Hadoop Streaming: Clustering

myMapper_cluster.py

```
#!/usr/bin/python
import sys
import math
import numpy as np

fd = open('centers.txt', 'r')
rows = fd.readlines()
fd.close()

centerDict = {}
for row in rows:
	split = row.split('\t')
	d_centerkey = split[0]
	val = split[1:]
	centerDict[d_centerkey]=[split[1],split[2],split[3],split[4],split[5],split[6],split[7],split[8],split[
9],split[10].strip()]

for line in sys.stdin:
	line = line.strip()
	split = line.split(' ')
	rval = split[0:]
	dist_list =[]
	cluster_key = None
	temp = 0
	rowval = ','.join(rval)
	for key, value in centerDict.items():
		x1, x2, x3, x4, x5, x6, x7, x8, x9, x10 = map(float, rowval.strip().split(','))
		y1 = value[0]
		y2 = value[1]
		y3 = value[2]
		y4 = value[3]
		y5 = value[4]
		y6 = value[5]
		y7 = value[6]
		y8 = value[7]
		y9 = value[8]
		y10 = value[9]
		dist = math.sqrt((float(y1)-x1)**2 + (float(y2)-x2)**2 + (float(y3)-x3)**2 + (float(y4)-
x4)**2 + (float(y5)-x5)**2 + (float(y6)-x6)**2 + (float(y7)-x7)**2 + (float(y8)-x8)**2 + (float(y9)-
x9)**2 + (float(y10)-x10)**2)
		dist_list.append([key,dist])
	a = np.array(dist_list)
	rows = a[np.argsort(a[:,1])][0]
	cluster_key = np.int64(rows[0])
	print "%d\t%.2f\t%.2f\t%.2f\t%.2f\t%.2f\t%.2f\t%.2f\t%.2f\t%.2f\t%.2f"%
(cluster_key,float(split[0]),float(split[1]),float(split[2]),float(split[3]),float(split[4]),float(split[5]),float(
split[6]),float(split[7]),float(split[8]),float(split[9]))
```

myReducer_cluster.py

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

curr_id = None
curr_cnt = 0
id = None
val1 = 0
val2 = 0
val3 = 0
val4 = 0
val5 = 0
val6 = 0
val7 = 0
val8 = 0
val9 = 0
val10 = 0
# The input comes from standard input (line by line)
for line in sys.stdin:
   line = line.strip()
   # parse the line and split it by '\t'
   ln = line.split('\t')    # [1, 5]
   # grab the key
   id = int(ln[0])   # current received key is lo_quantity
   if curr_id == id:
      curr_cnt += 1
      val1 = val1 + float(ln[1])
      val2 = val2 + float(ln[2])
      val3 = val3 + float(ln[3])
      val4 = val4 + float(ln[4])
      val5 = val5 + float(ln[5])
      val6 = val6 + float(ln[6])
      val7 = val7 + float(ln[7])
      val8 = val8 + float(ln[8])
      val9 = val9 + float(ln[9])
      val10 = val10 + float(ln[10])
   else:
      if curr_id: # output the count, single key completed
         # NOTE: Change this to '%s\t%d' if your key is a string
         print '%d\t%.2f\t%.2f\t%.2f\t%.2f\t%.2f\t%.2f\t%.2f\t%.2f\t%.2f\t%.2f' % (curr_id,
(val1/curr_cnt), (val2/curr_cnt), (val3/curr_cnt), (val4/curr_cnt), (val5/curr_cnt), (val6/curr_cnt),
(val7/curr_cnt), (val8/curr_cnt), (val9/curr_cnt), (val10/curr_cnt))  # print 1\t2 if you saw two 1s
      curr_id = id   # Reset the current key to the new key (e.g., 6)
      curr_cnt = 1
      val1 = 0
      val2 = 0
      val3 = 0
      val4 = 0
      val5 = 0
      val6 = 0
```

```
        val7 = 0
        val8 = 0
        val9 = 0
        val10 = 0
        val1 = val1 + float(ln[1])
        val2 = val2 + float(ln[2])
        val3 = val3 + float(ln[3])
        val4 = val4 + float(ln[4])
        val5 = val5 + float(ln[5])
        val6 = val6 + float(ln[6])
        val7 = val7 + float(ln[7])
        val8 = val8 + float(ln[8])
        val9 = val9 + float(ln[9])
        val10 = val10 + float(ln[10])

# output the last key
if curr_id == id:
    print '%d\t%.2f\t%.2f\t%.2f\t%.2f\t%.2f\t%.2f\t%.2f\t%.2f\t%.2f\t%.2f' % (curr_id, (val1/curr_cnt),
(val2/curr_cnt), (val3/curr_cnt), (val4/curr_cnt), (val5/curr_cnt), (val6/curr_cnt), (val7/curr_cnt),
(val8/curr_cnt), (val9/curr_cnt), (val10/curr_cnt))
```

1st iteration

Initial centers.txt

```
[ec2-user@ip-172-31-77-124 ~]$ cat centers.txt
1       75.0    75.0    0.0     0.0     0.0     0.0     0.0     0.0     0.0     0.0
2       0.0     0.0     150.0   150.0   0.0     0.0     0.0     0.0     0.0     0.0
3       0.0     0.0     0.0     0.0     225.0   225.0   0.0     0.0     0.0     0.0
4       0.0     0.0     0.0     0.0     0.0     0.0     300.0   300.0   0.0     0.0
5       0.0     100.0   240.0   0.0     0.0     0.0     0.0     0.0     370.0   370.0
```

hadoop jar hadoop-streaming-2.6.4.jar -input /data/kmeans/ -mapper myMapper_cluster.py -file ../myMapper_cluster.py -reducer myReducer_cluster.py -file ../myReducer_cluster.py -output /data/Kmeansoutput -file ../centers.txt

```
                Data-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=44113
                Total time spent by all reduces in occupied slots (ms)=4545
                Total time spent by all map tasks (ms)=44113
                Total time spent by all reduce tasks (ms)=4545
                Total vcore-milliseconds taken by all map tasks=44113
                Total vcore-milliseconds taken by all reduce tasks=4545
                Total megabyte-milliseconds taken by all map tasks=45171712
                Total megabyte-milliseconds taken by all reduce tasks=4654080
        Map-Reduce Framework
                Map input records=350000
                Map output records=350000
                Map output bytes=24428787
                Map output materialized bytes=25128799
                Input split bytes=204
                Combine input records=0
                Combine output records=0
                Reduce input groups=5
                Reduce shuffle bytes=25128799
                Reduce input records=350000
                Reduce output records=5
                Spilled Records=700000
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=192
                CPU time spent (ms)=25650
                Physical memory (bytes) snapshot=694505472
                Virtual memory (bytes) snapshot=6388142080
                Total committed heap usage (bytes)=490209280
        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=87504096
        File Output Format Counters
                Bytes Written=360
20/11/23 04:10:58 INFO streaming.StreamJob: Output directory: /data/Kmeansoutput
```

Output:

hadoop fs -ls /data/Kmeansoutput/

```
[ec2-user@ip-172-31-77-124 ~]$ hadoop fs -ls /data/Kmeansoutput/
Found 2 items
-rw-r--r--   2 ec2-user supergroup          0 2020-11-23 04:10 /data/Kmeansoutput/_SUCCESS
-rw-r--r--   2 ec2-user supergroup        360 2020-11-23 04:10 /data/Kmeansoutput/part-00000
```

hadoop fs -cat /data/Kmeansoutput/part-00000

```
[ec2-user@ip-172-31-77-124 ~]$ hadoop fs -cat /data/Kmeansoutput/part-00000
1       286.96  282.16  274.90  281.85  277.22  278.32  276.96  278.41  274.09  272.08
2       266.11  251.68  299.30  334.79  234.31  234.86  236.49  234.96  212.10  212.46
3       251.99  241.24  212.58  240.01  324.27  324.92  220.92  220.98  208.19  207.85
4       251.96  242.13  219.35  243.52  228.28  229.69  320.91  322.05  212.61  212.71
5       235.93  252.58  268.24  228.85  220.14  222.07  220.69  221.19  299.75  299.40
```

2nd iteration

centers.txt

```
[ec2-user@ip-172-31-77-124 ~]$ hadoop fs -cat /data/Kmeansoutput/part-00000
1       286.96  282.16  274.90  281.85  277.22  278.32  276.96  278.41  274.09  272.08
2       266.11  251.68  299.30  334.79  234.31  234.86  236.49  234.96  212.10  212.46
3       251.99  241.24  212.58  240.01  324.27  324.92  220.92  220.98  208.19  207.85
4       251.96  242.13  219.35  243.52  228.28  229.69  320.91  322.05  212.61  212.71
5       235.93  252.58  268.24  228.85  220.14  222.07  220.69  221.19  299.75  299.40
```

hadoop jar hadoop-streaming-2.6.4.jar -input /data/kmeans/ -mapper myMapper_cluster.py -file ../myMapper_cluster.py -reducer myReducer_cluster.py -file ../myReducer_cluster.py -output /data/Kmeansoutput2 -file ../centers.txt

```
                Data-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=34324
                Total time spent by all reduces in occupied slots (ms)=4220
                Total time spent by all map tasks (ms)=34324
                Total time spent by all reduce tasks (ms)=4220
                Total vcore-milliseconds taken by all map tasks=34324
                Total vcore-milliseconds taken by all reduce tasks=4220
                Total megabyte-milliseconds taken by all map tasks=35147776
                Total megabyte-milliseconds taken by all reduce tasks=4321280
        Map-Reduce Framework
                Map input records=350000
                Map output records=350000
                Map output bytes=24428787
                Map output materialized bytes=25128799
                Input split bytes=204
                Combine input records=0
                Combine output records=0
                Reduce input groups=5
                Reduce shuffle bytes=25128799
                Reduce input records=350000
                Reduce output records=5
                Spilled Records=700000
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=223
                CPU time spent (ms)=25770
                Physical memory (bytes) snapshot=714661888
                Virtual memory (bytes) snapshot=6388416512
                Total committed heap usage (bytes)=493355008
        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=87504096
        File Output Format Counters
                Bytes Written=360
20/11/23 04:15:56 INFO streaming.StreamJob: Output directory: /data/Kmeansoutput2
```

Output:

hadoop fs -ls /data/Kmeansoutput2/

```
[ec2-user@ip-172-31-77-124 ~]$ hadoop fs -ls /data/Kmeansoutput2/
Found 2 items
-rw-r--r--   2 ec2-user supergroup          0 2020-11-23 04:15 /data/Kmeansoutput2/_SUCCESS
-rw-r--r--   2 ec2-user supergroup        360 2020-11-23 04:15 /data/Kmeansoutput2/part-00000
```

hadoop fs -cat /data/Kmeansoutput2/part-00000

```
[ec2-user@ip-172-31-77-124 ~]$ hadoop fs -cat /data/Kmeansoutput2/part-00000
1       320.00  316.47  296.17  300.41  312.11  313.39  309.63  311.83  324.01  317.89
2       263.40  244.77  331.09  380.19  207.79  207.71  212.61  208.63  185.61  188.72
3       239.60  230.45  185.96  210.60  349.57  349.35  197.35  197.56  198.07  197.98
4       238.33  229.18  190.17  211.94  202.70  204.07  351.43  353.97  199.83  200.73
5       213.13  243.97  263.27  188.73  193.08  196.72  195.59  196.31  331.37  331.24
```

3rd iteration

centers.txt

```
[ec2-user@ip-172-31-77-124 ~]$ hadoop fs -cat /data/Kmeansoutput2/part-00000
1       320.00  316.47  296.17  300.41  312.11  313.39  309.63  311.83  324.01  317.89
2       263.40  244.77  331.09  380.19  207.79  207.71  212.61  208.63  185.61  188.72
3       239.60  230.45  185.96  210.60  349.57  349.35  197.35  197.56  198.07  197.98
4       238.33  229.18  190.17  211.94  202.70  204.07  351.43  353.97  199.83  200.73
5       213.13  243.97  263.27  188.73  193.08  196.72  195.59  196.31  331.37  331.24
```

hadoop jar hadoop-streaming-2.6.4.jar -input /data/kmeans/ -mapper myMapper_cluster.py -file ../myMapper_cluster.py -reducer myReducer_cluster.py -file ../myReducer_cluster.py -output /data/Kmeansoutput3 -file ../centers.txt

```
                Data-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=43565
                Total time spent by all reduces in occupied slots (ms)=7892
                Total time spent by all map tasks (ms)=43565
                Total time spent by all reduce tasks (ms)=7892
                Total vcore-milliseconds taken by all map tasks=43565
                Total vcore-milliseconds taken by all reduce tasks=7892
                Total megabyte-milliseconds taken by all map tasks=44610560
                Total megabyte-milliseconds taken by all reduce tasks=8081408
        Map-Reduce Framework
                Map input records=350000
                Map output records=350000
                Map output bytes=24428787
                Map output materialized bytes=25128799
                Input split bytes=204
                Combine input records=0
                Combine output records=0
                Reduce input groups=5
                Reduce shuffle bytes=25128799
                Reduce input records=350000
                Reduce output records=5
                Spilled Records=700000
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=236
                CPU time spent (ms)=26580
                Physical memory (bytes) snapshot=733491200
                Virtual memory (bytes) snapshot=6388178944
                Total committed heap usage (bytes)=530055168
        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=87504096
        File Output Format Counters
                Bytes Written=360
20/11/23 04:22:19 INFO streaming.StreamJob: Output directory: /data/Kmeansoutput3
```

Output:

hadoop fs -ls /data/Kmeansoutput3/

```
[ec2-user@ip-172-31-77-124 ~]$ hadoop fs -ls /data/Kmeansoutput3/
Found 2 items
-rw-r--r--   2 ec2-user supergroup          0 2020-11-23 04:22 /data/Kmeansoutput3/_SUCCESS
-rw-r--r--   2 ec2-user supergroup        360 2020-11-23 04:22 /data/Kmeansoutput3/part-00000
```

hadoop fs -cat /data/Kmeansoutput3/part-00000

```
[ec2-user@ip-172-31-77-124 ~]$ hadoop fs -cat /data/Kmeansoutput3/part-00000
1       315.33  312.23  294.04  301.06  313.02  314.88  309.96  311.28  326.59  318.04
2       260.13  244.00  327.69  375.49  205.18  205.00  210.03  205.58  185.49  190.55
3       236.49  229.24  184.78  203.83  346.77  345.46  194.97  195.49  199.23  199.87
4       235.79  227.18  187.69  204.50  198.97  200.22  347.18  350.65  200.54  201.65
5       209.60  241.12  258.49  179.62  189.29  193.99  192.29  193.18  330.96  330.85
```

4th iteration

centers.txt

```
[ec2-user@ip-172-31-77-124 ~]$ hadoop fs -cat /data/Kmeansoutput3/part-00000
1       315.33  312.23  294.04  301.06  313.02  314.88  309.96  311.28  326.59  318.04
2       260.13  244.00  327.69  375.49  205.18  205.00  210.03  205.58  185.49  190.55
3       236.49  229.24  184.78  203.83  346.77  345.46  194.97  195.49  199.23  199.87
4       235.79  227.18  187.69  204.50  198.97  200.22  347.18  350.65  200.54  201.65
5       209.60  241.12  258.49  179.62  189.29  193.99  192.29  193.18  330.96  330.85
```

hadoop jar hadoop-streaming-2.6.4.jar -input /data/kmeans/ -mapper myMapper_cluster.py -file ../myMapper_cluster.py -reducer myReducer_cluster.py -file ../myReducer_cluster.py -output /data/Kmeansoutput4 -file ../centers.txt

```
                Data-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=36978
                Total time spent by all reduces in occupied slots (ms)=4599
                Total time spent by all map tasks (ms)=36978
                Total time spent by all reduce tasks (ms)=4599
                Total vcore-milliseconds taken by all map tasks=36978
                Total vcore-milliseconds taken by all reduce tasks=4599
                Total megabyte-milliseconds taken by all map tasks=37865472
                Total megabyte-milliseconds taken by all reduce tasks=4709376
        Map-Reduce Framework
                Map input records=350000
                Map output records=350000
                Map output bytes=24428787
                Map output materialized bytes=25128799
                Input split bytes=204
                Combine input records=0
                Combine output records=0
                Reduce input groups=5
                Reduce shuffle bytes=25128799
                Reduce input records=350000
                Reduce output records=5
                Spilled Records=700000
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=333
                CPU time spent (ms)=24450
                Physical memory (bytes) snapshot=732733440
                Virtual memory (bytes) snapshot=6388256768
                Total committed heap usage (bytes)=510656512
        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=87504096
        File Output Format Counters
                Bytes Written=360
20/11/23 04:37:47 INFO streaming.StreamJob: Output directory: /data/Kmeansoutput4
```

Output:

hadoop fs -ls /data/Kmeansoutput4/

```
[ec2-user@ip-172-31-77-124 hadoop-2.6.4]$ hadoop fs -ls /data/Kmeansoutput4/
Found 2 items
-rw-r--r--   2 ec2-user supergroup          0 2020-11-23 04:37 /data/Kmeansoutput4/_SUCCESS
-rw-r--r--   2 ec2-user supergroup        360 2020-11-23 04:37 /data/Kmeansoutput4/part-00000
```

hadoop fs -cat /data/Kmeansoutput4/part-00000

```
[ec2-user@ip-172-31-77-124 hadoop-2.6.4]$ hadoop fs -cat /data/Kmeansoutput4/part-00000
1       310.82  307.90  292.55  303.27  312.75  314.84  309.89  310.22  327.43  316.66
2       258.80  244.61  325.17  374.41  204.47  204.14  209.51  204.61  185.94  193.06
3       236.10  229.69  185.40  200.48  346.81  344.29  194.22  195.07  199.27  200.62
4       236.30  227.13  187.67  200.99  197.35  198.62  345.51  350.21  200.71  202.07
5       209.32  240.81  257.25  174.89  187.00  192.96  190.79  191.66  331.86  331.19
```