**Course Project Report**

**Task 1:**

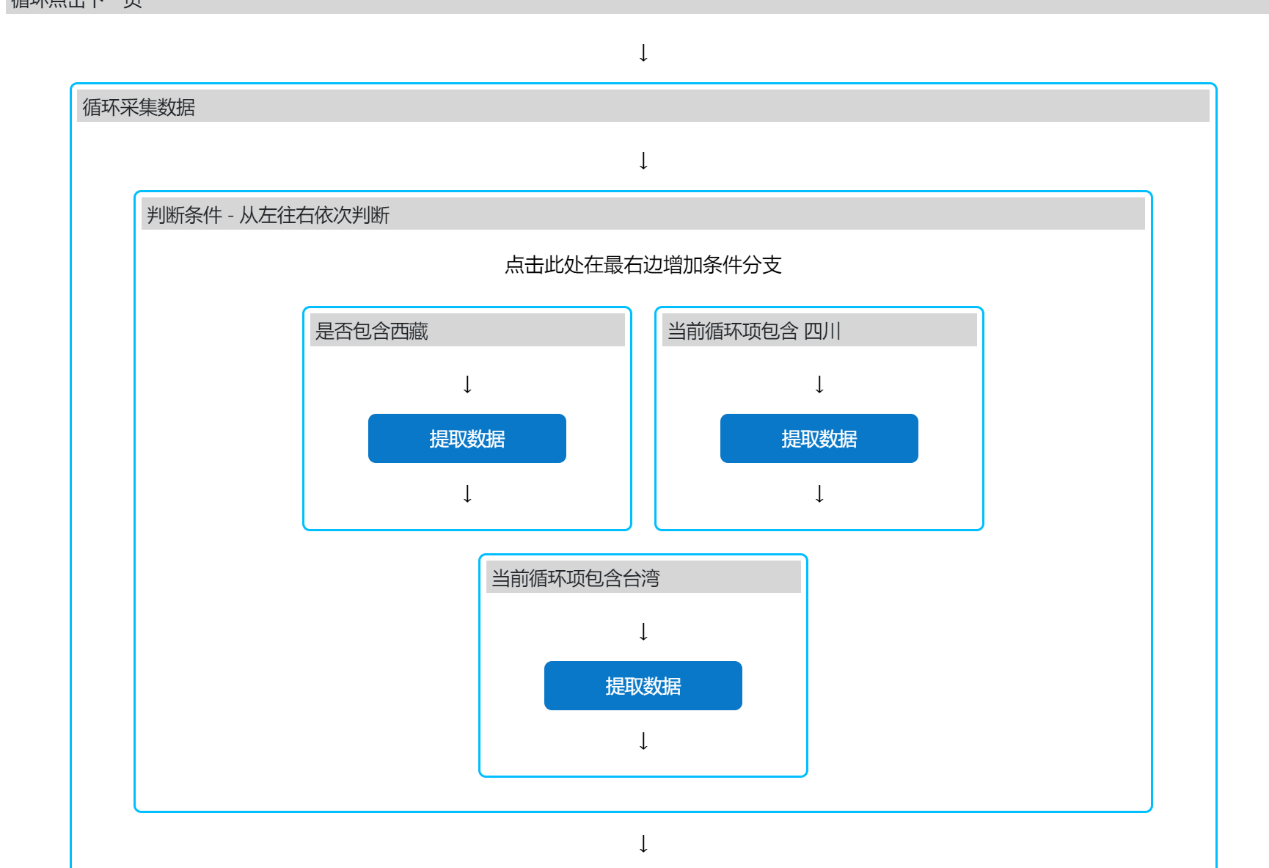
**爬取中国地震台网历史数据：**

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）。**

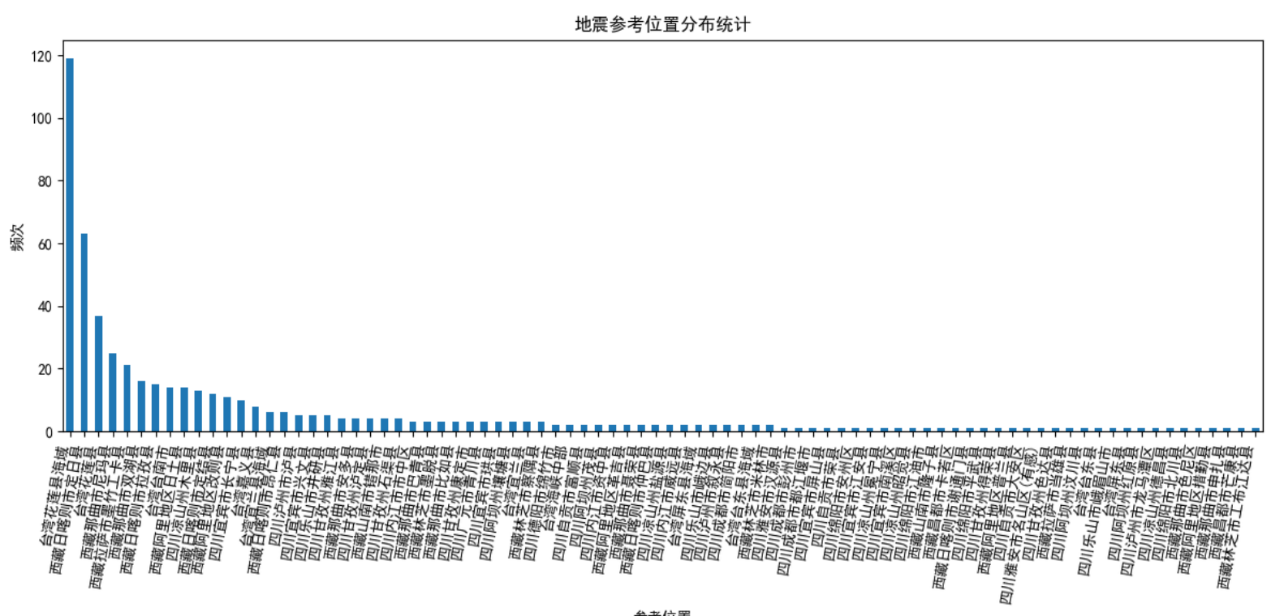
**报告中需要详细介绍设计思路和步骤（截图），并且展示可视化信息并进行分析。将爬取保存的数据文件命名为：小组名-Task1.csv（如有多个文件则分别命名后再压缩成一个压缩包，命名为小组名-Task1.zip） ，与报告一起上传至Moodle。**



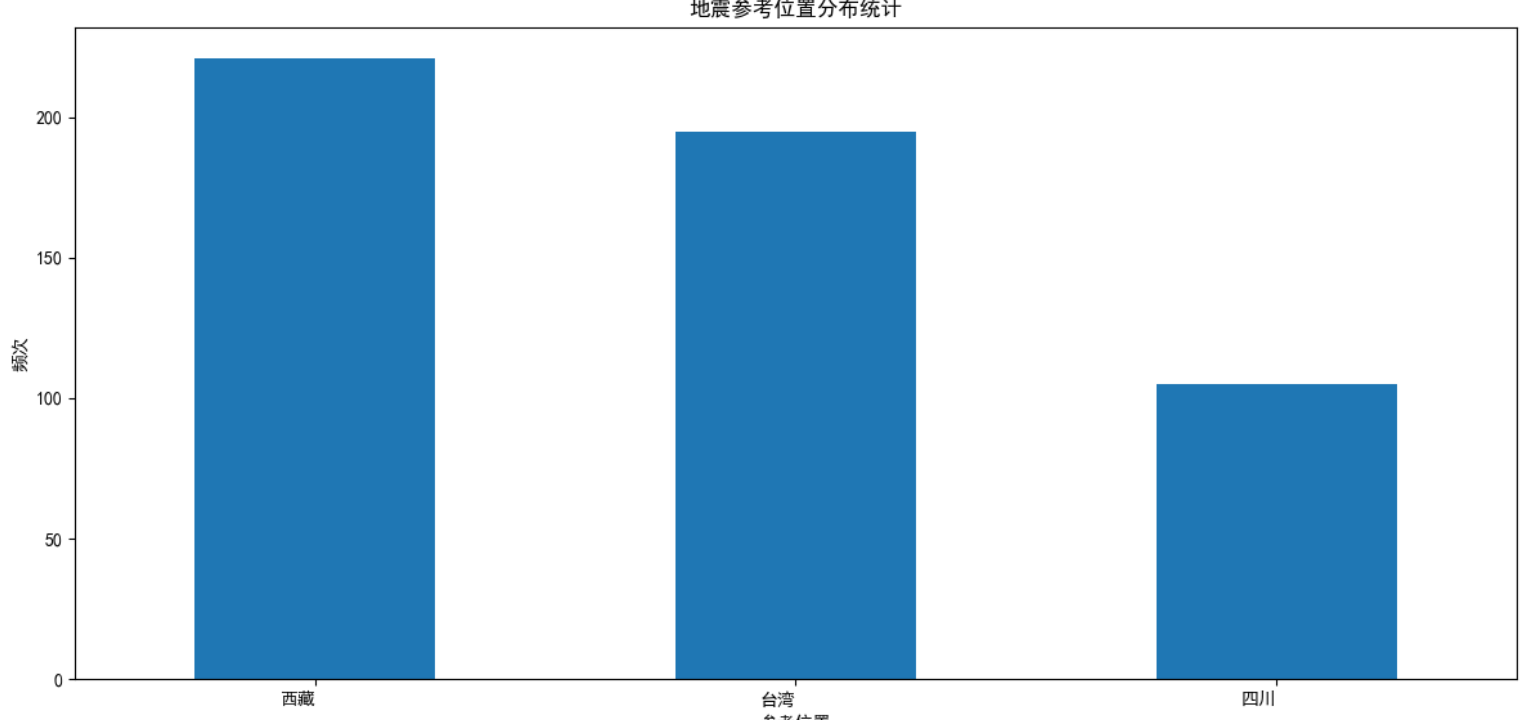


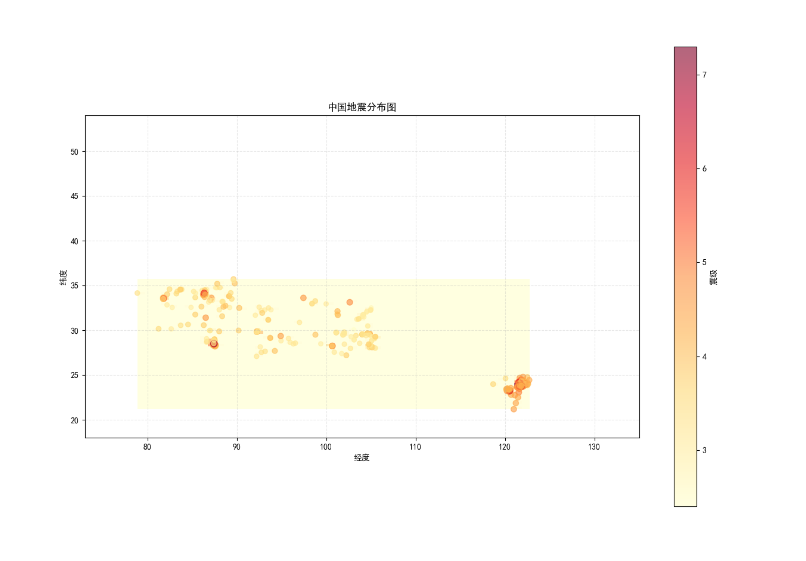


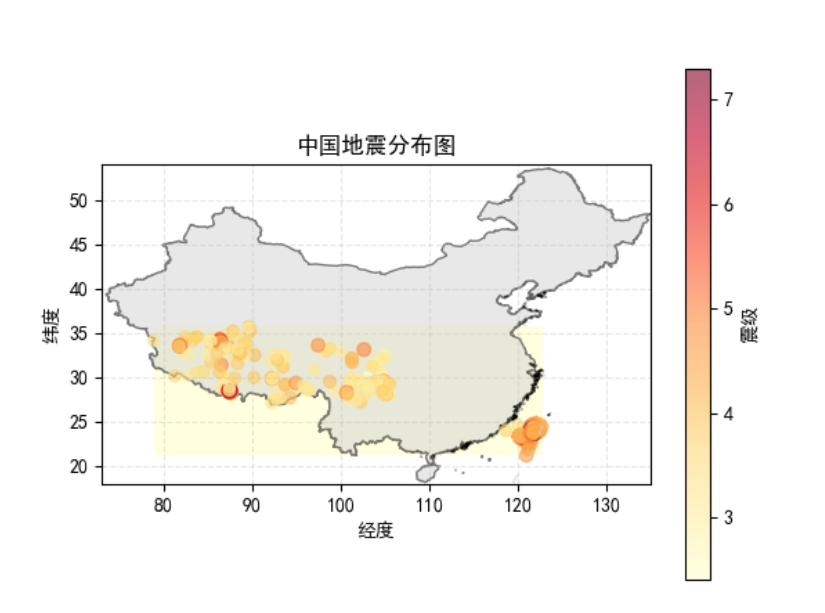


import pandas as pd  
import matplotlib.pyplot as plt  
  
# 读取数据  
df = pd.read\_csv("./中国地震台数据.csv")  
  
df = df.iloc[:,0:6]  
  
df.drop\_duplicates(inplace=True)  
# df.info()  
# print( df.head() )  
  
  
location\_counts = df['参考位置'].value\_counts()  
  
plt.figure(figsize=(12, 6))  
location\_counts.plot(kind='bar')  
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()



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()  
  


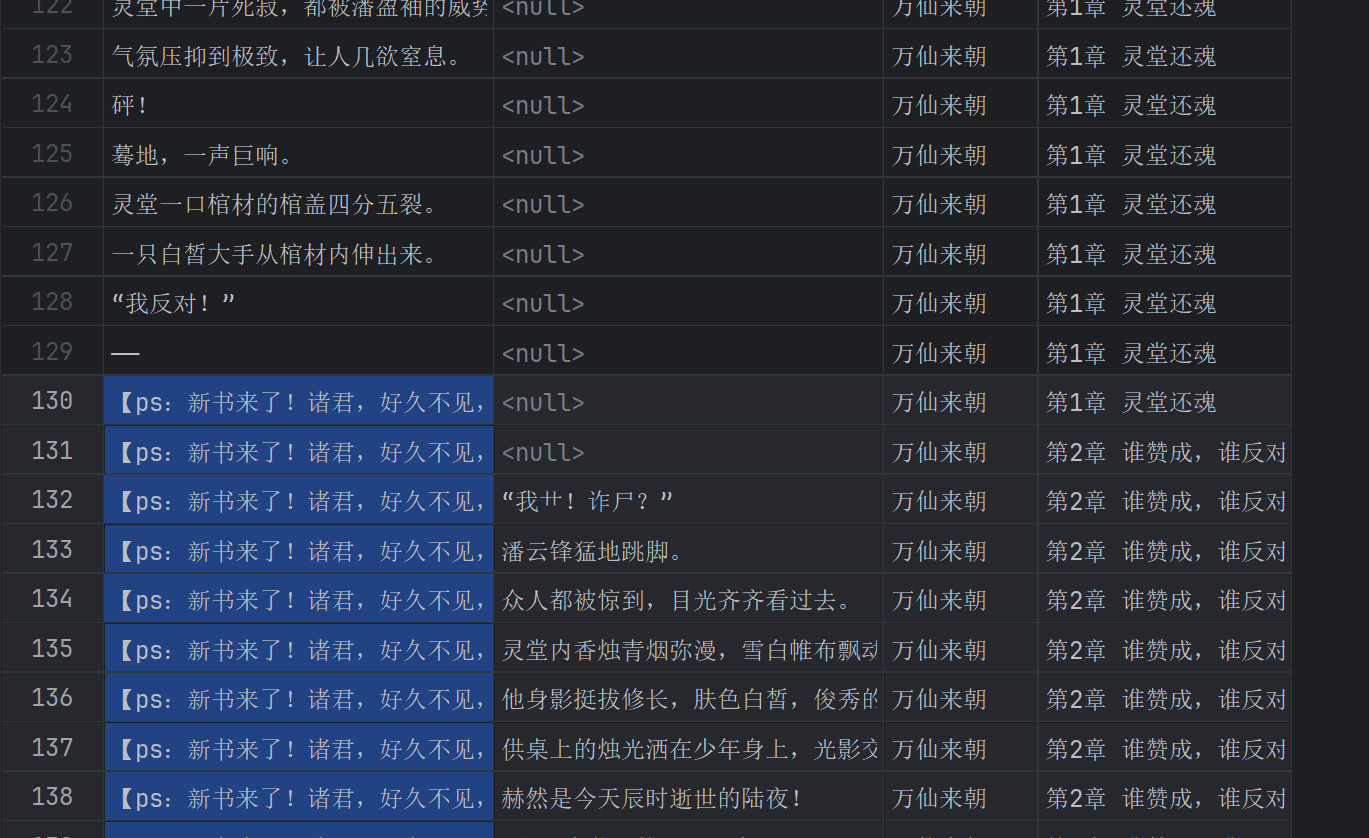
**Task 2:**

**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。**

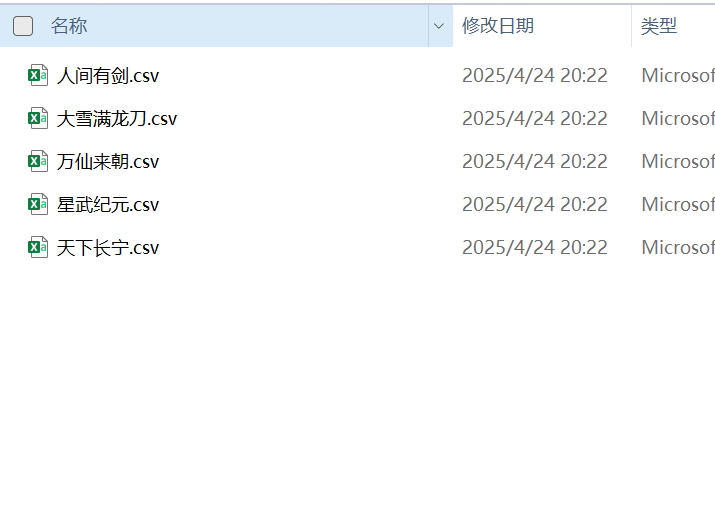
**报告中需要详细介绍设计思路和步骤（截图），并且展示5本小说的词云。将数据处理后的5本小说和词云图片文件一起打包为一个zip文件，文件名：小组名-Task2.zip，与报告一起上传至Moodle。**

**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()



**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/**](https://www.aqistudy.cn/)
   2. **浙江统计局：**[**http://tjj.zj.gov.cn**](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**](https://szzdjc.cnemc.cn:8070/GJZ/Business/Publish/Main.html)
   4. **浙江省住建厅：**[**https://jst.zj.gov.cn/**](https://jst.zj.gov.cn/)
   5. **国家气象科学数据中心：**[**http://data.cma.cn**](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% |
|  |  |  |