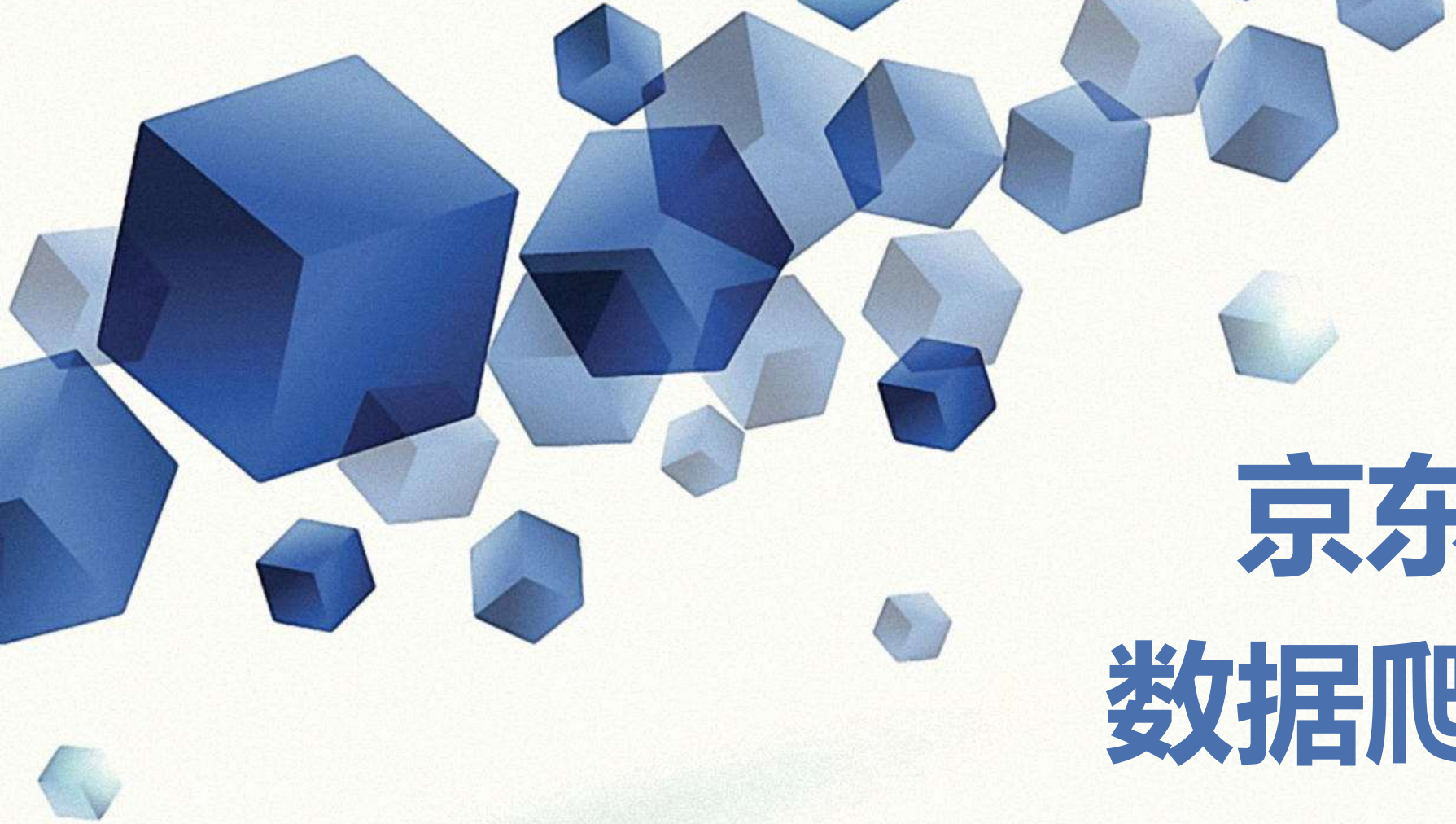


京东商品评论数据爬取与分析

2251730 刘淑仪

大数据与人工智能 | 2024/6/20

目录

CONTENTS

01 背景与分析

背景与分析

关注　　举报

企业购更优惠

京东超市 雀巢（Nestle）速溶咖啡粉1+2原味三合一南京十元咖啡冲调90条

京 东 价　¥ 109.00　降价通知
约 USD 15.00

累计评价
400万+

促　　销　PLUS专享

配 送 至　上海徐汇

支持　59

京东

由 京东

重　　量　1.63kg

选择颜色




1+2原味礼盒 15g 90条

冰袭拿铁 19g 10条 便捷条装

1+2原味90条*2 15g 2盒

1+2特浓组合装 13g 2盒

背景

爬取京东咖啡商品评论数据在技术上可行，资源成本低，但需遵守法律和道德规范；其目的在于通过市场分析、品牌管理、竞争分析、用户体验优化和数据科学研究，提供有价值的支持和洞察，从而提升产品质量、品牌声誉和市场竞争力，具有显著的可行性和实用价值。

背景与分析

数据抓取技术

使用如Python的Scrapy、BeautifulSoup等爬虫框架，可以高效地从网页中提取数据。这些工具成熟且有丰富的文档和社区支持，适合进行网页数据抓取。

网页结构分析

京东商品评论页面的结构相对固定，容易通过解析HTML和JavaScript来定位评论数据的位置。可以使用如Selenium来处理动态加载的内容。

数据存储与处理

抓取的数据可以存储在关系型数据库（如MySQL）或NoSQL数据库（如MongoDB）中。数据处理和分析可以使用Pandas、NumPy等数据处理库，结合自然语言处理库（如NLTK、spaCy）进行深入分析。

产品分析与企业改进

通过分析消费者对不同咖啡商品的评价，企业可以获得具体的改进建议，有助于提升产品质量和用户满意度。

舆情分析

实时监控评论数据，快速响应消费者的负面评价，及时采取措施，提升品牌形象；通过大规模评论数据分析，可以全面了解消费者对品牌的认知和态度，为品牌管理提供数据支持。

用户体验优化

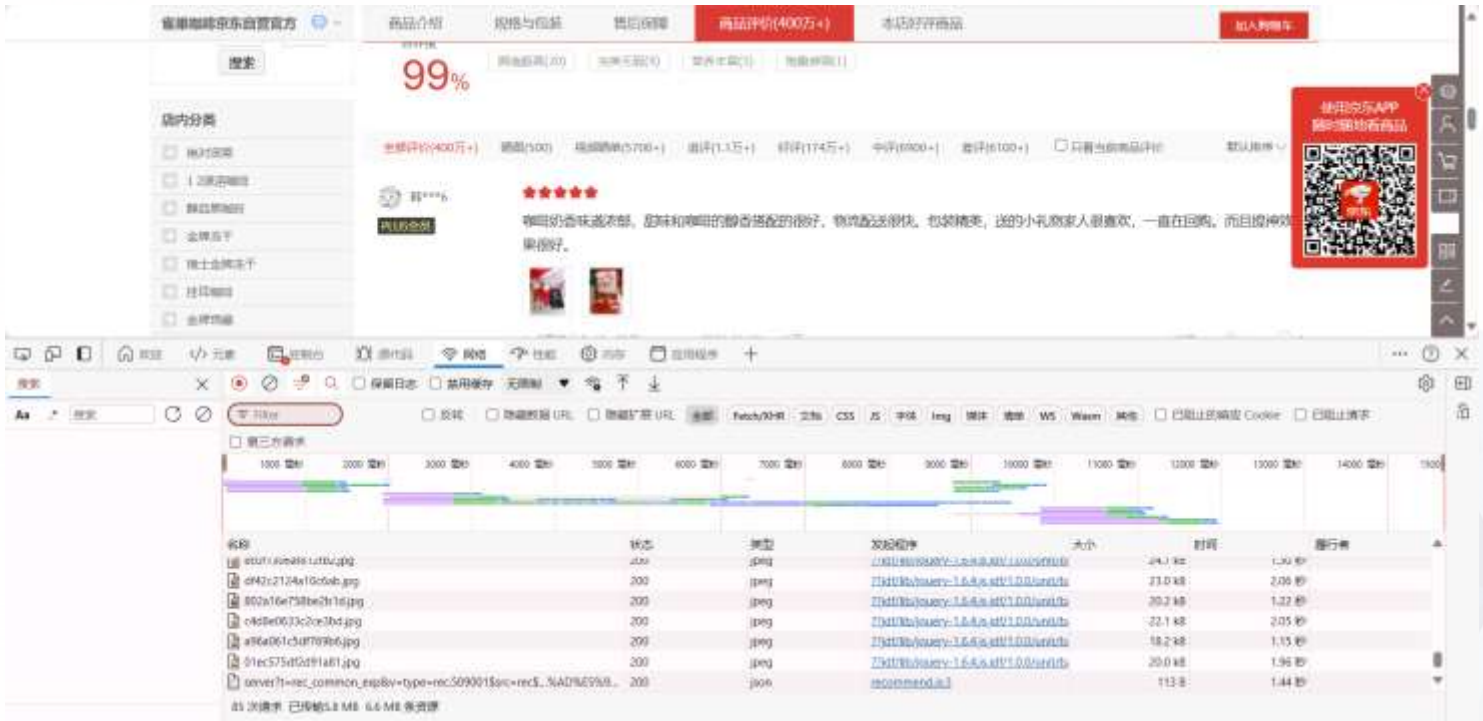
通过分析评论中的用户反馈，可以发现服务中的问题和不足，进行针对性的优化和改进；根据用户的评价和偏好，进行更有针对性的市场推广和广告投放，提高营销效果。

02 数据爬取

数据爬取

选取网页

并获取其评论的数据信息



1. 在开发者工具中，选择“网络”（Network）标签，显示网页加载过程中所有的网络请求。
2. 使用过滤器功能，输入“json”来筛选出所有的JSON请求，因为评论数据通常以JSON格式返回。
3. 在网络请求列表中，找到与评论相关的请求。通常，这些请求的URL中会包含“comment”或其他相关关键词。
4. 复制请求URL和参数，使用Python的requests库或者其他HTTP客户端工具发送相同的请求来获取评论数据。

数据爬取

在Python中爬取有效信息

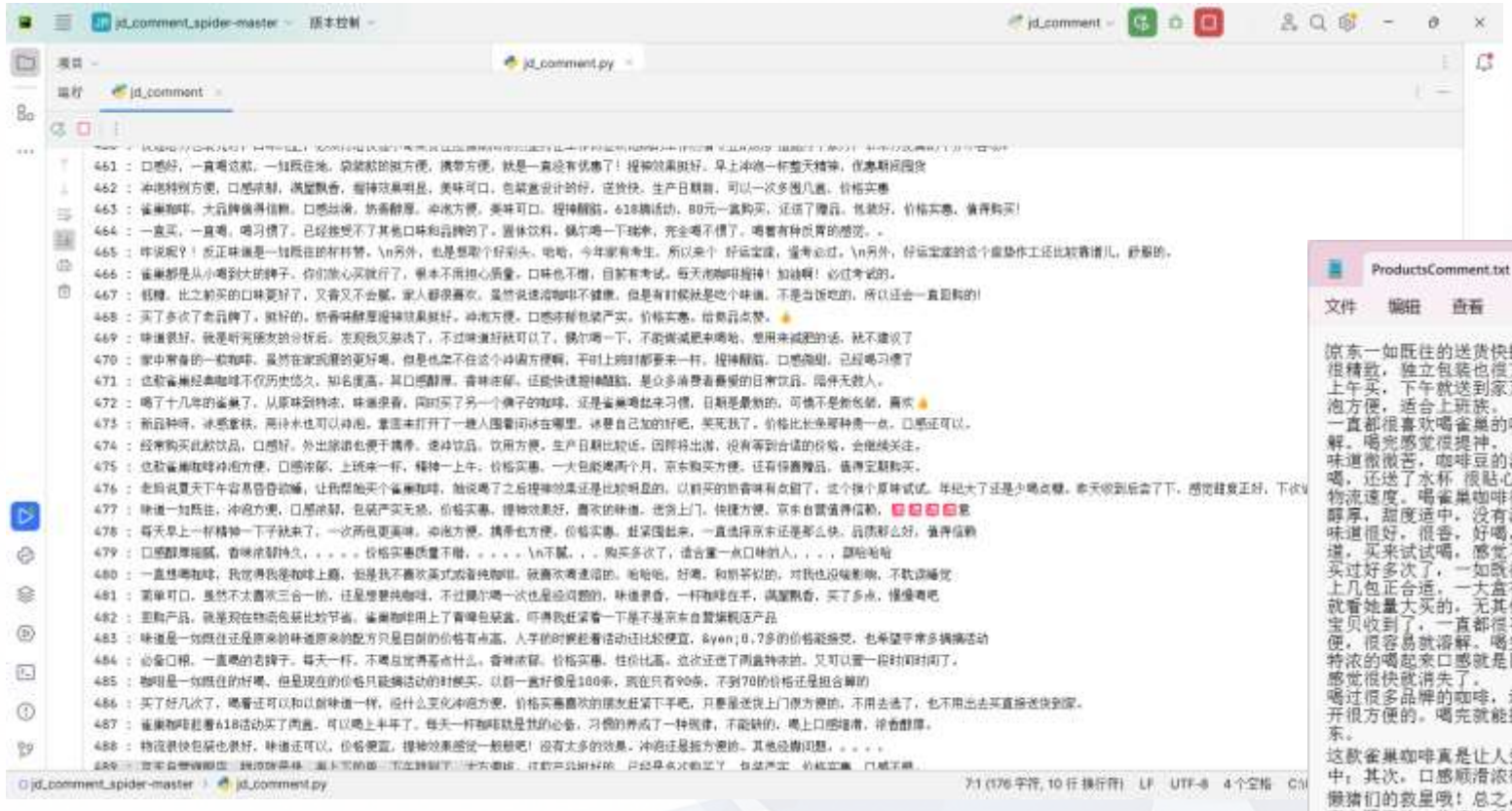
score=0： 是所有评论， score=1是差评，score=2是中评，score=3是好评，
page=0：代表的是评论的页数

```
'''
https://club.jd.com/comment/productPageComments.action?
callback=fetchJSON_comment98
&productId=1233203
&score=0
&sortType=5
&page=1
&pageSize=10
&isShadowSku=0
&fold=1
'''
```

```
import requests
import re
from wordcloud import WordCloud
import matplotlib.pyplot as plt

def fetch_comments():
    comments = []
    first = 1
    # 打开一个文件用于保存评论
        with open('ProductsComment.txt', 'w', encoding='utf-8') as file:
         for i in range(1, 50):
            url =
'https://club.jd.com/comment/productPageComments.action?callback=fetchJSON_comment98&productId=1233203&sco
re=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",
            }
            data = requests.get(url=finalurl, headers=header).text
            remodel_comment = re.compile(r'\"content\":\"([^"]+)\",\"(?:creationTime|vcontent)\"')  # 匹配评论
                    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
                        file.write(comment + '\n')  # 将评论写入文件，并添加换行符
        return comments
```

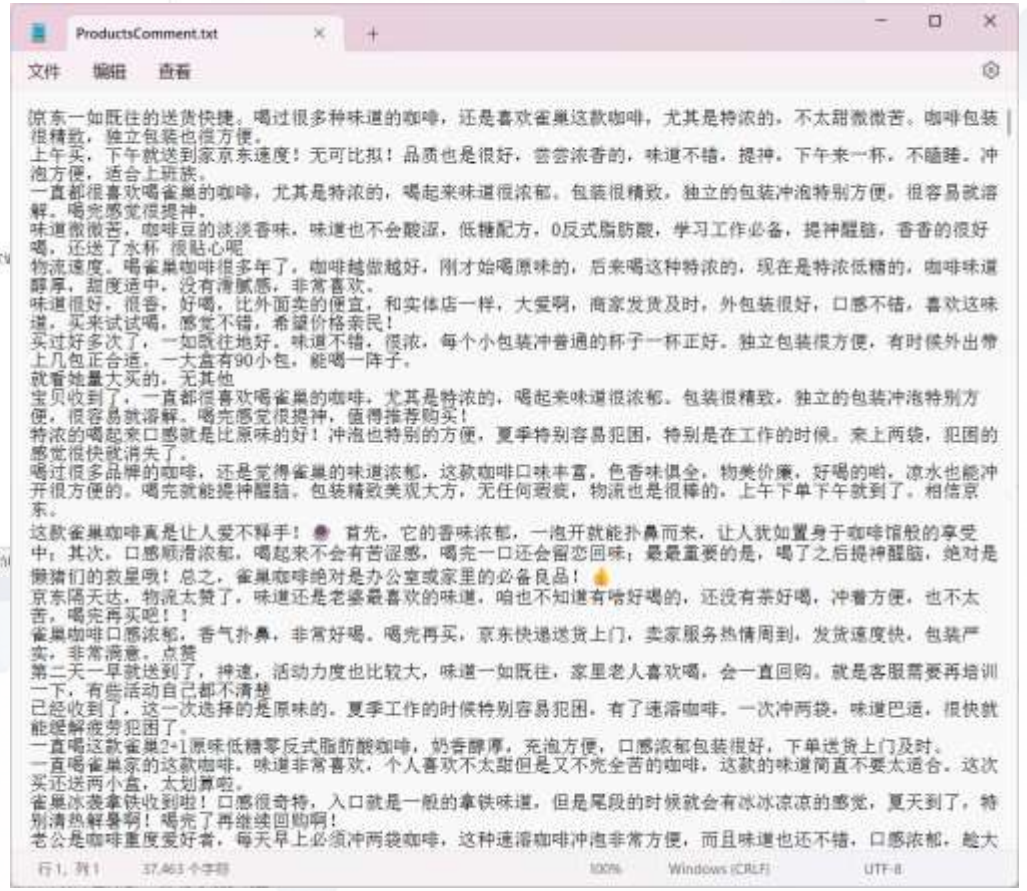
03 数据分析

数据分析



运行获得爬虫数据

并将其存入.txt文件中



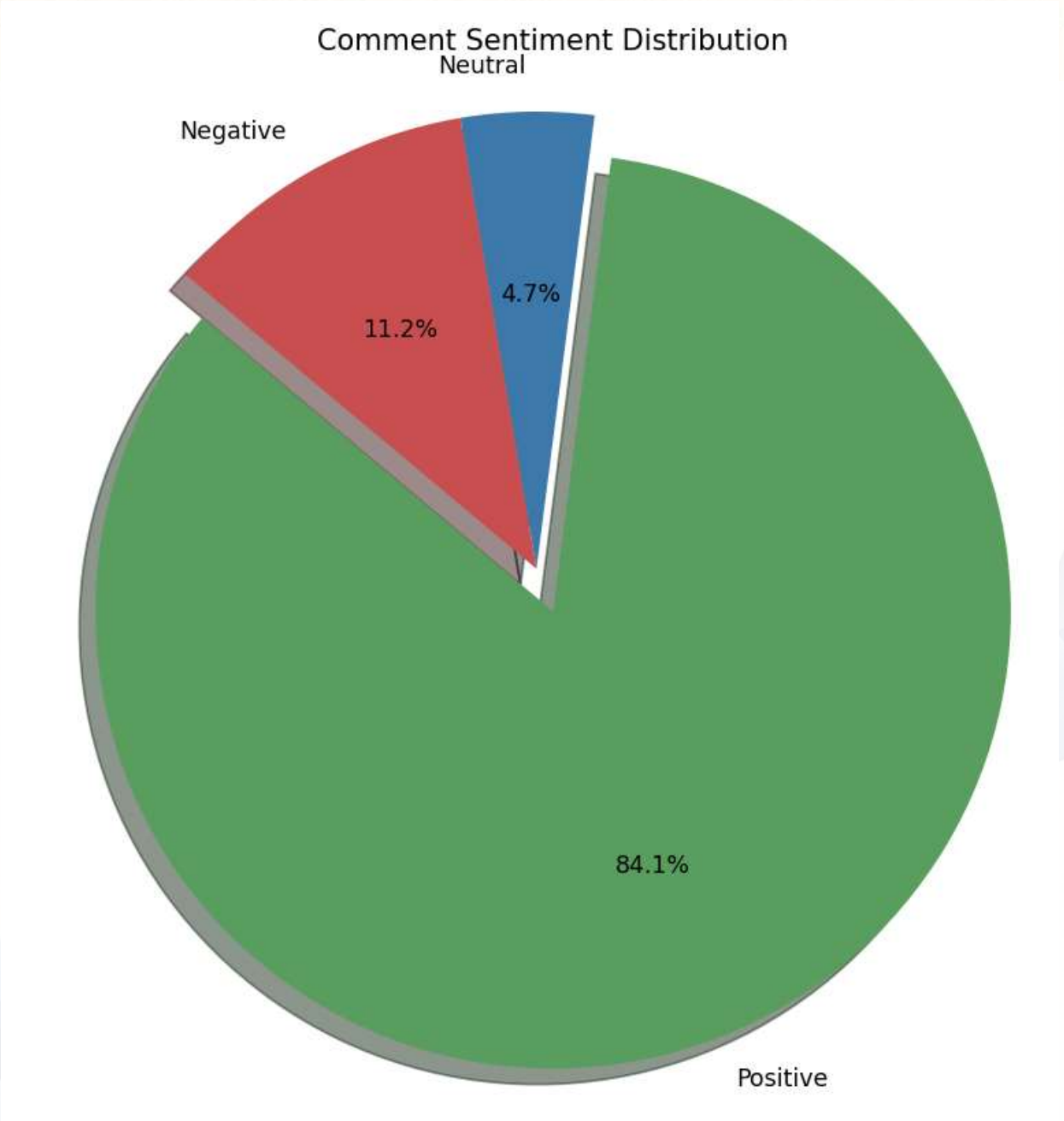
数据分析

制作词云，更好的分析数据

数据分析

制作饼状图

更好分析情感倾向



04 总结

总结

技术实现

1

数据获取

使用requests库向京东的评论API发起请求，获取评论数据。
构造URL进行分页处理，从而抓取多页的评论数据。

2

数据处理

使用re库（正则表达式）解析API返回的JSON格式字符串，提取出评论文本和评论日期。

3

数据存储

将评论数据写入本地文本文件，便于后续分析处理。

4

数据可视化

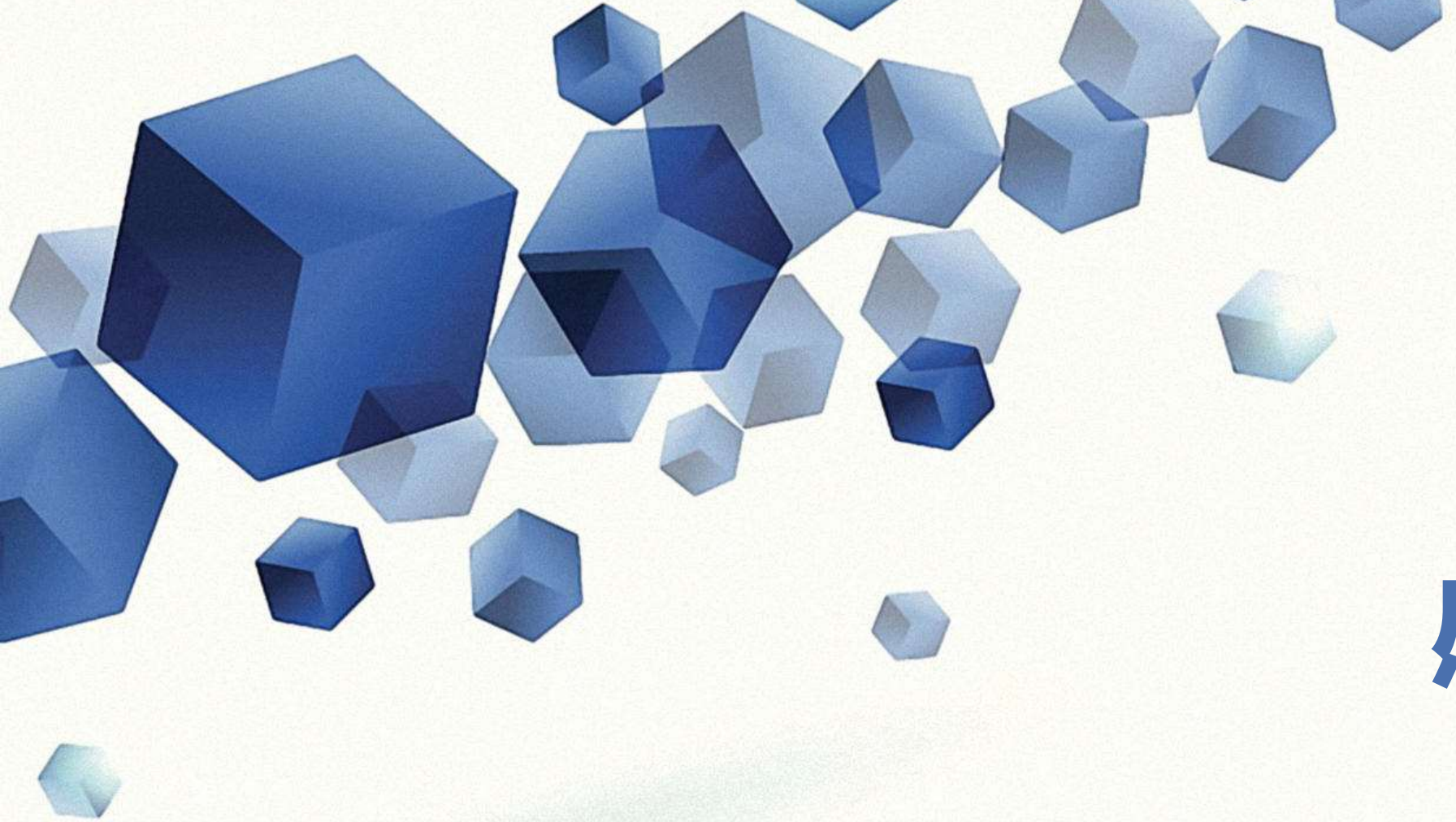
利用wordcloud库，将提取的评论文本生成词云，以直观显示评论中的高频词汇；
使用matplotlib库绘制情感极性和主观性的时间序列图，展示情感指标随时间的变化趋势。

总结

这个项目技术上涉及了爬虫的实现、数据处理与分析、文本处理、以及数据可视化等多个方面，是一个综合性的数据科学实践。从结果来看，通过生成的词云图和情感分析图表，可以有效地洞察消费者的观点和情绪，为商品改进或市场策略提供数据支持。这种方法不仅可以应用于京东，也可以扩展到其他电商平台或不同的数据分析场景。

词云图分析

1. 突出的关键词：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。
2. 产品特性：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。
3. 感情色彩：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。
4. 比较和期望：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。



感谢观看！

2251730 刘淑仪

大数据与人工智能 | 2024/6/20