**Course Project Report**

**Task 1:**

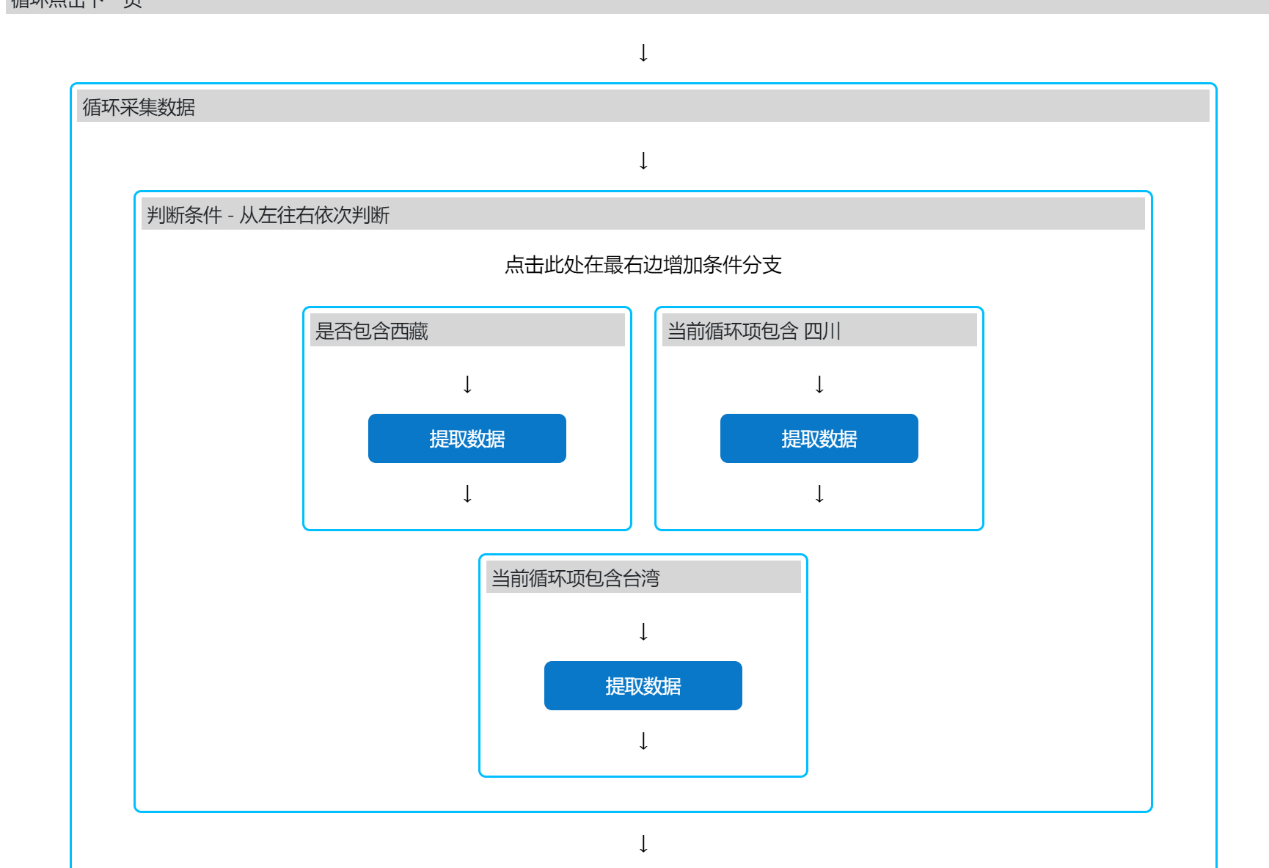
**爬取中国地震台网历史数据：**

1. **纬度分别大于10度、20度、30度、40度、50度区域的数据；**
2. **爬取查询结果页面前50页内容；**
3. **过滤结果，保存结果中含有西藏、四川、台湾的数据；**
4. **对保存的数据进行数据分析，可视化三地地震信息（可视化方式越丰富、信息表现越全面分数越高）。**

**报告中需要详细介绍设计思路和步骤（截图），并且展示可视化信息并进行分析。将爬取保存的数据文件命名为：小组名-Task1.csv（如有多个文件则分别命名后再压缩成一个压缩包，命名为小组名-Task1.zip） ，与报告一起上传至Moodle。**







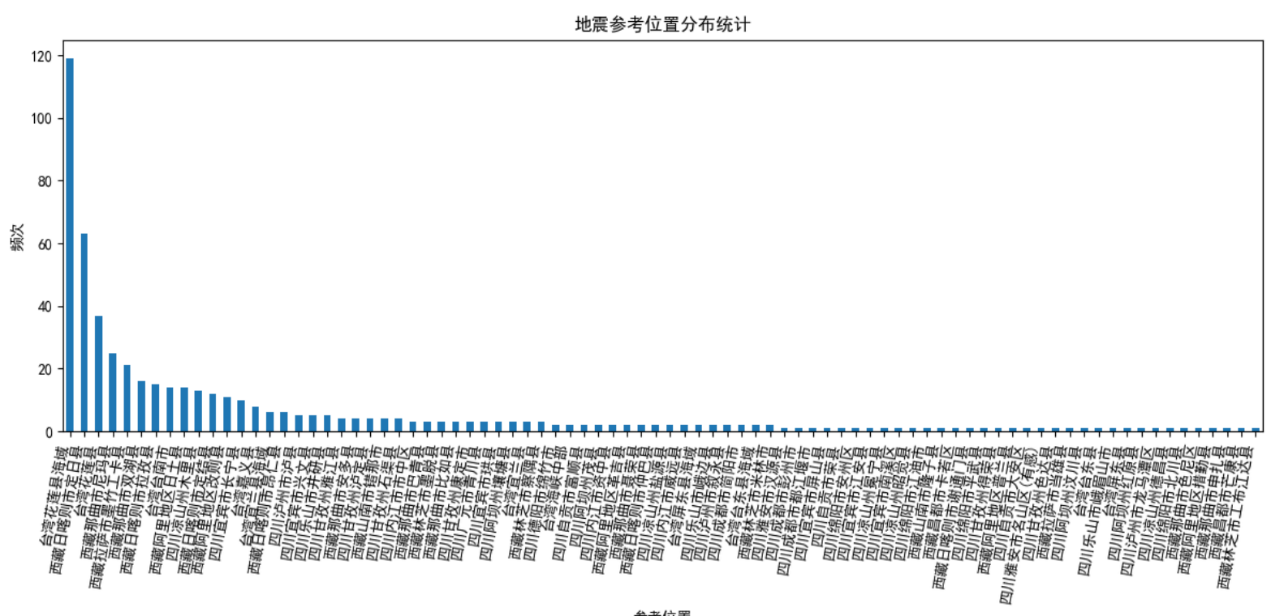




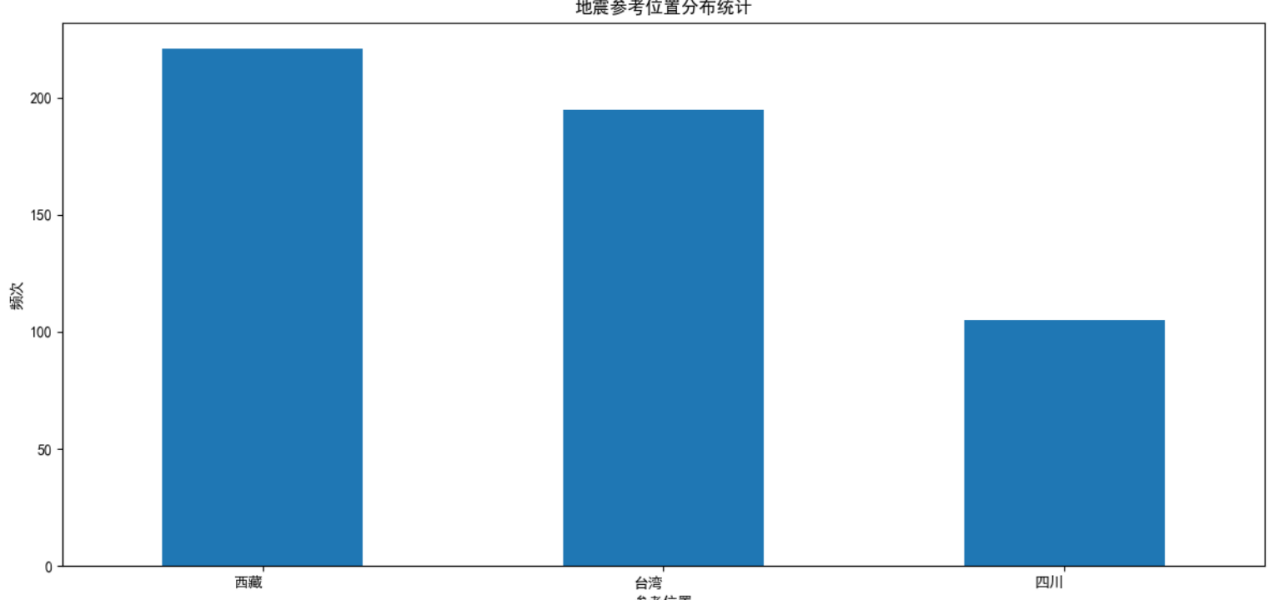


执行任务时，选择中国地震台网，然后直接执行就能得到对不同高纬度要求的结果，直至收集到五十条为止。

使用 easysider 设计数据爬取脚本。脚本逻辑为打开网页后，循环输入不同纬度（分别对应大于 10 度、20 度等要求 ）并点击查询。每次查询后，通过循环点击下一页按钮，对每一页进行数据采集。在采集过程中，检查数据字段，若包含西藏、四川或台湾相关信息，则进行采集。通过这样的循环操作 49 次（从第一页到第 50 页 ），来爬取 50 页的内容。通过这样的循环机制，重复操作 49 次（从第 2 页到第 50 页 ，第 1 页默认加载 ），从而实现 50 页内容的完整爬取。

import pandas as pd  
import matplotlib.pyplot as plt  
  
# 读取数据  
df = pd.read\_csv("./中国地震台数据.csv")  
  
df = df.iloc[:,0:6]  
  
df.drop\_duplicates(inplace=True)  
# df.info()  
# print( df.head() )  
  
  
location\_counts = df['参考位置'].value\_counts()  
  
plt.figure(figsize=(12, 6))  
location\_counts.plot(kind='bar')  
plt.rcParams['font.sans-serif'] = ['SimHei']  
plt.title('地震参考位置分布统计')  
plt.xlabel('参考位置')  
plt.ylabel('频次')  
plt.xticks(rotation=80, ha='right')  
plt.tight\_layout()  
plt.show()  
  
plt.close()

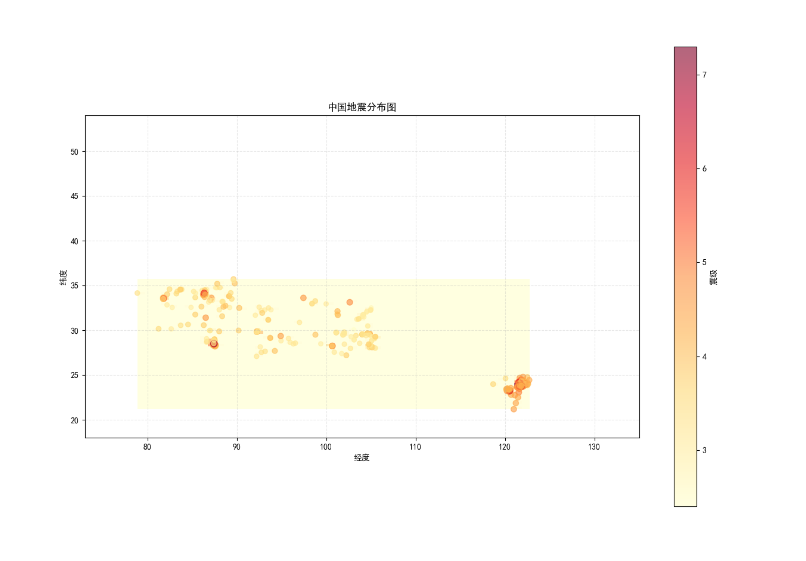
import pandas as pd  
import matplotlib.pyplot as plt  
  
df = pd.read\_csv("./中国地震台数据.csv")  
df = df.iloc[:,0:6]  
df.drop\_duplicates(inplace=True)  
# df.info()  
print( df.head() )  
location = df.pop('参考位置')  
location.astype(str)  
location=location.str[:2]   
df.insert(1,'参考位置',location)  
print(df.head())  
  
location\_counts = df['参考位置'].value\_counts()  
  
print("\n参考位置统计结果：")  
print(location\_counts)  
  
plt.figure(figsize=(12, 6))  
location\_counts.plot(kind='bar')  
plt.rcParams['font.sans-serif'] = ['SimHei']  
plt.title('地震参考位置分布统计')  
plt.xlabel('参考位置')  
plt.ylabel('频次')  
plt.xticks(rotation=0, ha='right')  
plt.tight\_layout()  
plt.show()  
  
plt.close()

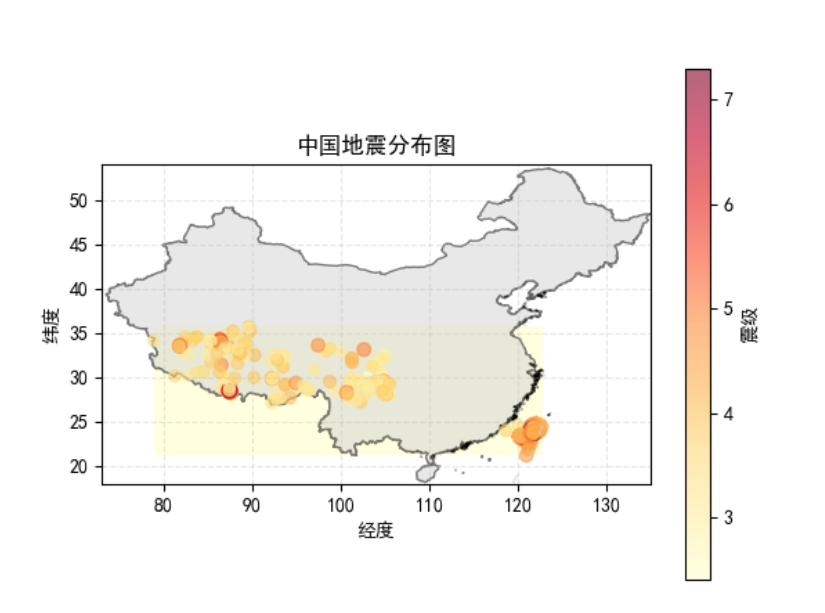


从代码 location\_counts = df['参考位置'].value\_counts() 和 location\_counts.plot(kind='bar') 生成的 “地震参考位置分布统计” 柱状图来看：

高频地震区域：通过代码统计出的结果直观显示，西藏、四川、台湾地区在柱状图中对应的柱子高度显著高于其他地区。这是因为 value\_counts 函数对 “参考位置” 列中的每个地区出现的次数进行了准确计数，反映在柱状图上就是高度代表频次。西藏位于印度洋板块与亚欧板块的强烈碰撞地带，板块的强烈挤压使得地壳岩石变形、破裂，从而引发频繁的地震活动。四川处于多个活跃的地质构造带，如龙门山断裂带等，板块间的应力积累和突然释放导致地震频发。台湾地处环太平洋地震带，太平洋板块与亚欧板块在此相互作用强烈，地壳运动极为活跃，因此地震发生的频次较高。

低频地震区域：除了上述三个地区外，其他地区的柱子高度相对较低。这表明这些地区的地震发生次数较少，但并不意味着它们没有地震风险。代码的统计结果清晰地展示了这种差异，有助于我们对不同地区的地震活动程度有一个初步的量化认识。

import pandas as pd  
import matplotlib.pyplot as plt  
import numpy as np  
from matplotlib.colors import LinearSegmentedColormap  
  
# 读取数据  
df = pd.read\_csv("./中国地震台数据.csv")  
df = df.iloc[:,0:6]  
df.drop\_duplicates(inplace=True)  
  
# 设置中文字体  
plt.rcParams['font.sans-serif'] = ['SimHei']  
plt.rcParams['axes.unicode\_minus'] = False  
  
# 创建图形  
plt.figure(figsize=(15, 10))  
  
# 绘制中国地图底图  
plt.plot([73.66, 135.05, 135.05, 73.66, 73.66], [53.55, 53.55, 18.15, 18.15, 53.55], 'k-', linewidth=1)  
=  
x = df['经度'].values  
y = df['纬度'].values  
heatmap, xedges, yedges = np.histogram2d(x, y, bins=50)  
extent = [xedges[0], xedges[-1], yedges[0], yedges[-1]]  
plt.imshow(heatmap.T, extent=extent, origin='lower', cmap='YlOrRd', alpha=0.6)  
=  
scatter = plt.scatter(df['经度'], df['纬度'],   
 c=df['震级'],   
 cmap='YlOrRd',  
 s=df['震级']\*10, # 点的大小根据震级变化  
 alpha=0.6)  
  
cbar = plt.colorbar(scatter)  
cbar.set\_label('震级')  
  
plt.title('中国地震分布图')  
plt.xlabel('经度')  
plt.ylabel('纬度')  
  
plt.xlim(73, 135)  
plt.ylim(18, 54)  
  
plt.grid(True, linestyle='--', alpha=0.3)  
plt.show()  
  
i=300, bbox\_inches='tight')  
  


import pandas as pd  
import matplotlib.pyplot as plt  
import numpy as np  
from matplotlib.colors import LinearSegmentedColormap  
import cnmaps  
import geopandas as gpd  
from shapely.geometry import shape  
  
# 读取数据  
df = pd.read\_csv("./中国地震台数据.csv")  
df = df.iloc[:,0:6]  
df.drop\_duplicates(inplace=True)  
  
# 设置中文字体  
plt.rcParams['font.sans-serif'] = ['SimHei']  
plt.rcParams['axes.unicode\_minus'] = False  
  
# 创建图形  
plt.figure(figsize=(15, 10))  
  
# 获取中国地图数据  
china\_shapes = cnmaps.get\_adm\_maps(level='国')  
# 提取geometry对象并转换为shapely几何对象  
geometries = [shape(item['geometry'].\_\_geo\_interface\_\_) for item in china\_shapes]  
china\_gdf = gpd.GeoDataFrame(geometry=geometries, crs="EPSG:4326")  
  
# 绘制中国地图底图  
china\_gdf.plot(color='lightgray', edgecolor='black', alpha=0.5)  
  
# 创建热力图  
x = df['经度'].values  
y = df['纬度'].values  
heatmap, xedges, yedges = np.histogram2d(x, y, bins=50)  
extent = [xedges[0], xedges[-1], yedges[0], yedges[-1]]  
plt.imshow(heatmap.T, extent=extent, origin='lower', cmap='YlOrRd', alpha=0.6)  
  
# 绘制散点图  
scatter = plt.scatter(df['经度'], df['纬度'],   
 c=df['震级'],   
 cmap='YlOrRd',  
 s=df['震级']\*10, # 点的大小根据震级变化  
 alpha=0.6)  
  
# 添加颜色条  
cbar = plt.colorbar(scatter)  
cbar.set\_label('震级')  
  
# 设置图表属性  
plt.title('中国地震分布图')  
plt.xlabel('经度')  
plt.ylabel('纬度')  
  
# 设置坐标轴范围  
plt.xlim(73, 135)  
plt.ylim(18, 54)  
  
# 添加网格  
plt.grid(True, linestyle='--', alpha=0.3)  
plt.show()  
  


地震经纬度与震级关系可视化结果分析

地震空间分布密度：对于代码 plt.hist2d(x=df['经度'].values, y=df['纬度'].values, bins=50) 生成的二维直方图，它将地震数据按照经度和纬度进行了分布统计。在我国西南地区（主要是西藏、四川一带 ）以及东南沿海的台湾地区，颜色较深（假设颜色深代表地震分布密度高 ）。这是因为 hist2d 函数将经纬度空间划分为 50×50 个区间，并统计每个区间内地震的数量，颜色的深浅反映了区间内地震数量的多少。西藏由于板块碰撞的强烈作用，地壳岩石破碎，形成了众多的地震活动区域，导致该地区在二维直方图上呈现出较高的地震分布密度。四川的多个活跃断裂带，如龙门山断裂带、鲜水河断裂带等，使得地震在该地区频繁发生，从而在直方图上显示出较高的密度。台湾处于板块边界，板块的强烈运动导致地震在空间上密集分布，因此在图中颜色较深。

震级与地理位置关联：由代码 scatter = plt.scatter(df['经度'], df['纬度'], c=df['震级'], cmap='viridis', s=df['震级']\*10, alpha=0.6) 生成的散点图中，台湾地区的点相对较大且颜色较深（假设颜色深、点大代表震级高 ）。这是因为散点图的点大小由 s=df['震级']\*10 决定，颜色由 c=df['震级'] 通过 viridis 颜色映射表确定，震级越高，点越大且颜色越深。台湾处于环太平洋地震带这一全球最活跃的地震构造区域，板块间的强烈碰撞和俯冲作用使得该地区能够积累巨大的能量，从而引发较高震级的地震。在西藏和四川地区，也存在一些较大且颜色较深的点，说明这些地区也会发生较高震级的地震。西藏的地震活动与板块碰撞的远程效应以及地壳内部的复杂构造有关，而四川的地震震级分布则受到其活跃断裂带的控制，不同断裂带的力学性质和活动方式导致了震级在空间上的差异。

**Task 2:**

**爬取纵横中文网小说：**

1. **TOP10月票榜前五的小说；**
2. **每本小说爬取前50章（少于50章的爬取全部章节）；**
3. **将爬取的数据进行数据清洗和整理，以方便阅读的格式和文件分别保存5本小说；**
4. **为5本小说分别绘制词云并保存成图片格式。**

**报告中需要详细介绍设计思路和步骤（截图），并且展示5本小说的词云。将数据处理后的5本小说和词云图片文件一起打包为一个zip文件，文件名：小组名-Task2.zip，与报告一起上传至Moodle。**

**设计思路**

数据获取思路明确从纵横中文网爬取小说数据，目标是获取 TOP10 月票榜前五的小说，且每本小说爬取前 50 章（少于 50 章则全部爬取 ）。需要利用网络爬虫技术，模拟浏览器请求，解析网页结构来定位小说列表及章节内容。考虑到网站可能存在反爬机制，可能需要设置合理的请求头、控制请求频率等。

选择合适的爬虫框架或库，如 BeautifulSoup 结合 requests，或者 Scrapy 框架，以高效地提取所需信息。

数据清洗和整理思路爬取到的数据可能存在格式不规范、缺失值、重复值等问题。通过代码中的数据处理步骤，如使用 Pandas 库的函数，先对数据进行读取（pd.read\_csv ），然后针对不同列的缺失值，使用 fillna 方法进行填充，保证数据的完整性。

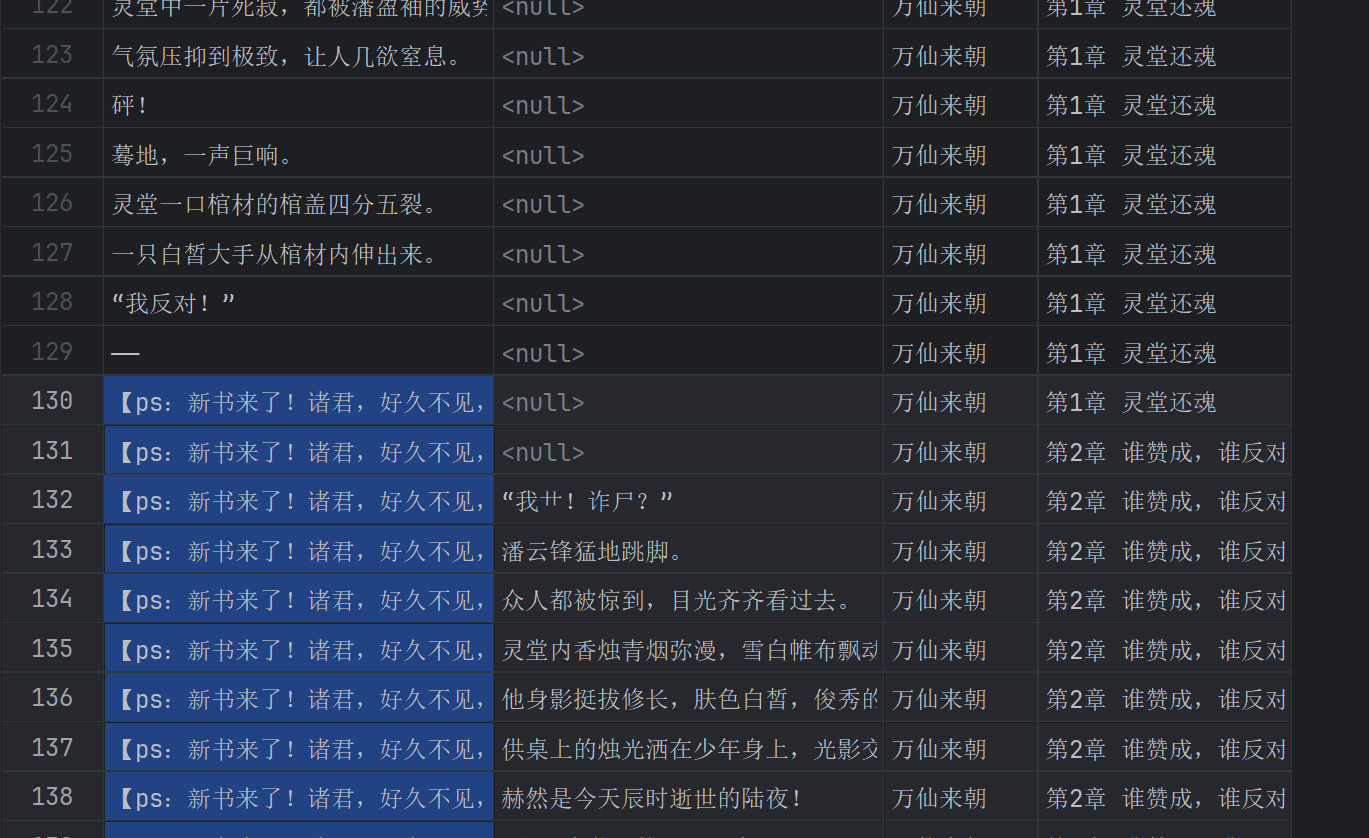
识别并删除重复记录（drop\_duplicates ），避免数据冗余对后续分析造成干扰。同时，根据数据特点，删除不必要的列（drop 函数 ），使数据结构更加简洁，便于后续操作。

数据保存思路为方便后续使用和管理，将处理后的数据以合适的格式保存。代码中根据作者名称进行分组，清理文件名中的非法字符（通过字符串操作筛选合法字符 ），然后将每组数据保存为单独的 CSV 文件，确保数据存储的规范性和可读性。

词云制作思路词云可以直观展示小说文本中的高频词汇。利用 Python 的 WordCloud 库，先将 CSV 文件中的文本内容合并为一个字符串（通过 Pandas 读取文件并处理文本列 ），再使用结巴分词（jieba 库 ）对文本进行分词处理，将文本转化为适合生成词云的词汇列表。

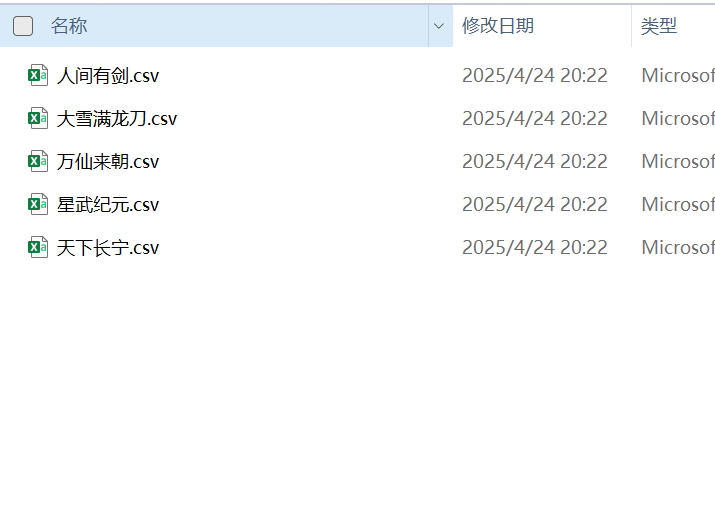
设置词云的相关参数，如字体（确保支持中文显示 ）、背景颜色、最大词汇数等，然后生成词云并保存为图片格式，实现对小说文本内容的可视化呈现。

**数据预处理部分：**









import numpy as np  
import pandas as pd  
import os  
  
# 读取CSV文件  
df = pd.read\_csv('小说.csv')  
  
# 将第一列为空的部分用第二列补充  
df.iloc[:, 0] = df.iloc[:, 0].fillna(df.iloc[:, 1])  
  
# 将第三列为空的部分用第四列补充  
df.iloc[:, 2] = df.iloc[:, 2].fillna(df.iloc[:, 3])  
  
# 删除第二列和第四列  
df = df.drop(df.columns[[1, 3]], axis=1)  
  
# 对数据进行去重  
df = df.drop\_duplicates()  
  
# 根据观察可以得出，从第二章开始第一列内容为一些广告和其他的奇怪内容  
# 所以我们直接将其赋值并且将 第二列内容删除  
# 最后形成的格式就是第一列为 内容  
# 而第二列为作评名称  
# 第三列是章节名称  
df.iloc[130:,0] = df.iloc[130:,1]  
df = df.drop(df.columns[[1]],axis=1)  
df = df.drop\_duplicates()  
  
# 创建保存文件的目录  
output\_dir = 'processed\_novels'  
if not os.path.exists(output\_dir):  
 os.makedirs(output\_dir)  
  
# 根据作者名称分组并分别保存  
for author, group in df.groupby(df.columns[1]): # 按第二列（作品名称）分组  
 # 清理文件名中的非法字符  
 safe\_author\_name = "".join(x for x in author if x.isalnum() or x in (' ', '-', '\_', '@'))  
 output\_file = os.path.join(output\_dir, f'{safe\_author\_name}.csv')  
 group.to\_csv(output\_file, index=False, encoding='utf-8-sig')  
 print(f'已保存作者 {author} 的作品到 {output\_file}')  
  
print("所有数据处理完成！")

**数据词云：**

import pandas as pd  
import matplotlib.pyplot as plt  
from wordcloud import WordCloud  
import jieba  
import os  
import numpy as np  
from PIL import Image  
  
def generate\_wordcloud(csv\_file, output\_dir):  
 # 提取文件名（不含扩展名）作为保存图片的名称  
 file\_name = os.path.splitext(os.path.basename(csv\_file))[0]  
   
 # 读取CSV文件  
 df = pd.read\_csv(csv\_file, encoding='utf-8-sig')  
   
 # 将所有文本内容合并成一个字符串  
 # 假设第一列是内容列  
 text = ' '.join(df.iloc[:, 0].dropna().astype(str).tolist())  
   
 # 使用jieba进行中文分词  
 words = ' '.join(jieba.cut(text))  
   
 # 生成词云  
 wordcloud = WordCloud(  
 font\_path='simhei.ttf', # 使用黑体字体，确保能显示中文  
 width=800,  
 height=600,  
 background\_color='white',  
 max\_words=200,  
 collocations=False,  
 contour\_width=1,  
 contour\_color='steelblue'  
 )  
   
 # 生成词云图  
 wordcloud.generate(words)  
   
 # 显示词云图  
 plt.figure(figsize=(10, 8))  
 plt.imshow(wordcloud, interpolation='bilinear')  
 plt.axis('off')  
 plt.tight\_layout()  
   
 # 保存图片  
 output\_file = os.path.join(output\_dir, f'{file\_name}\_wordcloud.png')  
 plt.savefig(output\_file, dpi=300)  
 plt.close()  
   
 print(f'已生成 {file\_name} 的词云图，保存为 {output\_file}')  
  
def main():  
 # 当前脚本所在目录  
 current\_dir = os.path.dirname(os.path.abspath(\_\_file\_\_))  
   
 # 创建输出目录  
 output\_dir = os.path.join(current\_dir, 'wordcloud\_images')  
 if not os.path.exists(output\_dir):  
 os.makedirs(output\_dir)  
   
 # 获取当前目录下所有CSV文件  
 csv\_files = [f for f in os.listdir(current\_dir) if f.endswith('.csv')]  
   
 # 为每个CSV文件生成词云  
 for csv\_file in csv\_files:  
 csv\_file\_path = os.path.join(current\_dir, csv\_file)  
 generate\_wordcloud(csv\_file\_path, output\_dir)  
  
if \_\_name\_\_ == '\_\_main\_\_':  
 main()

**步骤说明**

数据爬取首先确定纵横中文网的小说排行榜页面 URL，使用网络请求库（如 requests ）发送 GET 请求获取网页内容。

使用网页解析库（如 BeautifulSoup ）解析页面 HTML 结构，定位到 TOP10 月票榜前五小说的链接。

依次访问每本小说的页面，找到章节列表，根据规则（前 50 章或全部章节 ）获取各章节内容，将内容整理为结构化数据（如列表或字典形式 ），方便后续保存为 CSV 文件。

数据清洗与整理读取保存的 CSV 文件（df = pd.read\_csv('小说.csv') ），加载到 Pandas 数据框中。

对数据框中可能存在缺失值的列进行填充，如 df.iloc[:, 0] = df.iloc[:, 0].fillna(df.iloc[:, 1]) 和 df.iloc[:, 2] = df.iloc[:, 2].fillna(df.iloc[:, 3]) ，分别对第一列和第三列进行缺失值填充。

删除不必要的列（df = df.drop(df.columns[[1, 3]], axis=1) ），简化数据结构。

去除重复记录（df = df.drop\_duplicates() ），保证数据的唯一性。

根据观察对数据进一步处理，如发现某些章节存在广告等异常内容，通过数据筛选和处理操作（如代码中对特定行的处理 ）进行清理，最终整理出规范的小说章节数据。

数据保存创建保存文件的目录（output\_dir = 'processed\_novels' ，并通过 os.makedirs 函数创建目录 ）。

按作者名称对数据进行分组（for author, group in df.groupby(df.columns[1]) ），清理作者名称中的非法字符（safe\_author\_name = "".join(x for x in author if x.isalnum() or x in (' ', '\_', '-', '@')) ）。

将每组数据保存为单独的 CSV 文件（group.to\_csv(output\_file, index=False, encoding='utf-8-sig') ），确保数据按作者分类保存且文件编码正确。

词云制作定义生成词云的函数（generate\_wordcloud ），在函数内读取 CSV 文件（df = pd.read\_csv(csv\_file, encoding='utf-8-sig') ）。

将文本内容合并为一个字符串（text = " ".join(df.iloc[:, 0].dropna().astype(str).tolist()) ），并使用结巴分词进行分词（words = " ".join(jieba.cut(text)) ）。

设置词云参数，如字体（font\_path="simhei.ttf" ）、宽度、高度、背景颜色、最大词汇数等，生成词云（wordcloud = WordCloud(...) ，wordcloud.generate(words) ）。



**数据可视化结果结合代码分析**

词云展示高频词汇：从生成的词云图片可以看出，不同小说词云的词汇分布和大小不同。代码中通过对小说文本的分词和词云生成过程，将高频词汇以较大字体展示。例如，如果某本小说中 “修炼”“法宝” 等词汇字体较大，说明在该小说中这些词汇出现频率较高，可能反映出这本小说的题材偏向玄幻修仙类，经常涉及修炼体系和法宝相关情节。

反映小说内容侧重：通过对比不同小说的词云，能发现它们在内容上的差异。代码中对每本小说单独生成词云，使得这种对比成为可能。比如一本小说词云里 “江湖”“侠义” 等词汇突出，而另一本是 “星际”“科技” 等，说明前者可能是武侠题材，后者是科幻题材，帮助读者快速了解小说的大致内容走向和风格特点。

**Task 3:**

**尝试用大数据分析技术分析出杭州、宁波、温州哪个城市更宜居？**

1. **本任务要求运用大数据分析技术，通过收集、处理和分析多源数据，构建一个科学的评价体系，找出杭州、宁波和温州三个城市中哪个更宜居。需要完成从数据采集、清洗、分析到可视化的全流程工作，并最终形成分析报告；**
2. **需要用到的技术：数据采集（爬虫）、数据预处理、数据分析、数据可视化等；**
3. **需要分析的数据维度参考：**
   1. **经济维度：人均GDP、就业率、房价收入比、**消费水平指数**等；**
   2. **环境维度：空气质量指数（AQI）、水质状况、绿化覆盖率、年平均气温、极端天气频率等；**
   3. **基础设施维度：公共交通便利性（地铁里程、公交线路）、医疗资源(每千人床位数、三甲医院数量)、教育资源(中小学质量、高校数量)、商业设施密度等；**
   4. **社会文化维度：文化场馆数量(博物馆、图书馆、剧院)、休闲娱乐设施、城市安全指数(犯罪率)、人口密度等；**
   5. **生活便利维度：外卖/快递便利度、便利店/超市密度、社区服务设施、数字政务便利度等。**
4. **任务要求：**
   1. **数据收集：从多个不同来源收集数据，确保数据的全面性和代表性**
   2. **数据处理：清洗不一致、缺失的数据，进行必要的标准化处理，构建综合评价指标体系；**
   3. **分析建模：设计合理的权重分配方案，应用适当的分析模型，进行敏感性分析验证结果稳健性；**
   4. **可视化呈现：制作交互式可视化图表，创建城市宜居度地图，展示关键指标对比。**
5. **推荐数据来源网站：**
   1. **空气质量在线监测分析平台：<https://www.aqistudy.cn/>**
   2. **浙江统计局：<http://tjj.zj.gov.cn>**
   3. **国家地表水质自动监测实时数据发布系统：<https://szzdjc.cnemc.cn:8070/GJZ/Business/Publish/Main.html>**
   4. **浙江省住建厅：<https://jst.zj.gov.cn/>**
   5. **国家气象科学数据中心：<http://data.cma.cn>**

**报告中需要详细介绍设计思路和步骤（截图），并且用数据可视化方式展示分析和展示最后的结论。将所有爬取的数据、数据分析代码、可视化图表文件一起打包为一个zip文件，文件名：小组名-Task3.zip，与报告一起上传至Moodle。**

|  |  |  |
| --- | --- | --- |
| 组员姓名 | 分工 | 工作量比例 |
| 张三 | Task1 代码编写,数据解析 | 98% |
| 李四 | 跑腿买水 | 2% |
|  |  |  |