221275028-张伊璐-作业5

1.Task1

任务：统计数据集上市公司股票代码（ “stock” 列）的出现次数，按出现次数从大到小输出，输出格式为 "< 排名 >： < 股票代码 >， < 次数 > “；

1.1 设计思路

整体架构

- Mapper（TokenizerMapper）：负责解析输入的文本行，提取股票代码。
- Reducer（IntSumReducer）：负责汇总Mapper发送过来的数据，计算每个股票代码的出现次数。
- 辅助数据结构（StockCountPair）：用于存储股票代码及其对应的出现次数。
- 主程序（main）：配置和启动MapReduce作业。

Mapper类

- 功能：解析输入数据，提取股票代码。
- 实现细节：
  - 读取输入的文本行（每行是一个CSV记录）。
  - 使用逗号（`,`）作为分隔符，将行分割成多个字段。
  - 检查分割后的数组长度是否大于3，确保数据行包含足够的字段。
  - 提取最后一个字段作为股票代码（这里假设股票代码总是在每行的最后一个位置）。
  - 将股票代码作为key，值设为1（`IntWritable` 类型），发送给Reducer。

Reducer类

- 功能：接收来自Mapper的键值对，统计每个股票代码的出现次数。
- 实现细节：
  - 对于每个键（股票代码），累加其出现的次数。
  - 将股票代码及其出现次数存储在 `stockCounts` 列表中。
  - 在 `cleanup` 方法中，对 `stockCounts` 列表进行排序，按照出现次数从高到低。
  - 输出排序后的股票代码及其出现次数。

辅助类

- 功能：存储股票代码及其出现次数。

- 实现细节：
  - 包含两个字段：`stockCode`（字符串类型）和`count`（整型）。
  - 提供一个构造函数，用于初始化这两个字段。

1.2 问题与解决

问题：起初采用逗号分割，获取第四个元素的方式存储stock列数据，输出结果如下：

```
Portugal Plans Bills With Maturities Of More Than 1 Year"       1
Portugal to Portfolio of IAD Networks"  1
Position As An Intermediate Term Winner In Post-Industry Consolidation  1
Positioned to Compete on the Internet"  1
Positioning"    1
Positive Data From Its Affiliate Company        1
Positive Fundamentals"  1
Positive On Impacts Of 'hot weather'"   1
Positive On Outlook In Lieu Of Production Growth And Normalizing Oil Prices"
Positive On Upside Potential"   1
Positive Readout For Amgen's Blood Cancer Drug" 5
Positive Results For ViiV"      4
Positive Safety Review For Genfit's NASH Drug"  8
Possible FY17 Guidance Moderation"      1
Possible Strategic Combinations"        1
Post    2
Post-IPO TX Investment Now Nearly $300M"        1
Post-Weetabix   1
Pot Stocks And More"    1
Potash 8M-8.75M Tonnes  1
Potash Corp.    1
Potash Up 0.6%" 3
Potash with Overweight  1
Potbelly's      1
```

在任意搜索一个统计项后，发现统计项在csv文件中的位置如下：

```
855346,Royal Dutch Shell Forced to Pause Arctic Drilling Once Again,2012-09-17 00:00:00,MKO
11065,45 Biggest Movers From Yesterday,2018-06-08 00:00:00,ABM
187758,Bio-Reference Labs Reports Q2 EPS $0.33 vs $0.32 Est; Revenues $163.4M vs $161.48M Est,2012-06-07 00:00:00,BRLI
280020,"CyrusOne Reports Expected Closing of Exterior Shells for Houston West III, San Antonio II Projects, Post-IPO TX Investment Now Nearly $300M",2014-10-01 00:00:00,CONE
476827,Stocks That Hit 52-Week Lows On Thursday,2020-03-19 10:49:42-04:00,FBC
173419,Badger Meter Raises Qtr. Dividend from $0.17 to $0.18/Share,2013-08-09 00:00:00,BMI
292801,Benzinga's Top Initiations,2015-01-21 00:00:00,CPTA
132425,Alibaba Option Alert: Oct 20 $185 Calls Above Ask!: 11368 @ $1.95 vs 8174 OI; Ref=$165.1632,2017-08-17 00:00:00,BABA
```

可以发现，由于headline中也包含逗号，所以在用逗号分割后采用第四项时， 把headline中的内容当做了stock进行统计。

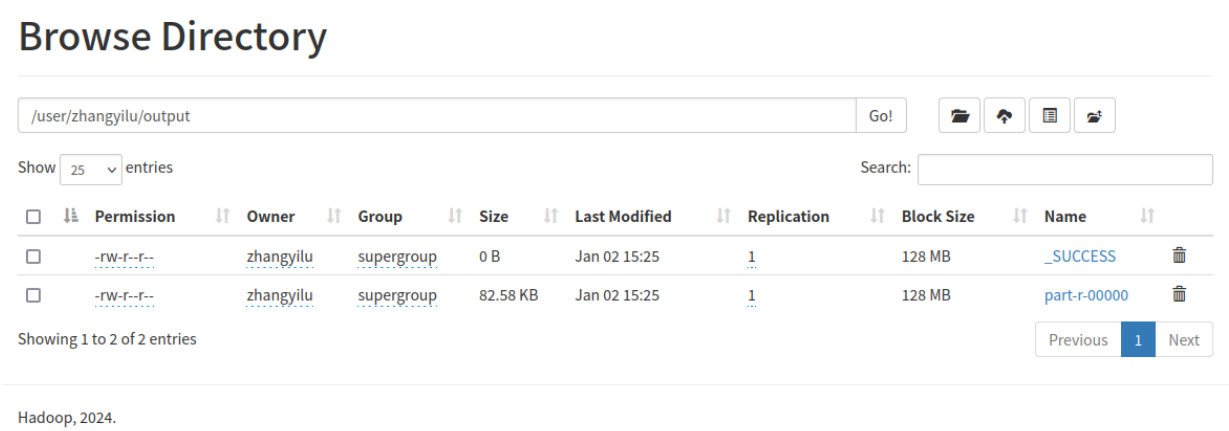
解决：把原来的采取第四项，改为采取最后一项，即stock项，代码如下

```
if (tokens.length > 3) {
                stockCode.set(tokens[tokens.length - 1].trim());
                context.write(stockCode, one);
            }
```

1.3 输出

```
                bytes written=84561
zhangyilu@zhangyilu-VMware-Virtual-Platform:/usr/local/hadoop$ ./bin/hdfs dfs -c
at output/*
1: MS   1174
2: MRK  1141
3: MU   1096
4: NVDA 1091
5: VZ   1080
6: NFLX 1078
7: QCOM 1051
8: BABA 1044
9: GILD 1041
```

web网页截图：



2.Task2

2.1 设计思路

整体架构

- 配置：首先，创建一个 `Configuration` 对象，用于存储 Hadoop 作业的配置信息。
- 设置停用词文件路径：通过 `conf.set("stopwords.file", "input/stop-word-list.txt");` 设置停用词文件的 HDFS 路径。
- 作业实例：使用 `Job.getInstance(conf, "top words")` 创建一个 MapReduce 作业实例。
- 设置作业参数：设置作业的 Jar 文件、Mapper 类、Reducer 类以及输出键值对的类型。
- 设置输入输出路径：使用 `FileInputFormat` 和 `FileOutputFormat` 分别设置作业的输入路径和输出路径。
- 启动作业：最后，调用 `job.waitForCompletion(true)` 启动作业，并等待其完成。

Mapper 类

- 成员变量：定义了一个 `HashSet` 来存储停用词，以及 `Text` 和 `IntWritable` 类型的变量用于输出键值对。
- `setup` 方法：在 Mapper 任务开始前，从 HDFS 加载停用词文件，并填充停用词集合。
- `map` 方法：对输入的每条记录进行处理。假设输入记录是由逗号分隔的值，程序只处理第二个字段（即标题）。对标题进行分词，移除非字母数字字符，然后检查分词是否为停用词。如果不是停用词，则输出该词和计数1。

Reducer 类

- 成员变量：定义了一个 `ArrayList` 来存储每个单词及其出现次数的 `WordCountPair` 对象。
- `reduce` 方法：对 Mapper 输出的相同单词的计数进行累加。
- `cleanup` 方法：在 Reduce 任务结束前，对所有单词按出现次数进行排序，并输出前100个最频繁出现的单词及其计数。

辅助类

- 成员变量：存储单词和计数。
- 构造函数：初始化单词和计数。

2.2 问题与解决

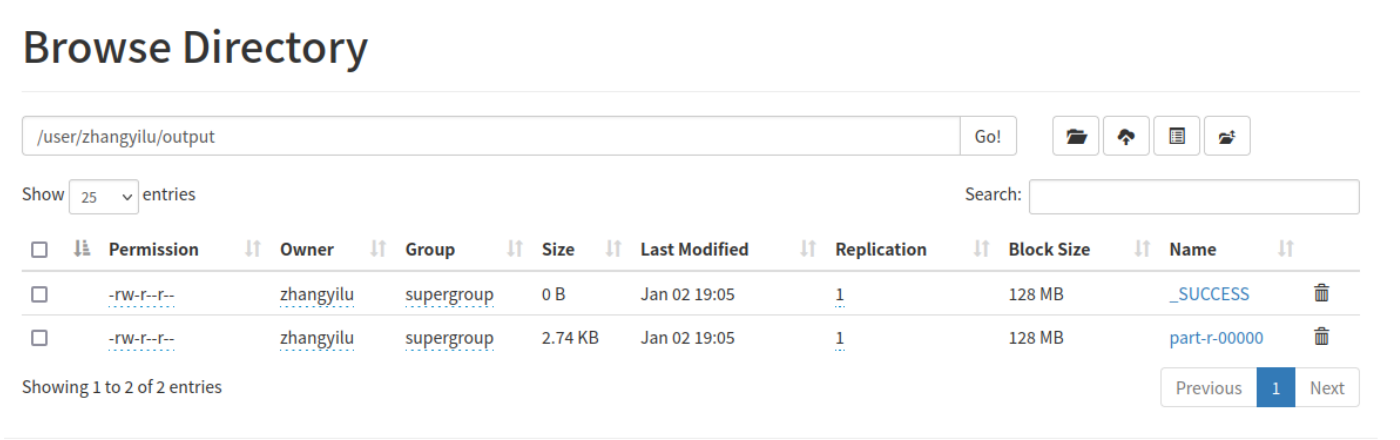
问题：不要起和文件名一样的包名否则会不幸，如The declared package "task2.example" does not match the expected package "task2.src.main.java.task2.example"

解决：把声明变成"task2.src.main.java.task2.example"再改回task2.example就好了

2.3 输出

```
zhangyilu@zhangyilu-VMware-Virtual-Platform:/usr/local/hadoop$ ./bin/hdfs dfs -c
at output/*
排名：1: stocks, 54724
排名：2: eps, 37999
排名：3: vs, 36769
排名：4: shares, 36249
排名：5: reports, 33652
排名：6: update, 31528
排名：7: est, 30372
排名：8: market, 29509
排名：9: earnings, 27749
排名：10: trading, 20508
排名：11: benzingas, 20077
排名：12: buy, 19952
排名：13: upgrades, 19498
```

web截图


Browse Directory
/user/zhangyilu/output
Go!
Show 25 entries
Search:
Permission
Owner
Group
Size
Last Modified
Replication
Block Size
Name
-rw-r--r--
zhangyilu
supergroup
0 B
Jan 02 19:05
1
128 MB
_SUCCESS
-rw-r--r--
zhangyilu
supergroup
2.74 KB
Jan 02 19:05
1
128 MB
part-r-00000
Showing 1 to 2 of 2 entries
Previous
1
Next