadoop-3.3.0$ sbin/start-all.sh
WARNING: Attempting to start all Apache Hadoop daemons as saiprakashlshetty in 10 seconds.
WARNING: This is not a recommended production deployment configuration.
WARNING: Use CTRL-C to abort.
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [LAPTOP-VO4EBJ1S]
Starting resourcemanager
Starting nodemanagers
saiprakashlshetty@LAPTOP-VO4EBJ1S:~/hadoop/hadoop-3.3.0$ jps
5507 DataNode
5290 NameNode
5772 SecondaryNameNode
6028 ResourceManager
6220 NodeManager
6589 Jps
saiprakashlshetty@LAPTOP-VO4EBJ1S:~/hadoop/hadoop-3.3.0$
```

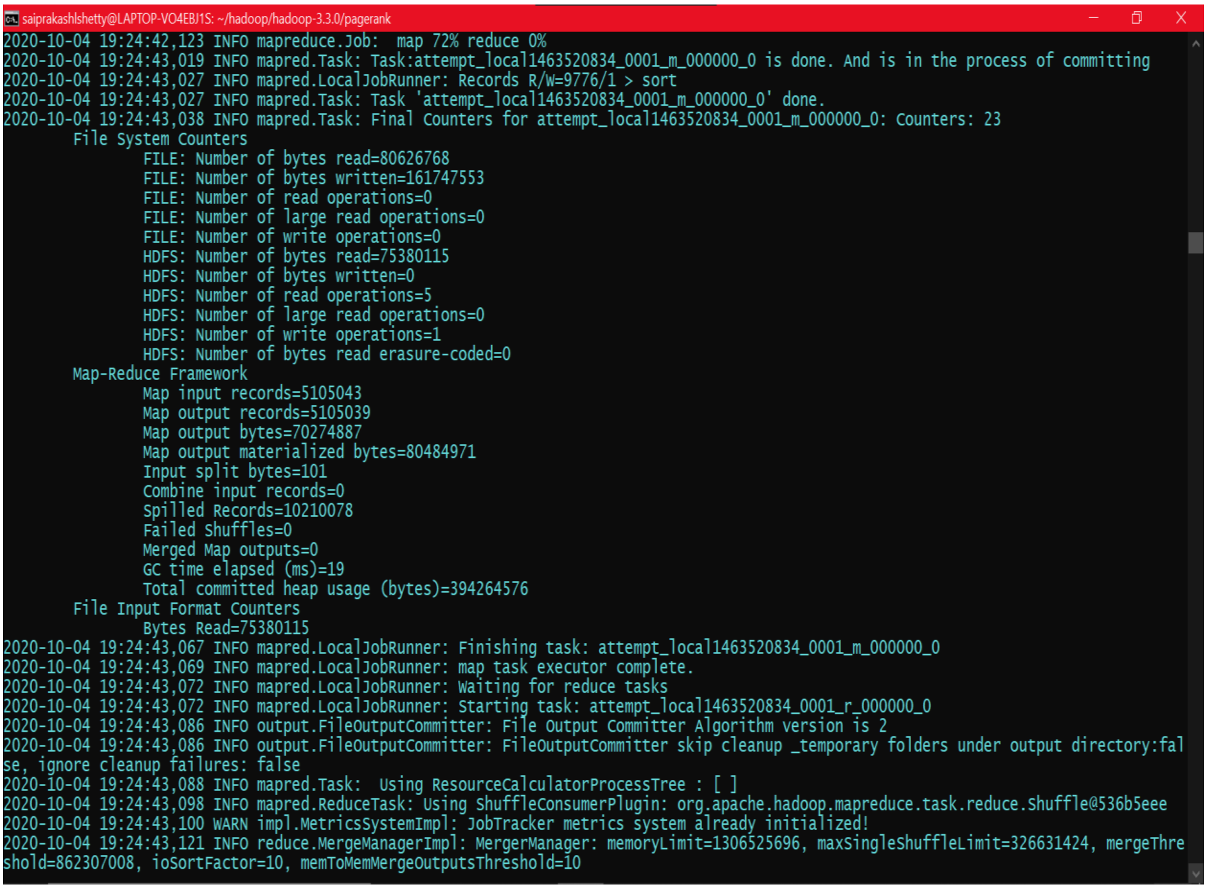


```
saiprakashlshetty@LAPTOP-VO4EBJ1S:~/hadoop/hadoop-3.3.0$ hdfs dfs -ls /
Found 4 items
drwxr-xr-x   - saiprakashlshetty supergroup          0 2020-10-04 18:57 /output1
drwxr-xr-x   - saiprakashlshetty supergroup          0 2020-10-04 18:57 /output2
drwxr-xr-x   - saiprakashlshetty supergroup          0 2020-10-04 18:57 /pagerankdata
drwxr-xr-x   - saiprakashlshetty supergroup          0 2020-10-04 18:52 /user
saiprakashlshetty@LAPTOP-VO4EBJ1S:~/hadoop/hadoop-3.3.0$
```



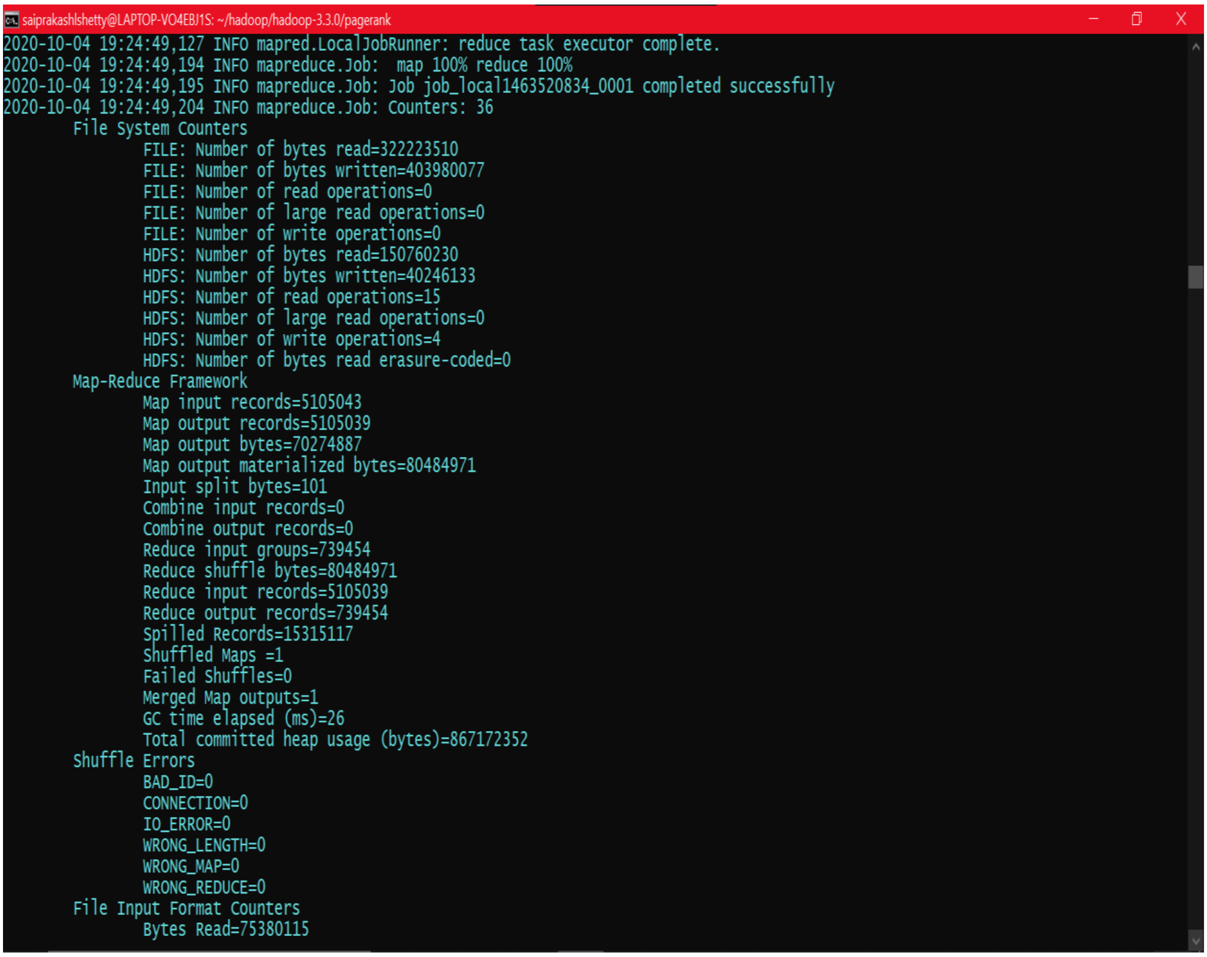
saiprakashlshetty@LAPTOP-VO4EBJ1S: ~/hadoop/hadoop-3.3.0/pagerank
Type here to search
ENG
07:34 PM
04-10-2020

```
saiprakashlshetty@LAPTOP-VO4EBJ1S:~/hadoop/hadoop-3.3.0/pagerank$ ./iterate-hadoop.sh
rm: cannot remove 'log*': No such file or directory
WARNING: Use of this script to execute dfsadmin is deprecated.
WARNING: Attempting to execute replacement "hdfs dfsadmin" instead.

Safe mode is OFF
Deleted /output1
2020-10-04 19:24:28,244 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2020-10-04 19:24:28,306 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
2020-10-04 19:24:28,306 INFO impl.MetricsSystemImpl: JobTracker metrics system started
2020-10-04 19:24:28,322 WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
2020-10-04 19:24:28,589 INFO mapred.FileInputFormat: Total input files to process : 1
2020-10-04 19:24:28,671 INFO mapreduce.JobSubmitter: number of splits:1
2020-10-04 19:24:28,806 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1463520834_0001
2020-10-04 19:24:28,807 INFO mapreduce.JobSubmitter: Executing with tokens: []
2020-10-04 19:24:28,973 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
2020-10-04 19:24:28,976 INFO mapred.LocalJobRunner: OutputCommitter set in config null
2020-10-04 19:24:28,978 INFO mapreduce.Job: Running job: job_local1463520834_0001
2020-10-04 19:24:28,981 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputCommitter
2020-10-04 19:24:28,986 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2020-10-04 19:24:28,986 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:fal
se, ignore cleanup failures: false
2020-10-04 19:24:29,061 INFO mapred.LocalJobRunner: Waiting for map tasks
2020-10-04 19:24:29,066 INFO mapred.LocalJobRunner: Starting task: attempt_local1463520834_0001_m_000000_0
2020-10-04 19:24:29,107 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2020-10-04 19:24:29,108 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:fal
se, ignore cleanup failures: false
2020-10-04 19:24:29,139 INFO mapred.Task:  Using ResourceCalculatorProcessTree : [ ]
2020-10-04 19:24:29,152 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/pagerankdata/web-Google.txt:0+75380115
2020-10-04 19:24:29,186 INFO mapred.MapTask: numReduceTasks: 1
2020-10-04 19:24:29,338 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
2020-10-04 19:24:29,338 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
2020-10-04 19:24:29,339 INFO mapred.MapTask: soft limit at 83886080
2020-10-04 19:24:29,340 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
2020-10-04 19:24:29,340 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
2020-10-04 19:24:29,346 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
2020-10-04 19:24:29,362 INFO streaming.PipeMapRed: PipeMapRed exec [/usr/bin/python3, /home/saiprakashlshetty/hadoop/hadoop-3.3.0/pager
ank/mapper_t1.py]
2020-10-04 19:24:29,372 INFO Configuration.deprecation: mapred.work.output.dir is deprecated. Instead, use mapreduce.task.output.dir
2020-10-04 19:24:29,373 INFO Configuration.deprecation: map.input.start is deprecated. Instead, use mapreduce.map.input.start
2020-10-04 19:24:29,374 INFO Configuration.deprecation: mapred.task.is.map is deprecated. Instead, use mapreduce.task.ismap
2020-10-04 19:24:29,374 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead, use mapreduce.task.attempt.id
```



saiprakashlshetty@LAPTOP-VO4EBJ1S: ~/hadoop/hadoop-3.3.0/pagerank

```
2020-10-04 19:24:42,123 INFO mapreduce.Job:  map 72% reduce 0%
2020-10-04 19:24:43,019 INFO mapred.Task: Task:attempt_local1463520834_0001_m_000000_0 is done. And is in the process of committing
2020-10-04 19:24:43,027 INFO mapred.LocalJobRunner: Records R/W=9776/1 > sort
2020-10-04 19:24:43,027 INFO mapred.Task: Task 'attempt_local1463520834_0001_m_000000_0' done.
2020-10-04 19:24:43,038 INFO mapred.Task: Final Counters for attempt_local1463520834_0001_m_000000_0: Counters: 23
        File System Counters
                FILE: Number of bytes read=80626768
                FILE: Number of bytes written=161747553
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=75380115
                HDFS: Number of bytes written=0
                HDFS: Number of read operations=5
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=1
                HDFS: Number of bytes read erasure-coded=0
        Map-Reduce Framework
                Map input records=5105043
                Map output records=5105039
                Map output bytes=70274887
                Map output materialized bytes=80484971
                Input split bytes=101
                Combine input records=0
                Spilled Records=10210078
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=19
                Total committed heap usage (bytes)=394264576
        File Input Format Counters
                Bytes Read=75380115
2020-10-04 19:24:43,067 INFO mapred.LocalJobRunner: Finishing task: attempt_local1463520834_0001_m_000000_0
2020-10-04 19:24:43,069 INFO mapred.LocalJobRunner: map task executor complete.
2020-10-04 19:24:43,072 INFO mapred.LocalJobRunner: Waiting for reduce tasks
2020-10-04 19:24:43,072 INFO mapred.LocalJobRunner: Starting task: attempt_local1463520834_0001_r_000000_0
2020-10-04 19:24:43,086 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2020-10-04 19:24:43,086 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:fal
se, ignore cleanup failures: false
2020-10-04 19:24:43,088 INFO mapred.Task:  Using ResourceCalculatorProcessTree : [ ]
2020-10-04 19:24:43,098 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.mapreduce.task.reduce.Shuffle@536b5eee
2020-10-04 19:24:43,100 WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
2020-10-04 19:24:43,121 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=1306525696, maxSingleShuffleLimit=326631424, mergeThre
shold=862307008, ioSortFactor=10, memToMemMergeOutputsThreshold=10
```

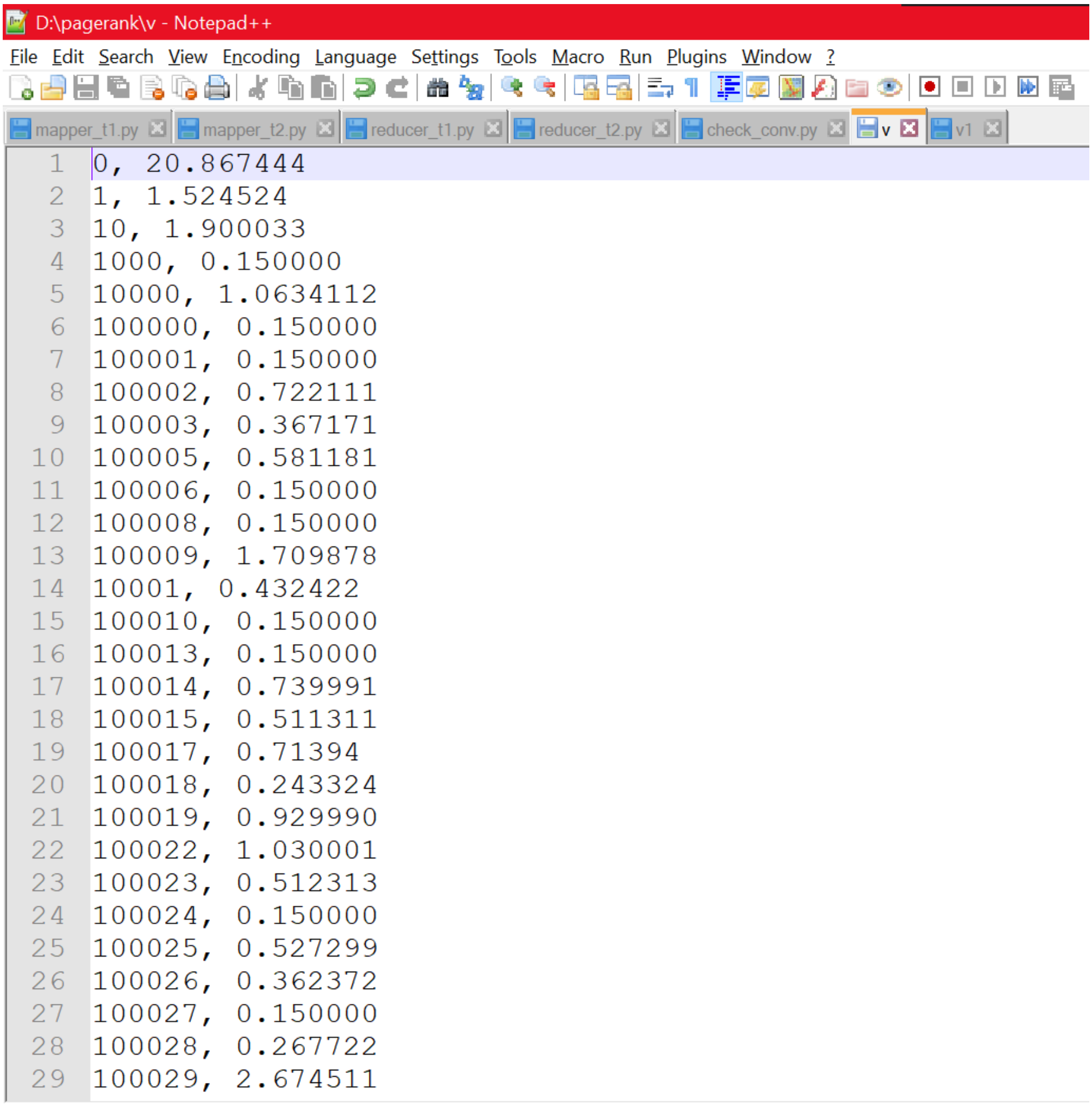


saiprakashlshetty@LAPTOP-VO4EBJ1S: ~/hadoop/hadoop-3.3.0/pagerank

```
2020-10-04 19:24:49,127 INFO mapred.LocalJobRunner: reduce task executor complete.
2020-10-04 19:24:49,194 INFO mapreduce.Job:  map 100% reduce 100%
2020-10-04 19:24:49,195 INFO mapreduce.Job: Job job_local1463520834_0001 completed successfully
2020-10-04 19:24:49,204 INFO mapreduce.Job: Counters: 36
        File System Counters
                FILE: Number of bytes read=322223510
                FILE: Number of bytes written=403980077
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=150760230
                HDFS: Number of bytes written=40246133
                HDFS: Number of read operations=15
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=4
                HDFS: Number of bytes read erasure-coded=0
        Map-Reduce Framework
                Map input records=5105043
                Map output records=5105039
                Map output bytes=70274887
                Map output materialized bytes=80484971
                Input split bytes=101
                Combine input records=0
                Combine output records=0
                Reduce input groups=739454
                Reduce shuffle bytes=80484971
                Reduce input records=5105039
                Reduce output records=739454
                Spilled Records=15315117
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=26
                Total committed heap usage (bytes)=867172352
        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=75380115
```

OUTPUT:

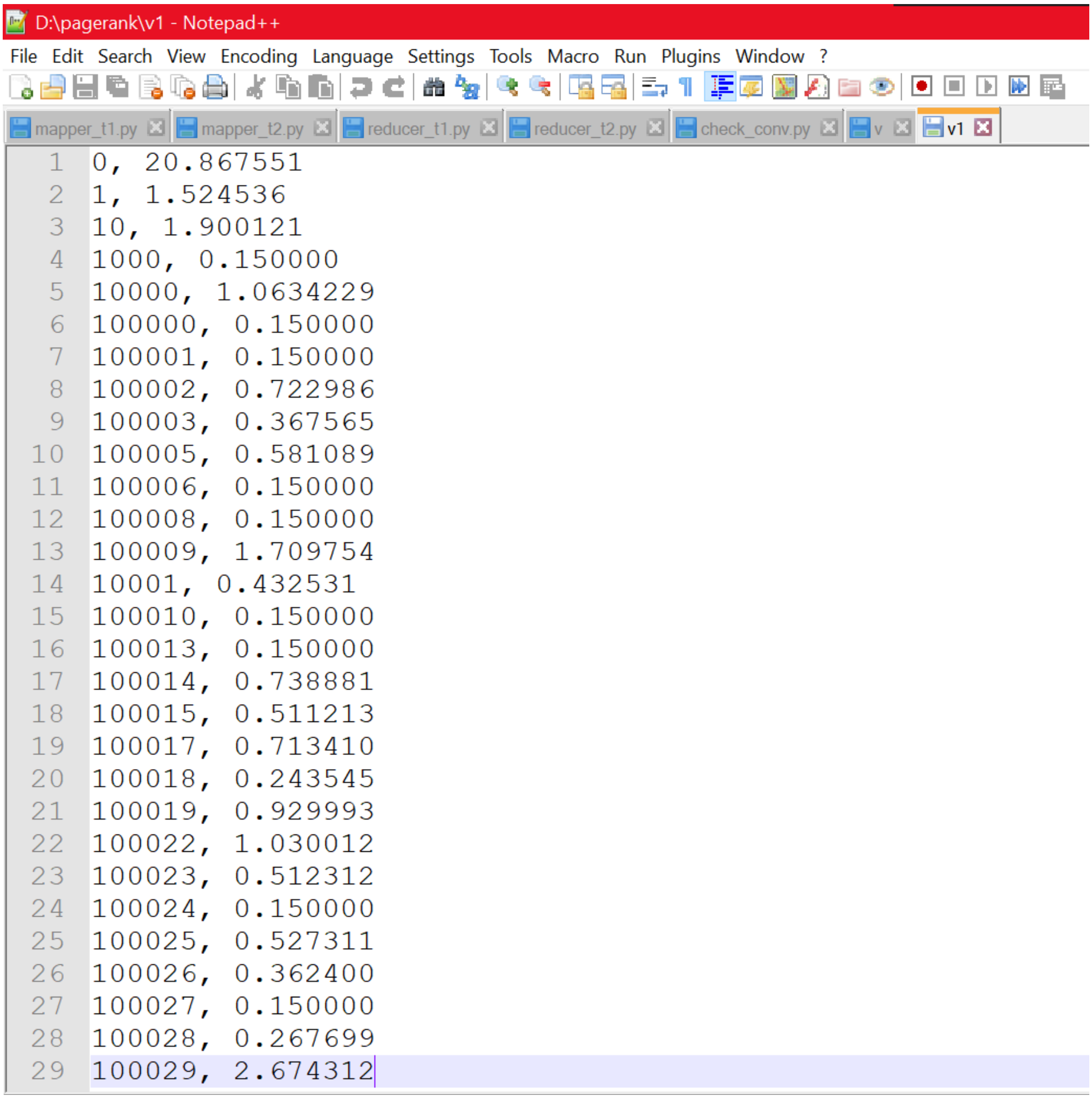
After 40 iterations, output file “v” is



D:\pagerank\v - Notepad++
File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?
mapper_t1.py
mapper_t2.py
reducer_t1.py
reducer_t2.py
check_conv.py
v
v1

```
0, 20.867444
1, 1.524524
10, 1.900033
1000, 0.150000
10000, 1.0634112
100000, 0.150000
100001, 0.150000
100002, 0.722111
100003, 0.367171
100005, 0.581181
100006, 0.150000
100008, 0.150000
100009, 1.709878
10001, 0.432422
100010, 0.150000
100013, 0.150000
100014, 0.739991
100015, 0.511311
100017, 0.71394
100018, 0.243324
100019, 0.929990
100022, 1.030001
100023, 0.512313
100024, 0.150000
100025, 0.527299
100026, 0.362372
100027, 0.150000
100028, 0.267722
100029, 2.674511
```

- After 40 iterations, output file “v1” is


D:\pagerank\v1 - Notepad++
File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?
mapper_t1.py
mapper_t2.py
reducer_t1.py
reducer_t2.py
check_conv.py
v
v1

```
0, 20.867551
1, 1.524536
10, 1.900121
1000, 0.150000
10000, 1.0634229
100000, 0.150000
100001, 0.150000
100002, 0.722986
100003, 0.367565
100005, 0.581089
100006, 0.150000
100008, 0.150000
100009, 1.709754
10001, 0.432531
100010, 0.150000
100013, 0.150000
100014, 0.738881
100015, 0.511213
100017, 0.713410
100018, 0.243545
100019, 0.929993
100022, 1.030012
100023, 0.512312
100024, 0.150000
100025, 0.527311
100026, 0.362400
100027, 0.150000
100028, 0.267699
100029, 2.674312
```

- Using sum of all the probabilities, convergence at nearly ***0.7*** after 40 iteration for page rank.