


```

tring line = value.toString();
String[] columns = line.split(","); // Assuming CSV format with ',' as
delimiter
if (columns.length >= 4) { // Ensure there are enough columns
    stockCode.set(columns[columns.length - 1].trim()); // Assuming stock
code is in the last column
    context.write(stockCode, one);
}

```

1.2、Reducer的设计思路

Reducer 包含有 reduce 函数和 cleanup 函数。

1. reduce 函数就是实现对于同一个key的 Iterable<IntWritable> values 中的每一个元素也就是统计数求和，最后再将求和结果存入一个叫做 TreeMap<Integer, String> countMap = new TreeMap<>() 的类里面以待 cleanup 函数实现对输出格式的处理。

注意：这里其实将 key 和 value 翻了过来，这样可以更加方便的直接在后面使用关于 value 也就是词频的倒序排序。

以下是 reduce 函数的主要功能语句：

```

int sum = 0;
for (IntWritable val : values) {
    sum += val.get();
}
countMap.put(sum, key.toString());

```

2. cleanup 函数就是实现对输出结果的格式控制。因为在 reduce 函数中将键值颠倒存储了，所有可以直接调用 Map.Entry<Integer, String> entry : countMap.descendingMap().entrySet() 方法实现关于 key 也就是词频的倒序排列；

该方法返回的是一个倒序列表，所以对于该倒序列表只需要 context.write 前100个字段即可，设置一个 for 循环，rank=1，最终只输出前100个即可退出循环，输出的每一个字段的 key 是 new Text(rank + ": " + entry.getValue() + ", " + entry.getKey()), value 为 null 即可完成格式修改。

以下是 cleanup 函数的主要功能语句：

```

int rank = 1;
for (Map.Entry<Integer, String> entry : countMap.descendingMap().entrySet())
{
    if (rank > 100) break;
    context.write(new Text(rank + ": " + entry.getValue() + ", " +
entry.getKey()), null);
    rank++;
}

```

注意：main 函数接口中需要添加 job.addCacheFile(new Path(args[2]).toUri()); 实现将缓存文件的参数也能传入至该程序的运行。

1.3、项目运行的配置设计

- 此次项目主要使用 Maven 进行项目管理，通过编辑 pom.xml 文件对该项目进行配置。pom.xml 文件的配置信息包含有该项目需要哪些库文件需要下载，该项目的项目文件有哪些。

- 依次使用 `mvn clean install` 进行配置，同时还可以使用 `mvn compile` 对 `.class` 文件进行生成，`mvn package` 实现对项目文件的 `.class` 文件打包成 `jar` 文件。
- 将 `analyst_ratings.csv` 上传至HDFS的 `/input` 文件夹里面，`stop-word-list.txt` 存入HDFS的 `/user/root` 文件夹里面，最后运行该项目的 `jar` 文件，运行命令为：

```
./hadoop jar /home/njucs/zuoye5_2/target/zuoye5_2-1.0-SNAPSHOT.jar  
HighFrequencyWords2 /input /output2 /user/root/stop-word-list.txt
```

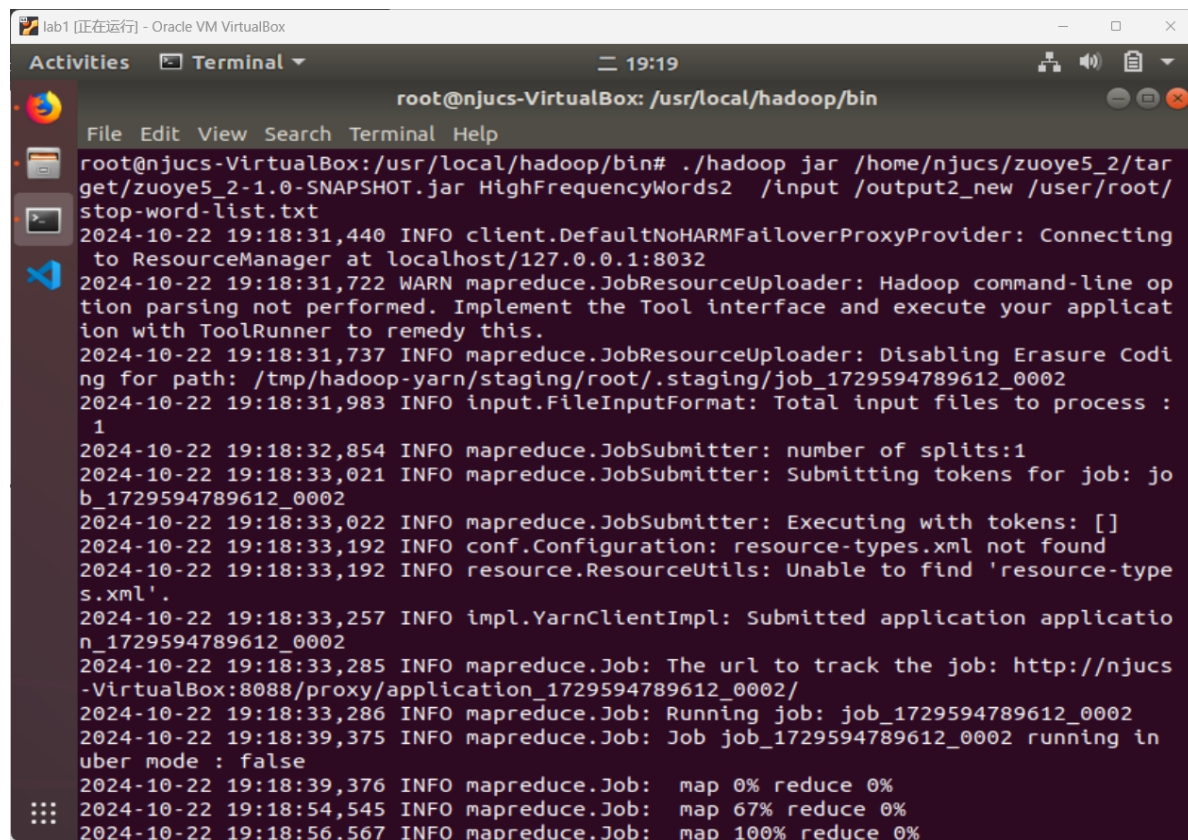
注意要把导出出来的 `part-r-00000` 解锁，以实现普通用户可以打开，命令如下：

```
sudo chown $USER part-r-00000
```

2、程序运行结果

以下即为 `HighFrequencyWords2.java` 程序执行的任务二（统计数据集热点新闻标题（“headline”列）中出现的前100个高频单词，按出现次数从大到小

输出。要求忽略大小写，忽略标点符号，忽略停词（`stop-word-list.txt`）。输出格式为“<排名>: <单词>, <次数>”的运行结果：



```
lab1 [正在运行] - Oracle VM VirtualBox
Activities Terminal 19:19
root@njucs-VirtualBox: /usr/local/hadoop/bin

File Edit View Search Terminal Help

root@njucs-VirtualBox: /usr/local/hadoop/bin# ./hadoop jar /home/njucs/zuoye5_2/target/zuoye5_2-1.0-SNAPSHOT.jar HighFrequencyWords2 /input /output2_new /user/root/stop-word-list.txt
2024-10-22 19:18:31,440 INFO client.DefaultNoHARMAFailoverProxyProvider: Connecting to ResourceManager at localhost/127.0.0.1:8032
2024-10-22 19:18:31,722 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
2024-10-22 19:18:31,737 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/root/.staging/job_1729594789612_0002
2024-10-22 19:18:31,983 INFO input.FileInputFormat: Total input files to process : 1
2024-10-22 19:18:32,854 INFO mapreduce.JobSubmitter: number of splits:1
2024-10-22 19:18:33,021 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1729594789612_0002
2024-10-22 19:18:33,022 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-10-22 19:18:33,192 INFO conf.Configuration: resource-types.xml not found
2024-10-22 19:18:33,192 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2024-10-22 19:18:33,257 INFO impl.YarnClientImpl: Submitted application application_1729594789612_0002
2024-10-22 19:18:33,285 INFO mapreduce.Job: The url to track the job: http://njucs-VirtualBox:8088/proxy/application_1729594789612_0002/
2024-10-22 19:18:33,286 INFO mapreduce.Job: Running job: job_1729594789612_0002
2024-10-22 19:18:39,375 INFO mapreduce.Job: Job job_1729594789612_0002 running in uber mode : false
2024-10-22 19:18:39,376 INFO mapreduce.Job: map 0% reduce 0%
2024-10-22 19:18:54,545 INFO mapreduce.Job: map 67% reduce 0%
2024-10-22 19:18:56,567 INFO mapreduce.Job: map 100% reduce 0%
```

```
lab1 [正在运行] - Oracle VM VirtualBox
Activities Terminal 19:19
root@njucs-VirtualBox: /usr/local/hadoop/bin

File Edit View Search Terminal Help

2024-10-22 19:18:39,375 INFO mapreduce.Job: Job job_1729594789612_0002 running in
uber mode : false
2024-10-22 19:18:39,376 INFO mapreduce.Job: map 0% reduce 0%
2024-10-22 19:18:54,545 INFO mapreduce.Job: map 67% reduce 0%
2024-10-22 19:18:56,567 INFO mapreduce.Job: map 100% reduce 0%
2024-10-22 19:19:04,623 INFO mapreduce.Job: map 100% reduce 100%
2024-10-22 19:19:05,637 INFO mapreduce.Job: Job job_1729594789612_0002 completed s
uccessfully
2024-10-22 19:19:05,706 INFO mapreduce.Job: Counters: 54
    File System Counters
        FILE: Number of bytes read=112393618
        FILE: Number of bytes written=169209999
        FILE: Number of read operations=0
        FILE: Number of large read operations=0
        FILE: Number of write operations=0
        HDFS: Number of bytes read=52463065
        HDFS: Number of bytes written=1719
        HDFS: Number of read operations=8
        HDFS: Number of large read operations=0
        HDFS: Number of write operations=2
        HDFS: Number of bytes read erasure-coded=0
    Job Counters
        Launched map tasks=1
        Launched reduce tasks=1
        Data-local map tasks=1
        Total time spent by all maps in occupied slots (ms)=15528
        Total time spent by all reduces in occupied slots (ms)=5288
        Total time spent by all map tasks (ms)=15528
        Total time spent by all reduce tasks (ms)=5288
```

```
lab1 [正在运行] - Oracle VM VirtualBox
Activities Terminal 19:19
root@njucs-VirtualBox: /usr/local/hadoop/bin

File Edit View Search Terminal Help

    Reduce shuffle bytes=56196806
    Reduce input records=4365731
    Reduce output records=100
    Spilled Records=13097193
    Shuffled Maps =1
    Failed Shuffles=0
    Merged Map outputs=1
    GC time elapsed (ms)=129
    CPU time spent (ms)=20650
    Physical memory (bytes) snapshot=875847680
    Virtual memory (bytes) snapshot=5244391424
    Total committed heap usage (bytes)=760217600
    Peak Map Physical memory (bytes)=491737088
    Peak Map Virtual memory (bytes)=2620211200
    Peak Reduce Physical memory (bytes)=384110592
    Peak Reduce Virtual memory (bytes)=2624180224
    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=52462953
    File Output Format Counters
        Bytes Written=1719
root@njucs-VirtualBox: /usr/local/hadoop/bin#
```

```
lab1 [正在运行] - Oracle VM VirtualBox
Activities Terminal 19:08
root@njucs-VirtualBox: /usr/local/hadoop/bin

File Edit View Search Terminal Help

root@njucs-VirtualBox:/usr/local/hadoop/bin# ./hadoop fs -cat /output1_new/part-r-00000
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
10: EBAY, 1037
11: QQQ, 1029
12: M, 1022
13: DAL, 1011
14: JNJ, 1007
15: KO, 960
16: FDX, 951
17: AA, 941
18: WFC, 926
19: ORCL, 914
20: EWU, 914
21: BMY, 882
22: HD, 875
23: JCP, 865
24: BBRY, 858
25: EWJ, 848
26: AGN, 833
27: GPRO, 827
```

3、WEB页面截图

lab1 [正在运行] - Oracle VM VirtualBox

Activities Firefox Web Browser 22:46

All Applications - Mozilla Firefox

All Applications

localhost:8088/cluster 50%

loop All Applications

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Used Resources
2	0	0	2	0	<memory:0 B, vCores:0>

Cluster Nodes Metrics

Active Nodes	Decommissioning Nodes	Decommissioned Nodes	Lost Nodes
1	0	0	0

Scheduler Metrics

Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocation	Maximum Cluster Application P
Capacity Scheduler	[memory-mb (unit=Mi), vcores]	<memory:1024, vCores:1>	<memory:8192, vCores:4>	0

Show: 20 entries

ID	User	Name	Application Type	Application Tags	Queue	Application Priority	StartTime	LaunchTime	FinishTime	State	FinalStatus	Running Containers
application_1729608133545_0002	root	high frequency words	MAPREDUCE		root.default	0	Tue Oct 22 22:45:20 +0800 2024	Tue Oct 22 22:45:21 +0800 2024	Tue Oct 22 22:45:50 +0800 2024	FINISHED	SUCCEEDED	N/A
application_1729608133545_0001	root	stock count	MAPREDUCE		root.default	0	Tue Oct 22 22:44:23 +0800 2024	Tue Oct 22 22:44:24 +0800 2024	Tue Oct 22 22:44:41 +0800 2024	FINISHED	SUCCEEDED	N/A

Showing 1 to 2 of 2 entries