1. /shared-data/reviews_Books_5.json 의 전체 상품의 평균 “overall” 점수는?

```
                Bytes Written=27
root@e9a7de4b866a:~# hdfs dfs -cat /home/19/reviews_Books_5_output/part-r-00000
overall 4.2499322041784255
root@e9a7de4b866a:~#
```

```java
    @Override
    public void map(LongWritable key, Text value, Context context) throws IOException, InterruptedException {

      for (String token : value.toString().split(",")) {
        if (token.startsWith(" \"overall\"")) {
          word.set(token);
        }
      }

      for(String token: word.toString().split("\\s")) {
        word1.set(token);
      }
      double overallVal = 0;

      try{
        overallVal = Double.parseDouble(word1.toString());
      } catch(NumberFormatException nfe){
      }
      context.write(new Text("overall"),new DoubleWritable(overallVal));
    }
  }

  public static class Reduce extends Reducer<Text, DoubleWritable, Text, DoubleWritable> {
    @Override
    public void reduce(Text key, Iterable<DoubleWritable> values, Context context)
        throws IOException, InterruptedException {
      int sum = 0;
      double count = 0;
      for (DoubleWritable val : values) {
        sum += val.get();
        count+=1;
      }

      context.write(key, new DoubleWritable(sum / count));
    }
  }
}
```

처음에 리뷰 샘플을 보니 , 앞뒤로 분류가 되어있는 것을 알게 되었습니다. 그래서 처음에 split을 이용해서 구분을 해주었습니다. 그 이후에 자른 것을 바탕으로 앞에 공백이 있어서 그것을 처리 해주지 않으면 제대로 인식을 안하는 것을 알게되었습니다. 그래서 공백을 없애준 뒤에 double값으로 overallValue값을 저장하는 변수를 하나 만들어 주고 NumberFormatException을 익셉션 처리를 해주어서 제대로 숫자 변환이 일어나면 overallValue에 값을 저장하고 아니면 아무것도 해주지 않는 식으로 처리를 했습니다. 그 이후에 reduce에 넘겨주기 위해 키 값은 overall로 지정해 주고 value값은 overallVal값을 전달해 주었습니다.

다음은 reduce부분 입니다. reduce는 가져온 value값을 전부 더해주고 overall이 얼마나 많이 나왔는지 count로 센다음에 그 값으로 나눠주어 평균값을 구하는 역할을 수행합니다.

```
root@e9a7de4b866a:~# javac -classpath /usr/local/hadoop/share/hadoop/common/hadoop-common-2.8.0.jar:/usr/local/hadoop/share/hadoop/mapreduce/hadoop-mapreduce-client-core-2.8.0.jar  -d reviews_Books_5_output  WordCount1.java
root@e9a7de4b866a:~# jar -cvf WordCount1.jar -C reviews_Books_5_output/ .
added manifest
adding: kr/(in = 0) (out= 0)(stored 0%)
adding: kr/ac/(in = 0) (out= 0)(stored 0%)
adding: kr/ac/kookmin/(in = 0) (out= 0)(stored 0%)
adding: kr/ac/kookmin/cs/(in = 0) (out= 0)(stored 0%)
adding: kr/ac/kookmin/cs/bigdata/(in = 0) (out= 0)(stored 0%)
adding: kr/ac/kookmin/cs/bigdata/WordCount1$Map.class(in = 2396) (out= 1055)(deflated 55%)
adding: kr/ac/kookmin/cs/bigdata/WordCount1$Reduce.class(in = 1722) (out= 735)(deflated 57%)
adding: kr/ac/kookmin/cs/bigdata/WordCount1.class(in = 2176) (out= 1023)(deflated 52%)
root@e9a7de4b866a:~# hadoop jar WordCount1.jar kr.ac.kookmin.cs.bigdata.WordCount1 /shared-data/reviews_Books_5.json /home/19/reviews_Books_5_output
[/shared-data/reviews_Books_5.json, /home/19/reviews_Books_5_output]
[/shared-data/reviews_Books_5.json, /home/19/reviews_Books_5_output]
18/04/14 06:08:22 INFO client.RMProxy: Connecting to ResourceManager at master/10.100.100.2:8050
18/04/14 06:08:22 INFO input.FileInputFormat: Total input files to process : 1
18/04/14 06:08:23 INFO mapreduce.JobSubmitter: number of splits:71
18/04/14 06:08:23 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1522171139460_0527
18/04/14 06:08:23 INFO impl.YarnClientImpl: Submitted application application_1522171139460_0527
18/04/14 06:08:23 INFO mapreduce.Job: The url to track the job: http://master:8088/proxy/application_1522171139460_0527/
18/04/14 06:08:23 INFO mapreduce.Job: Running job: job_1522171139460_0527
18/04/14 06:08:29 INFO mapreduce.Job: Job job_1522171139460_0527 running in uber mode : false
18/04/14 06:08:29 INFO mapreduce.Job:  map 0% reduce 0%
18/04/14 06:08:36 INFO mapreduce.Job:  map 1% reduce 0%
18/04/14 06:08:37 INFO mapreduce.Job:  map 11% reduce 0%
18/04/14 06:08:38 INFO mapreduce.Job:  map 20% reduce 0%
18/04/14 06:08:39 INFO mapreduce.Job:  map 30% reduce 0%
18/04/14 06:08:40 INFO mapreduce.Job:  map 46% reduce 0%
18/04/14 06:08:41 INFO mapreduce.Job:  map 58% reduce 0%
18/04/14 06:08:42 INFO mapreduce.Job:  map 72% reduce 0%
18/04/14 06:08:43 INFO mapreduce.Job:  map 94% reduce 0%
18/04/14 06:08:44 INFO mapreduce.Job:  map 97% reduce 0%
18/04/14 06:08:45 INFO mapreduce.Job:  map 100% reduce 0%
18/04/14 06:08:53 INFO mapreduce.Job:  map 100% reduce 100%
18/04/14 06:08:53 INFO mapreduce.Job: Job job_1522171139460_0527 completed successfully
18/04/14 06:08:53 INFO mapreduce.Job: Counters: 51
        File System Counters
                FILE: Number of bytes read=160164744
                FILE: Number of bytes written=330192906
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=9458312483
                HDFS: Number of bytes written=27
                HDFS: Number of read operations=216
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Killed map tasks=1
                Launched map tasks=71
                Launched reduce tasks=1
                Data-local map tasks=56
                Rack-local map tasks=15
                Total time spent by all maps in occupied slots (ms)=655892
                Total time spent by all reduces in occupied slots (ms)=25386
                Total time spent by all map tasks (ms)=655892
                Total time spent by all reduce tasks (ms)=12693
                Total vcore-milliseconds taken by all map tasks=655892
                Total vcore-milliseconds taken by all reduce tasks=12693
                Total megabyte-milliseconds taken by all map tasks=671633408
                Total megabyte-milliseconds taken by all reduce tasks=25995264
        Map-Reduce Framework
```

```
                Total megabyte-milliseconds taken by all reduce tasks=25995264
        Map-Reduce Framework
                Map input records=8898041
                Map output records=8898041
                Map output bytes=142368656
                Map output materialized bytes=160165164
                Input split bytes=8236
                Combine input records=0
                Combine output records=0
                Reduce input groups=1
                Reduce shuffle bytes=160165164
                Reduce input records=8898041
                Reduce output records=1
                Spilled Records=17796082
                Shuffled Maps =71
                Failed Shuffles=0
                Merged Map outputs=71
                GC time elapsed (ms)=166029
                CPU time spent (ms)=822930
                Physical memory (bytes) snapshot=42708770816
                Virtual memory (bytes) snapshot=202004533248
                Total committed heap usage (bytes)=58948845568
        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=9458304167
        File Output Format Counters
                Bytes Written=27
oot@e9a7de4b866a:~# hdfs dfs -cat /home/19/reviews_Books_5_output/part-r-00000
```

2. /shared-data/reviews_Books_5.json 에서 가장 많은 리뷰를 남긴 사용자의 아이디(“reviewerID”) 및 리뷰 횟수는 ?

```
[root@e9a7de4b866a:~# hdfs dfs -cat /home/19/reviews_Books_5_num2/part-r-00000
"AFVQZQ8PW0L"    23222
"A00050443V8RIN8AP25G8" 5
```

```
root@e9a7de4b866a:~# javac -classpath /usr/local/hadoop/share/hadoop/common/hadoop-common-2.8.0.jar:/usr/local/hadoop/share/hadoop/mapreduce/hadoop-mapreduce-client-core-2.8.0.jar  -d reviews_Books_5_output/  ReviewsCount.java
root@e9a7de4b866a:~# jar -cvf ReviewsCount.jar -C reviews_Books_5_output/ .
added manifest
adding: kr/(in = 0) (out= 0)(stored 0%)
adding: kr/ac/(in = 0) (out= 0)(stored 0%)
adding: kr/ac/kookmin/(in = 0) (out= 0)(stored 0%)
adding: kr/ac/kookmin/cs/(in = 0) (out= 0)(stored 0%)
adding: kr/ac/kookmin/cs/bigdata/(in = 0) (out= 0)(stored 0%)
adding: kr/ac/kookmin/cs/bigdata/ReviewsCount.class(in = 2181) (out= 1023)(deflated 53%)
adding: kr/ac/kookmin/cs/bigdata/WordCount1$Map.class(in = 2396) (out= 1055)(deflated 55%)
adding: kr/ac/kookmin/cs/bigdata/ReviewsCount$Reduce1.class(in = 1959) (out= 845)(deflated 56%)
adding: kr/ac/kookmin/cs/bigdata/ReviewsCount$Reduce.class(in = 2352) (out= 958)(deflated 59%)
adding: kr/ac/kookmin/cs/bigdata/WordCount1$Reduce.class(in = 1722) (out= 735)(deflated 57%)
adding: kr/ac/kookmin/cs/bigdata/ReviewsCount$Map.class(in = 2116) (out= 915)(deflated 56%)
adding: kr/ac/kookmin/cs/bigdata/WordCount1.class(in = 2176) (out= 1023)(deflated 52%)
root@e9a7de4b866a:~# hadoop jar ReviewsCount.jar kr.ac.kookmin.cs.bigdata.ReviewsCount /shared-data/reviews_Books_5.json /home/19/reviews_Books_5_num2
[/shared-data/reviews_Books_5.json, /home/19/reviews_Books_5_num2]
[/shared-data/reviews_Books_5.json, /home/19/reviews_Books_5_num2]
18/04/14 10:13:33 INFO client.RMProxy: Connecting to ResourceManager at master/10.100.100.2:8050
18/04/14 10:13:34 INFO input.FileInputFormat: Total input files to process : 1
18/04/14 10:13:34 INFO mapreduce.JobSubmitter: number of splits:71
18/04/14 10:13:34 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1522171139460_0589
18/04/14 10:13:34 INFO impl.YarnClientImpl: Submitted application application_1522171139460_0589
18/04/14 10:13:34 INFO mapreduce.Job: The url to track the job: http://master:8088/proxy/application_1522171139460_0589/
18/04/14 10:13:34 INFO mapreduce.Job: Running job: job_1522171139460_0589
18/04/14 10:13:42 INFO mapreduce.Job: Job job_1522171139460_0589 running in uber mode : false
18/04/14 10:13:42 INFO mapreduce.Job:  map 0% reduce 0%
18/04/14 10:13:54 INFO mapreduce.Job:  map 3% reduce 0%
18/04/14 10:13:55 INFO mapreduce.Job:  map 11% reduce 0%
18/04/14 10:13:56 INFO mapreduce.Job:  map 38% reduce 0%
18/04/14 10:13:57 INFO mapreduce.Job:  map 51% reduce 0%
18/04/14 10:13:58 INFO mapreduce.Job:  map 52% reduce 0%
18/04/14 10:13:59 INFO mapreduce.Job:  map 56% reduce 0%
18/04/14 10:14:00 INFO mapreduce.Job:  map 85% reduce 0%
18/04/14 10:14:01 INFO mapreduce.Job:  map 90% reduce 0%
18/04/14 10:14:02 INFO mapreduce.Job:  map 93% reduce 0%
18/04/14 10:14:04 INFO mapreduce.Job:  map 100% reduce 0%
18/04/14 10:14:12 INFO mapreduce.Job:  map 100% reduce 84%
18/04/14 10:14:14 INFO mapreduce.Job:  map 100% reduce 100%
18/04/14 10:14:14 INFO mapreduce.Job: Job job_1522171139460_0589 completed successfully
18/04/14 10:14:14 INFO mapreduce.Job: Counters: 51
        File System Counters
                FILE: Number of bytes read=202322713
                FILE: Number of bytes written=414508916
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=9458312403
                HDFS: Number of bytes written=46
                HDFS: Number of read operations=216
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Killed map tasks=1
                Launched map tasks=71
                Launched reduce tasks=1
                Data-local map tasks=56
                Rack-local map tasks=15
                Total time spent by all maps in occupied slots (ms)=1003415
                Total time spent by all reduces in occupied slots (ms)=31786
                Total time spent by all map tasks (ms)=1003415
                Total time spent by all reduce tasks (ms)=15893
```

```
                Total vcore-milliseconds taken by all map tasks=1003415
                Total vcore-milliseconds taken by all reduce tasks=15893
                Total megabyte-milliseconds taken by all map tasks=1027496960
                Total megabyte-milliseconds taken by all reduce tasks=32548864
        Map-Reduce Framework
                Map input records=8898041
                Map output records=8898041
                Map output bytes=184526625
                Map output materialized bytes=202323133
                Input split bytes=8236
                Combine input records=0
                Combine output records=0
                Reduce input groups=603668
                Reduce shuffle bytes=202323133
                Reduce input records=8898041
                Reduce output records=2
                Spilled Records=17796082
                Shuffled Maps =71
                Failed Shuffles=0
                Merged Map outputs=71
                GC time elapsed (ms)=165616
                CPU time spent (ms)=773080
                Physical memory (bytes) snapshot=41866350592
                Virtual memory (bytes) snapshot=201962962944
                Total committed heap usage (bytes)=58684604416
        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=9458304167
        File Output Format Counters
                Bytes Written=46
root@e9a7de4b866a:~# hdfs dfs -cat /home/19/reviews_Books_5_num2/part-r-00000
```

```
53    public static class Map extends Mapper<LongWritable, Text, Text, IntWritable> {
54    private final static IntWritable ONE = new IntWritable(1);
55    private Text word = new Text();
56
57    @Override
58    public void map(LongWritable key, Text value, Context context) throws IOException, InterruptedException {
59
60            for (String token : value.toString().split(",")) {
61                    if (token.startsWith("{\"reviewerID\"")) {
62                            for(String token1: token.split(" ")) {
63                                    word.set(token1);
64                            }
65                    }
66            }
67
68            context.write(word,ONE);
69    }
70    }
```

map에서는 먼저 , 단위로 텍스트 파일을 잘라준 후{“reviewerID” 를 기준으로 또 한번 짤라준다.
그 이후에 Text에 set 시킨 후 reduce로 넘겨준다. 그 때 valuer값으로 1을 넣어준다.

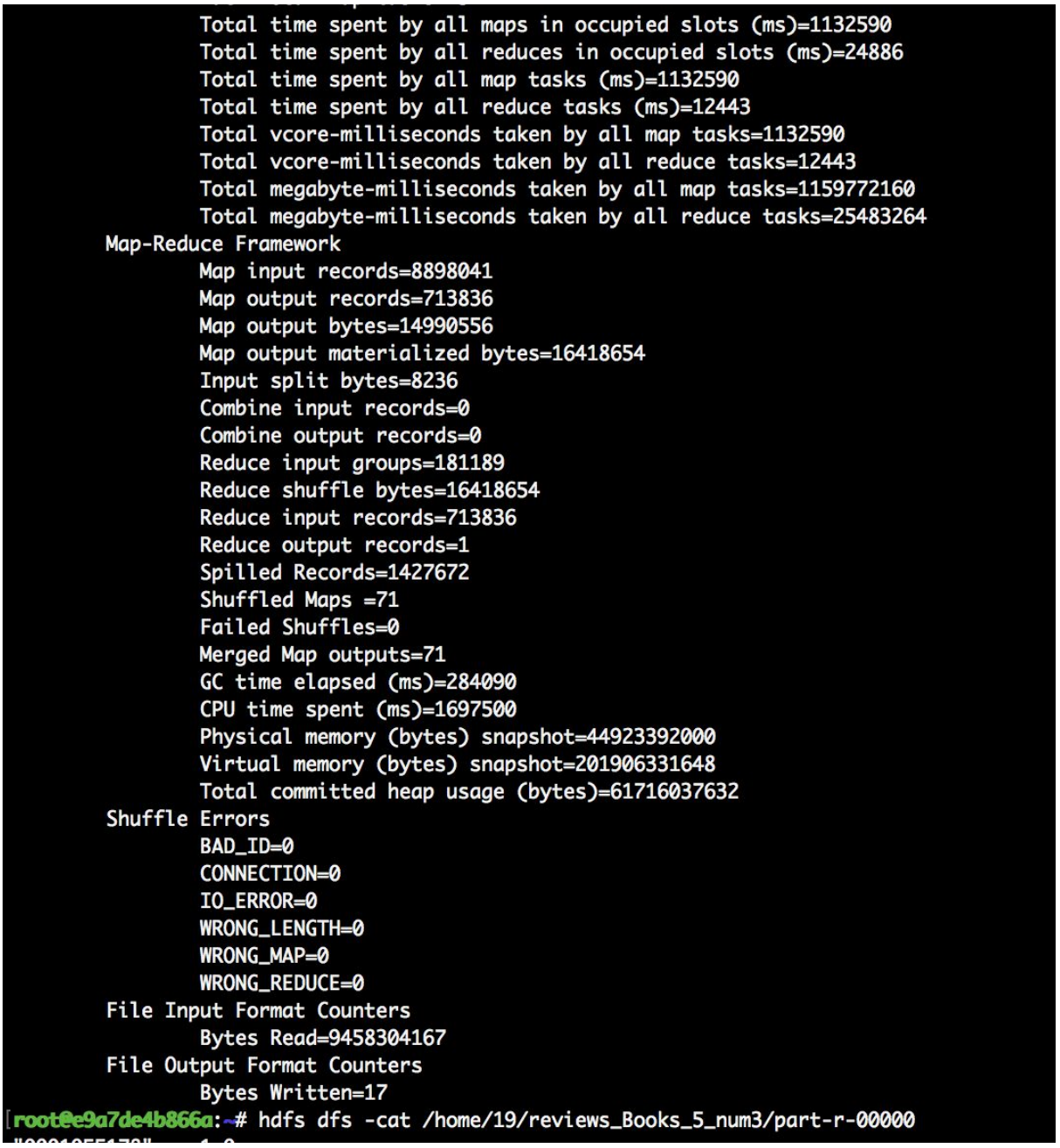
```
72    public static class Reduce extends Reducer<Text, IntWritable, Text, IntWritable> {
73    private Text maxID = new Text();
74    private int maxVal = 0;
75    private Text minID = new Text();
76    private int minVal = Integer.MAX_VALUE;
77
78
79    @Override
80    public void reduce(Text key, Iterable<IntWritable> values, Context context)
81    throws IOException, InterruptedException {
82
83            int sum = 0;
84            for (IntWritable val : values) {
85                    sum += val.get();
86            }
87
88            if(sum < minVal) {
89                    minVal = sum;
90                    minID.set(key);
91            }
92
93            if(sum > maxVal) {
94                    maxVal = sum;
95                    maxID.set(key);
96            }
97
98
99    }
100
101   @Override
102   protected void cleanup(Context context) throws IOException, InterruptedException {
103           context.write(maxID, new IntWritable(maxVal));
104           context.write(minID, new IntWritable(minVal));
105   }
106
107   }
```

reduce부분에서는 max값에 해당하는 아이디를 저장하기 위해 maxID, 값을 저장하기 위해 maxVal을 선언해주었다. 같은 이유로 minID, minVal을 선언해주었다. 이때 minVal은 처음에 들어오는 값보다 작아지기 위하여 Integer.MAX_VALUE값으로 할당해 주었다. 그 이후 먼저

얼마나 리뷰를 썻는지 알아내기 위해 sum을 이용하여 리뷰 쓴 값을 알아내준다. 그리고 다음에 들어오는 값과 이전에 들어오는 값을 비교하여 min, max값을 알아내주는 if문을 작성해주었다.
마지막으로 reduce함수 안에서 context.write를 해버리면 모든 아이디에 해당하는 maxID,maxVal,minID,minVal값이 나오기 때문에 reduce작업이 끝난다음에 값을 보여줄 수 있도록 cleanup함수안에 context.write를 작성해주어 원하는 값만 보이도록 해주었다.

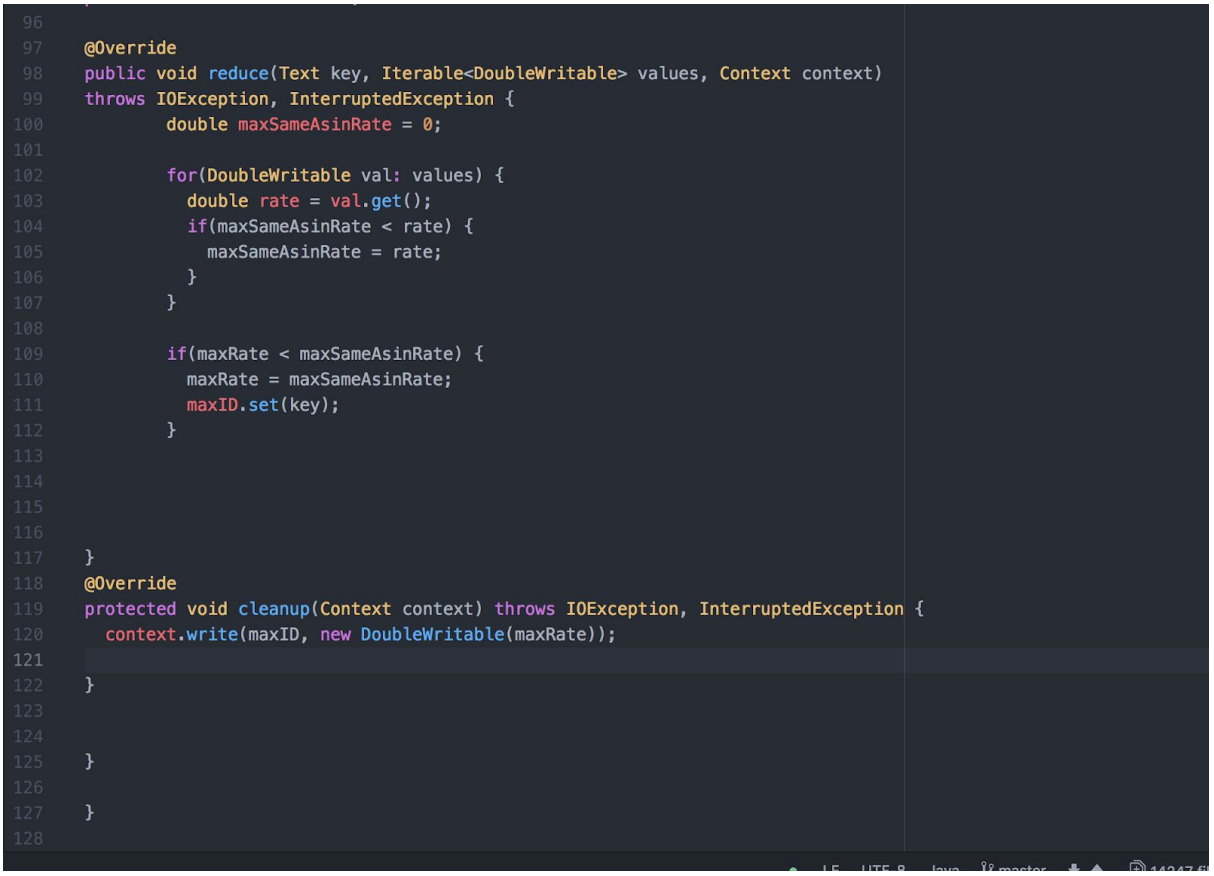
3. /shared-data/reviews_Books_5.json 에서 helpful 필드는 [a,b] 형식을 가지며, b 명의 사용자가 해당 리뷰가 도움이 되는지 투표했다는 의미이며, 이중 a 명의 사용자가 도움이 된다는 의견을 남겼다는 의미입니다. b 값이 10보다 큰 사용자 중에서 도움이 된다고 하는 사람들의 비율 (a/b) 이 가장 높은 아이템 (asin) 의 아이디는?


```
root@e9a7de4b866a:~# javac -classpath /usr/local/hadoop/share/hadoop/common/hadoop-common-2.8.0.jar:/usr/local/hadoop/share/hadoop/mapreduce/hadoop-mapreduce-client-core-2.8.0.jar  -d reviews_Books_5_output/  Helpful.java
root@e9a7de4b866a:~# jar -cvf Helpful.jar -C reviews_Books_5_output/ .
added manifest
adding: kr/(in = 0) (out= 0)(stored 0%)
adding: kr/ac/(in = 0) (out= 0)(stored 0%)
adding: kr/ac/kookmin/(in = 0) (out= 0)(stored 0%)
adding: kr/ac/kookmin/cs/(in = 0) (out= 0)(stored 0%)
adding: kr/ac/kookmin/cs/bigdata/(in = 0) (out= 0)(stored 0%)
adding: kr/ac/kookmin/cs/bigdata/Helpful.class(in = 2164) (out= 1021)(deflated 52%)
adding: kr/ac/kookmin/cs/bigdata/ReviewsCount.class(in = 2181) (out= 1023)(deflated 53%)
adding: kr/ac/kookmin/cs/bigdata/WordCount1$Map.class(in = 2396) (out= 1055)(deflated 55%)
adding: kr/ac/kookmin/cs/bigdata/Helpful$Reduce.class(in = 2216) (out= 892)(deflated 59%)
adding: kr/ac/kookmin/cs/bigdata/ReviewsCount$Reduce1.class(in = 1959) (out= 845)(deflated 56%)
adding: kr/ac/kookmin/cs/bigdata/ReviewsCount$Reduce.class(in = 2352) (out= 958)(deflated 59%)
adding: kr/ac/kookmin/cs/bigdata/WordCount1$Reduce.class(in = 1722) (out= 735)(deflated 57%)
adding: kr/ac/kookmin/cs/bigdata/ReviewsCount$Map.class(in = 2116) (out= 915)(deflated 56%)
adding: kr/ac/kookmin/cs/bigdata/Helpful$Map.class(in = 2362) (out= 1064)(deflated 54%)
adding: kr/ac/kookmin/cs/bigdata/WordCount1.class(in = 2176) (out= 1023)(deflated 52%)
root@e9a7de4b866a:~# hadoop jar Helpful.jar kr.ac.kookmin.cs.bigdata.Helpful /shared-data/reviews_Books_5.json /home/19/reviews_Books_5_num3
[/shared-data/reviews_Books_5.json, /home/19/reviews_Books_5_num3]
[/shared-data/reviews_Books_5.json, /home/19/reviews_Books_5_num3]
18/04/15 10:12:43 INFO client.RMProxy: Connecting to ResourceManager at master/10.100.100.2:8050
18/04/15 10:12:44 INFO input.FileInputFormat: Total input files to process : 1
18/04/15 10:12:44 INFO mapreduce.JobSubmitter: number of splits:71
18/04/15 10:12:44 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1522171139460_1061
18/04/15 10:12:44 INFO impl.YarnClientImpl: Submitted application application_1522171139460_1061
18/04/15 10:12:44 INFO mapreduce.Job: The url to track the job: http://master:8088/proxy/application_1522171139460_1061/
18/04/15 10:12:44 INFO mapreduce.Job: Running job: job_1522171139460_1061
18/04/15 10:12:50 INFO mapreduce.Job: Job job_1522171139460_1061 running in uber mode : false
18/04/15 10:12:50 INFO mapreduce.Job:  map 0% reduce 0%
18/04/15 10:13:00 INFO mapreduce.Job:  map 1% reduce 0%
18/04/15 10:13:01 INFO mapreduce.Job:  map 3% reduce 0%
18/04/15 10:13:02 INFO mapreduce.Job:  map 11% reduce 0%
18/04/15 10:13:03 INFO mapreduce.Job:  map 21% reduce 0%
18/04/15 10:13:04 INFO mapreduce.Job:  map 24% reduce 0%
18/04/15 10:13:07 INFO mapreduce.Job:  map 28% reduce 0%
18/04/15 10:13:09 INFO mapreduce.Job:  map 32% reduce 0%
18/04/15 10:13:10 INFO mapreduce.Job:  map 73% reduce 0%
18/04/15 10:13:11 INFO mapreduce.Job:  map 82% reduce 0%
18/04/15 10:13:12 INFO mapreduce.Job:  map 91% reduce 0%
18/04/15 10:13:13 INFO mapreduce.Job:  map 97% reduce 0%
18/04/15 10:13:14 INFO mapreduce.Job:  map 100% reduce 0%
18/04/15 10:13:18 INFO mapreduce.Job:  map 100% reduce 100%
18/04/15 10:13:20 INFO mapreduce.Job: Job job_1522171139460_1061 completed successfully
18/04/15 10:13:20 INFO mapreduce.Job: Counters: 51
        File System Counters
                FILE: Number of bytes read=16418234
                FILE: Number of bytes written=42699094
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=9458312403
                HDFS: Number of bytes written=17
                HDFS: Number of read operations=216
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Killed map tasks=1
                Launched map tasks=71
                Launched reduce tasks=1
                Data-local map tasks=58
                Rack-local map tasks=13
```


Total time spent by all maps in occupied slots (ms)=1132590
Total time spent by all reduces in occupied slots (ms)=24886
Total time spent by all map tasks (ms)=1132590
Total time spent by all reduce tasks (ms)=12443
Total vcore-milliseconds taken by all map tasks=1132590
Total vcore-milliseconds taken by all reduce tasks=12443
Total megabyte-milliseconds taken by all map tasks=1159772160
Total megabyte-milliseconds taken by all reduce tasks=25483264
Map-Reduce Framework
Map input records=8898041
Map output records=713836
Map output bytes=14990556
Map output materialized bytes=16418654
Input split bytes=8236
Combine input records=0
Combine output records=0
Reduce input groups=181189
Reduce shuffle bytes=16418654
Reduce input records=713836
Reduce output records=1
Spilled Records=1427672
Shuffled Maps =71
Failed Shuffles=0
Merged Map outputs=71
GC time elapsed (ms)=284090
CPU time spent (ms)=1697500
Physical memory (bytes) snapshot=44923392000
Virtual memory (bytes) snapshot=201906331648
Total committed heap usage (bytes)=61716037632
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=9458304167
File Output Format Counters
Bytes Written=17
root@e9a7de4b866a:~# hdfs dfs -cat /home/19/reviews_Books_5_num3/part-r-00000

```
58    @Override
59    public void map(LongWritable key, Text value, Context context) throws IOException, InterruptedException {
60
61            for(String token: value.toString().split("\n")){
62              word.set(token.replaceAll(" ", ""));
63            }
64
65            String[] strArr = word.toString().split(",");
66            String asin = strArr[1];
67            String[] asinArr = asin.split(":");
68            asin = asinArr[1];
69            asinText.set(asin);
70
71            String[] helpArr = word.toString().split("\"helpful\"");
72            helpArr= helpArr[1].split(":");
73            helpArr = helpArr[1].split("\\[");
74            helpArr = helpArr[1].split("\\]");
75            helpArr = helpArr[0].split(",");
76
77            int a = Integer.parseInt(helpArr[0]);
78            int b = Integer.parseInt(helpArr[1]);
79
80            double rate = 0;
81
82            if(b == 0) {
83              rate = 0;
84            } else {
85              rate = a / (double)b;
86            }
87            if(b > 10) {
88              context.write(asinText,new DoubleWritable(rate));
89            }
90    }
91    }
92
93    public static class Reduce extends Reducer<Text, DoubleWritable, Text, DoubleWritable> {
94    private Text maxID = new Text();
95    private double maxRate = 0;
```

먼저 맵 부분에서 helpful에 해당하는 값과 asin에 해당하는 값을 알아내기 위하여 텍스트 값을 String배열에 split해서 넣은다음 asin에 맞는 아이디 값을 asinText에 저장을 해주었다. 이후에 helpful의 값에 해당하는 부분을 찾기위해 여러번 split을 통해 a,b값을 각각 따로 지정해 주었다.
그리고 rate값을 a/b를 통해 넣어주었고 b값이 10이 넘을 경우에만 reduce로 넘어갈 수 있게 설정해주었다. 이때 예외처리를 해준 부분이 b값이 0이라면 rate를 계산하는 부분에서 오류가 날 수 있기때문에 b값이 0이라면 rate값을 아예 0으로 처리해주었다.



```
96
97    @Override
98    public void reduce(Text key, Iterable<DoubleWritable> values, Context context)
99    throws IOException, InterruptedException {
100           double maxSameAsinRate = 0;
101
102           for(DoubleWritable val: values) {
103             double rate = val.get();
104             if(maxSameAsinRate < rate) {
105               maxSameAsinRate = rate;
106             }
107           }
108
109           if(maxRate < maxSameAsinRate) {
110             maxRate = maxSameAsinRate;
111             maxID.set(key);
112           }
113
114
115
116
117   }
118   @Override
119   protected void cleanup(Context context) throws IOException, InterruptedException {
120     context.write(maxID, new DoubleWritable(maxRate));
121
122   }
123
124
125   }
126
127   }
128
```

reduce부분에서는 먼저 같은 asin아이디를 가지고 있는 부분에서 먼저 max값을 구해준 다음에 다른 asin값에서도 비교를 해주어서 제일 큰 값에 해당하는 asin값은 maxID에, rate값은 maxRate에 넣어주었다.
여기도 2번과 마찬가지로 cleanup함수를 사용해서 마지막 값에 해당하는 것만 표현해 주도록 해주었다.

4. /shared-data/reviews_Books_5.json 에서 각 reviewer 별로 helpful 필드 값을 모았을때 ([a,b] 를 a 값과 b 값 으로 각각 더함), 가장 높은 a 값을 가지는 사용자의 (reviewerID) 아이디는?

```
root@e9a7de4b866a:~# hdfs dfs -cat /home/19/reviews_Books_5_num4/part-r-00000
"AFVQZQ8PW0L"    95402
root@e9a7de4b866a:~#
```

```
root@e9a7de4b866a:~# hadoop jar HelpfulReviewer.jar kr.ac.kookmin.cs.bigdata.HelpfulReviewer /shared-data/reviews_Books_5.json /home/19/reviews_Books_5_num4
[/shared-data/reviews_Books_5.json, /home/19/reviews_Books_5_num4]
[/shared-data/reviews_Books_5.json, /home/19/reviews_Books_5_num4]
18/04/15 10:42:08 INFO client.RMProxy: Connecting to ResourceManager at master/10.100.100.2:8050
18/04/15 10:42:09 INFO input.FileInputFormat: Total input files to process : 1
18/04/15 10:42:09 INFO mapreduce.JobSubmitter: number of splits:71
18/04/15 10:42:09 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1522171139460_1080
18/04/15 10:42:09 INFO impl.YarnClientImpl: Submitted application application_1522171139460_1080
18/04/15 10:42:09 INFO mapreduce.Job: The url to track the job: http://master:8088/proxy/application_1522171139460_1080/
18/04/15 10:42:09 INFO mapreduce.Job: Running job: job_1522171139460_1080
18/04/15 10:42:15 INFO mapreduce.Job: Job job_1522171139460_1080 running in uber mode : false
18/04/15 10:42:15 INFO mapreduce.Job:  map 0% reduce 0%
18/04/15 10:42:25 INFO mapreduce.Job:  map 1% reduce 0%
18/04/15 10:42:26 INFO mapreduce.Job:  map 3% reduce 0%
18/04/15 10:42:27 INFO mapreduce.Job:  map 8% reduce 0%
18/04/15 10:42:28 INFO mapreduce.Job:  map 18% reduce 0%
18/04/15 10:42:29 INFO mapreduce.Job:  map 20% reduce 0%
18/04/15 10:42:31 INFO mapreduce.Job:  map 25% reduce 0%
18/04/15 10:42:32 INFO mapreduce.Job:  map 32% reduce 0%
18/04/15 10:42:33 INFO mapreduce.Job:  map 49% reduce 0%
18/04/15 10:42:34 INFO mapreduce.Job:  map 54% reduce 0%
18/04/15 10:42:35 INFO mapreduce.Job:  map 75% reduce 0%
18/04/15 10:42:36 INFO mapreduce.Job:  map 78% reduce 0%
18/04/15 10:42:37 INFO mapreduce.Job:  map 84% reduce 0%
18/04/15 10:42:38 INFO mapreduce.Job:  map 97% reduce 0%
18/04/15 10:42:39 INFO mapreduce.Job:  map 100% reduce 0%
18/04/15 10:42:44 INFO mapreduce.Job:  map 100% reduce 68%
18/04/15 10:42:48 INFO mapreduce.Job:  map 100% reduce 100%
18/04/15 10:42:49 INFO mapreduce.Job: Job job_1522171139460_1080 completed successfully
18/04/15 10:42:49 INFO mapreduce.Job: Counters: 51
        File System Counters
                FILE: Number of bytes read=202322713
                FILE: Number of bytes written=414509564
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=9458312403
                HDFS: Number of bytes written=20
                HDFS: Number of read operations=216
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Killed map tasks=1
                Launched map tasks=71
                Launched reduce tasks=1
                Data-local map tasks=60
                Rack-local map tasks=11
                Total time spent by all maps in occupied slots (ms)=1099723
                Total time spent by all reduces in occupied slots (ms)=36770
                Total time spent by all map tasks (ms)=1099723
```

```
        Total time spent by all reduce tasks (ms)=18385
        Total vcore-milliseconds taken by all map tasks=1099723
        Total vcore-milliseconds taken by all reduce tasks=18385
        Total megabyte-milliseconds taken by all map tasks=1126116352
        Total megabyte-milliseconds taken by all reduce tasks=37652480
    Map-Reduce Framework
        Map input records=8898041
        Map output records=8898041
        Map output bytes=184526625
        Map output materialized bytes=202323133
        Input split bytes=8236
        Combine input records=0
        Combine output records=0
        Reduce input groups=603668
        Reduce shuffle bytes=202323133
        Reduce input records=8898041
        Reduce output records=1
        Spilled Records=17796082
        Shuffled Maps =71
        Failed Shuffles=0
        Merged Map outputs=71
        GC time elapsed (ms)=247236
        CPU time spent (ms)=1635230
        Physical memory (bytes) snapshot=45115965440
        Virtual memory (bytes) snapshot=201911402496
        Total committed heap usage (bytes)=61716561920
    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=9458304167
    File Output Format Counters
        Bytes Written=20
```

```java
@Override
public void map(LongWritable key, Text value, Context context) throws IOException, InterruptedException {

        for(String token: value.toString().split("\n")){
          word.set(token.replaceAll(" ", ""));
        }

        String[] strArr = word.toString().split(",");
        String reviewer = strArr[0];
        String[] reviewerArr = reviewer.split(":");
        reviewer = reviewerArr[1];
        reviewerText.set(reviewer);

        String[] helpArr = word.toString().split("\"helpful\"");
        helpArr= helpArr[1].split(":");
        helpArr = helpArr[1].split("\\[");
        helpArr = helpArr[1].split("\\]");
        helpArr = helpArr[0].split(",");

        int a = Integer.parseInt(helpArr[0]);

        context.write(reviewerText, new IntWritable(a));


  }
}
```

```java
public static class Reduce extends Reducer<Text, IntWritable, Text, IntWritable> {
private Text maxID = new Text();
private int maxHelpful = 0;

@Override
public void reduce(Text key, Iterable<IntWritable> values, Context context)
throws IOException, InterruptedException {
        int sum = 0;

        for(IntWritable val: values) {
          sum += val.get();
        }

        if(maxHelpful < sum) {
          maxHelpful = sum;
          maxID.set(key);
        }
      }
      @Override
      protected void cleanup(Context context) throws IOException, InterruptedException {
        context.write(maxID, new IntWritable(maxHelpful));
      }
    }
}
```

맵 부분에서는 3번과 크게 달라질게 없다. 여기서 달라진 점은 asin값을 구하는 것에서 reviewer의 아이디를 구하는 것이다. reviewer의 아이디를 구한다음 helpful에서의 a값을 구한다. 그 이후 reducer로 reviewer의 아이디와 helpful의 a값을 context에 넣어서 reducer로 넘겨준다.

reducer에서는 넘겨온 int값인 a를 모두 더해주는 작업을 하고 있다. 이 작업은 같은 아이디를 가진 reveiwer의 helpful에서 a값을 모두 더해주는 역할을 한다. 그 이후 이 전과 동일한 방법으로 다른 아이디를 가진 reviewer들끼리의 a값의 합을 비교해서 최대값을 찾아주는 역할을 하고 있다.

cleanup함수를 또 만들어서 최대값을 가진 reviewer아이디와 최대값 만을 화면에 출력해준다.

5. reviewer 아이디를 통해 그 사람이 얼마나 물건에 대한 리뷰를 작성했는지 파악하여 최대로 작성한 사람의 아이디가 무엇인지 출력한다..(원래는 작성한 물건에 대한 asin값을 출력하고 싶었으나 그 값을 모두 출력하는데 오래 걸리므로 내용을 바꿨다.)

```
root@e9a7de4b866a:~# hdfs dfs -cat /home/19/reviews_Books_5_num5/part-r-00000
"AFVQZQ8PW0L"   23222
root@e9a7de4b866a:~#
```

```
root@e9a7de4b866a:~# hadoop jar MyMapReducer.jar kr.ac.kookmin.cs.bigdata.MyMapReducer /shared-data/reviews_Books_5.json /home/19/reviews_Books_5_num5
[/shared-data/reviews_Books_5.json, /home/19/reviews_Books_5_num5]
[/shared-data/reviews_Books_5.json, /home/19/reviews_Books_5_num5]
18/04/15 12:13:42 INFO client.RMProxy: Connecting to ResourceManager at master/10.100.100.2:8050
18/04/15 12:13:43 INFO input.FileInputFormat: Total input files to process : 1
18/04/15 12:13:43 INFO mapreduce.JobSubmitter: number of splits:71
18/04/15 12:13:43 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1522171139460_1153
18/04/15 12:13:43 INFO impl.YarnClientImpl: Submitted application application_1522171139460_1153
18/04/15 12:13:43 INFO mapreduce.Job: The url to track the job: http://master:8088/proxy/application_1522171139460_1153/
18/04/15 12:13:43 INFO mapreduce.Job: Running job: job_1522171139460_1153
18/04/15 12:13:49 INFO mapreduce.Job: Job job_1522171139460_1153 running in uber mode : false
18/04/15 12:13:49 INFO mapreduce.Job:  map 0% reduce 0%
18/04/15 12:14:01 INFO mapreduce.Job:  map 3% reduce 0%
18/04/15 12:14:02 INFO mapreduce.Job:  map 7% reduce 0%
18/04/15 12:14:03 INFO mapreduce.Job:  map 20% reduce 0%
18/04/15 12:14:04 INFO mapreduce.Job:  map 35% reduce 0%
18/04/15 12:14:05 INFO mapreduce.Job:  map 38% reduce 0%
18/04/15 12:14:06 INFO mapreduce.Job:  map 55% reduce 0%
18/04/15 12:14:07 INFO mapreduce.Job:  map 75% reduce 0%
18/04/15 12:14:08 INFO mapreduce.Job:  map 98% reduce 0%
18/04/15 12:14:09 INFO mapreduce.Job:  map 100% reduce 0%
18/04/15 12:14:18 INFO mapreduce.Job:  map 100% reduce 100%
18/04/15 12:14:18 INFO mapreduce.Job: Job job_1522171139460_1153 completed successfully
18/04/15 12:14:18 INFO mapreduce.Job: Counters: 51
        File System Counters
                FILE: Number of bytes read=282405082
                FILE: Number of bytes written=574673150
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=9458312403
                HDFS: Number of bytes written=20
                HDFS: Number of read operations=216
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Killed map tasks=1
                Launched map tasks=71
                Launched reduce tasks=1
                Data-local map tasks=56
                Rack-local map tasks=15
                Total time spent by all maps in occupied slots (ms)=985153
                Total time spent by all reduces in occupied slots (ms)=28124
                Total time spent by all map tasks (ms)=985153
                Total time spent by all reduce tasks (ms)=14062
                Total vcore-milliseconds taken by all map tasks=985153
                Total vcore-milliseconds taken by all reduce tasks=14062
                Total megabyte-milliseconds taken by all map tasks=1008796672
                Total megabyte-milliseconds taken by all reduce tasks=28798976
        Map-Reduce Framework
                Map input records=8898041
                Map output records=8898041
                Map output bytes=264608994
                Map output materialized bytes=282405502
                Input split bytes=8236
                Combine input records=0
                Combine output records=0
                Reduce input groups=603668
                Reduce shuffle bytes=282405502
                Reduce input records=8898041
                Reduce output records=1
                Spilled Records=17796082
```

```
                Shuffled Maps =71
                Failed Shuffles=0
                Merged Map outputs=71
                GC time elapsed (ms)=264087
                CPU time spent (ms)=1577100
                Physical memory (bytes) snapshot=44623368192
                Virtual memory (bytes) snapshot=201859448832
                Total committed heap usage (bytes)=61418242048
        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=9458304167
        File Output Format Counters
                Bytes Written=20
```

```java
public static class Map extends Mapper<LongWritable, Text, Text, Text> {
private Text word = new Text();
private Text reviewerText = new Text();
private Text asinText = new Text();

@Override
public void map(LongWritable key, Text value, Context context) throws IOException, InterruptedException {

        for(String token: value.toString().split("\n")){
          word.set(token.replaceAll(" ", ""));
        }

        String[] strArr = word.toString().split(",");
        String reviewer = strArr[0];
        String[] reviewerArr = reviewer.split(":");
        reviewer = reviewerArr[1];
        reviewerText.set(reviewer);

        String asin = strArr[1];
        String[] asinArr = asin.split(":");
        asin = asinArr[1];
        asinText.set(asin);

        context.write(reviewerText, asinText);
  }
}
```

```java
public static class Reduce extends Reducer<Text, Text, Text, IntWritable> {
private int maxCount = 0;
private Text maxCountReviewerID = new Text();

@Override
public void reduce(Text key, Iterable<Text> values, Context context)
throws IOException, InterruptedException {
         int count = 0;

         for(Text asin: values) {
           count++;
         }

         if(maxCount < count) {
           maxCount = count;
           maxCountReviewerID.set(key);
         }

        }

        @Override
        protected void cleanup(Context context) throws IOException, InterruptedException {
          context.write(maxCountReviewerID, new IntWritable(maxCount));
        }
      }
}
```

먼저 맵에서 asin의 이름과 reviewer의 아이디 값을 찾아준다. 그 이후 그 값을 텍스트에 넣어준다음 reduce로 넘겨준다.

reduce에서는 asin의 값이 얼마나 들어왔는지 count해준다. 그 이후 다른 아이디를 가진 reviewer와 비교해서 가장 큰 값을 찾아내서 저장한다.

cleanup을 이용해서 가장 큰 값을 가진 reviewer의 아이디와 그 횟수 만을 출력해준다.