基于Python爬虫的绍兴旅游数据可视化分析

0 引言

绍兴，地处长江三角洲南翼，有着 “鉴湖越台名士乡” 的美誉，是首批国家历史文化名城。这里历史底蕴深厚，名人辈出，鲁迅、王羲之、陆游等众多文化巨匠皆诞生于此，丰富的人文景观与秀美的自然风光相得益彰，名胜古迹星罗棋布。绍兴拥有众多文物旅游景点，其中国家级、省级重点文物保护单位数量可观，但网络上有关绍兴旅游的数据繁杂多样，如何借助网络数据快速了解绍兴旅游的发展状况，成为本文研究的重要意义所在。

本文运用 Python 语言，围绕数据爬取和可视化展开研究。首先，通过网络爬虫采集数据，利用 requests 获取网页源码，借助 xpath、re 和 BeautifulSoup 提取目标数据，并存储为 csv 文件；其次，对爬取的数据进行清洗，运用 Matplotlib 库绘制柱状图、饼图等，使用 wordcloud 库制作词云，实现数据的可视化分析。

1 网络爬虫与数据可视化

1.1 网络爬虫技术介绍

网络爬虫，是按照一定的规则和策略自动爬取网站某些数据信息的程序。网络爬虫是模拟浏览器的过程，通过模拟浏览器进行网站的访问，访问到浏览器后将网页进行下载爬取，从下载的网页中提取目标数据，然后进行解析，从而达到自动爬取的目的。Python语言是网络爬虫中使用最为广泛的语言。

1.2 可视化技术介绍

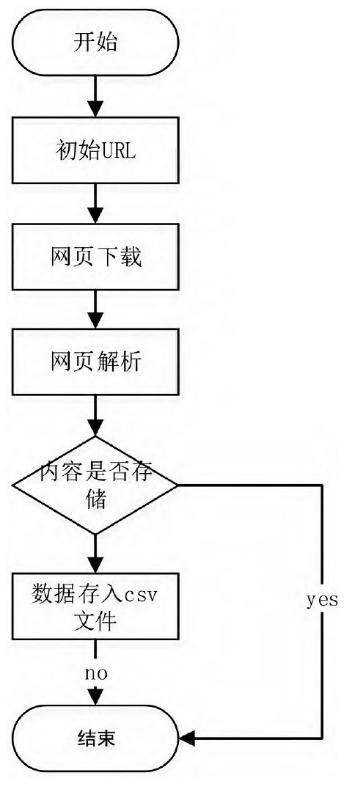
可视化技术是数据分析的一种，是将复杂且繁琐的大量数据经过可视化处理，形成直观、高效、容易理解的图表，分析得出数据中有价值的信息，然后更深层次的理解数据。

Matplotlib属于Python中的一个可视化绘图库，其功能强大且具有丰富的扩展，够实现柱状图、饼图、散点图、折线图等的绘制，从而实现快捷的可视化。

词云通过Python中的第三方库使用wordcloud库实现，是对文本中高频关键词进行可视化的展现，将多数低频、质量低、与主题无关的词语过滤，将质量高且频次高的筛选出进行展示，方便使用者对长文本的主旨快速掌握

2 数据爬取过程

绍兴旅游数据可视化分析的数据采集模块，主要负责对旅游网站中绍兴相关旅游数据的爬取、解析与存储。数据采集涵盖网页分析和数据爬取两个环节，具体流程如图 1 所示。



**图1 数据的采取流程图**

2.1 网页选取分析

在国内众多旅游网站中，途牛、驴妈妈、同程等网站在绍兴旅游的行程路线、游记、点评等方面存在不同程度的信息缺失。相较之下，“携程” 和 “去哪儿” 旅游网站信息更为全面，关于绍兴的旅游数据丰富，能有效保障数据的完整性与真实性。​

以 “携程” 网为例，使用 Edge 浏览器打开其绍兴地区主页，按下 F12 键或右键点击选择 “检查”，即可在调试窗口查看网页源代码。在爬虫过程中，为应对反爬虫机制，需模拟请求头，通过调试窗口获取浏览器请求头信息

在后续爬虫过程中要进行反爬虫机制，要进行请求头的模拟，通过调试窗口查找浏览器的请求头。

2.2 数据爬取

网页爬取借助 requests 库下载网页源代码。确定网页初始链接后，使用 requests 库的 get 方法发起 HTTP 请求，但直接访问会因被服务器识别非浏览器访问而遭拒。因此，需采取反爬虫策略，通过创建请求头模拟浏览器访问，在插入 URL 链接时填入准备好的请求头，防止爬虫被拦截。

获取网站源代码后，得到的 HTML 文档包含众多标签，目标数据蕴含其中，可使用 BeautifulSoup 库，或 re 正则、Xpath 对其进行解析提取。以 “携程” 网为例，经分析发现，该网页中各景点信息位于特定模块<div class="titleModule\_name\_\_Li4Tv"> </div>下，调用 BeautifulSoup 的 find\_all 方法，传入参数获取列表，遍历列表即可提取每个景点的链接信息。将爬取的数据，包括用户 ID、景点编号、评论分数、评论时间及内容等，存储至 csv 文件。



**图2 景点详情链接所在模块**

3 数据可视化分析

3.1 数据处理

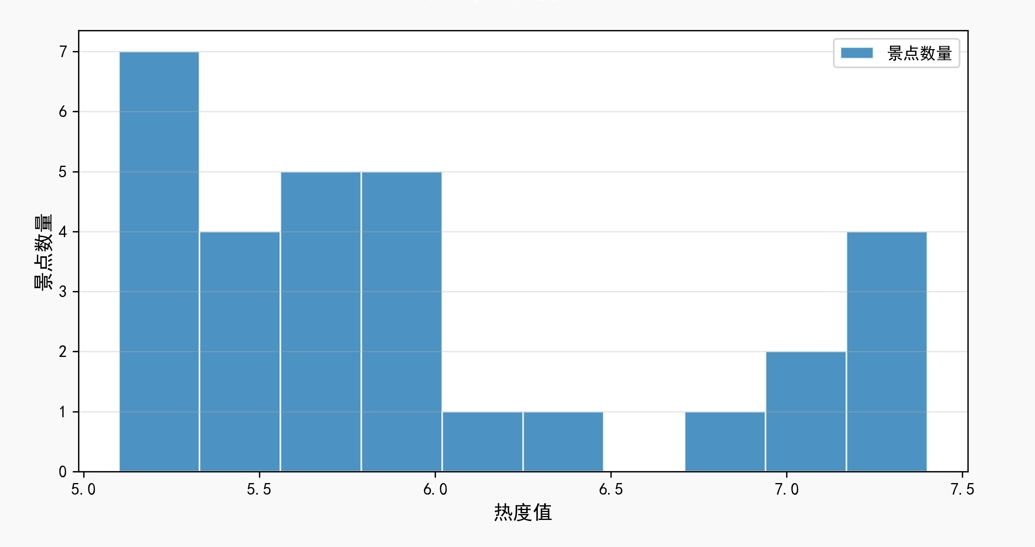
数据分析主要针对绍兴旅游数据中的用户评论，这些非结构化数据需借助 Jieba 库进行分词和去除停用词处理。

使用 Jieba 分词时，先初始化字典，再切分短语，通过正则将文本分割成语句，构建分词的有向无环图（DAG），计算各汉字节点到语句结尾路径的最大概率，确定字词结束位置，完成分词，分词结果以列表形式逐个返回。分词后，需过滤 “的”“了”“是” 等无意义的停用词，使用 wordcloud 的 stopwords 方法编辑停用词表，降低数据噪声，提升分析效率。

3.2 数据可视化分析

数据可视化主要包括Matplotlib图表的生成和wordcloud词云制作。

(1)景点热度值分析

利用 Pandas 从 csv 文件提取所需数据，传入 Matplotlib 绘制图表。对绍兴各区县的景点热度值进行爬取分析，绘制条形图，如图 3 所示。

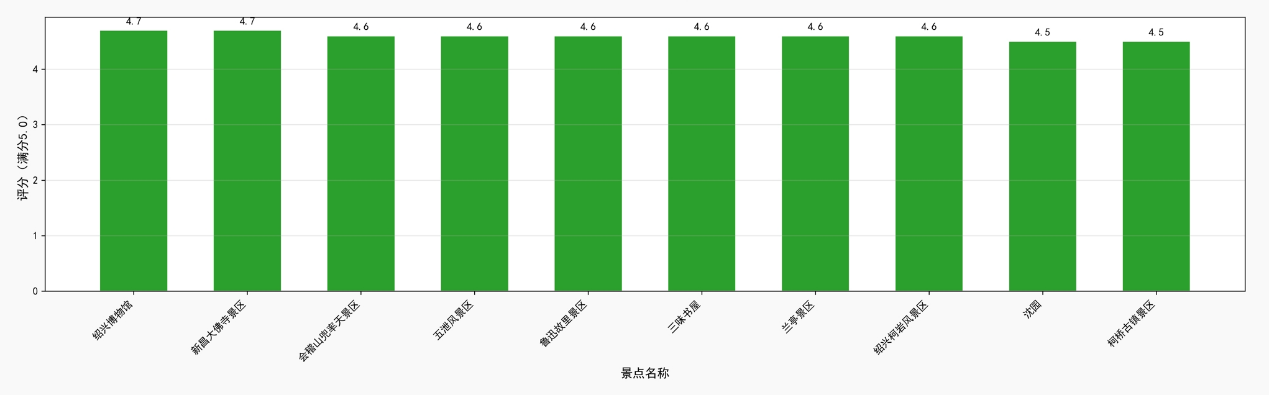
**图3 绍兴景点热度值分布**

热度值集中区间：从 “绍兴景点热度值分布” 直方图来看，多数绍兴景点的热度值集中在 5.0 - 6.0 这个区间。热度值在 5.0 - 5.5 的景点数量最多，有 7 个 ，其次是 5.5 - 6.0 区间，有 5 个景点。这表明绍兴大部分景点的热度处于中等偏上水平，具备一定的吸引力。

低热度与高热度景点情况：热度值较低（6.5 - 7.5）的景点数量相对较少。低热度可能意味着这些景点相对小众，或者宣传推广力度不足；而热度值高的景点虽然数量不多，但说明绍兴存在一些备受游客关注的热门景点，是游客出行的重点选择对象。

(2)景点评分分析

利用 Pandas 从 csv 文件提取所需数据，传入 Matplotlib 绘制图表。对绍兴各区县的景点评分进行爬取分析，获取绍兴评分最高的10大景点，绘制条形图，如图 4 所示。



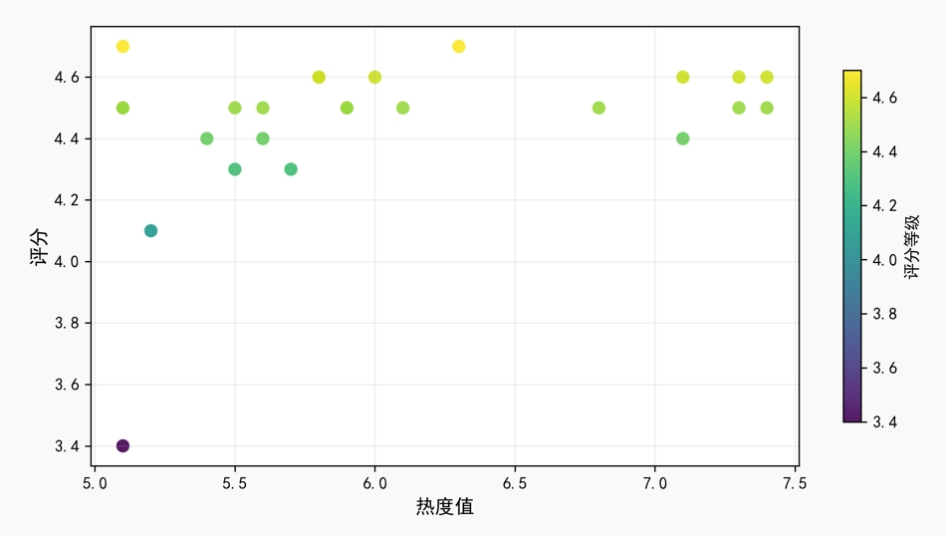
**图4 绍兴评分最高的10大景点**

高评分景点集中：从 “评分最高的 10 大景点” 柱状图可知，绍兴评分最高的 10 大景点评分都在 4.5 及以上，其中绍兴博物馆和新昌大佛寺景区评分最高，达到 4.7 。这表明绍兴在文化、宗教等类型的景点建设和运营方面表现出色，能够满足游客对于文化体验和精神需求的追求。

景点类型参考：这些高评分景点涵盖了博物馆、佛寺、自然风景区等多种类型，反映出绍兴旅游资源的丰富性和多样性，无论是对历史文化感兴趣的游客，还是喜欢自然风光的游客，都能在绍兴找到心仪的景点。

(3)景点评分与热度值的相关性分析

利用 Pandas 从 csv 文件提取所需的数据，传入 Matplotlib 绘制图表。对绍兴各区县的景点评分和热度值同时爬取，进行评分和热度值的相关性分析，绘制成散点图，如图 5 所示。



**图5 热度与评分相关性分析**

正相关趋势：在 “热度与评分相关性分析” 散点图中，可以观察到热度值和评分之间存在一定的正相关趋势。即随着热度值的增加，评分也有升高的趋势。例如，热度值较高的景点，其评分大多集中在 4.2 - 4.6 之间 ，说明热门景点在游客体验方面表现较好，具备较高的品质。

离散点分析：不过，散点图中也存在一些离散点，部分热度值较低的景点评分却相对较高，可能是一些小众但体验感很好的景点，只是宣传不足尚未被大众广泛知晓；而部分热度高但评分一般的景点，可能是因为游客期望过高，实际体验未达预期，后续需要进一步提升服务质量和游览体验。

(4)游客评论生成的词云图分析

运用 Jieba 分词处理绍兴旅游评论文本，筛选停用词后，传入 wordcloud 生成词云，如图 5所示。

**图6 绍兴游客评论生成的词云图**

高频关键词体现景点特色：词云图中，“鲁迅”“绍兴”“景区”“历史”“文化”“乌篷船”“沈园”“百草园” 等词汇字体较大且显眼。这表明鲁迅相关景点（如鲁迅故里等）在游客心中知名度极高，是绍兴旅游的重要名片；同时 “历史”“文化” 高频出现，说明绍兴深厚的历史文化底蕴是吸引游客的关键因素；“乌篷船” 作为绍兴独特的水乡元素，也备受游客关注和提及。

旅游体验相关词汇：“值得”“好玩”“免费”“美食” 等词汇反映出游客对于绍兴旅游体验的评价和关注点。“值得” 说明大部分游客对绍兴旅游的整体满意度较高；“好玩” 强调了景点的趣味性；“免费” 可能涉及到部分景点的优惠政策或免费开放区域，吸引了游客；“美食” 则体现出绍兴的饮食文化也是旅游体验的重要组成部分。

4 结论

本文围绕绍兴旅游数据，运用 Python 技术展开了全面的研究与分析。通过对网络爬虫和可视化技术的合理运用，成功实现了数据的有效爬取与直观呈现。

在数据爬取环节，对比众多旅游网站后，选定 “携程” 和 “去哪儿” 作为数据来源，因其信息全面，能保障数据质量。利用 requests 库下载网页源码，并通过模拟请求头巧妙应对反爬虫机制，结合 BeautifulSoup、re 正则或 Xpath 等工具，精准提取景点的各类信息并存储为 csv 文件，为后续分析奠定了坚实基础。

数据可视化分析部分，借助 Jieba 库对用户评论进行分词及去除停用词处理，有效降低数据噪声。通过 Matplotlib 绘制多种图表，以及利用 wordcloud 生成词云，从多个维度对绍兴旅游数据进行剖析。景点热度值分析表明，多数景点热度处于中等偏上，部分热门景点备受关注，而小众景点宣传有待加强；景点评分分析显示，绍兴在文化、宗教类景点建设运营出色，且旅游资源丰富多样；景点评分与热度值的相关性分析发现，二者呈正相关趋势，但也存在离散点，提示部分景点需提升服务体验；词云图则直观展现出鲁迅相关景点、历史文化元素及特色体验在游客心中的重要地位，以及游客对旅游体验的多方面关注。

总体而言，本文通过 Python 爬虫和可视化技术，深入挖掘了绍兴旅游数据背后的信息，为了解绍兴旅游发展状况提供了清晰视角。这些分析结果，一方面能为游客规划绍兴之旅提供参考，助其更好地选择景点、体验当地特色；另一方面，也为绍兴旅游管理部门和相关从业者提供了决策依据，可据此优化景点建设与运营、加强宣传推广，进一步提升绍兴旅游的吸引力与游客满意度，推动绍兴旅游业的高质量发展。

本文数据可视化的呈现，可为绍兴未来旅游业的发展提供一定的参考价值，但是还有很多不足之处需要改进，例如数据的获取量并不是十分充足，还有诸多有关绍兴旅游的独特之处没有体现出来。可以通过获取旅游网站不断增加的数据来弥补现有的不足之处，从而实现更加全面深入的分析。

参考文献

[1]王冬旭.基于Python的旅游网站数据爬虫研究[D].沈阳理工大学,2020

[2]薛丹丹.“旅游信息数据可视化”中的多媒体应用[D].黑龙江:哈尔滨师范大学,2020

[3]徐志,金伟.Python爬虫技术的网页数据抓取与分析[J].数字技术与应用,2020,38(10)

[4]徐博龙.应用Jieba和Wordcloud库的词云设计与优化[J].福建电脑,2019,35(6)

[5]张华.基于景点评论的京津冀旅游发展影响因素分析[D].河北:河北经贸大学,2019

[6]杨国军.基于Python的数据爬虫的设计与实现[J].数字技术与应用,2020,38(10)

[7] 张乐,孙怡芳.基于Python的运城旅游数据可视化分析,2022