**任课教师：张涵翠**

《Python程序设计》

**(2024-2025第1学期)**

**大作业实验报告**

**题目名（题号）：票房解读（5）**

**姓名（学号）：张雅瑞（2022334323029）、詹硕 （2022331201094）、闫可（2022331201066）、李婉洁（2022330300274）、任欣菡（2022330300093）**

**班级：*计算机科学与技术22(1)班、计算机科学与技术22(3)班、计算机科学与技术22(4)班***

# 摘要

随着互联网的不断发展和数据量的迅猛增长，网络爬虫技术已成为获取各类信息和大规模数据的关键手段。Python作为当前最受欢迎的编程语言，凭借其丰富的第三方库和框架，已成为开发网络爬虫的首选语言。Python网络爬虫技术在搜索引擎、电商平台、社交媒体、新闻报道等多个领域得到了广泛应用，推动了信息科技和数字化时代的快速发展。

本课题旨在分析并挖掘近年电影票房市场的趋势与特征。数据来源于The Numbers电影网站，该网站以系统化、算法化的方式追踪票房数据，具有较高权威性。研究从该网站获取了2020至2022年每年Top 100电影的详细数据，包括电影名称、上映时间、类型和票房收入等内容。

课题通过Python的Requests和BeautifulSoup库实现爬虫，以获取网站数据，并运用pandas对数据进行清洗、去重、分类和统计分析。主要分析内容包括票房总和、类型分布、票房排名、上映日期与票房的关联等，确保数据的全面性和有效性。

进一步利用OpenAI GPT模型对处理后的电影数据进行智能分析，并总结了电影市场的整体趋势，探索未来票房的热点。通过连接AI API与Python脚本，系统能够自动分析数据表、解读图表内容，并生成电影市场的季度票房总结和未来趋势预测。同时，系统支持用户输入自定义问题，调用AI提供基于数据的响应，从而实现人机交互。

**关键词**：电影票房；数据分析；大语言模型；爬虫；未来票房预测

# 课题执行（文档修订）记录表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 日期 | 版本 | 修订内容 | 执行人 | 修订人 |
| *2024.10.23* | *1.0* | *完成课题框架搭建和分工部署* | *张雅瑞* | *张雅瑞* |
| *2024.11.8* | *1.1* | *完成数据获取代码实现和报告撰写* | *詹硕* | *詹硕* |
| *2024.11.10* | *1.2* | *完成数据筛选代码实现和报告撰写* | *闫可* | *闫可* |
| *2024.11.10* | *1.3* | *完成数据分析代码实现和报告撰写* | *任欣菡* | *任欣菡* |
| *2024.11.13* | *1.4* | *完成报告摘要总结并整合* | *李婉洁* | *李婉洁* |

目录

[1.引言 6](#_Toc182421148)

[1.1．课题名称及其描述 6](#_Toc182421149)

[1.2. 课题目的及意义 6](#_Toc182421150)

[1.3．课题应用场景及特点 6](#_Toc182421151)

[2．数据获取 7](#_Toc182421152)

[2.1. 数据来源 7](#_Toc182421153)

[2.2. 爬虫基本工作原理 7](#_Toc182421154)

[2.2.1 爬虫工具 7](#_Toc182421155)

[2.2.2 爬虫基本工作原理 7](#_Toc182421156)

[2.3. 爬取流程 8](#_Toc182421157)

[2.3.1 确定目标界面和数据 8](#_Toc182421158)

[2.3.2 请求网页阶段 8](#_Toc182421159)

[2.3.3 获取网页内容 8](#_Toc182421160)

[2.3.4 解析HTML文档 8](#_Toc182421161)

[2.3.5提取所需数据 8](#_Toc182421162)

[2.3.6存储数据 8](#_Toc182421163)

[3．数据筛选 9](#_Toc182421164)

[3.1. 数据筛选背景和目标 9](#_Toc182421165)

[3.2. 数据筛选内容与流程 9](#_Toc182421166)

[3.2.1 数据读取与初步处理 9](#_Toc182421167)

[3.2.2 数据去重与整合操作 9](#_Toc182421168)

[3.2.3 数据分析与统计分析 9](#_Toc182421169)

[3.2.4 数据可视化与结果输出 11](#_Toc182421170)

[4.数据分析 11](#_Toc182421171)

[4.1 数据分析背景与目标 11](#_Toc182421172)

[4.2 数据分析内容与流程 11](#_Toc182421173)

[4.2.1 Python与AI的连接设置 11](#_Toc182421174)

[4.2.2 数据分析流程 12](#_Toc182421175)

[4.2.3 自动回答与用户交互 12](#_Toc182421176)

[4.3 代码说明 12](#_Toc182421177)

[5. 结束语 13](#_Toc182421178)

[5.1. 总结 13](#_Toc182421179)

[5.2. 不足与展望 13](#_Toc182421180)

[6. 参考文献 13](#_Toc182421181)

# 1.引言

## 1.1．课题名称及其描述

课题主题为“基于大语言模型的电影票房数据分析系统设计与实现”。本课题旨在设计一个系统化的数据分析流程，使用爬虫技术获取电影网站票房相关数据，并结合大语言模型对电影市场的整体数据进行分析和解读。通过对不同年份票房前100电影的标题、票房收入、上映日期和类型等数据进行深入分析，探讨电影票房趋势和观众偏好，挖掘影响票房表现的潜在因素。

## 1.2. 课题目的及意义

随着互联网的不断发展和数据量的迅猛增长，网络爬虫技术已成为获取各类信息和大规模数据的关键手段。Python凭借其丰富的第三方库和框架，已成为开发网络爬虫的首选语言。Python网络爬虫技术在搜索引擎、电商平台、社交媒体、新闻报道等多个领域得到了广泛应用，推动了信息科技和数字化时代的快速发展。[1]

本课题的目的在于通过数据分析技术与人工智能模型的结合，对电影票房数据进行系统化分析，以帮助研究者和从业者洞察电影市场的发展趋势和观众的观影需求。通过对票房、类型、上映日期等多维度的分析，可以识别电影市场中的高票房类型和上映档期的特点，为电影发行方提供决策支持。同时，利用大语言模型自动生成数据分析总结和预测，有助于提升数据处理效率，为电影市场研究提供一种智能化解决方案。

## 1.3．课题应用场景及特点

本系统适用于电影数据分析、市场研究和电影投资决策等领域。其特点包括：

1.自动化数据获取与处理：通过爬虫技术自动采集数据并清洗、去重，为后续分析提供高质量数据集；

2.智能化分析与解读：结合大语言模型对数据进行专业分析和总结，支持自动生成市场分析报告和未来趋势预测；

3.人机交互功能：系统支持用户提出自定义问题，由AI提供基于历史数据的回答，从而提升数据应用的灵活性和实用性。

本课题为电影行业提供了一个创新的数据驱动分析工具，不仅提高了数据处理效率，还增强了市场分析的智能化程度，对电影行业的发展有积极的推动作用。

## 2．数据获取

## 2.1. 数据来源

数据来源是 The number电影网站，该网站在电影票房数据方面具有一定的权威性。The number是一家位于[美国](https://zh.wikipedia.org/wiki/%E7%BE%8E%E5%9B%BD" \o "美国)[加利福尼亚州](https://zh.wikipedia.org/wiki/%E5%8A%A0%E5%88%A9%E7%A6%8F%E5%B0%BC%E4%BA%9E%E5%B7%9E)[比佛利山](https://zh.wikipedia.org/wiki/%E6%AF%94%E4%BD%9B%E5%88%A9%E5%B1%B1_(%E5%8A%A0%E5%88%A9%E7%A6%8F%E5%B0%BC%E4%BA%9E%E5%B7%9E))的[电影行业](https://zh.wikipedia.org/w/index.php?title=%E7%94%B5%E5%BD%B1%E8%A1%8C%E4%B8%9A&action=edit&redlink=1)[数据](https://zh.wikipedia.org/wiki/%E6%95%B0%E6%8D%AE)[网站](https://zh.wikipedia.org/wiki/%E7%BD%91%E7%AB%99)，以[系统化](https://zh.wikipedia.org/w/index.php?title=%E7%B3%BB%E7%BB%9F%E5%8C%96&action=edit&redlink=1)、[算法化](https://zh.wikipedia.org/w/index.php?title=%E7%AE%97%E6%B3%95%E5%8C%96&action=edit&redlink=1)的方式追踪[票房](https://zh.wikipedia.org/wiki/%E7%A5%A8%E6%88%BF)[收入](https://zh.wikipedia.org/wiki/%E6%94%B6%E5%85%A5)[2]。它专注于提供电影市场相关的数据和分析,数据比较完整性，适合用爬虫爬取该网站上特定年份的票房排行榜，可以获取到电影的排名、标题、票房收入、上映日期、类型等内容。

项目中具体数据来源是来自The number网站2020年到2022年每一年top前100的电影，其中内容包括电影名称、发行时间、电影类型、票房。

## 2.2. 爬虫基本工作原理

### 2.2.1 爬虫工具

项目主要使用了Requests库和BeautifulSoup库。Requests技术属于爬虫基础性工具包，它模拟人输入网址向服务器递交网络请求，实现自动爬取HTML网页页面信息的功能[3]。在代码中，通过设置User-Agent请求头来模拟浏览器行为，增加请求的真实性，降低被网站识别为爬虫而被拒绝访问的概率。BeautifulSoup技术负责解析HTML和XML文档，提取和处理网页中的数据，对于复杂的 HTML 文档，尤其是包含多层嵌套结构的页面，BeautifulSoup库能够轻松地处理，帮助开发者准确地提取所需的数据。

### 2.2.2 爬虫基本工作原理

项目先使用Requests库发送 HTTP 请求到指定的网站特定页面，并设置User-Agent请求头以模拟浏览器行为，避免被网站识别为爬虫而被拒绝访问。等到获取到网页的响应内容后，再使用BeautifulSoup库将 HTML 文档解析成一个结构，以便于从中提取所需的数据。通过定位到包含电影数据的表格元素，并遍历表格的行和列，提取出各个电影的相关信息包括排名、标题、票房收入等。

## 2.3. 爬取流程

### 2.3.1 确定目标界面和数据

确定爬取的目标页面的网址，需要爬取不同年份的数据，然后再确定爬取具体数据内容，包括电影标题、票房、上映日期、类型等。

### 2.3.2 请求网页阶段

项目使用Requests库向目标网页发送HTTP GET请求，请求中包含用户代理（User-Agent），这可以帮助模拟真实用户的浏览器，从而避免被网站的反爬虫机制阻止，降低降低被网站识别为爬虫的概率。

### 2.3.3 获取网页内容

当服务器接收到请求，它会返回一个HTTP响应，其中包含网页的HTML内容。如果请求成功，代码会获取到该网页的内容；如果请求失败，则获取不到网页内容，就会运行失败。

### 2.3.4 解析HTML文档

项目使用BeautifulSoup技术将获取到的网页响应内容解析成一个 BeautifulSoup对象，可以将HTML文档转换为一个结构化的对象，方便后续的数据提取。

### 2.3.5提取所需数据

项目通过find()或find\_all()等方式提取各个数据字段的值，包括排名、电影标题、票房收入、上映日期、类型等，并将这些数据保存到一个列表中，易于获取。

### 2.3.6存储数据

项目使用pandas库将数据保存到Excel文件中并且命名为top\_movies.xlsx，这使得数据可以方便地进行后续分析和查看。

# 3．数据筛选

## 3.1. 数据筛选背景和目标

项目主要工作是爬取近3~5年某电影网的Top100电影信息，分析电影数据、票房情况。数据来源为 The number 电影网站，数据筛选工作主要负责对获取到的票房原始数据进行处理与分析，为后续研究提供精准、有效的数据支持。

## 3.2. 数据筛选内容与流程

### 3.2.1 数据读取与初步处理

根据数据获取阶段存储为的 Excel 文件 top\_movies.xlsx，使用 pandas 库的 read\_excel 函数读取该文件，获取所有工作表的数据。对表中数据进行预处理，将 Box Office 列中的逗号去除，同时将该列转换为数值类型，确保数据在计算和分析时的准确性。[4]

### 3.2.2 数据去重与整合操作

因为有多部电影在读取的三年票房数据中多次出现，为保证数据的唯一性，避免重复数据影响分析结果，以 Title 列为依据对数据进行分组。计算每组中 Box Office 列的总和，得到每个电影的总票房。

然后，将去重后的原始数据与分组求和后的数据进行合并，确保在去除重复记录的同时，保留了准确的票房汇总信息，得到一个更干净、完整的数据集 data。

### 3.2.3 数据分析与统计分析

#### 3.2.3.1 按年份统计票房总和

计算每年的票房总和，得到 yearly\_box\_office 数据。有助于观察不同年份电影市场的总体票房表现，为后续分析票房的年度变化趋势提供基础数据。

#### 3.2.3.2分析电影类型分布

得到电影类型的分布情况，这使我们能够清晰了解市场上各类电影的占比，从而洞察观众对不同电影类型的喜好倾向。

#### 3.2.3.3 筛选票房排名前五十电影

对 data 数据集进行降序排序，选取前五十部电影，并为其添加排名信息，生成 top\_50\_movies 数据集。该数据集涵盖了排名、电影名称、票房、上映日期及年份等关键信息，为进一步研究高票房电影的特征和市场表现提供重要的数据样本。

#### 3.2.3.4 探究电影上映日期与票房关系

以“电影上映日期”列为分组依据，对“票房”列进行求和操作，得到 release\_date\_box\_office 数据。通过分析这组数据，可以探索电影上映日期与票房之间是否存在某种关联或规律，为电影发行策略的研究提供参考。[5]

#### 3.2.3.5 对电影票房进行分类统计

根据票房数值大小将电影票房分为不同类别。将该分类规则应用到 data 数据集的 Box Office\_y 列，新增 Box Office Class 列记录分类结果，并通过统计各分类的数量，得到 box\_office\_classification 数据。这有助于从票房量级的角度分析电影市场的结构和分布特点。

#### 3.2.3.6 对电影发布日期进行分类统计

根据电影上映日期中的月份信息将其划分为四个季度。将该分类规则应用到 data 数据集的 Release Date 列，新增 Release Date Class 列记录分类结果，并统计各季度电影数量，得到 release\_date\_classification 数据。这有助于研究电影上映时间的季节性分布及其对票房的潜在影响。

#### 3.2.3.7 电影发布日期与电影类型联合分类统计

对 data 数据集按照 Release Date Class 和 Genres 两列进行分组，统计每组的数量，得到 release\_date\_genre\_classification 数据。这组数据能够呈现不同季度下各类电影的分布情况，为深入分析电影市场的细分特征和规律提供有力支持。

### 3.2.4 数据可视化与结果输出

为了更直观地展示数据处理和分析结果，创建一个新的 Excel 文件 top\_movies\_analysis\_charts.xlsx。

利用 pandas 的 ExcelWriter 功能将各类数据分别写入相应的工作表，包括 yearly\_box\_office（每年票房总和）、genre\_distribution（电影类型分布情况）、top\_50\_movies（票房排名前五十的电影）、release\_date\_box\_office（电影上映日期与票房的关系）、box\_office\_classification（电影票房分类）、release\_date\_classification（电影发布日期分类）、release\_date\_genre\_classification（电影发布日期和电影类型分类）。

同时，使用 xlsxwriter 引擎在 Excel 文件中创建名为 “图表” 的工作表，绘制多种数据可视化图表，包括每年票房总和折线图，电影类型分布情况饼图，电影票房分类饼图，电影发布日期分类柱状图，电影发布日期和电影类型分类柱状图。

# 4.数据分析

## 4.1 数据分析背景与目标

本环节旨在通过ai分析的形式，通过已经筛选出的近年来电影票房数据来洞察电影市场的趋势和规律。具体而言，我们将利用Python与OpenAI的API链接，结合数据分析手段，对电影票房、类型、上映日期等多方面的数据进行综合研究，以便深入了解电影市场的走向，同时允许用户对ai进行有关内容的自由发问，辅助用户更加明确的了解信息，并为未来的市场策略提供数据和分析支持。数据来源于指定电影网站的Top100电影，已经通过筛选、去重、分类和可视化处理，以清晰呈现近年来电影市场的主要特征和趋势。

## 4.2 数据分析内容与流程

### 4.2.1 Python与AI的连接设置

通过openai库将Python与OpenAI的API连接，实现对电影数据的智能分析。设置API密钥、基础URL和请求头信息，为后续数据交互奠定基础。

### 4.2.2 数据分析流程

#### 4.2.2.1分析票房数据

使用Excel文件作为输入，提取并整理票房数据，调用OpenAI的GPT模型，要求其在输入的表格数据基础上分析电影票房的变化趋势，提供专业见解。

#### 4.2.2.2图表数据解读

对于Excel文件中的图表，AI模型根据图表内容解读票房趋势和各类电影在市场中的表现，深入分析数据的关联性。

#### 4.2.2.3 多表格综合总结

将多个表格的数据综合输入AI，以总结电影市场的整体趋势。AI根据历史数据和当前表格的内容总结行业变化情况，为预测未来市场趋势提供参考。

### 4.2.3 自动回答与用户交互

程序首先会逐个系统分析一下提供的excel的数据。直到系统分析和分析总结完后，用户可以输入与2020至2022年电影票房相关的问题，系统会调用AI模型提供基于数据的回答。用户的输入和数据内容会作为背景信息传递给AI，使其回答更符合实际情况。

## 4.3 代码说明

**数据分析函数**：analyze\_data(data\_str)调用GPT模型分析输入的电影数据。

**图表分析函数**：analyze\_img(img)负责对图表数据进行解读。

**数据总结函数**：sum\_data(df)使用AI总结不同表格中的电影趋势数据，帮助更好地了解市场走向。

**回答问题函数**：answer\_data(df)为用户提供自定义问题的回答。

**Excel数据读取与图像提取**：excel\_to\_json(path)函数将Excel数据转换为JSON格式；extract\_images(xls)函数提取Excel文件中的图像信息，供AI进行图表解读。

# 5. 结束语

## 5.1. 总结

本课题通过构建基于爬虫技术和大语言模型的电影票房数据分析系统，实现了对近几年电影市场的深入分析。系统自动获取和处理电影网站的数据，涵盖电影名称、票房、类型等多维度信息；并在此基础上利用GPT模型进行智能分析，生成市场趋势总结和票房预测结果。分析结果为研究者提供了对电影市场的全面理解，揭示了不同类型电影的票房表现和观众偏好，为发行方和决策者提供了参考依据。本系统展示了数据挖掘和人工智能在电影行业中的应用潜力，为未来电影市场的智能化研究提供了新的视角。

## 5.2. 不足与展望

尽管系统实现了对电影数据的有效分析，但仍存在一些不足之处。例如，由于爬虫技术依赖目标网站的稳定性，因此在获取数据时可能会受限于目标网站的访问权限和反爬机制。此外，大语言模型的分析结果受限于数据的准确性和全面性，对于数据的长远预测仍有待提升。未来，系统可以结合更多的数据来源，融入深度学习模型进行数据分析，以提升预测的精确性和稳健性。

# 6. 参考文献

[1]盛景,徐超,周涛.Python网络爬虫技术分析[J].中国信息界,2024,(06):210-212.

[2] 维基百科. The Numbers [EB/OL].[2024-07-23] https://zh.wikipedia.org/zh-cn/The\_Numbers.

[3]杨健,陈伟.基于Python的三种网络爬虫技术研究[J].软件工程,2023,26(02):24-27+19.DOI:10.19644/j.cnki.issn2096-1472.2023.002.005.

[4]McKinney, W. Python for Data Analysis[M]. O'Reilly Media, 2017.

[5]VanderPlas, J. Python Data Science Handbook[M]. O'Reilly Media, 2016.