作业5

目的

掌握MapReduce编程方法，掌握用MapReduce解决常见的数据处理问题。

平台

已经配置完成的Hadoop伪分布式环境或集群环境。

要求

在HDFS上加载上市公司热点新闻标题数据集（analyst_ratings.csv），该数据集收集了部分上市公司的热点财经新闻标题。编写MapReduce程序完成以下两项任务：

1. 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>“；
2. 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>“。

数据集

数据集格式：<索引>，<标题>，<发布时间戳记>，<股票代码>

数据文件：analyst_ratings.csv

停词文件：stop-word-list.txt

提交方式

git仓库地址或者相关文件的zip包，包含源代码和输出文件。git仓库目录组织建议：

```
.(Project Name)
├── src
├── target   (只保留jar文件，并忽略其它无关文件)
├── output
|   └── part-r-00000 (输出结果文件)
├── pom.xml
├── .gitignore
└── README.md (对设计思路，程序运行结果等给出说明，并给出提交作业运行成功的WEB页面截图。可以进一步对性能、扩展性等方面存在的不足和可能的改进之处进行分析。)
```