作业5

目的

掌握MapReduce编程方法，学习用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上加载股票数据集（stock_data.csv），该数据集收集了部分上市公司的热点财经新闻标题及该新闻对应的情感标签，1表示正面，-1表示负面。编写伪代码描述MapReduce程序完成下述任务，并编写程序实现。

任务说明：统计数据集里正面和负面新闻标题（“Text”列）中分别出现的前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，忽略数字，忽略停词（stop-word-list.txt）。按照正面和负面分别输出统计结果文件，输出格式为"<单词> \TAB <次数>“。

数据集

数据集格式：<标题>，<情感标签>

数据文件：stock_data.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。)
```