山西应用科技学院
SHANXI COLLEGE OF APPLIED SCIENCE AND TECHNOLOGY

大数据综合课程设计

题　　目:	基于 Django 的电影评论数据可视化分析
学院:	信息工程学院
专业班级:	大数据管理与应用 2101 班
姓　　名:	郭未
学　　号:	2113051011
指导教师:	高美玲

山西应用科技学院

二〇二四年六月

目　录

基于 Django 的电影评论数据可视化分析

1 绪论

随着大数据技术的不断发展和普及，人们在日常生活中产生的数据量呈爆炸性增长。电影评论数据作为一种丰富的信息源，包含了观众对电影的各种评价和喜好。在这个信息爆炸的时代，如何从海量的电影评论中提炼有价值的信息，为用户提供更智能、个性化的电影推荐服务成为一个备受关注的问题。

本项目选取豆瓣作为数据源，结合 Python 和 Django 等先进技术，构建了一个综合性的豆瓣电影评论可视化分析推荐系统。通过对大规模评论数据的采集和处理，我们能够深入挖掘用户的观影趋势、口碑评价等信息。在这个基础上，利用数据可视化技术，以直观的图表和图形展示用户的观影偏好，为用户提供了更深入的电影分析服务。

该项目旨在结合大数据、可视化和推荐系统的技术优势，为电影爱好者提供一种全新的电影探索和选择方式，提升用户体验。通过对豆瓣电影评论数据的深度挖掘，我们能够更好地理解用户的需求，为他们提供更精准、个性化的电影推荐，推动了电影推荐系统的发展和创新。同时，项目的实施也展示了 Python/Django 等技术在构建复杂大数据系统中的卓越应用，为相关领域的研究和应用提供了有益的经验。

2 研究目的

1.深入挖掘电影评论数据：通过构建基于 Python/Django 的豆瓣电影评论可视化分析推荐系统，旨在深入挖掘电影评论数据中蕴含的用户偏好、口碑评价等信息。通过对评论数据的系统性分析，揭示用户对电影的喜好和趋势。

2.构建全面的电影信息数据库：通过爬取豆瓣电影评论数据，进行数据清洗和处理，构建一个全面而准确的电影信息数据库。该数据库将包含丰富的电影元数据，为系统提供充足的信息基础，支持后续的分析和推荐。

3.实现数据可视化展示：利用 Python 中强大的数据处理和可视化库，将分析结果

以直观的图表、图形展示给用户。通过直观的可视化展示，使用户更容易理解电影数据背后的信息，为用户提供更深入的电影分析服务。

4.设计智能化的电影推荐算法：基于对电影评论数据的深度分析，设计智能化的推荐算法。通过考虑用户的历史喜好、观影习惯等因素，为用户提供个性化、精准的电影推荐服务，提升用户体验。

5.展示 Python/Django 在大数据应用中的优越性：通过该项目的实施，展示 Python 和 Django 等先进技术在大数据应用中的卓越性能[1]。强调这些技术在构建复杂系统、处理大规模数据时的高效性和可扩展性，为相关领域的研究和应用提供实用经验[2]。

总体而言，研究旨在通过构建综合性的电影评论可视化分析推荐系统，挖掘电影评论数据的潜在价值，提升用户对电影的选择和理解体验，同时突显 Python/Django 等技术在大数据领域的应用前景。

3 开发技术介绍

Django（发音为"jan-go"）是一个高级的 Pythonweb 框架，它鼓励快速开发和干净、可重用的设计。以下是 Django 框架的一些详细介绍：

系统整体采用 Django 的框架来进行前后端的开发，Python 下有 Flask、Django、Tornado 等许多款不同可供选择的 Web 框架,其中 Django 框架是主流框架中最有代表性的一位，它最早是被用于开发管理劳伦斯集团下的一些以新闻内容为主网站的内容管理系统软件。Django 创造了许多成功的网站和应用，使用该框架可以使 Web 开发能够以最小的代价构建 Web 应用并且使其维护变的方便。同时 Django 能为频繁进行编程的模块提供了快速的解决方案[3]。

Django 是基于 MVC 的架构进行开发的，MVC 即为 Model-View-Controller（模型-视图-控制器），各部分含义如下：Model（模型）代表一个存取数据的对象及其数据模型；View（视图）代表模型包含的数据的表达方式，一般表达为可视化的界面接口；Controller（控制器）作用于模型和视图上，控制数据流向模型对象，并在数据变化时更新视图。控制器可以使视图与模型分离开解耦合。最简单的 MVC 原理图可表示如下：

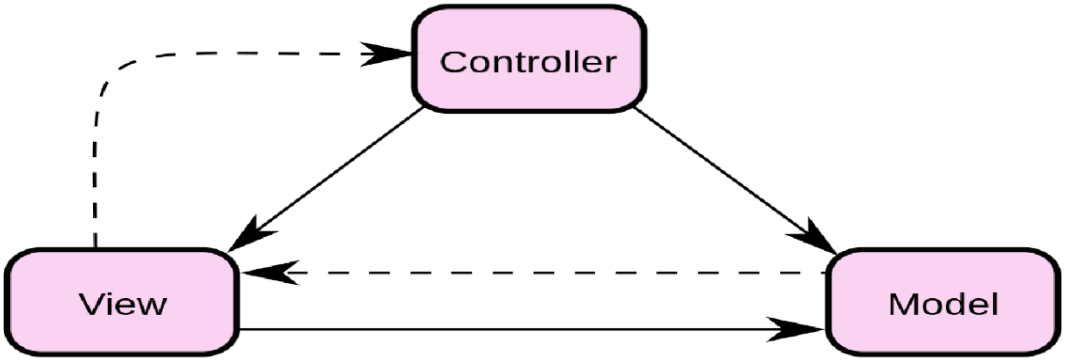

图 3-1MVC 原理图

其中包含的动作流程有：控制器创建模型；控制器创建一个或多个视图，并将它们与模型相关联；控制器负责改变模型的状态；当模型的状态发生改变时，模型会通知与之相关的视图进行更新。这是目前大型系统开发较为主流的一个过程，最直接的优点就是视图层和业务层分离，这样就允许更改视图层代码而不用重新编译模型和控制器代码，同样，一个应用的业务流程或者业务规则的改变只需要改动 MVC 的模型层即可。因为模型与控制器和视图相分离，所以很容易改变应用程序的数据层和业务规则，还有重用性高、部署快、可维护性高等优点[4]。在具体的 Django 开发中，对应的 MVC 应用可表示如下：

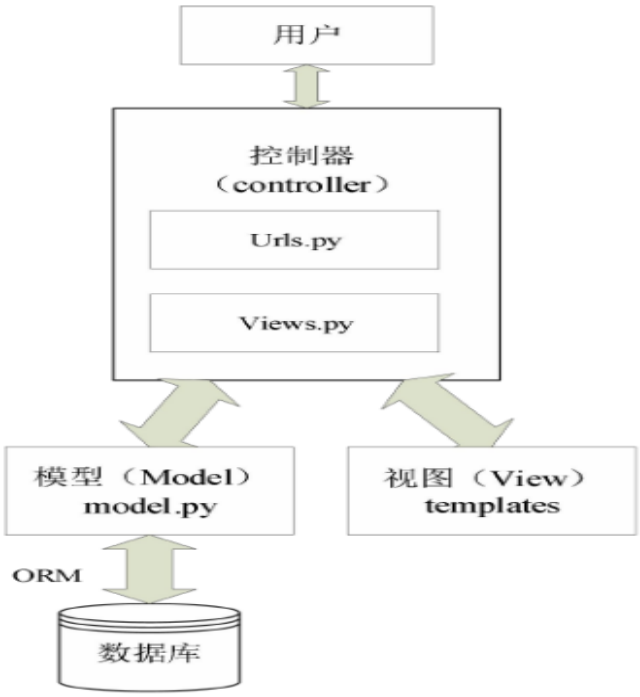

图 3-2MVC 原理图

这个图中，我们可以很清晰的看出基本符合 MVC 模式的特点，同时 Django 的体量较小，对于轻量级的系统开发和部署，具有很好的效果，不仅可以最大程度的提升开发速度，同时也可以很方便进行测试和维护。

由于采用的是 Django 的框架来进行前后端的开发，所以，本系统采用的是 B/S(Browser/Server)架构，即浏览器/服务器架构，是在 C/S（Client/Service，客户机/服务器）模式的基础上发展起来的一种体系结构，在开发 Web 应用时有明显的技术优势。针对本系统而言，可以使得系统的扩展较为容易且不需要安装专门的软件，使用也较为方便[5]。

4．软件架构的描述

软件架构模型是通过一组关键视图来描述的，同一个软件架构，由于选取的视角（Perspective）和抽象层次不同可以得到不同的视图，这样一组关键视图搭配起来可以完整地描述一个逻辑自洽的软件架构模型。一般来说，我们常用的几种视图有分解视图、依赖视图、执行视图、实现视图和工作任务分配视图。下面我们来进行具体的分析：

4．1 分解视图

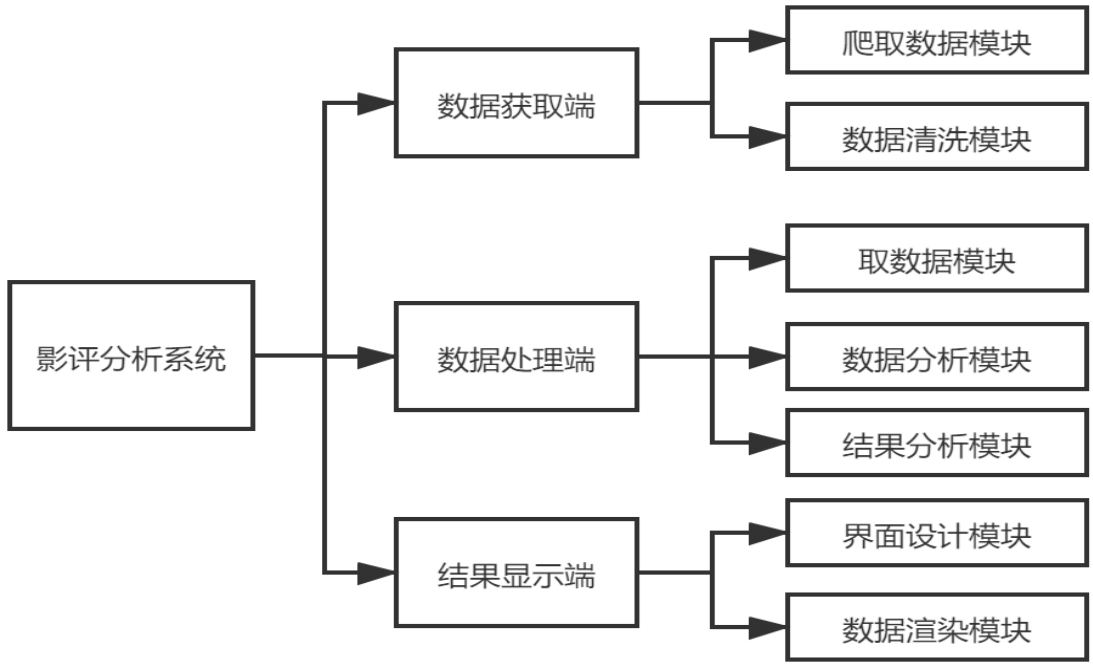

图 4-1 分解视图

分解视图也是描述软件架构模型的关键视图，一般分解视图呈现为较为明晰的分解结构（breakdownstructure）特点。简单的说也就是将系统的功能进行分解，形成几个小的部分，然后这些小部分又包含各自所要处理的业务，我们所要做的就是对这些具体业务的设计和开发。本系统中，即分解为数据获取端、数据处理端、结果显示端三个大模块，然后在各个模块内部细分为一些小的功能类，这样使得系统的结构较为清晰，同时也能看清楚各层次之间的联系，使得开发更加系统化。

4．2 依赖视图

依赖视图展现了软件模块之间的依赖关系。比如一个软件模块 A 调用了另一个软件模块 B，那么我们说软件模块 A 直接依赖软件模块 B。如果一个软件模块依赖另一个软件模块产生的数据，那么这两个软件模块也具有一定的依赖关系。本系统中，三大模块之间均存在数据的传递和调用，所以产生如下的依赖视图：

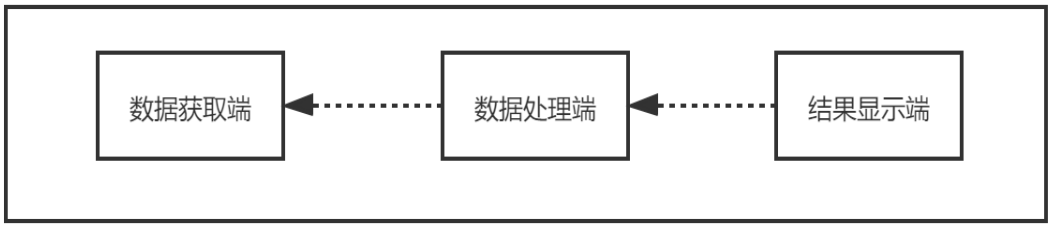

图 4-2 依赖视图

即数据处理端依赖于数据获取端从影评网站上爬取的数据来进行具体的数据分析，并给出最终生成的结果传递给结果显示端；而结果显示端也依赖于数据处理端传递来的数据来进行页面的数据渲染。

4．3. 执行视图

执行视图展示了系统运行时的时序结构特点，比如流程图、时序图等。执行视图中的每一个执行实体，一般称为组件（Component），都是不同于其他组件的执行实体。执行实体可以最终分解到软件的基本元素和软件的基本结构，因而与软件代码具有比较直接的映射关系。在设计与实现过程中，我们一般将执行视图转换为伪代码之后，再进一步转换为实现代码。根据系统的各部分的功能实现，我们画出系统的整体流程图如下：

可见，上图中基本概括了系统的执行时序和整体的动作，作为偏算法类的项目，所以具体的分析处理模块未明确列出，但是对于数据的传递路径以及结果的回调等，

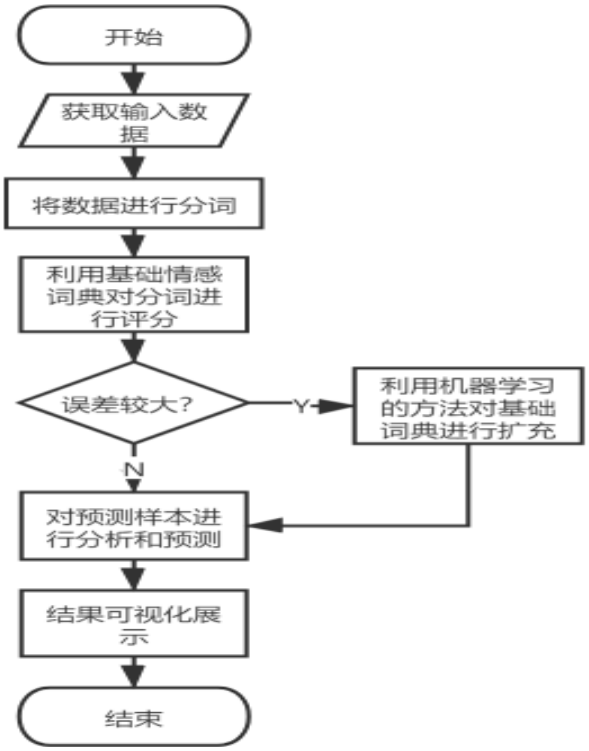

图 4-3 系统的整体流程图

均有了一个简单的执行顺序。

4．4. 实现视图

实现视图是描述软件架构与源文件之间的映射关系。比如软件架构的静态结构以

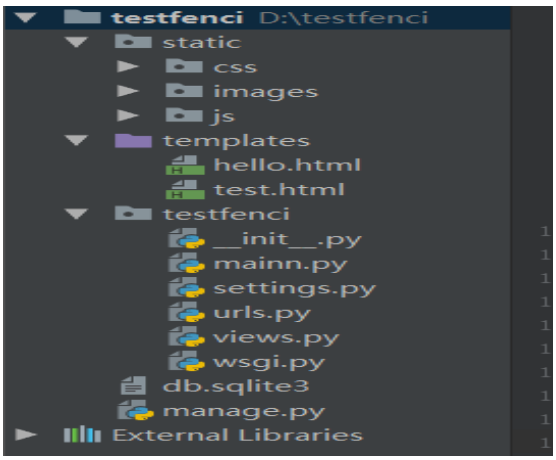

图 4-4 系统源代码的目录文件结构图

包图或设计类图的方式来描述，但是这些包和类都是在哪些目录的哪些源文件中具体实现的呢？一般我们通过目录和源文件的命名来对应软件架构中的包、类等静态结构单元，这样典型的实现视图就可以由软件项目的源文件目录树来呈现。

上图即为系统源代码的目录文件结构，从上往下依次为 css 文件，图片文件和 js

文件，其中分别对应着样式、图片和网页动作等；下面的包即为模板包，内含各个具体的前端页面，即相当于 MVC 中的 V 即视图模块，能将数据进行具体的展示；再下一个文件夹为 Django 框架中的处理相关的文件，内含默认初始化文件、测试文件、默认设置文件、url 路径文件、逻辑实现文件以及部署文件,url.py 和 views.py 两个文件的功能即相当于 MVC 中的 controller 的作用，其中具体的数据分析模型和取数据的模块嵌入在 views.py 当中；再往下即为数据库和全局管理文件。

实现视图有助于我们在海量源代码文件中找到具体的某个软件单元的实现，因为会使得具体的实现代码的层次更加清晰，对源代码的某个模块的定位也会更加精准。实现视图与软件架构的静态结构之间映射关系越是对应的一致性高，越有利于软件的维护，因此实现视图是一种非常关键的架构视图。

4．5. 工作分配视图

工作分配视图将系统分解成可独立完成的工作任务，以便分配给各项目团队和成员。工作分配视图有利于跟踪不同项目团队和成员的工作任务的进度，也有利于在个项目团队和成员之间合理地分配和调整项目资源，甚至在项目计划阶段工作分配视图对于进度规划、项目评估和经费预算都能起到有益的作用。由于本系统的模块区分较为明显，即为数据获取、数据处理、结果展示三大模块，所以组内的三名成员分工也较为明确：

表 4-1 成员分工表

数据获取和清洗	成员 A
数据处理和分析	成员 B
结果可视化和系统维护	成员 C

表 4-2 系统开发的进度安排表

时间	工作	阶段成果
1个月	查找文献资料、完成相关技术研究	概要设计报告
1.5个月	完成评分系统的详细设计	详细设计报告
3.5个月	进行系统的的代码编写，测试	系统代码，完整的系统
1个月	整理文档，撰写工程实践论文，进行最终答辩	工程实践论文

同时，针对本系统开发的进度安排，小组进度计划如下表所示：

商标即为整体规划，整个系统大概 7 个月的时间完成，包括系统的源代码和伴随的说明文档以及开发文档等。

5．实验步骤

5.1 数据获取

为了支持对于豆瓣电影中的电影、用户以及评分评论等信息的分析与可视化，我们需要从豆瓣电影网站中获取大量相关数据。经过对豆瓣电影网站的内容和框架的分析与探索，我们发现了网页中数据出现的规律，并最终通过网络爬虫爬取的方式获取所需数据。

表 5-1 相关数据分布表

A 电影列表	B 电影的详细信息；
C 用户信息	D 用户影评列表

首先，对于电影数据，我们期望选取豆瓣电影中的不同类型、不同地区和不同地区的具有代表性的电影作品。对于代表性，什么样的电影具有代表性？由于本文不仅仅只是分析电影的内容数据，还需要将电影的内容与用户的交互行为相结合，得到用户对于电影的兴趣爱好变化以及对影评内容分析。因此，这里的代表性，不仅仅指的是电影的精彩程度、票房数量等指标，而且包含了豆瓣电影社区中用户对于该电影的关注程度。为了进一步分析用户的行为，每部电影都应该具有一定数量的用户评分、短评以及影评基础，这样才能够正确的进行分析而不会因为样本太少产生偏差。下面对文中的数据获取方法进行详细的讲解。

我们对数据的爬取主要分为一下三个步骤：

（1）获取各个时期的具有代表性的电影在该步骤中，我们首先对豆瓣电影中的电影进行筛选。这里主要是根据电影评分、用户短评数量、用户评论数量等定量指标来进行筛选的。除此之外，我们还对不同时间段的电影进行挑选，以满足每个时间段中都包含一定数量的电影。

（2）获取选取电影中具有代表性的用户代表性用户的选择基于选出的代表性电影。对于第一步中选择的每一部电影，我们分析并获取对电影具有评分或评价的用户列表。这样，对于每一部电影，我们都得到一个用户的列表，通过将所有电影的用户

列表进行整合，我们得到了一个所有用户的活跃度列表。根据该用户活跃度列表，我们选择适量的用户进行分析。

（3）获取选取电影的影评数据在获取选取电影评论数据的过程中，我们需要保证电影是在第一步中的选取电影列表中，同时影评的用户在第二步中的用户列表中。由于每部电影的影评数据量较大，可以考虑对其进行筛选。在数据的获取过程中，最理想的状况是获取豆瓣电影中的所有电影以及用户数据，并以此获得所有的用户评论数据。这样便能够毫无误差的对于这些数据进行基于电影以及基于用户的分析，符合大数据时代的数据分析特点。但是，由于人力、时间、数据获取途径等诸多因素的限制，我们并不能够获得上述理想数据集。在进行电影筛选时，我们最初使用人工的手动选择。该方法具有很强的局限性，需要大量的人力，并且耗时较长。

在对手动挑选的电影数据进行分析后，我们发现：满足我们挑选条件的电影都是豆瓣评分较高的热门电影。据此，我们直接选取了豆瓣电影 TOP250 作为数据集的一部分。在获取代表性用户的过程中，我们发现每部电影的评价和评论人数都是极为庞大的。例如 1994 年电影肖申克的救赎，其中包含了 62 万多人的评论，14 万余条短评数据以及 3391 条的影评数据。为了简化数据获取难度，我们只对影评用户进行筛选，最终获得了前 100 名活跃用户。3.数据处理在获取到豆瓣电影的部分数据后，我们对这些数据进行了加工和处理，以方便之后的分析以及可视化。3.1 电影类别划分首先，获取的数据包含了上映时间、类别、国家地区、用户、评分、影评等信息。在经过观察后，我们发现，数据中的类别和国家地区信息包含了太多不同的属性，这导致了很多类别中只包含一部甚至不包含相关的电影，这给之后的可视化产生了一定的影响，为此，我们将主要的类别进行筛选，使得类别的数量在一个合适的范围。

首先，对于电影类型，我们选取了剧情、爱情、喜剧、犯罪、冒险、奇幻、动作、动画、悬疑、家庭、惊悚、科幻和战争总共 13 个类型。这样，每部电影都包含了其中一个或多个类型标签，不包含在类型中用其他类型表示。对于国家和地区属性，我们采用相同的方法。我们使用了数据集中包含的几个常见国家或地区：美国、英国、法国、日本、香港、德国、中国大陆、意大利、韩国。其余地区的电影被划分为其他类型。对于电影的年代属性，我们一共设了 9 个类型，1930s 到 2010s。经过分类后，数据中包含的信息更加紧密，同时也为之后的分类显示以及相似度计算提供了方便。3.2 情感分析在情感分析过程中，我们通过词典构造与统计方法相结合的情感词选择方法，

解决了计算复杂度高，通常没有考虑种子词的强度两个问题。我们使用的影评数据共有 3525 条，通过构建字典，以及使用结巴分词工具，我们统计了 3525 条影评中出现频率高的词语，并过滤掉了一些介词，标点符号等，我们留下了其中词频最高的 1000 个词并按照从词频从达到小排列出来，最后通过人工的筛选出情感词。我们使用词库里的情感词用来遍历所有的影评数据，对每条影评都统计影评中情感词语出现的个数程度。最终，我们得到了所有用户影评的感性理性分类信息，为之后的情感可视化提供了数据基础。4.可视化交互方案设计和效果在获取数据并对数据进行处理之后，便进入了可视化与交互方案设计这一核心环节。在该环节中，我们根据可视化目标，设计了三个可视化任务，分别展示所有用户之间的关系，用户与其邻居关系的可视化和用户信息可视化。

5.2 导入

首先，确保您已经安装了 Python 解释器。您可以从 Python 官方网站下载适合您操作系统的版本，并按照安装指南进行安装。在安装完 Python 后，我们需要安装一些必要的库，包括 requests 和 BeautifulSoup。这些库可以通过 pip 包管理器进行安装。在命令行中运行以下命令来安装这些库：

```
pip install requests
pip install beautifulsoup4
```

安装完成后，我们可以在 Python 脚本中导入这些库：

5.3 分析目标网页

```
import requests

from bs4 import BeautifulSoup
```

我们将使用 requests 库发起 HTTP 请求，获取豆瓣电影页面的 HTML 内容，并使用 BeautifulSoup 库解析 HTML 结构。

首先，我们需要指定目标电影的 URL 地址，并使用 requests 库发送 GET 请求以

```
url = 'https://movie.douban.com/subject/26631790/?from=showing'

response = requests.get(url)
```

获取页面的 HTML 内容：

上述代码中，我们使用 requests 库的 get 函数发送 GET 请求，并将返回的响应

保存在 response 对象中。

接下来，我们需要使用 BeautifulSoup 库解析 HTML 内容，以便提取我们所需的评论数据。我们可以使用 html.parser 解析器来解析 HTML 文档：

```
soup = BeautifulSoup(response.text, 'html.parser')
```

上述代码中，我们将响应的文本内容传递给 BeautifulSoup 的构造函数，并指定解析器为 html.parser。

现在，我们已经成功解析了 HTML 文档，并可以继续提取评论数据。

5.4 提取评论数据

通过分析豆瓣电影页面的 HTML 结构，我们可以定位到包含评论的 HTML 元素，并使用 BeautifulSoup 提供的方法提取用户名和评论内容。

首先，我们可以通过查看页面的源代码或使用浏览器的开发者工具来定位评论所在的 HTML 元素。在豆瓣电影页面中，评论通常位于具有 comment-item 类名的 div 元素中。

```
comments = soup.find_all(class_='comment-item')
```

上述代码使用 BeautifulSoup 的 find_all 方法根据类名查找所有包含评论的 div 元素，并将结果保存在 comments 列表中。

接下来，我们可以使用循环遍历 comments 列表，并提取每个评论的用户名和评论内容：

```
for comment in comments:
    username = comment.find(class_='comment-info').find('a').get_text(strip=True)
    content = comment.find(class_='comment-content').get_text(strip=True)
    print(f'Username: {username}')
    print(f'Content: {content}')
    print('---')
```

上述代码中，我们使用 find 方法和类名 comment-info、comment-content 来定位用户名和评论内容的 HTML 元素。然后，我们使用 get_text 方法提取元素的文本内容，并使用 strip 参数去除首尾的空白字符。在循环中，我们打印每个评论的用户名和内容，并使用分隔线进行分隔。

现在，我们已经成功提取了豆瓣电影页面中的评论数据。

5.5 错误处理和异常处理

在爬虫过程中，可能会遇到请求错误、页面解析错误等异常情况。为了确保爬虫的稳定性和可靠性，我们需要合理处理这些异常。我们可以使用 try-except 语句捕获异常，并添加相应的错误处理逻辑。在遇到异常时，我们可以记录错误日志、重新发送请求或采取其他措施来处理异常情况。

```
try:
    response = requests.get(url)
    response.raise_for_status()
    soup = BeautifulSoup(response.text, 'html.parser')
    # 提取评论数据的代码
except requests.exceptions.RequestException as e:
    print(f'Request error: {e}')
except Exception as e:
    print(f'Error: {e}')
```

5.6 数据处理和存储

在提取评论数据后，我们可能需要进行一些数据处理和清洗操作，例如去除空白字符和特殊符号，过滤和去重数据。根据实际需求，您可以自定义数据处理的逻辑。

此外，我们还可以选择将提取的评论数据存储到本地文件或数据库中，以便后续分析和使用。可以使用 Python 提供的文件操作函数，如 open、write，将数据写入文

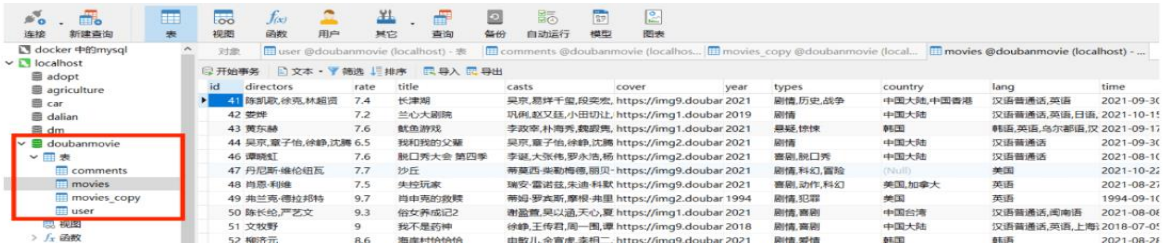

图 5-6 保存到数据库

本文件。或者，可以使用数据库库，如 SQLite 或 MySQL，将数据存储到数据库中。

5.7 数据分析与可视化

可视化是数据分析的重要部分，它可以帮助我们更好地理解和解释数据。在 Python 中，有多个库可以用来进行数据可视化，如 matplotlib、seaborn 和 plotly 等。

对于豆瓣电影 TOP250 的列表，我们可以创建一个条形图来展示前 250 部电影的

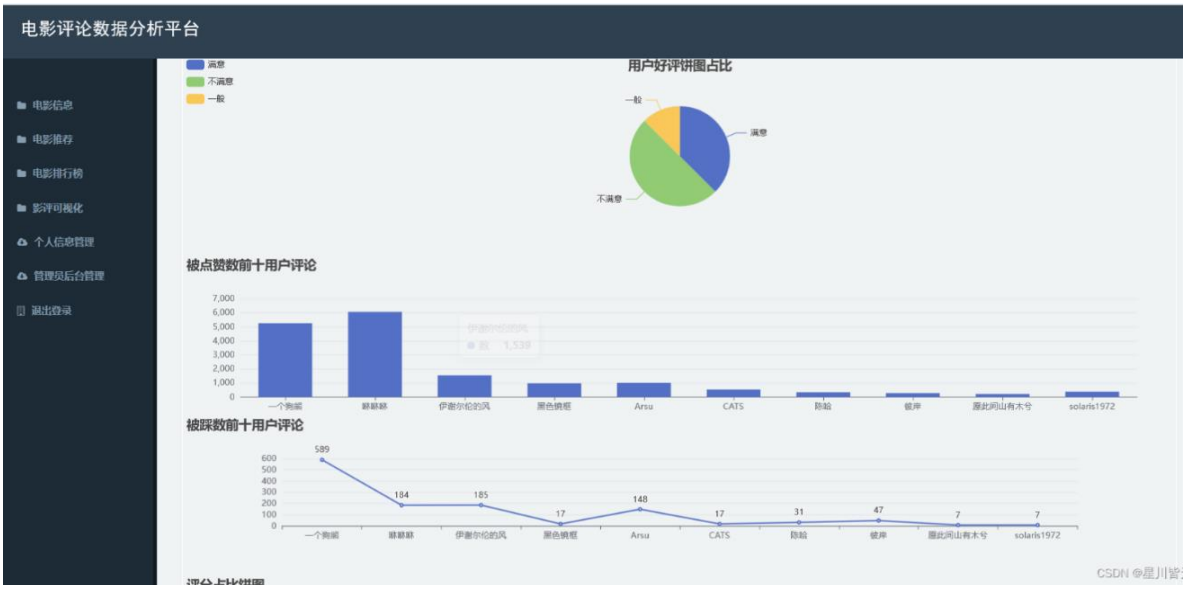

图 5-1 分析结果图

排名。我们还可以使用折线图柱状图来展示电影的平均评分和评分人数的关系。此外，我们还可以创建词云图来显示最常被提及的电影主题和关键字。

图 7-2 评分前十电影评论词云图

6. 概念原型的工作机制

经过上述的分析，可以总结出概念原型的工作机制，简而言之就是用户在系统中选择自己想看的电影，系统根据用户选择的电影从数据库中找到对应的评论，进而对这些评论进行分析，给出最后分析的电影评分和建议，供用户观影之前的参考；我们开发小组也会对该系统进行不断地测试和维护，当系统的评分模型出现较大的误差时，我们便会从数据和算法模型入手，去做数据清洗工作或者算法的改进工作，提高评分的准确率。

7 结论

经过本次的对软件系统的结构特点和架构风格的分析，我对自己的工程实践的开发又有了一些新的思路和认识，尤其是利用各种视图来进行软件系统概念原型的描述时，会让我们更加深入的去分析我们这个系统的目标是什么、我们怎么实现这个目标，尤其是算法类的项目，我们要达到的精度以及完整度才是我们应该追求的，而不仅仅是说为了去分析而分析。经过两次的概念原型的建模和分析，我对软件工程的方法也有了一定的深入了解，现在也会慢慢的思考这些方法的目的和优点所在，知其然也要知其所以然，在以后的需求分析和原型设计中，也要时刻想着利用这些常用的方法去进行分析和设计。

参考文献

[1]蔡文乐,周晴晴,刘玉婷,等基于 Python 爬虫的豆瓣电影影评数据可视化 2021,5(18):5DOI:1019850/j.cnki.2096-4706.2021.18.022.

[2]蔡文乐等."基于 Python 爬虫的豆瓣电影影评数据可视化分析."现代信息科技 5.18(2021):5.

[3]蔡文乐,周晴晴,刘玉婷,&秦立静.(2021).基于 python 爬虫的豆瓣电影影评数据可视化分析.现代信息科技,5(18),5.

[4]高巍,孙盼盼,and 李大舟."基于 Python 爬虫的电影数据可视化分析."沈阳化工大学学报 34.1(2020):6.

[5]孙建立&贾卓生.(2017).基于 Python 网络爬虫的实现及内容分析研究.中国计算机用户协会网络应用分会 2017 年第二十一届网络新技术与应用年会论文集.

...

致谢

感谢山西应用科技学院对我几年的培养！

感谢信息工程学院的各位老师各位领导对学子的教导，让学生掌握了基本的专业知识与技能！

感谢高美玲老师对我在学术上的谆谆教诲。让我不仅学到了知识，而且学到了做人的准则和严谨的治学作风。

在此，我表示衷心的感谢和崇高的敬意！

附录

```
#根据用户推荐信息给其他人
defrecommend(self,user):
try:
#相似度最高的用户
top_sim_user=self.top10_simliar(user)[0][0]
print(top_sim_user)
#相似度最高的用户的观影记录
items=self.data[top_sim_user]
recommendations=[]
#筛选出该用户未观看的信息并添加到列表中
foriteminitems.keys():
ifitemnotinself.data[user].keys():
recommendations.append((item,items[item]))
recommendations.sort(key=lambdaval:val[1],reverse=True)#按照评分排序
#返回评分最高的 10 部信息
iflen(recommendations)==1:
recommendations=[]
lists=[]
forkey,valueinself.data.items():
forkeys,valuesinvalue.items():
lists.append((keys,values))
foriinrange(4):
recommendations.append(random.choice(lists))
recommendations=list(set(recommendations))
returnrecommendations[:10]except:return"
```