IERG 4300 Spring 2023 Homework #1

Every Student MUST include the following statement, together with his/her signature in the submitted homework. I declare that the assignment submitted on Elearning system is original except for source material explicitly acknowledged, and that the same or related material has not been previously submitted for another course. I also acknowledge that I am aware of University policy and regulations on honesty in academic work, and of the disciplinary guidelines and procedures applicable to breaches of such policy and regulations, as contained in the website http://www.cuhk.edu.hk/policy/academichonesty/.

Name Yoo Hyun Jun SID 1155100531

Date 19/02/2023 Signature

Task a.

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


for line in sys.stdin:
    line = line.strip()
    cur_followee, follower = line.split(' ')
    print("%s\t%s" % (cur_followee, follower))
```

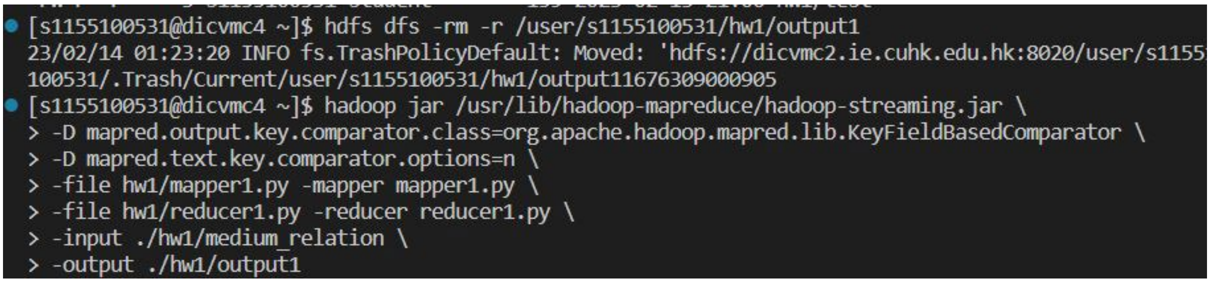
Mapper1.py

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

last_followee = None
follower_list = []

for line in sys.stdin:
    line = line.strip()
    cur_followee, follower = line.split('\t')


    if cur_followee == last_followee:
        follower_list.append(follower)
    else:
        if last_followee is not None:
            follower_list.sort()
            print("%s\t%s" % (last_followee, follower_list))
            follower_list = []
            last_followee = cur_followee
            follower_list.append(follower)
        elif last_followee is None:
            follower_list.append(follower)
            last_followee = cur_followee

print("%s\t%s" %(last_followee, follower_list))
```

Reducer.1py

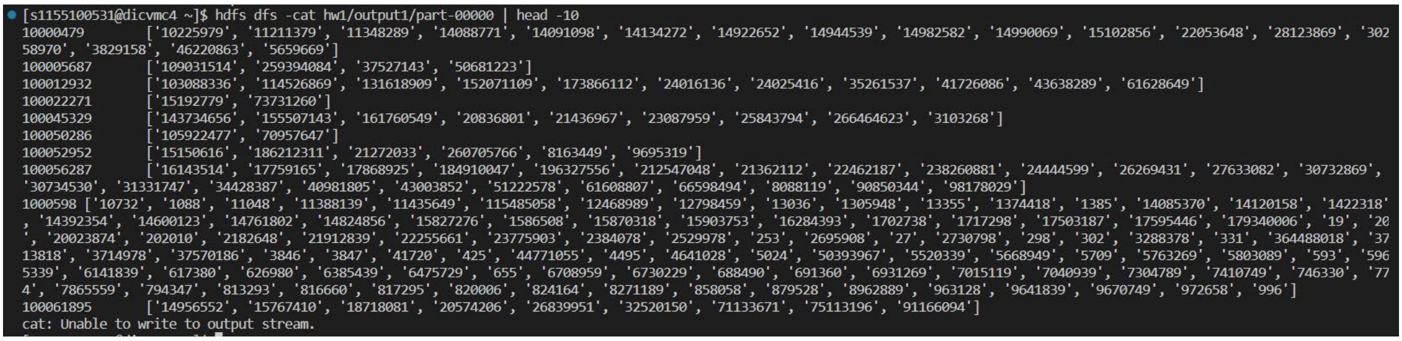
⇨ **In mapper1.py, it is for sorting according to the followee.**

⇨ **In reducer1.py, it makes to create a follower list according to the followee.**



```
[s1155100531@dicvmc4 ~]$ hdfs dfs -rm -r /user/s1155100531/hw1/output1
23/02/14 01:23:20 INFO fs.TrashPolicyDefault: Moved: 'hdfs://dicvmc2.ie.cuhk.edu.hk:8020/user/s1155
100531/.Trash/Current/user/s1155100531/hw1/output11676309000905
[s1155100531@dicvmc4 ~]$ hadoop jar /usr/lib/hadoop-mapreduce/hadoop-streaming.jar \
> -D mapred.output.key.comparator.class=org.apache.hadoop.mapred.lib.KeyFieldBasedComparator \
> -D mapred.text.key.comparator.options=n \
> -file hw1/mapper1.py -mapper mapper1.py \
> -file hw1/reducer1.py -reducer reducer1.py \
> -input ./hw1/medium_relation \
> -output ./hw1/output1
```

Cmd for running mapper1, reducer1



```
[s1155100531@dicvmc4 ~]$ hdfs dfs -cat hw1/output1/part-00000 | head -10
10000479        ['10225979', '11211379', '11348289', '14088771', '14091098', '14134272', '14922652', '14944539', '14982582', '14990069', '15102856', '22053648', '28123869', '302
58970', '3829158', '46220863', '5659669']
100005687       ['109031514', '259394084', '37527143', '50681223']
100012932       ['103088336', '114526869', '131618909', '152071109', '173866112', '24016136', '24025416', '35261537', '41726086', '43638289', '61628649']
100022271       ['15192779', '73731260']
100045329       ['143734656', '155507143', '161760549', '20836801', '21436967', '23087959', '25843794', '266464623', '3103268']
100050286       ['105922477', '70957647']
100052952       ['15150616', '186212311', '21272033', '260705766', '8163449', '9695319']
100056287       ['16143514', '17759165', '17868925', '184910047', '196327556', '212547048', '21362112', '22462187', '238260881', '24444599', '26269431', '27633082', '30732869',
'30734530', '31331747', '34428387', '40981805', '43003852', '51222578', '61608807', '66598494', '8088119', '90850344', '98178029']
1000598 ['10732', '1088', '11048', '11388139', '11435649', '115485058', '12468989', '12798459', '13036', '1305948', '13355', '1374418', '1385', '14085370', '14120158', '1422318'
, '14392354', '14600123', '14761802', '14824856', '15827276', '1586508', '15870318', '15903753', '16284393', '1702738', '1717298', '17503187', '17595446', '179340006', '19', '20
', '20023874', '202010', '2182648', '21912839', '22255661', '23775903', '2384078', '2529978', '253', '2695908', '27', '2730798', '298', '302', '3288378', '331', '364488018', '37
13818', '3714978', '37570186', '3846', '3847', '41720', '425', '44771055', '4495', '4641028', '5024', '50393967', '5520339', '5668949', '5709', '5763269', '5803089', '593', '596
5339', '6141839', '617380', '626980', '6385439', '6475729', '655', '6708959', '6730229', '688490', '691360', '6931269', '7015119', '7040939', '7304789', '7410749', '746330', '77
4', '7865559', '794347', '813293', '816660', '817295', '820006', '824164', '8271189', '858058', '879528', '8962889', '963128', '9641839', '9670749', '972658', '996']
100061895       ['14956552', '15767410', '18718081', '20574206', '26839951', '32520150', '71133671', '75113196', '91166094']
cat: Unable to write to output stream.
```

Output of reducer1.py

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

for line in sys.stdin:
    line = line.strip()
    cur_followee, follower_list = line.split('\t')
    follower_list = follower_list[1:-1].split(',')
    for i in follower_list:
        for j in follower_list:
            if i != j:
                pair1 = i.strip()
                pair2 = j.strip()
                p1 = int(pair1[1:len(pair1)-1])
                p2 = int(pair2[1:len(pair2)-1])
                print("%d\t%d\t%s" % (p1, p2, cur_followee))
```

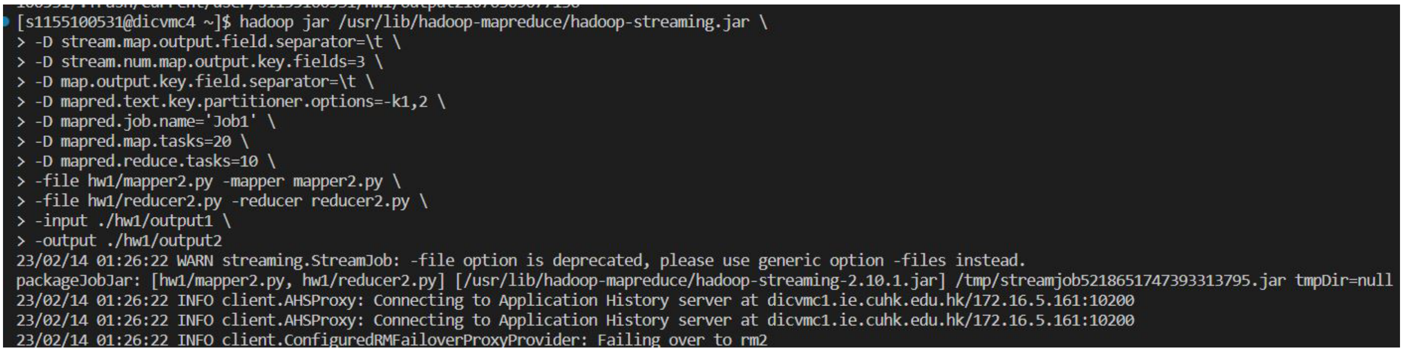
Mapper2.py

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

import sys

last_pair = None
followee_list = []

for line in sys.stdin:
    line = line.strip()
    pair1, pair2, followee = line.split('\t')
    pair = pair1 + ":" + pair2
    if pair == last_pair:
        followee_list.append(followee)
    else:
        if last_pair is not None:
            print("%s\t%s\t%s\t%d" % (pair1, pair2, followee_list, len(followee_list)))
            followee_list = []
            last_pair = pair
            followee_list.append(followee)
        elif last_pair is None:
            followee_list.append(followee)
            last_pair = pair

print("%s\t%s\t%s\t%d" %(pair1, pair2, followee_list, len(followee_list)))
```

Reducer2.py

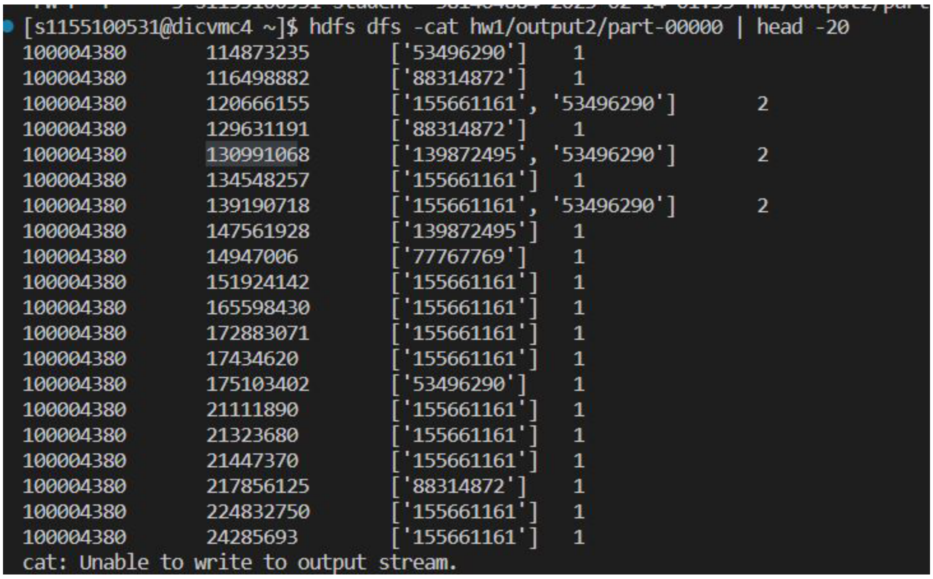
In mapper2.py, create each follower pair form the follower list organized according to cur_followee.

In reducer2.py, for each follower pair, obtain common followee list and obtain the number of common followee.



```
[s1155100531@dicvmc4 ~]$ hadoop jar /usr/lib/hadoop-mapreduce/hadoop-streaming.jar \
> -D stream.map.output.field.separator=\t \
> -D stream.num.map.output.key.fields=3 \
> -D map.output.key.field.separator=\t \
> -D mapred.text.key.partitioner.options=-k1,2 \
> -D mapred.job.name='Job1' \
> -D mapred.map.tasks=20 \
> -D mapred.reduce.tasks=10 \
> -file hw1/mapper2.py -mapper mapper2.py \
> -file hw1/reducer2.py -reducer reducer2.py \
> -input ./hw1/output1 \
> -output ./hw1/output2
23/02/14 01:26:22 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [hw1/mapper2.py, hw1/reducer2.py] [/usr/lib/hadoop-mapreduce/hadoop-streaming-2.10.1.jar] /tmp/streamjob5218651747393313795.jar tmpDir=null
23/02/14 01:26:22 INFO client.AHSProxy: Connecting to Application History server at dicvmc1.ie.cuhk.edu.hk/172.16.5.161:10200
23/02/14 01:26:22 INFO client.AHSProxy: Connecting to Application History server at dicvmc1.ie.cuhk.edu.hk/172.16.5.161:10200
23/02/14 01:26:22 INFO client.ConfiguredRMFailoverProxyProvider: Failing over to rm2
```

Cmd for running mapper2, reducer2



```
[s1155100531@dicvmc4 ~]$ hdfs dfs -cat hw1/output2/part-00000 | head -20
100004380       114873235       ['53496290']    1
100004380       116498882       ['88314872']    1
100004380       120666155       ['155661161', '53496290']       2
100004380       129631191       ['88314872']    1
100004380       130991068       ['139872495', '53496290']       2
100004380       134548257       ['155661161']   1
100004380       139190718       ['155661161', '53496290']       2
100004380       147561928       ['139872495']   1
100004380       14947006        ['77767769']    1
100004380       151924142       ['155661161']   1
100004380       165598430       ['155661161']   1
100004380       172883071       ['155661161']   1
100004380       17434620        ['155661161']   1
100004380       175103402       ['53496290']    1
100004380       21111890        ['155661161']   1
100004380       21323680        ['155661161']   1
100004380       21447370        ['155661161']   1
100004380       217856125       ['88314872']    1
100004380       224832750       ['155661161']   1
100004380       24285693        ['155661161']   1
cat: Unable to write to output stream.
```

Output of reducer2.py

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

import sys


for line in sys.stdin:
    line = line.strip()
    pair1, pair2, followee, cnt = line.split('\t')
    p1 = int(pair1)
    p2 = int(pair2)
    print("%d\t%d\t%s\t%s" % (p1, p2, followee, cnt))
```

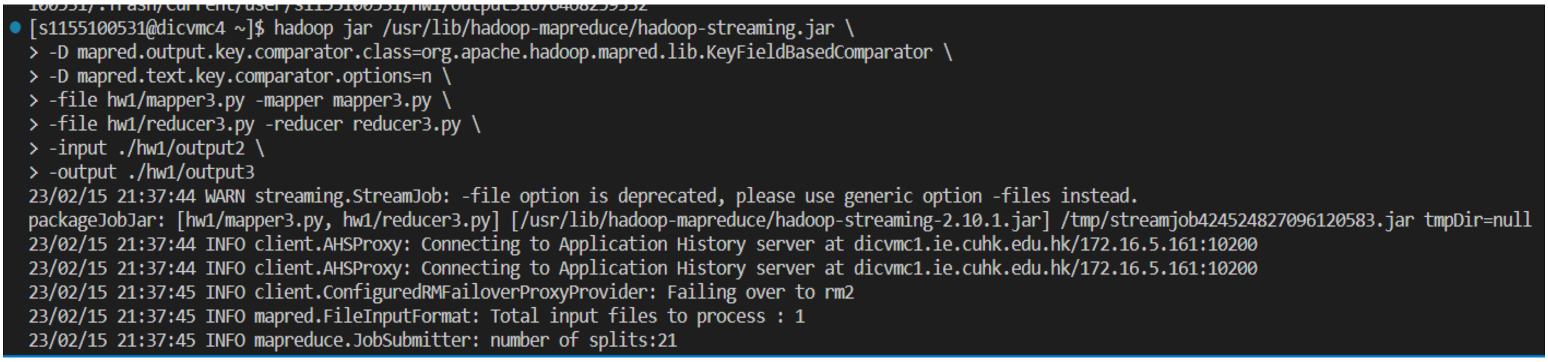
Mapper3.py

```python
#!/usr/bin/env python
import sys
last_pair = None
max_followee_list = None
max_pair1 = None
max_pair2 = None
max = 0
for line in sys.stdin:
    line = line.strip()
    pair1, pair2, followee, cnt = line.split('\t')
    pair = pair1 + ":" + pair2
    if pair1 == last_pair:
        if cnt > max:
            max = cnt
            max_pair1 = pair
            max_followee_list = followee[:]
        elif cnt == max:
            max = cnt
            if max_pair2 < pair2:
                max_pair1 = pair
                max_followee_list = followee[:]
    else:
        if last_pair is not None:
            print("%s\t%s\t%s" % (max_pair1, max_followee_list, max))
            max = cnt
            max_followee_list = followee[:]
            max_pair1 = pair
            last_pair = pair1
            max_pair2 = pair2
        elif last_pair is None:
            max = cnt
            max_followee_list = followee[:]
            max_pair1 = pair
            last_pair = pair1
            max_pair2 = pair2

print("%s\t%s\t%s" % (max_pair1, max_followee_list, max))
```

Reducer3.py

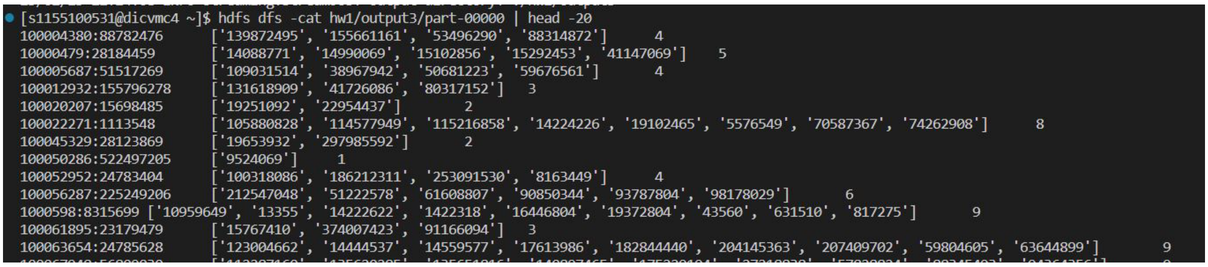
In mapper3.py, it is for sorting according to follower pair.

In reducer3.py, it is for obtaining blog pair with the maximal number of common followees and with the largest number ID pair.



```
● [s1155100531@dicvmc4 ~]$ hadoop jar /usr/lib/hadoop-mapreduce/hadoop-streaming.jar \
 > -D mapred.output.key.comparator.class=org.apache.hadoop.mapred.lib.KeyFieldBasedComparator \
 > -D mapred.text.key.comparator.options=n \
 > -file hw1/mapper3.py -mapper mapper3.py \
 > -file hw1/reducer3.py -reducer reducer3.py \
 > -input ./hw1/output2 \
 > -output ./hw1/output3
23/02/15 21:37:44 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [hw1/mapper3.py, hw1/reducer3.py] [/usr/lib/hadoop-mapreduce/hadoop-streaming-2.10.1.jar] /tmp/streamjob4245248270961205 83.jar tmpDir=null
23/02/15 21:37:44 INFO client.AHSProxy: Connecting to Application History server at dicvmc1.ie.cuhk.edu.hk/172.16.5.161:10200
23/02/15 21:37:44 INFO client.AHSProxy: Connecting to Application History server at dicvmc1.ie.cuhk.edu.hk/172.16.5.161:10200
23/02/15 21:37:45 INFO client.ConfiguredRMFailoverProxyProvider: Failing over to rm2
23/02/15 21:37:45 INFO mapred.FileInputFormat: Total input files to process : 1
23/02/15 21:37:45 INFO mapreduce.JobSubmitter: number of splits:21
```

Cmd for running mapper3, reducer3



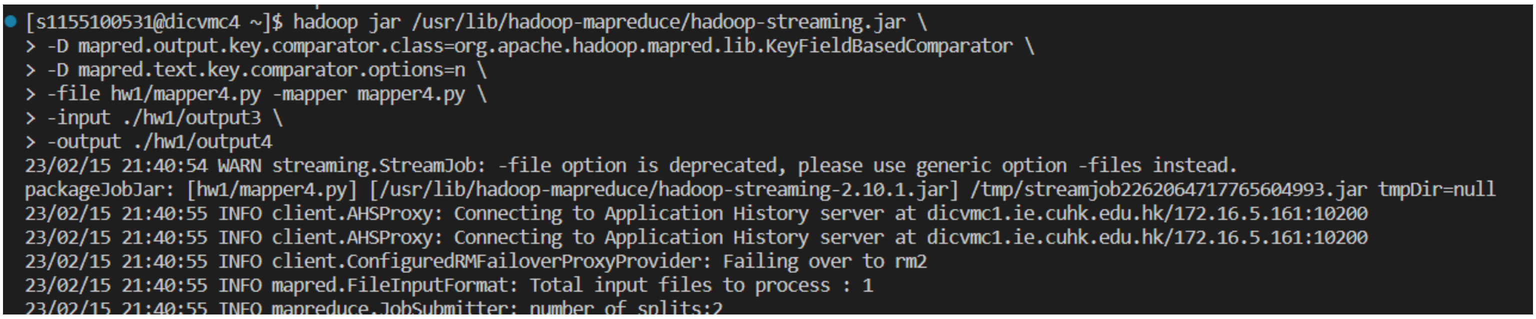
```
[s1155100531@dicvmc4 ~]$ hdfs dfs -cat hw1/output3/part-00000 | head -20
100004380:88782476      ['139872495', '155661161', '53496290', '88314872']      4
10000479:28184459       ['14088771', '14990069', '15102856', '15292453', '41147069']    5
100005687:51517269      ['109031514', '38967942', '50681223', '59676561']       4
100012932:155796278     ['131618909', '41726086', '80317152']   3
100020207:15698485      ['19251092', '22954437']        2
100022271:1113548       ['105880828', '114577949', '115216858', '14224226', '19102465', '5576549', '70587367', '74262908']      8
100045329:28123869      ['19653932', '297985592']       2
100050286:522497205     ['9524069']     1
100052952:24783404      ['100318086', '186212311', '253091530', '8163449']      4
100056287:225249206     ['212547048', '51222578', '61608807', '90850344', '93787804', '98178029']       6
1000598:8315699 ['10959649', '13355', '14222622', '1422318', '16446804', '19372804', '43560', '631510', '817275']       9
100061895:23179479      ['15767410', '374007423', '91166094']   3
100063654:24785628      ['123004662', '14444537', '14559577', '17613986', '182844440', '204145363', '207409702', '59804605', '63644899']       9
```

Output of reducer3.py

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

import sys


for line in sys.stdin:
    line = line.strip()
    pair1, followee, cnt = line.split('\t')
    id = pair1.split(":")[0]
    if id[-3:] == "531":
        print("%s, {%s}, %s" % (pair1, followee, cnt))
```

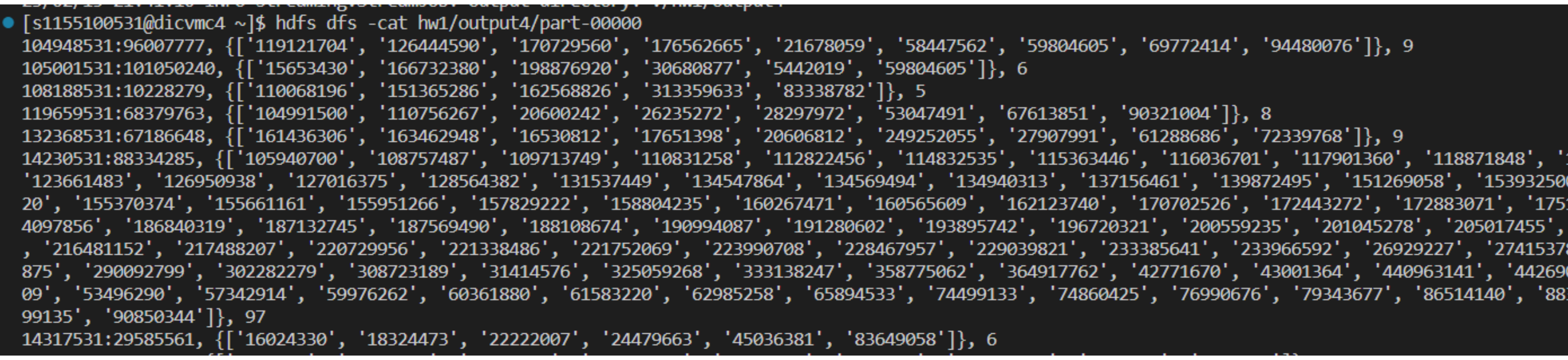
Mapper4.py

In mapper4.py, it is for obtaining community members for blogs with ID 531, 1531,.... As my SID is 1155100531.



```
[s1155100531@dicvmc4 ~]$ hadoop jar /usr/lib/hadoop-mapreduce/hadoop-streaming.jar \
> -D mapred.output.key.comparator.class=org.apache.hadoop.mapred.lib.KeyFieldBasedComparator \
> -D mapred.text.key.comparator.options=n \
> -file hw1/mapper4.py -mapper mapper4.py \
> -input ./hw1/output3 \
> -output ./hw1/output4
23/02/15 21:40:54 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [hw1/mapper4.py] [/usr/lib/hadoop-mapreduce/hadoop-streaming-2.10.1.jar] /tmp/streamjob2262064717765604993.jar tmpDir=null
23/02/15 21:40:55 INFO client.AHSProxy: Connecting to Application History server at dicvmc1.ie.cuhk.edu.hk/172.16.5.161:10200
23/02/15 21:40:55 INFO client.AHSProxy: Connecting to Application History server at dicvmc1.ie.cuhk.edu.hk/172.16.5.161:10200
23/02/15 21:40:55 INFO client.ConfiguredRMFailoverProxyProvider: Failing over to rm2
23/02/15 21:40:55 INFO mapred.FileInputFormat: Total input files to process : 1
23/02/15 21:40:55 INFO mapreduce.JobSubmitter: number of splits:2
```

Cmd for running mapper4



```
[s1155100531@dicvmc4 ~]$ hdfs dfs -cat hw1/output4/part-00000
104948531:96007777, {['119121704', '126444590', '170729560', '176562665', '21678059', '58447562', '59804605', '69772414', '94480076']}, 9
105001531:101050240, {['15653430', '166732380', '198876920', '30680877', '5442019', '59804605']}, 6
108188531:10228279, {['110068196', '151365286', '162568826', '313359633', '83338782']}, 5
119659531:68379763, {['104991500', '110756267', '20600242', '26235272', '28297972', '53047491', '67613851', '90321004']}, 8
132368531:67186648, {['161436306', '163462948', '16530812', '17651398', '20606812', '249252055', '27907991', '61288686', '72339768']}, 9
14230531:88334285, {['105940700', '108757487', '109713749', '110831258', '112822456', '114832535', '115363446', '116036701', '117901360', '118871848', '
'123661483', '126950938', '127016375', '128564382', '131537449', '134547864', '134569494', '134940313', '137156461', '139872495', '151269058', '15393250
20', '155370374', '155661161', '155951266', '157829222', '158804235', '160267471', '160565609', '162123740', '170702526', '172443272', '172883071', '175
4097856', '186840319', '187132745', '187569490', '188108674', '190994087', '191280602', '193895742', '196720321', '200559235', '201045278', '205017455', 
, '216481152', '217488207', '220729956', '221338486', '221752069', '223990708', '228467957', '229039821', '233385641', '233966592', '26929227', '2741537
875', '290092799', '302282279', '308723189', '31414576', '325059268', '333138247', '358775062', '364917762', '42771670', '43001364', '440963141', '44269
09', '53496290', '57342914', '59976262', '60361880', '61583220', '62985258', '65894533', '74499133', '74860425', '76990676', '79343677', '86514140', '88
99135', '90850344']}, 97
14317531:29585561, {['16024330', '18324473', '22222007', '24479663', '45036381', '83649058']}, 6
```

Output of mapper4.py

Task b.

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

for line in sys.stdin:
        line = line.strip()
        followee, follower = line.split(' ')
        print("%s\t%s" %(follower, followee))
```

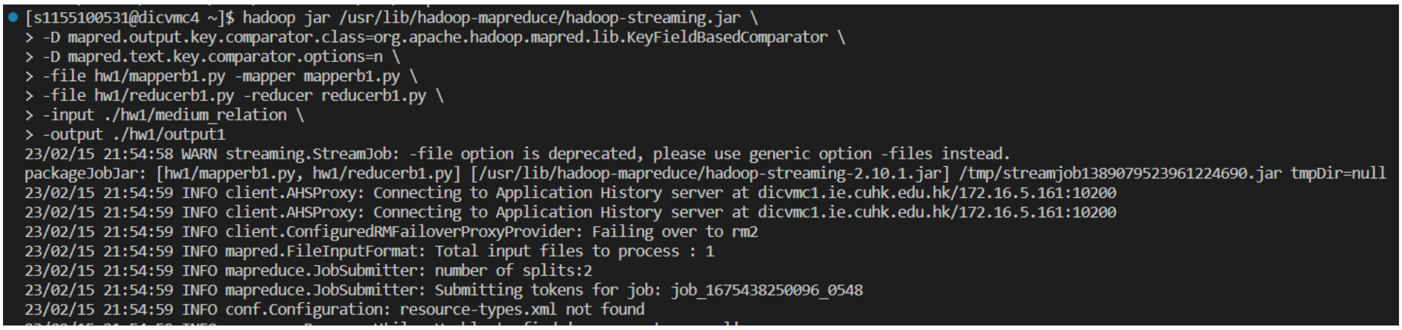
Mapperb1.py

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

last_follower = None
followee_list = []

for line in sys.stdin:
    line = line.strip()
    cur_follower, followee = line.split('\t')
    if cur_follower == last_follower:
        followee_list.append(followee)
    else:
        if last_follower is not None:
            print("%s\t%s" % (last_follower, followee_list))
            followee_list = []
            last_follower = cur_follower
            followee_list.append(followee)
        elif last_follower is None:
            followee_list.append(followee)
            last_follower = cur_follower

print("%s\t%s" % (last_follower, followee_list))
```

reducerb1.py

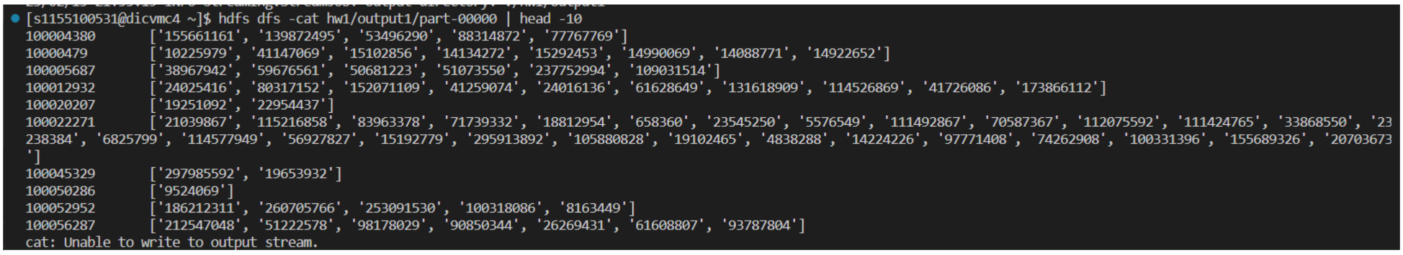
In mapperb1.py, it will switch the follower and followee, and sort according to follower.

In reducerb1.py, it will obtain the list of followees of each follower.



```
[s1155100531@dicvmc4 ~]$ hadoop jar /usr/lib/hadoop-mapreduce/hadoop-streaming.jar \
> -D mapred.output.key.comparator.class=org.apache.hadoop.mapred.lib.KeyFieldBasedComparator \
> -D mapred.text.key.comparator.options=n \
> -file hw1/mapperb1.py -mapper mapperb1.py \
> -file hw1/reducerb1.py -reducer reducerb1.py \
> -input ./hw1/medium_relation \
> -output ./hw1/output1
23/02/15 21:54:58 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [hw1/mapperb1.py, hw1/reducerb1.py] [/usr/lib/hadoop-mapreduce/hadoop-streaming-2.10.1.jar] /tmp/streamjob1389079523961224690.jar tmpDir=null
23/02/15 21:54:59 INFO client.AHSProxy: Connecting to Application History server at dicvmc1.ie.cuhk.edu.hk/172.16.5.161:10200
23/02/15 21:54:59 INFO client.AHSProxy: Connecting to Application History server at dicvmc1.ie.cuhk.edu.hk/172.16.5.161:10200
23/02/15 21:54:59 INFO client.ConfiguredRMFailoverProxyProvider: Failing over to rm2
23/02/15 21:54:59 INFO mapred.FileInputFormat: Total input files to process : 1
23/02/15 21:54:59 INFO mapreduce.JobSubmitter: number of splits:2
23/02/15 21:54:59 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1675438250096_0548
23/02/15 21:54:59 INFO conf.Configuration: resource-types.xml not found
```

Cmd for running reducerb1.py



```
[s1155100531@dicvmc4 ~]$ hdfs dfs -cat hw1/output1/part-00000 | head -10
100004380       ['155661161', '139872495', '53496290', '88314872', '77767769']
10000479        ['10225979', '41147069', '15102856', '14134272', '15292453', '14990069', '14088771', '14922652']
100005687       ['38967942', '59676561', '50681223', '51073550', '237752994', '109031514']
100012932       ['24025416', '80317152', '152071109', '41259074', '24016136', '61628649', '131618909', '114526869', '41726086', '173866112']
100020207       ['19251092', '22954437']
100022271       ['21039867', '115216858', '83963378', '71739332', '18812954', '658360', '23545250', '5576549', '111492867', '70587367', '112075592', '111424765', '33868550', '23
238384', '6825799', '114577949', '56927827', '15192779', '295913892', '105880828', '19102465', '4838288', '14224226', '97771408', '74262908', '100331396', '155689326', '20703673
']
100045329       ['297985592', '19653932']
100050286       ['9524069']
100052952       ['186212311', '260705766', '253091530', '100318086', '8163449']
100056287       ['212547048', '51222578', '98178029', '90850344', '26269431', '61608807', '93787804']
cat: Unable to write to output stream.
```

Output of reducerb1.py

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

dictionary = {}

for line in sys.stdin:
    line = line.strip()
    follower, followee_list = line.split('\t')
    followee_list = ast.literal_eval(followee_list)
    dictionary[follower] = followee_list

for k in dictionary:
    for j in dictionary:
        if k != j:
            if len(dictionary[k]) == 0 or len(dictionary[j]) == 0:
                continue
            else:
                common = list(set(dictionary[k]) & set(dictionary[j]))
                len_c = len(common)
                total = len(list(set(dictionary[k]) | set(dictionary[j])))
                print("%s\t%s\t%s\t%d\t%d" % (k, j, common, len_c, total))
```

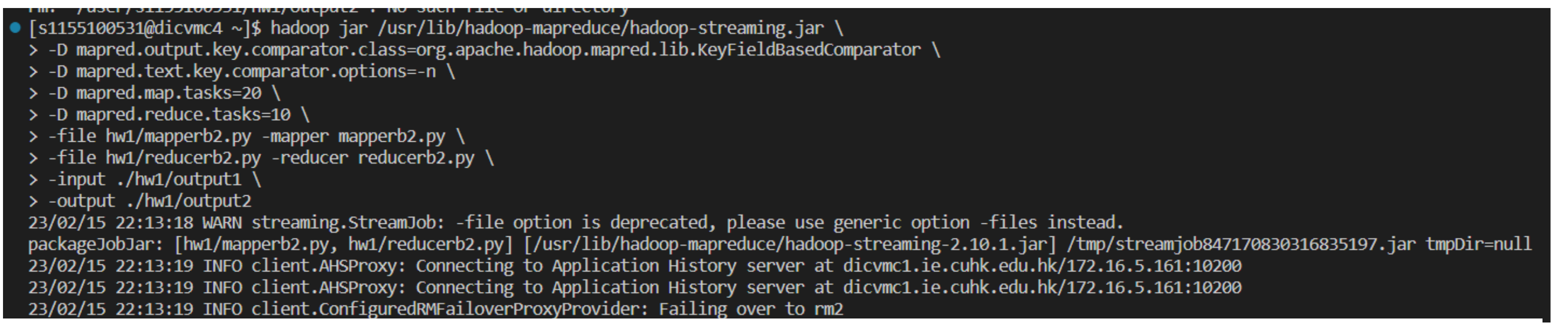
Mapperb2.py

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

for line in sys.stdin:
    line = line.strip()
    pair1, pair2, common_list, len_c, total = line.split('\t')
    similarity = float(len_c) / float(total)
    if similarity > 0:
        print("%s\t%s\t%s\t%s" % (pair1, similarity, pair2, common_list))
```

Reducerb2.py

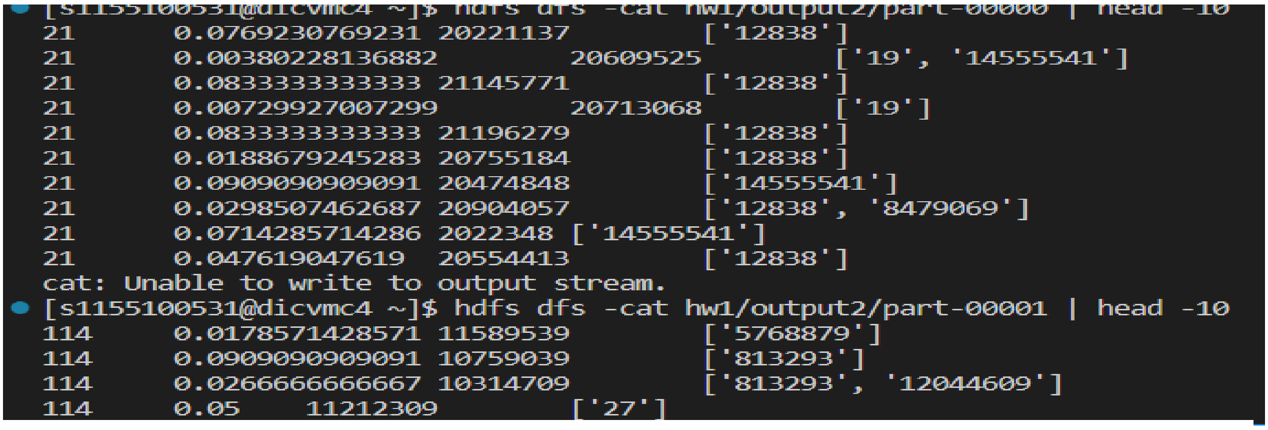
In mapperb2.py, it will obtain the intersection and union of two followee lists according to two followers.

In reducerb2.py, it will calculate the similarity for each follower pair.



```
rm: `/user/s1155100531/hw1/output2': No such file or directory
[s1155100531@dicvmc4 ~]$ hadoop jar /usr/lib/hadoop-mapreduce/hadoop-streaming.jar \
> -D mapred.output.key.comparator.class=org.apache.hadoop.mapred.lib.KeyFieldBasedComparator \
> -D mapred.text.key.comparator.options=-n \
> -D mapred.map.tasks=20 \
> -D mapred.reduce.tasks=10 \
> -file hw1/mapperb2.py -mapper mapperb2.py \
> -file hw1/reducerb2.py -reducer reducerb2.py \
> -input ./hw1/output1 \
> -output ./hw1/output2
23/02/15 22:13:18 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [hw1/mapperb2.py, hw1/reducerb2.py] [/usr/lib/hadoop-mapreduce/hadoop-streaming-2.10.1.jar] /tmp/streamjob847170830316835197.jar tmpDir=null
23/02/15 22:13:19 INFO client.AHSProxy: Connecting to Application History server at dicvmc1.ie.cuhk.edu.hk/172.16.5.161:10200
23/02/15 22:13:19 INFO client.AHSProxy: Connecting to Application History server at dicvmc1.ie.cuhk.edu.hk/172.16.5.161:10200
23/02/15 22:13:19 INFO client.ConfiguredRMFailoverProxyProvider: Failing over to rm2
```

Cmd for Reducerb2.py

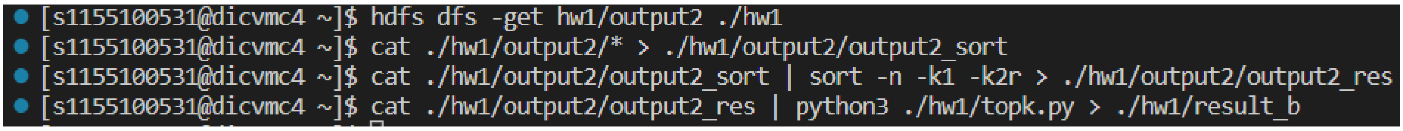


```
[s1155100531@dicvmc4 ~]$ hdfs dfs -cat hw1/output2/part-00000 | head -10
21      0.0769230769231 20221137        ['12838']
21      0.00380228136882        20609525        ['19', '14555541']
21      0.0833333333333 21145771        ['12838']
21      0.00729927007299        20713068        ['19']
21      0.0833333333333 21196279        ['12838']
21      0.0188679245283 20755184        ['12838']
21      0.0909090909091 20474848        ['14555541']
21      0.0298507462687 20904057        ['12838', '8479069']
21      0.0714285714286 2022348 ['14555541']
21      0.047619047619  20554413        ['12838']
cat: Unable to write to output stream.
[s1155100531@dicvmc4 ~]$ hdfs dfs -cat hw1/output2/part-00001 | head -10
114     0.0178571428571 11589539        ['5768879']
114     0.0909090909091 10759039        ['813293']
114     0.0266666666667 10314709        ['813293', '12044609']
114     0.05    11212309        ['27']
```

Output of Reducerb2.py

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

last_pair = None
cnt = 0
for line in sys.stdin:
    line = line.strip()
    pair1, sim,  pair2, followee_list   = line.split('\t')
    followee_list = ast.literal_eval(followee_list)
    if pair1 == last_pair:
        if cnt < 3:
            if pair1[-3:] =="531":
                print("%s:%s, {%s}, %s" % (pair1, pair2, followee_list, sim))
            cnt += 1
    else:
        if last_pair is not None:
            cnt = 0
            if pair1[-3:] =="531":
                print("%s:%s, {%s}, %s" % (pair1, pair2, followee_list, sim))
            last_pair = pair1
            cnt += 1

        elif last_pair is None:
            if pair1[-3:] =="531":
                print("%s:%s, {%s}, %s" % (pair1, pair2, followee_list, sim))
            last_pair = pair1
            cnt += 1
```

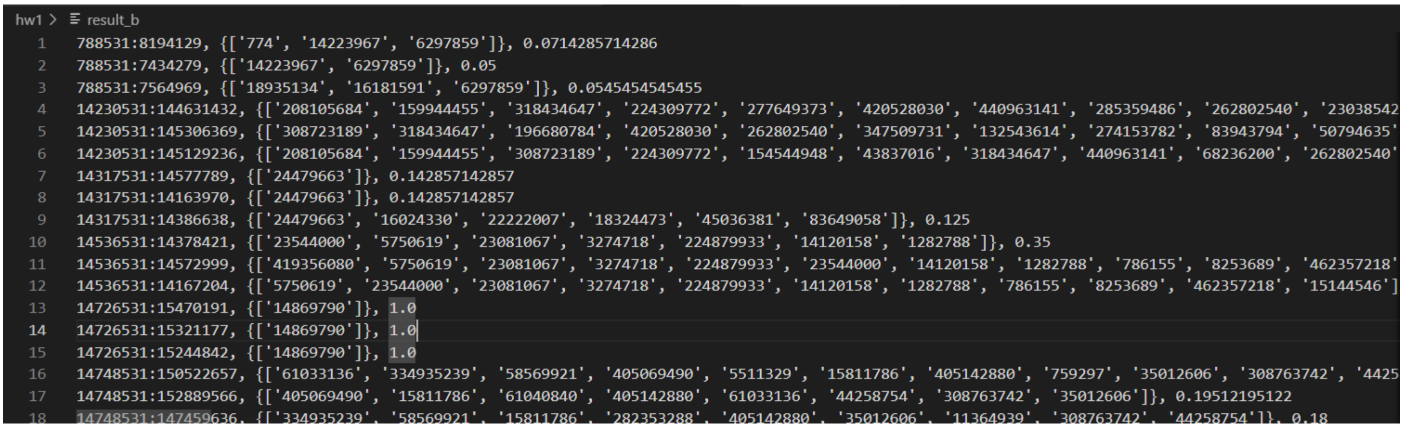
topk.py



```
[s1155100531@dicvmc4 ~]$ hdfs dfs -get hw1/output2 ./hw1
[s1155100531@dicvmc4 ~]$ cat ./hw1/output2/* > ./hw1/output2/output2_sort
[s1155100531@dicvmc4 ~]$ cat ./hw1/output2/output2_sort | sort -n -k1 -k2r > ./hw1/output2/output2_res
[s1155100531@dicvmc4 ~]$ cat ./hw1/output2/output2_res | python3 ./hw1/topk.py > ./hw1/result_b
```

Cmd for copying and sorting the output of reducerb2.py

After copying the output of reducerb2.py, sorting it according to follower and similarity. Then, printing the top 3 of each blog according to my last num digit SID “531”.



```
788531:8194129, {['774', '14223967', '6297859']}, 0.0714285714286
788531:7434279, {['14223967', '6297859']}, 0.05
788531:7564969, {['18935134', '16181591', '6297859']}, 0.0545454545455
14230531:144631432, {['208105684', '159944455', '318434647', '224309772', '277649373', '420528030', '440963141', '285359486', '262802540', '23038542
14230531:145306369, {['308723189', '318434647', '196680784', '420528030', '262802540', '347509731', '132543614', '274153782', '83943794', '50794635'
14230531:145129236, {['208105684', '159944455', '308723189', '224309772', '154544948', '43837016', '318434647', '440963141', '68236200', '262802540'
14317531:14577789, {['24479663']}, 0.142857142857
14317531:14163970, {['24479663']}, 0.142857142857
14317531:14386638, {['24479663', '16024330', '22222007', '18324473', '45036381', '83649058']}, 0.125
14536531:14378421, {['23544000', '5750619', '23081067', '3274718', '224879933', '14120158', '1282788']}, 0.35
14536531:14572999, {['419356080', '5750619', '23081067', '3274718', '224879933', '23544000', '14120158', '1282788', '786155', '8253689', '462357218'
14536531:14167204, {['5750619', '23544000', '23081067', '3274718', '224879933', '14120158', '1282788', '786155', '8253689', '462357218', '15144546']
14726531:15470191, {['14869790']}, 1.0
14726531:15321177, {['14869790']}, 1.0
14726531:15244842, {['14869790']}, 1.0
14748531:150522657, {['61033136', '334935239', '58569921', '405069490', '5511329', '15811786', '405142880', '759297', '35012606', '308763742', '4425
14748531:152889566, {['405069490', '15811786', '61040840', '405142880', '61033136', '44258754', '308763742', '35012606']}, 0.19512195122
14748531:147459636, {['334935239', '58569921', '15811786', '282353288', '405142880', '35012606', '11364939', '308763742', '44258754']}, 0.18
```

Output of topk.py

Task c.

It will use the output of reducer1.py which contains the list of follower of each followee.

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

check = ""

commu_list = {} # to save community info
#./hw1/small/small_label
with open('medium_label') as f:
        for line in f.readlines():
                commu = line.strip()
                blog, commu = commu.split(' ')
                commu_list.update({blog: commu})

for line in sys.stdin:
    line = line.strip()
    followee, follower_list = line.split('\t')
    follower_list = ast.literal_eval(follower_list)
    community = community_list[followee]
    if len(follower_list) > 1:
        check = "o"
    else:
        check = "x"
    print("%s\t%s\t%s" % (community, followee, check))
```

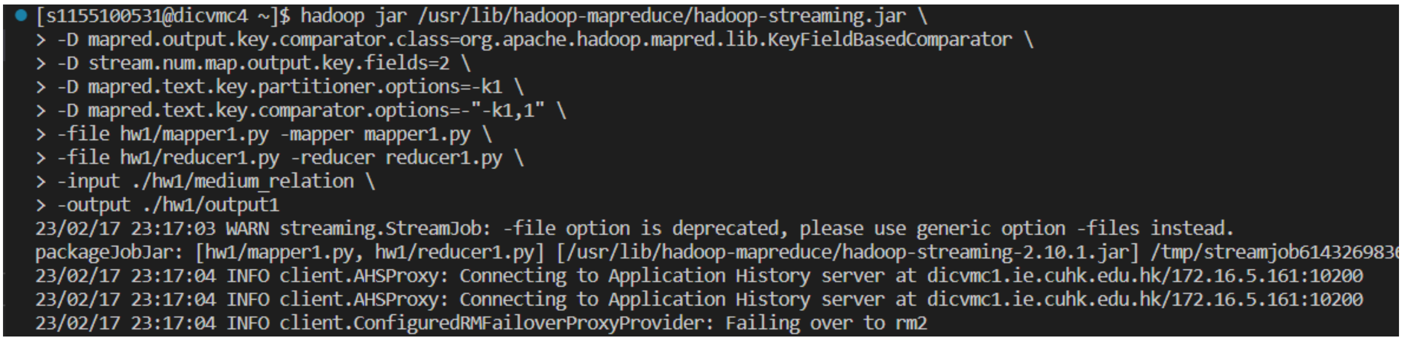
Mapperc1.py

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

last_commu = None
followee_list = {}
cnt = 0

for line in sys.stdin:
    line = line.strip()
    cur_commu, followee, check = line.split('\t')
    # print("%s\t%s\t%s" % (cur_commu, followee, check))
    if cur_commu == last_commu:
        followee_list[followee] = check
        if check == 'o':
            cnt += 1
    else:
        if last_commu is not None:
            print("Community %s:\t%s" % (last_commu, cnt))
            cnt = 0
            followee_list = {}
            last_commu = cur_commu
            followee_list[followee] = check
            if check == 'o':
                cnt += 1
        elif last_commu is None:
            followee_list[followee] = check
            if check == 'o':
                cnt += 1
            last_commu = cur_commu

print("Community %s:\t%s" % (last_commu, cnt))
```

Reducerc1.py

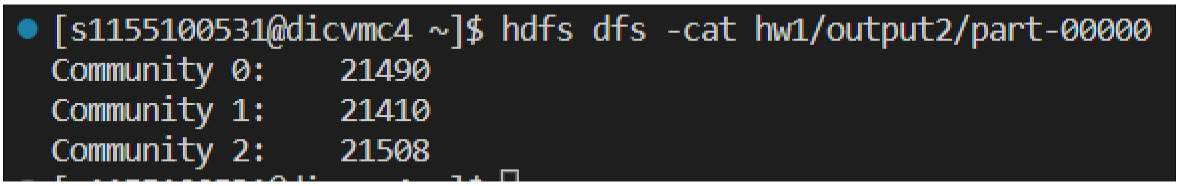
In mapperc1.py, it allows the label file and the relation file to be merged based on blog id. And check whether it has at least two followers.

In reducerc1.py, it will sum up the number of blog which has check “o” according to the community number.



```
[s1155100531@dicvmc4 ~]$ hadoop jar /usr/lib/hadoop-mapreduce/hadoop-streaming.jar \
 > -D mapred.output.key.comparator.class=org.apache.hadoop.mapred.lib.KeyFieldBasedComparator \
 > -D stream.num.map.output.key.fields=2 \
 > -D mapred.text.key.partitioner.options=-k1 \
 > -D mapred.text.key.comparator.options=-"-k1,1" \
 > -file hw1/mapper1.py -mapper mapper1.py \
 > -file hw1/reducer1.py -reducer reducer1.py \
 > -input ./hw1/medium_relation \
 > -output ./hw1/output1
23/02/17 23:17:03 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [hw1/mapper1.py, hw1/reducer1.py] [/usr/lib/hadoop-mapreduce/hadoop-streaming-2.10.1.jar] /tmp/streamjob614326983
23/02/17 23:17:04 INFO client.AHSProxy: Connecting to Application History server at dicvmc1.ie.cuhk.edu.hk/172.16.5.161:10200
23/02/17 23:17:04 INFO client.AHSProxy: Connecting to Application History server at dicvmc1.ie.cuhk.edu.hk/172.16.5.161:10200
23/02/17 23:17:04 INFO client.ConfiguredRMFailoverProxyProvider: Failing over to rm2
```

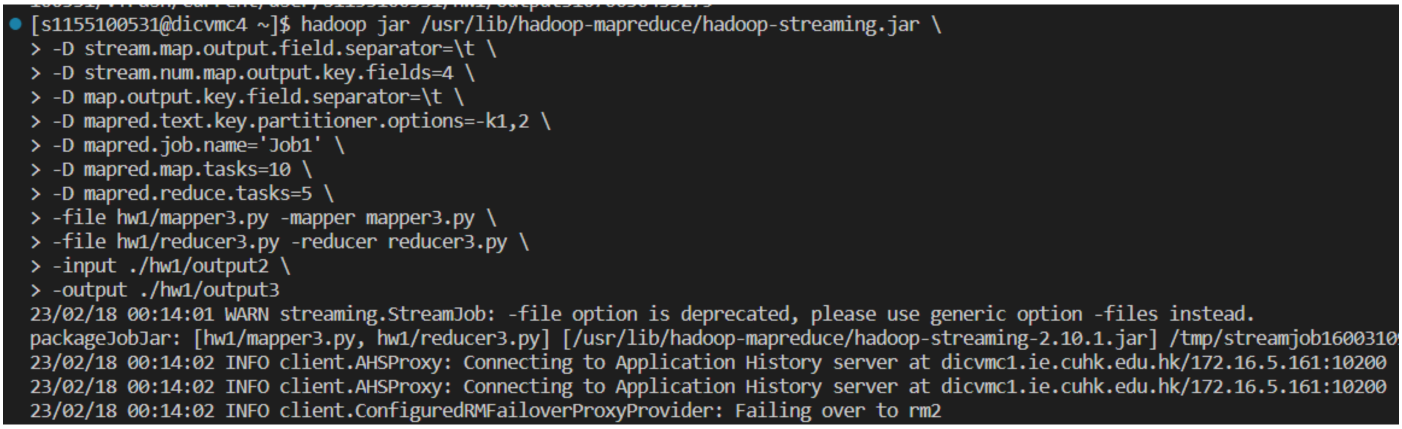
Cmd for mapperc1.py and reducerc1.py



```
[s1155100531@dicvmc4 ~]$ hdfs dfs -cat hw1/output2/part-00000
Community 0:    21490
Community 1:    21410
Community 2:    21508
```

Output of redcerc1.py

Task d.



```
[s1155100531@dicvmc4 ~]$ hadoop jar /usr/lib/hadoop-mapreduce/hadoop-streaming.jar \
 > -D stream.map.output.field.separator=\t \
 > -D stream.num.map.output.key.fields=4 \
 > -D map.output.key.field.separator=\t \
 > -D mapred.text.key.partitioner.options=-k1,2 \
 > -D mapred.job.name='Job1' \
 > -D mapred.map.tasks=10 \
 > -D mapred.reduce.tasks=5 \
 > -file hw1/mapper3.py -mapper mapper3.py \
 > -file hw1/reducer3.py -reducer reducer3.py \
 > -input ./hw1/output2 \
 > -output ./hw1/output3
 23/02/18 00:14:01 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
 packageJobJar: [hw1/mapper3.py, hw1/reducer3.py] [/usr/lib/hadoop-mapreduce/hadoop-streaming-2.10.1.jar] /tmp/streamjob1600310
 23/02/18 00:14:02 INFO client.AHSProxy: Connecting to Application History server at dicvmc1.ie.cuhk.edu.hk/172.16.5.161:10200
 23/02/18 00:14:02 INFO client.AHSProxy: Connecting to Application History server at dicvmc1.ie.cuhk.edu.hk/172.16.5.161:10200
 23/02/18 00:14:02 INFO client.ConfiguredRMFailoverProxyProvider: Failing over to rm2
```

Cmd format for task d

Mapper num	Reducer num	Max mapper time	Min mapper time	Avg mapper time	Max reducer time	Min reducer time	Avg reducer time	Total job time
10	1	23sec	14sec	17sec	1min 44sec	1min 44sec	1min 44sec	4min 42sec
10	5	23sec	16sec	19sec	21sec	20sec	21sec	3min 46sec
10	10	21sec	16sec	18sec	10sec	10sec	10sec	3min 47sec
10	20	22sec	16sec	19sec	6sec	5sec	5sec	3min 56sec
50	10	15sec	11sec	12sec	10sec	10sec	10sec	4min 12sec
100	10	9sec	6sec	7sec	10sec	10sec	10sec	5min 8sec

Result table of task d

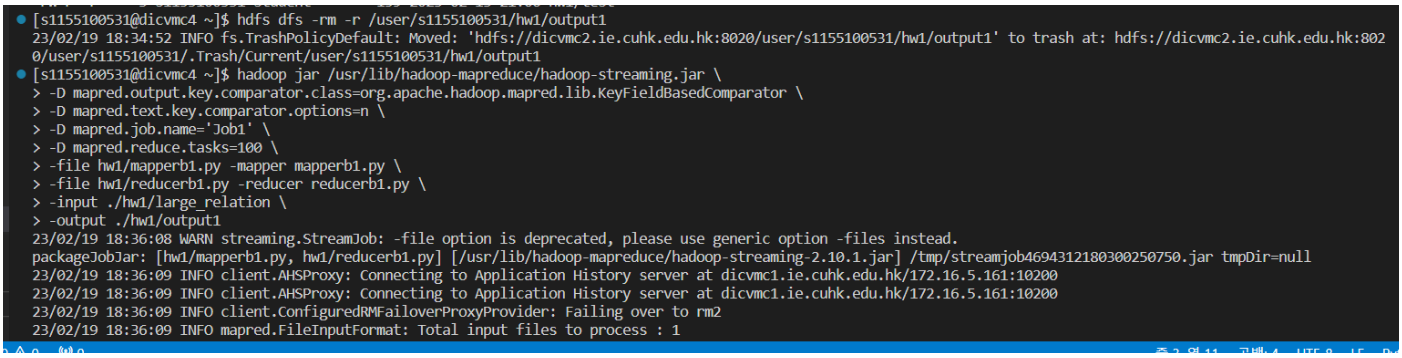
1 Increasing the number of reducer

When I increase the number of reducer, the reduce time decreases, but the total job running time does not change significantly after a certain amount of time. The reason is that most of the time is spent on shuffle and merge in the first reducer. After that, it doesn't spend as much time on reduce time. Therefore, no matter how much you increase the number of reducer, the reducer time will decrease, but since most of the time will be used in shuffle and merge so the impact on the total running time is insignificant.

2. Increasing the number of mapper

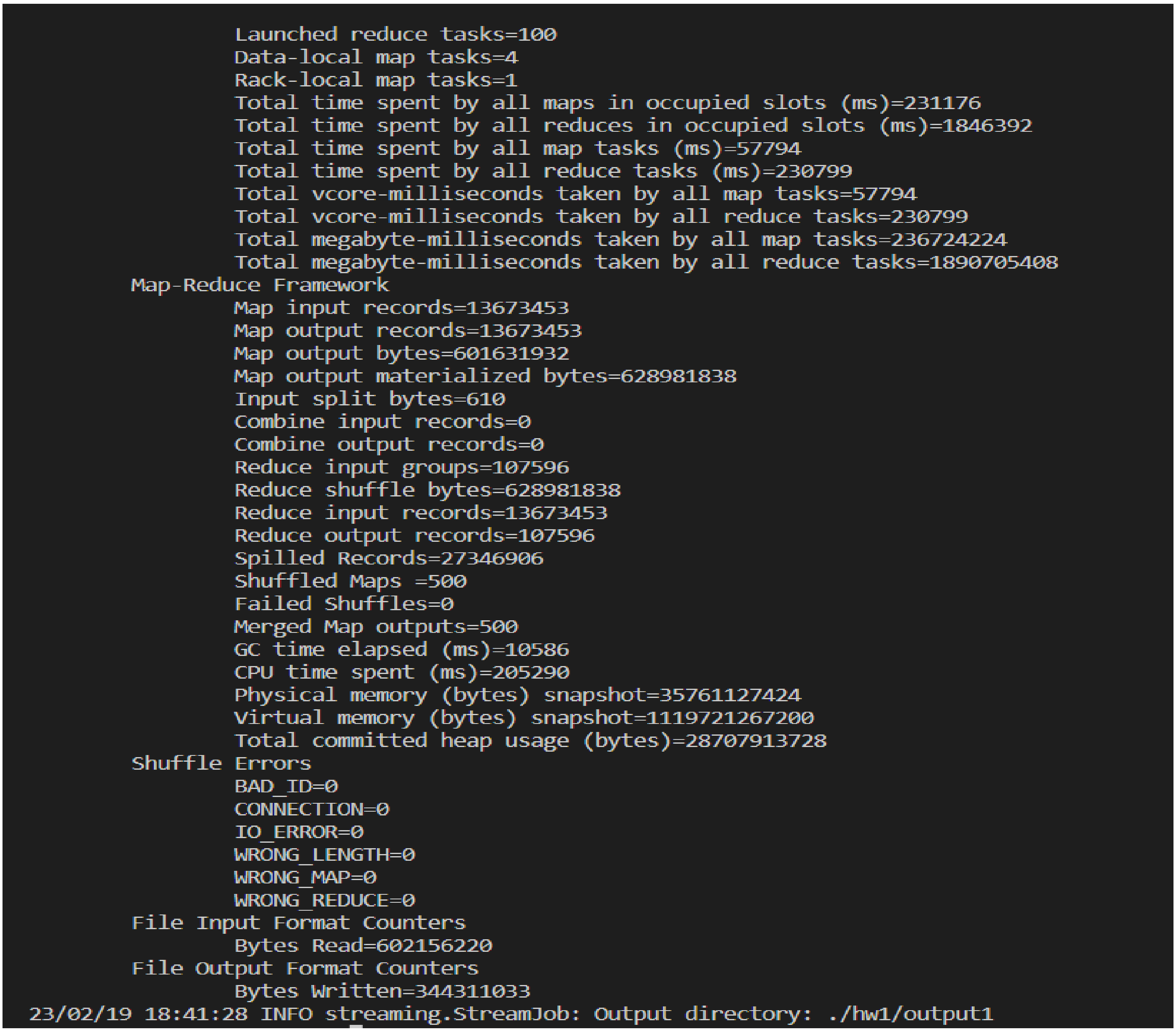
When I increase the number of mapper, the mapper time is significantly decreased. However, the total job running time is increased. For this reason, form my point of view, if I create too many mappers, it seems to reduce the efficiency in dividing data nodes.

Task e.



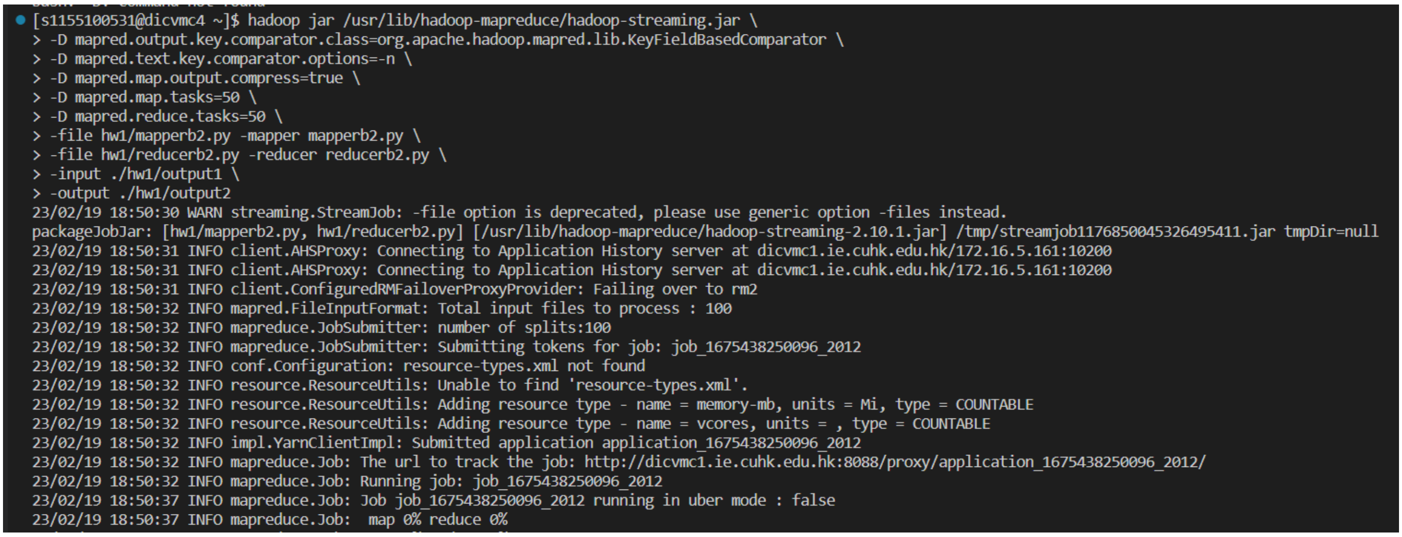
```
● [s1155100531@dicvmc4 ~]$ hdfs dfs -rm -r /user/s1155100531/hw1/output1
  23/02/19 18:34:52 INFO fs.TrashPolicyDefault: Moved: 'hdfs://dicvmc2.ie.cuhk.edu.hk:8020/user/s1155100531/hw1/output1' to trash at: hdfs://dicvmc2.ie.cuhk.edu.hk:802
  0/user/s1155100531/.Trash/Current/user/s1155100531/hw1/output1
● [s1155100531@dicvmc4 ~]$ hadoop jar /usr/lib/hadoop-mapreduce/hadoop-streaming.jar \
  > -D mapred.output.key.comparator.class=org.apache.hadoop.mapred.lib.KeyFieldBasedComparator \
  > -D mapred.text.key.comparator.options=n \
  > -D mapred.job.name='Job1' \
  > -D mapred.reduce.tasks=100 \
  > -file hw1/mapperb1.py -mapper mapperb1.py \
  > -file hw1/reducerb1.py -reducer reducerb1.py \
  > -input ./hw1/large_relation \
  > -output ./hw1/output1
  23/02/19 18:36:08 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
  packageJobJar: [hw1/mapperb1.py, hw1/reducerb1.py] [/usr/lib/hadoop-mapreduce/hadoop-streaming-2.10.1.jar] /tmp/streamjob4694312180300250750.jar tmpDir=null
  23/02/19 18:36:09 INFO client.AHSProxy: Connecting to Application History server at dicvmc1.ie.cuhk.edu.hk/172.16.5.161:10200
  23/02/19 18:36:09 INFO client.AHSProxy: Connecting to Application History server at dicvmc1.ie.cuhk.edu.hk/172.16.5.161:10200
  23/02/19 18:36:09 INFO client.ConfiguredRMFailoverProxyProvider: Failing over to rm2
  23/02/19 18:36:09 INFO mapred.FileInputFormat: Total input files to process : 1
```

Cmd for running taskb(mapperb1,reducerb1) in large relation



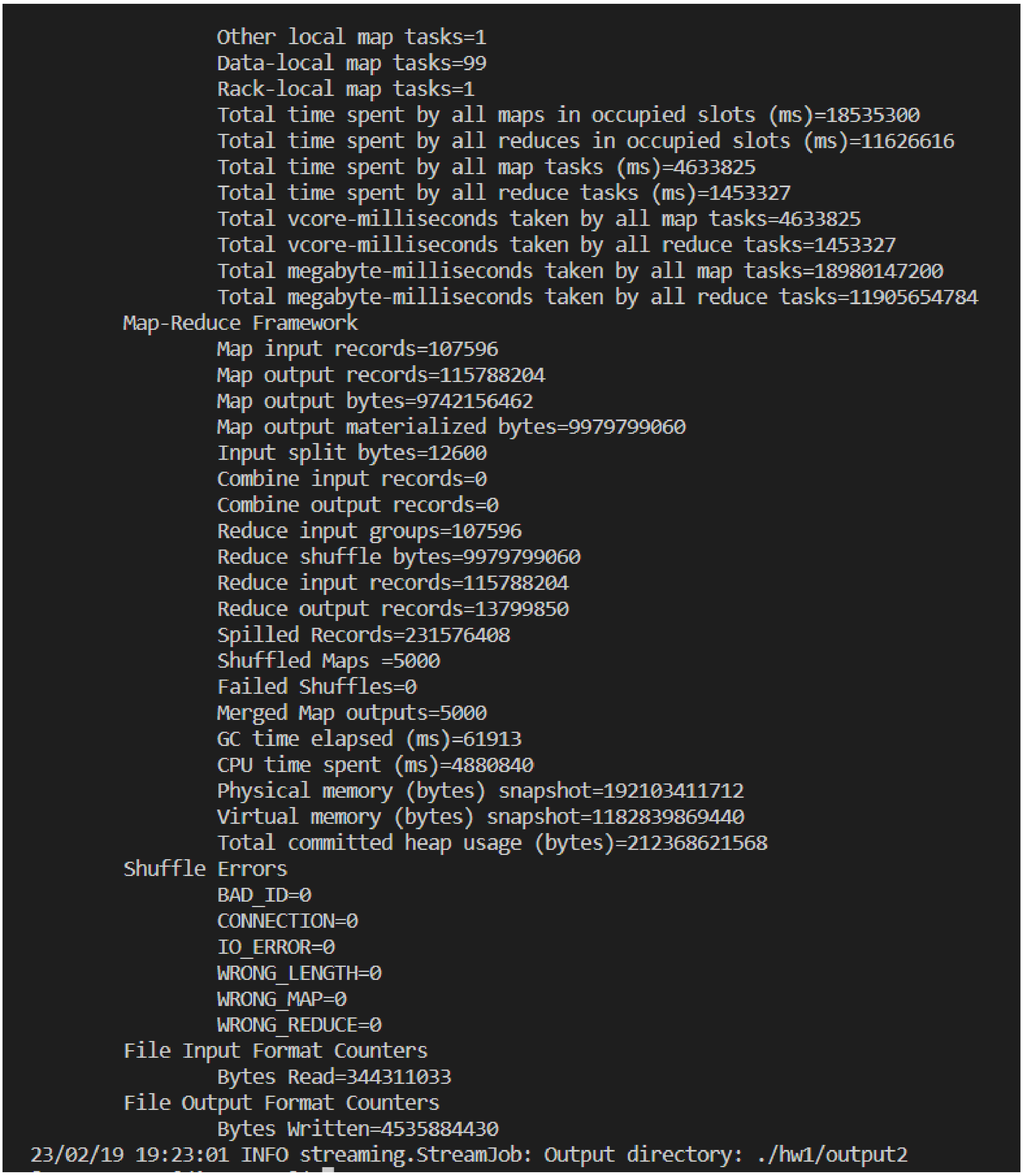
```
                Launched reduce tasks=100
                Data-local map tasks=4
                Rack-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=231176
                Total time spent by all reduces in occupied slots (ms)=1846392
                Total time spent by all map tasks (ms)=57794
                Total time spent by all reduce tasks (ms)=230799
                Total vcore-milliseconds taken by all map tasks=57794
                Total vcore-milliseconds taken by all reduce tasks=230799
                Total megabyte-milliseconds taken by all map tasks=236724224
                Total megabyte-milliseconds taken by all reduce tasks=1890705408
        Map-Reduce Framework
                Map input records=13673453
                Map output records=13673453
                Map output bytes=601631932
                Map output materialized bytes=628981838
                Input split bytes=610
                Combine input records=0
                Combine output records=0
                Reduce input groups=107596
                Reduce shuffle bytes=628981838
                Reduce input records=13673453
                Reduce output records=107596
                Spilled Records=27346906
                Shuffled Maps =500
                Failed Shuffles=0
                Merged Map outputs=500
                GC time elapsed (ms)=10586
                CPU time spent (ms)=205290
                Physical memory (bytes) snapshot=35761127424
                Virtual memory (bytes) snapshot=1119721267200
                Total committed heap usage (bytes)=28707913728
        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=602156220
        File Output Format Counters
                Bytes Written=344311033
23/02/19 18:41:28 INFO streaming.StreamJob: Output directory: ./hw1/output1
```

Result of first mapper, reducer



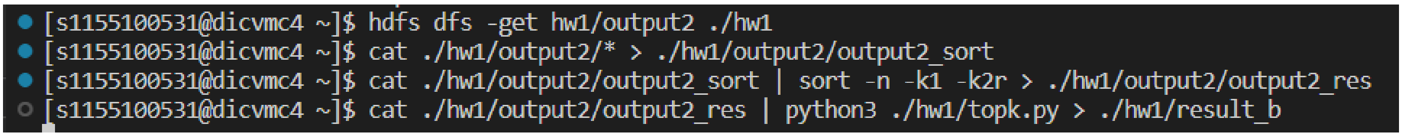
```
● [s1155100531@dicvmc4 ~]$ hadoop jar /usr/lib/hadoop-mapreduce/hadoop-streaming.jar \
  > -D mapred.output.key.comparator.class=org.apache.hadoop.mapred.lib.KeyFieldBasedComparator \
  > -D mapred.text.key.comparator.options=-n \
  > -D mapred.map.output.compress=true \
  > -D mapred.map.tasks=50 \
  > -D mapred.reduce.tasks=50 \
  > -file hw1/mapperb2.py -mapper mapperb2.py \
  > -file hw1/reducerb2.py -reducer reducerb2.py \
  > -input ./hw1/output1 \
  > -output ./hw1/output2
  23/02/19 18:50:30 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
  packageJobJar: [hw1/mapperb2.py, hw1/reducerb2.py] [/usr/lib/hadoop-mapreduce/hadoop-streaming-2.10.1.jar] /tmp/streamjob1176850045326495411.jar tmpDir=null
  23/02/19 18:50:31 INFO client.AHSProxy: Connecting to Application History server at dicvmc1.ie.cuhk.edu.hk/172.16.5.161:10200
  23/02/19 18:50:31 INFO client.AHSProxy: Connecting to Application History server at dicvmc1.ie.cuhk.edu.hk/172.16.5.161:10200
  23/02/19 18:50:31 INFO client.ConfiguredRMFailoverProxyProvider: Failing over to rm2
  23/02/19 18:50:32 INFO mapred.FileInputFormat: Total input files to process : 100
  23/02/19 18:50:32 INFO mapreduce.JobSubmitter: number of splits:100
  23/02/19 18:50:32 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1675438250096_2012
  23/02/19 18:50:32 INFO conf.Configuration: resource-types.xml not found
  23/02/19 18:50:32 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
  23/02/19 18:50:32 INFO resource.ResourceUtils: Adding resource type - name = memory-mb, units = Mi, type = COUNTABLE
  23/02/19 18:50:32 INFO resource.ResourceUtils: Adding resource type - name = vcores, units = , type = COUNTABLE
  23/02/19 18:50:32 INFO impl.YarnClientImpl: Submitted application application_1675438250096_2012
  23/02/19 18:50:32 INFO mapreduce.Job: The url to track the job: http://dicvmc1.ie.cuhk.edu.hk:8088/proxy/application_1675438250096_2012/
  23/02/19 18:50:32 INFO mapreduce.Job: Running job: job_1675438250096_2012
  23/02/19 18:50:37 INFO mapreduce.Job: Job job_1675438250096_2012 running in uber mode : false
  23/02/19 18:50:37 INFO mapreduce.Job:  map 0% reduce 0%
```

Cmd for running taskb(mapperb2,reducerb2) in large relation



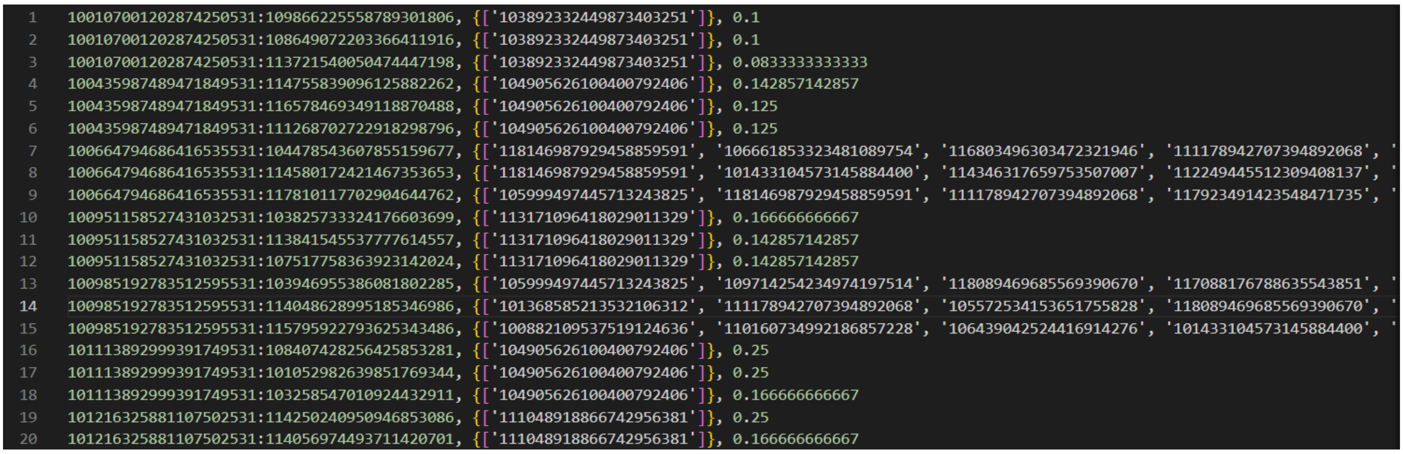
```
                Other local map tasks=1
                Data-local map tasks=99
                Rack-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=18535300
                Total time spent by all reduces in occupied slots (ms)=11626616
                Total time spent by all map tasks (ms)=4633825
                Total time spent by all reduce tasks (ms)=1453327
                Total vcore-milliseconds taken by all map tasks=4633825
                Total vcore-milliseconds taken by all reduce tasks=1453327
                Total megabyte-milliseconds taken by all map tasks=18980147200
                Total megabyte-milliseconds taken by all reduce tasks=11905654784
        Map-Reduce Framework
                Map input records=107596
                Map output records=115788204
                Map output bytes=9742156462
                Map output materialized bytes=9979799060
                Input split bytes=12600
                Combine input records=0
                Combine output records=0
                Reduce input groups=107596
                Reduce shuffle bytes=9979799060
                Reduce input records=115788204
                Reduce output records=13799850
                Spilled Records=231576408
                Shuffled Maps =5000
                Failed Shuffles=0
                Merged Map outputs=5000
                GC time elapsed (ms)=61913
                CPU time spent (ms)=4880840
                Physical memory (bytes) snapshot=192103411712
                Virtual memory (bytes) snapshot=1182839869440
                Total committed heap usage (bytes)=212368621568
        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=344311033
        File Output Format Counters
                Bytes Written=4535884430
23/02/19 19:23:01 INFO streaming.StreamJob: Output directory: ./hw1/output2
```

Result of second mapper, reducer



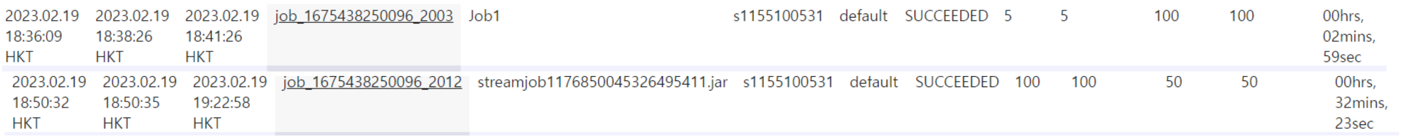
```
[s1155100531@dicvmc4 ~]$ hdfs dfs -get hw1/output2 ./hw1
[s1155100531@dicvmc4 ~]$ cat ./hw1/output2/* > ./hw1/output2/output2_sort
[s1155100531@dicvmc4 ~]$ cat ./hw1/output2/output2_sort | sort -n -k1 -k2r > ./hw1/output2/output2_res
[s1155100531@dicvmc4 ~]$ cat ./hw1/output2/output2_res | python3 ./hw1/topk.py > ./hw1/result_b
```

Cmd for copying and sorting the output of reducerb2.py for large dataset



```
1001070012028742505311:1098662255587893018O6, {['1038923324498734O3251']}, 0.1
```

Output of topk.py for large dataset



2023.02.19 18:36:09 HKT	2023.02.19 18:38:26 HKT	2023.02.19 18:41:26 HKT	job_1675438250096_2003	Job1	s1155100531	default	SUCCEEDED	5	5	100	100	00hrs, 02mins, 59sec
2023.02.19 18:50:32 HKT	2023.02.19 18:50:35 HKT	2023.02.19 19:22:58 HKT	job_1675438250096_2012	streamjob1176850045326495411.jar	s1155100531	default	SUCCEEDED	100	100	50	50	00hrs, 32mins, 23sec

Total Running time for task e (35min 22sec)