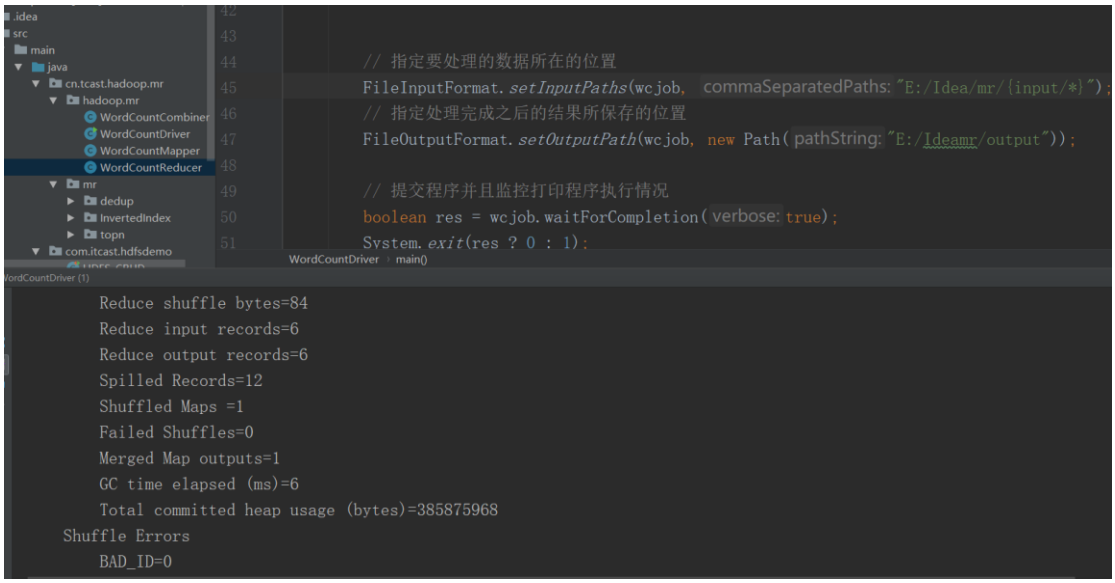
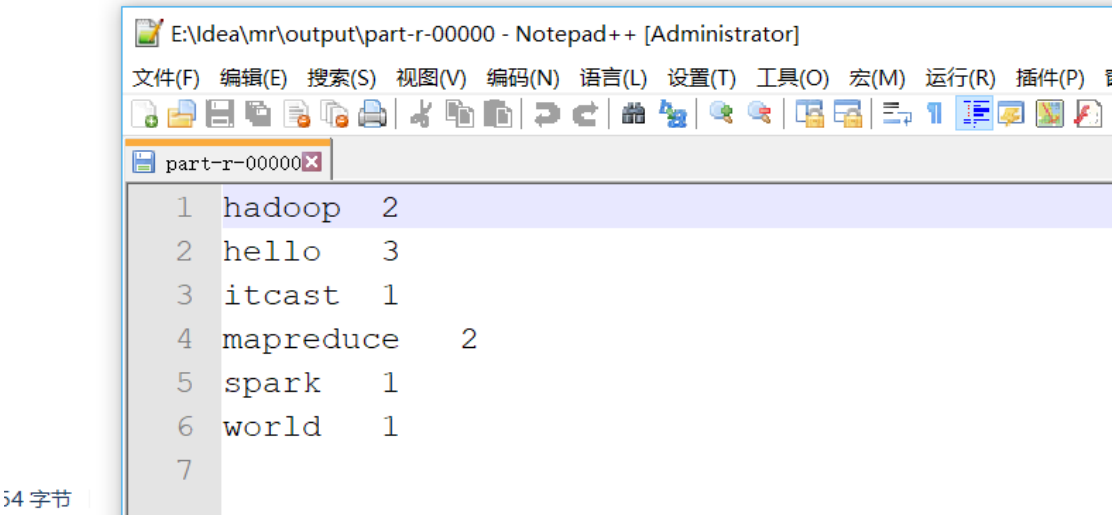
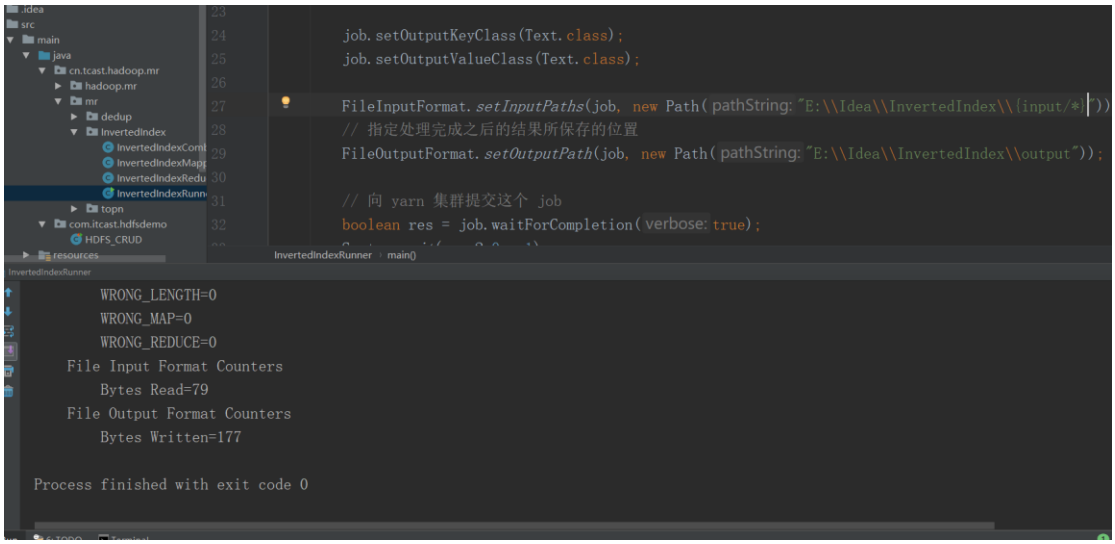
词频统计

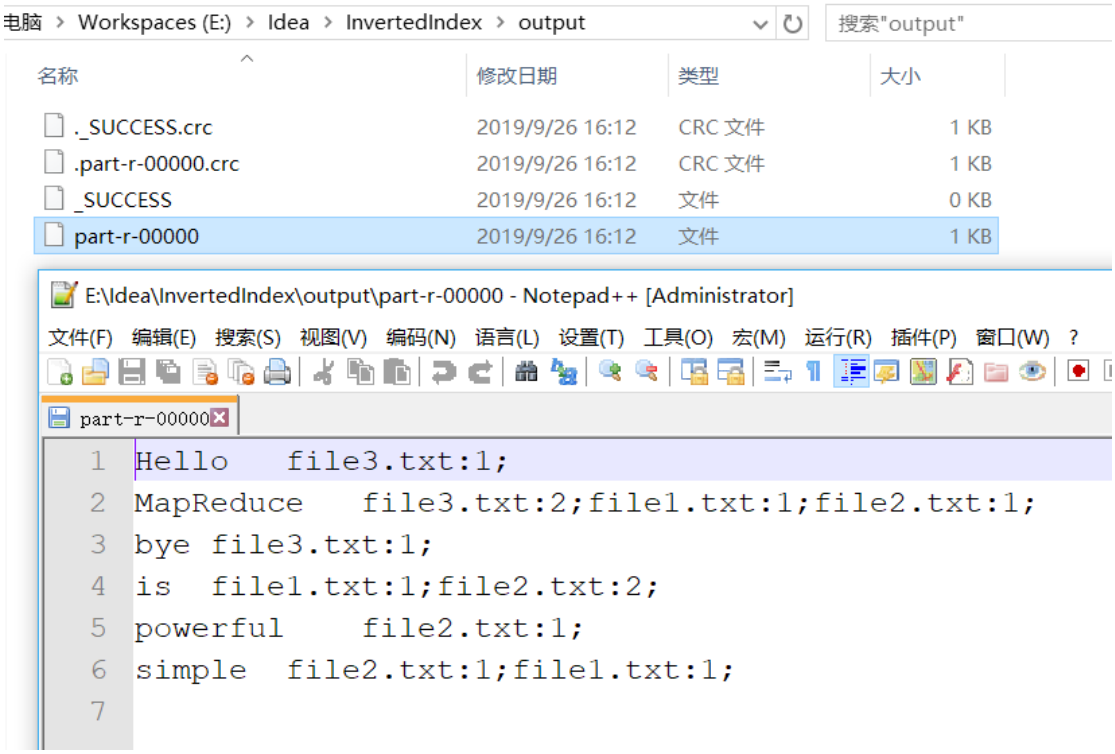

idea
src
main
java
cn.tcast.hadoop.mr
hadoop.mr
WordCountCombiner
WordCountDriver
WordCountMapper
WordCountReducer
mr
dedup
InvertedIndex
topn
com.itcast.hdfsdemo
WordCountDriver (1)
// 指定要处理的数据所在的位置
FileInputFormat.setInputPaths(wcjob, commaSeparatedPaths: "E:/Idea/mr/{input/*}");
// 指定处理完成之后的结果所保存的位置
FileOutputFormat.setOutputPath(wcjob, new Path(pathString: "E:/Ideamr/output"));
// 提交程序并且监控打印程序执行情况
boolean res = wcjob.waitForCompletion(verbose: true);
System.exit(res ? 0 : 1);
WordCountDriver › main()
Reduce shuffle bytes=84
Reduce input records=6
Reduce output records=6
Spilled Records=12
Shuffled Maps =1
Failed Shuffles=0
Merged Map outputs=1
GC time elapsed (ms)=6
Total committed heap usage (bytes)=385875968
Shuffle Errors
BAD_ID=0


电脑 › Workspaces (E:) › Idea › mr › output
搜索"output"
名称
修改日期
类型
大小
._SUCCESS.crc 2019/9/26 15:56 CRC 文件 1 KB
.part-r-00000.crc 2019/9/26 15:56 CRC 文件 1 KB
_SUCCESS 2019/9/26 15:56 文件 0 KB
part-r-00000 2019/9/26 15:56 文件 1 KB

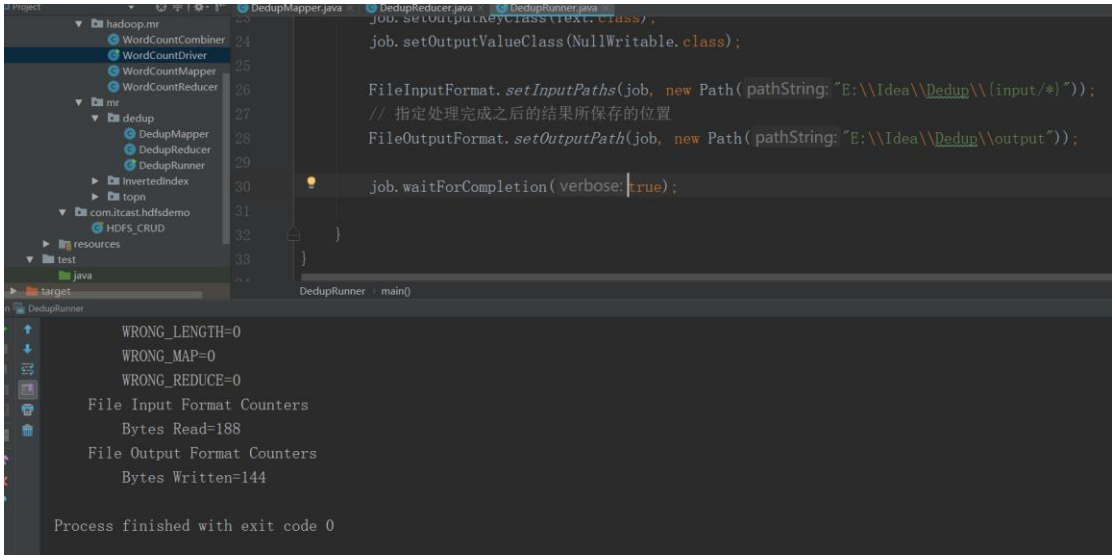

E:\Idea\mr\output\part-r-00000 - Notepad++ [Administrator]
文件(F) 编辑(E) 搜索(S) 视图(V) 编码(N) 语言(L) 设置(T) 工具(O) 宏(M) 运行(R) 插件(P)
part-r-00000
1 hadoop 2
2 hello 3
3 itcast 1
4 mapreduce 2
5 spark 1
6 world 1
7
54 字节

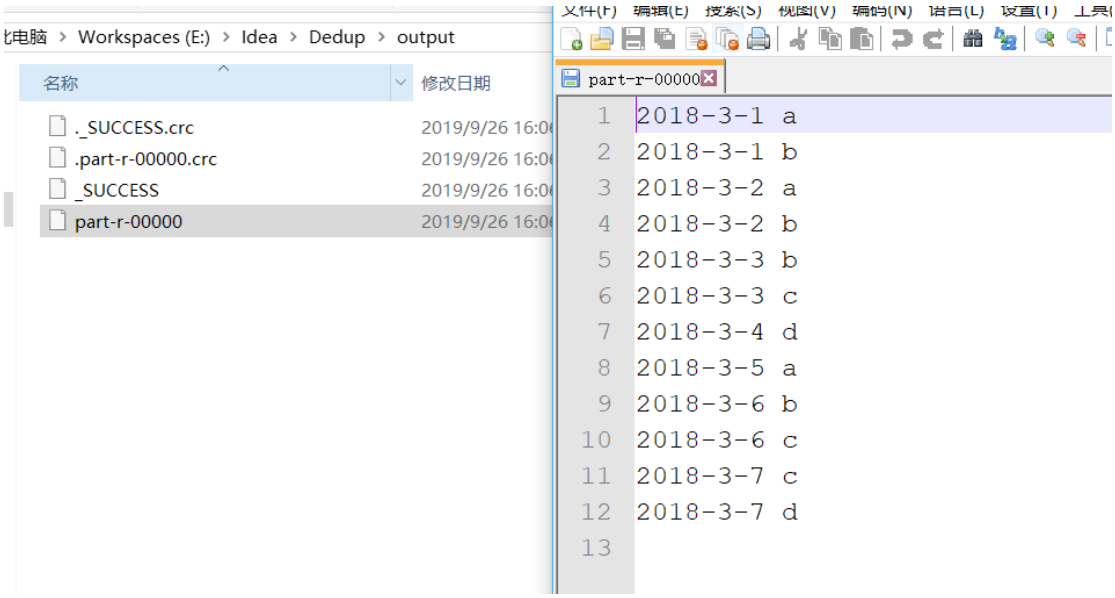
倒排索引


job.setOutputKeyClass(Text.class);
job.setOutputValueClass(Text.class);
FileInputFormat.setInputPaths(job, new Path(pathString: "E:\\Idea\\InvertedIndex\\input/*"));
// 指定处理完成之后的结果所保存的位置
FileOutputFormat.setOutputPath(job, new Path(pathString: "E:\\Idea\\InvertedIndex\\output"));
// 向 yarn 集群提交这个 job
boolean res = job.waitForCompletion(verbose: true);
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_REDUCE=0
File Input Format Counters
Bytes Read=79
File Output Format Counters
Bytes Written=177
Process finished with exit code 0

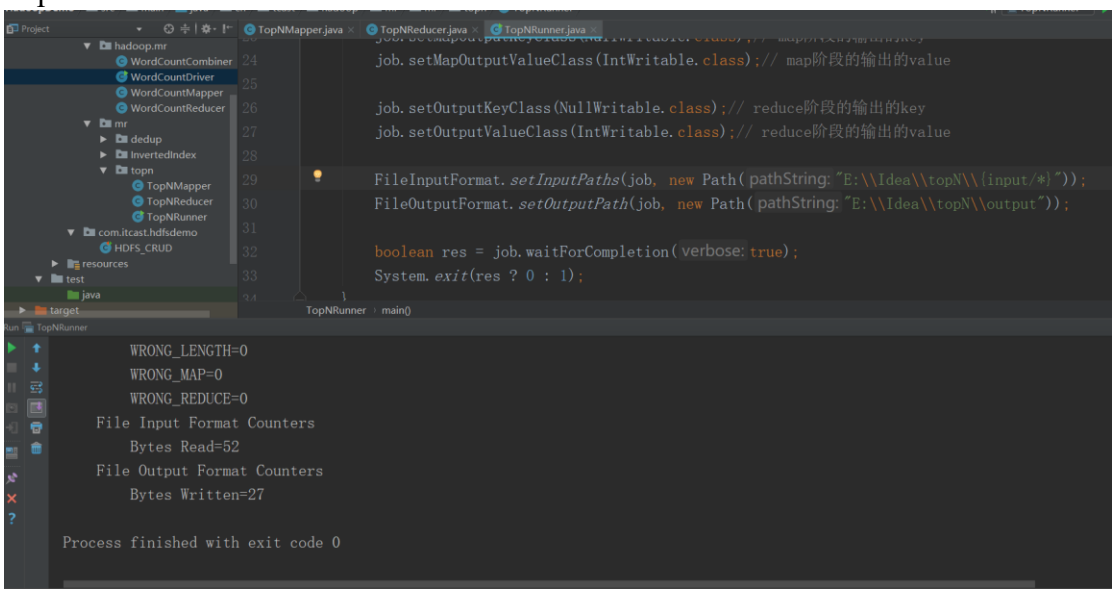

电脑 > Workspaces (E:) > Idea > InvertedIndex > output
搜索"output"
名称
修改日期
类型
大小
._SUCCESS.crc 2019/9/26 16:12 CRC 文件 1 KB
.part-r-00000.crc 2019/9/26 16:12 CRC 文件 1 KB
_SUCCESS 2019/9/26 16:12 文件 0 KB
part-r-00000 2019/9/26 16:12 文件 1 KB
E:\Idea\InvertedIndex\output\part-r-00000 - Notepad++ [Administrator]
文件(F) 编辑(E) 搜索(S) 视图(V) 编码(N) 语言(L) 设置(T) 工具(O) 宏(M) 运行(R) 插件(P) 窗口(W) ?
part-r-00000
1 Hello file3.txt:1;
2 MapReduce file3.txt:2;file1.txt:1;file2.txt:1;
3 bye file3.txt:1;
4 is file1.txt:1;file2.txt:2;
5 powerful file2.txt:1;
6 simple file2.txt:1;file1.txt:1;
7

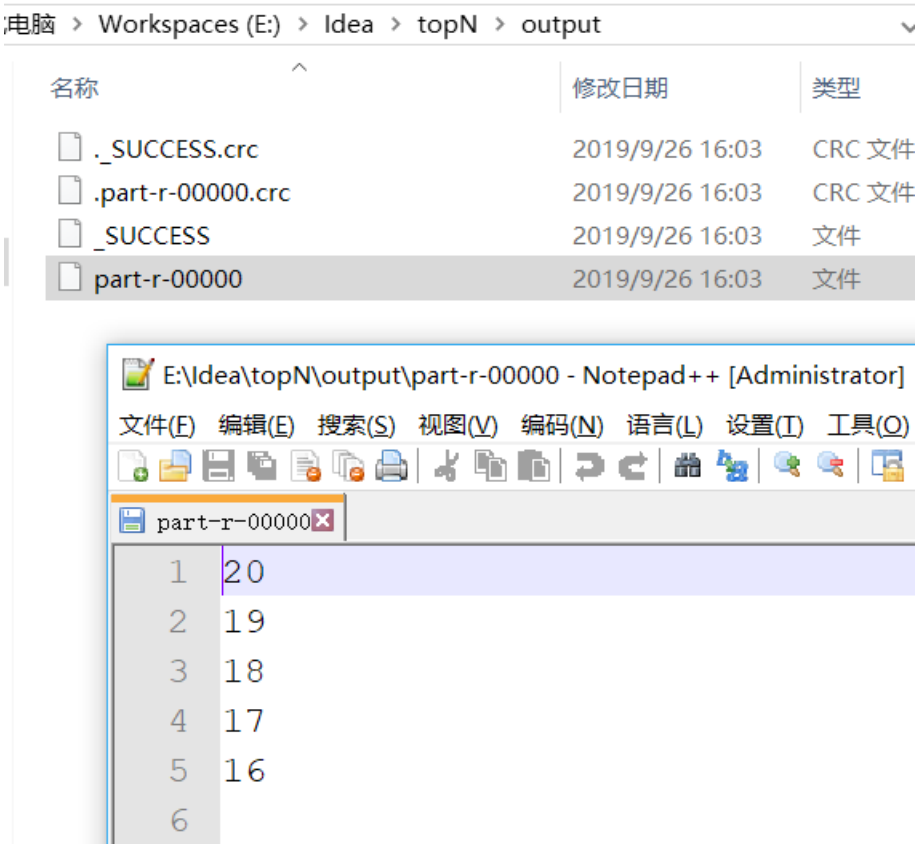
数据去重


job.setOutputValueClass(NullWritable.class);
FileInputFormat.setInputPaths(job, new Path(pathString: "E:\\Idea\\Dedup\\{input/*}"));
// 指定处理完成之后的结果所保存的位置
FileOutputFormat.setOutputPath(job, new Path(pathString: "E:\\Idea\\Dedup\\output"));
job.waitForCompletion(verbose: true);
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_REDUCE=0
File Input Format Counters
Bytes Read=188
File Output Format Counters
Bytes Written=144
Process finished with exit code 0


Workspaces (E:) › Idea › Dedup › output
名称
修改日期
._SUCCESS.crc
.part-r-00000.crc
_SUCCESS
part-r-00000
part-r-00000
2018-3-1 a
2018-3-1 b
2018-3-2 a
2018-3-2 b
2018-3-3 b
2018-3-3 c
2018-3-4 d
2018-3-5 a
2018-3-6 b
2018-3-6 c
2018-3-7 c
2018-3-7 d

TopN


job.setMapOutputValueClass(IntWritable.class);// map阶段的输出的value
job.setOutputKeyClass(NullWritable.class);// reduce阶段的输出的key
job.setOutputValueClass(IntWritable.class);// reduce阶段的输出的value
FileInputFormat.setInputPaths(job, new Path(pathString: "E:\\Idea\\topN\\{input/*}"));
FileOutputFormat.setOutputPath(job, new Path(pathString: "E:\\Idea\\topN\\output"));
boolean res = job.waitForCompletion(verbose: true);
System.exit(res ? 0 : 1);
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_REDUCE=0
File Input Format Counters
Bytes Read=52
File Output Format Counters
Bytes Written=27
Process finished with exit code 0


电脑 › Workspaces (E:) › Idea › topN › output
名称
修改日期
类型
._SUCCESS.crc
2019/9/26 16:03
CRC 文件
.part-r-00000.crc
2019/9/26 16:03
CRC 文件
_SUCCESS
2019/9/26 16:03
文件
part-r-00000
2019/9/26 16:03
文件
E:\Idea\topN\output\part-r-00000 - Notepad++ [Administrator]
文件(F) 编辑(E) 搜索(S) 视图(V) 编码(N) 语言(L) 设置(T) 工具(O)
part-r-00000
1 20
2 19
3 18
4 17
5 16
6