金融大数据处理技术Project2

151220105 王舸帆

1. 实验内容

上市公司财经新闻情感分析

2. 设计思路&算法

模型建立

使用chi_words，效果不佳；
在整个训练集里，利用频繁挖掘模式，找出权值最高的前n（n=1000）个词语。这n个词语构成情感分析的训练模型（该权值与该词出现的频次以及其对于该文本特征的贡献正相关，与wordcount有相似之处）。
模型的部分词语：

```
股份
行业
证券
品牌
培训
投资
业务
研发
艺术
生产
汽车
亿元
万元
销售
国内
未来
集团
机构
天际
提升
产业
合作
平台
上市
智能
领域
凤形
电梯
直升机
质量
优势
需求
创新
资金
预计
自主
```

分词

对待分类文本和训练文本进行分词。在project1的基础上做了改进，清洗了年份，符号和字母等无用的部分。

tfidf

利用tfidf将每个文本的每个词赋予一个权值，每个词的权值与它在该文本中出现的次数正相关，与在整个文件夹出现的次数反相关。
得到两个文件夹：`training_tfidfout`（训练集）&`stock_tfidfout`（待分类文本）

因为得到的tfidf均为小数，所以对所有值进行了统一倍数的放大。

文本向量化

对于每一个文本（包括训练集和待分类文本），建立一个k维数组（与模型维度对应），初始化所有值均为零。
遍历文本中的词语，如果模型中存在该word，将k维数组对应位置的元素改为该word在该文本的tfidf值，没有则为零。
这样每一个文本就可以用一个k维数组来表示

```
while ((temp0 = br1.readLine()) != null) {
    temp1 = temp0.split( regex: " ");
    k = dic.indexOf(temp1[0]);
    if (k != -1) {
        vec[k] = Double.parseDouble(temp1[1]);
    }
}
```

比如`52 0 0 34.....`代表在该文本中股份的tf值为52，品牌的tf值为34，没有行业和证券。

knn分类

原理：

在训练集中找出与待分类文本距离最近的k个点，根据这k个点的类别决定待分类文本的类别。

数据格式：

前n-1列为属性值，最后一列为类别

```
0.2607    0.3502    1
0.2875    0.6713    -1
0.916    0.7363    -1
0.1615    0.2564    1
0.2653    0.9452    1
0.0911    0.4386    0
0.0012    0.3947    -1
0.4253    0.8419    1
0.0067    0.4424    -1
0.8244    0.2089    1
0.3868    0.3592    0
0.9174    0.216    -1
```

待分类文本格式：
每一行为属性值

```
0.9516    0.0326
0.9203    0.5612
0.0527    0.8819
0.7379    0.6692
0.2691    0.1904
0.4228    0.3689
0.5479    0.4607
0.9427    0.9816
```

pro2中的格式在此基础上扩大规模。

朴素贝叶斯

原理：

- 设每个数据样本用一个n维特征向量来描述n个属性的值，即：$X=\{x1,x2,...,xn\}$，假定有m个类，分别用$Y1,Y2,...Ym$表示
- 给定一个未分类的数据样本X，若朴素贝叶斯分类将未知的样本X分配给类Yi，则一定有$P(Yi|X)>P(Yj|X)$，$1\le j\le m,\ j\ne i$
- 根据贝叶斯定理
- $P(Yi|X)=\frac{P(Yi)*P(X|Yi)}{\sum_{i=1}^{k}P(Yi)*P(X|Yi)}=\frac{P(Yi)*P(X|Yi)}{P(X)}$
- 由于$P(X)$对于所有类为常数，概率$P(Yi|X)$可转化为概率$P(X|Yi)*P(Yi)$。

- 如果训练数据集中有很多具有相关性的属性，计算$P(X|Yi)$将非常复杂，为此，通常假设各属性是互相独立的，这样$P(X|Yi)$的计算可简化为求$P(x1|Yi)$，$P(x2|Yi)$，...，$P(xn|Yi)$之积；而每个$P(xj|Yi)$可以从训练数据集近似求得。

n个属性：k维向量的每一维就是一个向量

m类：`neu` `neg` `pos`

数据格式：

待分类样本：每一行一个属性值

训练集：第一行为类名，其余行为属性值

3. 运行过程

DataDealing.java预处理数据

对新闻文本进行分词，并清洗分词后的结果
结果：文件夹`Data1` `stock`
注意：在不同感情色彩的新闻标题文件名之前分别加上前缀：`neu` `pos` `neg`

Tfidf.java

输入：`Data1`以及`stock`
输出：`traing_tfidfout` `stock_tfidfout`

```
合规 2207.0
中投 2060.0
检查 1683.0
零七 1390.0
研 1243.0
证监局 1212.0
因在 1002.0
放卫星 907.0
强推 852.0
监管 832.0
次数 812.0
事件 811.0
头条 782.0
尘埃落定 782.0
过高 782.0
研究报告 757.0
报 742.0
内部 727.0
责令 701.0
提交 674.0
深圳 666.0
通报 567.0
增加 565.0
证券 550.0
备受 546.0
```

(tfidf值均扩大了一万倍)

TxtVec.java

将上一步的结果进行向量化处理
输入：`traing_tfidfout` `stock_tfidfout`
输出：`data` `train`
data:
每一行是代表一个待分类文本的k维数组
train：
每一行是代表一个训练文本的k维数组加上它的类别

knn分类

输入：
`conf`:配置文件
`train`：训练集

`test`：待分类文本

输出：

`knn_out`

大部分是positive，说明我国经济目前发展良好。

knn_out

```
1:neg   2:pos   3:neu
sh603096新经典.txt 1（测试组：1）类别为：2
sz002631德尔未来.txt 2（测试组：2）类别为：2
sh600137浪莎股份.txt 3（测试组：3）类别为：2
sh600510黑牡丹.txt 4（测试组：4）类别为：2
sz300588熙菱信息.txt 5（测试组：5）类别为：2
sh600057象屿股份.txt 6（测试组：6）类别为：2
sh603167渤海轮渡.txt 7（测试组：7）类别为：2
sz000786北新建材.txt 8（测试组：8）类别为：2
sh603569长久物流.txt 9（测试组：9）类别为：2
sh600035楚天高速.txt 10（测试组：10）类别为：2
sz000707双环科技.txt 11（测试组：11）类别为：1
sh600790轻纺城.txt 12（测试组：12）类别为：2
sz000656金科股份.txt 13（测试组：13）类别为：2
sz002174游族网络.txt 14（测试组：14）类别为：2
sz002026山东威达.txt 15（测试组：15）类别为：2
sz300405科隆股份.txt 16（测试组：16）类别为：2
sz002790瑞尔特.txt 17（测试组：17）类别为：1
sh600604市北高新.txt 18（测试组：18）类别为：1
sz002846英联股份.txt 19（测试组：19）类别为：2
sh600807天业股份.txt 20（测试组：20）类别为：2
sh600683京投发展.txt 21（测试组：21）类别为：2
```

双环科技（neg）：

sz000707双环科技.txt — 已编辑

```
sz000707  2017-08-11  19:21  湖北宜化年内出4起事故受重罚 实控人急速“跑路”  http://
finance.sina.com.cn/roll/2017-08-11/doc-ifyixcaw4258032.shtml
sz000707  2017-08-08  16:47  双环科技实控人或变更  http://finance.sina.com.cn/stock/t/
2017-08-08/doc-ifyitayr9818328.shtml
sz000707  2017-08-08  16:44  8月8日上市公司晚间公告速递  http://finance.sina.com.cn/stock/s/
2017-08-08/doc-ifyitapp2824662.shtml
sz000707  2017-08-08  16:36  双环科技：宜化及其控股股东正筹划控股权转让事宜  http://
finance.sina.com.cn/stock/s/2017-08-08/doc-ifyitamv7244496.shtml
sz000707  2017-08-08  11:48  宜化集团或将筹划重大事项？旗下两上市公司临停  http://
finance.sina.com.cn/stock/s/2017-08-08/doc-ifyitamv7106214.shtml
sz000707  2017-05-20  01:06  湖北双环科技股份有限公司2017年第四次临时股东大会决议公告  http://
finance.sina.com.cn/stock/t/2017-05-20/doc-ifyfkqwe0379570.shtml
sz000707  2017-05-10  07:15  5月10日上市公司重要公告集锦  http://cj.sina.com.cn/article/detail/
2311077472/244234
sz000707  2017-05-10  03:16  湖北双环科技股份有限公司公告（系列）  http://finance.sina.com.cn/
stock/t/2017-05-10/doc-ifyexxhw3005675.shtml
sz000707  2017-05-04  02:20  湖北双环科技股份有限公司公告（系列）  http://finance.sina.com.cn/
stock/t/2017-05-04/doc-ifyeycte8520866.shtml
sz000707  2017-04-06  04:10  湖北双环科技股份有限公司八届三十三次董事会决议公告  http://
finance.sina.com.cn/stock/t/2017-04-06/doc-ifyecfnu7391846.shtml
sz000707  2017-03-31  15:31  湖北宜化化工股份有限公司2016年度报告摘要  http://
finance.sina.com.cn/stock/t/2017-03-31/doc-ifycwyxr8962799.shtml
sz000707  2017-03-29  13:55  化工行业：双环科技大幅亏损 涤纶长丝产品价差收窄  http://
cj.sina.com.cn/article/detail/3787884325/201865
sz000707  2017-03-29  09:07  双环科技：盈利尚可 食用盐受益专营放开  http://finance.sina.com.cn/
stock/t/2017-03-29/doc-ifycsukm4083144.shtml
```

市北高新（neg）：

sh600604市北高新.txt

```
sh600604  2017-04-27  04:50  上海市北高新股份有限公司2017第一季度报告  http://
finance.sina.com.cn/stock/t/2017-04-27/doc-ifyetstt3508306.shtml
sh600604  2017-04-26  18:52  证监会公布李健操纵市北高新案处罚决定书 罚没3750.95万  http://
finance.sina.com.cn/stock/t/2017-04-26/doc-ifyepsra5647786.shtml
sh600604  2017-04-26  18:46  证监会公布李健操纵市北高新案处罚决定书 罚没3750.95万元  http://
finance.sina.com.cn/stock/s/2017-04-26/doc-ifyetwsm0440893.shtml
sh600604  2017-04-26  18:28  中国证监会行政处罚决定书（李健）  http://finance.sina.com.cn/stock/
t/2017-04-26/doc-ifyetstt3453246.shtml
sh600604  2017-04-26  05:10  上海数据港股份有限公司2016年度报告摘要  http://finance.sina.com.cn/
stock/t/2017-04-26/doc-ifyepsch3377352.shtml
sh600604  2017-04-21  07:34  上海市北高新股份有限公司公告（系列）  http://finance.sina.com.cn/
stock/t/2017-04-21/doc-ifyepsra4947557.shtml
sh600604  2017-04-11  03:06  上海市北高新股份有限公司关于2016年度网上业绩说明会召开情况的公告  http://
finance.sina.com.cn/stock/t/2017-04-11/doc-ifyeayzu7464306.shtml
sh600604  2017-04-08  04:31  上海市北高新股份有限公司关于召开2016年度网上业绩说明会的预告公告  http://
finance.sina.com.cn/stock/t/2017-04-08/doc-ifyeayzu7179800.shtml
sh600604  2017-03-28  21:38  市北高新2016年度拟10转10  http://finance.sina.com.cn/stock/t/
2017-03-28/doc-ifycsukm4026440.shtml
sh600604  2017-03-28  21:18  市北高新：去年净利增16% 拟10转10派0.2  http://finance.sina.com.cn/
stock/t/2017-03-28/doc-ifycsukm4024088.shtml
sh600604  2017-03-20  06:48  股海导航 3月20日沪深股市公告提示  http://finance.sina.com.cn/stock/
s/2017-03-20/doc-ifycnpit2365275.shtml
sh600604  2017-03-17  19:43  市北高新：2799万元认购华建集团定增股份  http://finance.sina.com.cn/
stock/t/2017-03-17/doc-ifycnpvh4811697.shtml
sh600604  2017-03-17  17:42  3月17日上市公司晚间公告速递  http://finance.sina.com.cn/stock/s/
2017-03-17/doc-ifycnpiu8944088.shtml
```

新经典（pos）：

sh603096新经典.txt

```
sh603096  2017-08-14  06:41  股海导航 8月14日沪深股市公告提示  http://finance.sina.com.cn/stock/
s/2017-08-14/doc-ifyixtym3261967.shtml
sh603096  2017-08-11  17:09  新经典半年报预增50%到65%  http://finance.sina.com.cn/stock/t/
2017-08-11/doc-ifyixcaw4229157.shtml
sh603096  2017-06-21  16:54  6月21日上市公司晚间公告速递  http://finance.sina.com.cn/stock/s/
2017-06-21/doc-ifyhfnrf9438347.shtml
sh603096  2017-06-01  00:51  今年新股“掌门”60后是主流 九成实控人为自然人  http://
finance.sina.com.cn/roll/2017-06-01/doc-ifyfuvpm6908796.shtml
sh603096  2017-05-19  02:51  新经典文化股份有限公司公告（系列）  http://finance.sina.com.cn/
stock/t/2017-05-19/doc-ifyfkkmc9687646.shtml
sh603096  2017-05-18  20:27  新经典：拟10派3.6元  http://finance.sina.com.cn/stock/t/
2017-05-18/doc-ifyfkqwe0211281.shtml
sh603096  2017-05-18  16:41  5月18日上市公司晚间公告速递  http://finance.sina.com.cn/stock/s/
2017-05-18/doc-ifyfkqks4282003.shtml
sh603096  2017-05-10  07:30  新经典营收小步走净利增幅大 图书存货节节攀升  http://
finance.sina.com.cn/roll/2017-05-10/doc-ifyeycte9357864.shtml
sh603096  2017-04-29  07:05  新经典文化股份有限公司股票交易异常波动公告  http://
finance.sina.com.cn/stock/t/2017-04-29/doc-ifyetstt3846133.shtml
sh603096  2017-04-25  13:36  红杉中国投资“新经典”成2017首家登陆沪市主板的优质内容企业  http://
finance.sina.com.cn/stock/t/2017-04-25/doc-ifyepsra5518776.shtml
```

朴素贝叶斯（Mybayes.java）

输入：

bayes_train.txt

bayes_test.txt

输出：

bayes.out

```
sh603096新经典.txt 1 属于类别pos
sz002631德尔未来.txt 2 属于类别pos
sh600137浪莎股份.txt 3 属于类别pos
sh600510黑牡丹.txt 4 属于类别pos
sz300588熙菱信息.txt 5 属于类别pos
sh600057象屿股份.txt 6 属于类别pos
sh603167渤海轮渡.txt 7 属于类别pos
sz000786北新建材.txt 8 属于类别pos
sh603569长久物流.txt 9 属于类别pos
sh600035楚天高速.txt 10 属于类别pos
sz000707双环科技.txt 11 属于类别pos
sh600790轻纺城.txt 12 属于类别pos
sz000656金科股份.txt 13 属于类别pos
sz002174游族网络.txt 14 属于类别pos
sz002026山东威达.txt 15 属于类别neg
sz300405科隆股份.txt 16 属于类别pos
sz002790瑞尔特.txt 17 属于类别neg
sh600604市北高新.txt 18 属于类别neg
sz002846英联股份.txt 19 属于类别pos
sh600807天业股份.txt 20 属于类别pos
sh600683京投发展.txt 21 属于类别pos
sz300513恒泰实达.txt 22 属于类别pos
sz000701厦门信达.txt 23 属于类别neg
sz002337赛象科技.txt 24 属于类别pos
```

结果分析：

knn和bayes不完全相同，但有重叠。可能是因为一个股票里的信息既有正面的，又有负面的。分类方法不同导致结果不同。如果两种方法的结果都是neg，说明这只股票确实比较走势比较差。反之亦然。

4.不足&改进

一、特征词库的选择：一开始用来chi_words，发现效果不好，后来直接用来wordcount+chi_words的方式。
改进：信息增益
二、希望尝试一下其他的分类方法。

参考：
http://blog.csdn.net/wustzbq0713/article/details/46004875
http://blog.csdn.net/huludan/article/details/51674524