目 录

[第1章 概 述 1](#_Toc140570321)

[1.1 系统开发目的和意义 1](#_Toc140570322)

[1.2 开发环境及技术 1](#_Toc140570323)

[2.1问题描述 3](#_Toc140570324)

[2.2 问题分析 4](#_Toc140570325)

[第3章 数据初探与数据预处理 7](#_Toc140570326)

[3.1 数据初步探索 7](#_Toc140570327)

[3.2 数据预处理 8](#_Toc140570328)

[第4章 数据分析 11](#_Toc140570329)

[4.1 分析思路 11](#_Toc140570330)

[4.2数据分析 12](#_Toc140570331)

[4.3书籍推荐 24](#_Toc140570332)

[结 论 28](#_Toc140570333)

[参考文献 30](#_Toc140570334)

# 第1章 概 述

## 1.1 系统开发目的和意义

本次课程设计的题目是“经济管理学院图书大数据分析”，该设计是对《程序设计综合实践课程》中所学知识的巩固和深入应用。本次课程设计通过使用python语言分析已有的图书馆相关数据，通过对经济管理学院图书大数据项目的分析，利用Python编程语言处理和分析数据，以获得对该学院图书资源的深入理解和洞察[1]。

大数据分析分析可以为学院图书馆管理者、教职员工和学生提供有价值的信息，帮助他们做出更好的图书采购、管理和使用决策。当前我国图书馆基本实现了数字化图书馆，建立了电子图书资源库、文献检索系统、图书借阅系统等，为读者提供方便快捷的图书资料服务[2]。随着数字化图书馆的运行，图书馆积累了大量的读者借阅书籍、检索文献的数据[3]。这些数据经过长时间积累变成了数据库中冗余并且不能删除的垃圾数据。大数据技术的发展，这些原本的垃圾数据变成了优秀的信息资源，这些积累的信息可以通过数据分析技术得到客户需要的多种信息。在图书资源优化、学科需求分析、图书使用行为研究、资源投入评估、用户满意度提升等方面有重大意义[4]。

通过对经济管理学院图书大数据项目的分析，可以为学院图书馆管理者和学院师生提供宝贵的数据支持，帮助他们做出更明智的图书资源管理和使用决策，提升学院的教学质量和学术声誉。通过“经济管理学院图书大数据分析”，可以熟悉数据分析流程，了解数据分析的基本思想，将原本的垃圾数据变成有价值的数据，通过分析，了解当前经济管理学院图书大数据分析师生最急迫掌握的知识体系，对图书馆购买经济管理学院图书大数据分析所需要的图书提供决策参考[5]。

## 1.2 开发环境及技术

经济管理学院图书大数据分析项目使用了Python3开发，采用Anaconda作为集成开发环境。分析过程中使用了以下扩展包：

1、Numpy扩展库

NumPy（Numerical Python）是Python中最重要的科学计算库之一，它为Python提供了高效的多维数组对象和广泛的数学函数库。然而，NumPy的功能并不仅限于这些，还有一些非常有用的扩展库可以进一步扩展NumPy的功能。SciPy、Matplotlib、Pandas、Scikit-learn、OpenCV这些扩展库与NumPy紧密结合，为科学计算、数据分析和机器学习等领域提供了强大的工具和功能。它们进一步拓展了NumPy的应用范围，使得Python成为了进行科学计算和数据分析的首选语言之一。

2、Pandas扩展库

Pandas是一个基于NumPy的数据处理和分析库，它提供了高性能、易用的数据结构和数据操作工具。Pandas的核心数据结构是DataFrame，它是一个二维的表格型数据结构，类似于Excel中的表格或SQL中的数据库表。除了DataFrame，Pandas还提供了Series这一一维的标签型数据结构，用于存储一列数据。

3、Matplotlib扩展库

Matplotlib是一个用于绘制二维图表和可视化数据的库，它提供了广泛的绘图选项和定制功能。以下是一些常见的Matplotlib扩展库的简要介绍：

Seaborn：Seaborn是一个基于Matplotlib的统计数据可视化库，提供了更高级别的图表和统计图形，使得数据可视化变得更加简单和美观。Seaborn提供了一系列内置的主题和调色板，可以快速创建各种统计图形，如条形图、箱线图、散点图、热力图等。它还提供了对分类数据的专门支持，如分类散点图、分类箱线图等。

Plotly：Plotly是一个交互式的可视化库，可以生成高度可交互的图表和可视化效果。它支持绘制多种图表类型，如折线图、散点图、柱状图、地理地图等，并提供了丰富的交互功能，如缩放、平移、悬停等。Plotly还支持在Web上展示和共享可视化结果，使得团队合作和数据共享更加便捷。

这些扩展库为Matplotlib提供了更多的图表类型、样式和交互性，使得数据可视化更加灵活和丰富。它们扩展了Matplotlib的能力，使得用户可以根据需求创建出高质量、可交互的图表和可视化效果。

4、WordCloud扩展库

WordCloud是一个用于生成词云图的Python库，它可以根据文本数据中的词频信息创建具有视觉吸引力的词云图像。词云图是一种以词语为基本元素的可视化方式，通过将词语按照其重要性或频率展示为不同大小、颜色和排列方式的图形，形成一幅具有艺术性和信息呈现的图像。WordCloud库的主要特点和使用方法有以下：1. 词频统计：WordCloud可以根据给定的文本数据统计词语的频率信息。它会根据词语出现的频率大小，在词云图中以不同的字体大小来展示词语的重要程度。2. 视觉定制：WordCloud提供了丰富的视觉定制选项，可以通过设置字体、颜色、背景图像、形状等参数来创建自定义的词云图。用户可以根据需求调整词云图的外观，使其更符合预期的视觉效果。3. 词语形状：除了基本的词云形状（如矩形），WordCloud还支持使用自定义的形状作为词云的基础形状。用户可以选择使用心形、星形、动物形象等各种形状来呈现词云图，增加视觉的趣味性和个性化。

第2章 问题描述与问题分析

## 2.1问题描述

在大数据环境下，图书馆及其服务也必将产生新的巨大变化，深层次的服务功能可以通过大数据技术的运用加以实现。主要体现在两个方面：

1．个性化服务：大数据技术可以帮助图书馆实现个性化的服务。通过收集和分析用户的阅读记录、搜索行为、兴趣偏好等大数据信息，图书馆可以为每位用户提供定制化的图书推荐、阅读建议和学术支持。通过深入了解用户的需求和兴趣，图书馆可以提供更加精准和个性化的图书馆服务，提升用户体验和满意度[6]。

2．智能决策支持：大数据技术可以帮助图书馆进行智能决策和资源优化。通过对大数据进行分析和挖掘，图书馆可以了解图书馆资源的使用情况、流通规律和需求趋势，从而优化图书采购、馆藏管理和图书分发等决策。例如，根据读者的借阅行为和评论数据，图书馆可以预测图书需求，合理规划馆藏，减少库存积压和资源浪费。此外，大数据还可以帮助图书馆进行预测分析和需求预测，提前做好资源准备，优化服务流程[7]。

经济管理学院图书大数据分析项目就是结合图书馆大数据针对经济管理学院师生的借阅信息完成以下任务：

1．对提供的图书大数据做初步探索性分析：通过对数据表的基础统计，分析表的结构，关联关系，错误查询。查看数据表的结构，了解每个表包含的字段（列）以及字段的数据类型。这可以通过查看数据表的列名和数据类型来完成。例如，可以使用Pandas库的info()函数或SQL查询语句中的DESCRIBE命令来获取有关数据表结构的信息。如果存在多个数据表，我们可以分析表之间的关联关系。通过查看各个表之间的共享字段或键，我们可以确定哪些字段可以用于连接多个表。这有助于理解数据之间的关联性，例如图书与作者之间的关系、图书与借阅记录之间的关系等。

2．对给定数据做数据预处理：1. 处理缺失值：检查数据中是否存在缺失值，即空白或未记录的数据。对于缺失值，可以考虑删除缺失值所在的行或列，或者使用合适的填充方法进行填补。对于借阅信息，可能存在缺失的学生或图书信息，需要相应处理。2. 处理异常值：检测和处理异常值，这些值可能是数据记录中的离群值或错误值。对于借阅信息，可以检查借阅日期、归还日期等是否符合合理的范围，排除异常值或者进行修正。3. 处理异常数据：对于异常数据或不符合业务规则的数据，根据实际情况进行处理。例如，借阅时长超过规定范围的记录可能需要进行修正或排除。

3．保留需要的数据：读者信息：保留读者的读者ID、读者类型、性别、单位。图书目录：保留书籍图书ID、书名、图书分类号。图书借还信息：操作时间、操作类型、图书ID、读者ID。

4．经济管理学院教师借阅情况：2014年，2015年，2016年，2017年所借书籍类别占前10的都是什么类别。最喜欢看的小说分别是那一本。一共借了多少书，专业书籍多少本。一共有多少本书没有归还。没有归还的书籍哪类书籍最多。

5．经济管理学院学生借阅情况：2014年，2015年，2016年，2017年所借书籍类别占前10的都是什么类别。最喜欢看的小说分别是那一本。一共借了多少书，专业书籍多少本。一共有多少本书没有归还。没有归还的书籍哪类书籍最多。

6．数据可视化：用折线图画出你分配的那个学院的教师2014年，2015年，2016年，2017年所借书籍类别占前10的书籍的走向图。学生所借书籍类别占前10的书籍的走向图。教师所喜欢的排名前10的小说的走向图。学生所喜欢的排名前10的小说的走向图。

7．结果分词：将所有借阅的图书的书名通过jieba分词，分词后使用停用词库排除停用词

8．词云制作：加载停用词库文件，并将其中的停用词添加到stopwords集合中。

接下来，对书名进行分词，并使用停用词库排除停用词，将分词结果存储在words\_list列表中。统计每个词的词频，并将结果存储在word\_counts字典中。使用WordCloud类创建词云对象，并调用generate\_from\_frequencies()方法传入词频数据。使用plt.imshow()函数绘制词云图。

9．图书推荐：为你所在院系的师生购书（应该增加什么样的书籍）提供建议。根据数据分析结果。

## 2.2 问题分析

经济管理学院图书大数据分析项目针对经济管理学院2014年，2015年，2016年的图书借阅数据进行分析。

数据初次探索与分析：当初次拿到数据时，由于这是一份完全陌生的数据，不能直接对数据做分析处理，需要针对拿到的数据做初步探索。要准确地分析数据就需要先了解数据，了解数据的过程称为探索性数据分析。数据探索的目的：

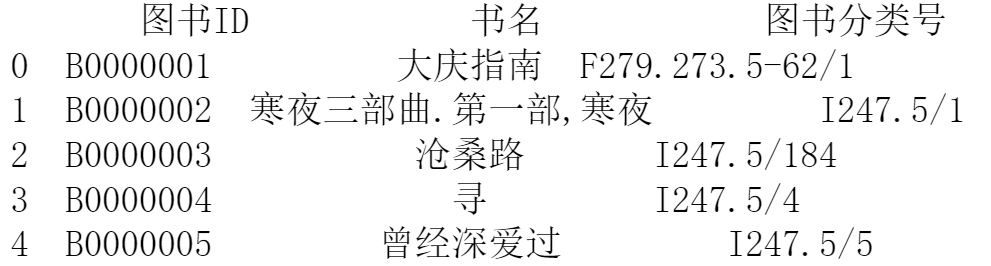
1）了解数据里有什么，以及数据里没有什么。这将决定数据是否满足分析的要求。通过数据的结构了解数据的含义。

2）了解数据的质量。是否缺少数据？是否有错误的数据？以此决定在分析前如何对数据进行数据清洗，需要补充哪些数据、过滤哪些数据。

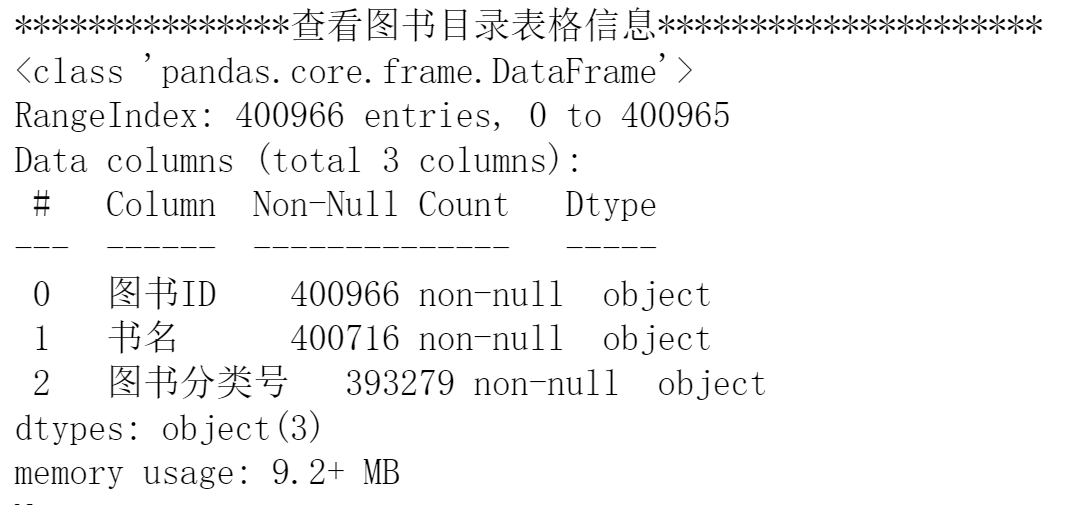
3）数据类型是否符合分析模型，并进行数据转换。转换包括数据类型转换，数据编码等。

4）了解数据的基本统计情况，以确定一些分析模型的参数。

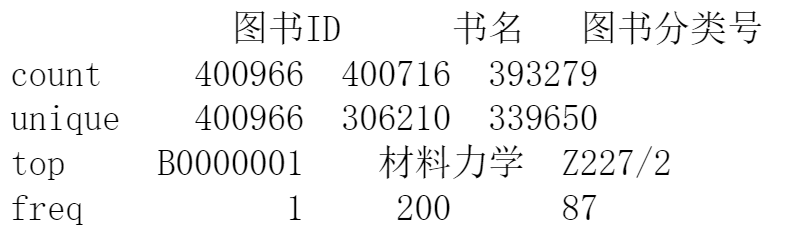
对给定的数据做预处理：可以看到数据中包含图书目录信息、读者信息、图书借还2014年到2017年。图书目录数据的前5行信息为：



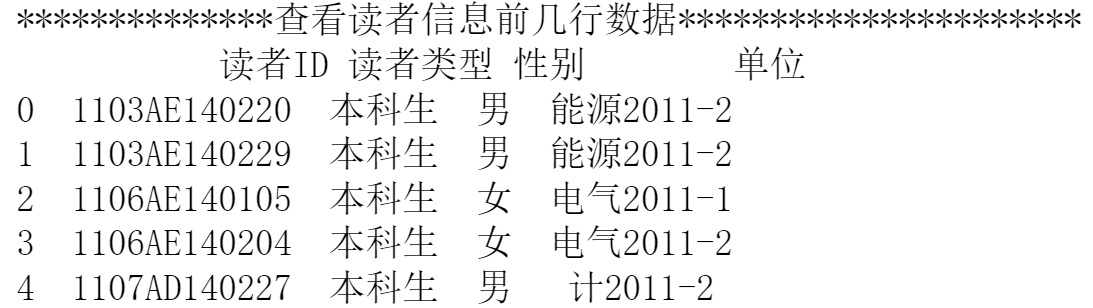
表格信息为：



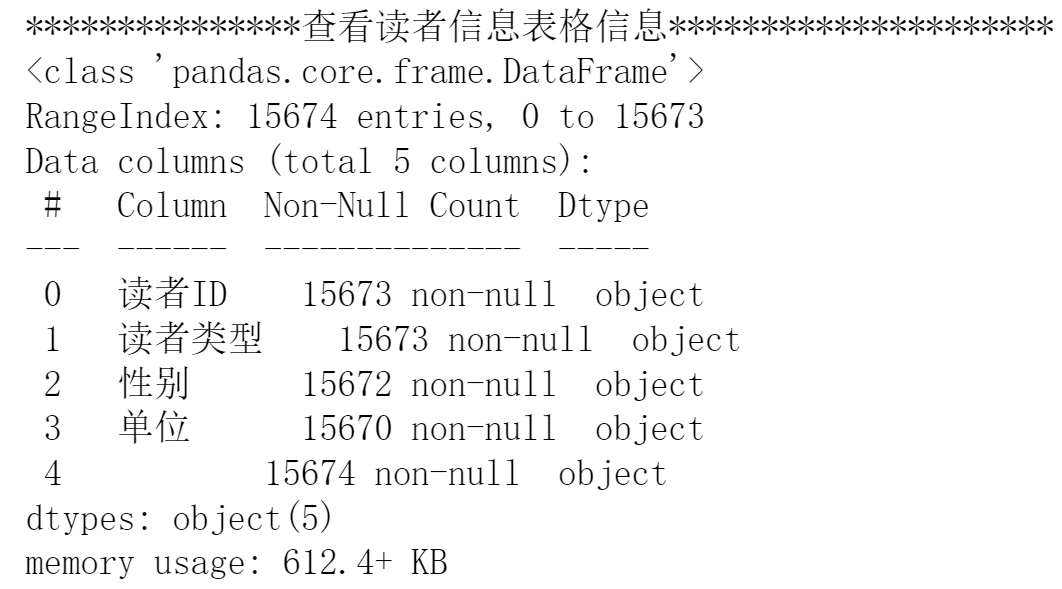
统计摘要为：



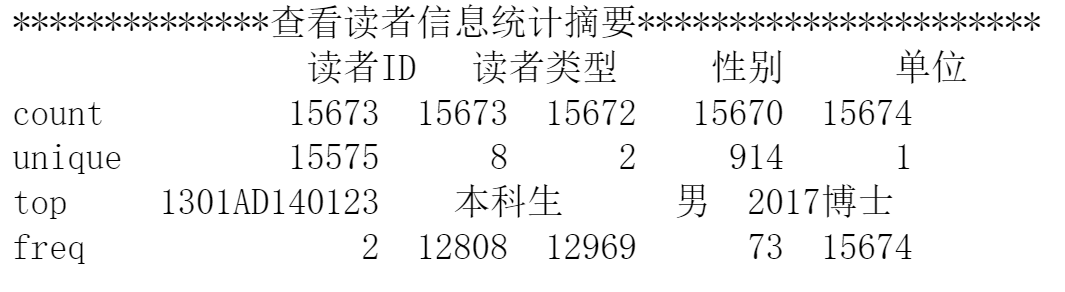
读者信息.xlsx数据的前5行信息为：



表格信息为：



统计摘要为：



保留所需的数据：读者信息：保留读者的读者ID、读者类型、性别、单位。图书目录：保留书籍图书ID、书名、图书分类号。图书借还信息：操作时间、操作类型、图书ID、读者ID。

数据存在一定数据缺失，重复数据没有发现，对于缺失数据，采用赋值处理，对于NaN数据，赋值“”空字符处理。

经济管理学院师生图书推荐：针对师生图书借阅量走势预测哪一类别的书数量增多，从而推荐该类别的书籍。从普遍书到小说，分别分析推荐。

# 第3章 数据初探与数据预处理

## 3.1 数据初步探索

经济管理学院图书大数据分析项目针对图书馆提供的基础数据进行了数据初步探索性分析，经过分析图书馆提供的基础数据包含：

1、《中国图书馆图书分类法》简表。

2、读者信息表

3、常用停用词表（stopwords.txt）

4、图书目录表

5、图书借还2014-2017表

6、敏感词词表

经过数据表的探索性分析的出结果：

1、《中国图书馆图书分类法》简表的作用是为所有的馆藏书籍分类，其内容如表3-1所示：

表3-1 中国图书馆图书分类法数据构成

|  |  |  |  |
| --- | --- | --- | --- |
| 列名 | 说明 | 数据类型 | 是否项目需要 |
| 图书分类号 | 图书分类号 | 字符串 | 需要 |
| 书籍类别名称 | 书籍类别名称 | 中文字符 | 需要 |

2、读者信息表，其内容如表3-2所示：

表3-2 读者信息表数据构成

|  |  |  |  |
| --- | --- | --- | --- |
| 列名 | 说明 | 数据类型 | 是否项目需要 |
| 读者ID | 读者ID | 字符串 | 需要 |
| 读者类型 | 读者类型 | 中文字符 | 需要 |
| 性别 | 性别 | 中文字符 | 不需要 |
| 单位 | 单位 | 中文字符 | 不需要 |

3、图书目录表内容如表3-3所示：

表3-3图书目录表数据构成

|  |  |  |  |
| --- | --- | --- | --- |
| 列名 | 说明 | 数据类型 | 是否项目需要 |
| 图书ID | 图书ID | 字符串 | 需要 |
| 书名 | 书名 | 中文字符 | 需要 |
| 图书分类号 | 图书分类号 | 字符串 | 需要 |

3、图书借还表内容如表3-3所示：

表3-3图书借还表数据构成

|  |  |  |  |
| --- | --- | --- | --- |
| 列名 | 说明 | 数据类型 | 是否项目需要 |
| 操作时间 | 操作时间 | Datatime | 不需要 |
| 操作类型 | 操作类型 | 中文字符 | 需要 |
| 图书ID | 图书ID | 字符串 | 需要 |
| 读者ID | 读者ID | 字符串 | 需要 |

下面针对分析后项目需要的数据进行分析

1）书籍分类号：

该数据项是将图书按其学科内容分成22个基本大类，每一大类下分许多小类，每一小类下再分子小类。最后，每一种书都可以分到某一个类目下，每一个类目都有一个类号。例如I表示文学，I24表示小说，I247表示当代作品，TP表示自动化技术、计算机技术，TP3表示计算技术、计算机技术。

经过分析该数据项内容正确可以满足项目需要。在借阅中存在大量在图书分类号数据中不存在的分类。如图书分类号:O13-44/210、图书分类号:A122/2、图书分类号:I267/468、图书分类号:TE311/17、图书分类号:G254.9/9-2、图书分类号:O241/77、图书分类号:O241/79、图书分类号:O17-44/18、图书分类号:TE311/17。

本项目分析目标为经济管理学院的图书数据，经过分析该数据项符合项目需求不需要进行数据过滤。

2）书籍类别名称：

书籍类别名称没有直接提供，可以通过《中国图书馆图书分类法》简表对其图书ID进行类别解析

2、读者信息表的作用是存储所有的读者信息

该表包含15673条信息。最多的人群为本科生、男生。包含读者ID 、读者类型、性别、单位。其中性别单位不需要，读者类型有的数据项为空，有1项，需要去掉。

经过对原始数据的初步探索，发现原始数据并不足以满足分析项目的要求，原始数据缺少针对书籍类别的描述。需要在数据预处理阶段增加书籍类别表格来描述书籍类别。

此外，由于查询分析需要联合三个表，因为建立起读者信息、图书信息、借阅信息三表的联合表。

## 3.2 数据预处理

根据数据分析结果针对原始数据做数据预处理工作，工作包含取出重复值，去Null值，修改读者信息表，图书目录中的错误数据，增加完整信息表来完成联合分析工作，教师学生借阅分析工作，筛选读者信息表，借阅信息表中经济管理学院数据。

### 3.2.1 去重复值去NULL值

根据数据初步探索性分析结果：《中国图书馆图书分类法》简表，图书借还2014表中存在着重复值，《中国图书馆图书分类法》简表，图书借还2014表中存在Null值。

去重的实现代码为：

Jieyue2014['图书分类号'] = Jieyue2014 ['图书分类号'].fillna('')

Jieyue2014 = Jieyue2014 ['图书分类号'].drop\_duplicates()

### 3.2.2 修改错误值

根据数据初步探索性分析结果：《中国图书馆图书分类法》简表，图书借还2014表中存在图书分类号字段值错误，错误字段包括

"O13-44/210",

"A122/2",

"I267/468",

"TE311/17",

"G254.9/9-2",

"O241/77",

"O241/79",

"O17-44/18",

"TE311/17"

修改错误值的实现代码为：

# 例如：将错误的图书分类号 " O13-44/210" 修改为 "O13-44"

borrow\_return\_df.loc[borrow\_return\_df['图书分类号'] == ' O13-44/210', '图书分类号'] = ' O13-44'

### 3.2.3 增加数据表格

根据数据初步探索性分析结果：现有数据不能完成分析项目，需要增加图书分类号映射表数据。经过分析需要增加以下表格：

1、图书分类号映射表，表格包含字段图书分类号、图书分类名称，其意义是图书在中国图书馆图书分类法中的编号、按照中国图书馆图书分类法编号的名称，表格样式如表3-2所示：

表3-2 \*\*\*表数据构成

|  |  |  |  |
| --- | --- | --- | --- |
| 列名 | 说明 | 数据类型 | 是否项目需要 |
| 图书分类号 | 图书在中国图书馆图书分类法中的编号 | 字符串 | 需要 |
| 图书分类名称 | 按照中国图书馆图书分类法编号的名称 | 字符串 | 需要 |

### 3.2.4 数据过滤

根据数据初步探索性分析结果：图书借还2014、图书借还2015、图书借还2016，图书借还2017、表中数据可以按照项目需求新型过滤，留下计算机专业的借阅信息。实现过滤计算机专业的借阅信息，我们需要对图书借还表和读者信息表进行联结（Join），根据读者ID和计算机专业的条件进行过滤。

过滤的实现代码为：

# 将读者信息表中计算机专业的读者筛选出来

computer\_major\_readers = reader\_info[reader\_info['专业'].str.startswith('计')]

# 使用merge函数将图书借还表和计算机专业读者信息表进行联结，过滤出计算机专业的借阅信息

filtered\_2014 = borrow\_return\_2014.merge(computer\_major\_readers, on='读者ID')

filtered\_2015 = borrow\_return\_2015.merge(computer\_major\_readers, on='读者ID')

filtered\_2016 = borrow\_return\_2016.merge(computer\_major\_readers, on='读者ID')

filtered\_2017 = borrow\_return\_2017.merge(computer\_major\_readers, on='读者ID')

# 第4章 数据分析

## 4.1 分析思路

通过对原始数据初探与数据预处理，得到了可以用于数据分析的数据，现针对经济管理学院图书大数据分析项目的各个任务目标制定分析工作思路。

1、经济管理学院的教师2014年，2015年，2016年，2017年所借书籍类别占前10的都是什么类别的分析思路：统计每一年经济管理学院教师借阅的图书，将借阅最多的图书的类别分析统计出来，寻找累计最多的10种类别，如果不足10种，输出所有类别。

2、统计经济管理学院的教师2014年，2015年，2016年，2017年所最喜欢看的小说分别是那一本。将教师2014年，2015年，2016年，2017年cat到一起，联合读者信息与图书目录表，寻找累计最多的10种类别。查找小说需要查看图书分类号里是否包含小说的的图书分类代号。

3、统计经济管理学院的教师2014年，2015年，2016年，2017年一共有多少本书没有归还。没有归还的书籍哪类书籍最多。先筛选出所有老师借阅的记录，再筛选出操作类型为借的，筛选出操作类型为还的。将借的数量减去还的数量就是还没有还的书籍的数量。

4、统计经济管理学院的教师2014年，2015年，2016年，2017年一共借了多少书，专业书籍多少本。统计总共借阅数量只需要筛选学院与教师两个信息，统计借的数据一共有多少条。统计专业书籍的数量需要统计其中图书分类号以'A', 'TB', 'S', 'R', 'C', 'F', 'G', 'K', 'I', 'N'这些字符开头的图书分类号。

5. 对于以上信息，在学生群体里处理方法相似，需要将教师更换为学生。需要注意的是，学生群体包括博士研究生、研究生、本科生。

6. 用折线图画出经济管理学院的教师或学生的2014年，2015年，2016年，2017年所借书籍类别占前10的书籍的走向图。先将四年中前10的书籍类别找出来，再统计四年前十书籍的数量变化。以年份为横轴，数量为纵轴，书籍类别为标签，画出折线图。

7. 用折线图画出经济管理学院的教师2014年，2015年，2016年，2017年所喜欢的排名前10的小说的走向图。先将四年中前10的小说书籍类别找出来，再统计四年前十书籍的数量变化。以年份为横轴，数量为纵轴，书籍类别为标签，画出折线图。

8. 分析经济管理学院系师生所借阅的所有图书什么知识是最迫切需要学习的。将所有借阅的图书的书名通过jieba分词，分词后使用停用词库排除停用词，然后制作词云，并将词云中出现最高的10个词拿出作为该专业师生最迫切需要学习的知识。

## 4.2数据分析

### 4.2.1教师借阅专业书籍分析

根据任务统计经济管理学院的教师2014年，2015年，2016年，2017年所借书籍类别占前10的都是什么类别的分析思路完成以下分析工作。

1、统计每一年经济管理学院教师借阅的图书数量，经过统计经济管理学院教师2014年度，借阅的图书最多的是I209/7 类别： '文学史、文学思想史'，占比前3的图书如表4-1所示。

表4-1 2014年经济管理学院教师借阅图书统计表

|  |  |  |  |
| --- | --- | --- | --- |
| 图书分类号 | 图书名 | 类别 | 数量 |
| I247.5/3244 | 武林外史 | 当代作品 | 12 |
| I247.57/2175 | 皇妃出阁 | 当代作品 | 8 |
| J632.326/3 | 古筝必学教程 | 中国民族器乐 | 8 |

本过程实现代码是：

for year in jiehuan:

borrow\_df = jiehuan[year]

merge\_df = pd.merge(borrow\_df, content, on='图书ID')

merge\_df = pd.merge(merge\_df, reader, on='读者ID')

teacher\_df = merge\_df[merge\_df['读者类型'] == '教师']

category\_counts = teacher\_df['图书分类号'].value\_counts().head(10)

print("教师{}年借阅的书籍类别占前10的类别为：".format(year))

for cls, num in category\_counts.items():

print(parse\_classification\_number(cls), cls, num, content[content['图书分类号']==cls]['书名'],sep='\t')

# print(category\_counts)

2015年度，借阅的图书最多的是潜意识与人格，占比前3位的图书如表4-2所示。

表4-2 2015年经济管理学院教师借阅图书统计表

|  |  |  |  |
| --- | --- | --- | --- |
| 图书分类号 | 图书名 | 类别 | 数量 |
| B848/66 | 潜意识与人格 | 个性心理、人格心理学 | 32 |
| J648.32/1 | 中国传统古筝曲大全 | 民族器乐曲 | 10 |
| I247.5/2035 | 侠客行 | 当代作品（1949-- | 10 |

本过程实现代码是：

for year in jiehuan:

borrow\_df = jiehuan[year]

merge\_df = pd.merge(borrow\_df, content, on='图书ID')

merge\_df = pd.merge(merge\_df, reader, on='读者ID')

teacher\_df = merge\_df[merge\_df['读者类型'] == '教师']

category\_counts = teacher\_df['图书分类号'].value\_counts().head(10)

print("教师{}年借阅的书籍类别占前10的类别为：".format(year))

for cls, num in category\_counts.items():

print(parse\_classification\_number(cls), cls, num, content[content['图书分类号']==cls]['书名'],sep='\t')

# print(category\_counts)

2016年度，借阅的图书最多的是中国文学史，占比前3位的图书如表4-3所示。

表4-3 2016年经济管理学院教师借阅图书统计表

|  |  |  |  |
| --- | --- | --- | --- |
| 图书分类号 | 图书名 | 类别 | 数量 |
| I209/7 | 中国文学史 | 文学史、文学思想史 | 15 |
| I247.48/100 | 射雕英雄传 | 当代作品（1949-- | 12 |
| I206.09/1 | 当代文艺思潮小史 | 文学评论和研究 | 10 |

本过程实现代码是：

for year in jiehuan:

borrow\_df = jiehuan[year]

merge\_df = pd.merge(borrow\_df, content, on='图书ID')

merge\_df = pd.merge(merge\_df, reader, on='读者ID')

teacher\_df = merge\_df[merge\_df['读者类型'] == '教师']

category\_counts = teacher\_df['图书分类号'].value\_counts().head(10)

print("教师{}年借阅的书籍类别占前10的类别为：".format(year))

for cls, num in category\_counts.items():

print(parse\_classification\_number(cls), cls, num, content[content['图书分类号']==cls]['书名'],sep='\t')

# print(category\_counts)

2017年度，借阅的图书最多的是刘心武评点《金瓶梅》，占比前3位的图书如表4-4所示。

表4-4 2016年经济管理学院教师借阅图书统计表

|  |  |  |  |
| --- | --- | --- | --- |
| 图书分类号 | 图书名 | 类别 | 数量 |
| I207.419/66 | 刘心武评点《金瓶梅》 | 各体文学评论和研究 | 11 |
| C913.2-64/1 | 善谋者成:图解最有效的职业规划 | 社会生活和社会问题 | 10 |
| J292.113.5/1 | 经典行书教案 | 中国书法、篆刻 | 8 |

本过程实现代码是：

for year in jiehuan:

borrow\_df = jiehuan[year]

merge\_df = pd.merge(borrow\_df, content, on='图书ID')

merge\_df = pd.merge(merge\_df, reader, on='读者ID')

teacher\_df = merge\_df[merge\_df['读者类型'] == '教师']

category\_counts = teacher\_df['图书分类号'].value\_counts().head(10)

print("教师{}年借阅的书籍类别占前10的类别为：".format(year))

for cls, num in category\_counts.items():

print(parse\_classification\_number(cls), cls, num, content[content['图书分类号']==cls]['书名'],sep='\t')

# print(category\_counts)

2、统计每一年经济管理学院教师借阅最多的小说

教师2014年最喜欢看的小说为：

测量学 16

Name: 书名, dtype: int64

教师2015年最喜欢看的小说为：

潜意识与人格 32

Name: 书名, dtype: int64

教师2016年最喜欢看的小说为：

西游记 20

Name: 书名, dtype: int64

教师2017年最喜欢看的小说为：

西游记 11

Name: 书名, dtype: int64

代码：

for year in jiehuan:

borrow\_df = jiehuan[year]

merge\_df = pd.merge(borrow\_df, content, on='图书ID')

merge\_df = pd.merge(merge\_df, reader, on='读者ID')

teacher\_df = merge\_df[(merge\_df['读者类型'] == '教师')]# & ('I207.4' in merge\_df['图书分类号'])]

category\_counts = teacher\_df['书名'].value\_counts().head(1)

print("教师{}年最喜欢看的小说为：".format(year))

print(category\_counts)

3、教师一共借阅：25097本书

book\_sum = 0

all\_teacher\_df = pd.DataFrame()

for year in jiehuan:

borrow\_df = jiehuan[year]

merge\_df = pd.merge(borrow\_df, content, on='图书ID')

merge\_df = pd.merge(merge\_df, reader, on='读者ID')

teacher\_df = merge\_df[(merge\_df['读者类型'] == '教师')]

all\_teacher\_df = pd.concat([all\_teacher\_df, teacher\_df], ignore\_index=True)

print("教师一共借阅：{}本书".format(len(all\_teacher\_df['操作类型']=='借')))

4、 教师借阅专业书5974本

zhuanye\_df = zhuanye\_df[zhuanye\_df['图书分类号'].str.startswith(('A', 'TB', 'S', 'R', 'C', 'F', 'G', 'K', 'I', 'N'))]

print("教师借阅专业书{}本".format(len(zhuanye\_df)))

5、没有归还的书籍数量： 420。没有归还的书籍中数量最多的图书分类号： I209/7 类别： ('文学史、文学思想史', None)

borrowed\_books = all\_teacher\_df[all\_teacher\_df['操作类型'] == '借']

returned\_books = all\_teacher\_df[all\_teacher\_df['操作类型'] == '还']

books\_not\_returned = borrowed\_books['图书ID'].value\_counts() - returned\_books['图书ID'].value\_counts()

num\_books\_not\_returned = len(books\_not\_returned[books\_not\_returned > 0])

print("没有归还的书籍数量：", num\_books\_not\_returned)

not\_returned\_categories = borrowed\_books[borrowed\_books['图书ID'].isin(books\_not\_returned[books\_not\_returned > 0].index)]['图书分类号']

category\_counts = not\_returned\_categories.value\_counts()

most\_common\_category = category\_counts.idxmax()

6、学生2014年最喜欢看的小说为：

平凡的世界 65

Name: 书名, dtype: int64

学生2015年最喜欢看的小说为：

平凡的世界 191

Name: 书名, dtype: int64

学生2016年最喜欢看的小说为：

活着 191

Name: 书名, dtype: int64

学生2017年最喜欢看的小说为：

白鹿原 235

Name: 书名, dtype: int64

代码：for year in jiehuan:

borrow\_df = jiehuan[year]

merge\_df = pd.merge(borrow\_df, content, on='图书ID')

merge\_df = pd.merge(merge\_df, reader, on='读者ID')

student\_condition = ((merge\_df['读者类型'] == '博士研究生') | (merge\_df['读者类型'] == '本科生') | (merge\_df['读者类型'] == '研究生')

student\_df = merge\_df[student\_condition & merge\_df['图书分类号'].str.contains('|'.join(xiaoshuo), case=False)]# & ('I207.4' in merge\_df['图书分类号'])]

category\_counts = student\_df['书名'].value\_counts().head(1)

print("学生{}年最喜欢看的小说为：".format(year))

print(category\_counts)

7、学生一共借阅：617147本书

book\_sum = 0

all\_student\_df = pd.DataFrame()

for year in jiehuan:

borrow\_df = jiehuan[year]

merge\_df = pd.merge(borrow\_df, content, on='图书ID')

merge\_df = pd.merge(merge\_df, reader, on='读者ID')

student\_condition = ((merge\_df['读者类型'] == '博士研究生') | (merge\_df['读者类型'] == '本科生') | (merge\_df['读者类型'] == '研究生'))

student\_df = merge\_df[student\_condition]

all\_student\_df = pd.concat([all\_student\_df, student\_df], ignore\_index=True)

print("学生一共借阅：{}本书".format(len(all\_student\_df['操作类型']=='借')))

8、学生借阅专业书142756本

student\_jieyue = all\_student\_df[all\_student\_df['操作类型']=='借']

student\_jieyue['图书分类号'] = student\_jieyue['图书分类号'].fillna('')

student\_zhuanye\_df = student\_jieyue[student\_jieyue['图书分类号'].str.startswith(('A', 'TB', 'S', 'R', 'C', 'F', 'G', 'K', 'I', 'N'))]

print("学生借阅专业书{}本".format(len(student\_zhuanye\_df)))

9. 学生没有归还的书籍数量： 9735。学生没有归还的书籍中数量最多的图书分类号： I313.45/514 类别： ('日本文学', None)。

student\_borrowed\_books = all\_student\_df[all\_student\_df['操作类型'] == '借']

student\_returned\_books = all\_student\_df[all\_student\_df['操作类型'] == '还']

student\_books\_not\_returned = student\_borrowed\_books['图书ID'].value\_counts() - student\_returned\_books['图书ID'].value\_counts()

student\_num\_books\_not\_returned = len(student\_books\_not\_returned[student\_books\_not\_returned > 0])

print("学生没有归还的书籍数量：", student\_num\_books\_not\_returned)

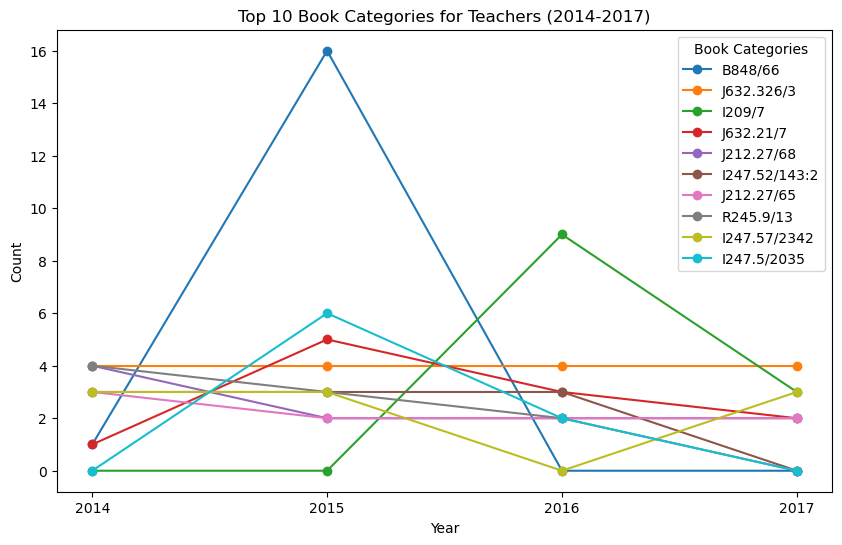
studetn\_not\_returned\_categories = student\_borrowed\_books[student\_borrowed\_books['图书ID'].isin(student\_books\_not\_returned[student\_books\_not\_returned > 0].index)]['图书分类号']

sutdent\_category\_counts = studetn\_not\_returned\_categories.value\_counts()

student\_most\_common\_category = sutdent\_category\_counts.idxmax()

print("学生没有归还的书籍中数量最多的图书分类号：", student\_most\_common\_category,"类别：", parse\_classification\_number(student\_most\_common\_category))

10． 根据以上数据分析结果，画出经济管理学院的教师2014年，2015年，2016年，2017年所借书籍类别占前10的类别书籍走向图，如图所示。



核心代码：

teacher\_points = {}

for year in jiehuan:

borrow\_df = jiehuan[year]

merge\_df = pd.merge(borrow\_df, content, on='图书ID')

merge\_df = pd.merge(merge\_df, reader, on='读者ID')

merge\_df = merge\_df[(merge\_df['操作类型']=='借') & (merge\_df['读者类型'] == '教师')]

for bid, bnum in teacher\_top\_10.items():

if bid not in teacher\_points:

teacher\_points[bid]=[]

teacher\_points[bid].append(len(merge\_df[merge\_df['图书分类号']==bid]))

print(teacher\_points)

plt.figure(figsize=(10, 6))

years = [2014, 2015, 2016, 2017]

np.array(years, dtype=int)

for key in teacher\_points:

plt.plot(years, teacher\_points[key], marker='o', label=key)

plt.title(f"Top 10 Book Categories for Teachers (2014-2017)")

plt.xlabel("Year")

plt.ylabel("Count")

plt.xticks(years)

plt.legend(title="Book Categories")

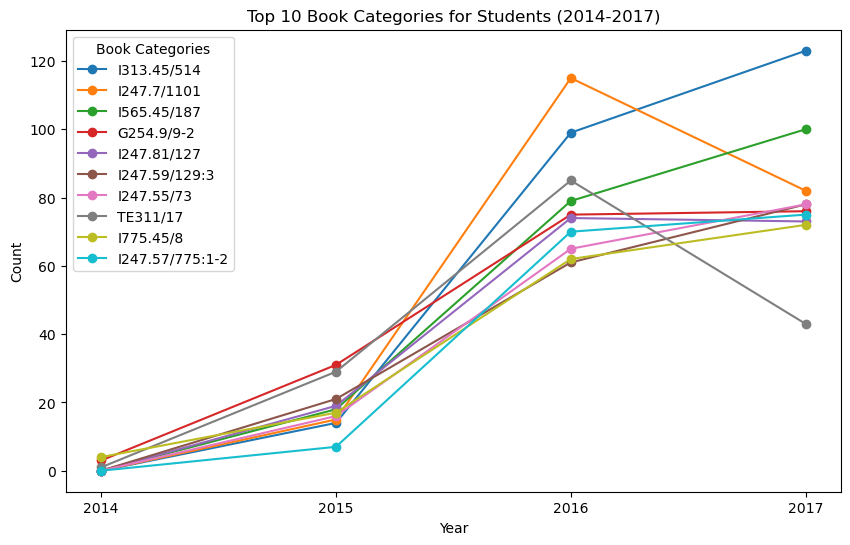
plt.xticks(years)

plt.show()

从图形结果分析来看经济管理学院教师最关注的知识是中国民族器乐，热度增加最快的知识是神医傻妃，推荐图书馆购入的图书是中国民族器乐类别的图书。

### 4.2.2任务2分析

11. 根据以上数据分析结果，画出经济管理学院的学生2014年，2015年，2016年，2017年所借书籍类别占前10的类别书籍走向图，如图所示。



核心代码：

student\_points = {}

for year in jiehuan:

borrow\_df = jiehuan[year]

merge\_df = pd.merge(borrow\_df, content, on='图书ID')

merge\_df = pd.merge(merge\_df, reader, on='读者ID')

student\_condition = ((merge\_df['读者类型'] == '博士研究生') | (merge\_df['读者类型'] == '本科生') | (merge\_df['读者类型'] == '研究生'))

merge\_df = merge\_df[(merge\_df['操作类型']=='借') & student\_condition]

for bid, bnum in student\_top\_10.items():

if bid not in student\_points:

student\_points[bid]=[]

student\_points[bid].append(len(merge\_df[merge\_df['图书分类号']==bid]))

print(student\_points)

import matplotlib.pyplot as plt

import numpy as np

plt.figure(figsize=(10, 6))

years = [2014, 2015, 2016, 2017]

years = np.array(years, dtype=int)

for key in student\_points:

plt.plot(years, student\_points[key], marker='o', label=key)

plt.title(f"Top 10 Book Categories for Students (2014-2017)")

plt.xlabel("Year")

plt.ylabel("Count")

plt.xticks(years)

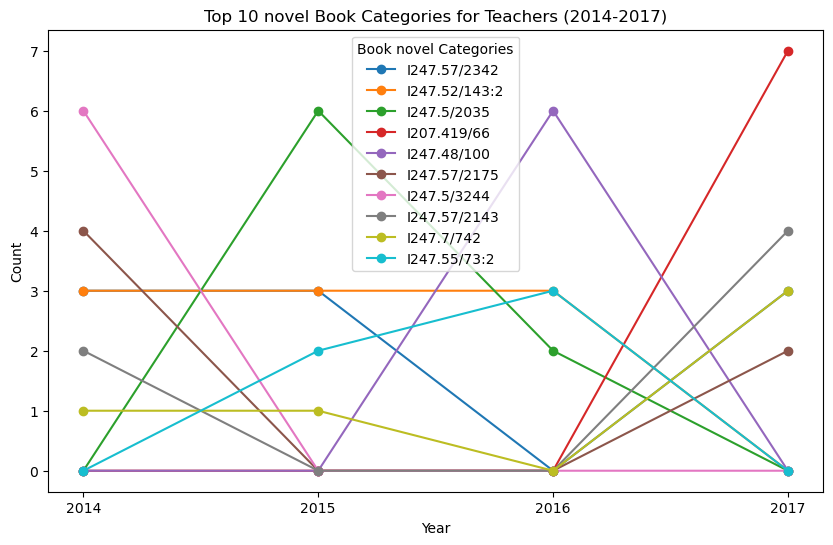
plt.legend(title="Book Categories")

plt.show()

从图形结果分析来看经济管理学院学生最关注的知识是日本文学，热度增加最快的知识是日本文学，推荐图书馆购入的图书是日本文学类别的图书。

### 4.2.3任务3分析

根据以上数据分析结果，画出经济管理学院的教师2014年，2015年，2016年，2017年所借小说书籍类别占前10的类别书籍走向图，如图所示。



核心代码：

xs\_teacher\_points = {}

for year in jiehuan:

borrow\_df = jiehuan[year]

merge\_df = pd.merge(borrow\_df, content, on='图书ID')

merge\_df = pd.merge(merge\_df, reader, on='读者ID')

merge\_df = merge\_df[(merge\_df['操作类型']=='借') & (merge\_df['读者类型'] == '教师')]

for bid, bnum in xs\_teacher\_top\_10.items():

if bid not in xs\_teacher\_points:

xs\_teacher\_points[bid]=[]

xs\_teacher\_points[bid].append(len(merge\_df[merge\_df['图书分类号']==bid]))

print(xs\_teacher\_points)

plt.figure(figsize=(10, 6))

years = [2014, 2015, 2016, 2017]

np.array(years, dtype=int)

for key in xs\_teacher\_points:

plt.plot(years, xs\_teacher\_points[key], marker='o', label=key)

plt.title(f"Top 10 novel Book Categories for Teachers (2014-2017)")

plt.xlabel("Year")

plt.ylabel("Count")

plt.legend(title="Book novel Categories")

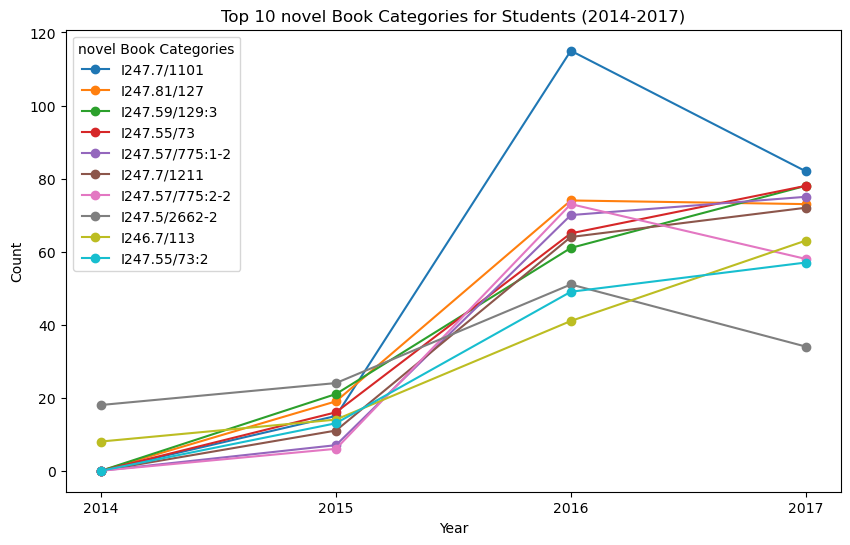
plt.xticks(years)

plt.show()

从图形结果分析来看经济管理学院老师最关注的小说是刘心武评点《金瓶梅》，热度增加最快的知识是各体文学评论和研究，推荐图书馆购入的图书是各体文学评论和研究类别的图书。

### 4.2.4任务4分析

根据以上数据分析结果，画出经济管理学院的教师2014年，2015年，2016年，2017年所借小说书籍类别占前10的类别书籍走向图，如图所示。



核心代码：

xs\_student\_points = {}

for year in jiehuan:

borrow\_df = jiehuan[year]

merge\_df = pd.merge(borrow\_df, content, on='图书ID')

merge\_df = pd.merge(merge\_df, reader, on='读者ID')

student\_condition = ((merge\_df['读者类型'] == '博士研究生') | (merge\_df['读者类型'] == '本科生') | (merge\_df['读者类型'] == '研究生'))

merge\_df = merge\_df[(merge\_df['操作类型']=='借') & student\_condition]

for bid, bnum in xs\_student\_top\_10.items():

if bid not in xs\_student\_points:

xs\_student\_points[bid]=[]

xs\_student\_points[bid].append(len(merge\_df[merge\_df['图书分类号']==bid]))

print(xs\_student\_points)

all\_df = pd.DataFrame()

for year in jiehuan:

borrow\_df = jiehuan[year]

merge\_df = pd.merge(borrow\_df, content, on='图书ID')

merge\_df = pd.merge(merge\_df, reader, on='读者ID')

all\_df = pd.concat([all\_df, teacher\_df], ignore\_index=True)

从图形结果分析来看经济管理学院学生最关注的小说是从你的全世界路过:让所有人心动的故事，热度增加最快的知识是当代作品（1949--，推荐图书馆购入的图书是当代作品（1949--类别的图书。

## 4.3书籍推荐

### 4.3.1使用jieba分词

将DataFrame all\_df 中的书名进行中文分词，并统计分词后每个词汇出现的次数（词频）。在统计之前，对词汇进行了停用词过滤，去掉了常见的无意义词汇。

核心代码：

import jieba

from wordcloud import WordCloud

stopwords\_files = ['data/停用词及敏感词库/baidu\_stopwords.txt', 'data/停用词及敏感词库/cn\_stopwords.txt',

'data/停用词及敏感词库/hit\_stopwords.txt', 'data/停用词及敏感词库/scu\_stopwords.txt',

'data/停用词及敏感词库/stop-words.txt']

stopwords = set()

for file in stopwords\_files:

with open(file, 'r', encoding='utf-8') as f:

for line in f:

stopwords.add(line.strip())

stopwords.add('版')

words\_list = []

for book\_name in all\_df['书名']:

words = jieba.cut(str(book\_name))

words\_filtered = [word for word in words if word not in stopwords and word.strip()!='']

words\_list.extend(words\_filtered)

word\_counts = {}

for word in words\_list:

if word in word\_counts:

word\_counts[word] += 1

else:

word\_counts[word] = 1

### 4.3.2制作词云

使用WordCloud类生成词云图，并使用matplotlib库将词云图绘制出来。在代码中，我们通过generate\_from\_frequencies方法将字典中的词频信息传递给WordCloud类，并设置词云图的尺寸和背景颜色。然后，使用plt.imshow()方法绘制词云图，并使用plt.axis('off')关闭坐标轴，以便更好地展示词云效果。

核心代码：

font\_path = r'C:\Users\dulun\Desktop\jiedan2\NotoSerifSC-Regular.otf'

wordcloud = WordCloud(width=800, height=400, max\_words=50, background\_color='white', font\_path=font\_path)

wordcloud.generate\_from\_frequencies(word\_counts)

plt.figure(figsize=(10, 6))

plt.imshow(wordcloud, interpolation='bilinear')

plt.axis('off')

plt.show()

top\_words = sorted(word\_counts.items(), key=lambda x: x[1], reverse=True)[:10]

for word, count in top\_words:

print(f'{word}: {count}')

### 4.3.3图书推荐



根据经济管理学院图书馆借阅书籍名称的词云图统计词频，为该学院系的师生购书，我会提出以下建议。这些建议是基于数据分析结果，并考虑到学校思政教育的主要目的，以便给出明确的图书推荐：

1. 经济学：考虑购买更多关于经济学原理、宏观经济学和微观经济学的经典教材，以满足学生对经济学基础知识的学习需求。同时，推荐购买一些当代经济学理论研究和经济趋势分析的专业书籍，帮助学生了解最新的经济发展和政策动向。

2. 方法：增加购买关于研究方法和数据分析的教材和参考书，包括统计学方法、经济学实证研究方法、社会科学研究方法等。这将有助于学生在学术研究和数据处理方面提高能力。

3. PETS：鉴于PETS可能是某个特定学科、项目或领域的缩写，建议与学院相关的教师和学生进一步明确PETS代表的具体内容，并根据需求购买相关的专业书籍或教材，满足学习和研究需求。

4. 商业模式：考虑购买关于商业模式设计和创新管理的书籍，帮助学生了解商业模式的构建和管理，以应对不断变化的商业环境。

5. 技术：增加购买关于信息技术、数字化转型、人工智能等方面的书籍，帮助学生了解技术对经济和商业的影响，提高数字化时代的竞争力。

6. 统计分析：推荐购买关于统计分析和数据处理的实用手册和案例教程，帮助学生掌握数据分析技能，为经济和商业决策提供可靠支持。

7. 研究：增加购买关于学术研究方法、论文写作和学术规范的书籍，帮助学生提高学术写作和研究水平。

9. 社会：推荐购买关于社会科学和社会经济问题的书籍，帮助学生深入了解社会变革、社会问题和社会经济发展。

这些建议是基于数据分析结果和学校思政教育目标进行的图书推荐，目的是满足经济管理学院师生对不同领域知识和能力的需求，提升学术研究和实践水平。购书时，图书馆可以参考学院师生的需求和课程设置，选择适合学院教学和研究的优质图书资源。

# 结 论

经济管理学院图书大数据分析项目使用了Python3开发，采用Anaconda作为集成开发环境。分析过程中使用了Numpy扩展库、Pandas扩展库、Matplotlib扩展库。

该项目通过经济管理学院师生2014年、2015年、2016年的图书借阅数据。统计分析了如下任务：

1、统计经济管理学院的教师2014年，2015年，2016年，2017年所借书籍类别占前10的类别分别为文学史、文学思想史、当代作品、中国民族器乐、美国文学、日本文学 、法国文学，经济管理学院的师生关注的知识是当代作品，推荐图书馆购入的相关书籍是中国民族器乐，中国民族器乐类别类别。

2、统计经济管理学院的教师2014年，2015年，2016年，2017年所最喜欢看的小说分别是平凡的世界、平凡的世界、活着、白鹿原。

3、统计经济管理学院的教师2014年，2015年，2016年，2017年一共有617147本书没有归还。没有归还的书籍文学史、文学思想史最多。

4、统计经济管理学院的学生2014年，2015年，2016年，2017年一共有9735本书没有归还。没有归还的书籍日本文学最多。

5、从经济管理学院的教师2014年，2015年，2016年，2017年所借书籍类别占前10的类别书籍走向图结果分析来看经济管理学院教师最关注的知识是中国民族器乐，热度增加最快的知识是神医傻妃，推荐图书馆购入的图书是中国民族器乐类别的图书。

6、从经济管理学院的学生2014年，2015年，2016年，2017年所借书籍类别占前10的类别书籍走向图结果分析来看经济管理学院教师最关注的知识是中国民族器乐，热度增加最快的知识是神医傻妃，推荐图书馆购入的图书是中国民族器乐类别的图书。

7、从经济管理学院的教师2014年，2015年，2016年，2017年所借小说书籍类别占前10的类别书籍走向图结果分析来看经济管理学院老师最关注的小说是刘心武评点《金瓶梅》，热度增加最快的知识是各体文学评论和研究，推荐图书馆购入的图书是各体文学评论和研究类别的图书。

8、经济管理学院的教师2014年，2015年，2016年，2017年所借小说书籍类别占前10的类别书籍走向图结果分析来看经济管理学院学生最关注的小说是从你的全世界路过:让所有人心动的故事，热度增加最快的知识是当代作品（1949--，推荐图书馆购入的图书是当代作品（1949--类别的图书。

9、根据经济管理学院图书馆借阅书籍名称的词云图统计词频，为该学院系的师生购书，建议购买以下图书：增加经济学原理、宏观经济学和微观经济学的经典教材，以满足学生对经济学基础知识的学习需求；提供关于研究方法和数据分析的教材，培养学生在学术研究和数据处理方面的能力；根据PETS代表的具体内容，购买相关专业书籍，满足学习和研究需求；推荐商业模式设计和创新管理的书籍，帮助学生了解商业环境；增加信息技术、数字化转型和人工智能方面的书籍，提高学生在数字化时代的竞争力；推荐统计分析和数据处理的实用手册，帮助学生掌握数据分析技能；增加学术研究方法、论文写作和学术规范的书籍，提高学生的学术写作和研究水平；根据具体含义，购买相关专业书籍，满足学习和研究需求；推荐社会科学和社会经济问题的书籍，深入了解社会变革和经济发展。这些建议基于数据分析结果和学校思政教育目标，旨在满足经济管理学院师生对不同领域知识和能力的需求，提升学术研究和实践水平。

# 参考文献

[1]许辉.基于大数据分析的高校图书馆服务系统设计研究[J].信息与电脑(理论版),2023,35(07):158-160.

[2]王世华.基于用户需求的高校图书馆智慧服务大数据分析系统构建——以上海大学为例[J].高校图书馆工作,2022,42(06):43-47.

[3]赵楠.以大数据为驱动的智慧图书馆构建与服务模式[J].图书馆学刊,2022,44(01):63-68.DOI:10.14037/j.cnki.tsgxk.2022.01.012.

[4]黄富英.大数据在公共图书馆精准服务中的应用[J].电子技术,2021,50(12):228-229.

[5]陈艳,丁晓梅.高等院校图书馆大数据技术的应用模式分析[J].衡阳师范学院学报,2021,42(06):117-121.DOI:10.13914/j.cnki.cn43-1453/z.2021.06.020.

[6]郭惠.基于大数据分析技术的数字图书馆信息检索模型设计[J].电子技术与软件工程,2021(19):200-201.

[7]吴荣.基于大数据分析技术的数字图书馆信息检索模型设计[J].数字技术与应用,2021,39(05):121-123.DOI:10.19695/j.cnki.cn12-1369.2021.05.40.