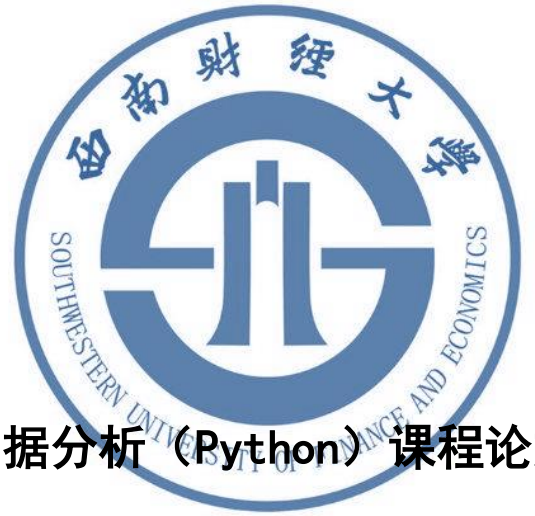


数据分析（Python）课程论文

中国股灾前后股民情感变化分析与预测

专业	金融工程
姓名	郑宇浩
学号	41621101
指导教师	郑海超

中国股灾前后股民情感变化分析与预测

摘要：本文为研究中国股灾前后股民的情感变化情况，拟通过对 2015 年股灾前后一年内股民所有微博内容进行情感分析，获取微博情感得分时序。研究首先运用随机森林算法计算出多个金融检索词的百度指数变化量对沪深 300 指数涨跌幅的特征重要性，结合二者之间的时差相关系数，选取出了“股票市场”、“股票代码”和“股票交易”三个对中国股市变化具有强解释力的金融词汇作为微博检索词以获取全部股民微博内容。利用朴素贝叶斯分类器对带有“买入”与“卖出”标签的多份个股研报进行文本分析并搭建分类模型，模型的测试集精准度高达 91.3%，依此，研究得到了金融领域专用的情感正负向词袋。在微博中搜索上述三个金融关键词，再利用金融情感词袋对所有的微博内容进行情感分析，最终获得了中国股灾前后一年内股民的情感得分时序。将其与沪深 300 指数周收盘价进行相关性分析，研究发现股民情感变化对中国股市涨跌幅具有严重的滞后性与弱相关性，故依此构建的交易策略在理论上是无效的。但研究发现了股民微博情感在股灾发生前会出现异常高涨且临近时有略微下滑的特点，故可用于股灾预测。

关键词：百度指数 随机森林 金融词袋 贝叶斯分类器 情感分析 股灾预测

一、引 言

伴随着 2017 年诺贝尔经济学奖再一次出人意料地颁发给了行为金融领域的学者理查德•H•泰勒，我们对于全球股票市场的动荡研究已经不能再简单地忽略人们的心理因素影响了。在国外，人们的情感隐藏着他们对股市的看空与看涨预期，并指导着他们采取的交易策略。所以，只需要提前了解到大部分股民对股市的态度，就可以大致地预测出他们未来会采取的交易策略，在此之前，先于大多数人而进行股票的买卖，不失为一种有效的择时手段。此外，提前捕捉到人们情感波动的异常值，甚至有可能可以推断出未来发生股灾的可能性。

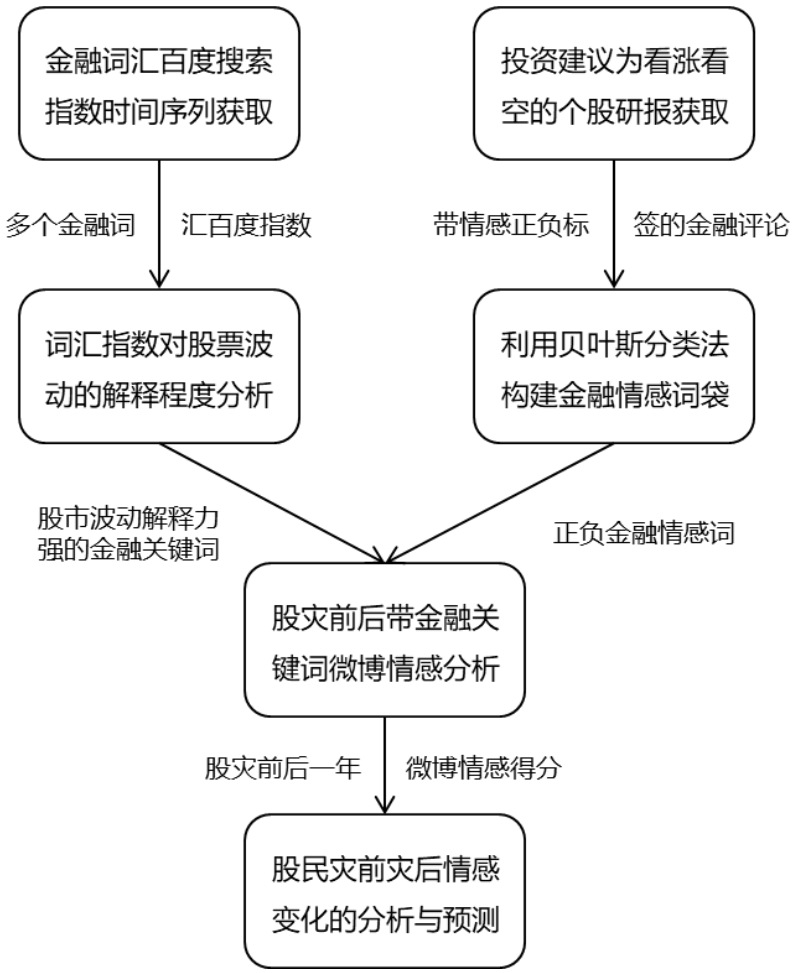
早在 2013 年，华威大学的行为金融学教授 Tobias Preis 就对这一猜想进行了实证分析。他认为，人们在对股市看好的情况下会频繁地访问与金融有关的网页内容，反之，当人们对未来的行情失去信心的时候则会减少对财经的关注度。基于这一假设，Tobias Preis 收集了大量金融词汇在谷歌中的搜索次数，依据其变化幅度，构造了相应的交易策略，并发现结果显著地优于随机策略。这一结果，有效地论证了情感分析在交易策略中的实用性。

反观国内，倘若要复刻这一交易策略，我们必须要先解决两个问题。一是确定大部分股民们的情感状态是否真的会主导着他们买卖股票，即是说，我们要先思考，情感这一影响因素是否应该被作为判断股票涨跌的依据。二是找到与金融关键词的搜索指数相类似的，能够反映大部分中国股民情感变化的参考指标。曾有创新新公司专门挖掘搜索引擎中各类关键词的搜索次数，研究其与国内股票价格走势之间的关系，依此来构造交易策略寻求套利机会。然而不幸的是，该公司于成立不久后就以失败告终。这一案例警示着我们，或许搜索指数这一指标，在中国股票市场上并不是反映股民情感变化的可靠依据。究其原因，有学者曾提到过，中国股票市场上的大部分散户的情感变化与国外的投资者恰好相反，即人们在买入股票后才会对股市抱有信心，而在购买之前却很少关注市场行情。这就意味着，追踪某些金融关键词的搜索指数变化趋势可能不能帮助投资者超前地预测未来股价的走向，甚至还有可能出现滞后性。

为了很好地解决上述两个问题，本课程论文将围绕 2015 年中国股灾发生前后，股民们

的情感变化展开研究，探讨在行情大好和股市崩溃时人们情感的不同之处，并研究在股灾发生之前是否有足够多的负向情感指标可以用来预示未来不好情形的发生，从而作为较为可靠的评判标准，在下一次股灾到来之前能够提前通过人们情感变化的异常值大概地判断出股灾的来临。具体的研究流程如下图 1：

【图 1】研究流程



二、文献综述

2013 年，华威大学的 Tobias Preis 在《Scientific Reports》上发表的《Quantifying Trading Behavior in Financial Markets Using Google Trends》一文中提出了搜索引擎中的金融词汇搜索次数变化可以反映未来股价波动的猜想[1]，据此，他利用部分金融词汇在谷歌的搜索次数变化数据构造了交易策略并成功获利，对该猜想给予了充分的论证。反观国内学者，也有不少人利用百度指数来研究股票的价格走向，如安徽新华学院财金学院的谢明柱就曾在《投资者有限关注与股票收益关系研究——基于百度指数的实证分析》一文中提到：“投资者关注和股票的市场表现之间是双向引导的关系,投资者在周末对股票的关注对下周一股票价格跳跃和收益率的跳跃都有显著的正向影响。”[2] 此外，北京工业大学经济与管理学院的王耀君等人也在《基于网络搜索指数的股票市场微观结构特征》中对中国近三千只股票在 2013 年到 2016 年间的收益率表现进行研究，指出了：“股票的关注度增加,会提高股票市场交易的流动性和收益率。”[3] 浙江大学的庞云枫更是在其研究生论文中利用互联网的关注度构造了

股票成交量的预测模型[4]。从对前人的研究中可以看出，中国股票市场的流动性、收益率、成交量都会相应地超前反映在互联网的关注度上，依此来构造可盈利的交易策略似乎是有可能实现的。

但近几年来，互联网关注度与股价变动的高度相关这一理论在中国被广泛地运用在交易策略中，利用搜索指数企图先于市场变化采取行动以取得套利的机会已经变得少之又少了，这也很好地说明了前文提到的某家创新型公司倒闭的原因。即便该策略的广泛运用导致了套利机会的消失，某些与金融相关的关键词的搜索指数的预测能力会随之失效，但他们与股价变动之间的强依赖性仍然是有实用价值的。我们可以利用微博等社交平台，分析一段时间内人们发表的含有这些与股价变动具有强依赖性的关键词的相关评论，来判断出该时间段内，大众所呈现出来的情感正负向，由此来判断大多数股民对未来股市的看空与看涨情况。这一分析思路，成功地避免了去寻找能够替代搜索指数的情感反映指标，能够更好地论证情感分析在中国股票市场上的实用性。如清华大学电子系的肖婷就曾发表过一篇名为《一种基于股票情感分析的股市趋势预测方法》的论文，论述了基于对大量专业人士发表的股评进行情感分析搭建起的股票预测模型的有效性[5]。

二、研究方法

根据图 1 的研究流程所示，为了能够获取中国股民的微博，我们必须先做好两项准备工作：一是微博检索词的确定；二是金融情感词袋的构建。

针对工作一，本研究先利用百度指数的需求图谱，找到了与“股票市场”的搜索相关性最强的其余七个金融词汇。通过获取上述八个金融词汇在 2014 年 3 月 3 日至 2016 年 2 月 29 日之间的周百度指数，我们可以研究其与沪深 300 指数周收盘价的时差相关系数。据此，我们可以选取出具有提前相关性的金融词汇。除此之外，本研究还将沪深 300 指数周收盘价的涨跌幅作为类别变量，而上述八个金融词汇的周百度指数变化量则作为其特征变量，将二者代入到随机森林分类器中，求算出八个金融词汇各自的特征重要性。最终，我们选取出时差相关系数最高，且特征重要性得分排名前三的金融词汇：“股票市场”、“股票代码”和“股票交易”作为后续的微博检索词。

针对工作二，本研究从东方财富网上批量下载了多份个股研报，由于其标签中带有证券公司给投资者们的投资建议，分别为：“买入”、“增持”、“中性”、“减持”和“卖出”五大类别，我们可以将其作为各份研报的类别变量。利用 Python 的 jieba 包，我们可以对各份研报逐一进行切词，再删去无意义的词汇，利用 Python 的 sklearn 文本特征提取器，得到未分类的金融词库。最后，根据研报自带的分类标签，我们可以用朴素贝叶斯分类模型将上述过程中所得到的未分类的金融词库划分为“看涨”和“看跌”两类金融情感词袋，作为后续微博情感评分的依据。

完成了上述两项准备工作之后，我们就可以开始微博评论的获取与评分了。先在微博中搜索工作一中得到的三个金融关键词：“股票市场”、“股票代码”和“股票交易”，设置检索日期为 2014 年 3 月 3 日至 2016 年 2 月 29 日，但由于微博平台对同一 IP 地址的访问次数设有上限，每十分钟只能够抓取微博信息四十次，且在未登陆情况下，只允许访问评论的第一页。为了解决上述两个问题，本研究对代码进行了睡眠时间的设置，每访问微博平台四十次，程序就会自动休息十分钟，此外，我们缩短了每次检索的时间间隔，从一天改为了两小时，从而保证了每次检索的结果都能够在第一页中就显示完整，而不需要进行模拟登陆后自动翻页等复杂操作。至此，我们已经成功地解决了微博评论获取的所有问题，关于其情感评分模型，我们将在后续的模型构建章节中进行详细的介绍。

在得到了 2015 年股灾前后一年内的所有股民微博评论及其情感得分之后，本研究将其

与沪深 300 指数周收盘价进行对比，研究其趋势之间的相关性以及情感得分变化量对股价涨跌幅的预测能力。详细的研究内容与分析结果将会在第五章节中进行具体地阐述。

四、数据获取与模型构建

4.1 金融关键词获取

微博上的评论千千万万，但并不是每一条都能够反映股民们对股市的情感变化，我们必须通过检索与中国股市相关的金融关键词，获取带有股民个人情绪的微博。因此，金融关键词的选取就变得尤为重要了。有效的金融关键词，其百度搜索指数时间序列与中国股市的波动会有较强的相关性[6]，因此我们需要先获取多个金融词汇的百度指数时序，甄选出对股市波动具有强解释力的几个金融词汇作为主要的微博检索词。

通过查找“股票市场”的百度指数，根据百其提供的需求图谱，我们得到其他七个与“股票市场”检索相关性较高的金融词汇。由于百度指数无法通过免费下载获得，且通过观察，我们发现，关键词每一周的搜索指数是根据鼠标所指向的指数趋势变化图所在位置而动态加载浮现的，无法一次性完整抓取，故采用鼠标模拟移动与图像识别的方法，利用 Python 实现自动截图与指数识别，以构建多个金融词汇的百度指数周时序[7]。为削弱突发事件，如爆炸性新闻报道等对检索指数的影响，我们对原始指数周时序进行了四周的移动平均处理。随后调用 Python 的 tushare 包，导入 2014 年 3 月 3 日至 2016 年 2 月 29 日的沪深 300 指数数据，分析上述 8 个金融词汇的百度指数与股票周收盘价之间的相关性。通过公式 1 计算所得的时差相关系数如下表 1：

$$\rho_d = \frac{\sum_{t=1}^{n}\left(index_{t+d} - \overline{index}\right)\left(stock_t - \overline{stock}\right)}{\sum_{t=1}^{n}\left(index_{t+d} - \overline{index}\right)^2 \sum_{t=1}^{n}\left(stock_t - \overline{stock}\right)^2}, d = 0, \pm1, \pm2, \cdots, \pm D \qquad (1)$$

【表 1】时差相关系数

金融词汇	时差相关系数	提前期（周）
股票代码	0.81601037	1
股票交易	0.78669451	1
股票市场	0.75229532	1
股价	0.65106262	1
股票行情	0.62396327	1
行情	0.51228060	1
股市	0.42041746	1
交易	0.26713189	5

为了更好地分析出金融词汇对股票波动的解释程度强弱，我们对沪深 300 指数的每周波动率进行分类作为因变量，具体的分类方法如下表 2：

【表 2】随机森林类别变量

类别	-2	-1	0	1	2
波动率	≤-6%	≤-2%	≤2%	≤6%	＞6%

自变量则为金融词汇百度指数的周增量，我们将其作为股市波动的特征代入到随机森林分类

模型中，利用 Python 的 sklearn 分类包中的随机森林分类器，计算得到各个金融词汇的特征重要性。随机森林中某个特征重要性的具体计算方法如下[8]：

Ⅰ. 使用袋外数据计算对于随机森林中的每一颗决策树的袋外数据误差,记为 $ErrOBB_1$；

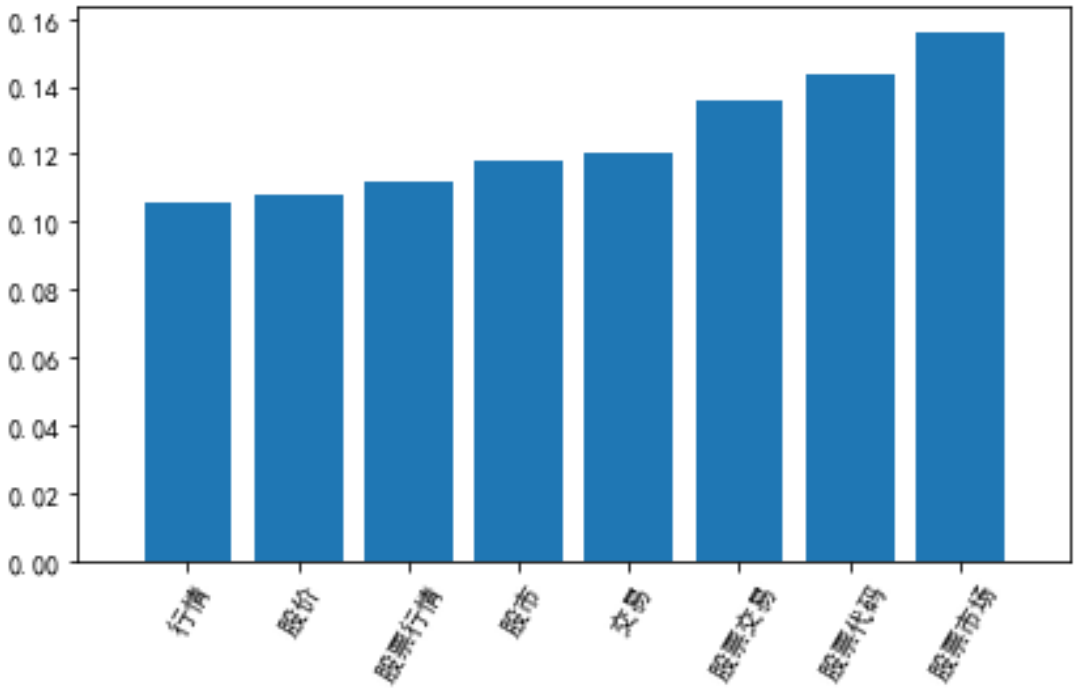
Ⅱ. 随机地对特征加入噪声干扰，再次计算袋外数据误差，记为 $ErrOBB_2$；

Ⅲ. 若对某个特征随机地加入噪声之后，袋外的准确率大幅度降低，则说明这个特征对样本的分类结果影响很大，即重要程度高，则有如下公式 2：

$$Feature\ Importance = \frac{\sum(ErrOBB_2 - ErrOBB_1)}{N_{tree}} \tag{2}$$

将各个金融词汇的特征重要性绘制为柱状图如下图 2：

【图 2】特征重要性



结合各个金融词汇的时差相关系数及其对股票波动率的特征重要程度分析，我们可知，关键词“股票代码”、“股票交易”及“股票市场”的百度指数对股价都具有超前一周预测的能力，相关性高达 0.750，且其增量对股票波动分类的特征重要性都在 0.135 以上，即对股市波动的解释程度较强，故我们选取“股票市场”、“股票代码”与“股票交易”作为第三步中微博检索的三个关键词。

4.2 金融情感词袋构建

在分析微博的情感正负向之前，我们必须要有与金融相关的情感词袋。清华大学的李军与台湾大学的自然语言处理实验室都曾做过中文的情感词分类收集，但其并不适用于分析股民对未来股市看法的情绪变化，因为一般的情感词袋并未纳入一些金融市场专有的术语，如负向词“杀跌”等等，这就有可能导致我们无法正确地归类股民的情感。事实上，还需要注意的是，在本项研究中，我们并不需要准确地知道股民们的情感是好是坏，只需通过特定的

金融情感词分析出股民对未来股市是看涨还是看空即可。由于中国股市不存在做空机制，所以我们通常把看空态度视为负向情感，也为了后续说明的方便，在本研究中，股民对股市的看涨看空即视为是其情感态度的正负向。

为了构建看涨看空，即金融情感正负向词袋，我们必须获取带有正负向标签的金融相关评论进行切词与分类，而证券公司为个股撰写的研报正好满足了我们的需求，每一篇研报的末尾都有研究员对个股的投资建议，分为“买入”、“增持”、“中性”、“减持”和“卖出”，可作为分类模型的类别变量。基于朴素贝叶斯算法的分类模型构建具体过程如下[9]：

Ⅰ. 从东方财富网上批量下载个股研报，对导入的多份研报进行文本清洗，删除标点及特殊符号等之后，利用 Python 的 jieba 包进行切词，同时留下一半的数据量作为测试集；

Ⅱ. 引用哈尔滨工业大学创建的中文停词库，删去无实际意义的评论词，再利用 Python 的 sklearn 文本特征提取器，得到未分类的金融词库，部分结果如下表 3：

【表 3】未分类金融词库的数目向量矩阵

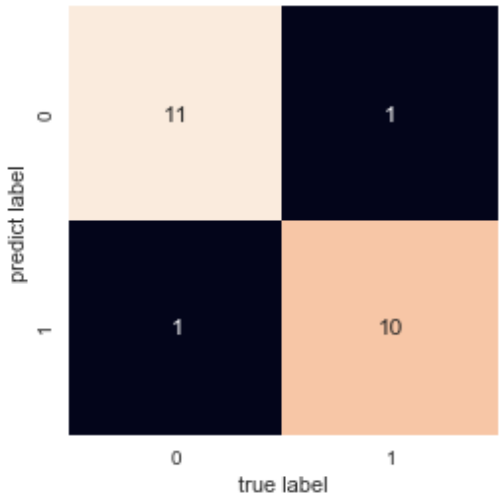
词汇 / 文本	驱动	驾驶	高于	高位	高端	高达	高速	龙头
1	0	0	0	0	0	1	0	1
2	0	0	0	0	0	0	0	1
3	0	4	0	0	0	0	1	0
4	0	0	0	0	0	2	0	0
5	0	0	0	0	0	0	0	0

Ⅲ. 利用朴素贝叶斯算法计算得到每一个词汇的正负向情感分类概率，正向概率值大于负向时，将该词标记为金融正向情感词，即纳入看涨态度的词袋中。朴素贝叶斯分类概率算法如下公式 3：

$$\tilde{P}(word|attitude) = \frac{count(word|attitude) + 1}{\sum_{word \in test} count(word|attitude) + |test|} \tag{3}$$

Ⅳ. 对所建立的分类模型进行样本集内交叉验证，其准确度得分为 0.821。再代入测试集进行检验，真实性得分为 0.913，得到混沌矩阵如下图 3：

【图 3】朴素贝叶斯分类模型混沌矩阵



根据分类模型的准确度得分与测试集的混沌矩阵可知，我们所构建的金融情感词袋对证券机构的投资建议预测效果非常良好，可作为第四步中微博评论的情感分析依据。

4.3 微博评论获取

依次将第一步中所得到的三个金融关键词：“股票市场”、“股票代码”和“股票交易”作为微博检索词，搜索 2014 年 3 月 3 日至 2016 年 2 月 29 日每日 7 时至 24 时的微博，按每两小时为一个时间段进行划分，不光可研究股民在股灾前后的情感变化，还为后续研究日内股民情绪波动提供了数据基础。

为获取所有的微博信息，我们必须登陆才能够翻页查阅完整的检索结果。但由于平台对用户的账号和密码加密方式与以往不同，故无法利用 Python 进行模拟登陆，于是会出现信息不完整的问题。本研究将每次检索的时间段从一天缩短为两小时，成功使得所有结果能够在首页显示完整，即便牺牲了大量时间去多次访问微博，但却保证了信息的完整性，将误差降至最小。同时，微博设置每十分钟同一 IP 地址只能访问平台四十次，故设置程序的睡眠时间为十分钟，使其自动循环访问微博获取评论，缩短总的数据获取时间[10]。

4.4 微博情感评分

利用第二步得到的金融情感词袋，我们可以构建微博评论情感分析模型[11]如下公式 4：

$$score^{*}=\begin{cases}1 & if\ count(word\ in\ positive)>0\ and\ count(word\ in\ negative)=0\\ \lambda=\dfrac{count(word\ in\ positive)}{count(word\ in\ negative)}=\begin{cases}1\ if\ \lambda>1.5\\ 0\ else\\ -1\ if\ \dfrac{1}{\lambda}>1.5\end{cases} & else\\ -1 & if\ count(word\ in\ negative)>0\ and\ count(word\ in\ positive)=0\end{cases}\quad (4)pt.1$$

$$score=\begin{cases}score^{*} & if\ count(word\ in\ deny)\ mod\ 2=0\\ -1\times score^{*} & if\ count(word\ in\ deny)\ mod\ 2=1\end{cases}\quad (4)pt.2$$

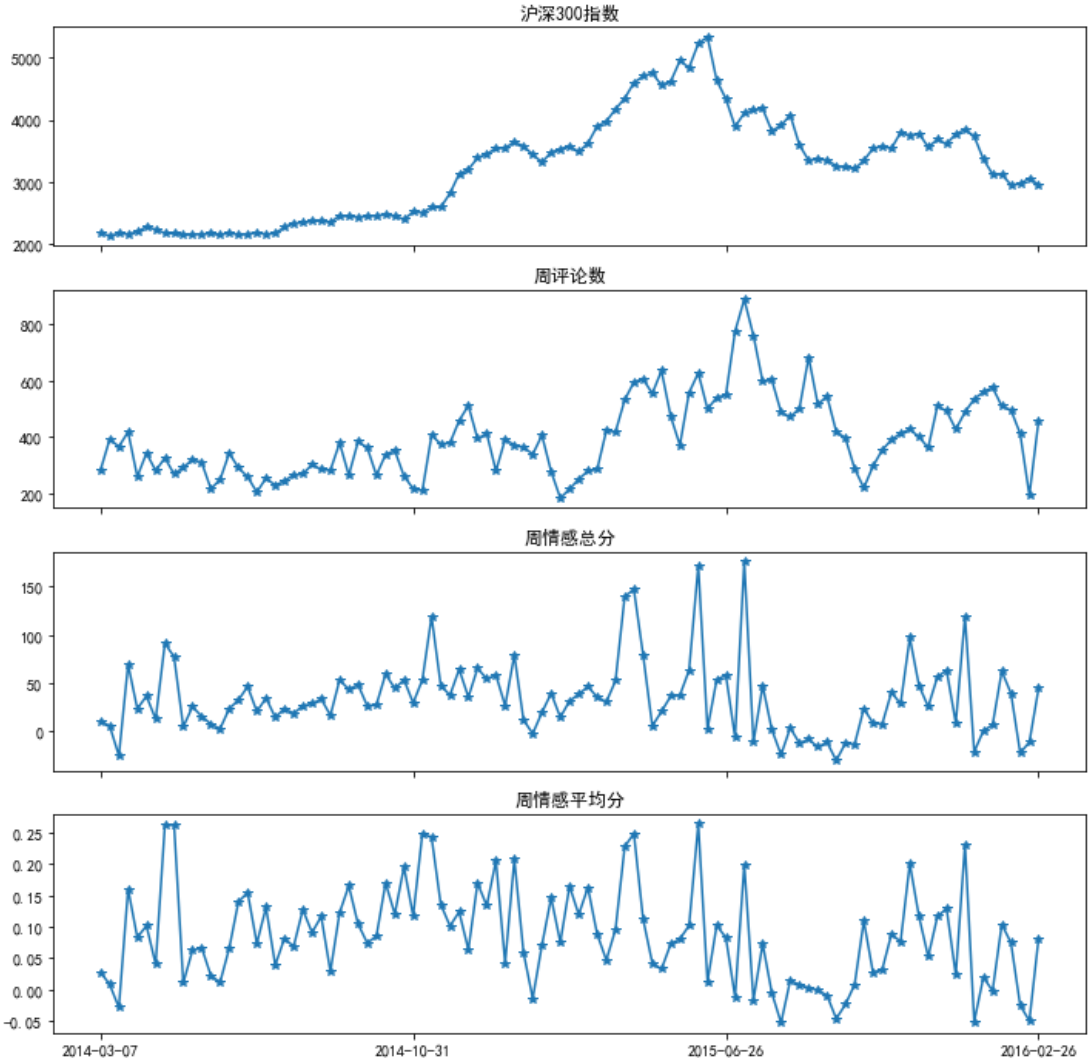
在将原始数据带入模型之前，我们必须对评论进行去除标点符号、停词之类的无效信息等数据清洗工作。同时引入台湾大学自然语言处理实验室创建的否定词库，注意到当一个评论中否定词为奇数个时，该条微博的情感得分应该取相反数。

五、情感分析与预测

5.1 微博情感变化趋势分析

在得到股灾前后一年内股民们的微博评论情感得分之后，我们可以绘制得分的趋势变化图[12]，与沪深 300 指数周收盘价的 K 线图进行对比，直观地看出二者之间是否具有趋势相关性，对比图如下图 4（下图仅展示检索词为“股票市场”的微博情感得分结果，检索词为“股票代码”与“股票交易”的情感得分趋势图见文末附录处）：

【图 4】情感得分与沪深 300 指数周收盘价趋势对比图



我们可以直观地看到，事实上即便本研究绕过了使用百度搜索指数这一类代表股民情感变化的指标，直接探索股民微博中透露出的真实情感，其对股价涨跌的反映仍然存在一定的滞后性，甚至呈现出较弱的相关性，这可从如下表 4 中看出：

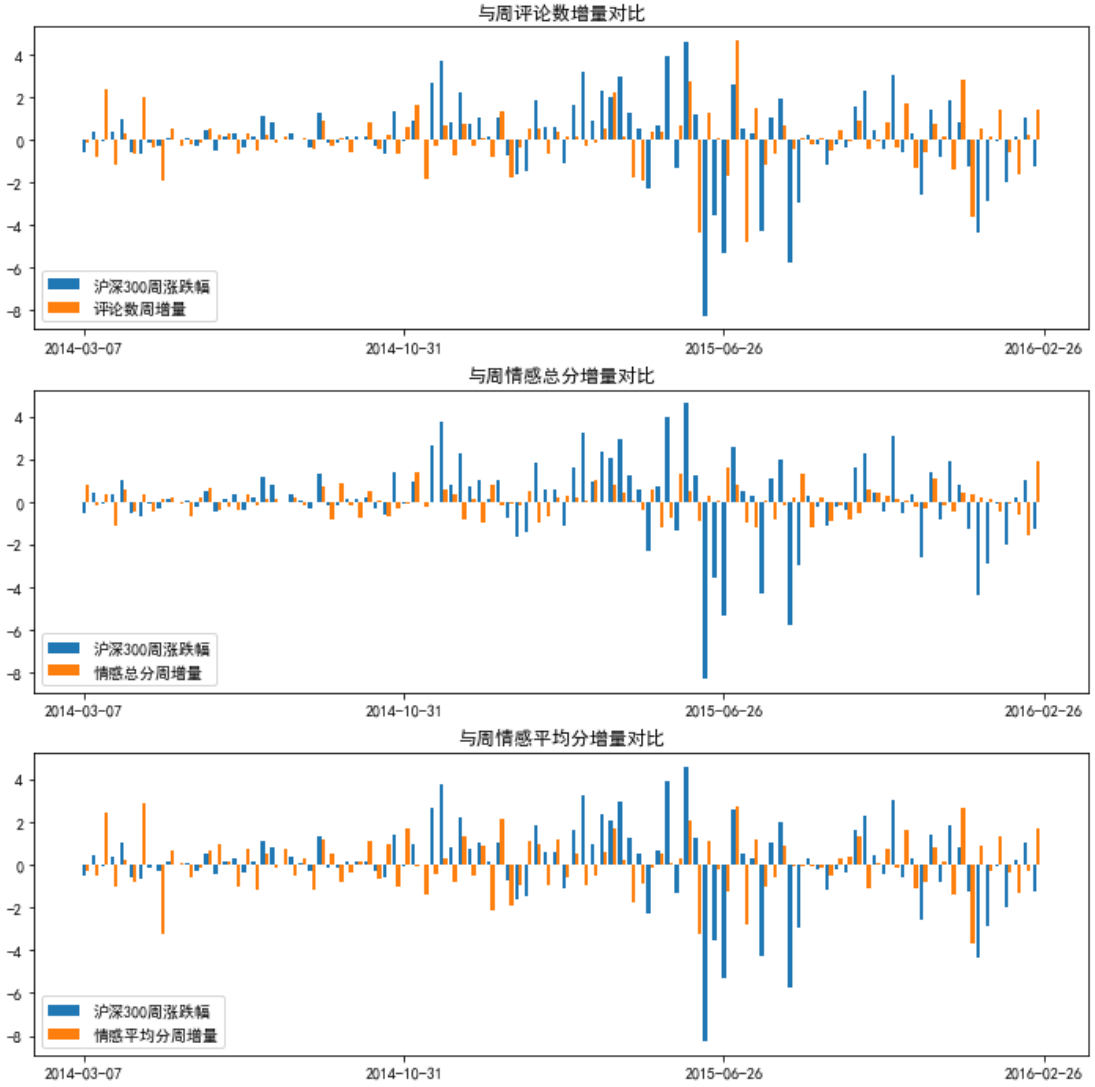
【表 4】情感得分与沪深 300 指数周收盘价相关性

指标 / 检索词	(周总评论数,提前期)	(周情感总分,提前期)	(周情感均分,提前期)
股票市场	(0.61607098,1)	(0.24708212,1)	(0.01911257,4)
股票代码	(0.04812777,1)	(-0.05882334,4)	(0.04265431,3)
股票交易	(0.65162103,3)	(0.62955772,4)	(0.49291898,2)

为了能够更进一步地探究股民微博评论情感指数的有效性，我们从另一个角度入手，研究股价的周涨跌幅与股民情感周变化量之间的关系，可以通过绘制得分增量柱状图进行比较。在计算增量之前，我们先对股票及情感得分运用公式 5 进行标准化处理以去除量纲，方便作图对比，最终结果如下图 5（下图仅展示检索词为“股票市场”的微博情感得分结果，检索词为“股票代码”与“股票交易”的情感得分趋势图见文末附录处）：

$$normalize(x_i) = \frac{x_i - mean(X)}{std(X)} \tag{5}$$

【图 5】情感得分周增量与沪深 300 指数周涨跌幅柱状对比图



如我们先前所预期的，股民们的情感变化确实与股票的涨跌并无太大关系，即便中国股市在 2015 年的大崩盘让很多股民的情绪在一周内出现了明显的大起大落，但想要依此来构造长期的套利策略恐怕并不非常明智，我们可以从如下表 5 中看出，利用股民微博情感得分来预测股市涨跌的正确性并不可观。情感得分的预测准确度算法如下公式 6：

$$accuracy = \frac{count(\Delta score_t \times \Delta stock_{t+1} \geq 0)}{T_{backtesting}} \tag{6}$$

【表 5】情感得分预测准确度

准确度 / 检索词	周总评论数	周情感总分	周情感均分
股票市场	0.5196	0.5686	0.4902
股票代码	0.4804	0.5882	0.5098
股票交易	0.5588	0.5490	0.5588

对于胜率只有一半的交易策略来讲，我们不应当运用它来寻找套利机会，这无异于是在赌博，同时它也在一定程度上解释了为什么曾有利用搜索指数构建交易策略的公司最终会走向倒闭的悲剧。

5.2 利用微博情感变化预测中国股灾发生的可行性分析

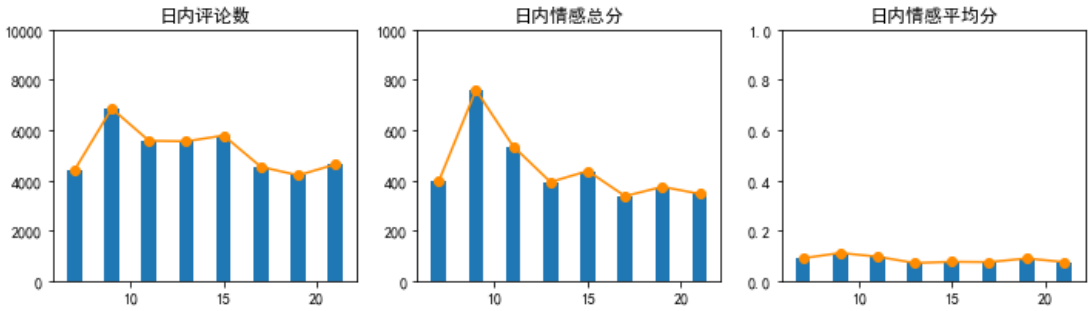
虽然我们不能用情感得分来谋取利益，但股民们情绪的大起大落能够反映在他们的微博中，倒是给了我们提前大概预知股灾发生的能力。由上图的分析可知，不论是情感得分还是微博评论数，都在 2015 年股灾发生前后出现了剧烈的波动，这说明了股市的变动还是能够在一定程度上反映在股民的情绪变化当中，这也符合在行为金融学领域上的解释——自我归因偏差：大部分股民们在获取了很可观的超额收益后，往往会将这部分成果归结于自己的能力，而忽视了整个大盘的涨跌情况实际上是一直在上升且过度虚高的。事实上这并不是股民们自己的能力，而只是他们随波逐流获取到的短期利益，但人们的自我归因偏差就会导致多数人在微博上分享自己所谓的经验之道。除此之外，媒体在 2015 年上半年快结束时大肆地宣传中国牛市地到来，广告商们也疯狂地向散户推荐各种股票，整个微博平台充斥着各种高转发量的股市分析文章……种种网络上的正向言论理所当然地将股民们微博的情感得分拉到了意想不到的至高点，微博的热度也随之上升。同理，在中国股市崩盘后，负面情绪充斥着整个微博平台也是意料之中的。

如果我们能够捕捉到微博平台上，股民们异常高涨的情绪变化，或许就能够提前预判股市危机的到来。而这一理论的依据，就来源于行为金融学对人们过度自信的合理分析。我们知道，大多数股民在获利之后会由于自我归因偏差而产生异于常态的自信，而这一心理会让大多数人疯狂买入股票追涨而忘记了止盈的重要性。中国股市上充斥着大量的散户，他们集体的这种由于羊群效应造成的大规模买入会迷惑很多较为理性的投资者甚至是资深的基金经理，使得股价持续上涨，中国股市的天花板在短期内无法被迅速地推断得出，即便有可能次日就是股市的大崩盘。但与之不同的是，微博的情感得分不会一直处于持续上涨的趋势，在中国股市大涨一段时间后，依旧会有理性的投资者站出来对宏观环境进行仔细的分析。他们在微博上发布较为负面的警告讯息，使得微博评论的情感得分呈现出略微的下滑趋势，加上微博转发的放大效应，部分股民们的负向情绪将能够被及时得捕捉到。因此，正如上图所分析的，当我们发现股民的微博情感得分在出现异常高涨之后又有着些许地回落，这很有可能就是股市过分虚高的前兆，应当采取恰当的行动，谨慎处理以减少可能发生的股灾会带来的损失。

5.3 股民日内情感变化分析

在分析完微博情感得分的股灾预测可行性之后，本研究还收集了两年来微博股民一天中在不同时间段里的情感波动情况，如下图 6（下图仅展示检索词为“股票市场”的微博情感得分结果，检索词为“股票代码”与“股票交易”的情感得分趋势图见文末附录处）：

【图 6】微博股民日内情感变化图



可以看出，在每日早晨九点开盘前，股民们普遍情绪高涨，这也与我们对大部分股民的心理状况预期相一致：一般采取短期交易策略的投资者，都会选择在每日早晨进行股票的买卖操作，将持有隔夜的股票进行卖空套利，而这一过程通常是令人感到乐观的。但随着一天中交易市场的变化，股民们的情感得分逐渐下滑，特别是在午间休市的 2 个小时里，股民们的情感持续低迷。由此可以看出，大多数股民对自己在一天的交易过程中的表现还是不太满意的。或者是说，事实上大部分人还是无法准确地预测股市的走向，才会导致他们进行了错误的操作，使得自己心情低落。最后下午休市的时候，人们还是无法扭转自己的负向情绪，这一悲观心态一直持续到了夜间。由此看来，行为金融学当中的损失厌恶理论在中国股民的日内交易表现中得到了充分的验证。但总的来说，两年间中国股民的微博情感平均得分都处于 0.1 左右，这说明了事实上人们在交易过程中还是相对理性的，能够尽量减少自己的情绪化波动，以免带来股票交易操作上的失误。

六、模型缺陷的合理化解释与改进方法

很明显，本模型存在的最大缺陷就是，它使得分析得到的中国股民情感得分不具有普适性，当我们将其运用于交易策略的构建时，根据其所计算得到的情感变化并不能超前地反映股价未来的波动，导致策略的胜率仅有 50%左右。虽然这在很大程度上是因为原本股民们的情感对股市波动就具有滞后性，但模型仅采用微博股民代表所有人所带来的误差也可能是相当巨大的。未来可以改进的点在于寻找除微博以为更能够反映大量股民情绪变化的社交平台，如东方财富网的股吧、百度贴吧等等。

同时，本项研究基于“具有代表性的微博金融检索词的百度指数与股票波动具有强相关性”这一假设，仅考虑了“股票市场”、“股票代码”和“股票交易”这三个金融关键词，事实上这很有可能进一步削弱了研究所用数据样本的代表性。但由于针对每一个关键词我们都要花费大量的时间去重新检索微博，所以在未来希望弥补此缺陷时，最应当解决的是利用 Python 实现微博的模拟登陆，以达到快速批量获取微博的目的，进而扩大有限时间内可搜索的关键词数目，以此来完善样本的代表性。

除了样本具有较弱的代表性之外，微博检索结果中重复出现的大量广告也可能会扩大单条微博内容的情感得分，导致计算结果大于或者小于真实值。在改进过程中，应当在模型中加入删除重复微博的功能，因为几乎不可能在两篇原创微博里出现相同的内容，由此我们可以甄别出哪些是被恶意重复发送的广告，减少无效信息对模型评分的影响。

此外，金融情感词袋的非专业性也有可能造成微博情感得分的误差。由于本研究在构建金融情感词袋的过程中，受限制于投资建议为负向的个股研报数量过少，我们无法很好地训练分类模型在负向情感词汇上的区分。同时，训练集样本数过少也是导致金融情感词袋不完善的原因之一。但因为是中国股市缺乏做空机制，大量证券公司都偏好于研究预期大概率会

上涨的个股，所以对于推荐卖空的个股研报，我们只能够在未来的研究中不断地收集，以此来完善现阶段才初步形成的金融情感词袋。

七、结论

本研究成功地绕开了寻找能够替代搜索指数的情感反映指标的困难，通过获取股民在股灾前后发布的多条微博，直接对其进行文本分析，获取股民真是的情感变化数据。在对中国股民的情感趋势变化与沪深 300 指数周收盘价的涨跌幅进行了相关性分析之后，本研究发现了股民情感变化具有严重滞后于股价波动的特点，且在大多数时候，二者的波动完全不具有相关性。依此，我们解释了利用搜索指数构建交易策略无法在中国股市上获利的原因。同时，本研究还论证了提前捕捉微博股民情感异常高涨的现象，能够有效地预测股灾的来临，以做好防范措施。这是因为在股市过度虚高的同时，不同于股价的持续无理由上涨难以被预测，微博情感得分会在达到异常制高点之后，提前于股市崩盘出现小幅度的情感下滑，这得益于理性人的负向情感能够在其微博中得以体现，并通过转发放大当天的负向情感得分，从而使得其情感的异常变化能够被轻易且及时地捕捉到，以作为股灾发生前的预警信号。除此之外，本研究还额外发现了中国股民日内交易的情感变化，在早上开盘前大多数人都呈现积极乐观的态度，到了中午休市之后就一直以逐渐低落的情绪持续到夜间结束，这也能侧面说明大部分中国股民都具有自我归因偏差带来的过度自信，在经历了多数情况下都会发生的股市变动与预期不符合之后，由于损失厌恶心理，人们会一直持续心情低落直至下一个开盘点企图扭亏为盈，然而事实往往是事与愿违的。当然，中国股市波动与股民们的预期相反是常态，总的来说，股灾前后一年内股民们的情感得分都维持在 0.1 左右，可以看出大部分人还是能够在多数时候保持理性，避免失误操作而引起不必要的损失与负向情感。

八、学期收获、挑战与未来打算

本学期最大的收获就是能够形成一套完整的文本分析流程，从数据的挖掘、清洗到文本中信息的提取与分析，再到最后利用机器学习的方法将文本分析得到的结果运用到模型训练当中去，得到一个具有预测能力的模型。期间我的 Python 编程能力也得到了很大的提升，尤其是在 pandas 与 numpy 包的使用上，甚至到了学期末能够接触到与机器学习有关的 scipy 包真的是受益匪浅。我也将我在课上所学的内容运用到了期末论文的研究当中去，从微博数据的获取与清洗，到情感得分模型的构建，再到最后的情感分析，期间也运用了机器学习的方法实现了金融词汇的特征重要性求算与金融情感词袋的构建。一整个研究过程都由我自己结合课上所学内容一个人完成，虽然免不了遇到一些棘手的困难，但在克服了其中的挑战之后，也帮助我对文本分析的流程有了进一步的理解，在日后遇到相同困难时我也知道应该从哪个角度入手去解决问题。

在整个学期的学习过程中，最大的挑战就是跳出老师上课给的具体且简单的例子，运用到我们自己的课程论文当中去。这一个过程看似简单，但实际上要我们自己从头去设计一整套实验流程，还有很多步骤需要我们一点点地去细化，且在实施过程中，总会有问题不断地跳出来，一些在课上看似简单的操作，运用到实际当中去却会遇到意料之外的小麻烦。如，微博的网页不能够直接抓取全部信息而需要模拟登陆，且利用 Python 实现自动翻页的功能在课堂上老师也没有提及过，但它却是在我们进行网页信息获取的过程中不可避免的挑战。当我们全部克服了这些困难以后，实际上提升的不光是我们的编程能力，也锻炼了我们从不同角度看待问题的能力。有时我们并不需要正面地去解决问题，而可以从侧面绕开这个问题，比如微博的模拟登陆会耗费我很长时间去学习，但如果我能够让微博信息全部显示在第一页，

我就不需要去模拟登陆后再用 Python 实现自动翻页的功能。于是在我换个角度思考问题以后，我果断地缩短了每次检索的时间间隔，使得微博都能够在一页内就加载完毕，通过更多次的访问就可以获取到完整的信息了，并不必去翻页。

未来我可能不会仅仅停留在文本分析的研究上，我会借着老师在最后几周讲解的神经网络知识，作为我入门机器学习的开始，深入地研究各种不同的非线性回归模型，并探索其在金融领域上的运用，例如，新型的金融防诈骗手段就是利用分类模型训练得到的。当然，老师传授的文本分析的内容也会对我未来的学习有很大帮助，如我前文所提到的，重要的是我习得了一套完整的研究思路，借着这一套流程，我就可以继续更深入地学习不同领域的知识。

参考文献

[1] Tobias Preis，《Quantifying Trading Behavior in Financial Markets Using Google Trends》，《Scientific Reports》，2013 年

[2] 谢明柱，《投资者有限关注与股票收益关系研究——基于百度指数的实证分析》，《江南大学学报(人文社会科学版)》，2018 年 05 期

[3] 王耀君，《基于网络搜索指数的股票市场微观结构特征》，《北京理工大学学报(社会科学版)》，2018 年 05 期

[4] 庞云枫，《基于互联网关注度的股票成交量模型的策略设计》，2018 年 6 月

[5] 肖亭，《一种基于股票情感分析的股市趋势预测方法》，2018 年 12 月 30 日

[6] 赵明清，武圣强，《基于微博情感分析的股市加权预测方法研究》，《数据分析与知识发现》，2019 年 02 期

[7] 相关程序位于“百度指数获取”文件夹下

[8] 人若无名，《随机森林之特征选择》，博客园，https://www.cnblogs.com/justcxtoworld/p/3447231.html，2013 年 11 月 28 日，相关程序位于“百度指数获取”文件夹下

[9] 蓝色枫魂，《朴素贝叶斯分类器》，CSDN 博客，https://blog.csdn.net/qq_32690999/article/details/78737393，2017 年 12 月 07 日，相关程序位于“金融情感词袋构建”文件夹下

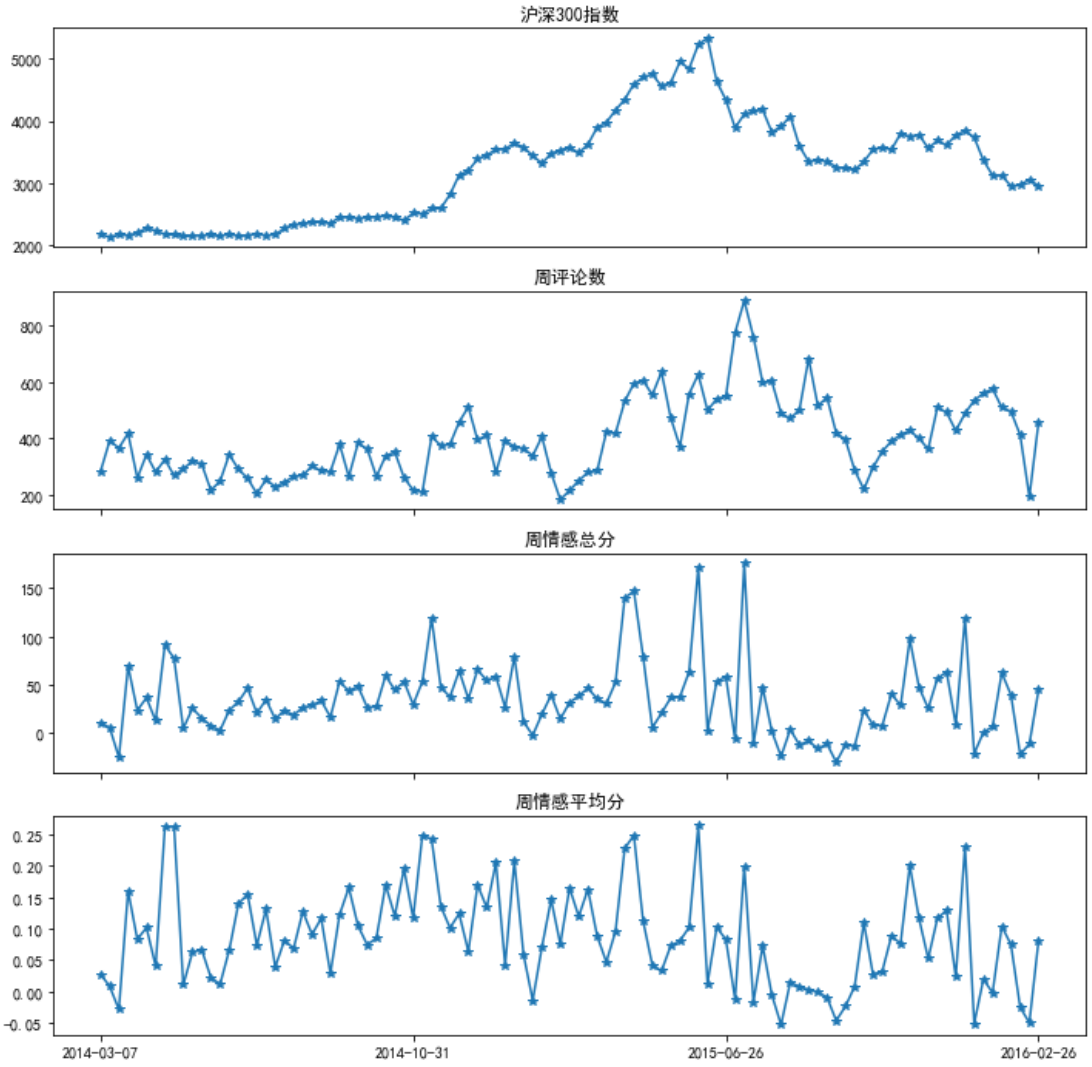
[10] 相关程序位于“微博评论获取”文件夹下

[11] 相关程序位于“微博评论获取”文件夹下

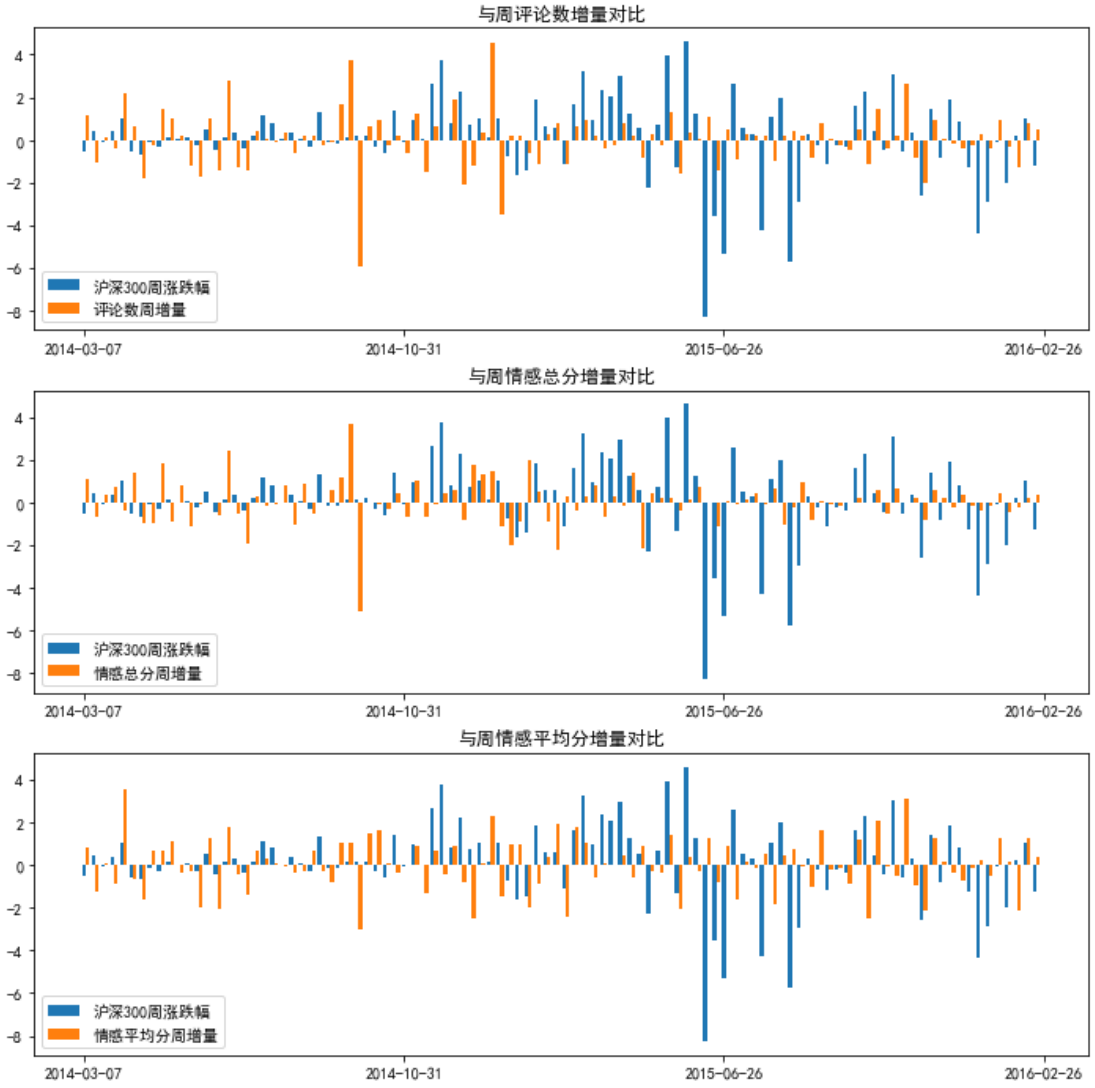
[12] 相关程序位于“情感分析与预测”文件夹下

附录

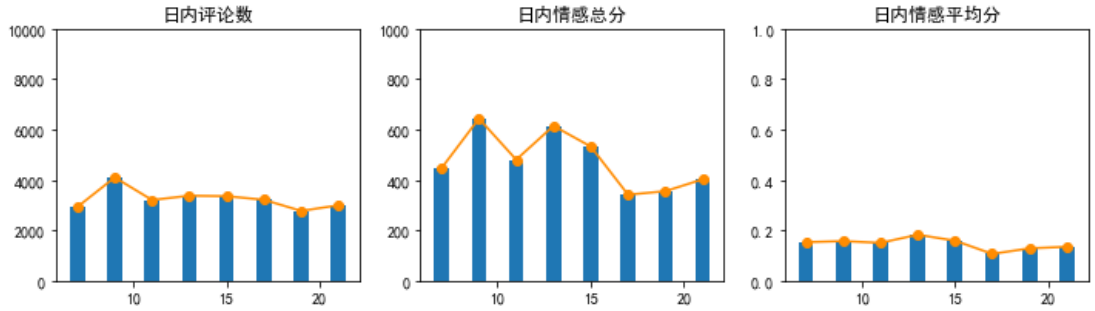
【图 4.2】检索词为“股票代码”的情感得分趋势图



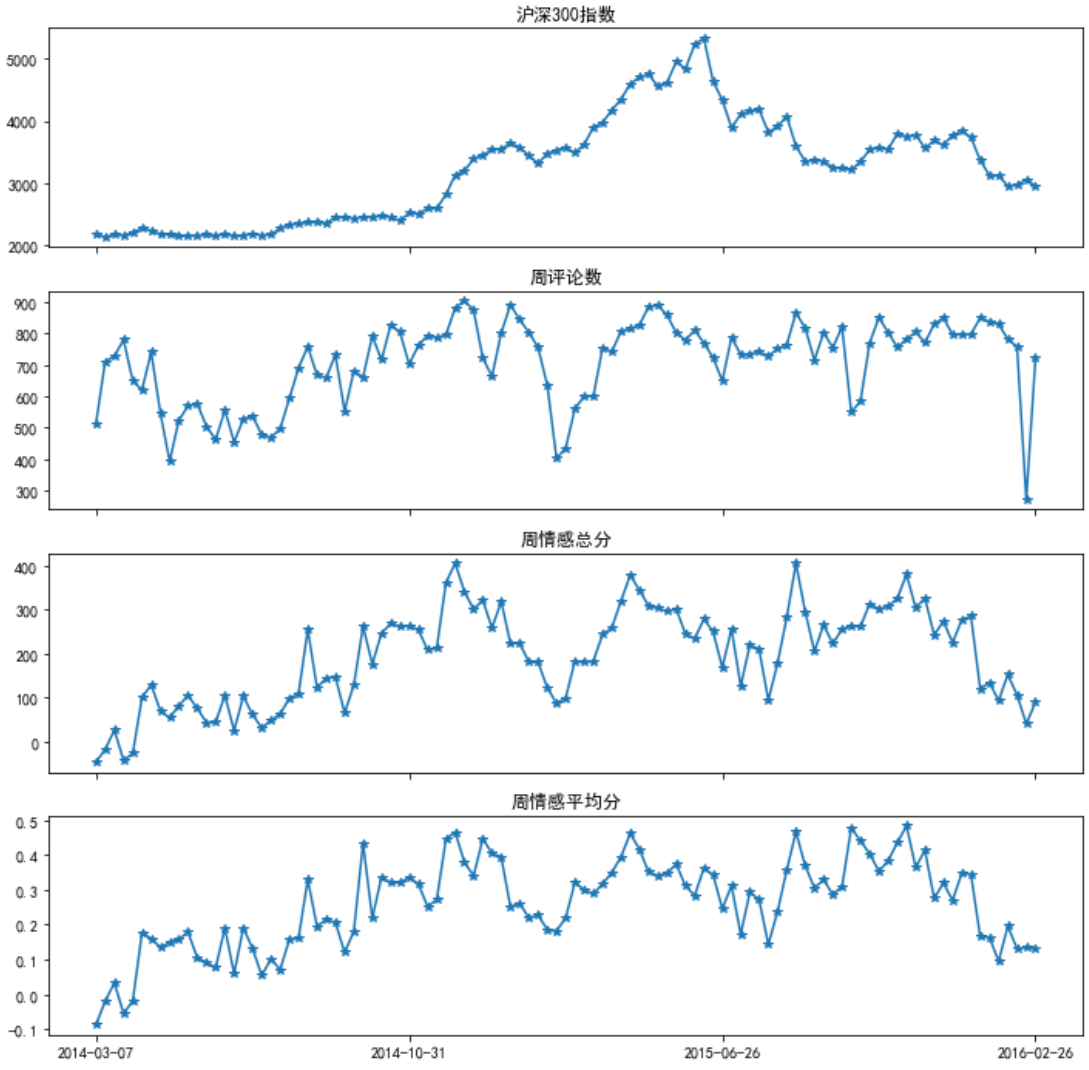
【图 5.2】检索词为“股票代码”的情感得分周增量图



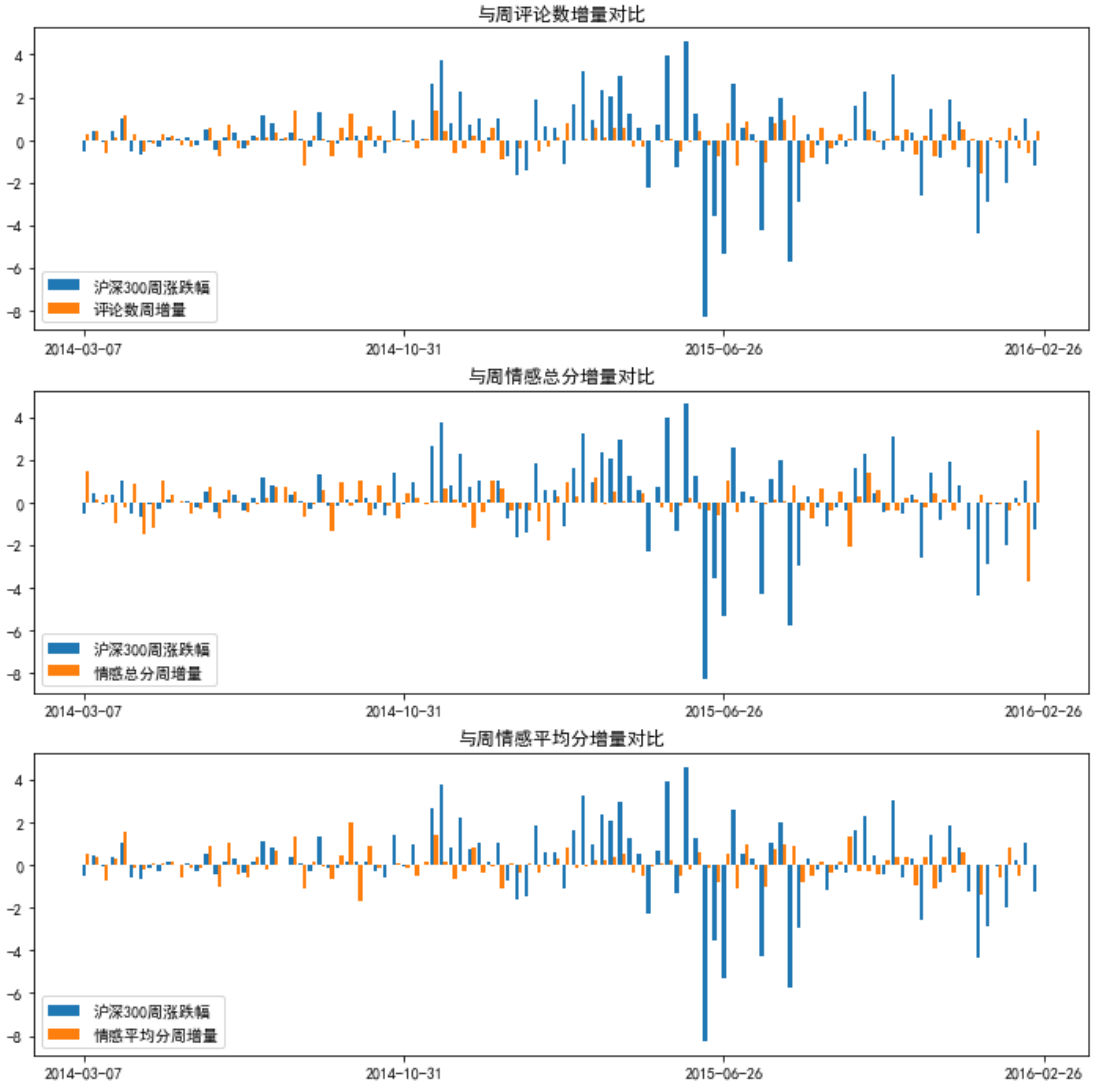
【图 6.2】检索词为“股票代码”的日内情感变化图



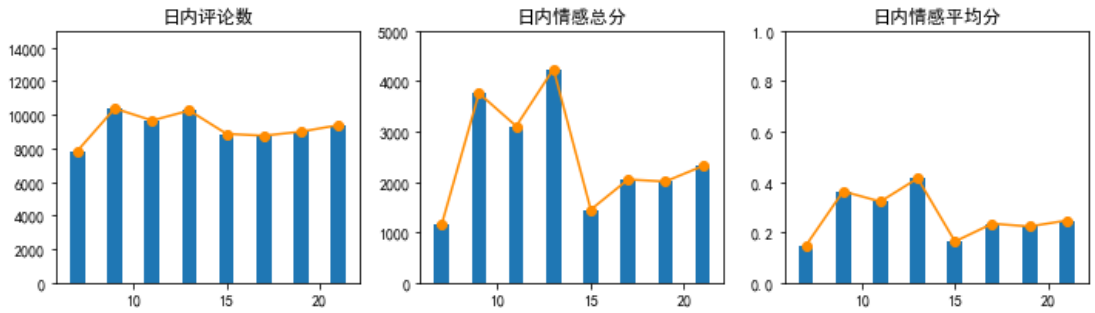
【图 4.3】检索词为“股票交易”的情感得分趋势图



【图 5.3】检索词为“股票交易”的情感得分周增量图



【图 6.3】检索词为“股票交易”的日内情感变化图



经世济民

致致以求