分类号　　密级

中国地质大学（北京）

本 科 毕 业 论 文

题　　目　新冠疫情视角下的我国股市与投资者情感联动性分析

英文题目　Analysis on the Linkage Between Chinese Stock Market and Investor Sentiment from the Perspective of COVID-19

学生姓名　王禹川　　学　　号　1008171223

学　　院　经济管理学院　　专　　业　信息管理与信息系统

指导教师　李华姣　　职　　称　副教授

2021 年 5 月

中国地质大学（北京）
本科毕业设计（论文）原创性声明和使用授权的说明

学院	经济管理学院	专业	信息管理与信息系统	班级	10071781
学号	1008171223	姓名	王禹川	指导教师	李华姣
设计（论文）题目	新冠疫情视角下的我国股市与投资者情感联动性分析				

原创性声明

本人声明所呈交的论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得中国地质大学或其它教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

学生签名：王禹川　　日　期：2021 年 5 月 31 日

关于论文使用授权的说明

本人完全了解中国地质大学有关保留、使用学位论文的规定，即：学校有权保留送交论文的复印件，允许论文被查阅和借阅；学校可以公布论文的全部或部分内容，可以采用影印、缩印或其他复制手段保存论文。

☑公开　　☐保密（____年）　（保密的论文在解密后应遵守此规定）

学生签名：王禹川　　导师签名：李华姣　　日　期：2021 年 5 月 31 日

中国地质大学（北京）本科毕业设计（论文）任务书

学院	经济管理学院	专业	信息管理与信息系统	班级	10071781
学号	1008171223	姓名	王禹川	指导教师	李华姣
设计（论文）题目	新冠疫情视角下的我国股市与投资者情感联动性分析				

毕业设计（论文）主要内容和要求：

1.主要内容

挖掘股吧评论文本中的投资者情感，爬取相应的股票数据建立股票指数，通过上述时间序列，建立情感-股指联动模态，构建联动模态复杂网络模型。通过分析联动模态网络模型的网络拓扑指标和网络特征，进一步新冠疫情影响下的我国股市投资者情绪与股价之间的关联波动关系，并为突发性社会危机事件影响下的我国股市和投资者如何应对提出建议。

2.要求

（1）查阅文献，学习相关理论与技术；

（2）股吧评论文本与股票数据的获取；

（3）对数据进行处理与建模；

（4）构建联动模态网络，计算网络拓扑指标；

（5）分析网络特征，得出结论。

毕业设计（论文）主要参考资料：

（1）文本挖掘相关研究；

（2）股票相关研究；

（3）复杂网络相关研究。

毕业设计（论文）应完成的主要工作：

（1）评论情感值的提取；

（2）股票指标的建立；

（3）联动模态网络建模；

（4）网络指标的分析与结果。

毕业设计（论文）进度安排：

序号	毕业设计（论文）各阶段内容	时间安排	备注
1	**查阅文献，确定选题**	**2020.12.20 -2020. 12.28**	
2	**开题**	**2020.12.29**	
3	**数据获取与处理**	**2020.12.30 - 2021.2.14**	
4	**情感序列与股指序列的计算**	**2021.2.15 - 2021.3.14**	
5	**中期检查**	**2021.4.5 -2021.4.11**	
6	**复杂网络建模与指标计算**	**2021.4.12 - 2021.5.24**	
7	**结题答辩**	**5.25**	

课题信息：

课题性质： 设计　　论文☑

课题来源： 科学研究　　生产/社会实际　　自拟课题　　其他☑

发出任务书日期：2020 年 12 月 22 日

指导教师签名：

2020 年 12 月 22 日

教研室意见：

教研室主任签名：

年　　月　　日

学生签名：

摘要

行为经济学的研究表明投资者的情感与股票之间存在联动关系，而突发性社会事件的发生，通常引起股民情感的变化，进而对股票价格产生影响。2019 年 12 月，中国新冠疫情爆发。疫情之下，我国股票市场的稳定发展受到冲击。为研究社会危机事件下的投资者情感与我国股市之间的联动关系，本研究以新冠疫情为研究背景，以上证指数为实证对象，采用复杂网络为主要实验方法，对疫情期间我国的股市和投资者之间的联动关系进行实证研究。

在具体研究中，本文首先爬取了上证指数在疫情期间的股票数据和对应的股吧评论数据。随后，应用主成分分析法对股票数据降维处理，构建了股价综合指数；应用基于词典的情感分析法计算每日的股吧评论情感值。在这之后，根据计算得到的股价综合指数和情感值，本文采用粗粒化方法得到情感和股票指数之间的联动模态。最后，我们根据联动模态的传递关系建立了复杂网络模型，并选取网络的加权出度、特征向量中心度、紧密中心度、中介中心度、模块度等 5 个网络拓扑指标识别了网络中的关键联动模态和主要传输路径，并分析了网络的社团化特征。

研究发现，在新冠疫情期间，受突发性社会事件的新闻和网上论坛的舆论影响，我国股市股中民的情绪出现不同程度的起伏，并对股票的价格、成交量等指标产生影响。疫情期间投资者情绪与股票以同向联动为主，并在少数交易日反向联动。联动模态的加权出度等指标均符合幂律分布。通过对加权出度、特征向量中心度、和紧密中心度、中介中心度网络等指标的共同分析，确定网络中存在着关键模态和关键枢纽。关键模态连接了网络中的主要传导路径，成为了网络中的枢纽。网络中以一些关键模态为中心存在着社团现象，网络可通过模块度进行社团划分。疫情期间可根据股民情绪对股指进行监控，进行适当的市场调控。本研究建立了股指与股民情感联动的复杂网络模型，并提出一些分析方法，期望在未来为由突发性社会事件引起的股民情绪和能源股价变化提供预警机制，并期待不断完善研究。

关键词：复杂网络，股票市场，投资者情感，主成分分析

Abstract

Behavioral economics shows that linkage exists between investors and stocks, and emergent social events usually causes impact on investors' sentiment as well as stock prices. Under the COVID-19 epidemic, China's stock market has been impacted greatly. To study the relationship between investor sentiment and China's stock market under the COVID-19 background, we took the Shanghai Composite Index as the empirical object, and adopted complex network as main method to conduct an empirical study on the linkage between stock market and investors.

In the specific research, this paper first applied principal component analysis and dictionary-based sentiment analysis to calculate the daily stock price index and the sentiment value of stock bar comments. After that, we used the coarse-grained method to obtain the linkage mode between sentiment and stock index. Finally, we established a complex network model and selected five topological indicators of the network, which including weighted outdegree, eigenvector centrality, closeness centrality, betweeness centrality and modularity class. We identified the key linkage modes and main transmission paths in the network.

The research found that investor sentiment and stocks were mainly linked in the same direction during the epidemic. The weighted outdegree of the linkage modes conforms to the power law distribution, and there are key modes exist in the network. Those critical modes connect the main conduction paths in the network and become hubs in the network. There is a community phenomenon centered on the key modes, and other modes transition to other modes through the key modes.

Keywords: complex network,Stock market,Investor sentiment,Principal component analysis

目 录

1 绪论

1.1 研究背景及意义

1.1.1 研究背景

2019 年 12 月，中国新冠疫情爆发。2020 年至 2021 年初，疫情已经得到基本控制，但仍在部分地方出现反复。疫情之下，各行业生产业务皆受到打击。新冠疫情也在一定程度上导致了我国股票市场的动荡，冲击了金融市场的稳定发展。在经济的高速奔跑历程中，我国的股票市场扮演着重要的角色。一方面，我国股票市场开放至今仅仅历时三十年左右，与欧美长达一、二百年之久的金融股票市场相比，仍处于其市场发展的幼稚时期，在股市机制、法律法规等诸多层面仍存在着明显的不足和待提升。然而，另一方面，在这三十年间，我国股票市场的发展速度却令诸多西方股市望尘莫及，并取得了卓越的成就。截至 2020 年 12 月 31 日，沪深两市共计有上市公司 4140 家，总市值达 79.72 万亿元。我国的股票市场仅用二、三十年的时间，就在规模方面达到了与欧美股市相近的水平。股市的快速发展和和与之巨大规模相伴随的效益，也在我国引起了炒股热潮。截止到 2020 年最后一个交易日，A 股投资者已逾 1.8 亿户；2020 年 A 股的总成交额达到 206 万亿元，总成交量达到 16.7 万亿股[1]。我国股市在促进我国经济发展的道路上发挥着无法替代的重要影响。具体而言，股票交易和股票市场为民众优化个人储蓄的效益提供了方式，为企业融资、扩充资金提供了平台，也使社会上的资资金资源实现合理、有效地配置。由于金融市场的规律和趋势难以预测，股市交易通常伴随着风险。因此，如何防范和化解金融市场的风险，对于我国国民经济的发展意义非常。

在以往针对股市的研究中，研究者们依照传统金融学理论，通常选择结构化的金融数据，如股价、交易量、交易额等数值型数据，而对于非结构化的数据如文本数据等的应用相对较少。然而，行为经济学的研究显示，除了整体的市场行情、政策和国内外市场等因素，一些异常的市场现象也由投资者的情感引起[2]。进入 21 世纪后，随着互联网的高速发展和 Web2.0 时代的到来，各式各样的社交媒体不断涌现，例如国内的微博、微信，国外的 Twitter、Facebook 等。由于其便捷、传播信息快速的特点，社交媒体已经成为人们获取信息的重要渠道。股民

也开始在股吧或其他主流社交媒体上针对个股或股市整体发表自己的见解。这些评论中往往包含着其自身的情感，并随着互联网的连通、传递使其包含的情绪在股民间发散，进而影响他们的行为和投资决策，对投资者的市场行为产生更加广泛的影响。具体而言，乐观积极的市场情绪可能引起股票价格和成交量的上升，悲观的股民情绪则可能导致股价下跌[3]。随着互联网爬虫、网页文本采集等技术的发展和以自然语言处理技术为基础的文本挖掘理论的不断丰富，这些情感资源已经可以为研究者所捕捉。在社交媒体上对网民的评论文本采集，并挖掘其中的舆情信息，已经成为行为经济学对股市研究的重要方向。

投资者情绪既由市场本身的变化引起，也和社会事件相关。当一件突发性社会危机爆发，往往引起股民情绪的复杂变化，进而对股票价格产生冲击。具体而言，某一领域内的社会危机往往首先引起相关行业股票的巨大波动，进而通过其社会影响，对投资者的情感产生影响[4]。当某一社会事件影响范围较大时，其辐射范围则可能作用于整个股市。新冠疫情影响范围广且历时持久，疫情出现多次反复，对于我国股市和投资者的情感产生着持久的影响。研究新冠疫情背景下的投资者情感与股票指数之间的联动关系，对于社会危机事件下的我国股市维持稳定具有重要意义，也对于在社会突发事件下投资者进行自我调节以适应市场变化具有借鉴作用。

1.1.2 研究意义

本文应用文本挖掘技术进行情感分析，并构建情感-股票指数联动模态复杂网络模型，研究在新冠疫情背景下的我国股市投资者情感与股票之间的波动关联，对于金融股票研究具有显著的理论意义和实践意义。

（1）理论意义

传统金融理论的股票研究多通过数值型的股票统计指标对股市进行研究，而往往忽视了投资者与股市之间的关联和作用。近些年的研究表明，投资者的情感表达也可用于股市研究。本研究应用文本情感分析技术，挖掘股吧评论中的投资者情绪，可以丰富和完善情感分析应用于股市研究的相关理论。

此外，本研究构建了一种情感-股票指数联动模态复杂网络模型，将股票数值型数据的波动与投资者情感的波动进行联动比较。对于经济时间序列复杂网络在股票研究和投资者情感研究，提供了一种新的思路。

（2）实践意义

股票市场的发展状况反映着国家经济的运作水平，对于股市与投资者情感的研究对于政府经济部门和个体投资者的决策具有重要的借鉴作用。本研究结合文本情绪挖掘新冠疫情背景下的投资者情感对股市的影响，有助于政府和投资者量化股民情绪，为面对持久性的社会危机事件时如何做出金融决策提供依据。

1.2 研究现状

1.2.1 文本情感分析研究现状

文本情感分析又被称之为文本意见挖掘，它是指依据文本挖掘技术和理论，对于包含有情感色彩的文本进行分析和处理，挖掘文本中所隐含的主观情感信息。文本情感分析有效地利用了非格式化的文本数据，挖掘其中不能直接被人们所利用的情感信息并提炼成格式化的数据，对于非格式化的数据挖掘具有重要意义。

作为文本挖掘下面的一个分支，文本情感分析在二十一世纪初随着 Web2.0 时代的到来而兴起。针对文本情感分析，国外的研究起步相对较早。在文本情感分析的技术思路方面，J.H.Yi 和 W.Niblack 提出使用自然语言处理技术来确定每个主题引用的情感，而不是对整个文档的主题进行分类，并描述了功能完备的系统环境和算法[5]。S.Matsumoto 和 H.Takamura 等应用支持向量机在电影评论数据集上进行实验，根据对文档的意见的正反极性(赞成或反对)对文档进行分类，并提出利用句子中词语之间的句法关系进行文档情感分类的方法[6]。在对多种语言的文本进行情感分析方面，A.Abbasi 和 H.C.Chen 等提出利用情感分析方法对多语言网络论坛的观点进行分类，针对阿拉伯语的语言特征，集成了特定的特征提取组件，提出了一种融合信息增益启发式的混合遗传算法[7]。文本情感分析通常以需求为导向，应用在电商、股市、舆情监控等诸多领域。作为早期将情感信息挖掘应用于商业和经济学方面的研究，M.Gamon 和 A.Aue 等提出了一个代号为 Pulse 的原型系统，用于从汽车评论数据文本中联合挖掘主题和情感倾向性[8]。A.Archak 和 A.Ghoose 等将亚马逊电商网站的评论分解成评估产品个体特征，开发了一种结合文本挖掘和计量经济学的新型混合技术，将消费产品评论建模为特征和评价空间张量积中的元素[9]。该研究表明，与单纯依赖数字数据的基线技术相比，评论的文本部分可以改善产品销售预测。

随着对于文本挖掘应用于情感分析的实际价值的深入认识，国内相关研究也

逐渐丰富。在文本情感分析的理论层面，赵妍妍等按照情感信息抽取、分类和检索这三项主要任务对文本情感分析的研究现状和应用方式进行了梳理和介绍[10]。考虑到传统 LSTM 方法的不足，梁军等提出了一种基于长短时记忆扩展得到的树结构的递归神经网络，用于文本深层次的语义信息挖掘[11]。在文本情感分析的应用研究方面，文本情感分析被广泛应用于网络评论的挖掘，已得到客户的反馈信息。赵春艳等应用文本对西部某旅游产品的网络点评数据进行高频词、语义网络等方面的挖掘，探求顾客对旅游产品的满意程度，以便得到改进措施[12]。许欣等应用 BERT 分析电商评论中隐藏的情感信息，并解决数据的不平衡问题[13]。

1.2.2 投资者情绪研究现状

股票市场是我国经济的重要组成部分，对股票和金融市场的研究对于政府和投资者都具有重要意义。传统金融学理论主要通过统计学和计量手段，对基本面和技术面进行分析。近些年，行为经济学研究表明，投资者的情感表达也可能对股票市场产生影响，股民情绪的变化与股价变化往往关联波动。投资者情绪相关研究可以探究二者之间的演变规律[14]。情感与股票价格的相关关系和如何应用投资者情绪对市场进行预测是投资者情绪研究的核心问题，许多学者利用文本挖掘技术和统计学习、计量模型、机器学习等方法，对于二者的联动效果进行了研究。

为了检验网络舆情和股票的回报是否之间是否具有联动关系，国内外研究者均进行了大量的研究工作。PK.Narayan 和 D.Bannigidadmath 构建了金融新闻时间序列数据集进行实验。实验发现正面和负面新闻对股票收益的预测均有效果，正面新闻对股票收益的影响更大；对一些股票来说，金融新闻对回报的冲击只有部分逆转[15]。许启发、伯仲璞、蒋翠侠等分别运用均值格兰杰因果和分位数格兰杰因果检验，探讨股民情绪波动与股票收益之间的因果关系，并发现尽管在均值检验下情绪波动与收益之间因果关系并不明显，但基于分位数的因果分析却表明二者在极端分位点区间处存在广泛且显著的因果关系[16]。

在验证二者存在因果关系后，进一步得出情感、行为因素如何影响股票投资成为研究的主要问题。A.Siganos, A 和 E.Vagenas-Nanos 等利用 Facebook 的国民幸福总值指数，研究了 20 个国际市场内的每日情绪和交易行为之间的关系，并发现情绪与股票收益呈同时期正相关，投资者的情感与股市之间存在因果关系。研究还观察到市场人气与回报率之间的关系在接下来几周出现逆转，进一步表明，

负面情绪与交易量和回报波动性的增加有关[17]。易洪波、赖娟娟、董大勇等运用关键词词典衡量投资者情绪，并运用 VAR 模型考察网络论坛投资者情绪与交易市场指标的关系。结果表明，投资者的空多双方情绪对市场成交量和收益率存在着非对称影响；非交易时段情况下多方情绪对市场未来收益率存在影响[18]。

考虑到投资者情绪对于股票市场的关联机制和影响作用，应用舆情信息对股市进行预测成为投资者情绪研究在实际应用上的重要内容。董理、王中卿、熊德意等运用支持向量回归研究了股民情绪对股票价格的预测机制[19]。为了得出较好的预测模型，L.Liu 和 J.Wu 等提出了一种新的模型，根据公司特定的社交媒体指标预测股票波动，并揭示社交媒体指标对股票回报波动研究的影响。通过分析来自纽约证券交易所和纳斯达克证券交易所的样本，该研究发现拥有官方 Twitter 账户的公司比没有官方 Twitter 账户的公司有更高的波动，公司的关注者数量和发送的推文数量等指标不仅可以预测股票的波动，而且可以显著提高波动预测的准确性[20]。

总而言之，在投资者情绪方面的相关研究，涉及股市研究、行为经济学，以及应用统计学和机器学习方法进行有效预测，是一个多学科交叉的研究领域，仍有大量研究方向等待细化和挖掘。

1.2.3 股市的联动性研究

联动性，是指不同对象之间因为各种内外因素而拥有相同的运动趋势的现象。对于联动性的研究被广泛应用于股票市场的研究中，其联动的范围从个股之间的联动、板块之间的联动，到不同股市、不同国家之间的联动，具体可以表现为某两只股票或某两个股市的总体走势同涨或同跌等。对于联动性的研究，往往针对两个方面，即联动性是否存在和联动性产生的原因。近年来，国内外针对股市联动性相关的研究不断展开。

对于股市中的联动性最基本的研究方向是对股市本身进行研究，研究的对象往往是不同股市或不同个股之间的联动。在对国内与发达资本主义国家的股市联动性研究方面, 韩非和肖辉以上证 A 指和美标准普尔指数为例，通过研究得出结论，国内股市与美国股市之间的相关性较弱。具体对于两国股市的影响方向,国内股市的收盘收益率对美国股市的开盘价格存在影响,但是影响较为微弱；而美国股市的收盘表现则对中国股市的开盘价格几乎不存在影响[21]。周珺利用协整分

析和格兰杰因果检验对上证市场与周边主要证券市场的长期动态均衡关系进行了分析。结果发现,上海的证券市场与台湾、日本的证券市场之间均没有协整关系,但与香港证券市场在某一样本区间内不仅存在协整关系,而且存在单边性引导关系[22]。

除了股票市场本身，对于其他经济市场与股票市场的联动也是重要的研究方向。Mensi.W 和 Hkiri.B 等人考察了金砖四国的股票收益率与原油价格和黄金价格之间的共同运动，结果表明，金砖四国指数收益率与西德克萨斯中质原油价格在低频率(长期)下有共同变动，但没有发现金砖国家股票市场和黄金价格之间的共同运动的证据[23]。Bashir.U 和 Yu.YG 等人通过时变协同运动研究了外汇市场和股票市场之间的动态关系，他们分析了拉丁美洲国家从 1991 年到 2015 年的月度时间序列，并应用格兰杰因果关系来验证外汇市场和股票市场之间的因果方向。其实证结果表明，所有拉美国家的汇率和股价之间存在正的相互关系[24]。

股票市场与其他贸易市场的联动性研究表明，时间序列之间可以广泛的开展联动性研究。因此，在前文提到的投资者情绪研究领域，也有学者根据情感时间序列和股票时间序列进行股民情感与股票价格之间的联动性研究。余秋玲基于我国 A 股市场的面板数据进行分析，探讨投资者情绪与证券市场的股价联动。该研究发现，我国证券市场中存在明显的股价联动效应，投资者的情绪对股价联动有负向影响，个股投资者的情绪对股价联动的影响大于市场层面的综合情绪变量，并且随着公司规模增大，情绪对股价的影响逐渐减小[25] 。

对于股价的联动性研究被经常应用于某一社会事件的金融影响上。以新冠肺炎为例，钟熙维和吴莹丽以上证指数、道琼斯指数和恒生指数为例，利用 VAR 模型分析全球股市收益率之间的联动性，其研究证明了此次由新冠疫情引起的金融危机在三大指数之间的传递路径为由上证指数至道琼斯指数，道琼斯指数与恒生指数之间相互传导[26]。

由于类似的社会事件往往引起投资者情感的变化，因此对于某一社会事件背景下的情感与股票价格的联动性研究具有可行性。

1.2.4 复杂网络在股市中的应用

复杂网络是一种通过网络研究高度复杂的系统的有效途径，其网络中的节点表示系统中的个体，边表示个体之间的联系。近些年，复杂网络不断被应用于贸

易网络研究中。董志良和杨巧然采用复杂网络方法研究国际粮食贸易网络的网络特性，计算出网络的参数特征，并分析国际粮食贸易的特点，随后分析在不同攻击模式下的粮食贸易网络的鲁棒性[27]。刘羿和余航将我国 24 家上市银行机构组成一个金融复杂网络，随后以改进三因子定价模型为基准，获得各机构的时变违约概率并将之深度融合进复杂网络分析方法，最终构造出了一种可度量银行机构系统性风险和重要性指标的新的计量模型[28]。H.Zhang 和 Y.Wang 以 2000 - 2016 年全球样本为数据，从复杂网络视角分析能源进出口的贸易模式，并探讨了不同类型的风险对贸易模式的影响[29]。

在贸易网络研究中，复杂网络常常时间序列演变的研究相结合。Huang X 等结合事件分析法，运用复杂网络研究了国际反倾销事件对我国光伏发电产业股票的影响，并发现不同事件和股价的影响程度以及敏感程度均不同[30]。Gao XY 等应用复杂网络研究原油价格时间序列的波动机制，并探究了不同汇率下的影响[31]。

应用复杂网络对时间序列进行研究，可以探究股票之间的联动关系。刘超和郭亚东通过定义股票间相互影响的联动模式，构建加权有向网络图以分析股市间的联动关系，该研究表明，不同股市之间有明显的传染效应,且传染的速度和持续时长均不同，传染效应在联动网络中表现为联动模式的高聚类性和高联动性[32]。因此，类比的看，采用复杂网络方法，可以有效挖掘投资者情绪和股价序列的关联和联动关系。

1.3 研究内容与亮点

1.3.1 研究内容

本研究挖掘股吧评论文本中的投资者情感，爬取相应的股票数据建立股票指数，通过上述情感波动时间序列和股票指数波动时间序列，建立情感-股指波动的联动模态，构建联动模态复杂网络模型。通过分析联动模态网络模型的网络拓扑指标和网络特征，进一步新冠疫情影响下的我国股市投资者情绪与股价之间的关联波动关系，并为突发性社会危机事件影响下的我国股市和投资者如何应对提出建议。

如图 1-1 所示，本研究具体实验方法通过以下步骤实现。首先，确定研究对象和时间序列的区间，收集相应的股票数据和股民评论数据。随后，对于股票数据和评论数据，分别进行建模前的数据预处理。对于股票数据，进行 Z 标准化

处理，随后进行 KMO 检验和巴特利球形度检验；确定检验通过后，使用主成分分析对多个指标进行降维，得到股票数据综合指数。对于文本评论数据，我们应用基于词典的情感分析，计算每日的情感值。这一步完成后，分别计算二者的日波动值值。应用处理好的股票综合指数波动时间序列和情感值波动时间序列，我们进行平稳性检验并计算皮尔逊相关系；确定序列平稳及二者存在相关关系后，我们采用粗粒化方法，根据两个变量间的同向或异向波动方向，分割联动模态。最后，我们应用联动模态的传递关系建立复杂网络模型，并分析模态特征，识别关键模态。

本文设置章节内容如下。

第一章是本研究整个论文的绪论部分。该章节首先明确了研究背景，随后论述研究意义；通过阅读文献资料和比较前人的研究成果，对文本情感分析、投资者情绪、股票联动性、复杂网络理论以及复杂网络在股票市场方面的国内外研究成果和现状进行了总结；最后，根据前人的研究经验与不足，提出了本文的研究内容和研究亮点，明确了技术路线图。

第二章根据本研究的技术路线介绍了本研究主要应用的理论知识和相关技术，共分为三部分，分别是文本情感分析方法、股票数据降维方法和复杂网络拓扑指标。在文本情感分析方法上选取基于情感词典的情感分析方法。对于冗余、复杂的股票数据，使用主成分分析法对其降维。最后，介绍了用于分析结果的 5 个复杂网络拓扑指标节点加权出度紧密中心度、特征向量中心度、中介中心度、模块化系数。

第三章基于复杂网络方法初步构建了一个情感-股票指数波动联动模态网络，并对构建的过程进行了解释。包括通过文本情感分析计算情感值，主成分分析构建上证综合股票指数，以及如何运用粗粒化方法切割联动模态。

第四章是结果分析部分，是论文的核心部分。本章选取了 5 个复杂网络拓扑指标，即加权出度、特征向量中心度、紧密中心度、中介中心度、模块化系数。根据这 5 个拓扑特征的数值，对情感-股票指数联动网络进行了网络特征分析，并识别了关键路径和关键模式，发现了隐藏在网络中的核心子网络。本章内容也是得出本文后续结论的基础。

第五章是结论部分，根据第四章的结果对于本文的分析内容进行了总结，并

提出了针对重大突发事件下对投资者情绪进行研究的角度的建议。

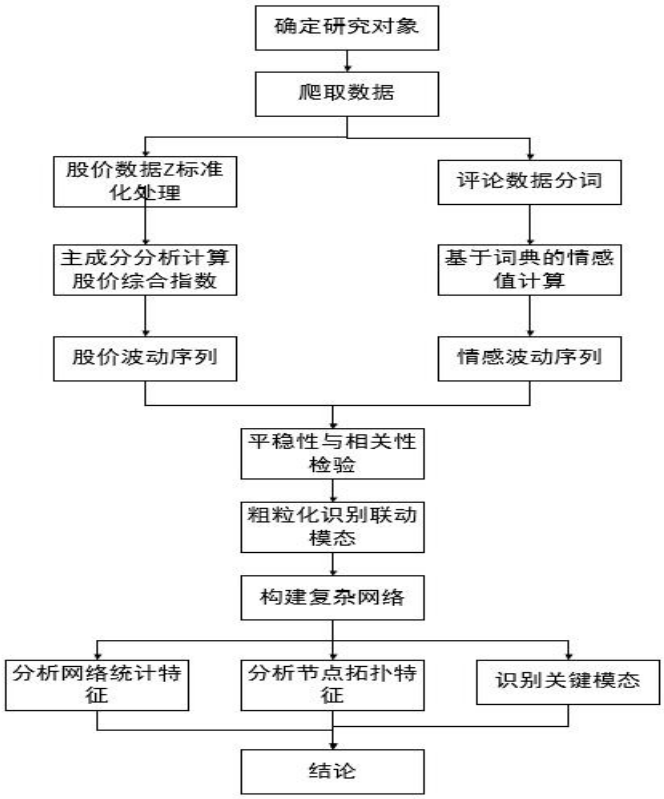

图 1-1 研究过程

1.3.2 本文亮点

本文的亮点如下：

（1）应用文本情感分析，考虑情感对股票走势的影响因素，将投资者的情感与股票数据结合进行建模。没有依据传统金融学的经典统计方法，而是依据行为金融学理论，研究投资者的情感和股票市场之间的联动关系。

（2）构建了一种股票综合指标。传统应用股票数据的研究往往直接选用单一的数值型股票数据指标，例如收盘价或最高价等等。本研究为了避免单一的股票数值型数据带来的预测不准确和误差影响，选择了多个股票数据指标，并通过降维方法主成分分析法构建了一个股票数据综合指标，以代替单一的股票数据，增加了分析的准确性。

（3）构建了一种情感-股指波动时间序列的联动模态，并由此建立复杂网络模型。以往的研究在应用联动模态进行股票市场研究时，大多是研究不同股票间的关联波动关系或情感值之间的联动和传递，本研究则定义了一种情感和股指间

的联动模态，用来研究股票走势和投资者情感之间的关联和传递关系。

（4）以新冠疫情背景作为研究案例，研究突发性社会危机事件对于股票和金融市场及投资者情感的持久性影响，为金融市场和投资者应对此类突发性事件提供借鉴和思考。

1.4 本章小结

本章是本研究整个论文的绪论部分。该章节首先在论述了研究背景后明确了研究意义；随后通过阅读文献资料和比较前人的研究成果，对文本情感分析和复杂网络在股票市场方面的国内外研究成果和现状进行了总结；最后，根据前人的研究经验与不足，提出了本文的研究内容和研究亮点，明确了技术路线图。

2 研究数据及方法

2.1 数据来源

本次新冠疫情的最早案例于 2019 年 12 月初被发现；2020 年 1 月和 2 月，国内疫情达到顶峰；3 月开始，国内疫情逐渐好转；至 5 月，我国全部省份均降至突发性公共卫生事件二级响应以上。但 5 月开始经过 2020 年底至 2021 年初，全国范围内均有部分省市出现疫情的不断反复。本研究全国范围内疫情对于金融市场均可能产生影响，因此选取的研究时间为 2020 年 1 月 1 日至 2021 年 1 月 31 日。

对于研究对象，本文选取上海证券综合指数（简称上证指数）作为实证研究的对象。上证指数是我国股市的主要综合指标，能较好地反映我国股市和金融市场的状况。为避免单一股票数据指标带来的预测不准确，本研究选取上证指数在 2020 年 1 月 1 日至 2021 年 1 月 31 日期间每日的开盘价、最高价、最低价、收盘价、成交量、成交额作为变量，剔除非交易日，共包含 263 个交易日的数据[33]。

本研究的股民情感值由股吧评论计算而来。评论数据来自东方财富网，该网站的股吧为网民提供了发帖交流和互动的平台，是国内财经板块最热门且最活跃的股吧，能够充分反映舆论所关注的金融热点。以 Jupyter Notebook 为平台，使用 Python 编写爬虫程序。根据网址的 url 结构对 2020 年 1 月 1 日至 2021 年 1 月 31 日时间段内上证指数股吧每日的评论进行爬取标题，内容包括标题内容、发表作者、发表时间、评论和阅读数量等。剔除非交易日评论和空标题、标点标题、转发等不能表达股民情绪的无意义评论数据，共得到 263 个交易日内的 104876 条评论数据[34]。

2.2 研究方法

本研究分析股吧评论文本情感值，并将股票数值型数据降维后得到股票综合指标，随后构建联动模态复杂网络模型。使用的主要技术和方法包括如下：

（1）数据获取技术：用于获取股吧评论数据的网页文本采集技术；

（2）数据处理技术：对股吧评论进行清洗、整理、分词等操作的文本预处理技术，用于计算股吧文本每日情感值的基于词典的情感分析方法，用于构建股票数据综合指数的降维方法主成分分析法；

（3）数据检验方法：用于检验数据是否能进入因子分析的 KMO 检验和巴特利球形检验，用于检验时间序列平稳性的 ADF 根检验法，用于检验数据相关关系的皮尔逊相关系数计算方法。

（4）建模方法：用于构建情感-股指联动模态的粗粒化方法，复杂网络建模方法。

本章介绍本研究采用的一些主要技术、方法和分析指标，包括情感分析方法、输入特征降维方法，以及选取的用于结果分析的复杂网络拓扑指标。

2.2.1 基于词典的情感分析方法

对于评论文本的情感分析法主要有基于词典的情感分析与基于机器学习的情感分析。本研究采用基于词典的情感分析。具体的步骤如下：

step1. 读取评论数据，对评论进行分词。

step2. 查找分词中的情感词，主要考虑否定词和程度副词的影响，记录积极还是消极，以及位置。

step3. 往情感词前查找程度词，找到就停止搜寻。程度副词和情感词的组合中在语句中起到加强情感词的作用，如“特别没意思”，例子中情感词“没意思”表示负面情感，程度副词“特别”对情感词“没意思”起到了加强作用。为程度词设权值，乘以情感值。

step4. 往情感词前查找否定词多个否定词及情感词组合在情感句中保持极性不变或极性相反，如“并非不实惠”，例子中的“实惠”表示正面情感，否定词“不”对“实惠”情感词起到相反作用，否定词“并非”对情感表达“不实惠”起到了再次相反作用，此情感表达极性不变。若否定词为双数，情感极性不变，否则，极性相反。

step5. 对于一个句子的情感值，由下式计算：

$$S = w_{adv} \times (-1)^n \times w_{senti} \qquad (1)$$

其中 S 代表每条评论的情感值，w_{adv} 代表程度副词的权重，n 代表否定词的次数方，w_{senti} 代表情感词的权重。表 2-1 以依据某一的评论为例，计算该句子的情感值。

表 2-1 评论情感值计算

日期	序号	内容	程度副词权重	否定词次数	情感词权重	得分
2020-12-03	1	美股走势太垃圾了	太 2	0	垃圾 -2	$2*(-1)^0*(-2)=-4$
2020-12-03	2	今天不能说不给力	很 1	2	给力 2	$1*(-1)^2*2=2$

step6. 统计情感分析结果，将每一天的所有文本的情感值求其平均值作为该天的情感得分，计算公式如下：

$$DS_i = \frac{1}{N}\sum_{j=1}^{N} S_j \qquad (2)$$

其中 DS_i 为第 i 日情感值，S_j 为该日第 j 条评论的情感值，N 为该日的评论总数[35]。

2.2.2 主成分分析法

研究的系统较为复杂时，变量较多会增加分析的复杂性与难度。应用主成分分析法，将有相关性的变量转化为彼此独立的新的变量，可以实现对数据降维。本研究所选取的多个股票指标如开盘价、最高价、最低价、收盘价、成交量、成交额之间有一定的相关性，所包含的信息之间具有一定的重叠性，因此，我们使用主成分分析法处理这些指标，得到一个由包含这些股票指标的新的变量组成的股价综合指标[36]。应用这个综合指标，可以在基本保留原始变量内部信息的基础上，更好地揭示一支股票的每日表现。应用主成分分析法建立股价综合指标的算法步骤如下：

step1. 构造包含 n 个研究对象和 m 个指标的样本矩阵 $X_{(n\times m)}$，对样本进行标准化处理，形成标准化样本矩阵根据样本矩阵 $a_{(n\times m)}$，计算相关系数矩阵 $R_{(m\times m)}$；

step2. 计算 KMO 检验和巴特利球形检验的值，检验是否能进入因子分析；

step3. 应用雅可比法求出相关系数矩阵 $R_{(m\times m)}$ 的特征值 λ，计算对应于每个

特征值λ_i的特征向量u_i，使$\|u_i\|=1$，由特征向量组成m个新的指标变量$[y_1,\cdots,y_m]$；

step4. 计算每个主成分y_i的负载$yi=\sum_{j=1}^{m}u_{ij}a_{ij}$；

step5. 计算主成分y_i的信息贡献率$b_i=\frac{\lambda_i}{\sum_{k=1}^{m}\lambda_k}$和累计方差贡献率$\alpha_i=\frac{\sum_{k=1}^{i}\lambda_k}{\sum_{k=1}^{m}\lambda_k}$，根据累计方差贡献率选取前q个主成分$[y_1,\cdots,y_q]$作为新的变量代替原来的变量。

step6. 计算综合得分$Score=\sum_{i=1}^{q}b_iy_i$，作为股票综合指标[37]。

2.2.3 复杂网络

本研究通过可视化的复杂网络对探究不同联动模态在时间序列中的特性和地位。复杂网络是一种将现实世界中的各种复杂系统通过抽象的方式进行网络模型化的方法。在网络中以节点代表现实世界中的个体，以联结节点的边代表个体间的关系。在联动模态网络中，不同的模态即为不同的节点，模态间依据时间序列的传递关系即体现为边。复杂网络的拓扑特征是指能体现网络拓扑结构的一些特征，拓扑结构是能够展现网络性质的结构，选择合适的复杂网络拓扑特征进行分析，可以更好的挖掘模态特征。本研究选择加权出度、特征向量中心度、紧密中心度、中介中心度、和模块化系数作为5个指标来研究节点的拓扑特性。

（1）加权出度

在有向加权网络中，节点的加权出度度是从一个节点发出的边的权重之和，且除了悬挂节点，加权出度等于加权入度。节点的加权出度值越大，该节点在网络中的的作用与影响力越大。节点度的计算公式为：

$$K_i=\sum_{j=1}^{N}k(j,i) \tag{3}$$

其中K_i为节点i的加权出度值。i和j表示网络中的节点。N为网络中的节点总数。$k(j,i)$为表示从节点i到节点j的边的权重。在联动网络中，一个节点

的加权出度度表示由该模态可以转换到的模态数，反过来更多的模态是由这种模态转换而来，证明其与众多节点相关联。节点的加权出度值越大，在网络中的地位和影响力越高，代表该情感与股票的这种联动关系在网络中出现频率高[38]。

（2）特征向量中心度

特征向量中心度体现一个节点的邻居节点的重要性，当节点的特征向量中心度高时，其邻居节点为网络中重要的点，该节点因此传播更为重要的信息。]特征向量中心度的计算公式为：

$$EC_i = \lambda^{-1}\sum_{j=1}^{N} a_{ij}e_j \tag{4}$$

a_{ij} 为表示节点 i 和 j 之间是否存在边连接的 0-1 变量，若边存在则为 1，否则为 0。设 A 为网络的邻接矩阵，$[\lambda_1,\lambda_2,\cdots,\lambda_N]$ 为矩阵 A 的特征值，且 λ_i 对应的特征向量为 $a=(e_1,e_2,\cdots e_N)$[38]。在本研究中，特征向量中心度体现一个模态的重要性。特征向量中心度高的模态更多地与其他重要模态相连，代表情感与股票的这种联动状态在网络中与众多重要节点相连接，因此成为主要的联动模态，并成为关键节点。

（3）紧密中心度

紧密中心度体现一个节点与其他节点的接近程度。节点的紧密中心度越大，越容易到达其他节点[38]。节点紧密中心度的计算公式为：

$$CC_i = \frac{N-1}{\sum_{j\neq i} d_{ij}} \tag{5}$$

CC_i 为节点的紧密中心度，d_{ij} 为节点 i 到 j 的最短路径数。在联动模态网络中，一个模态的紧密中心度越大，越容易到达其他模态，则该节点越重要。

（4）中介中心度

在网络中一个节点也可能出现在其他节点之间的最短路径上，中介中心度指的是一个节点出现在其他节点之间的最短路径上的个数。节点的中介中心度越高，

对网络中信息传播的控制力越强。节点的中介中心度计算公式为：

$$BC_i = \sum_{j \neq i} \frac{d_{ij}()}{d_{ij}} \tag{6}$$

BC_i 为节点 i 的中介中心度。d_{ij} 为节点 i 到 j 的最短路径数，$d_{ij}()$ 为节点 i 到 j 的最短路径经过的节点数[38]。在联动模态复杂网络中，一个节点的中介中心度体现节点作为枢纽的能力。中介中心数高的节点，有更多的转换模式通过，更有可能成为网络中的关键节点。

（5）模块化系数

网络的模块化系数描述的是网络整体的集团化程度，网络中节点的模块化系数描述的是节点与网络中集团的联系紧密程度。模块化系数高的节点与其他节点之间联系更紧密，往往容易形成小集团。节点的聚类系数的计算公式为：

$$Q = \sum_{i=1}^{N} \left(\frac{1}{E} - (\frac{2I + O}{2E})^2 \right) \tag{7}$$

Q 为节点的聚类系数。I 为两节点均在同一社区的边的数目。O 为其中一个端点在而另一个端点不在同一个社区的边数[38]。节点的模块化系数较大，说明该模态处在某一社团区域；节点的模块化系数较小，则该节点可能是网络中的悬挂节点，所处位置网络稀疏。

2.3 本章小结

本章根据本文的技术路线介绍了本研究主要应用的理论知识和相关技术，共分为三部分，分别是文本情感分析方法、股票数据降维方法和复杂网络拓扑指标。在文本情感分析方法上选取基于情感词典的情感分析方法。对于冗余、复杂的股票数据，使用主成分分析法对其降维。最后，介绍了用于分析结果的 5 个复杂网络拓扑指标节点加权出度紧密中心度、特征向量中心度、中介中心度、模块化系数。

3 联动模态的建立

3.1 基于词典的股吧文本情感分析

使用 Python 爬虫程序，爬取 2020 年 1 月 1 日至 2021 年 1 月 31 日的上证指数股吧每日的评论标题，剔除非交易日数据和无意义评论，得到 263 个交易日内的 104876 条评论数据。部分原始评论数据展示如表 3-1。

表 3-1 股吧评论原始数据

阅读	评论	标题	作者	最后更新
167	0	明天雄起了	朝花夕来	2021/1/29
146	0	天黑了，大家请闭眼，放松自己。	学而思思	2021/1/29
666	2	耐心等待底部走平，机会在后面。	慢慢的看多	2021/1/29
158	0	牛市正在进行中	趋势 v	2021/1/29
162	0	华龙一号并网度电	n1396687330a5b05	2021/1/29
.....	.....	.....	.....	.....
174	0	明天回到 3515 吧	富富先生	2020/1/2
359	0	字如千金，送给天下股民：做投资一定要路子对，“路子	点石钵满得陈家洛	2020/1/2
84	0	心理学分析，过年之前会涨，因为要给韭菜们吃口糖，好	多财善贾地袁承志	2020/1/2
277	0	本周操作：买入国恩股份 30.3，买入沪电股份 17.9，买入	独行小兵	2020/1/2
219	0	任何依靠公开信息炒股的行为都是不可取的，上周末公布	热板龙头顺主逆散	2020/1/2

在获得评论原始数据后，应用基于词典的情感分析方法计算每日情感值。该方法的步骤为首先构造构造词典；随后对具体的评论文本进行分词，并识别切分出的词语在情感词典中的成分；最后根据情感词典中不同词的权重计算每条评论的最终得分。由于自己构造词典工程量巨大，且不易包括全部词语，本研究选用的情感词典为知网 HOWNET 词典，该词典包括程度副词词典、否定词词典、情感词词典等部分。按照 HowNet 情感词典的分类方法，将程度副词分为“极其|extreme / 最|most”，“很|very”，“较|more”，“稍|-ish”，“欠|insufficiently”，对应的权重分别为 2.0，1.5，1.25，0.5，0.25，对于否定副词“不，不大，不必，没，没有……”，其权重为-1。具体的权重设置如表 3-2 所示。

文本预处理的方式，其中主要分为中文的编码、分词、过滤停用词。如图 3-1 使用 Python 中的 jieba 分词包对评论数据进行分词处理，设置正则表达式以过滤掉文本中的数字。将分词结果与情感词典对照，部分评论的情感值计算结果如表 3-4 所示。计算每日情感值，得到交易日情感值时间序列，部分展示如表 3-3 所示。

```
明天把剩下的钱转出来，免得给我亏完了[大哭]
去除停用词结果 ['明天', '剩下', '钱', '转', '免得', '亏', '完', '大哭']
情感得分：-1

股债收益率模型图，大家可以看看。目前已经和上两次牛
去除停用词结果 ['股债', '收益率', '模型', '图', '上', '两次', '牛']
情感得分：0

现在市场迫切需要改革实实在在举措且能够改变现状的长
去除停用词结果 ['市场', '迫切需要', '改革', '实实在在', '举措', '改变现状', '长']
情感得分：1.2

做了二十多年股票了，今天算开了眼了，4个板块在上涨
去除停用词结果 ['做', '二十多年', '股票', '算开', '眼', '4', '板块', '上涨']
情感得分：1.2
```

图 3-1 分词与计算实验界面

表 3-2 程度副词权重设置

程度副词	权重值
半点，不大，不甚，不怎么……	0.25
稍，有点，略，略微，稍微……	0.5
较，比较……	1.25
实在，特别，很，尤其，太……	1.5
极，十分，绝对，非常……	2.0
不，不大，不必，没，没有……	-1
没有程度副词的正向情感词	1
没有程度副词的负向情感词	-1
没有程度副词的正向情感词（句首、句尾）	1.2
没有程度副词的负向情感词（句首、句尾）	-1.2

表 3-3 情感值时间序列

date	sentiment
1/2/2020	0.18
1/3/2020	0.16
1/6/2020	0.07
1/7/2020	0.17
1/8/2020	0.09
.....	.....
1/25/2021	0.14
1/26/2021	0.01
1/27/2021	0.1
1/28/2021	-0.01
1/29/2021	0.02

表 3-4 部分评论情感值计算

标题	作者	最后更新	分词结果	情感得分
明天雄起了	朝花夕来	2021/1/29	明天 雄起	1.2
天黑了，大家请闭眼，放松自己。	学而思思	2021/1/29	天黑 请 闭眼 放松	0
耐心等待底部走平，机会在后面。	慢慢的看多	2021/1/29	耐心 等待 底部 走平 机会	1.2
牛市正在进行中	趋势 v	2021/1/29	牛市 中	0
华龙一号并网度电	n1396687330a5b05	2021/1/29	华龙 一号 网度 电	0
.....	.....	.....	.....	.....
明天回到 3515 吧	富富先生	2020/1/2	明天 回到 3515	0
字如千金，送给天下股民：做投资一定要路子对，“路子	点石钵满得陈家洛	2020/1/2	字 千金 送给 天下 股民 做 投资 路子 路子	0
心理学分析，过年之前会涨，因为要给韭菜们吃口糖，好	多财善贾地袁承志	2020/1/2	心理学 分析 过年 会涨 韭菜 吃 口 糖 好	2
本周操作：买入国恩股份 30.3，买入沪电股份 17.9，买入	独行小兵	2020/1/2	本周 操作 买入 国恩 股份 30.3 买入 沪 电 股份 17.9 买入	2
任何依靠公开信息炒股的行为都是不可取的，上周末公布	热板龙头顺主逆散	2020/1/2	公开 信息 炒股 都 不 可取 上周末 公布	1

3.2 股票综合指标的建立

以 Jupyter Notebook 为平台，Python 为语言工具，通过 Tushare 金融股票数据接口连接到东方财富网，爬取上证指数（股票代码 000001）在 2020 年 1 月 1 日至 2021 年 1 月 31 日的股票历史数据，选取开盘价、最高价、最低价、收盘价、成交量、成交额 6 个数值型指标，去除非交易日数据，得到 263 个交易日的数据。部分原始数据如表 3-5 所示。

由于这些指标的量纲并不统一，为避免指标量纲不均带来的计算误差，因此首先采用 Z-Score 对数据进行标准化处理，将数据变为 (-1, 1) 之间的数值。处理公式如下：

$$x' = \frac{x - \mu}{\sigma} \tag{8}$$

其中 x' 为处理后的数据，x 为原始数据，μ 为原始数据均值，σ 为原始数据标准差。

表 3-5 部分上证指数原始数据

date	close	high	low	open	volume	amount
2020/1/2	3085.1976	3098.1001	3066.3357	3066.3357	292470208	3.27E+11
2020/1/3	3083.7858	3093.8192	3074.5178	3089.022	261496667	2.90E+11
2020/1/6	3083.4083	3107.2032	3065.3088	3070.9088	312575842	3.31E+11
2020/1/7	3104.8015	3105.4507	3084.329	3085.4882	276583111	2.88E+11
2020/1/8	3066.8925	3094.2389	3059.1313	3094.2389	297872553	3.07E+11
.....	.....	.....	.....	.....	.....	.....
2021/1/25	3624.2381	3637.1002	3591.0204	3605.3611	327341070	5.27E+11
2021/1/26	3569.429	3610.9686	3564.7415	3610.9686	278139884	4.36E+11
2021/1/27	3573.3412	3578.7968	3546.492	3567.5485	264107161	3.96E+11
2021/1/28	3505.1759	3549.544	3496.8788	3534.6685	270862461	3.92E+11
2021/1/29	3483.0692	3531.5981	3446.5471	3521.7175	293662664	4.17E+11

处理后的数据均值为 0，标准差为 1。标准化处理后 6 个变量的时间序列如

图 3-2 所示。由图可以看出，Z 标准化处理后的数据的范围均在-1 和 1 之间，可以从图上较为直观地看出各组数据虽然数值上存在差异，但变化趋势相似，图像的轨迹大部分相同。由此可见，这些数据间隐藏着差异化的信息，但也存在着重叠的部分。使用降维方法对数据进行降维处理，能够去除数据重叠、冗余的部分，使数据的使用更加高效。

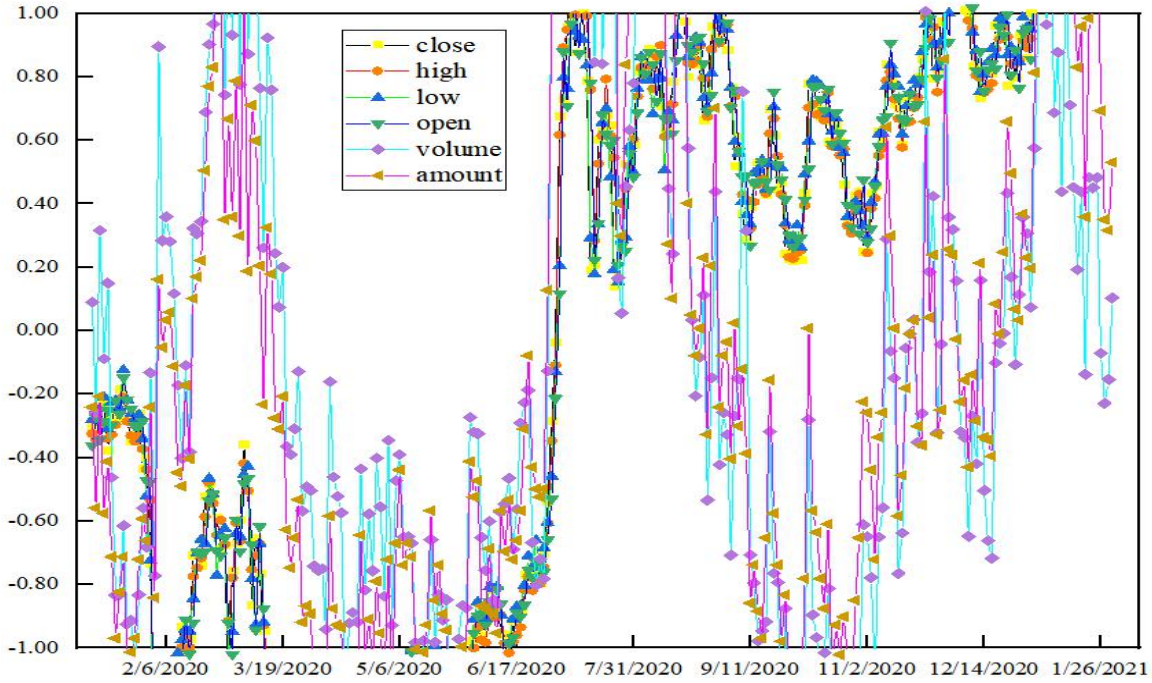

图 3-2 股票数据时间序列

表 3-6 相关性系数矩阵

		close	high	low	open	volume	amount
相关性	close	1	0.996	0.996	0.991	0.316	0.529
	high	0.996	1	0.996	0.996	0.336	0.548
	low	0.996	0.996	1	0.997	0.283	0.5
	open	0.991	0.996	0.997	1	0.3	0.516
	volume	0.316	0.336	0.283	0.3	1	0.944
	amount	0.529	0.548	0.5	0.516	0.944	1

借助 SPSS 软件对这 6 个股票数值型变量进行因子分析，通过计算得出主成分的方式对数据进行降维处理，以得出股票综合指数。首先构造相关系数矩阵，

并通过相关系数矩阵进行 KMO 检验和巴特利球形度检验的计算。KMO 检验和巴特利球形度检验用于检验数据间的相关性，从而判断数据是否适合进行因子分析。相关性系数矩阵、KMO 检验和巴特利球形度检验的结果分别如表 3-6、表 3-7 所示。

表 3-7　KMO 检验和巴特利球形度检验结果

KMO 取样适切性量数		0.683
巴特利特球形度检验	**近似卡方**	5112.496
	自由度	15
	显著性	0.000

结果显示，KMO 检验的测度为 0.683，巴特利球形度检验的近似卡方值为 5112.496，自由度为 15，显著性为 0.000，变量之间相关性较强，拒绝了原假设，可以使用主成分分析法。主成分分析的结果如表 3-8 所示。

表 3-8　主成分分析结果

	u_i	α_i（%）	b_i（%）
y_1	4.541	75.678	76.175
y_2	1.42	23.67	23.826
y_3	0.027		
y_4	0.009		
y_5	0.003		
y6	0.001		

根据第一次主成分分析的结果得到选取特征值大于 1 的前两个变量$[y_1, y_2]$作为主成分。根据上述结果得到股价综合指标的计算公式为：

$$Score = 76.175\% \times y_1 + 23.826\% \times y_2 \tag{9}$$

3.3 联动模态的识别

设 p_t 为和 p_{t-1} 分别为第 t 天和第 $t-1$ 天的股票指数和情感得分，则第 t 天相对于第 $t-1$ 天的波动值为

$$\Delta p = p_t - p_{t-1} \tag{10}$$

波动值为正表示今天的股票指数或情感值相对于昨天有正向变化，波动值为正表示今天的股价指标或情感值相对于昨天有负向变化，波动值为 0 则表示今天值相对于昨天无变化。用此方法计算得到股票指数和情感分在交易日的波动时间序列如图 3-3 和图 3-4。

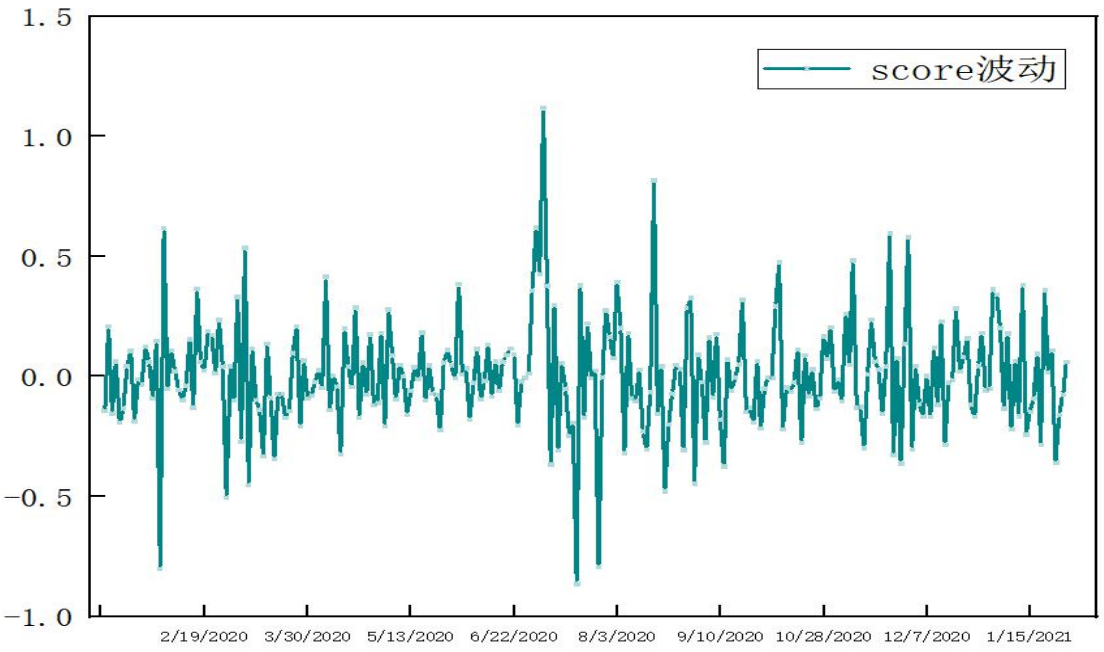

图 3-3 股票指数波动时间序列

为避免虚假的相关影响对二者联动模态的判断，需首先保证时间序列的平稳性。使用 ADF 根检验验证两个时间序列的平稳性，结果显示股价指数波动时间序列和情感波动时间序列均在 1%水平下显著，证明序列平稳。

我们通过皮尔逊相关系数检验上证股票指数波动序列和投资者情感波动序列的相关性。皮尔逊相关系数用于度量两个变量之间的相关性，是判断两组数据与某一直线拟合程度的一种度量，被广泛应用于股票价格和其影响因素之间的相

关性检验。其值介于-1 与 1 之间，值越大则说明相关性越强。

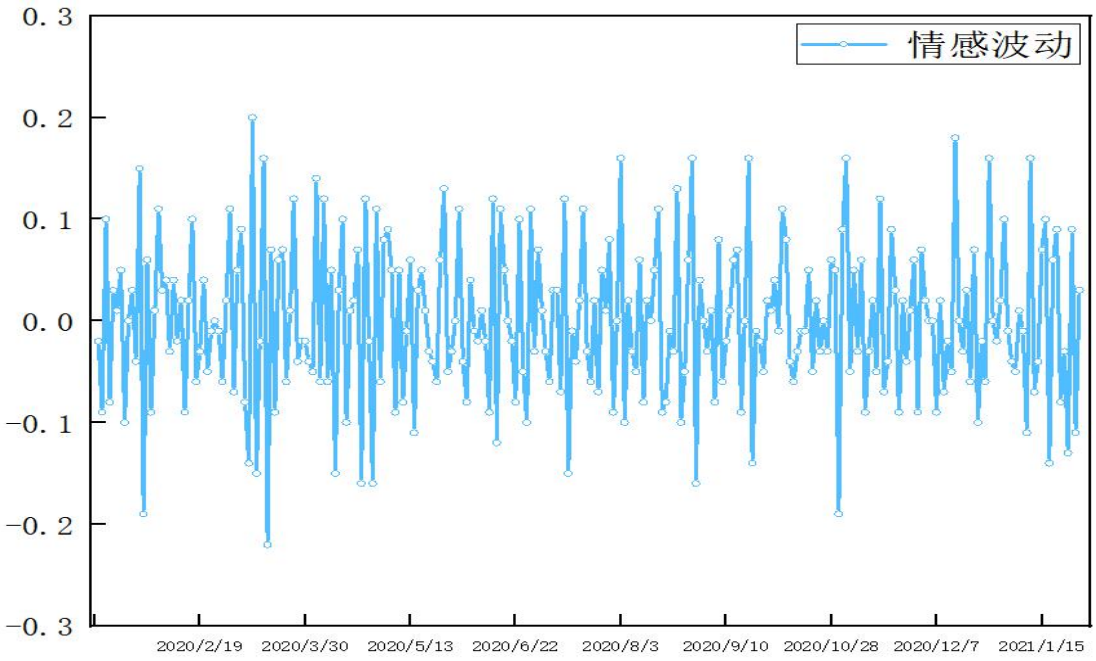

图 3-4 情感波动时间序列

它在数据不是很规范或相对于平均水平偏离很大时，会倾向于给出更好的结果。如果某组数据总是倾向于给出比另一组数据高的数值，而二者的分差又始终保持一致，也即二者呈线性关系，则会得到较高的皮尔逊相关度。两个变量之间的皮尔逊相关系数定义为两个变量之间的协方差和标准差的商

$$\rho(x,y) = \frac{\mathrm{cov}(x,y)}{\sigma(x)\sigma(y)} \qquad (11)$$

其中，$\mathrm{cov}(x,y)$表示两组数据 x 和 y 的协方差，$\sigma(x)$和$\sigma(y)$分别表示两组数据的方差。当两个变量的标准差都不为零时，相关系数才有定义，皮尔逊相关系数适用于两个变量之间是线性关系且都是连续数据时，要求两个变量的总体是正态分布，或接近正态的单峰分布，且两个变量的观测值是成对的，每对观测值之间相互独立。通过上述公式在 SPSS 软件中计算出皮尔逊相关系数为 0.4，证明股价波动序列和投资者情绪波动序列间有一定的正相关。

在获得波动时间序列后，我们使用粗粒化方法描述两个变量间的联动性。粗粒化方法的本质是省略样本不重要的细节，促进局部整体特征的研究[39]。我们运用粗粒化定义 3 个联动性符号，分别是 Y，O，N。符号 Y 表示股票指数的波动

与情感波动同向联动，符号 N 表示股票指数的波动与情感波动反向联动，符号 O 表示股价指数的波动方向与情感波动方向无联动性。

$$l=\begin{cases}Y, \Delta p_{score} * \Delta p_{senti} > 0 \\ O, \Delta p_{score} * \Delta p_{senti} = 0 \\ N, \Delta p_{score} * \Delta p_{senti} < 0\end{cases} \quad (12)$$

Δp_{score} 代表股票指数波动，Δp_{senti} 代表情感波动。进行粗粒化处理后，我们将原本的两条波动时间序列变为一条联动性时间序列 L，即 $L=[l_1, l_2 \cdots, l_n](l_1, l_2 \cdots, l_n \in Y, N, O)$。一般情况下，股票每周的交易时间为 5 天，因此我们以连续的 5 个交易日为一个时间窗口，以 1 天为步长将整条联动时间序列分割为连续的滑动时间窗口，每个窗口的窗口期均为 5 天，并以这 5 天的联动性符号为一个组合，定义为一个窗口期下的联动模态。例如，2020 年 1 月 13 日到 17 日中的 5 个交易日的联动性符号分别为 Y，N，O，N，Y，则这一窗口的联动模态为 YNONY。应用这一方法，将整条联动性序列中的 262 个联动符号分割为 258 个连续的联动性模态。联动性模态识别过程如表 3-9。理论上所有联动性符号应总共可以构成 243 种不同的联动性模态，实际上本研究中联动性模态序列只包含 75 个不同的联动模态。

表 3-9 联动性模态识别

日期	联动性符号	联动模态
2020-01-13	Y	
2020-01-14	N	
2020-01-15	O	
2020-01-16	N	
2020-01-17	Y	YNONY
2020-01-20	Y	NONYY
2020-01-21	N	ONYYN

3.4 复杂网络模型的建立

本研究的核心部分在于建立情感-股指联动模态复杂网络模型，根据联动网络模型的拓扑指标对新冠疫情背景下的我国股市投资者情感与股票走势之间的联动关系进行分析与探究。根据 3.3 章节得出的联动模态和模态之间的关系，以不同的联动模态为节点，联动模态间是否存在随时间的连续转化关系定义是否存在边，在建立复杂网络模型，该模型中包含 75 个节点和 119 条边。应用 gephi 进行网络模型的可视化，结果如图 3-5 所示。初步观察该网络模型图可知，网络中节点间关联较为稀疏，存在着部分边的权重较大。根据公式计算该网络模型的拓扑指标，选取节点的加权出度、特征向量中心度、紧密中心度、中介中心度和模块化系数 5 个指标的计算结果，以备下一章的分析。

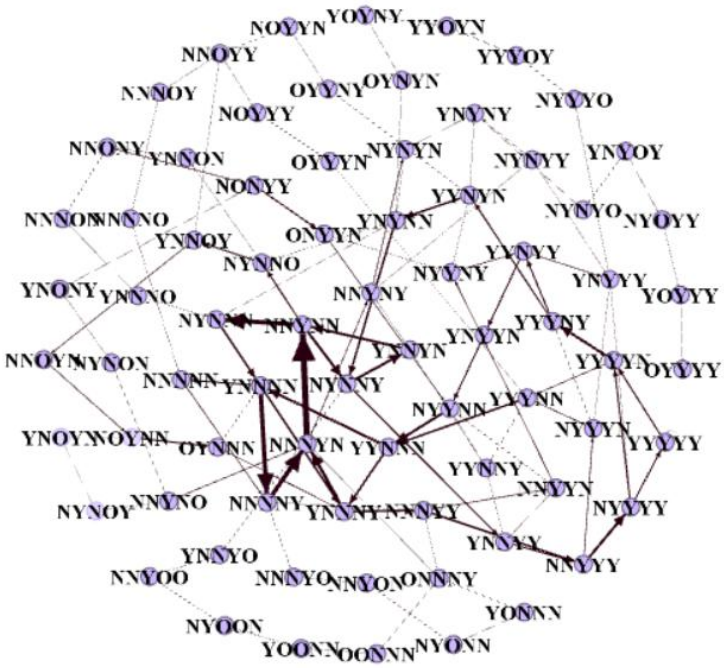

图 3-5 联动模态网络

3.5 本章小结

本章基于复杂网络方法初步构建了一个情感-股票指数波动联动模态网络，并对构建的过程进行了解释。包括通过文本情感分析计算情感值，主成分分析构建上证综合股票指数，以及如何运用粗粒化方法切割联动模态。

4 分析与结果

在本节，我们通过加权出度与特征向量中心度识别了投资者情绪与上证股票指数变化联动的关键模态，并通过边的权重排名确定网络主要传输路径。最后，我们分析了网络的模块化程度。

4.1 关键模态识别

为了确定情感与股票指数的主要联动关系，我们需要对关键模态进行识别。在复杂网络中，节点的加权出度表示节点向其他节点传导的频率。在情感-股指联动模态网络中，当一个节点的加权出度大，表明由这种模态向其他模态转换的次数多，因而这种模态在网络中总体的出现频率更高，地位更重要。表 4-1 列出了联动模态的加权出度及其累积分布排名。模式“YYYYN”出现的概率最大，这表明在连续的 5 个工作日内投资者的情感与股票指数均同向联动的概率最大。加权出度排名第二的模式和排名第三的模式分别是“YYYYY”和“NYYYY”，综合表明前三的模态来看，投资者情感与股指联动在多数交易日内均有同向联动。图 4-1 为加权出度累积百分比随节点排序名次的累计百分比的分布。图 4-1 显示，加权出度排名前 15 的模态，其加权出度累积占所有模态的达 53%，说明这些模态在情感股指联动网络中最活跃，而这些模态仅占所有模态的 20%，说明模态活跃度分布并不均匀，少数的联动模态主导着更多的交易日，而多数联动模态转换并不活跃。将所有加权出度值与其在所有模态中出现的概率放入双对数坐标如图 4-2 所示，二者存在 $y=-1.1603x-1.085$ 的线性回归关系，说明联动性网络整体服从幂律分布。

紧密中心度体现了节点到达其他节点的容易程度，中介中心度高的节点则更容易成为网络中的节点枢纽，因此紧密中心度和中心度高的节点，在网络中的地位也更加重要。图 4-3 和图 4-4 为联动模态的紧密中心度和中介中心度累积分布。由图可知，部分模态的紧密中心度和中介中心度值较高，40%左右的节点的紧密中心度和中介中心度累积占比达到约 80%，说明少数模态在网络中充当枢纽，成为一个模态向另一个模态过渡的中介，也说明网络中可能存在着部分社团，社团中的模态以某些节点为中心形成群簇，从而在各模态之间不断传导、演化。通过对网络的紧密中心度和中介中心度进行排序比较可知，在两个拓扑指标的排名前

三中均存在的节点为“YYYNY”。该节点的紧密中心度为所有节点中最高的 0.226，中介中心度为 1301.83，排在所有节点中的第二位。说明该节点可能成为网络的核心枢纽，其它模态围绕这一联动模态进行演化。该模态中 4 个工作日为情感与股票同向联动，进一步说明在新冠疫情背景下投资者的情感与股市趋势有相同的波动方向。

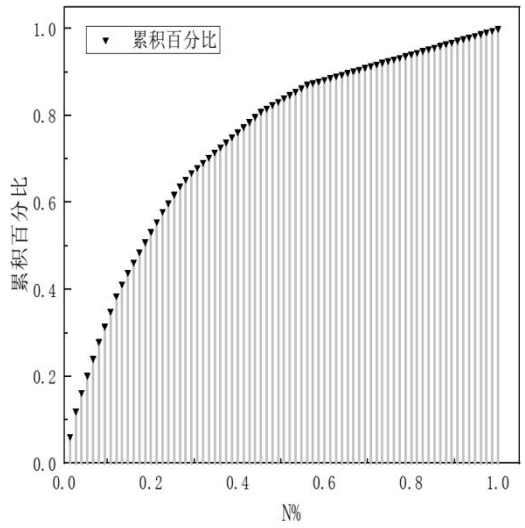

图 4-1 加权出度累积百分比

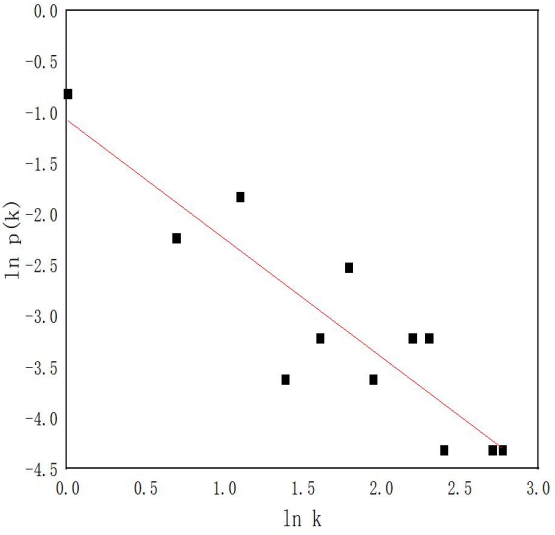

图 4-2 加权出度概率分布

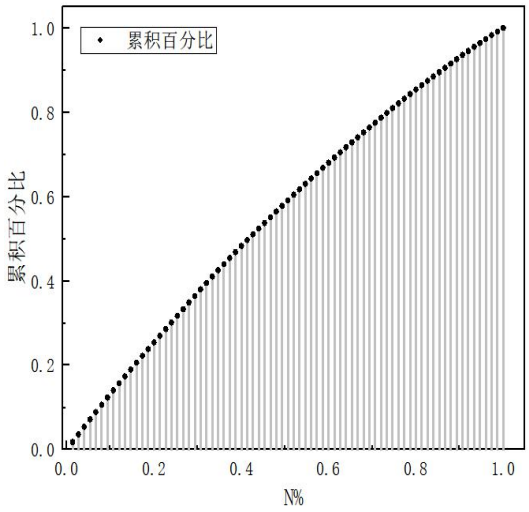

图 4-3 紧密中心度累积分布

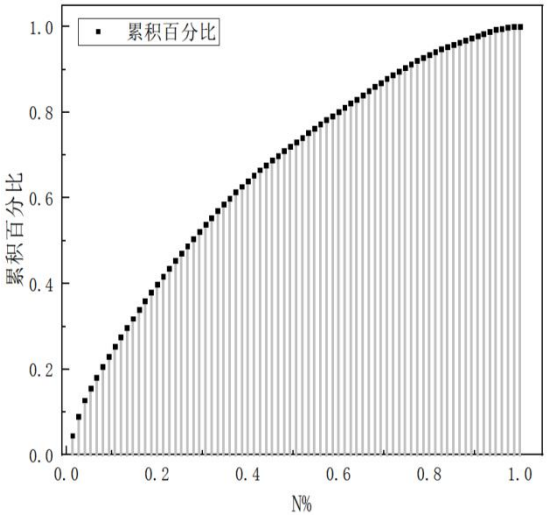

图 4-4 中介中心度累积分布

特征向量中心度可以用来找到系统中的核心参与者，特征向量中心度高的联动模态，其是网络中核心的潜在可能性更高。为了识别投资者情感与股票指数关联波动的关键模式，我们综合加权出度与特征向量中心度两个指标进行搜寻。如图 4-5 所示，该图的横坐标为模态的加权出度，纵坐标为模态的特征向量中心度，

二者围成面积大的点，其二者相乘之积更大，是网络中重要的联动模态。选取其中排名前三的模态如表 4-2。

表 4-1 加权出度及其累积分布

模态	排名	加权出度	累积百分比
YYYYN	1	16	6.23%
YYYYY	2	15	12.06%
NYYYY	3	11	16.34%
NNNNY	4	10	20.23%
YNNNY	5	10	24.12%
NNNYN	6	10	28.02%
YYNNN	7	9	31.52%
NYNNY	8	9	35.02%
NYNNN	9	9	38.52%
NNYYY	10	7	41.25%
NNYNN	11	7	43.97%
NYYNN	12	6	46.30%
YNYNN	13	6	48.64%
YNNYN	14	6	50.97%
NNNYY	15	6	53.31%

从表 7 可知，新冠疫情期间最关键的三个组合联动模态分别为“YYYYN”、“YYYYY”和“NYYYY”，表明这三个模态在网络中发挥着传动中枢的作用，相比于其它模态更加活跃，是投资者情感和股票指数联动性的核心模式。单模态而言，出现最多的为“Y”，即情感与股价同步波动，占所有模态的 52.29%，这也验证了之前的结论，即多数交易日内疫情期间投资者的情感与股票指数同方向变化。“O”模态出现次数仅占 4.96%，说明大多数交易日内，无论是正向或反向，

投资者情感与股票指数均存在波动，仅少数交易日内二者无关。

表 4-2 关键模态

排名	组合模态	加权出度	特征向量中心度	模态	次数	频率
1	YYYYN	16	0.979908	Y	137	52.29%
2	YYYYY	15	1	N	112	42.75%
3	NYYYY	11	0.996438	O	13	4.96%

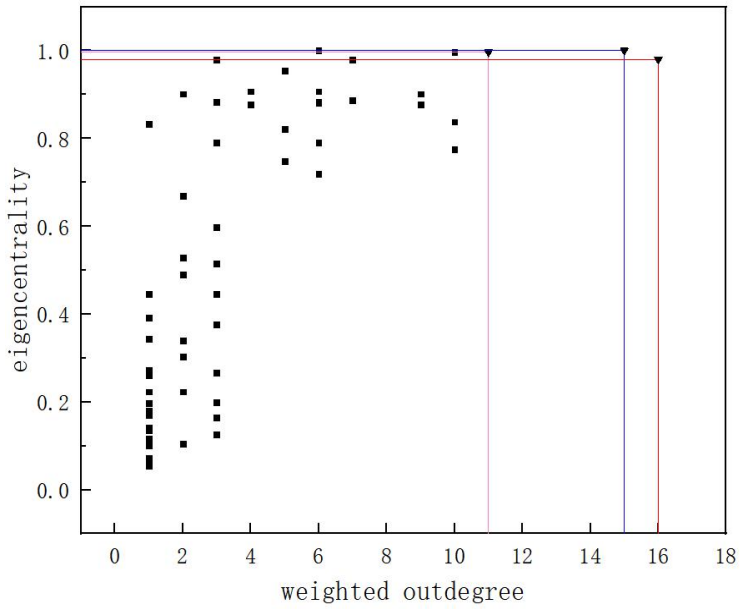

图 4-5 特征向量中心度与加权出度分布

4.2 主要传导路径识别

主要传导路径是网络中节点之间传递和联结最频繁的路径，该路径是网络中节点传输的主要方向。在情感-股指联动网络模型中，识别出节点之间的主要传导路径是我们判断联动模态波动方向，进而确定在新冠疫情背景下我国股市的投资者们的情感与股票市场的实际走势之间的关联关系的核心步骤。情感与股指关联波动网络是有向加权网络，在该网络中，若两节点间存在一条边，则两节点的距离为 1，网络的直径为距离最远的两节点间的距离，网络的平均最短路径为任意两节点间距离的均值。网络的直径和平均最短路径反映网络的连接度和稀疏程度，是研究网络物理拓扑结构的重要指标。情感与股指联动网络的网络直径为 15，平均最短路径为 6.313，说明网络总体较为稀疏，模态间的转换需要经过较多路径。为了进一步研究网络的物理结构，需识别网络的主要传导路径。主要传

输路径是网络中传输最频繁的路径，即联动模态最有可能沿着这一路径进行转换。在情感-股指网络中，这代表情感和股指之间的联动关系最可能朝着某一方向波动传递。在复杂网络中，边的权重表示两个节点之间的连接频率，即两个节点之间传输的概率较大。因此，可以通过对联动网络的边权重进行比较的方式来确定主要的传输路径。我们对情感股指联动网络的边的权重进行排名，结果如表 4-3.

取边权重排名的前十名，其排名结果显示了在网络中情感与股票指数的联动更容易通过这些路径进行转换。对上述所选取的权重排行前十的边，我们基于其中节点和边的传输关系重新构建了一个子网络，在该子网络中去除了与网络中大多数节点独立的边和节点，得到该网络的网络主体，该主体部分共包含 7 条边和 7 个节点，如图 4-6 所示。我们发现这个子网络主体并非闭环的，在其中存在 4 条路径，分别是“NYNNY--YNNYN--YYYYN--YYYYY--YNNNY”、“NYNNY--YNNYN--YYYYN--YYYYY--NNNNY”、“NYNNN--YYYYN--YYYYY--YNNNY”、“NYNNN--YYYYN--YYYYY--NNNNY”。这说明在网络中联动模态更可能直接或间接地沿着这些路径进行传输，即这些路径是大部分路径的真子集。这 4 条路径的共有部分为“YYYYN--YYYYY”，说明这条路径为联动模态演变的最主要路径，这也与节点加权出度和边权重的排名相符合。在这条路径中，包括了 4.1 节中的三个关键模态中的两个，说明股价和情感主要以同向联动为主，并在少数交易日反向交叉，印证了情感与股票市场之间的同向联动性，但也体现了情感会在某一时刻具有变化和波动性。

在子网络中，除了上述核心部分，还存在有单独的节点或者由少部分节点形成的路径，观察可知，包括“NNNNY--NYYYY”和“YNNNN--YYNNN--NYYNN”两条路径和一个单独的节点“NNNYN”。在上述联动模态中，“N”模态，即投资者的情感与股市走势相反占 63.33%，为大多数。由此可知，此次疫情中，投资者情感与股市的发展趋势也在一部分时间相反。这些情感与股市情绪的不一致可能来自于时间上的滞后性影响或考虑少数偶然性情况的影响。在 2020 年的某些时间，部分地区疫情出现短暂起伏或反弹，这些小范围事件也会引起股民情绪短暂的波动。这些情感与股市走势相反的时间虽然仍在主要路径上，但大多在小范围内引起，并没有形成核心的子网络，这也进一步表明投资者情感与股市的发

展方向在多数情况下是可预见的，二者更倾向于保持相同的发展方向，即突发性社会危机所引起的舆论情感的正负情况对于金融投资市场的正负走势具有直接或间接地引导作用。

表 4-3 边权重排名

排名	起点	终点	权重
1	YYYYY	YYYYN	11
2	NNNNY	YYYYY	8
3	NYYYY	NNNNY	8
4	YYYYN	NYNNN	8
5	NNNYN	NNNYN	7
6	YNNNY	YYYYY	6
7	NYYNN	YYNNN	5
8	YYNNN	YNNNN	5
9	NYNNY	YNNYN	5
10	YNNYN	YYYYN	5

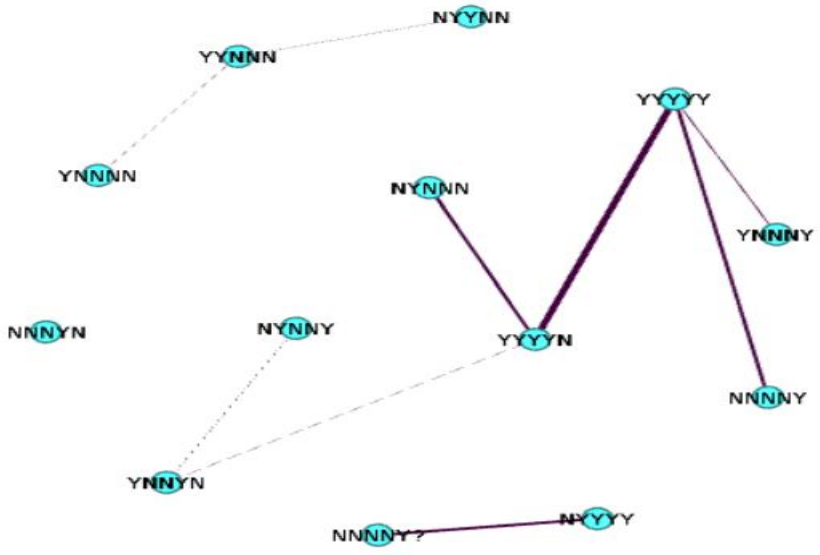

图 4-6 主要传输路径网络

4.3 网络社团化分析

对联动网络的网络社团化分析，是对复杂网络物理结构研究的重要内容，可以帮助我们对网络的传递的稳定性进行梳理。模块化系数描述网络中联结较为紧

密的点簇，如果一组节点拥有相同的模块度，则可能在同一社团。通过对模块度进行分析，来对我们所构建的情感-股指联动网络模型进行网络社团化分析。

图 4-7 中展现了联动模态的模块度分布，可直观看出，图中共有从 0 到 8 共 9 个节点团簇，因此网络可大致被分割为 9 个社区。在这些社区内部的节点联系更加紧密，而社区之间的节点联结则相对疏松。每个社团内的联动模态数量如表 4-4。

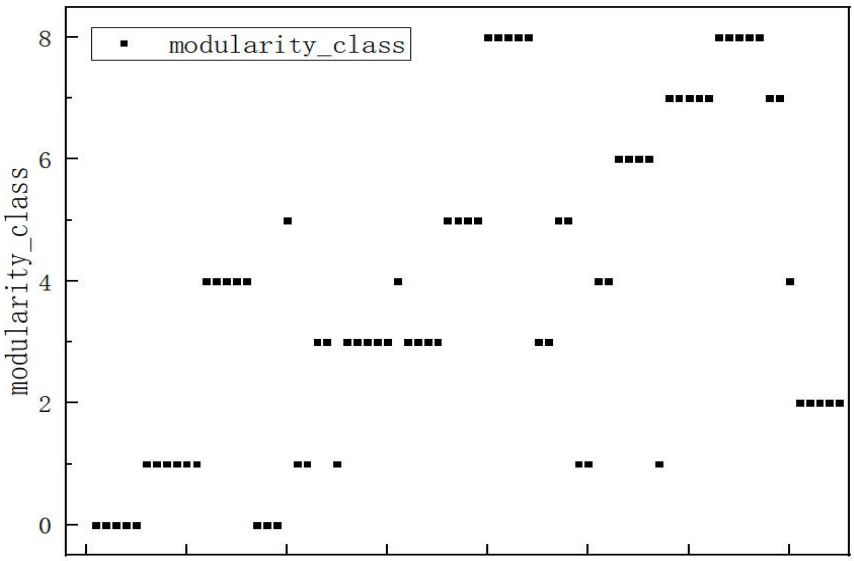

图 4-7 模块度分布

表 4-4 模块内联动模态数量

模块	联动模态数量
0	8
1	12
2	5
3	13
4	9
5	7
6	4
7	7
8	10

由表可知，模块 2 和模块 6 内的联动模态较少，这两个模块内部的节点相对

来说与网络整体的联系较为薄弱，可能属于悬挂节点。大部分联动模态集中在了模块 1、模块 3 和模块 8 内，在这三个模块内的节点则具有更加紧密的关系。模块 3 内包含有 13 个联动模态，为网络内最大的社团。通过对模块 3 内部的联动模态具体分析可知，其中大部分联动模态均为网络中的关键节点，如“YYYYY”、“YYYYN”，联动模态网络可能以这些模态为核心，在这些模态之间更加频繁地转换。该模块内的模态“Y”占比达到 67.7%，这也再次证实了在新冠疫情期间，投资者情感与股票市场的走势大部分情况下为同向波动。模块 1 中模态“N”占比 53%，说明投资者情感与股市走势相反的情况集中在某一社区内部出现。

4.4 本章小结

本章是论文的核心部分。本章选取了 5 个复杂网络拓扑指标，即加权出度、特征向量中心度、紧密中心度、中介中心度、模块化系数。根据这 5 个拓扑特征的数值，对情感-股票指数联动网络进行了网络特征分析，并识别了关键路径和关键模式，发现了隐藏在网络中的核心子网络。本章内容也是得出本文后续结论的基础。

5 结论

本文采用复杂网络方法，以上证综合指数为例，研究了新冠疫情视角下的我国股市投资者情绪与市场形势的联动模式。在具体实验中，首先，爬取上证指数在 2020 年 1 月 1 日至 2021 年 1 月 31 日的股票数据，应用主成分分析法实现对开盘价、最高价、最低价、收盘价、成交量、成交额等股票数值型指标的降维，计算得到股票综合指数；随后，使用 Python 爬虫技术爬取东方财富股吧的上证指数吧股民评论，在对文本进行清洗、分词等预处理后，使用基于词典的情感分析法计算每日的股吧评论情感值。在这之后，采用粗粒化方法，以 5 个工作日为一组，实现对情感、股票指数的联动模态构造。应用联动模态之间的传递关系，构建情感-股指联动模态复杂网络模型，进行网络可视化处理，选取并计算网络的加权出度、特征向量中心度、紧密中心度、中介中心度、模块化系数等拓扑指标，识别了网络中的关键联动模态，确认了主要传输路径，并分析了网络的社团化特征。经过上述工作，发现了以下结果：

（1）股票指数与情感关联模态的加权出度分布符合幂律分布，主要的联动模态仅为几种。有 20%加权出度累积占所有模态的 53%，说明少数的联动模态其重要作用，而大部分模态并不活跃。不活跃的模态间或在网络中出现，说明在疫情期间，投资者的情感与股价波动起伏，二者关联模式并非平稳不变，而是既有同向联动，也有反向联动或无联动。综合考虑新冠疫情背景，上述现象的原因可能由于疫情的反复起伏。

（2）紧密中心度和中介中心度体现了节点的在网络中的枢纽地位，分析联动模态的紧密中心度和中介中心度累积分布可知，部分模态的紧密中心度和中介中心度值较高，这些模态在网络中充当枢纽，成为一个模态向另一个模态过渡的介质，说明网络中可能存在着以某一节点为中心形成群簇。这说明疫情期间，某一种情感与股市的联动关系占主导地位。

（3）综合加权出度和特征向量中心度，对联动模态进行搜索，发现二者乘积之和最高的三个联动模态分别为“YYYYN”、“YYYYY”和“NYYYY”，表明这三个模态在网络中地位尤其重要，发挥着传动中枢的作用，是体现投资者情绪和股市总体行情联动性的关键模态。出现最多的模态符号为“Y”，占所有模态的

52.29%，“O”模态出现次数仅占 4.96%，说明大多数交易日内，投资者情绪与股指同向波动，情绪的正向或负向变化也会引起股价的升或降，但联动模态也并非一成不变，而是在一段时间序列中存在部分波动段。仅有极少数交易日内情感与股票指数无关。

（4）情感与股价联动网络的网络直径为 15，平均最短路径为 6.313，说明网络总体较为稀疏，模态间的转换需要经过较多路径。对情感股指联动网络的边的权重进行排名，结果显示了在网络中情感与股票指标的联动更容易通过某些路径进行转换。取权重前十的边，基于其中节点和边的传输关系形成一个字网络，子网络中存在 4 条路径，分别是“NYNNY--YNNYN--YYYYN--YYYYY--YNNNY”、“NYNNY--YNNYN--YYYYN--YYYYY--NNNNY”、“NYNNN--YYYYN--YYYYY--YNNNY”、“NYNNN--YYYYN--YYYYY--NNNNY”，4 条路径的共有部分为这条路径为联动模态演变的最主要路径。在这条路径中，包括了之前识别的三个关键模态中的两个，说明疫情股价和情感主要以同向联动为主，并在少数交易日反向。

（5）通过对模块度进行分析来进行网络社团化分析。网络可大致被分割为 9 个社区。在这些社区内部的节点联系更加紧密，而社区之间的节点联结则相对疏松。大部分联动模态集中在了三个模块内，其中模块 3 内包含有 13 个联动模态，为网络内最大的社团。模块 3 内部的大部分联动模态均为网络中的关键节点，如“YYYYY”、“YYYYN”，联动模态网络可能以这些模态为核心，在这些模态之间更加频繁地转换。该模块内的模态“Y”占比达到 67.7%，这也再次证实了在新冠疫情期间，投资者情感与股票市场的走势大部分情况下为同向波动。

（6）在新冠疫情期间，受突发性社会事件的新闻和网上论坛的舆论影响，我国股市股中民的情绪出现不同程度的起伏，并对股票的价格、成交量等指标产生影响。二者的联动性呈现同向为主、部分反向、少数模态在关键模态附近波动起伏的现象，说明疫情期间可根据股民情绪对股指进行监控，进行适当的市场调控。本研究建立了股指与股民情感联动的复杂网络模型，并提出一些分析方法，期望在未来为由突发性社会事件引起的股民情绪和能源股价变化提供预警机制，并期待不断完善研究。

参考文献

[1] 东方财富网. 数说 2020 年 A 股：总市值近 80 万亿全年募资总额 1.36 万亿元[EB/OL]. (2021-01-13)[2021-05-01].https://baijiahao.baidu.com/s?id=1688774366251464429&wfr=spider&for=pc.

[2] Nasreen S, Tiwari AK, et al. Dynamic connectedness between oil prices and stock returns of clean energy and technology companies[J]. JOURNAL OF CLEANER PRODUCTION. 2020, 260.

[3] He LT, Casey KM. Forecasting ability of the investor sentiment endurance index: The case of oil service stock returns and crude oil prices[J]. ENERGY ECONOMICS. 2014, 47:121-128.

[4] 陈赟,沈艳,王靖一.重大突发公共卫生事件下的金融市场反应[J].金融研究.2020,(06):20-39.

[5] Yi JH, Niblack W. Sentiment mining in Web Fountain[J]. IEEE International Conference on Data Engineering. 2005,1073-1083.

[6] Matsumoto S, Takamura H, Okumura M. Sentiment classification using word sub-sequences and dependency sub-trees[J]. Lecture Notes in Artificial Intelligence. 2005, 3518:301-311.

[7] Abbasi Ahmed, Chen Hsinchun, Salem Arab. Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums[J]. ACM TRANSACTIONS ON INFORMATION SYSTEMS. 2008, 26(3).

[8] Gamon M, Aue A, Corston-Oliver S, et al. Pulse: Mining customer opinions from free text[J]. Lecture Notes in Computer Science. 2005, 3646: 121-132.

[9] Archak Nikolay, Ghose Anindya, Ipeirotis Panagiotis G. Show me the Money! Deriving the Pricing Power of Product Features by Mining Consumer Reviews[J]. KDD-2007 PROCEEDINGS OF THE THIRTEENTH ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING. 2007,56-65.

[10] 赵妍妍,秦兵,刘挺.文本情感分析[J].软件学报.2010,21(08):1834-1848.

[11] 梁军,柴玉梅,原慧斌,高明磊,畅红英.基于极性转移和 LSTM 递归网络的情感分析[J].中文信息学报.2015,29(05):152-159.

[12] 赵春艳,王丽萍.基于网络文本分析的旅游体验感知研究——以贵州西江千户苗寨为例

[J].湖北理工学院学报(人文社会科学版).2021,38(02):7-12+19.

[13] 许欣,余杉.基于 BERT 与 Focal Loss 的电商平台评论情感研究[J].仪器仪表用户.2021,28(03):26-29.

[14] Baker Malcolm, Wurgler Jeffrey. Investor sentiment and the cross-section of stock returns[J]. JOURNAL OF FINANCE. 2006, 61(4): 1645-1680.

[15] Narayan PK, Bannigidadmath D. Financial News Predict Stock Returns? New Evidence from Islamic and Non-Islamic Stocks[J]. PACIFIC-BASIN FINANCE JOURNAL. 2017, 42: 24-25.

[16] 许启发,伯仲璞,蒋翠侠.基于分位数 Granger 因果的网络情绪与股市收益关系研究[J].管理科学.2017,30(03):147-160.

[17] Siganos A, Vagenas-Nanos E, ET AL. Facebook's daily sentiment and international stock markets[J]. JOURNAL OF ECONOMIC BEHAVIOR & ORGANIZATION. 2014, 107: 730-743.

[18] 易洪波,赖娟娟,董大勇.网络论坛不同投资者情绪对交易市场的影响——基于 VAR 模型的实证分析[J].财经论丛.2015,(01):46-54.

[19] 董理,王中卿,熊德意.基于文本信息的股票指数预测[J].北京大学学报(自然科学版).2017,53(02):273-278.

[20] Liu Ling, Wu Jing, et al. A social-media-based approach to predicting stock comovement[J]. EXPERT SYSTEMS WITH APPLICATIONS. 2015, 42(8): 3893-3901.

[21] 韩非,肖辉.中美股市间的联动性分析[J].金融研究.2005(11):117-129.

[22] 周珺.我国大陆股票市场与周边主要股票市场的联动分析[J].企业经济.2007(01):165-167.

[23] Mensi W, Hkiri B, et al. Analyzing time frequency co-movements across gold and oil prices with BRICS stock markets: A VaR based on wavelet approach[J]. INTERNATIONAL REVIEW OF ECONOMICS & FINANCE. 2018, 54:74-102.

[24] Bashir U, Yu, YG, et al. Do foreign exchange and equity markets co-move in Latin American region? Detrended cross-correlation approach[J]. PHYSICA A-STATISTICAL MECHANICS AND ITS APPLICATIONS. 2016, 462: 889-897.

[25] 余秋玲.投资者情绪与股价联动——基于 A 股市场的面板数据分析[J].西南交通大学学报(社会科学版),2015,16(02):109-117.

[26] 钟熙维,吴莹丽.新冠肺炎疫情下全球股票市场的联动性研究[J].工业技术经济,2020,39(10):29-37.

[27] 董志良,杨巧然.国际粮食贸易网络鲁棒性分析[J/OL].当代经济管理.2021,04(21):1-11.

[28] 刘羿,余航.金融网络视角下的中国上市银行机构系统性风险指标度量——基于改进后的多因子定价模型[J].未来与发展.2021,45(02):26-33+20.

[29] Zhang Hongwei,Wang Ying, et al. The impact of country risk on energy trade patterns based on complex network and panel regression analyses[J]. Energy,2021,222.

[30] Huang X, An HZ, et al. Impact assessment of international anti-dumping events on synchronization and comovement of the Chinese photovoltaic stocks[J]. RENEWABLE & SUSTAINABLE ENERGY REVIEWS. 2015, 59:459-469.

[31] Gao XY, Fang W. Detecting method for crude oil price fluctuation mechanism under different periodic time series[J]. APPLIED ENERGY. 2017, 192:20

[32] 刘超,郭亚东.金融风险在股票市场的传染效应及联动行为分析[J].运筹与管理.2020,29(10):198-211.

[33] 东方财富网. 中国石油吧 [EB/OL]. (2021-02-15)[2021-02-15]. http://guba.eastmoney.com/list,000001.html.

[34] 东方财富网. 中国石油 [EB/OL]. (2021-02-15)[2021-02-15]. http://quote.eastmoney.com/zssh000001.html.

[35] 李亚珍,李晓戈,于根.基于中文股票博客的情感分类[J].武汉大学学报(理学版).2015,61(02):163-168.

[36] 蔡红,陈荣耀.基于 PCA-BP 神经网络的股票价格预测研究[J].计算机仿真.2011,28(03):365-368.

[37] 司守奎，孙兆亮. 数学建模算法与应用[M]. 北京：国防工业出版社，2015,231-239.

[38] 汪小帆，李翔，陈关荣. 复杂网络理论及其应用[M]. 北京：清华大学出版社，2006.

[39] Qi YJ, Li HJ, et al. Transmission characteristics of investor sentiment for energy stocks from the perspective of a complex network[J]. JOURNAL OF STATISTICAL MECHANICS-THEORY AND EXPERIMENT. 2018.

致谢

六月将至，我的毕设亦即将完成。答辩在即，在此页特感谢在毕业论文写作过程中曾给予我帮助的老师、学姐、同学和家人们。

首先，我要感谢我的指导老师李华姣老师。自大二与李老师确认导师关系后，三年来受李老师教诲颇丰、影响颇深。这些影响既包括知识、技能上的积累，也有学习习惯上一些细节化的改变。可以说，能在相对较短的时间内完成毕设，并不止在于李老师几个月的指导，而是三年来潜移默化、步步攀登。除了李老师，同样要感谢的还有信管专业的其他老师们。能够完成毕设，老师们在大学四年的教授不可或缺。我要感谢安海忠老师、林文老师、高湘昀老师、刘海燕老师、涂庆老师、崔巍老师、黄书培老师、张龙老师、方伟老师、周进生老师，感谢老师们这些年的教诲和帮助。

除了老师们，我也要感谢在毕设写作过程中不辞辛劳为我指导、修改论文的齐亚杰学姐。学姐不但学识丰富，且耐心认真，堪称榜样。是学姐的细心指导让我的论文从雏形走向完全。

我还要感谢在这个过程中一起督促、一起勉励、互相帮助的同学们，愿你们都能收获属于自己的美丽毕设果实。

最后，我也要感谢我的家人，你们是我坚实的后盾，是你们为我提供生活的物质基础和学习的精神支撑。

附录

******爬取股票历史数据代码******

```
import tushare as ts
import os
import pandas as pd
pd.set_option('display.max_rows', None)
ts.set_token('9e9f3b15599850bd7646a6dc7556ecaee0ed037594ba8377e716a80b')
pro = ts.pro_api()
df = pro.daily(ts_code='000001.SH', start_date='20200102', end_date='20210129')
df.to_csv(u'E:\Jupyter\毕业论文\股价.csv',sep=',',index=True,header=True,encoding='utf_8_sig')
```

******爬取评论代码******

```
import requests
import pandas as pd
from lxml import etree
import csv
ab1=[]
n=16000   #定义爬取的页面
daima='zssh000001' #定义爬取的股票

for a in range(1,n):
    url = 'http://guba.eastmoney.com/list,'+daima+'_'+str(a)+'.html'
    headers = {'User-Agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36
(KHTML, like Gecko) Chrome/88.0.4324.150 Safari/537.36'}
    res = requests.get(url=url,headers=headers)
    res.encoding ='utf-8'
    tree = etree.HTML(res.text)
    li1 = tree.xpath('//*[@id="articlelistnew"]/div/span[1]/text()')
```

```
    li1.remove('阅读')
    li2 = tree.xpath('//*[@id="articlelistnew"]/div/span[2]/text()')
    li2.remove('评论')
    li3 = tree.xpath('//*[@id="articlelistnew"]/div/span[3]/a/text()')
    li4 = tree.xpath('//*[@id="articlelistnew"]/div/span[4]/a/font/text()')
    li5 = tree.xpath('//*[@id="articlelistnew"]/div/span[5]/text()')
    li5.remove('最后更新')

    for i in range(len(li1)):
            try:
                list1 = [li1[i],li2[i],li3[i],li4[i],li5[i]]
                ab1.append(list1)
            except:
                continue
filename='./'+daima+'.csv'
print(filename)
with open(filename, 'w',newline='',encoding = 'utf-8') as csvfile:
    writer = csv.writer(csvfile)
    writer.writerow(('阅读','评论','标题','作者','最后更新'))
    writer.writerows(ab1)
```

********情感分析代码********

```
import re, requests, codecs, time, random, jieba,tushare
import jieba.analyse
from lxml import html
import glob
import os
import numpy as np
whole_add = glob.glob('./数据_2/*_UTF-8.csv')
date_list = []
```

```
value_list = []
excludes = ['满仓','空仓',"涨", "跌", "看好", "积极", "垃圾", "看跌", '卖', '买', '看涨', '利好', "利空", '庄', '诱多', '诱空', '出货', '底', '顶','跳水','拉升']
excludes_num = {'满仓':3,'空仓':-3,"涨":9,"跌":-9,"看好":8,"积极":7,"垃圾":-7,"看跌":-8,'卖':2,'买':-2,'看涨':8,'利好':4,"利空":-4,'庄':0,'诱多':1,'诱空':-1,'出货':3,'底':-5,'顶':5,'跳水':-6,'拉升':6}
for k in range(len(whole_add)):
    address = whole_add[k]
    comment_list = []
    with open(address,mode='r',encoding='utf-8') as f:
        for comments in f.readlines():
            comment_list.extend(comments.split(',')[1:])
        date_list.append(comments.split(',')[0])
    comment_str = ''.join(comment_list)
    date_str = ''.join(date_list)
    #print(comment_str)
    #print(date_str)
    kwords = jieba.analyse.textrank(comment_str, topK=50, withWeight=True, allowPOS=('ns', 'n', 'vn', 'v'))
    print('关键词统计')
    for kword in kwords[:50]:
        print("{0:{2:}<6}{1:>6.4}".format(kword[0], kword[1], chr(12288)))

    words = jieba.lcut(comment_str)
    counts = {}
    for word in words:
        if word in excludes:
            counts[word] = counts.get(word, 0) + 1
    items = list(counts.items())
    items.sort(key=lambda x: x[1], reverse=True)
    print('\n\n 交易情绪')
```

```
        value = 0
        counts_all=0
        for i in range(len(items)):
            word, count = items[i]
            value += count*excludes_num[word]
            counts_all+=count
            print("{0:{2:}<6}{1:>3}".format(word, count, chr(12288)))
        if counts_all==0:
             value_list.append(0)
        else:
             value_list.append(value/counts_all)
import xlwt
workbook=xlwt.Workbook(encoding='utf-8')
booksheet=workbook.add_sheet('Sheet 1', cell_overwrite_ok=True)
for i in range(len(date_list)):
        booksheet.write(i,0,date_list[i])
        booksheet.write(i,1,value_list[i])
workbook.save('情感序列_2.xls')
```