lab4 实验报告 161278050

161278050 张旭

一、实验需求：

- 主题：上市公司财经新闻情感分析
- 背景：互联网技术不断发展，给人类带来了更快速的信息传播媒介。在这个互联网时代，不仅是时事新闻，股市新闻传播地也更加快速。股市新闻中往往包含了大量信息，除了上市公司的财务数据外，还包括经营公告、行业动向、国家政策等大量文本信息，这些文本信息中常常包含了一定的情感倾向，会影响股民对公司股票未来走势的预期，进一步造成公司的股价波动。如果能够挖掘出这些新闻中蕴含的情感信息，则可以对股票价格进行预测，对于指导投资有很大的作用。本实验尝试使用文本挖掘技术和机器学习算法，挖掘出新闻中蕴含的情感信息，分别将每条新闻的情感判别为“positive”、“neutral”、“negative”这三种情感中的一种，可根据抓取的所有新闻的情感汇总分析来对股票价格做预测。
- 实验目标：使用多种机器学习算法对文本进行情感判别，包括 KNN、决策树、朴素贝叶斯、支持向量机等，学习如何进行模型训练，如何进行分类预测。要求使用至少两种分类方法。
- 要求：核心程序在 MapReduce 上运行，要求使用至少两种分类方法。

二、实验设计思路：

三、代码及类设计：

1.Test.py:

- ✓ Step1:读取 fulldata.txt,分词提取新闻标题，生成新闻标题分词列表 docs[]
- ✓ Step2:读取特征词文件 chi_words.txt,生成特征词列表 words[]
- ✓ Step3:计算 cfs[]
- ✓ Step4:计算 tfs[]
- ✓ Step5:计算 idfs[]
- ✓ Step6:计算 tfidf[]，生成向量列表
- ✓ Step7: 按照要求格式 [title+tfidf[]+-1] 输出至 testData.txt 或 NBayes.test

2.train.py:

- ✓ Step1: 读取特征词文件 chi_words.txt,生成特征词列表 words[]
- ✓ Step2：循环遍历三个类别文件夹，读取训练集文件，中文分词，生成 docs[]，计算 tfidf[]向量，按要求格式[tfidf[]+label]逐行输出至 trainData.txt

  Label(0:negative,1:neutral,2:positive)

3.KNN0.java:

1) protected void setup(Context context)：读取文件，生成训练集列表
2) protected void map(LongWritable k1, Text v1,Context context)：计算欧式距离最近的 label，键值对（标题文本，标签）
3) public static class MyReducer extends Reducer<Text, Text, Text,

NullWritable>：计算出频率最高的的 label，键值对（标题文本，预测类别）输出。

4) public static class Distance：计算欧式距离

5) public static class Instance：生成训练集数据样例

6) public static class TestInstance：生成测试集数据样例

4. public class NaiveBayesMain：读取配置和输入文件，运行各种类文件

5. public class NaiveBayesTrain：并行化处理训练集文件

6. public class NaiveBayesConf：读取配置文件

7. public class NaiveBayesTrainData：读取并处理测试集数据

8. public class NaiveBayesTest：并行化处理根据朴素贝叶斯算法计算并输出（标题文本，标签类别）

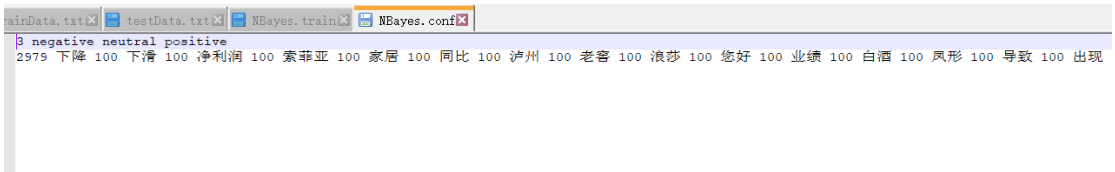
四、实验流程及结果截图：

1.数据预处理，用 tfidf 算法和特征词文件对测试集和训练集数据进行向量化处理，生成相应文件。

NBayes.conf

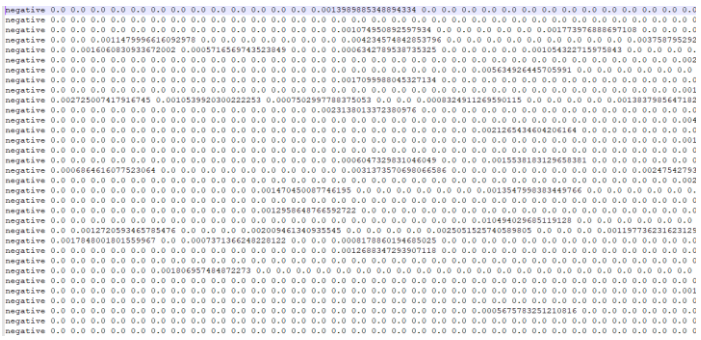
第一行：类别数量+标签名

第二行：特征属性数量+<特征名 最大值>



TestData.TXT

train.txt



2.编译运行 knn.java，NBayesMain.java 文件，对数据进行训练和分类预测。

```
[root@localhost src]# /opt/hadoop/hadoop-2.9.1/bin/hadoop jar nbm.jar NaiveBayesMain /naivebayes NBayes.conf NBayes.train NBayes.test NBayes.out
18/12/25 21:13:55 INFO client.RMProxy: Connecting to ResourceManager at localhost/127.0.0.1:8032
18/12/25 21:13:56 INFO input.FileInputFormat: Total input files to process : 1
18/12/25 21:13:56 INFO mapreduce.JobSubmitter: number of splits:1
18/12/25 21:13:57 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enabled
18/12/25 21:13:57 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1540904211849_0127
18/12/25 21:13:57 INFO impl.YarnClientImpl: Submitted application application_1540904211849_0127
18/12/25 21:13:57 INFO mapreduce.Job: The url to track the job: http://localhost:8088/proxy/application_1540904211849_0127/
18/12/25 21:13:57 INFO mapreduce.Job: Running job: job_1540904211849_0127
18/12/25 21:14:09 INFO mapreduce.Job: Job job_1540904211849_0127 running in uber mode : false
18/12/25 21:14:09 INFO mapreduce.Job:  map 0% reduce 0%
18/12/25 21:14:26 INFO mapreduce.Job:  map 100% reduce 0%
18/12/25 21:14:36 INFO mapreduce.Job:  map 100% reduce 100%
18/12/25 21:14:36 INFO mapreduce.Job: Job job_1540904211849_0127 completed successfully
18/12/25 21:14:36 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=711310
                FILE: Number of bytes written=1819061
```

```
[root@localhost hadoop-2.9.1]# bin/hadoop jar knn.jar KNN0 /input /output
18/12/24 21:16:51 INFO client.RMProxy: Connecting to ResourceManager at localhost/127.0.0.1:8032
18/12/24 21:16:52 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
18/12/24 21:16:53 INFO input.FileInputFormat: Total input files to process : 1
18/12/24 21:16:53 INFO mapreduce.JobSubmitter: number of splits:1
18/12/24 21:16:53 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enabled
18/12/24 21:16:53 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1540904211849_0124
18/12/24 21:16:54 INFO impl.YarnClientImpl: Submitted application application_1540904211849_0124
18/12/24 21:16:54 INFO mapreduce.Job: The url to track the job: http://localhost:8088/proxy/application_1540904211849_0124/
18/12/24 21:16:54 INFO mapreduce.Job: Running job: job_1540904211849_0124
18/12/24 21:17:06 INFO mapreduce.Job: Job job_1540904211849_0124 running in uber mode : false
18/12/24 21:17:06 INFO mapreduce.Job:  map 0% reduce 0%
18/12/24 21:17:14 INFO mapreduce.Job:  map 100% reduce 0%
```

低市盈率概念炒作延伸，中原高速直线涨停 2.0
供给端发力促涨价潮现部分资源品种进入中线横盘期 0.0
兖州煤业(01171)、华电国际（01071）被剔出中证恒生沪港通AH股精明指数 2.0
全国十大钢铁企业济钢钢铁生产线全线停产 2.0
八一钢铁披露重组标的拟购控股股东旗下资产 2.0
六机构齐呼以业绩估值为基准抓反弹284低估值股扎堆三行业 2.0
制造业公司中报已发布297份三维度掘金95家绩优公司 2.0
叫停10亿超短融！债市倒春寒波及中原高速 1.0
周期股引发业内人士 “大战”各路资金搏杀正酣 1.0
四利好助周期类股持续走强钢铁等3板块吸金逾49亿元 2.0
四行业紧锣密鼓央企重组 “神剌百家” 2.0
国家电力投资集团公司公开发行2017年公司债券（第二期）募集说明书摘要 2.0
地条钢整顿再度加码钢铁企业一改颓势 0.0
多重利好助推核电板块崛起机构预计6只个股涨幅超50% 2.0
宝钢包装半年度业绩预亏 2.0
山东钢铁上半年净利近60亿同比增长20倍 2.0
投资金额高达6000亿上市公司疯狂买理财 2.0
操盘手赠言：庄家如何做庄，读懂参透赚钱越来越轻松 2.0
散户谨记选股铁律，从此一买就涨，捕捉暴涨黑马股！ 2.0
新浪财经晚报：11月16日晚间影响市场消息汇总 2.0
新规下可转债优势凸显38公司终止定增改发可转债 2.0
民间股神10年总结的盈利秘笈，炒股赚钱，一招搞定 2.0
水煮旧闻丨新周期？十年前钢铁股狂飙，就这样解释！ 2.0
河南又一资产管理巨鳄浮出水面，九大豪华股东曝光 2.0

3.分类预测结果截图：

part-r-00000 part-r-00000

"一带一路"高峰论坛举办在即工程机械等领域迎上涨契机 1.0
"兜底式增持"江湖重现中国好老板"能否兑现承诺存疑 1.0
"双徐"对调落定一汽系公司与长安汽车股票一涨一跌 1.0
*ST建峰重组获批明年有望恢复上市 1.0
*ST爱富复牌涨停引爆上海国企国资改革进入亮剑时间 1.0
1019份年报泄露QFII动向重仓82家公司青睐钢铁银行 0.0
101只个股连续上涨超过五日 0.0
104只股短线走稳站上五日均线 1.0
105只股短线走稳站上五日均线 1.0
107只股中线走稳站上半年线 1.0
10只个股大宗交易超5000万元 1.0
10只分级B仍存下折风险部分重仓股存下跌压力 1.0
10日午间研报精选10股望爆发 1.0
10日财经要闻：神雾环保闪崩、融创收购万达项目等 1.0
10月12日晚间上市公司利好消息一览 1.0
11月11日上市公司晚间公告速递 1.0
11月22日上市公司晚间公告速递 1.0
11月23日上市公司重要公告集锦 1.0
11月2日上市公司晚间公告速递 2.0
123只个股连续上涨超过五日 1.0
12亿主力资金近三日撤出大宗商品概念股 1.0
12月13日涨停揭秘：股权板块活跃，抄底游资增多 2.0
12月23日涨停揭秘庄妖股齐杀跌，增强现实发红包 1.0
12月25日上市公司晚间公告速递 1.0
12月26日上市公司晚间公告速递 2.0
133只个股连续上涨超过五日 1.0
1346只个股获机构推荐集中看好三行业54只个股 1.0
137只个股连续上涨超过五日 0.0
139只股中线走稳站上半年线 1.0
139只股短线走稳站上五日均线 1.0
13只个股大宗交易超5000万元 1.0
13车企半年报净利增1倍西仪股份预计净利增长3倍 1.0
14只个股大宗交易超5000万元 1.0
14只被集中持有个股跑赢大盘有机构坚守茅台10年 1.0
14家投资者参与联通混改方案公布5小时后撤下 0.0

part-r-00000 part-r-00000

"双徐"对调落定一汽系公司与长安汽车股票一涨一跌 negative
1019份年报泄露QFII动向重仓82家公司青睐钢铁银行 negative
101只个股连续上涨超过五日 negative
107只股中线走稳站上半年线 negative
10只个股大宗交易超5000万元 negative
10只分级B仍存下折风险部分重仓股存下跌压力 negative
10日财经要闻：神雾环保闪崩、融创收购万达项目等 negative
11月22日上市公司晚间公告速递 negative
11月23日上市公司重要公告集锦 negative
123只个股连续上涨超过五日 negative
12月25日上市公司晚间公告速递 negative
12月26日上市公司晚间公告速递 negative
1346只个股获机构推荐集中看好三行业54只个股 negative
137只个股连续上涨超过五日 negative
13只个股大宗交易超5000万元 negative
14只个股大宗交易超5000万元 negative
14只个股大宗交易超5000万元 negative
14只被集中持有个股跑赢大盘有机构坚守茅台10年 negative
15.4亿溢价接盘14.7%股权中钰资本成金字火腿二股东 negative
156只个股连续上涨超过五日 negative
16只股中线走稳站上半年线 negative
17日晚间机构研报精选：10股有望爆发 negative
181只股短线走稳站上五日均线 negative
186上市公司中报出炉近八成净利增长48家增幅超100% negative
186上市公司中报出炉近八成净利增长48家增幅超100% negative
18只股中线走稳站上半年线 negative
19家名单已确定，第三批试点中，这些你不得不看 negative
19股业绩连续三年高增长股价却表现欠佳 negative
2013年湖北楚天高速公路股份有限公司公司债券（第一期）2017年付息公告 negative
2016"妖股"呈现六大套路今年播种关注短期题材股 negative
2016中国医药工业百强、商业50强榜单出炉！ negative
2016年报：5家上市公司员工平均薪酬超百万元 negative
2017年世界500强榜单发布：国家电网、中石化居二三位 negative
2017年金牌董秘获奖全名单 negative
2017证券公司分类出炉 最惨的券商被连降六级 negative

五、实验总结：

1.分类结果 knn 算法分类结果比较多元，但关于情感的分类并不准确，可能是因为训练集本身情感分类标签并不准确。贝叶斯算法结果全是 negative，可能是因为新闻标题内容较少，大部分特征值为 0 造成结果的偏差。

2.预处理阶段 tfidf 计算使用 python 进行的串行计算，效率较低。