191870294-朱云佳-作业6

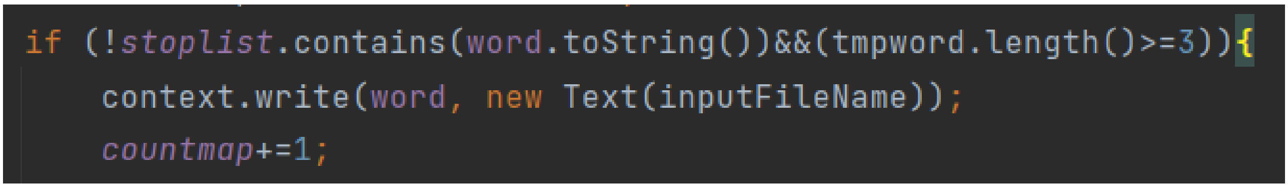
环境说明：**windows**使用**intellij**创建**maven**项目，环境配置与作业**5**相同，调试完毕后进入**WSL+hdfs**分布式环境运行

解题思路

标点、停词、大小写、数字和单词长度的处理沿用了作业5wordcount，此处不再赘述

Map

map输出的键值对是单词和所在文件名

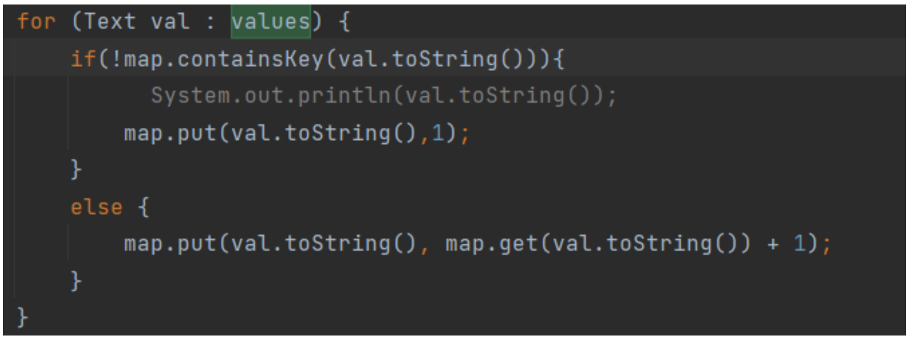


```
if (!stoplist.contains(word.toString())&&(tmpword.length()>=3)){
    context.write(word, new Text(inputFileName));
    countmap+=1;
```

Reduce

reduce函数里定义hashmap类型的变量，用于存放输入reduce的特定单词在各文件中的出现次数。hashmap的键为文件名，值为出现次数。

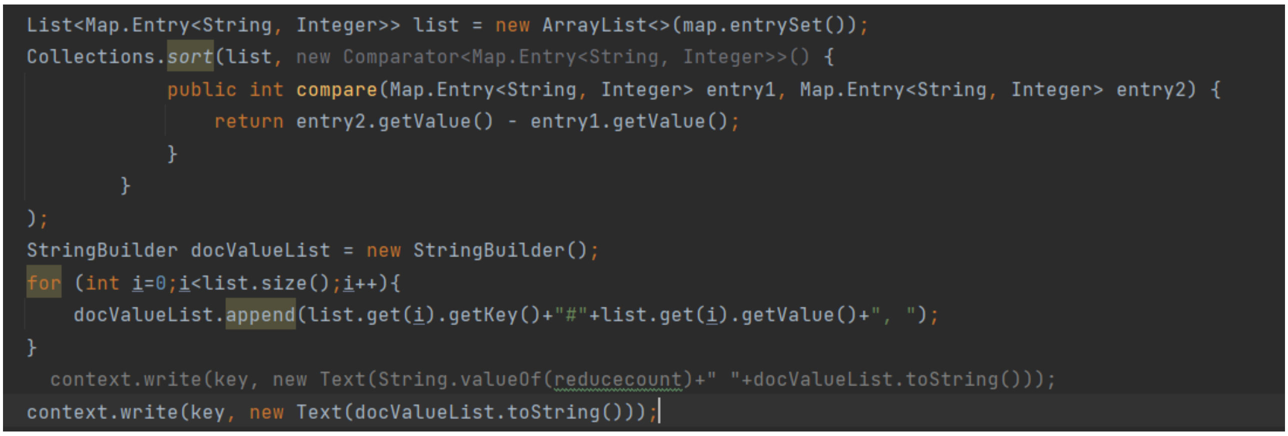
统计次数



```
for (Text val : values) {
    if(!map.containsKey(val.toString())){
          System.out.println(val.toString());
        map.put(val.toString(),1);
    }
    else {
        map.put(val.toString(), map.get(val.toString()) + 1);
    }
}
```

若键不存在于map中，则创建并初始化值为1；否则更新已有的值

按词频排序



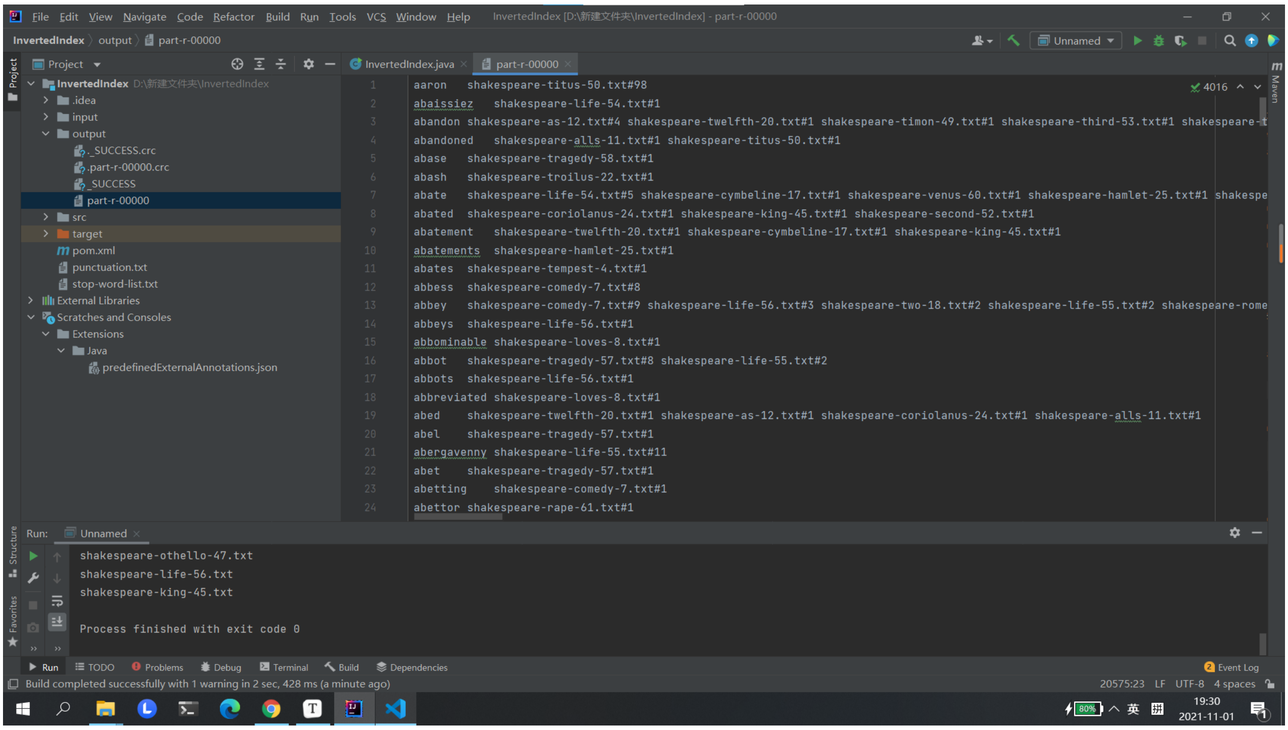
```
List<Map.Entry<String, Integer>> list = new ArrayList<>(map.entrySet());
Collections.sort(list, new Comparator<Map.Entry<String, Integer>>() {
            public int compare(Map.Entry<String, Integer> entry1, Map.Entry<String, Integer> entry2) {
                return entry2.getValue() - entry1.getValue();
            }
        }
);
StringBuilder docValueList = new StringBuilder();
for (int i=0;i<list.size();i++){
    docValueList.append(list.get(i).getKey()+"#"+list.get(i).getValue()+", ");
}
  context.write(key, new Text(String.valueOf(reducecount)+" "+docValueList.toString()));
context.write(key, new Text(docValueList.toString()));
```

map转entrySet之后，定义Collectionn.sort降序排列，最后用StringBuilder统一结果写入

实验结果

完整结果参见result文件夹part-r-00000，这里显示本地intellij和WSL分布式运行的结果

本地intellij


File Edit View Navigate Code Refactor Build Run Tools VCS Window Help
InvertedIndex [D:\新建文件夹\InvertedIndex] - part-r-00000
InvertedIndex 〉 output 〉 part-r-00000
Unnamed
Project
InvertedIndex D:\新建文件夹\InvertedIndex
.idea
input
output
._SUCCESS.crc
.part-r-00000.crc
_SUCCESS
part-r-00000
src
target
pom.xml
punctuation.txt
stop-word-list.txt
External Libraries
Scratches and Consoles
Extensions
Java
predefinedExternalAnnotations.json
InvertedIndex.java
part-r-00000
4016
Run: Unnamed
Run TODO Problems Debug Terminal Build Dependencies
Event Log
20575:23 LF UTF-8 4 spaces
80%
英
19:30
2021-11-01

```
aaron   shakespeare-titus-50.txt#98
abaissiez   shakespeare-life-54.txt#1
abandon shakespeare-as-12.txt#4 shakespeare-twelfth-20.txt#1 shakespeare-timon-49.txt#1 shakespeare-third-53.txt#1 shakespeare-t
abandoned   shakespeare-alls-11.txt#1 shakespeare-titus-50.txt#1
abase   shakespeare-tragedy-58.txt#1
abash   shakespeare-troilus-22.txt#1
abate   shakespeare-life-54.txt#5 shakespeare-cymbeline-17.txt#1 shakespeare-venus-60.txt#1 shakespeare-hamlet-25.txt#1 shakespe
abated  shakespeare-coriolanus-24.txt#1 shakespeare-king-45.txt#1 shakespeare-second-52.txt#1
abatement   shakespeare-twelfth-20.txt#1 shakespeare-cymbeline-17.txt#1 shakespeare-king-45.txt#1
abatements  shakespeare-hamlet-25.txt#1
abates  shakespeare-tempest-4.txt#1
abbess  shakespeare-comedy-7.txt#8
abbey   shakespeare-comedy-7.txt#9 shakespeare-life-56.txt#3 shakespeare-two-18.txt#2 shakespeare-life-55.txt#2 shakespeare-rome
abbeys  shakespeare-life-56.txt#1
abbominable shakespeare-loves-8.txt#1
abbot   shakespeare-tragedy-57.txt#8 shakespeare-life-55.txt#2
abbots  shakespeare-life-56.txt#1
abbreviated shakespeare-loves-8.txt#1
abed    shakespeare-twelfth-20.txt#1 shakespeare-as-12.txt#1 shakespeare-coriolanus-24.txt#1 shakespeare-alls-11.txt#1
abel    shakespeare-tragedy-57.txt#1
abergavenny shakespeare-life-55.txt#11
abet    shakespeare-tragedy-57.txt#1
abetting    shakespeare-comedy-7.txt#1
abettor shakespeare-rape-61.txt#1
```

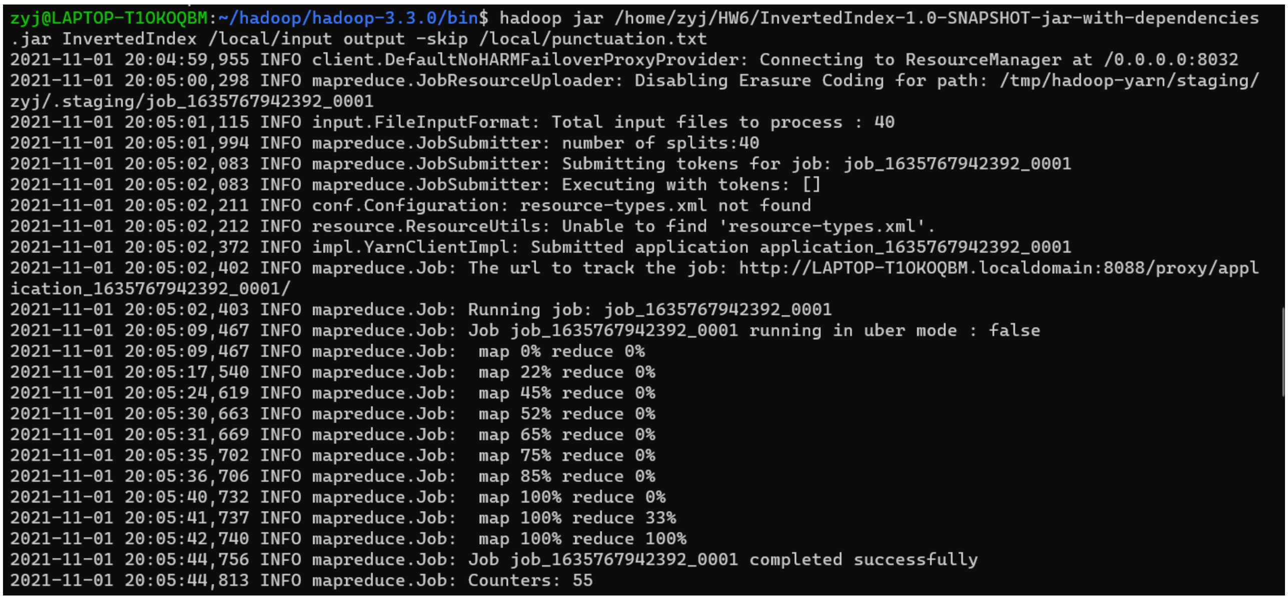
```
shakespeare-othello-47.txt
shakespeare-life-56.txt
shakespeare-king-45.txt

Process finished with exit code 0
```

Build completed successfully with 1 warning in 2 sec, 428 ms (a minute ago)

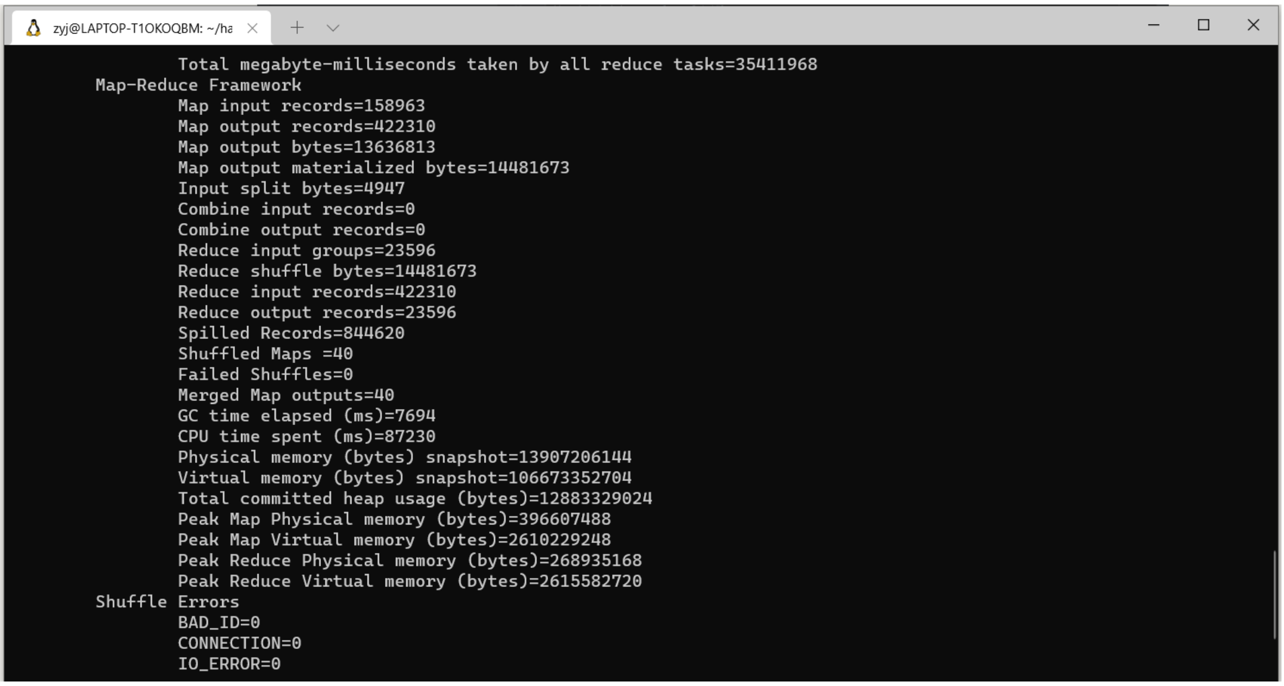
WSL分布式

用`hadoop jar`命令执行



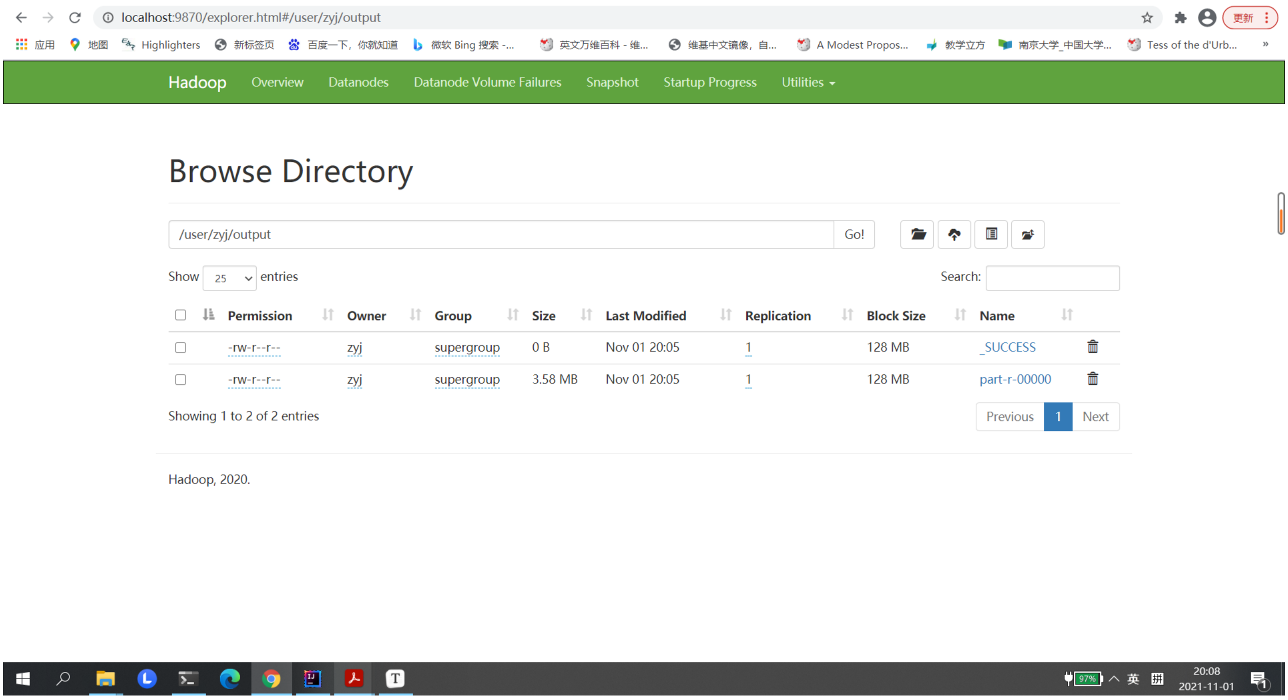
```
zyj@LAPTOP-T1OKOQBM:~/hadoop/hadoop-3.3.0/bin$ hadoop jar /home/zyj/HW6/InvertedIndex-1.0-SNAPSHOT-jar-with-dependencies
.jar InvertedIndex /local/input output -skip /local/punctuation.txt
2021-11-01 20:04:59,955 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2021-11-01 20:05:00,298 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/
zyj/.staging/job_1635767942392_0001
2021-11-01 20:05:01,115 INFO input.FileInputFormat: Total input files to process : 40
2021-11-01 20:05:01,994 INFO mapreduce.JobSubmitter: number of splits:40
2021-11-01 20:05:02,083 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1635767942392_0001
2021-11-01 20:05:02,083 INFO mapreduce.JobSubmitter: Executing with tokens: []
2021-11-01 20:05:02,211 INFO conf.Configuration: resource-types.xml not found
2021-11-01 20:05:02,212 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2021-11-01 20:05:02,372 INFO impl.YarnClientImpl: Submitted application application_1635767942392_0001
2021-11-01 20:05:02,402 INFO mapreduce.Job: The url to track the job: http://LAPTOP-T1OKOQBM.localdomain:8088/proxy/appl
ication_1635767942392_0001/
2021-11-01 20:05:02,403 INFO mapreduce.Job: Running job: job_1635767942392_0001
2021-11-01 20:05:09,467 INFO mapreduce.Job: Job job_1635767942392_0001 running in uber mode : false
2021-11-01 20:05:09,467 INFO mapreduce.Job:  map 0% reduce 0%
2021-11-01 20:05:17,540 INFO mapreduce.Job:  map 22% reduce 0%
2021-11-01 20:05:24,619 INFO mapreduce.Job:  map 45% reduce 0%
2021-11-01 20:05:30,663 INFO mapreduce.Job:  map 52% reduce 0%
2021-11-01 20:05:31,669 INFO mapreduce.Job:  map 65% reduce 0%
2021-11-01 20:05:35,702 INFO mapreduce.Job:  map 75% reduce 0%
2021-11-01 20:05:36,706 INFO mapreduce.Job:  map 85% reduce 0%
2021-11-01 20:05:40,732 INFO mapreduce.Job:  map 100% reduce 0%
2021-11-01 20:05:41,737 INFO mapreduce.Job:  map 100% reduce 33%
2021-11-01 20:05:42,740 INFO mapreduce.Job:  map 100% reduce 100%
2021-11-01 20:05:44,756 INFO mapreduce.Job: Job job_1635767942392_0001 completed successfully
2021-11-01 20:05:44,813 INFO mapreduce.Job: Counters: 55
```

显示运行成功



```
                Total megabyte-milliseconds taken by all reduce tasks=35411968
        Map-Reduce Framework
                Map input records=158963
                Map output records=422310
                Map output bytes=13636813
                Map output materialized bytes=14481673
                Input split bytes=4947
                Combine input records=0
                Combine output records=0
                Reduce input groups=23596
                Reduce shuffle bytes=14481673
                Reduce input records=422310
                Reduce output records=23596
                Spilled Records=844620
                Shuffled Maps =40
                Failed Shuffles=0
                Merged Map outputs=40
                GC time elapsed (ms)=7694
                CPU time spent (ms)=87230
                Physical memory (bytes) snapshot=13907206144
                Virtual memory (bytes) snapshot=106673352704
                Total committed heap usage (bytes)=12883329024
                Peak Map Physical memory (bytes)=396607488
                Peak Map Virtual memory (bytes)=2610229248
                Peak Reduce Physical memory (bytes)=268935168
                Peak Reduce Virtual memory (bytes)=2615582720
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
```

发现hdfs系统上已生成结果文件



Browse Directory

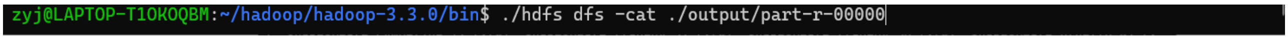
/user/zyj/output

□	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
□	-rw-r--r--	zyj	supergroup	0 B	Nov 01 20:05	1	128 MB	_SUCCESS
□	-rw-r--r--	zyj	supergroup	3.58 MB	Nov 01 20:05	1	128 MB	part-r-00000

Showing 1 to 2 of 2 entries

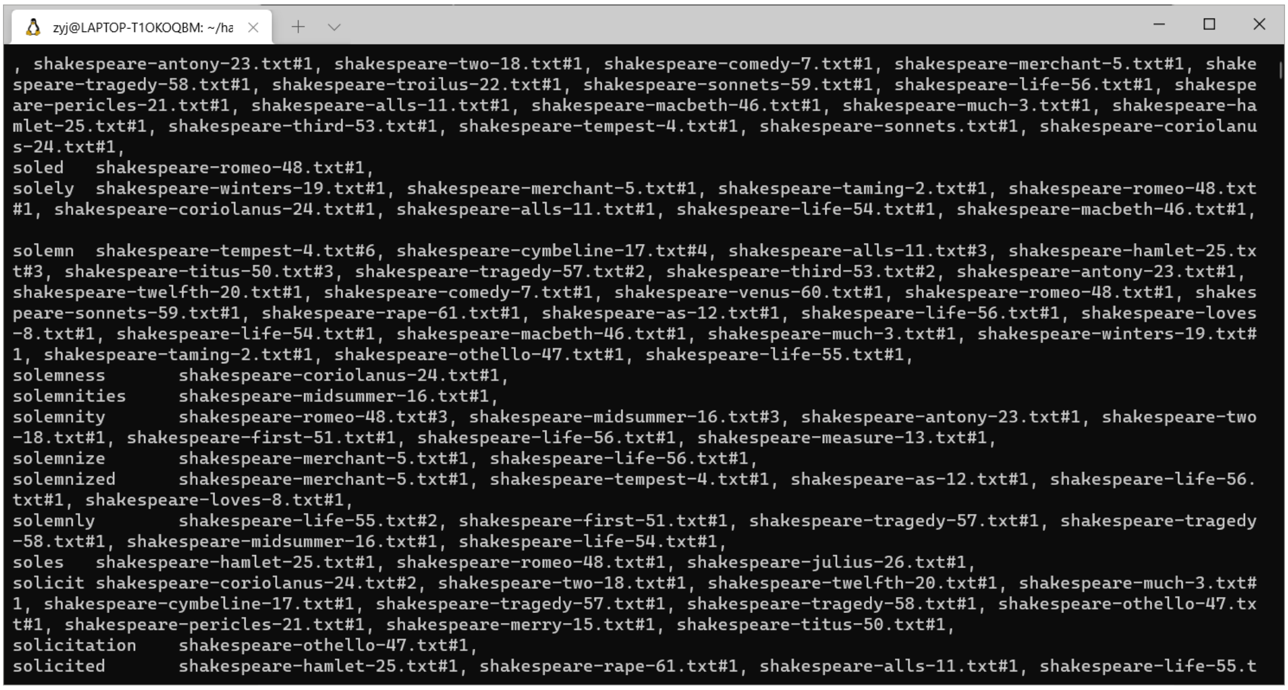
Hadoop, 2020.

`cat`查看生成结果:



```
zyj@LAPTOP-T1OKOQBM:~/hadoop/hadoop-3.3.0/bin$ ./hdfs dfs -cat ./output/part-r-00000
```

(虽然由于行宽限制输出有点混乱但是)结果与本地一致!

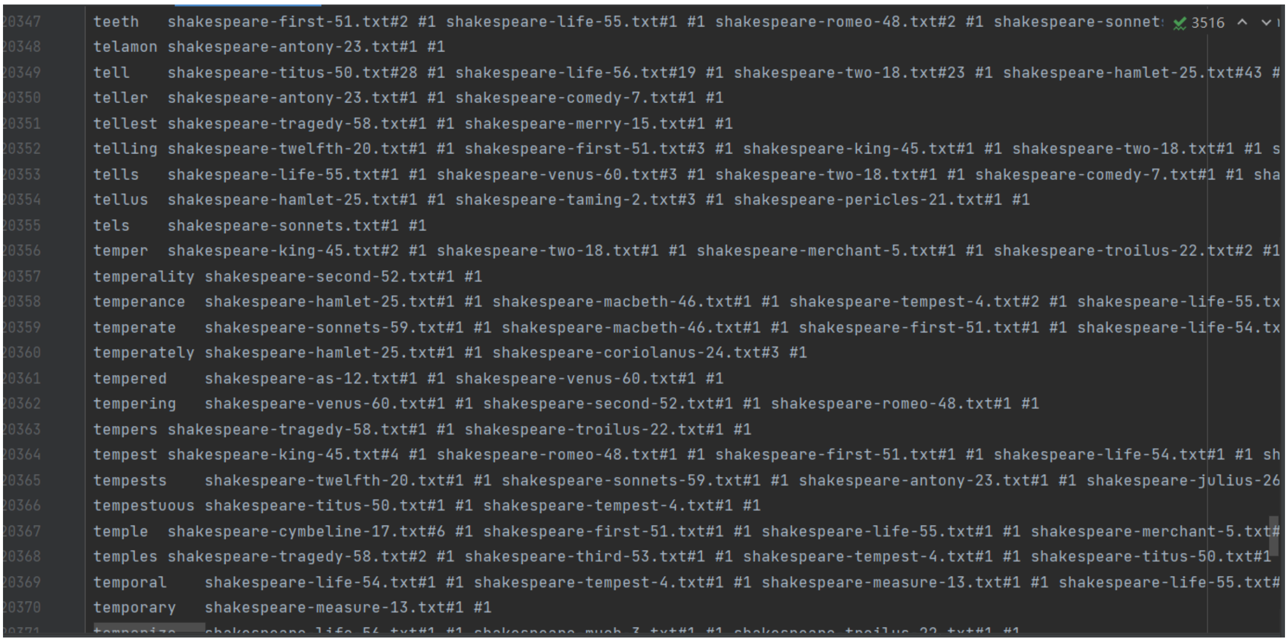


```
, shakespeare-antony-23.txt#1, shakespeare-two-18.txt#1, shakespeare-comedy-7.txt#1, shakespeare-merchant-5.txt#1, shake
speare-tragedy-58.txt#1, shakespeare-troilus-22.txt#1, shakespeare-sonnets-59.txt#1, shakespeare-life-56.txt#1, shakespe
are-pericles-21.txt#1, shakespeare-alls-11.txt#1, shakespeare-macbeth-46.txt#1, shakespeare-much-3.txt#1, shakespeare-ha
mlet-25.txt#1, shakespeare-third-53.txt#1, shakespeare-tempest-4.txt#1, shakespeare-sonnets.txt#1, shakespeare-coriolanu
s-24.txt#1,
soled   shakespeare-romeo-48.txt#1,
solely  shakespeare-winters-19.txt#1, shakespeare-merchant-5.txt#1, shakespeare-taming-2.txt#1, shakespeare-romeo-48.txt
#1, shakespeare-coriolanus-24.txt#1, shakespeare-alls-11.txt#1, shakespeare-life-54.txt#1, shakespeare-macbeth-46.txt#1,

solemn  shakespeare-tempest-4.txt#6, shakespeare-cymbeline-17.txt#4, shakespeare-alls-11.txt#3, shakespeare-hamlet-25.tx
t#3, shakespeare-titus-50.txt#3, shakespeare-tragedy-57.txt#2, shakespeare-third-53.txt#2, shakespeare-antony-23.txt#1,
shakespeare-twelfth-20.txt#1, shakespeare-comedy-7.txt#1, shakespeare-venus-60.txt#1, shakespeare-romeo-48.txt#1, shakes
peare-sonnets-59.txt#1, shakespeare-rape-61.txt#1, shakespeare-as-12.txt#1, shakespeare-life-56.txt#1, shakespeare-loves
-8.txt#1, shakespeare-life-54.txt#1, shakespeare-macbeth-46.txt#1, shakespeare-much-3.txt#1, shakespeare-winters-19.txt#
1, shakespeare-taming-2.txt#1, shakespeare-othello-47.txt#1, shakespeare-life-55.txt#1,
solemness       shakespeare-coriolanus-24.txt#1,
solemnities     shakespeare-midsummer-16.txt#1,
solemnity       shakespeare-romeo-48.txt#3, shakespeare-midsummer-16.txt#3, shakespeare-antony-23.txt#1, shakespeare-two
-18.txt#1, shakespeare-first-51.txt#1, shakespeare-life-56.txt#1, shakespeare-measure-13.txt#1,
solemnize       shakespeare-merchant-5.txt#1, shakespeare-life-56.txt#1,
solemnized      shakespeare-merchant-5.txt#1, shakespeare-tempest-4.txt#1, shakespeare-as-12.txt#1, shakespeare-life-56.
txt#1, shakespeare-loves-8.txt#1,
solemnly        shakespeare-life-55.txt#2, shakespeare-first-51.txt#1, shakespeare-tragedy-57.txt#1, shakespeare-tragedy
-58.txt#1, shakespeare-midsummer-16.txt#1, shakespeare-life-54.txt#1,
soles   shakespeare-hamlet-25.txt#1, shakespeare-romeo-48.txt#1, shakespeare-julius-26.txt#1,
solicit shakespeare-coriolanus-24.txt#2, shakespeare-two-18.txt#1, shakespeare-twelfth-20.txt#1, shakespeare-much-3.txt#
1, shakespeare-cymbeline-17.txt#1, shakespeare-tragedy-57.txt#1, shakespeare-tragedy-58.txt#1, shakespeare-othello-47.tx
t#1, shakespeare-pericles-21.txt#1, shakespeare-merry-15.txt#1, shakespeare-titus-50.txt#1,
solicitation    shakespeare-othello-47.txt#1,
solicited       shakespeare-hamlet-25.txt#1, shakespeare-rape-61.txt#1, shakespeare-alls-11.txt#1, shakespeare-life-55.t
```

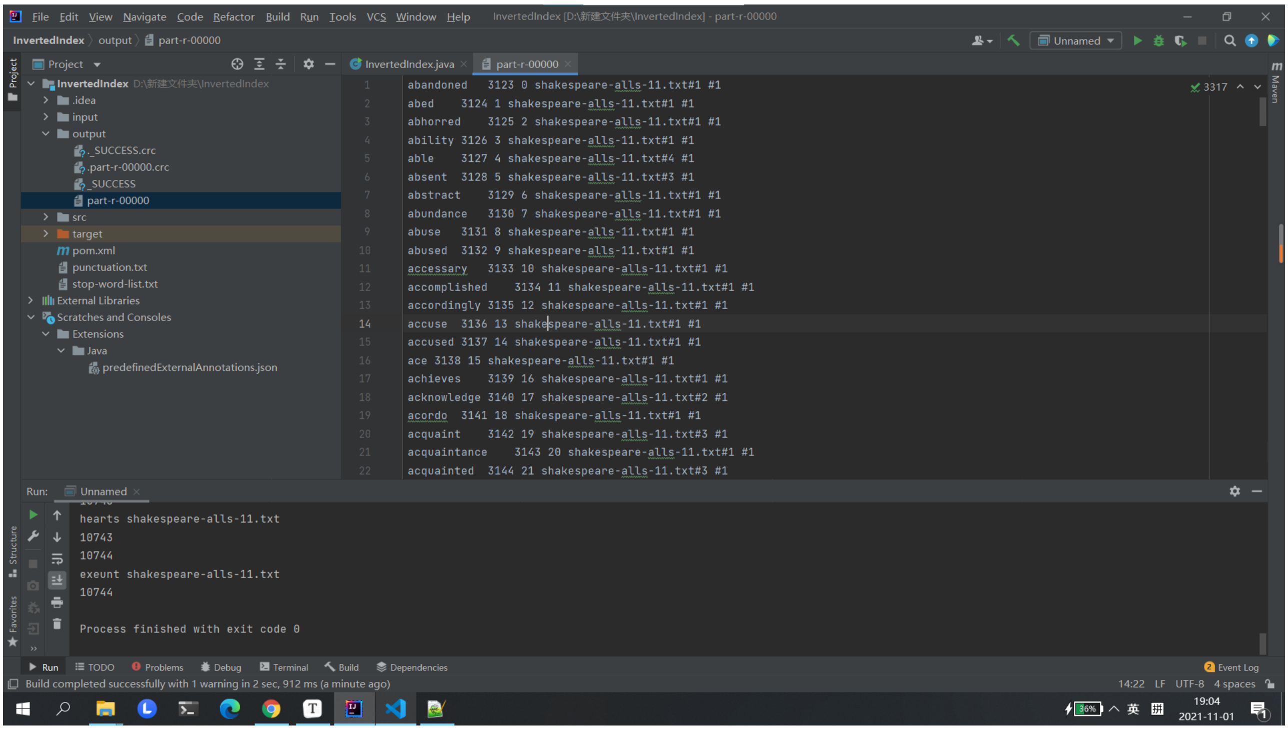
问题及解决方法

很快就完成了coding的框架，只是发现输出的时候：一是多余写入了"#1"，二是排序功能无法正常显示（后来推测与前者有关）



```
teeth   shakespeare-first-51.txt#2 #1 shakespeare-life-55.txt#1 #1 shakespeare-romeo-48.txt#2 #1 shakespeare-sonnet
telamon shakespeare-antony-23.txt#1 #1
tell    shakespeare-titus-50.txt#28 #1 shakespeare-life-56.txt#19 #1 shakespeare-two-18.txt#23 #1 shakespeare-hamlet-25.txt#43 #
teller  shakespeare-antony-23.txt#1 #1 shakespeare-comedy-7.txt#1 #1
tellest shakespeare-tragedy-58.txt#1 #1 shakespeare-merry-15.txt#1 #1
telling shakespeare-twelfth-20.txt#1 #1 shakespeare-first-51.txt#3 #1 shakespeare-king-45.txt#1 #1 shakespeare-two-18.txt#1 #1 s
tells   shakespeare-life-55.txt#1 #1 shakespeare-venus-60.txt#3 #1 shakespeare-two-18.txt#1 #1 shakespeare-comedy-7.txt#1 #1 sha
tellus  shakespeare-hamlet-25.txt#1 #1 shakespeare-taming-2.txt#3 #1 shakespeare-pericles-21.txt#1 #1
tels    shakespeare-sonnets.txt#1 #1
temper  shakespeare-king-45.txt#2 #1 shakespeare-two-18.txt#1 #1 shakespeare-merchant-5.txt#1 #1 shakespeare-troilus-22.txt#2 #1
temperality shakespeare-second-52.txt#1 #1
temperance  shakespeare-hamlet-25.txt#1 #1 shakespeare-macbeth-46.txt#1 #1 shakespeare-tempest-4.txt#2 #1 shakespeare-life-55.tx
temperate   shakespeare-sonnets-59.txt#1 #1 shakespeare-macbeth-46.txt#1 #1 shakespeare-first-51.txt#1 #1 shakespeare-life-54.tx
temperately shakespeare-hamlet-25.txt#1 #1 shakespeare-coriolanus-24.txt#3 #1
tempered    shakespeare-as-12.txt#1 #1 shakespeare-venus-60.txt#1 #1
tempering   shakespeare-venus-60.txt#1 #1 shakespeare-second-52.txt#1 #1 shakespeare-romeo-48.txt#1 #1
tempers shakespeare-tragedy-58.txt#1 #1 shakespeare-troilus-22.txt#1 #1
tempest shakespeare-king-45.txt#4 #1 shakespeare-romeo-48.txt#1 #1 shakespeare-first-51.txt#1 #1 shakespeare-life-54.txt#1 #1 sh
tempests    shakespeare-twelfth-20.txt#1 #1 shakespeare-sonnets-59.txt#1 #1 shakespeare-antony-23.txt#1 #1 shakespeare-julius-26
tempestuous shakespeare-titus-50.txt#1 #1 shakespeare-tempest-4.txt#1 #1
temple  shakespeare-cymbeline-17.txt#6 #1 shakespeare-first-51.txt#1 #1 shakespeare-life-55.txt#1 #1 shakespeare-merchant-5.txt#
temples shakespeare-tragedy-58.txt#2 #1 shakespeare-third-53.txt#1 #1 shakespeare-tempest-4.txt#1 #1 shakespeare-titus-50.txt#1
temporal    shakespeare-life-54.txt#1 #1 shakespeare-tempest-4.txt#1 #1 shakespeare-measure-13.txt#1 #1 shakespeare-life-55.txt#
temporary   shakespeare-measure-13.txt#1 #1
temporize   shakespeare-life-54.txt#1 #1 shakespeare-much-3.txt#1 #1 shakespeare-troilus-22.txt#1 #1
```

调整了context.write的输出（加入序列号）看起来，似乎是reduce一波之后，又来了一波：



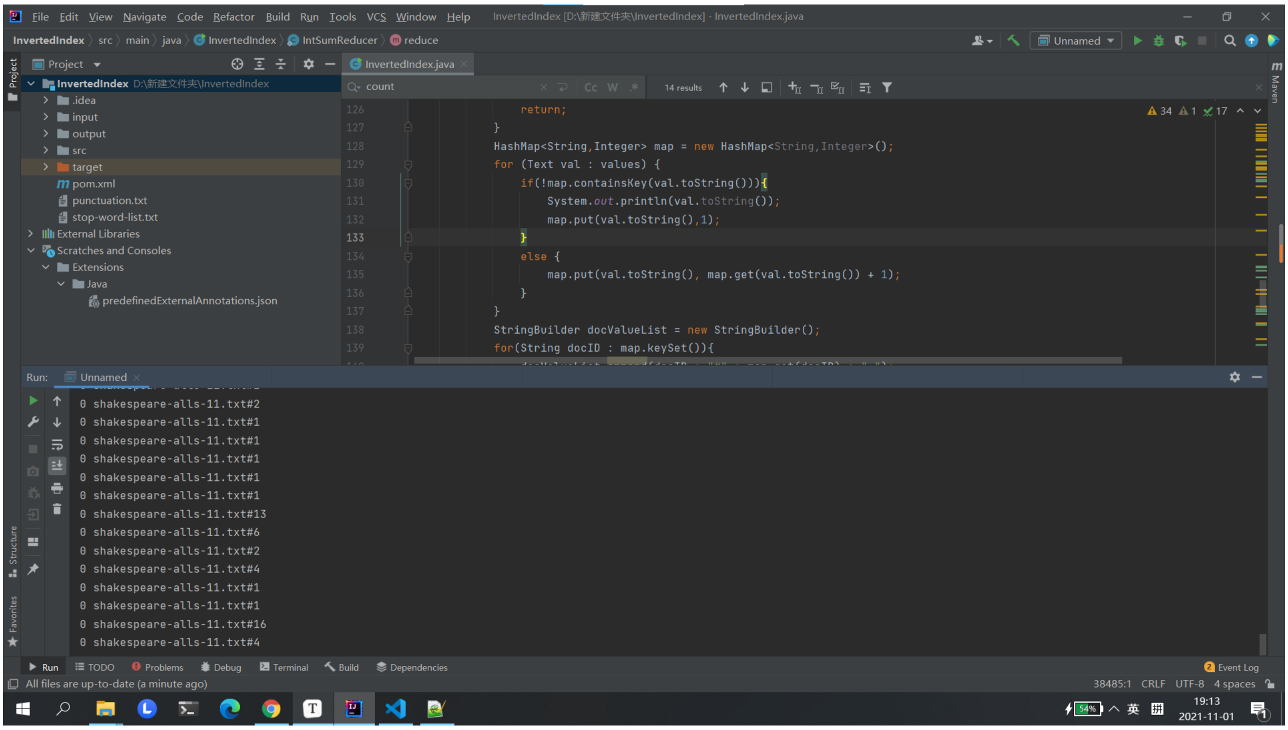
```
abandoned	3123 0 shakespeare-alls-11.txt#1 #1
abed	3124 1 shakespeare-alls-11.txt#1 #1
abhorred	3125 2 shakespeare-alls-11.txt#1 #1
ability 3126 3 shakespeare-alls-11.txt#1 #1
able	3127 4 shakespeare-alls-11.txt#4 #1
absent	3128 5 shakespeare-alls-11.txt#3 #1
abstract	3129 6 shakespeare-alls-11.txt#1 #1
abundance	3130 7 shakespeare-alls-11.txt#1 #1
abuse	3131 8 shakespeare-alls-11.txt#1 #1
abused	3132 9 shakespeare-alls-11.txt#1 #1
accessary	3133 10 shakespeare-alls-11.txt#1 #1
accomplished	3134 11 shakespeare-alls-11.txt#1 #1
accordingly 3135 12 shakespeare-alls-11.txt#1 #1
accuse	3136 13 shakespeare-alls-11.txt#1 #1
accused 3137 14 shakespeare-alls-11.txt#1 #1
ace 3138 15 shakespeare-alls-11.txt#1 #1
achieves	3139 16 shakespeare-alls-11.txt#1 #1
acknowledge 3140 17 shakespeare-alls-11.txt#2 #1
acordo	3141 18 shakespeare-alls-11.txt#1 #1
acquaint	3142 19 shakespeare-alls-11.txt#3 #1
acquaintance	3143 20 shakespeare-alls-11.txt#1 #1
acquainted	3144 21 shakespeare-alls-11.txt#3 #1
```

```
hearts shakespeare-alls-11.txt
10743
10744
exeunt shakespeare-alls-11.txt
10744

Process finished with exit code 0
```

Build completed successfully with 1 warning in 2 sec, 912 ms (a minute ago)

用println调试法--看起来它将之前的结果又统计了一遍（这里指的是shakespeare-alls-11.txt#2这种StringBuilder的内容，以下的print结果验证了猜想）

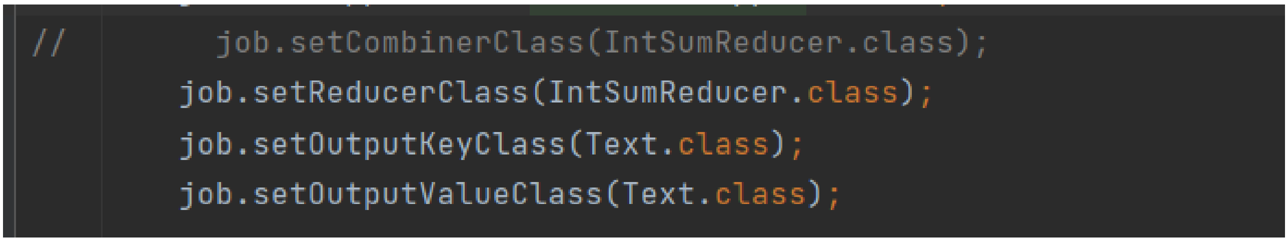


```
            return;
        }
        HashMap<String,Integer> map = new HashMap<String,Integer>();
        for (Text val : values) {
            if(!map.containsKey(val.toString())){
                System.out.println(val.toString());
                map.put(val.toString(),1);
            }
            else {
                map.put(val.toString(), map.get(val.toString()) + 1);
            }
        }
        StringBuilder docValueList = new StringBuilder();
        for(String docID : map.keySet()){
```

```
0 shakespeare-alls-11.txt#2
0 shakespeare-alls-11.txt#1
0 shakespeare-alls-11.txt#1
0 shakespeare-alls-11.txt#1
0 shakespeare-alls-11.txt#1
0 shakespeare-alls-11.txt#1
0 shakespeare-alls-11.txt#13
0 shakespeare-alls-11.txt#6
0 shakespeare-alls-11.txt#2
0 shakespeare-alls-11.txt#4
0 shakespeare-alls-11.txt#1
0 shakespeare-alls-11.txt#1
0 shakespeare-alls-11.txt#16
0 shakespeare-alls-11.txt#4
```

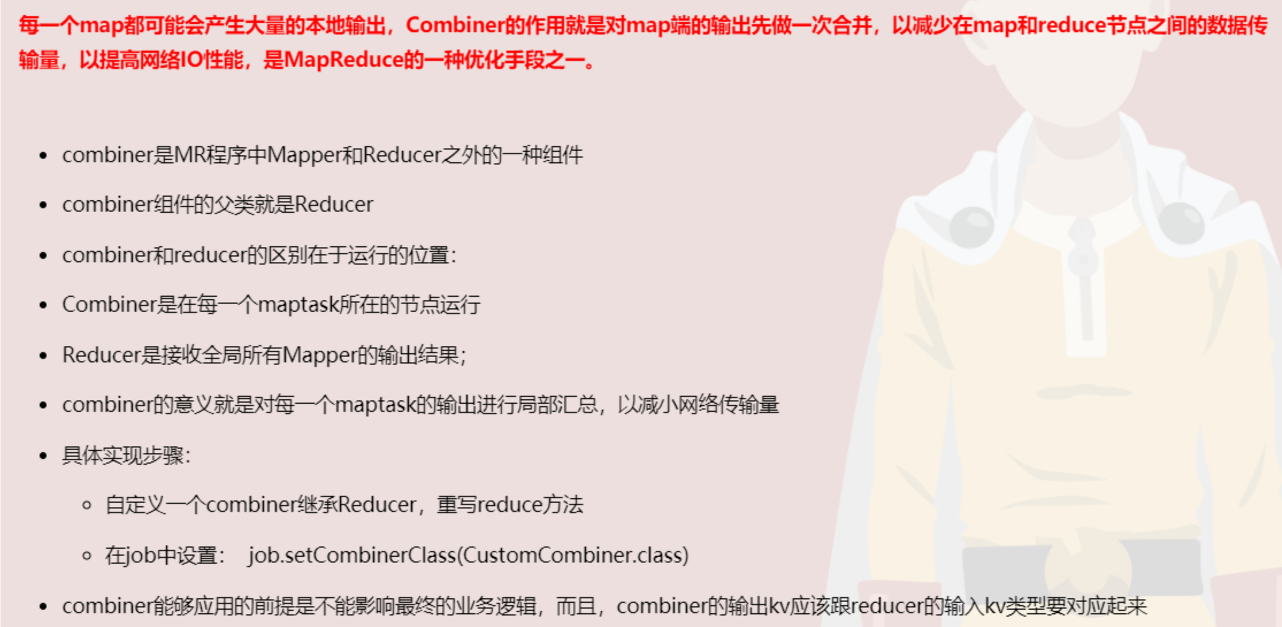
All files are up-to-date (a minute ago)

（其实花了9个小时断断续续调试才发现了上述端倪）...找来找去应该是combiner的锅（由于复用了wordcount的代码，没有去掉combiner，一开始也觉得不用去掉）



```
//        job.setCombinerClass(IntSumReducer.class);
        job.setReducerClass(IntSumReducer.class);
        job.setOutputKeyClass(Text.class);
        job.setOutputValueClass(Text.class);
```

于是复习了combiner的作用（看来在这个情景下使用combiner会导致reduce的时候再统计一遍，因为第一波combiner过后value类型是Text）



每一个map都可能会产生大量的本地输出，Combiner的作用就是对map端的输出先做一次合并，以减少在map和reduce节点之间的数据传输量，以提高网络IO性能，是MapReduce的一种优化手段之一。

- combiner是MR程序中Mapper和Reducer之外的一种组件
- combiner组件的父类就是Reducer
- combiner和reducer的区别在于运行的位置:
- Combiner是在每一个maptask所在的节点运行
- Reducer是接收全局所有Mapper的输出结果;
- combiner的意义就是对每一个maptask的输出进行局部汇总，以减小网络传输量
- 具体实现步骤:
  - 自定义一个combiner继承Reducer，重写reduce方法
  - 在job中设置： job.setCombinerClass(CustomCombiner.class)
- combiner能够应用的前提是不能影响最终的业务逻辑，而且，combiner的输出kv应该跟reducer的输入kv类型要对应起来

看来...要慎重考虑combiner的使用，解决方法就是注释掉combiner