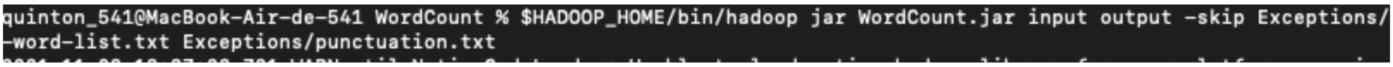
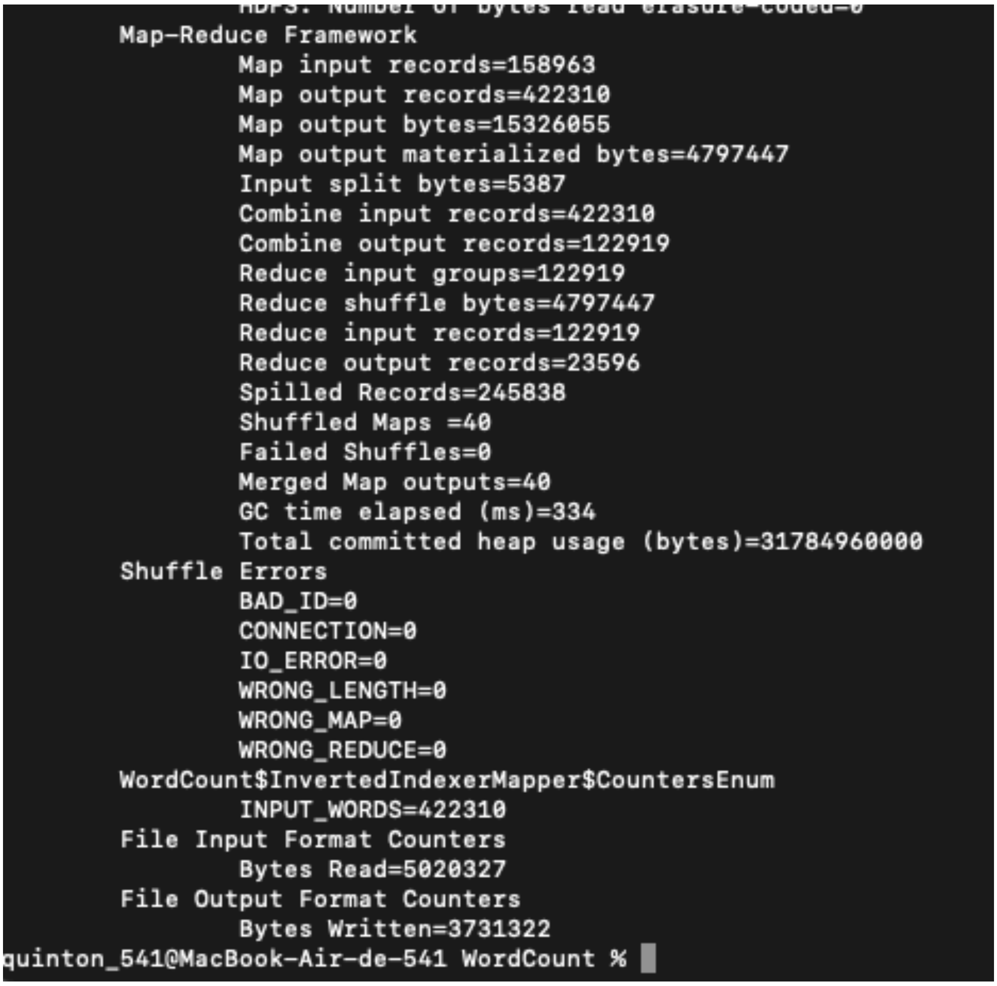
思路

这次只需要执行一个词频统计和一个排序输出的job。和上一次作业分文件词频统计一样，我们在Mapper中采用<word-file,1>的方式输出即可，Reducer分为两部分 先通过Combiner统计出每一个文件中单词的数量，然后分单词统计每个文件中出现的数量,通过List.sort函数将wordlist排序后写入输出文件。

运行截图

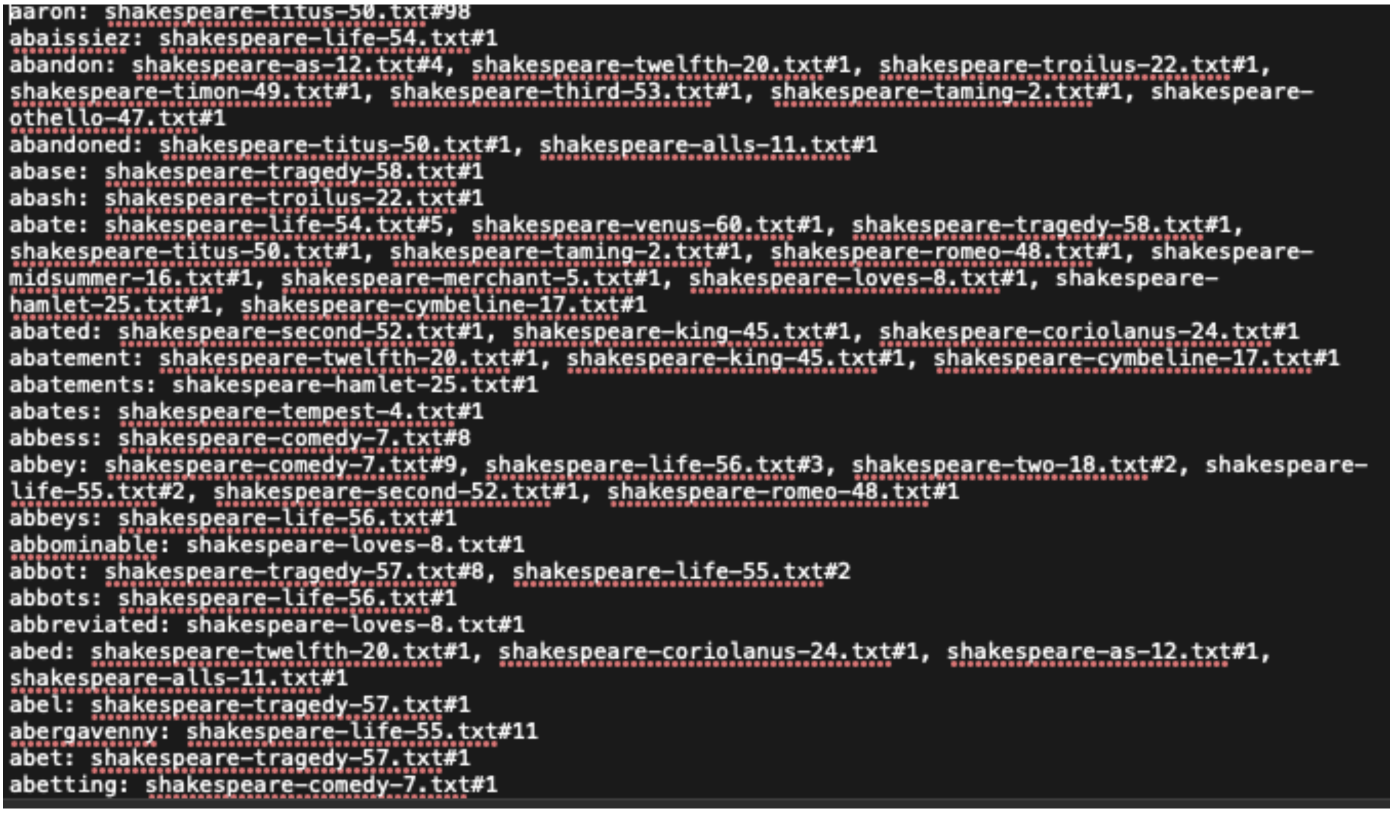


```
quinton_541@MacBook-Air-de-541 WordCount % $HADOOP_HOME/bin/hadoop jar WordCount.jar input output -skip Exceptions/
-word-list.txt Exceptions/punctuation.txt
```



```
                HDFS: Number of bytes read erasure-coded=0
        Map-Reduce Framework
                Map input records=158963
                Map output records=422310
                Map output bytes=15326055
                Map output materialized bytes=4797447
                Input split bytes=5387
                Combine input records=422310
                Combine output records=122919
                Reduce input groups=122919
                Reduce shuffle bytes=4797447
                Reduce input records=122919
                Reduce output records=23596
                Spilled Records=245838
                Shuffled Maps =40
                Failed Shuffles=0
                Merged Map outputs=40
                GC time elapsed (ms)=334
                Total committed heap usage (bytes)=31784960000
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        WordCount$InvertedIndexerMapper$CountersEnum
                INPUT_WORDS=422310
        File Input Format Counters
                Bytes Read=5020327
        File Output Format Counters
                Bytes Written=3731322
quinton_541@MacBook-Air-de-541 WordCount % 
```

输出文件：



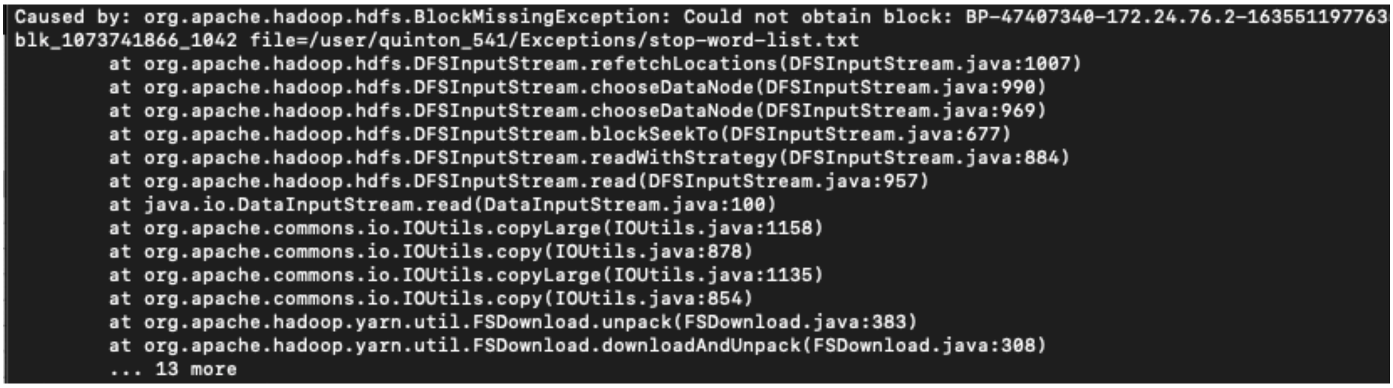
```
aaron: shakespeare-titus-50.txt#98
abaissiez: shakespeare-life-54.txt#1
abandon: shakespeare-as-12.txt#4, shakespeare-twelfth-20.txt#1, shakespeare-troilus-22.txt#1,
shakespeare-timon-49.txt#1, shakespeare-third-53.txt#1, shakespeare-taming-2.txt#1, shakespeare-
othello-47.txt#1
abandoned: shakespeare-titus-50.txt#1, shakespeare-alls-11.txt#1
abase: shakespeare-tragedy-58.txt#1
abash: shakespeare-troilus-22.txt#1
abate: shakespeare-life-54.txt#5, shakespeare-venus-60.txt#1, shakespeare-tragedy-58.txt#1,
shakespeare-titus-50.txt#1, shakespeare-taming-2.txt#1, shakespeare-romeo-48.txt#1, shakespeare-
midsummer-16.txt#1, shakespeare-merchant-5.txt#1, shakespeare-loves-8.txt#1, shakespeare-
hamlet-25.txt#1, shakespeare-cymbeline-17.txt#1
abated: shakespeare-second-52.txt#1, shakespeare-king-45.txt#1, shakespeare-coriolanus-24.txt#1
abatement: shakespeare-twelfth-20.txt#1, shakespeare-king-45.txt#1, shakespeare-cymbeline-17.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-second-52.txt#1, shakespeare-romeo-48.txt#1
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-coriolanus-24.txt#1, shakespeare-as-12.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
```

出现的问题

由于作业5已经做过 所以问题比较少 主要是这个问题折磨了很久

在执行程序时，一直提示 ”Could not contain block:...“

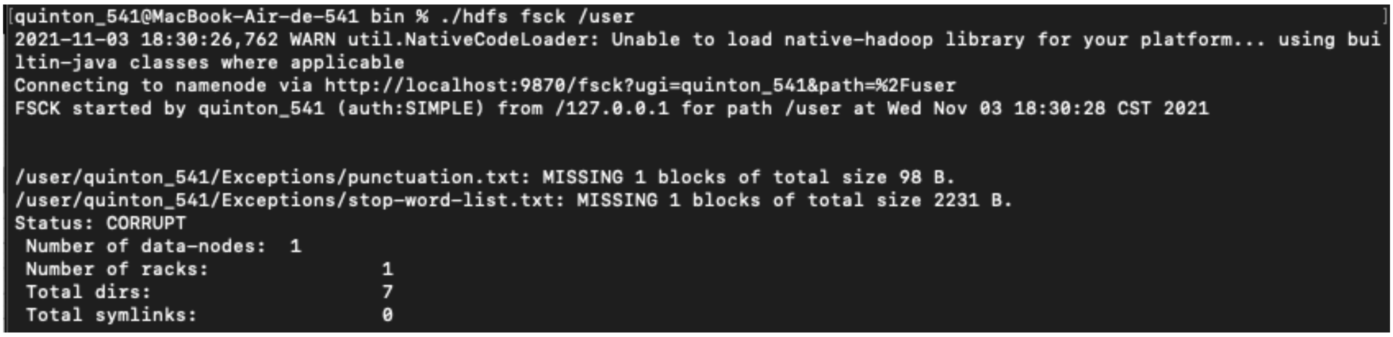
网上的教程非常不靠谱，全是提示我datanode出了问题 于是重启hadoop无数次也没解决



```
Caused by: org.apache.hadoop.hdfs.BlockMissingException: Could not obtain block: BP-47407340-172.24.76.2-163551197763
blk_1073741866_1042 file=/user/quinton_541/Exceptions/stop-word-list.txt
        at org.apache.hadoop.hdfs.DFSInputStream.refetchLocations(DFSInputStream.java:1007)
        at org.apache.hadoop.hdfs.DFSInputStream.chooseDataNode(DFSInputStream.java:990)
        at org.apache.hadoop.hdfs.DFSInputStream.chooseDataNode(DFSInputStream.java:969)
        at org.apache.hadoop.hdfs.DFSInputStream.blockSeekTo(DFSInputStream.java:677)
        at org.apache.hadoop.hdfs.DFSInputStream.readWithStrategy(DFSInputStream.java:884)
        at org.apache.hadoop.hdfs.DFSInputStream.read(DFSInputStream.java:957)
        at java.io.DataInputStream.read(DataInputStream.java:100)
        at org.apache.commons.io.IOUtils.copyLarge(IOUtils.java:1158)
        at org.apache.commons.io.IOUtils.copy(IOUtils.java:878)
        at org.apache.commons.io.IOUtils.copyLarge(IOUtils.java:1135)
        at org.apache.commons.io.IOUtils.copy(IOUtils.java:854)
        at org.apache.hadoop.yarn.util.FSDownload.unpack(FSDownload.java:383)
        at org.apache.hadoop.yarn.util.FSDownload.downloadAndUnpack(FSDownload.java:308)
        ... 13 more
```

然后在外网找到一个./hdfs fsck的指令，可以查看hdfs文件系统里面已经损毁的文件

一看就是两个停词的文件被损坏了



```
quinton_541@MacBook-Air-de-541 bin % ./hdfs fsck /user
2021-11-03 18:30:26,762 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using bui
ltin-java classes where applicable
Connecting to namenode via http://localhost:9870/fsck?ugi=quinton_541&path=%2Fuser
FSCK started by quinton_541 (auth:SIMPLE) from /127.0.0.1 for path /user at Wed Nov 03 18:30:28 CST 2021


/user/quinton_541/Exceptions/punctuation.txt: MISSING 1 blocks of total size 98 B.
/user/quinton_541/Exceptions/stop-word-list.txt: MISSING 1 blocks of total size 2231 B.
Status: CORRUPT
 Number of data-nodes:  1
 Number of racks:               1
 Total dirs:                    7
 Total symlinks:                0
```

重新上传之后 问题解决

（后来运行的时候发现有一个输入文件也损坏了 不知道为什么本机的hdfs文件系统为什么这么不靠谱