

**《大数据与人工智能》**

**期末报告**

标 题 京东商品评论数据爬取与分析

姓 名 刘淑仪

学 号 2251730

学 院 软件学院

专 业 软件工程

教 师 余 有 灵

二〇二四 年 六 月 二十 日

**京东商品评论数据爬取与分析**

刘淑仪 2251730 软件学院 软件工程

[摘 要] 本文通过分析京东上雀巢速溶咖啡的消费者评论，展示了数据科学在商业策略和产品优化中的应用。文章首先强调了分析评论数据的重要性，通过爬虫和自然语言处理技术获取和分析评论，以提供消费者洞察和市场趋势。文中描述了具体的技术实施过程，包括数据爬取、情感分析、数据可视化等，并提供了Python代码示例来实现这些功能。此外，文章还探讨了技术和方法的改进方向，如使用更先进的爬虫框架和情感分析模型，以及增加交互式的可视化功能。整体而言，该项目不仅技术实现上具有教学和研究价值，也为商业决策提供了实用的数据支持。

[关键词]Python爬虫 数据分析 数据可视化 自然语言处理

**一、背景介绍**

爬取京东上的雀巢速溶咖啡商品评论数据并进行分析，可以为企业提供全面的消费者洞察，帮助优化产品和服务，并制定精准的市场策略。这不仅有助于提升消费者满意度和品牌忠诚度，还能通过数据驱动的决策提高企业竞争力。从技术角度看，这一任务依赖于成熟的网络爬虫技术和自然语言处理工具，能够高效地获取并分析大量评论数据，提供深刻的市场见解。

1. 现实意义

通过分析京东上的消费者评论数据，企业能够深入了解用户对雀巢速溶咖啡的真实反馈，包括产品的优点和缺点、消费者的需求和期望等。这些信息对于企业改进产品质量、提升服务水平、进行精准营销等方面具有重要的指导意义。同时，分析评论数据还可以识别潜在的市场机会和趋势，帮助企业在激烈的市场竞争中保持优势。

2. 技术可行性

可以利用Python编程语言中的Scrapy、BeautifulSoup等网络爬虫工具，按照京东网站的页面结构，自动化地爬取评论数据。爬取到的数据可以存储在数据库中，随后利用自然语言处理（NLP）技术，如NLTK或spaCy，对评论进行情感分析、主题挖掘等处理。最后，使用Pandas进行数据整理和分析，结合Matplotlib等可视化工具展示分析结果。当前已有许多成功案例表明，以上技术手段成熟且可靠，能够有效地支持此类数据采集和分析任务。

**二、实现过程**

1. 网页数据获取

打开京东的商品页面。最好使用没有登录的浏览器窗口或隐私模式，以避免个人账户信息影响数据抓取。在网页完全加载后，打开浏览器的“开发者工具”（F12或右键选择“检查”），切换到“网络”(Network)标签页。这里显示了所有加载的资源和数据。在“网络”工具中，可以使用筛选功能来只显示JavaScript（js）类型的文件或请求。这样可以减少你需要查看的请求数量，快速定位到可能包含评论数据的请求。刷新页面，观察新加载的请求。寻找可能包含评论数据的请求。通常，这类数据会通过异步请求（如Ajax）加载。查看请求的响应内容，尤其是JSON格式的响应，因为这种格式常用于传输数据。在JSON响应中查找评论数据。你可能需要检查多个请求的响应内容，因为评论数据可能被分散在不同的请求中。使用JSON格式查看器可以更方便地分析这些数据。

2. Python代码设计

代码结构按功能模块化，主要包含以下几个部分：

(1) 数据获取 (fetch\_comments)：该函数通过循环请求京东商品评论页面，提取每页的评论内容，并将其存储在一个列表中。

(2) 情感分析 (analyze\_sentiments)：对抓取到的评论进行情感分析，将评论情感分为积极、中性和消极三类，并计数。

(3) 情感分布可视化 (create\_sentiment\_pie\_chart)：使用matplotlib生成情感分布的饼图，图中明确标出了各情感类别的比例。

(4) 词云生成 (create\_wordcloud)：将所有评论整合成一个字符串，并利用WordCloud库生成词云，以直观显示评论中常用词汇。

本代码使用SnowNLP库来进行情感分析。SnowNLP是一个中文自然语言处理库，能够对中文文本进行情感分析。本脚本中通过SnowNLP分析每条评论的情感值，将其分类为积极、中性或消极。一些英文库会对评论产生错误检测。

在数据可视化方面，首先整合所有评论文本，使用WordCloud生成词云，直观展示评论中高频出现的关键词，以图形方式反映出消费者关注的焦点。不仅如此，根据情感分析的结果，绘制饼图来分布展示评论的情感倾向比例，使得情感分布一目了然。

3. 运行结果分析

· 词云图

最大的词是“味道”，显示这是用户评论中提及最频繁的要素。接下来显著的词汇包括“香浓”和“不错”，表明普遍的正面评价。从“口感”，“香气”等词汇可以看出，用户在乎的不仅是咖啡的味道，还包括整体的感官体验。服务和体验：词云中包含“快递”、“包装”，这表明用户在评价中也考虑了购买流程和商品的接收状态。

正面词汇如“满意”，“喜欢”，“推荐”表明了客户的高满意度和推荐意向。然而，也有如“退货”这样的负面词汇，暗示了某些客户的不良体验。词汇“比较”、“超值”、“期待”等表明用户在进行评价时，有比较其他产品或以往经历的倾向，以及对产品性价比的评估。



图1 词云图

· 情感分析饼图

积极的评论占了大多数，达到了84.1%。负面的评论占了11.2%。中性的评论相对较少，只有4.7%。这种分布表明大部分人对于讨论的主题持积极态度。负面评论虽然存在，但相对较少，表明有一定的不满声音。中性评论最少，可能意味着大多数参与者对话题有明确的情感倾向。这种数据对于分析公众对某一话题的整体看法非常有用。

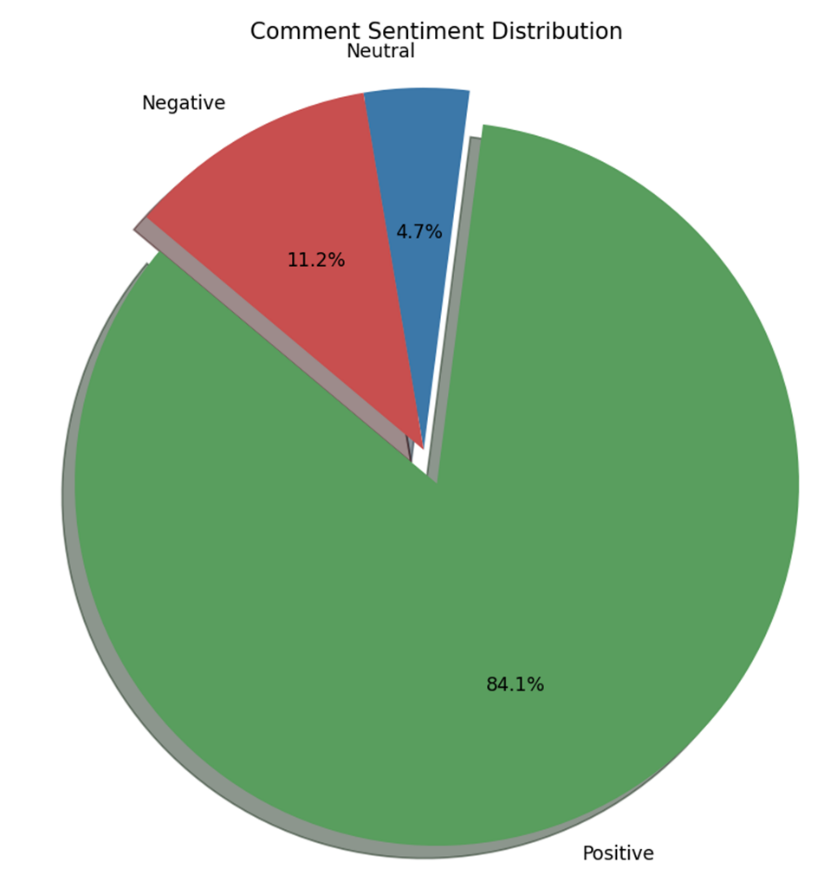


图2 情感饼图

4. 源代码

*# -\*- endoding: utf-8 -\*-*import requests *# 用于发送网络请求*import re *# 正则表达式库，用于文本匹配*from wordcloud import WordCloud *# 生成词云*import matplotlib.pyplot as plt *# 绘图库*from snownlp import SnowNLP *# 中文文本情感分析库*'''  
https://club.jd.com/comment/productPageComments.action?  
callback=fetchJSON\_comment98  
&productId=1233203  
&score=0  
&sortType=5  
&page=1  
&pageSize=10  
&isShadowSku=0  
&fold=1  
'''  
  
  
def fetch\_comments():  
 comments = [] *# 存储提取的评论* first = 1 *# 评论序号，用于打印* for i in range(1, 50): *# 循环50页评论  
 # 构建请求URL* url = 'https://club.jd.com/comment/productPageComments.action?callback=fetchJSON\_comment98&productId=1233203&score=0&sortType=5&pageSize=10&isShadowSku=0&fold=1&page='  
 finalurl = url + str(i)  
  
 header = {  
 'User-Agent': "Mozilla/5.0 (Windows NT 10.0; WOW64; rv:38.0) Gecko/20100101 Firefox/38.0",  
 } *# 请求头部，模仿浏览器行为* data = requests.get(url=finalurl, headers=header).text  
 remodel\_comment = re.compile(r'\"content\":\"([^"]+)\",\"(?:creationTime|vcontent)\"') *# 正则表达式匹配评论内容* comment\_list = remodel\_comment.findall(data)  
  
 for comment in comment\_list:  
 print(first, ":", comment)  
 first += 1  
 comments.append(comment) *# 将评论添加到列表中* return comments  
  
def analyze\_sentiments(comments):  
 sentiments = [] *# 存储情感分析结果* for comment in comments:  
 s = SnowNLP(comment) *# 使用SnowNLP进行情感分析* sentiments.append(s.sentiments)  
 positive\_count = sum(1 for x in sentiments if x > 0.6) *# 计算积极评论数* neutral\_count = sum(1 for x in sentiments if 0.4 <= x <= 0.6) *# 计算中性评论数* negative\_count = sum(1 for x in sentiments if x < 0.4) *# 计算消极评论数* return [positive\_count, neutral\_count, negative\_count]  
  
def create\_sentiment\_pie\_chart(sizes):  
 labels = ['Positive', 'Neutral', 'Negative']  
 colors = ['#599e5e', '#3c78a9', '#c84f4f']  
 explode = (0.1, 0, 0) *# 突出显示积极部分* plt.figure(figsize=(8, 8))  
 plt.pie(sizes, explode=explode, labels=labels, colors=colors, autopct='%1.1f%%', shadow=True, startangle=140)  
 plt.title('Comment Sentiment Distribution')  
 plt.axis('equal')  
 plt.show()  
  
def create\_wordcloud(comments):  
 text = " ".join(comments) *# 将所有评论合并为一个长字符串* font\_path = 'C:/Windows/Fonts/simhei.ttf'   
 wordcloud = WordCloud(width=800, height=400, background\_color='white', font\_path=font\_path).generate(text)  
 plt.figure(figsize=(10, 5))  
 plt.imshow(wordcloud, interpolation='bilinear')  
 plt.axis('off')  
 plt.show()  
  
def main():  
 comments = fetch\_comments()  
 sentiment\_counts = analyze\_sentiments(comments)  
 create\_sentiment\_pie\_chart(sentiment\_counts)  
 create\_wordcloud(comments)  
  
if \_\_name\_\_ == '\_\_main\_\_':  
 main()

**三、改进思路**

1. 改进数据抓取效率和稳定性：使用高级库代替requests和正则表达式的组合。可以使用如BeautifulSoup或Scrapy这样的爬虫框架来进行数据抓取。这些库不仅可以更高效地解析HTML，还可以更好地管理多页面的数据爬取，如自动遵循页面链接。在数据抓取过程中加入错误处理逻辑，如在请求失败时自动重试，这样可以提高数据完整性和程序的鲁棒性。

2. 优化情感分析精度：使用更先进的模型：虽然SnowNLP提供了基本的情感分析功能，但可能不足以处理更复杂的文本情绪。可以考虑使用基于深度学习的情感分析模型，如基于BERT的模型，这些模型通常能提供更高的准确率。

增加情感分类：目前的情感分为三类，可以考虑增加更多的情绪维度，如兴奋、愤怒、失望等，以提供更丰富的情感分析结果。

3. 提供结果的可视化功能：目前的可视化结果是静态的，可以考虑使用如Plotly或Bokeh等库生成交互式图表，这样用户可以通过交互操作来深入探索数据。除了基本的情感饼图和词云，可以考虑添加时间序列分析（评论随时间的情感变化）、用户评级与情感关联分析等多维度的数据展示。

**四、结论**

这个项目在技术上综合运用了多个数据科学的核心技术，包括爬虫的实现、数据处理与分析、文本处理、以及数据可视化等，展示了一个完整的数据科学实践流程。在爬虫技术方面，项目通过自动化脚本高效地从京东商品评论页面批量抓取数据，展示了网络数据采集的能力。在数据处理与分析部分，利用正则表达式进行文本提取并运用自然语言处理工具如SnowNLP对中文文本进行情感分析，这不仅加深了对文本数据结构的理解，也提升了处理和分析非结构化文本数据的技术能力。

文本处理环节中，通过对评论内容的解析和情感判断，有效地将文本数据转化为可量化的情感值，为后续的数据分析提供基础。数据可视化方面，通过生成词云图和情感分析图表，不仅直观地展示了消费者评论的关键词和情绪分布，而且更加形象地传达了数据背后的信息，增强了信息的表达效力和观众的理解深度。

从项目结果来看，通过这些分析和可视化方法，可以有效地洞察消费者的观点和情绪，进而为商品改进、营销策略调整、甚至市场趋势预测提供有力的数据支持。这种方法的应用不限于京东，其原理和技术也可以迁移到其他电商平台或不同的数据分析场景，如社交媒体情感分析、客户反馈分析等，显示了广泛的应用潜力和高度的适应性。

整体而言，本项目不仅在技术实现上展示了数据科学在实际应用中的多样性和复杂性，还在实践中探索了数据科学在商业决策支持中的实用价值，是数据科学领域实践教学和研究的优秀案例。这些经验和技术积累将为未来在更大规模或更复杂数据环境中的应用提供坚实的基础。

**五、参考文献**

ALL程序猿，JD京东爬虫-商品评论爬虫-----附源码.于 2021-03-08 08:51:15 发布

地址：<https://blog.csdn.net/qq_45687410/article/details/114517861>

**六、附录**

链接：https://pan.baidu.com/s/1PL0UMT1\_-lbpjmCBcwX1Fw?pwd=0620

提取码：0620