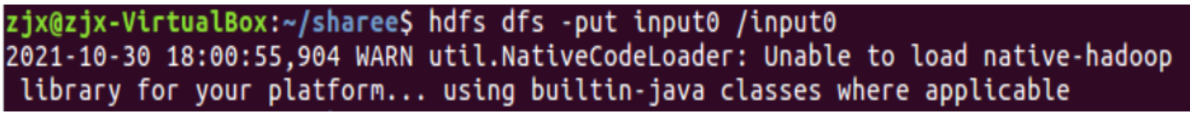
作业6

莎士比亚文集倒排索引

一、思路：

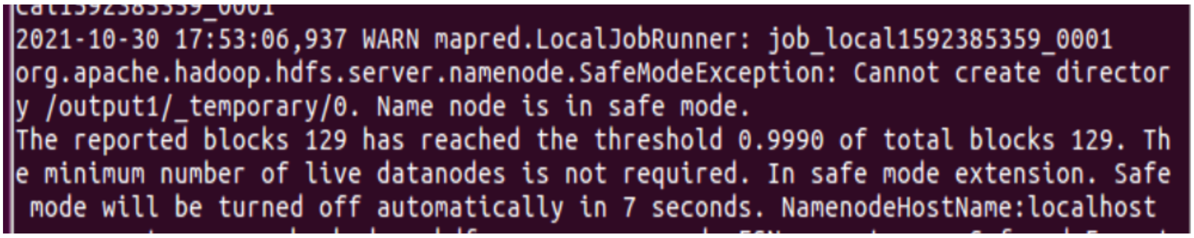
与作业5类似，先进行忽略大小写、忽略标点符号、忽略停词、忽略数字、单词长度>=3的词频统计，Map函数的输入键值对是行和文本，输出是单词和次数1，通过一个Reduce过程无法同时完成词频统计和生成文档列表，所以必须增加一个Combine过程完成词频统计。经过Combiner合并，Combine过程将key值相同的value值累加，得到一个单词在文档在文档中的词频，保证输出格式的key值按照word分发给Reducer，最后进行倒排索引写入到文件。

二、运行截图



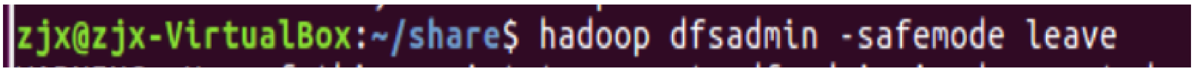
```
zjx@zjx-VirtualBox:~/sharee$ hdfs dfs -put input0 /input0
2021-10-30 18:00:55,904 WARN util.NativeCodeLoader: Unable to load native-hadoop
 library for your platform... using builtin-java classes where applicable
```

报错：



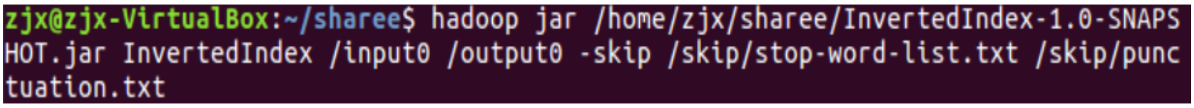
```
cal1592385359_0001
2021-10-30 17:53:06,937 WARN mapred.LocalJobRunner: job_local1592385359_0001
org.apache.hadoop.hdfs.server.namenode.SafeModeException: Cannot create director
y /output1/_temporary/0. Name node is in safe mode.
The reported blocks 129 has reached the threshold 0.9990 of total blocks 129. Th
e minimum number of live datanodes is not required. In safe mode extension. Safe
 mode will be turned off automatically in 7 seconds. NamenodeHostName:localhost
```

解决方法：强制退出安全模式

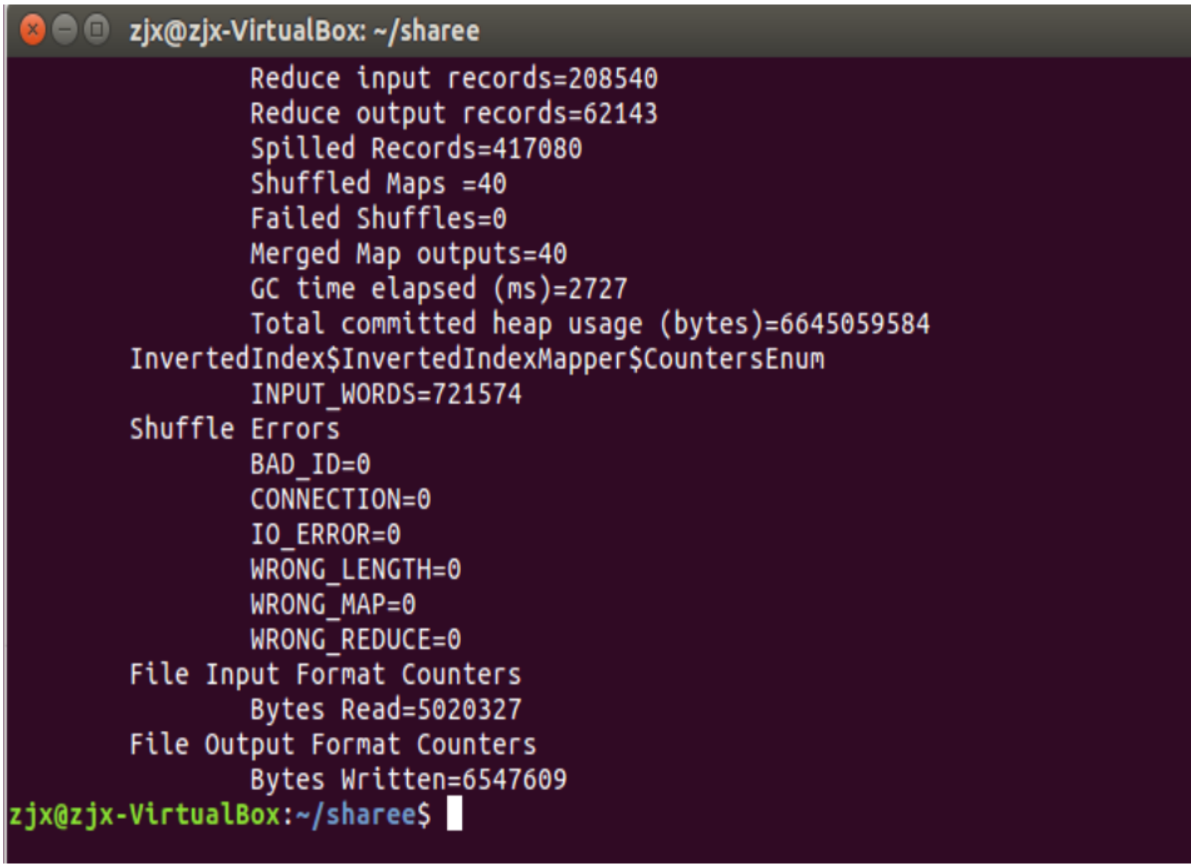


```
zjx@zjx-VirtualBox:~/share$ hadoop dfsadmin -safemode leave
```

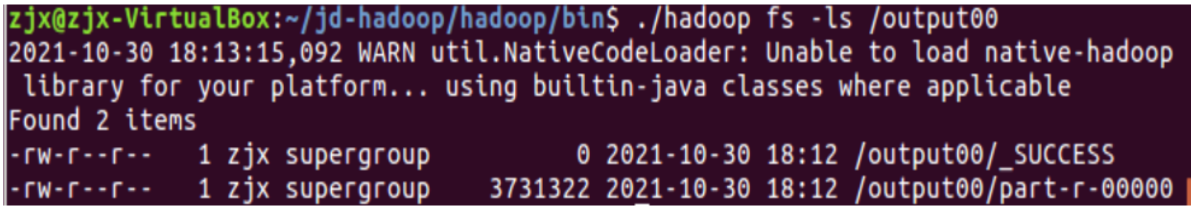
再次运行：



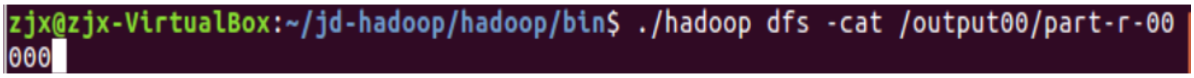
```
zjx@zjx-VirtualBox:~/sharee$ hadoop jar /home/zjx/sharee/InvertedIndex-1.0-SNAPS
HOT.jar InvertedIndex /input0 /output0 -skip /skip/stop-word-list.txt /skip/punc
tuation.txt
```



```
                Reduce input records=208540
                Reduce output records=62143
                Spilled Records=417080
                Shuffled Maps =40
                Failed Shuffles=0
                Merged Map outputs=40
                GC time elapsed (ms)=2727
                Total committed heap usage (bytes)=6645059584
        InvertedIndex$InvertedIndexMapper$CountersEnum
                INPUT_WORDS=721574
        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=5020327
        File Output Format Counters
                Bytes Written=6547609
zjx@zjx-VirtualBox:~/sharee$
```

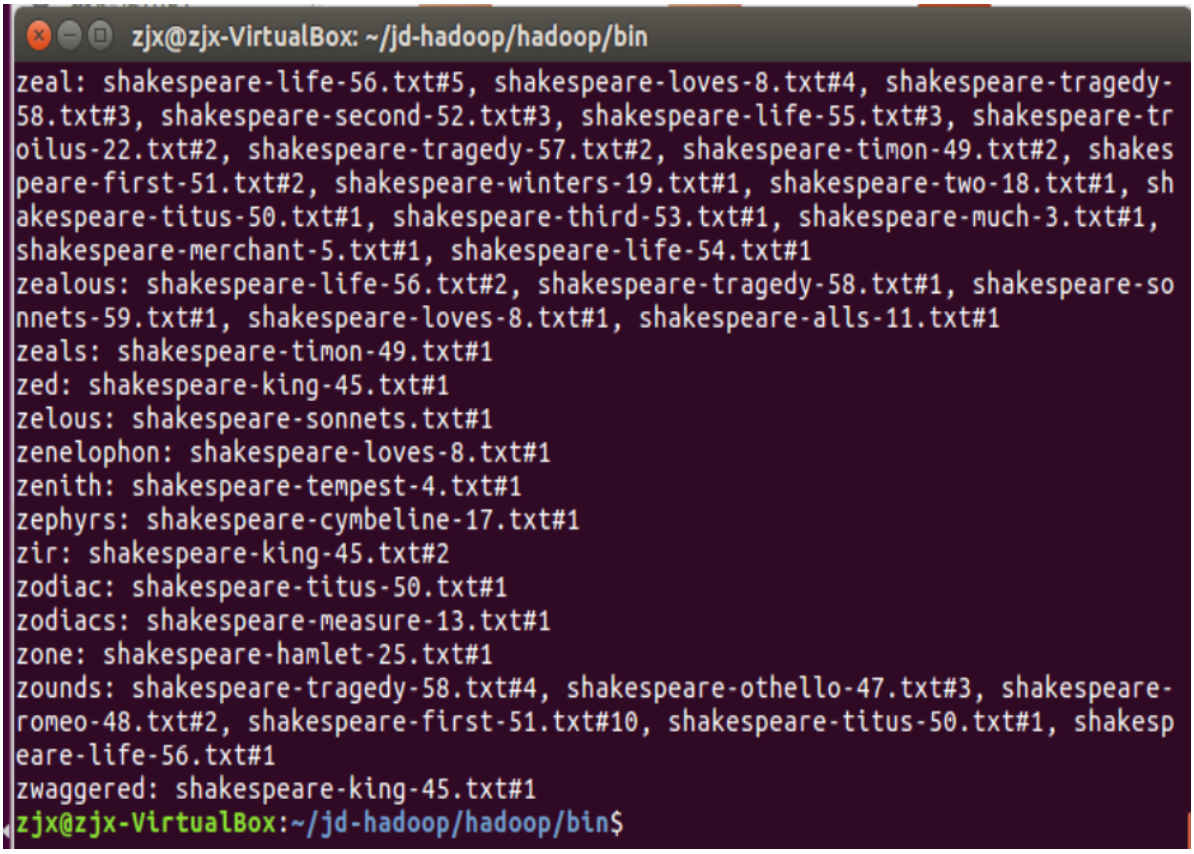


```
zjx@zjx-VirtualBox:~/jd-hadoop/hadoop/bin$ ./hadoop fs -ls /output00
2021-10-30 18:13:15,092 WARN util.NativeCodeLoader: Unable to load native-hadoop
 library for your platform... using builtin-java classes where applicable
Found 2 items
-rw-r--r--   1 zjx supergroup          0 2021-10-30 18:12 /output00/_SUCCESS
-rw-r--r--   1 zjx supergroup    3731322 2021-10-30 18:12 /output00/part-r-00000
```



```
zjx@zjx-VirtualBox:~/jd-hadoop/hadoop/bin$ ./hadoop dfs -cat /output00/part-r-00
000
```

结果截图：



```
zeal: shakespeare-life-56.txt#5, shakespeare-loves-8.txt#4, shakespeare-tragedy-
58.txt#3, shakespeare-second-52.txt#3, shakespeare-life-55.txt#3, shakespeare-tr
oilus-22.txt#2, shakespeare-tragedy-57.txt#2, shakespeare-timon-49.txt#2, shakes
peare-first-51.txt#2, shakespeare-winters-19.txt#1, shakespeare-two-18.txt#1, sh
akespeare-titus-50.txt#1, shakespeare-third-53.txt#1, shakespeare-much-3.txt#1,
shakespeare-merchant-5.txt#1, shakespeare-life-54.txt#1
zealous: shakespeare-life-56.txt#2, shakespeare-tragedy-58.txt#1, shakespeare-so
nnets-59.txt#1, shakespeare-loves-8.txt#1, shakespeare-alls-11.txt#1
zeals: shakespeare-timon-49.txt#1
zed: shakespeare-king-45.txt#1
zelous: shakespeare-sonnets.txt#1
zenelophon: shakespeare-loves-8.txt#1
zenith: shakespeare-tempest-4.txt#1
zephyrs: shakespeare-cymbeline-17.txt#1
zir: shakespeare-king-45.txt#2
zodiac: shakespeare-titus-50.txt#1
zodiacs: shakespeare-measure-13.txt#1
zone: shakespeare-hamlet-25.txt#1
zounds: shakespeare-tragedy-58.txt#4, shakespeare-othello-47.txt#3, shakespeare-
romeo-48.txt#2, shakespeare-first-51.txt#10, shakespeare-titus-50.txt#1, shakesp
eare-life-56.txt#1
zwaggered: shakespeare-king-45.txt#1
zjx@zjx-VirtualBox:~/jd-hadoop/hadoop/bin$
```