**信息管理与信息系统专业课程设计（2020级）**

**×××分析报告**

**学生姓名：**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**学生班级：**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**指导教师：**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**黑龙江科技大学管理学院**

课程设计指导教师成绩评定书

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 评分标准：   |  |  | | --- | --- | | 评定项目 | 成绩 | | 1.选题是否符合专业培养要求、难易度、工作量是否适当（10分） |  | | 2.报告结构是否合理，分析四部分是否全面，有应用价值（10分） |  | | 3.数据挖掘、结果分析是否准确，文档格式是否规范（20分） |  | | 4.学生运用专业知识的熟练程度，分析解决问题的能力（10分） |  | | 5.分析成果是否有独到见解，是否有创新或亮点（10分） |  | | 总分（60） |  |   指导教师（签字）： 年 月 日 |

课程设计答辩及综合成绩评定书

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 评分标准：   |  |  | | --- | --- | | 评定项目 | 成绩 | | 1.叙述是否条理清晰，层次分明，重点突出。（10分） |  | | 2.回答问题是否全面准确，理解深刻，是否语言表达清晰、流利，逻辑性强。（15分） |  | | 3.是否很好地掌握了本专业的理论和知识，是否具有非常强的分析和解决实际问题的能力。（15分） |  | | 总分（40） |  |   综合成绩： 年 月 日 |

目 录

[第1章 绪论 1](#_Toc4890)

[1.1选题背景 1](#_Toc5600)

[1.2目的与意义 2](#_Toc2735)

[1.2.1分析目的 2](#_Toc29641)

[1.2.2分析意义 2](#_Toc25880)

[第2章 数据准备 3](#_Toc16842)

[2.1 数据采集 3](#_Toc13956)

[2.1.1 数据来源 3](#_Toc8658)

[2.1.2 获取方法 3](#_Toc1448)

[2.2 数据整理 3](#_Toc8791)

[2.2.1整理工具 3](#_Toc1895)

[2.2.2 整理过程 3](#_Toc11217)

[2.2.3 整理结果 3](#_Toc29939)

[第3章 数据挖掘的模型构建与评估 4](#_Toc8923)

[3.1 理论基础 4](#_Toc14310)

[3.2模型构建 4](#_Toc11371)

[3.2.1 特征选择与数据探索性分析 4](#_Toc17434)

[3.2.2 算法选择与模型训练 4](#_Toc11461)

[3.3模型评估 4](#_Toc30027)

[第4章 可视化及结果分析 5](#_Toc7958)

[4.1 可视化需求分析 5](#_Toc14027)

[4.2 可视化工具选择 5](#_Toc4828)

[4.3可视化设计与展示 5](#_Toc11724)

[4.3.1 \*\*趋势分析 5](#_Toc17791)

[4.3.2 \*\*相关性分析 5](#_Toc8978)

[4.4 结论与建议 5](#_Toc16841)

[参考文献 6](#_Toc3842)

# 第1章 绪论

## 1.1选题背景

随着互联网技术的飞速发展和信息资源的日益丰富， 信息量的增长速度远远超过了我们所能处理的速度。如何在如此庞大的信息量中找出我们感兴趣的信息是非常困难的。推荐系统的研究和发展已经成为电子商务行业中不可缺少的方法。它能够通过对用户的历史行为数据进行分 析找到用户比较感兴趣的商品或项目，达到对用户实现个性化的推荐的目的。其中，使用协同过滤算法是推荐系统 中应用最成功、应用最广泛的技术之一。它的主要特点是 不依赖于产品的内容，而是完全取决于一类用户所表达的喜好。 目前旅游行业信息化程度越来越高，各个旅游景点的相关信息在各类系统中存储的数据量越来越大，用户查找 自己感兴趣的旅游景区信息越来越困难。在旅游行业使用推荐系统进行各种信息的推荐方法还没有广泛使用， 旅游景区网站或主流的旅游网只提供旅游景点信息发布、 查询以及实现了在线旅游景点门票的购买服务等，而对旅游景区景点、促销等服务的推荐工作还没有广泛的应用， 为此本文将推荐算法应用于旅游景区的推荐服务中。

自从新冠疫情结束以来，国内经济的快速发展使得居民的生活条件逐渐改善，我国居民人均可支配收入持续增多，居民消费能力和消费水平均同步提高。2023年一季度，我国居民人均可支配收入为10870元，较2022年同期增长了5.07%。物质生活条件的持续改善使得人们精神层面的需求加速释放，我国旅游需求迅速增多。居民可支配收入的增多，使得人们旅游消费能力增强，再加上我国庞大的人口基数，国内旅游业的增量空间巨大。近年来，我国居民人均可支配收入始终保持稳定增长的趋势，2022年至2023年，我国多次发布政策鼓励消费，以消费拉动内需，加快国内经济的恢复与稳健发展，2023年我国经济增长形势大好，居民的收入情况也将持续改善，居民人均可支配收入将继续保持增长趋势，同时拉动国内旅游需求的持续释放。



**图 1-1 2018-2023年第一季度我国居民人均可支配收入（元）**

在我国居民人均可支配收入快速增长时期，同时数据量如此之大的情况下，我们才更应该进行本文的研究。

## 1.2目的与意义

## 1.2.1分析目的

随着人们对旅游需求的不断提高，旅游推荐系统作为一种有效的旅游智能服务，其应用范围和意义也越来越显著。本文选择基于协同过滤算法的旅游推荐系统进行研究，旨在设计一种高效、准确的旅游推荐系统，为旅游者提供个性化的旅游信息与推荐服务。

本文的研究目标是设计一种基于物品的协同过滤算法的旅游推荐系统，并将其实现。通过对用户偏好和行为的分析，为用户提供个性化的旅游推荐服务，并探讨协同过滤算法在旅游推荐系统中的应用。

## 1.2.2分析意义

首先，对于旅游行业来说，旅游推荐系统的设计和实现具有很高的实用价值。通过对用户偏好和行为的分析，推荐系统可以更加准确地为用户提供相应的旅游信息和服务，从而提高用户出行的满意度和体验感。对于旅游企业来说，旅游推荐系统也可以为企业增强品牌形象、提高竞争力和用户黏性。

其次，本文的研究对于计算机科学和数据挖掘领域也有一定的价值。本文将应用协同过滤算法进行旅游推荐，通过对算法原理和实现方法的研究，可以为该领域相关研究提供借鉴和参考，进一步推进该领域的发展和创新。

最后，本文的研究对于个人和社会来说也有积极的影响。旅游推荐系统为旅游者提供了更加便捷、高效的旅游服务，可以降低用户的旅游成本和时间成本，提高旅游的品质和效率。同时，推荐系统的应用也可以为社会创造更多的就业机会和商业机会，促进经济发展和社会进步。

综上所述，本文将通过对基于协同过滤算法的旅游推荐系统进行设计和实现的研究，提高旅游行业的服务质量和效率，推进计算机科学和数据挖掘领域的发展和创新，对个人和社会产生积极的影响。

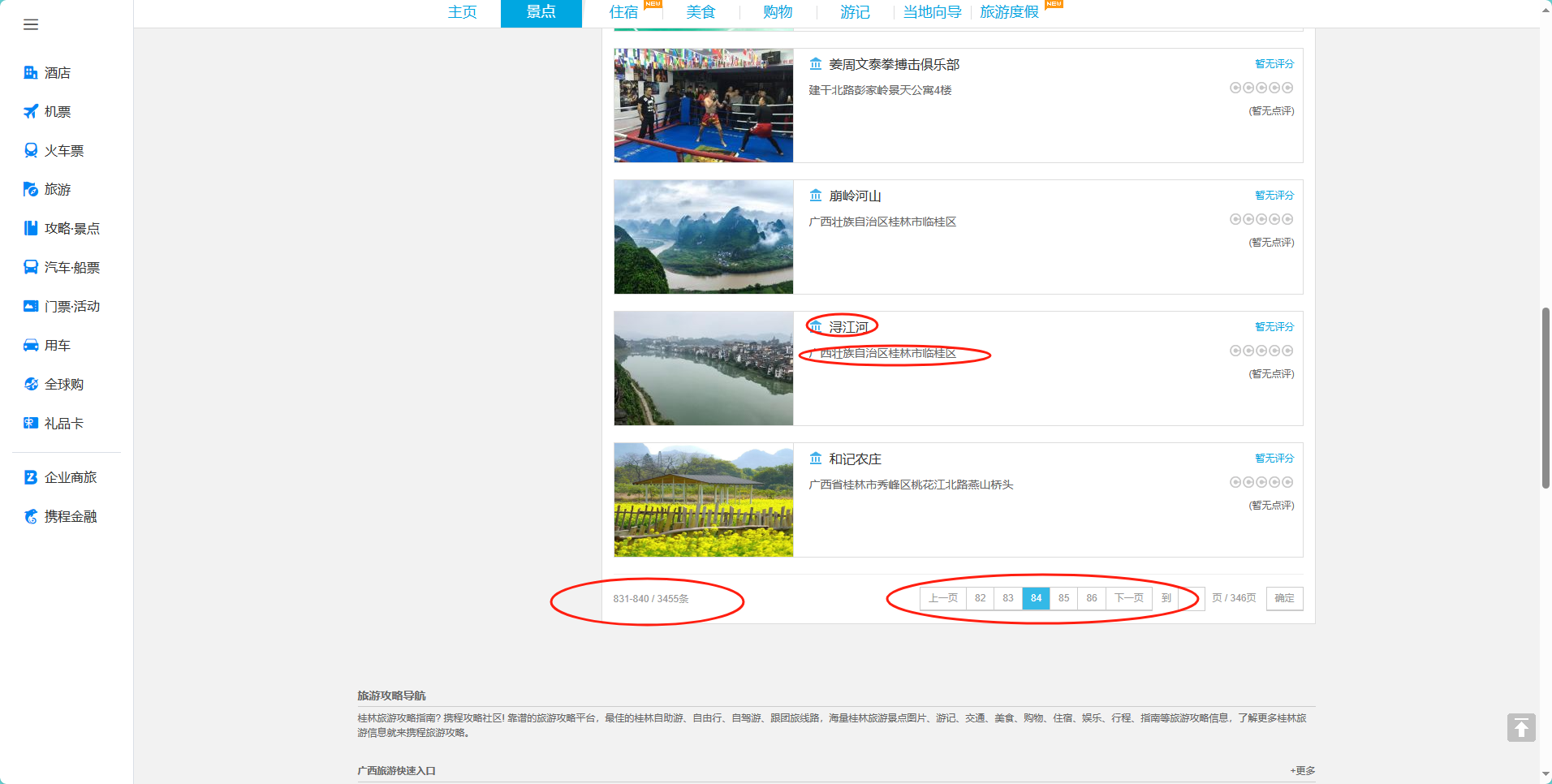
# 第2章 数据准备

## 2.1 数据采集

## 2.1.1 数据来源

数据来源是爬取“携程”网的数据，用到了requests库和beautifulsoup库。具体网址为[携程旅行网官网:酒店预订,机票预订查询,旅游度假,商旅管理 (ctrip.com)](https://www.ctrip.com/)。

## 2.1.2 获取方法

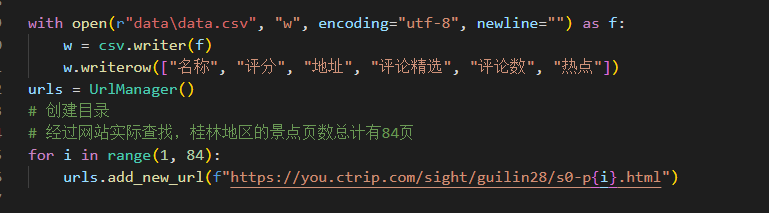
携程网址下方包含了全国各个城市的数据，从中选取了广西桂林，进行爬取。在信息栏现实的景点数有3455条，就我对常识的理解，桂林不可能有这么多的景点，经过我的一番尝试，发现从84页开始，所筛选出来的景点均不是桂林景点，如图2-2所示。

**图 2-2 携程网站列表页面**

在每一个具体景点信息中，爬取的内容为景区名称、景区所在地区、景区评分、景区热点、建议游玩时间，同时下方还储存这用户的评论信息，如图2-3所示。通过简单的查看，可以分析得到，该网站是静态网站，这就很适合使用python中的requests库来进行数据的发送与获取。

**图 2-3 景点的详细页面**

通过对比发现，携程网站的列表页面的url不是随机的数字.html的结构，而是https://you.ctrip.com/sight/guilin28/s0-p页数.html的格式，这里我们就可以选择使用for in循环来得到我们所需要的所有的详细页面的url，代码如图2-4。

**图 2-4 列表页面的获取**

当我们获取列表页面完成后，接下来就是对页面中所有详细页面的url的解析，部分代码如下图 2-5。

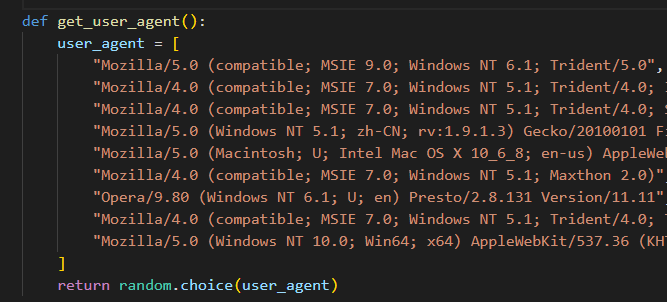
**图 2-5 解析获得超连接**

获取到所需要的url后，我又定义了一个函数，用来解析详细页面，以此来获取我们需要的数据，部分代码如下图2-6。

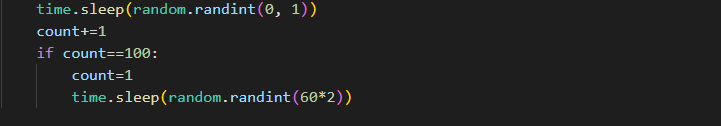


**图 2-6 详细页面的解析**

我在程序设计中遇到的问题有：1、许多景区的信息并不是完整的，可能会存在某些数据没有。这时，如果我依旧使用get\_text语句去强行获取数据，程序就会出现错误，而导致程序的中断。2、现在市面上的所有网站基本上都会有一定的反爬虫拦截，需要设计方法去避免被反爬虫识别。

我的解决措施：1、对于数据不完整的情况，我选择使用try语句去捕获这个异常，并且在异常出现时，使用null去填充不完整的数据流，上图中也有展示。2、对于反爬虫来说，我使用了两种方法来应对，第一就是，使用随机的用户代理，以此来模仿人为操作，这里，我定义了一个全新的函数get\_user\_agent，用来在每次访问网站时，获取一个随机的用户代理，如下图2-7。当然也可以使用fake\_useragent库中的UserAgent类来随机输出用户代理。第二是，在每次访问结束后，我都是用time库进行程序的整体暂停，并且在程序获取到100条数据时，将程序暂停2分钟，如下图2-8。

**图 2-7 get\_user\_agent函数**



**图 2-8 程序的暂停**

最终，我在程序运行完成后将所获得的数据写入了csv文件，以便我后续的处理

## 2.2 数据整理

## 2.2.1整理工具

使用传统的统计学软件（excel）进行数据整理工作，Microsoft Excel 是 Microsoft 为使用 Windows 和 Apple Macintosh 操作系统的电脑编写的一款电子表格软件。直观的界面、出色的计算功能和图表工具，再加上成功的市场营 销，使 Excel 成为最流行的个人计算机数据处理软件。

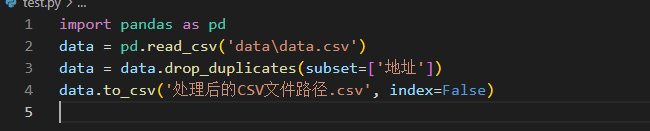
使用 python 进行大数据清洗工作，Pandas 是 Python 中很流行的类库，使 用它可以进行数据科学计算和数据分析，Pandas 提供功能强大的类库,不管数据处于什么状态，他可以帮助我们通过清洗数据，排序数据，最后得到清晰明 了的数据。除此以外，主要涉及应用的库还有 numpy，matplotlib 等。

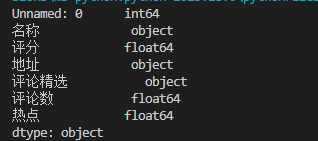
## 2.2.2 整理过程

（1）使用 Excel 对表格进行样式调整，删除多余的数据列。

（2）由于我在前文中，数据爬取的过程中，我已经对数据的缺失值做出了处理，这里的整理过程中，就不用再次对缺失值进行填充。

（3）去除数据流中可能出现的重复项，并且获取的数据中，有一些数据为，地址相同，名称不同。这些无用数据，都应该优先去除实现的代码如下图2-9。

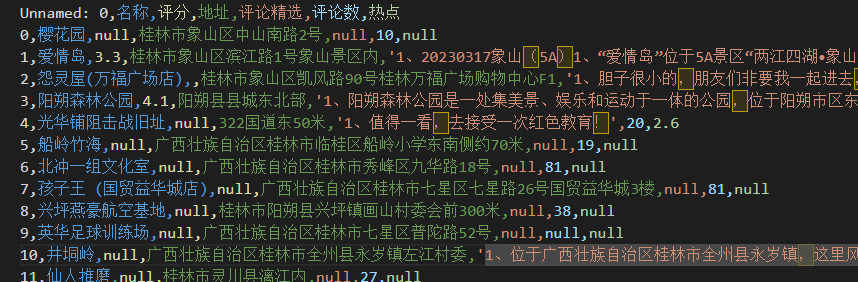
**图 2-9 数据的处理**

（4）数据类型的转换，在获取的数据中，很多数据都是杂乱无章的，这里需要使用.astype语句进行将数据的类型转换成我们需要可以使用的类型，数据的类型如下图2-10。

**图 2-10 数据的类型**

## 2.2.3 整理结果

在最后整理中，共获取如图 2-11 中的内容有七个字段。

**图 2-11 整理后的部分csv文件**

# 第3章 数据挖掘的模型构建与评估

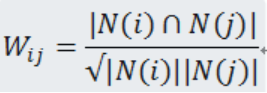
## 3.1 理论基础

协同过滤(Collaborative Filtering)作为推荐算法中最经典的类型，包括在线的协同和离线的过滤两部分。所谓在线协同，就是通过在线数据找到用户可能喜欢的物品，而离线过滤，则是过滤掉一些不值得推荐的数据，比比如推荐值评分低的数据，或者虽然推荐值高但是用户已经购买的数据。

协同过滤的模型一般为m个物品，m个用户的数据，只有部分用户和部分数据之间是有评分数据的，其它部分评分是空白，此时我们要用已有的部分稀疏数据来预测那些空白的物品和数据之间的评分关系，找到最高评分的物品推荐给用户。

基于物品的协同过滤算法。

通过查询一下资料,ItemCF的物品相似度计算模型如下：

**图 3-1 相似度计算**

公式中|N(i)|表示喜欢物品i的用户数，|N(j)|表示喜欢物品j的用户数， |N(i)∩N(j)|表示同时喜欢物品i和物品j的用户数。从上面的公式我们可以看出物品i和物品j相似是因为他们共同别很多的用户喜欢，相似度越高表示同时喜欢他们的用户数越多。  
下面举例讲解一下相似度的计算过程：  
假设用户A对物品a,b,d有过评价，用户B对物品b,c,e有过评价，如下表.

|  |  |  |  |
| --- | --- | --- | --- |
| A | a | b | d |
| B | b | c | e |
| C | c | d |  |
| D | b | c | d |
| E | a | d |  |

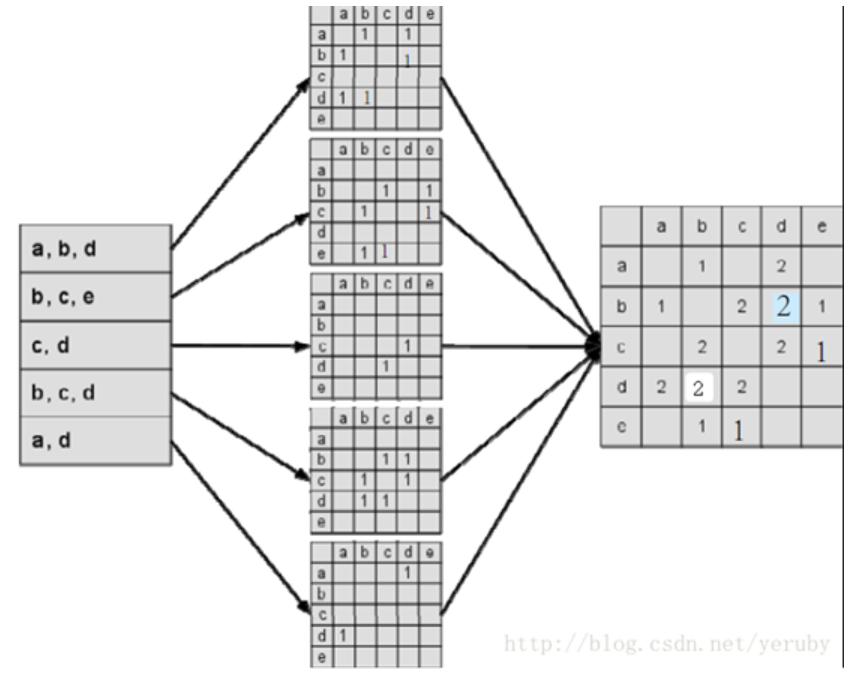
**表 3-1**

根据上面用户的行为构建:用户——物品倒排表：例如：物品a有用户A和E做过评价，如表3-2。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| a | A | E |  |  |
| b | A | B | D |  |
| c | B | C | D |  |
| d | A | C | D | E |
| e | b |  |  |  |

**表3-2**

根据上面的倒排表我们可以构建一个相似度矩阵，如图3-1：

**图3-1**

图中最左边的是用户输入的用户行为记录，每一行代表用户感兴趣的物品集合，然后对每个物品集合，我们将里面的物品两两加一，得到一个矩阵。最终将这些矩阵进行相加得到上面的C矩阵。其中Ci记录了同时喜欢物品i和j的用户数。这样我们就得到了物品之间的相似度矩阵W。

## 3.2模型构建

## 3.2.1 数据的准备

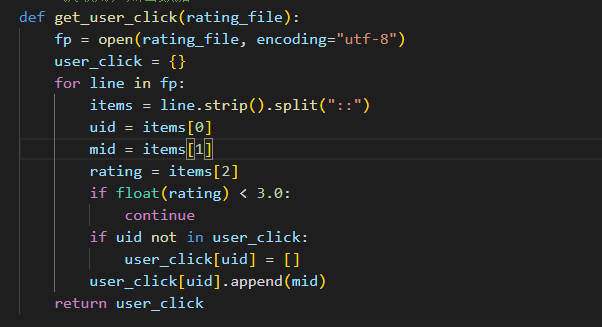
鉴于我们需要搭建基于物品的协同过滤算法实现，需要将爬取到的数据设计成矩阵的形式，以景区标号：：景区名称：：景区地址的格式进行处理，处理代码如下图3-2。

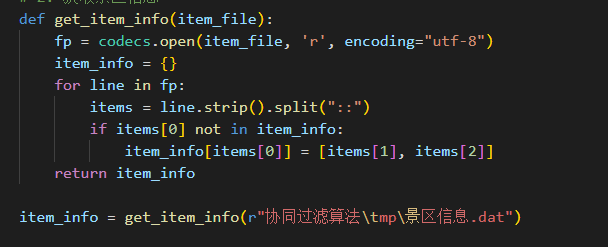


**图 3-2格式的完善**

## 3.3.2 计算物品得到用户的矩阵

我这里使用网络上开源的用户数据，用以计算矩阵模型，我定义了一个def get\_user\_click函数，用于从文件中的dat文件中获取用户数据。同时我还定义了一个get\_item\_info，用以获取，景区数据实现的代码，如下图。

**图 3-3 用户点击数据的获取**



**图 3-4 景区数据的获取**

## 3.2.3 计算物品的同现矩阵



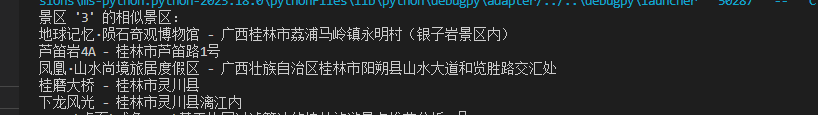
**图 3-5 同现矩阵**

同现矩阵其实就是物品与物品之间的关联度， 这个关联度由所有用户对所有物品的评分决定。

同现的意思就是同时出现的意思，就是说，喜欢 101 的时候同时喜欢 102 。

比如第一条， 说明同时有3个用户喜欢 101 与 102。意思就是说，在3个用户的各自喜好列表中 101 和 102 同时出现了。

## 3.3模型评估



最后我运行程序，输出对景点相似度最搞得几个景区的推荐，完成设计。

# 第4章 可视化及结果分析

## 4.1 可视化需求分析

可视化可以帮助研究者更好地理解数据。通过可视化图表、图像和图形，你可以观察数据的分布、趋势和关联关系。这对于了解桂林旅游景点评分数据、用户行为模式和景点之间的相似性至关重要。分析用户的行为模式。你可以创建用户行为路径图、点击热点图等可视化工具，以了解用户如何与桂林旅游景点进行互动，哪些景点更受欢迎，以及用户的旅游偏好。协同过滤算法产生的推荐结果通常需要解释给用户。可视化可以用来呈现推荐结果，以便用户更容易理解为什么会被推荐某个景点。

## 4.2 可视化工具选择

Pyecharts库是一个基于Web技术的Python数据可视化库，它提供了丰富的图表类型，使数据分析和展示变得更加直观和易于理解。与Jupyter Notebook相结合使用，可以轻松创建交互式数据可视化文档。以下是Pyecharts库和Matplotlib库的一些主要特点：

（1）Pyecharts库：Pyecharts库允许你创建各种类型的图表，如折线图、柱状图、饼图、地图等，以展示数据的不同方面。

（2）与Jupyter Notebook集成：你可以在Jupyter Notebook中使用Pyecharts库来创建图表，无需离开笔记本环境，这使得数据分析更加流畅和直观。

（3）富媒体展示：Pyecharts可以生成富媒体格式的图表，包括HTML，图片和SVG等。这使得你可以在文档中嵌入漂亮的图表，以更好地传达数据的含义。

（4）Markdown支持：与Jupyter Notebook一样，Pyecharts库也支持Markdown语法，这意味着你可以在图表旁边添加解释性的文本，帮助读者理解数据。

（5）互动性：Pyecharts库允许你创建交互式图表，通过鼠标悬停和点击等方式与数据进行互动，进一步增强了数据分析的效果。

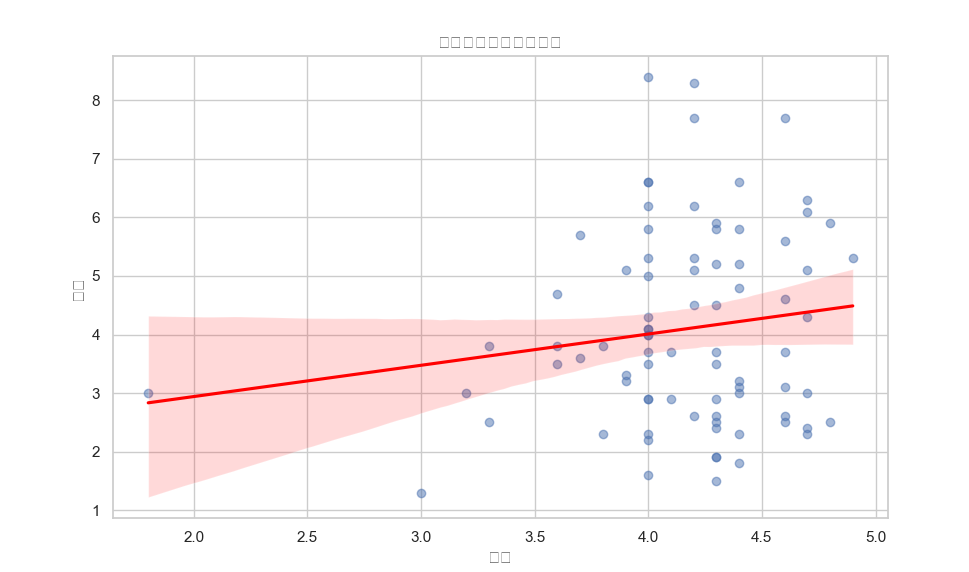
（1）Matplotlib库：Matplotlib库是Python的绘图库，用于生成高质量的图形，包括线图、散点图、直方图等。它提供了广泛的自定义选项，以确保你可以创建满足特定需求的图形。

（2）满足大部分绘图需求：Matplotlib和Seaborn库通常能够满足数据分析中的绝大多数绘图需求，无论是探索性数据分析还是最终的结果展示。

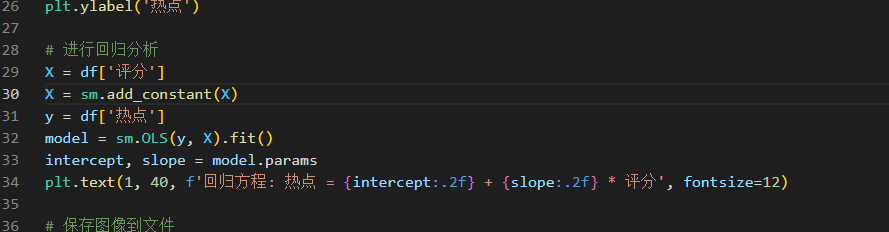
Pyecharts库和Matplotlib库都是数据可视化的有力工具，可以在数据分析和报告中起到关键作用。Pyecharts库特别适用于创建交互式和富媒体的图表，而Matplotlib和Seaborn库则适合生成各种类型的静态图形。

## 4.3可视化设计与展示

## 4.3.1 热点与评分回归分析



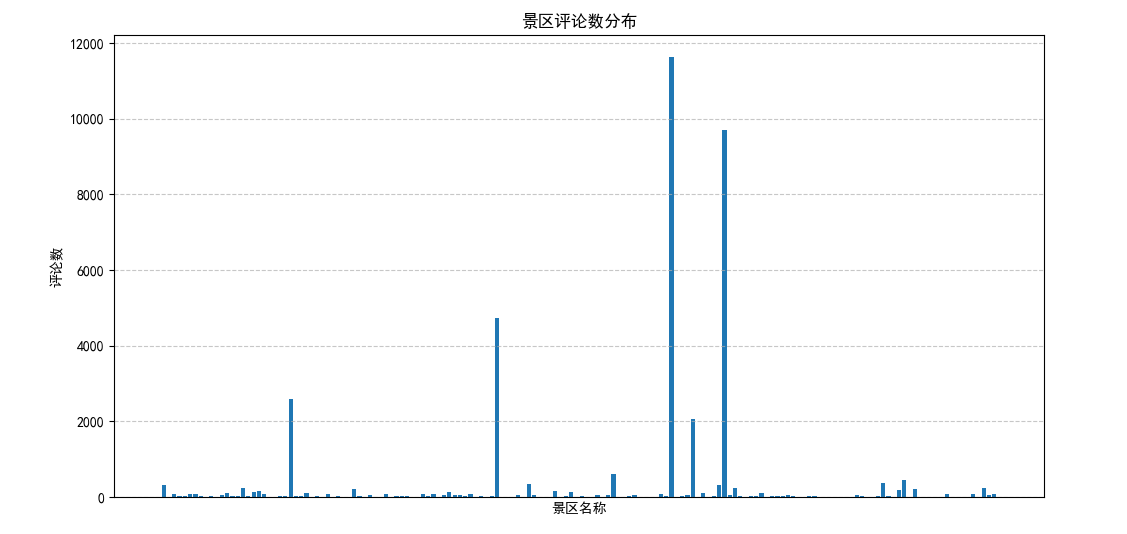
**图 4-1 热点与评分回归分析**



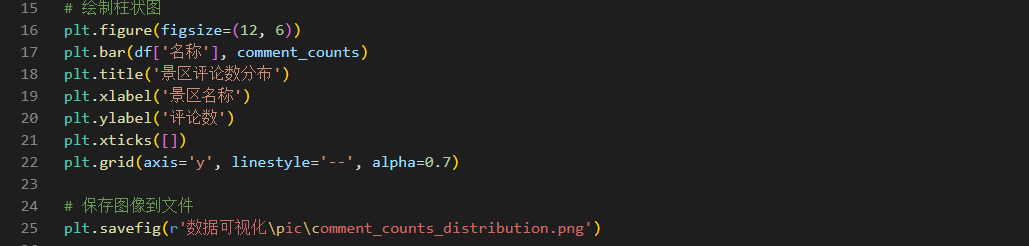
**图 4-2 实现代码**

根据图中所示，桂林该地的景区的热点与评分的线性关系不大。

## 4.3.2 评论数分析

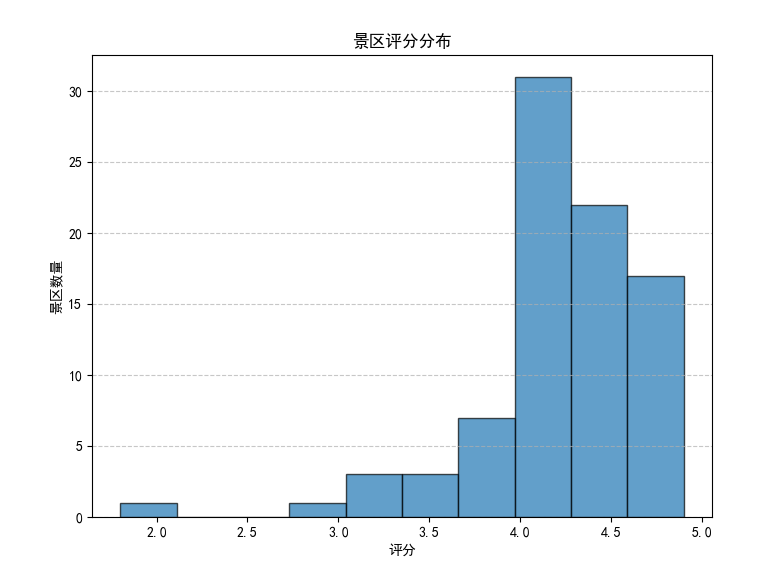


**图 4-3 景区评论数**



**4-4实现代码**

**4.3.3 景区评分分布**



**图 4-5 景区评分分布**

## 4.4 结论与建议

结论：

本文证实了协同过滤算法在桂林旅游景点推荐中的有效性。该算法利用用户的历史行为和物品之间的相似性来为游客提供个性化的景点推荐，有望提高游客的旅游体验。

个性化推荐的重要性： 个性化推荐在旅游领域尤为重要，因为每位游客的需求和兴趣都不同。协同过滤算法的使用可以帮助游客更好地选择适合他们的旅游景点，增加游客的满意度。

数据质量与性能： 为了获得更准确的推荐结果，必须确保输入数据的质量和完整性。数据的质量对于协同过滤算法的性能至关重要，包括用户评分数据和物品相似性数据。

用户隐私与数据安全： 在实际应用中，必须关注用户隐私和数据安全的问题。保护用户的个人信息是一个不可忽视的任务，需要采取适当的数据脱敏和安全措施。

建议：

1. 不断改进数据质量： 继续努力提高数据质量，包括用户评分数据和物品相似性数据。清洗和校验数据，确保数据的准确性和完整性。

2. 引入更多特征： 除了用户行为和物品相似性，可以考虑引入更多特征，如用户的个人偏好、地理位置和季节性因素等。这将进一步提升推荐的准确性。

3. 实时性的考虑： 旅游景点的热门度和用户需求可能会随时间变化。因此，可以考虑引入实时数据，以便更好地适应变化的需求。

4. 用户反馈： 收集和分析用户反馈和评价，以了解推荐系统的效果，并根据反馈进行改进。

5. 可视化和用户界面： 提供用户友好的界面和可视化工具，以便用户更好地理解推荐结果和选择旅游景点。

6. 深入研究： 进一步深入研究不同的推荐算法，如深度学习和自然语言处理，以探索更复杂和精细的推荐模型。

# 参考文献

Goldberg, D. E., Nichols, B. J., & Oki, B. M. (1992). Using collaborative filtering to weave an information tapestry. Communications of the ACM, 35(11), 61-70.

Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., & Riedl, J. (1994). GroupLens: an open architecture for collaborative filtering of netnews. Proceedings of the 1994 ACM conference on Computer supported cooperative work, 175-186.

Burke, R. (2002). Hybrid recommender systems: Survey and experiments. User Modeling and User-Adapted Interaction, 12(4), 331-370.

Sarwar, B., Karypis, G., Konstan, J., & Riedl, J. (2001). Item-based collaborative filtering recommendation algorithms. In Proceedings of the 10th international conference on World Wide Web, 285-295.

Adomavicius, G., & Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. IEEE transactions on knowledge and data engineering, 17(6), 734-749.

Linden, G., Smith, B., & York, J. A. (2003). Amazon.com recommendations: Item-to-item collaborative filtering. IEEE Internet computing, 7(1), 76-80.

Smyth, P., & Cunningham, H. (2000). Mining association rules between sets of items in large databases using a parallel tree project algorithm. Data mining and knowledge discovery, 4(4), 359-383.

Wang, X. Y. (2009). Collaborative filtering with privacy preservation: An overview of recent developments and challenges ahead. Information Fusion, 11(2), 165-177.

Zhao, Y. (2009). Research on personalized recommendation algorithm based on user interest modeling in social network service environment. Journal of Physics: Conference Series (Vol. 163, No. 1), 012088.

Zhang, L. (2015). Research and application of collaborative filtering recommendation algorithm based on big data analysis in e-commerce platform. International Journal of Business and Management Engineering Research and Applications (IJBMRE), 4(2), 25-33.

李晓光，陈华钧。协同过滤推荐系统研究综述[J].计算机学报，2006，39(8)：1747-1765。

杨林，李晓光。基于改进的协同过滤算法的网络信息资源推荐[J].小型微型计算机系统，2009，2(3)：456-459。

王亚南，张伟民，张文生。个性化推荐系统的关键技术研究[J].软件导刊，2008，(8)：4-6。

李晓光，陈华钧。一种基于用户兴趣模型的协同过滤推荐算法[J].计算机科学，2007，34(Sup)：245-248。

张建军，李晓光。基于内容分析与协同过滤的混合推荐模型研究[J].计算机工程与应用，2009，45(1)：1-4。

周志华，刘挺，程学旗。推荐系统的研究与应用[J].电脑编程技巧与维护，2007，(1)：36-38。

赵丽娟，胡瑞敏。基于用户行为分析的个性化推荐算法研究[J].科技信息，2009，(35)：368-369。

王亚南。基于协同过滤的个性化网络信息资源推荐系统设计[J].计算机知识与技术，2008，(1)：35-37。

李晓光，王亚南。一种基于用户兴趣模型的协同过滤算法[J].电子学报，2007，35(1)：166-170。

张伟民，张建军。基于用户行为的个性化信息推荐系统研究[J].计算机时代，2009，(8)：45-47。