**基于Python对电影《热辣滚烫》影评数据的调查**

**内容摘要**

在电影业快速发展的今天，越来越多的人喜欢在观影后，通过评论的方式在网络上留下自己的观影感悟。本调查针对的是近期比较火的一部电影《热辣滚烫》，首先使用Python库的爬虫从豆瓣网站上获取影评数据，然后利用Python相关知识对这些数据进行处理与分析，完成对《热辣滚烫》影评数据的调查工作。

**关键词：** 热辣滚烫 Python影评 数据处理与分析

1. **调查内容**

豆瓣是一个知名的中文电影评价网站，用户可以在上面发布对电影的评分和评论，这些评分和评论数据对于了解用户对电影的喜好和评价具有重要意义。本次调查针对近期比较火的一部电影《热辣滚烫》，通过对该电影的影评数据的分析，不仅了解到用户对电影的评价、评论内容的情感倾向以及影评的数量分布等信息，而且能帮助电影制作公司和影视从业者更好地了解用户的喜好和反馈，为电影的制作和推广提供参考。

1. **调查工具**

实现这次调查目的的Python语言有几种常用的工具和库，下面列举一下：

1、Requests：用于发送HTTP请求并获取网页内容，获取豆瓣网页上的HTML代码；

2、Pandas：用于数据清洗和处理，去除重复项、缺失值和异常值，统一数据格式，转换数据类型等；

3、Matplotlib：用于数据可视化，绘制折线图、柱状图等，直观地展现数据分布和趋势；

4、lxml：用于解析XML和HTML文档，从中提取所需的数据，提供了快速、灵活和简单的API；

5、openpyxl：用于操作Excel文件（.xlsx格式），包括读取、写入、修改和创建Excel文件；

6、jieba：用于对中文文本进行分词操作，方便后续的文本分析和挖掘；

7、PIL：用于加载、处理和保存图像数据，进行图像的剪裁、缩放、旋转等操作；

8、wordcloud：用于生成词云图，将文本中出现频率较高的词语以词云的形式展现出来，直观地展示文本的关键信息和特征。

这些工具和库在进行网页数据爬取、数据处理、文本处理和可视化分析等方面发挥着重要作用，能够帮助我们更高效地完成对影评数据的调查工作。

1. **调查方案**

3.1数据来源

本调查定义了一个URL列表，该列表的网页内容涵盖了豆瓣电影《热辣滚烫》200条最热门的影评数据，包含的数据属性有：用户名、评价、IP属地、发表时间、发表内容和支持人数。

3.2数据爬取

为了爬取豆瓣影评数据，可以使用Python编程语言来编写爬虫程序，爬虫程序实现模拟用户登录豆瓣网站，然后通过网页抓取技术获取电影《热辣滚烫》的全部影评数据的操作。

3.3数据处理与可视化分析

得到豆瓣影评数据后，使用Python中的数据处理和分析库（如Pandas、NumPy、Matplotlib等）对数据进行处理和分析。首先对影评数据进行清洗和整理，去除重复数据、缺失数据和异常数据，然后对清洗后的数据进行可视化分析，包括评价的分布、评论内容的情感倾向分析、评论数量的时间变化趋势等。

1. **调查过程**

4.1影评数据提取过程

首先进行《热辣滚烫》影评数据的提取，接下来说一下具体操作：

1、设置请求头：get\_headers()函数定义了一组请求头，这些请求头通常用于模拟浏览器的请求，以避免被网站识别为爬虫。

2、获取网页内容：get\_douban\_chart(url, headers)函数使用requests.get方法发送HTTP GET请求到指定的URL，并携带之前设置的请求头。如果请求成功，它将返回网页的HTML内容；如果请求失败，将打印错误信息并返回None。

3、解析HTML内容：在main()函数中，对于每个URL获取到的HTML内容，使用etree.HTML(data)将HTML字符串转换为一个可解析的HTML对象。

4、定位影评数据元素：使用XPath表达式（如'//\*[@class="comment-item "]'）来定位影评数据的元素，每个元素具有特定class属性。

5、存储影评数据：将提取到的影评数据存储在一个字典movie\_info中，然后将这个字典添加到一个列表movie\_list中，并使用openpyxl库创建一个新的Excel工作簿，将数据写入其中，并保存为文件。

6、执行主函数：通过调用main(urls)函数并传递一个包含豆瓣电影评论页面URL的列表来启动整个数据提取过程。

下面图1是部分Excel的《热辣滚烫》影评数据。



图1 部分《热辣滚烫》影评数据

4.2影评数据可视化分析过程

4.2.1词云图

对所爬取Excel文件中的“评价”和“评价内容”两列的数据采用词云图方式进行文本可视化分析，具体步骤如下：

1、读取Excel文件:使用pandas库读取之前存储影评数据的Excel文件，并将数据存储在df变量中；

2、加载停用词：load\_stopwords(file\_path)函数读取一个包含停用词的文本文件，并返回一个停用词集合；

3、清洗和分词：首先接收文本和停用词集合作为参数，然后使用jieba库进行中文分词，同时去除停用词；

4、分类评价并生成词云图：根据评价内容将评论分为好评和差评两类，接收处理后的文本、词云图标题、指定中文字体路径，生成好评和差评对应的词云图，将其保存为图片文件。

根据图2“好评”词云图，可以看出关键词为“贾玲”“女性”“导演”“减肥”。以下对关键词做一个解析：

1. 贾玲：作为《热辣滚烫》电影的导演和主演，贾玲的名字在好评词云图中的出现占比大表明观众对她的表现给予了高度认可。她的演技、导演才华以及个人魅力可能都得到了观众的赞赏；
2. 女性：这个关键词可能强调了贾玲作为女性在电影行业中的代表性和影响力，观众可能认为贾玲的成功为女性在电影行业中的地位和作用树立了积极的榜样。
3. 导演：贾玲作为《热辣滚烫》电影的导演身份得到了观众的认可，这可能意味着观众欣赏她在电影制作中的创意、导演技巧以及对电影艺术的贡献；

4、减肥：这个关键词可能指的是贾玲个人形象的变化，她为了在《热辣滚烫》中塑造拳击手角色而减肥，观众可能对贾玲的职业精神、决心和为角色所做的牺牲表示敬佩决心和毅力。而且在公众人物中，减肥常常是媒体和公众关注的焦点。

“好评”词云图中，这些关键词的出现揭示了观众对贾玲在电影《热辣滚烫》中的多方面贡献的认可，包括她的个人魅力、导演才华、对角色的全身心投入，以及作为女性在电影行业中的积极形象。这些正面评价可能还体现了观众对电影的情感共鸣、对贾玲个人故事的认同，以及对她在艺术和社会责任方面所做努力的赞赏。

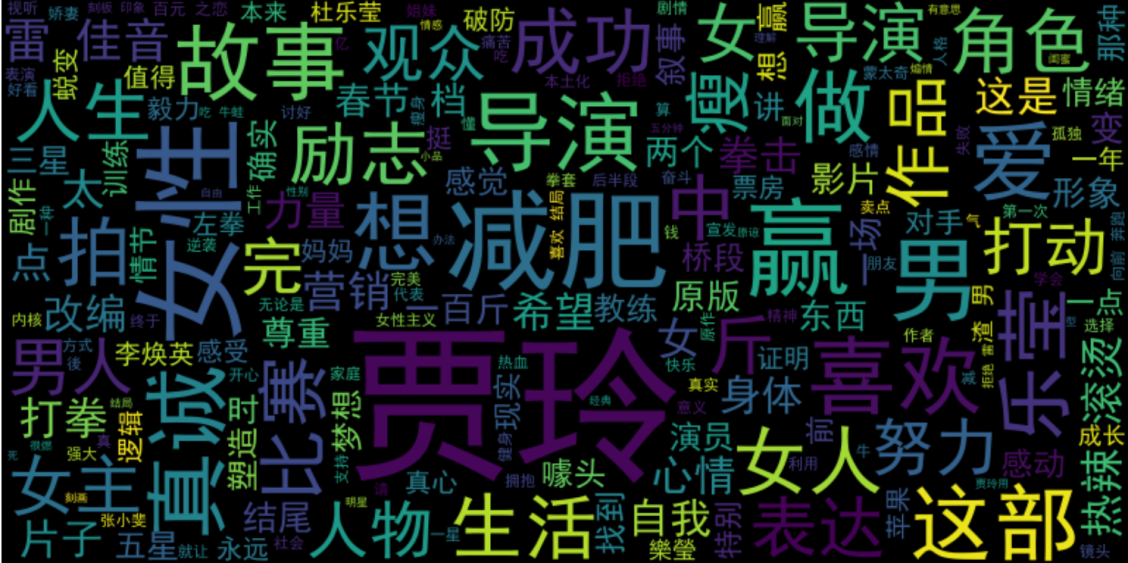


图2 “好评”词云图

根据图3“差评”词云图，可以看出关键词的是“营销”，“观众”，“故事”和“励志”。出现上述关键词的原因可能是：

1. 营销：贾玲的减肥新闻成为了《热辣滚烫》电影宣传的一大焦点，这种营销策略吸引了大量观众的注意，但也可能引起了部分观众的反感，认为电影过度依赖于营销手段而忽视了电影本身的质量，也可能有观众认为这种营销策略过于商业化，缺乏对艺术的尊重；
2. 观众：有可能部分观众对《热辣滚烫》电影的期待与实际观影体验不符，导致了不满。例如观众可能期待的是一部深度探讨女性自我成长和励志的故事，而《热辣滚烫》电影在某些方面未能满足这种期待，就可能会产生负面评价；

3、故事：差评中提到“故事”可能表示观众对《热辣滚烫》电影的情节、叙事结构或故事深度不满意。这可能涉及到故事的原创性、情感深度或情节的合理性；

4、励志：虽然“励志”是一个正面的词汇，但在差评中出现可能意味着观众对《热辣滚烫》电影的励志元素处理不当或过于刻意，可能缺乏真实感或深度。

“差评”词云图中，这些关键词的出现揭示了《热辣滚烫》电影在多方面因素存在的不足，包括宣传营销、叙事、角色塑造、主题表达和情感处理等，因此，电影制作方可能需要在这些方面进行反思和改进。

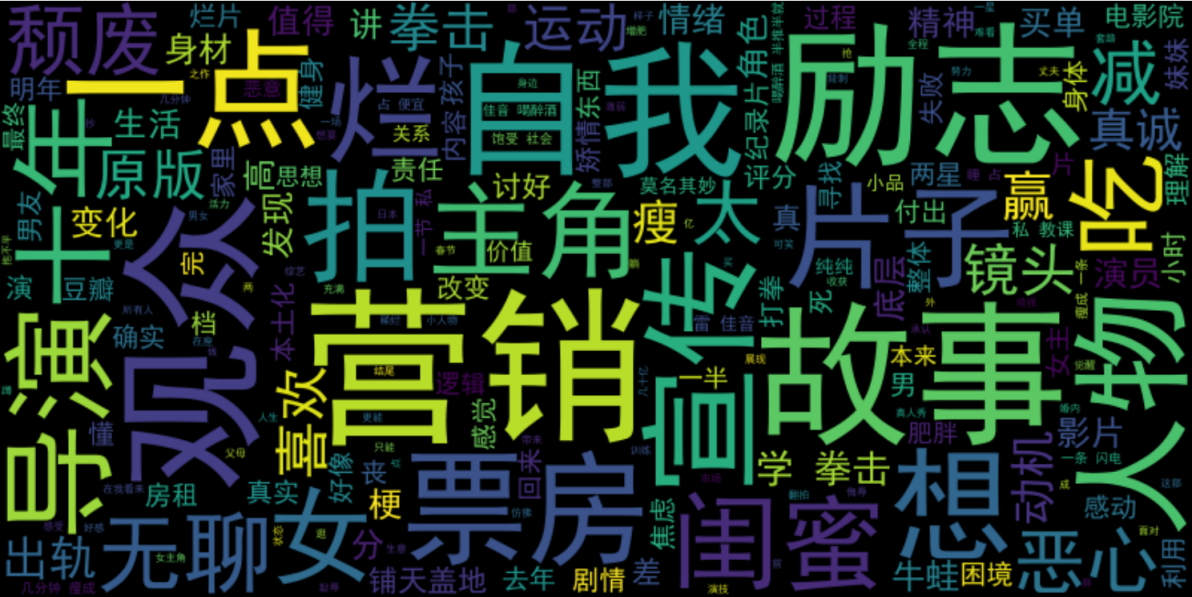


图3 “差评”词云图

4.2.2影评评分分布饼状图

对所爬取Excel文件中的“评价”列的数据采用饼状图的方式进行可视化分析，具体步骤如下：

1、过滤评价数据：定义一个包含有效评价的列表valid\_ratings。从df中读取评价列，去除空值，并过滤出有效评分；

2、统计评价数据：创建一个空列表ratings\_count用于存放每个评分的计数。遍历valid\_ratings中的每个评分，并计算其在filtered\_ratings中的出现次数，将结果添加到ratings\_count列表；

3、设置中文字体和绘制饼状图:设置Matplotlib的配置以支持中文显示，使用plt.pie函数绘制饼状图，展示不同评分的分布情况。

《热辣滚烫》影评评分饼状图分析 ：根据图 4 可知好评率为66.2%（评价为“还行”，“推荐”和“力荐”），这意味着在所有评价中，大约三分之二的观众给出了好评。这个比例相对较高，表明电影受到了大部分观众的喜爱，不过也有大约三分之一的观众给了差评，这意味着《热辣滚烫》可能在一些地方，比如故事情节、角色塑造、导演风格、视觉效果等还没有得到这些观众的认可，还有比较大的改进空间。

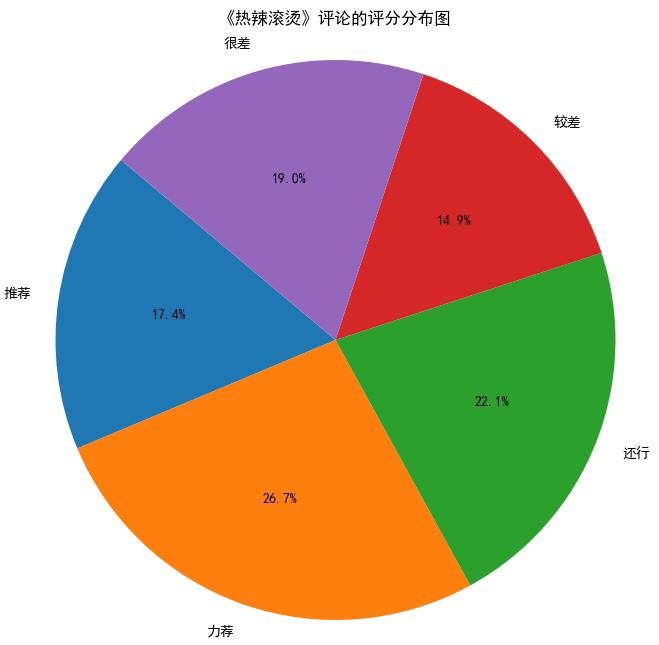


图4 影评评分饼状图

4.2.3 影评IP属地分布柱状图

对所爬取Excel文件中的“IP属地”列的数据采用柱状图的方式进行可视化分析，具体步骤如下：

1、去除空值: 通过dropna()函数去除“IP属地”列中的空值，确保分析的数据是完整的；

2、统计出现次数: 使用value\_counts()函数统计每个IP属地的出现次数，得到一个包含IP属地及其对应数量的序列；

3、绘制柱状图: 设置图形的大小，通过plot(kind='bar')函数绘制IP属地分布柱状图，展示不同IP属地的出现次数，并添加柱状图的标题、X轴标签和Y轴标签。

《热辣滚烫》电影评价的IP属地分布柱状图分析：排名前5名的地区分别是浙江、山东、江苏、北京、广东，由此可见，排名前5名的地区主要集中在经济比较发达的地区，这些地区的居民对于文化娱乐消费有着较高的需求和消费能力，也可能是因为电影制片方在这些地区进行了大力的宣传，增加了群众对该电影的期望值。

排名后5名的地区分别是新西兰、吉林、中国台湾、德国、美国，由此可见，排名后5名的地区主要在国外，说明这部电影在国际市场的宣传力度是不够的，如果后期相关电影制片方在电影制作上可以考虑符合当今国际市场对电影的需求，在上映前在国际市场中加大宣传力度，相信可以获得更多的经济效益。

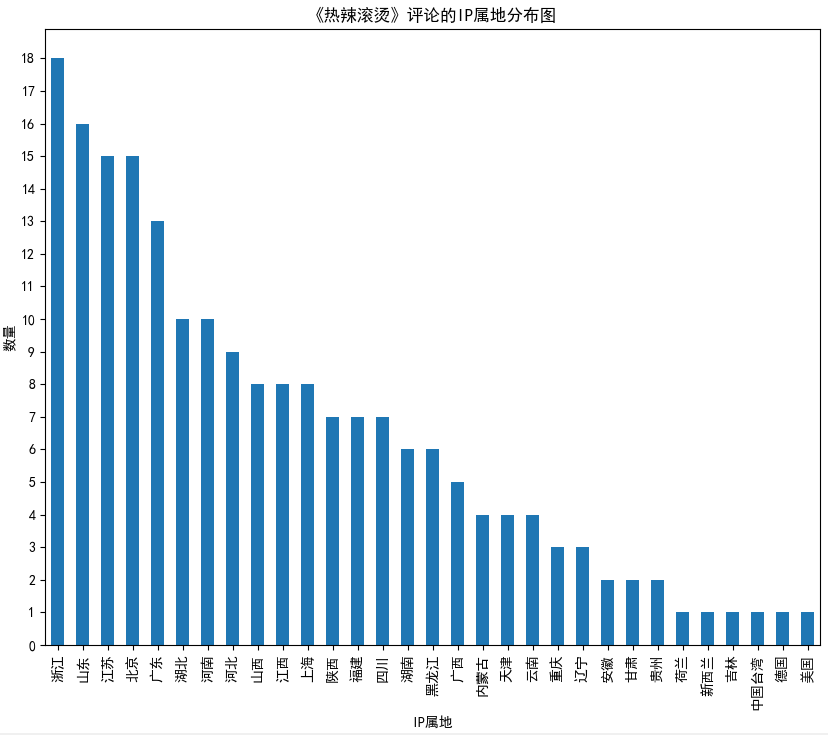


图5 影评IP属地分布柱状图

4.2.4影评发表时间分布折线图

对所爬取Excel文件中的“发表时间”列数据采用折线图的方式进行可视化分析，具体步骤如下：

1、处理发表时间的数据: 首先去除“发表时间”列中的空值，使用pd.to\_datetime()函数将时间字符串转换为datetime对象，并设置为数据框的索引；

2、重采样计数: 通过resample('D').count()函数按天对数据进行重采样，并计算每天的评论数量；

3、设置日期格式: 使用mdates.DateFormatter()设置日期格式，并使用set\_major\_formatter()和set\_major\_locator()函数设置X轴的日期格式和刻度间隔，并使用plt.xticks()函数旋转X轴的日期标签，以便于阅读；

4、绘制折线图：设置图形的大小，使用plt.plot()函数绘制折线图，展示每天的评论数量，并添加折线图标题、X轴标签和Y轴标签。

根据图 6可知，《热辣滚烫》的影评数量从2024年2月10日至2024年2月26日总体呈现出下降趋势，在2024年2月10日影评数量达到最大值，这天正是《热辣滚烫》电影的上映日，可以看出公众对该电影期望值较高，可能与前期制片方对该部电影进行了大力宣传和超前点映等工作有关。但在这部电影正式上映后，影评数量总体呈现为快速下降趋势，在短短半个月左右影评已经下降到接近0。这可能是因为公众前期对该电影的期望值过高，实际观影后未能达到公众预定的期望值，导致影评数量降低。也有可能是网上的讨论使得电影《热辣滚烫》口碑两极分化，这导致一部分观众在看到负面评价后选择不观看电影，从而减少了影评的产生。

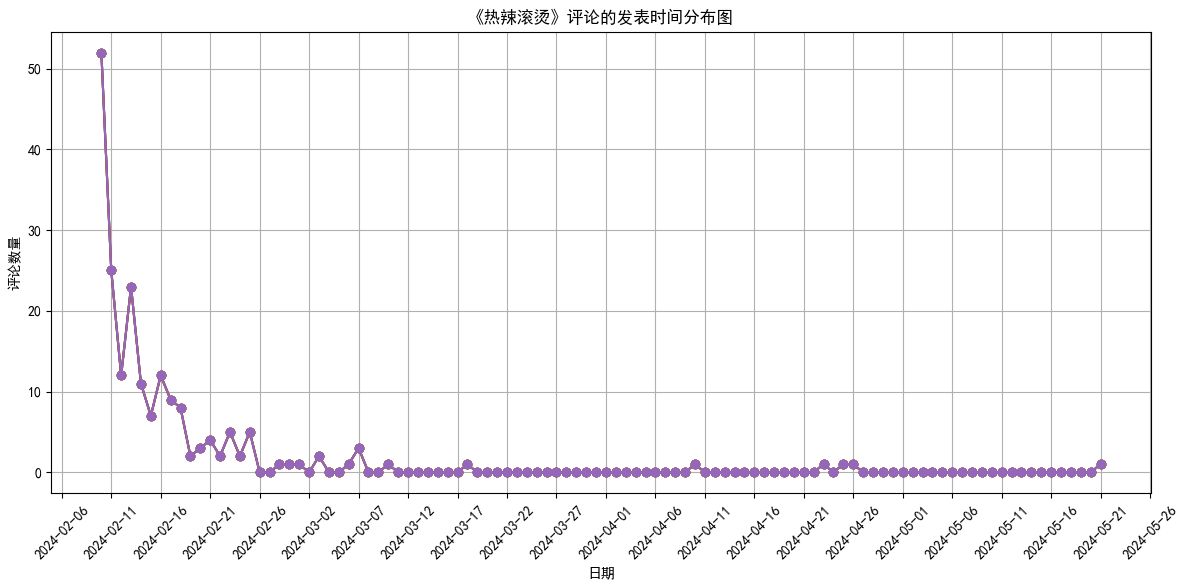


图6 影评发表时间分布图

1. **时间进度安排**

针对影评数据的爬取与分析的调查，以下是时间进度安排：

1、规划阶段：（1到2天）

（1）确定项目的目标：确定要调查的电影对应的评论及想要调查了解的问题等；

（2）确定数据源：确认从豆瓣网站爬取影评数据；

（3）制定详细的调查计划：包括爬取的影评数据范围、数量等；

（4）预估所需的时间和资源：考虑到影评数据量和分析复杂性，估计完成整个程序设计的时间，合理安排调查的工作。

2、数据收集阶段：（1到2天）

（1）确定数据爬取的工具和技术：选择合适的Python知识和解析库（如lxml）；

（2）编写爬虫程序：根据项目需求，编写爬虫程序以获取要调查的电影的影评数据，需要一些时间来调试和优化爬虫代码；

（3）测试运行爬虫程序：确保爬虫程序能够正常运行，并且能够获取所需的影评数据；

（4）监控数据收集过程：为确保爬虫程序能够正常运行，设计代码及时处理可能出现的错误或异常。

3、数据处理和分析阶段：（2到3天）

（1）数据清洗和整理：清洗爬取到的影评数据，处理缺失值、重复值和异常值等；

（2）文本分析：使用自然语言处理技术对影评数据进行分析，如情感分析、关键词提取等；

（3）数据可视化：利用Matplotlib、wordcloud等Python库对影评数据进行可视化分析结果。

4、报告撰写和呈现阶段：（1周左右）

（1）撰写调查报告：根据这次调查，完成调查报告的撰写，严格按照报告格式要求撰写，确保报告内容清晰易懂，语言简洁明确，资料数据详实可靠；

（2）呈现调查结果：通过图表和文字叙述等展示影评数据调查的结果，并进行总结。

1. **调查结果与结论**

本次调查以豆瓣电影《热辣滚烫》的影评数据为调查对象，使用Python语言对这部电影的影评数据进行了爬取，然后对所爬取的数据进行处理，利用“IP属地”、“发表时间 ”、“评价”和“评价内容”的数据元素进行调查分析。由前文的调查分析得出以下结论：1、《热辣滚烫》这部电影好评率较高，并且贾玲作为这部电影中的主演，她的个人魅力、导演才华、对角色的全身心投入等得到很多公众的认可；2、《热辣滚烫》电影还有比较多做得不足的地方，比如：电影制作方在电影宣传、故事叙述等方面可能做的工作不够，电影国际推广力度不够，电影影评热度下降过快。

通过对影评数据的分析，可以了解用户对当今热门电影的评价偏向、评分情况以及评论内容特点等信息，也为后期电影制作方提供了参考，这就是本次调查的目的。

因此，后期电影制作方和发行方应该从影评数据中总结经验，对现在电影行业发展形势和公众有深入的了解，并采取灵活多样的策略来应对不同的挑战。

**参考文献**

[1]李海荣,徐圣方,蒋晨曦,谌诗坤,耿一蕊. 基于Python的豆瓣影评数据的爬取与分析[J]. 电脑知识与技术,2024,20(09):75-79.