

基于情感词典的情感分类

使用文档

程序运行前需做好如下准备： 1.本程序是用python3.6编写，首先确保有python3环境。 2.句法分析的工具是使用哈工大ltp工具包，使用前请先安装，使用pip安装：

```
$pip install pyltp
```

其分词用的模型已保存在文件夹下，文件名为ltpdata3.4.0

程序配置

本实验采用挂载情感词典的方法，情感词典分为积极和消极，保存在dictionary文件夹下。

实体表保存在entity文件夹下的object文件。

程序信息

程序存储在代码文件夹下，文件名为demo_temp

输入文件格式

本方法用python编写，为便于操作的简单方便，参考前期提供的语料库格式，输入文档格式选用.csv格式，且需处理的文字内容为列名为content，样例如下：

content
' 变速箱的迅猛发展也与国家的政策支持关系密切。从《汽车产业发展政策》到《汽车产业调整和振兴规划》，在国家高度重视变速箱等关键零部件发展的前提下，技术含量较高的自动变速箱自然就成了企业竞逐的焦点。

样例文件在alldata文件夹中提供参考

输入参数

程序运行时，需提供如下几个参数：

- 1. 输入文件路径 （该路径下存有待处理的文件，文件格式上节已给出。）
- 2. 输出路径 （该路径存储程序运行完的结果）

输出结果

输出结果保存在输出路径下，名为`result`的文件名，文件夹下存储每篇文章的结果。

每篇文章的结果由三个文件组成，均为csv文件，分别是



entity_extraction



SentimentClassification



senWords_extraction

entity_extraction文件

该文件存放实体抽取结果,其结果样例如下：

entity_extraction				
sentence	entity	SenWords	pairs	polarity
0 而在第二季度时，郑素贞也是持有大恒科技、文峰股份两只个股，但其所持流通市值接近60亿元	大恒科技			
1 而在第二季度时，郑素贞也是持有大恒科技、文峰股份两只个股，但其所持流通市值接近60亿元	文峰股份			
2 郑素贞目前是大恒科技控股股东及实际控制人，其此前以受让股份的方式入主大恒科技	大恒科技			
3 不过在第三季度时，葛卫东退出了中国平安的前十大流通股东，仅现身平安银行第五大流通股东，持有2.35亿股，持股数量较上期略有减少，持股市值为24亿元	中国平安			
4 不过在第三季度时，葛卫东退出了中国平安的前十大流通股东，仅现身平安银行第五大流通股东，持有2.35亿股，持股数量较上期略有减少，持股市值为24亿元	平安银行			
5 不过，需要说明的是，葛卫东可能仍然持有中国平安的股份，但因为其三季度持股数量达不到进入前十大流通股东的“门槛”，所以无法计算其这一部分市值	中国平安			
6 葛卫东二季度持有近8000万股中国平安，为第七大流通股东，但中国平安三季度前十大流通股东的进入“门槛”提高至2.18亿股	中国平安			
7 何雪萍为A股市场上较为活跃的牛散，从持股情况分析，其三季度的布局中较为偏重医药股，有千山药机、紫鑫药业、康芝药业等三只个股，按持股市值来看，第一大重仓股为千山药机，持有流通市值超过5亿元，占其全部持	千山药机			
8 何雪萍为A股市场上较为活跃的牛散，从持股情况分析，其三季度的布局中较为偏重医药股，有千山药机、紫鑫药业、康芝药业等三只个股，按持股市值来看，第一大重仓股为千山药机，持有流通市值超过5亿元，占其全部持	紫鑫药业			
9 何雪萍为A股市场上较为活跃的牛散，从持股情况分析，其三季度的布局中较为偏重医药股，有千山药机、紫鑫药业、康芝药业等三只个股，按持股市值来看，第一大重仓股为千山药机，持有流通市值超过5亿元，占其全部持	康芝药业			
10 从持股时间来分析，何雪萍偏向于中长期持股，如千山药机为其去年第二季度进驻，目前已持有六个季度	千山药机			
11 黄木顺第三季度仍在坚守泛海控股，持有1.30亿股，较上期略有减少，截至三季度末持股市值约为12亿元	泛海控股			
12 黄木顺喜欢投资地产股，且偏向长期持股，资料显示，其自2006年起就现身泛海控股的前十大流通股东，随后长线持股，总体上不断加仓	泛海控股			
13 加上泛海控股，黄木顺近10年来共现身10只个股的前十大流通股东，其中，泛海控股、、均为地产股，天健集团、均为与地产相关度较大的建筑行业个股	泛海控股			
14 加上泛海控股，黄木顺近10年来共现身10只个股的前十大流通股东，其中，泛海控股、、均为地产股，天健集团、均为与地产相关度较大的建筑行业个股	天健集团			
15 在第二季度时，黄木顺也是仅现身泛海控股的前十大流通股东，不过持股市值超过20亿元	泛海控股			
16 今年2月份，沙钢股份公告，沙钢集团拟通过协议转让方式转让公司股份8.69亿股，占总股本的55.12％	沙钢股份			
17 沙钢股份股价此后一路上涨，截至6月份累计涨幅超过4倍，随后于6月25日起停牌，而9位接盘的自然人股东财富也暴增，停牌前的持股市值均超过26亿元	沙钢股份			

表格由sentence，entity，SenWords，pairs，polarity组成。由于是实体抽取，故后三列为空。

senWords— —extraction文件

该文件存放情感词抽取结果，其结果样例如下：

sentence	entity	SenWords	pairs	polarity
0 而在第二季度时，郑素贞也是持有大恒科技、文峰股份两只个股，但其所持流通市值接近60亿元	大恒科技			0
1 而在第二季度时，郑素贞也是持有大恒科技、文峰股份两只个股，但其所持流通市值接近60亿元	文峰股份			0
2 郑素贞目前是大恒科技控股股东及实际控制人，其此前以受让股份的方式入主大恒科技	大恒科技			0
3 不过在第三季度时，葛卫东退出了中国平安的前十大流通股东，仅现身平安银行第五大流通股东，持有2.35亿股，持股数量较上期略有减少，持股市值为24亿元	中国平安	['退出','仅']		
4 不过在第三季度时，葛卫东退出了中国平安的前十大流通股东，仅现身平安银行第五大流通股东，持有2.35亿股，持股数量较上期略有减少，持股市值为24亿元	平安银行	['退出','仅']		
5 不过，需要说明的是，葛卫东可能仍然持有中国平安的股份，但因为其三季度持股数量达不到进入前十大流通股东的“门槛”，所以无法计算其这一部分市值	中国平安	['无法']		
6 葛卫东二季度持有近8000万股中国平安，为第七大流通股东，但中国平安三季度前十大流通股东的进入“门槛”提高至2.18亿股	中国平安	['提高']		
7 何雪萍为A股市场上较为活跃的牛散，从持股情况分析，其三季度的布局中较为偏重医药股，有千山药机、紫鑫药业、康芝药业等三只个股，按持股市值来看，第一大重仓股为千山药机，持有流通市值超过5亿元，占其全部持	千山药机			0
8 何雪萍为A股市场上较为活跃的牛散，从持股情况分析，其三季度的布局中较为偏重医药股，有千山药机、紫鑫药业、康芝药业等三只个股，按持股市值来看，第一大重仓股为千山药机，持有流通市值超过5亿元，占其全部持	紫鑫药业			0
9 何雪萍为A股市场上较为活跃的牛散，从持股情况分析，其三季度的布局中较为偏重医药股，有千山药机、紫鑫药业、康芝药业等三只个股，按持股市值来看，第一大重仓股为千山药机，持有流通市值超过5亿元，占其全部持	康芝药业			0
10 从持股时间来分析，何雪萍偏向于中长期持股，如千山药机为其去年第二季度进驻，目前已持有六个季度	千山药机			0
11 黄木顺第三季度仍在坚守泛海控股，持有1.30亿股，较上期略有减少，截至三季度末持股市值约为12亿元	泛海控股			0
12 黄木顺喜欢投资地产股，且偏向长期持股，资料显示，其自2006年起就现身泛海控股的前十大流通股东，随后长线持股，总体上不断加仓	泛海控股			0
13 加上泛海控股，黄木顺近10年来共现身10只个股的前十大流通股东，其中，泛海控股、、均为地产股，天健集团、均为与地产相关度较大的建筑行业个股	泛海控股			0
14 加上泛海控股，黄木顺近10年来共现身10只个股的前十大流通股东，其中，泛海控股、、均为地产股，天健集团、均为与地产相关度较大的建筑行业个股	天健集团			0
15 在第二季度时，黄木顺也是仅现身泛海控股的前十大流通股东，不过持股市值超过20亿元	泛海控股	['仅']		
16 今年2月份，沙钢股份公告，沙钢集团拟通过协议转让方式转让公司股份8.69亿股，占总股本的55.12%	沙钢股份	['通过']		
17 沙钢股份股价此后一路上涨，截至6月份累计涨幅超过4倍，随后于6月25日起停牌，而9位接盘的自然人股东财富也暴增，停牌前的持股市值均超过26亿元	沙钢股份	['上涨','涨幅']		

未抽出情感词的句子，极性设为0，表示中性。

SentimentClassification文件

该文件存放情感分类结果，其结果样例如下：

SentimentClassific

	entity	score
0	大恒科技	0
1	文峰股份	0
2	中国平安	0
3	平安银行	0
4	千山药机	0
5	紫鑫药业	0
6	康芝药业	0
7	泛海控股	0
8	天健集团	0
9	沙钢股份	2

分数大于0的表示积极，小于0的表示消极，等于0表示中性。