## 作业1 数据收集与预处理

#### 一、引言​

文本数据收集是展开文本信息挖掘的首要任务。该类信息内容覆盖科研、生活、工作、娱乐等人类活动的各个方面，形式包含新闻、博客、论坛、微博、对话设计、学术期刊、商业单证等，可以满足文本挖掘等技术建模研究的需要。本次作业旨在通过收集个人文本数据，并对其进行预处理，从而掌握数据收集与预处理的基本流程和方法，同时探索数据背后隐藏的信息与规律。​

#### 二、数据收集​

##### 2.1 数据集内容与创意​

本次收集的个人文本数据集为本人过去的12篇旅行日记，记录了在不同城市、不同场景下的所见所闻、内心感受以及对当地文化、美食的体验。选择旅行日记作为数据集，是因为其内容丰富多样，涵盖了风景描写、人物故事、情感抒发等多种体裁，能够为后续的数据预处理和分析提供充足的素材，且具有独特的个人色彩和趣味性。​

##### 2.2 数据收集方式​

本次数据采用手动收集的方式，将分散记录在纸质笔记本和手机备忘录以及发布在社交平台中的旅行日记整理汇总到raw\_travel\_diaries.txt一个文本文件中。整个数据集包含 12 篇旅行日记，总字数约10000 字。​经过数据预处理后得到的数据集保存在raw\_travel\_diaries.txt文件

#### 三、数据预处理​

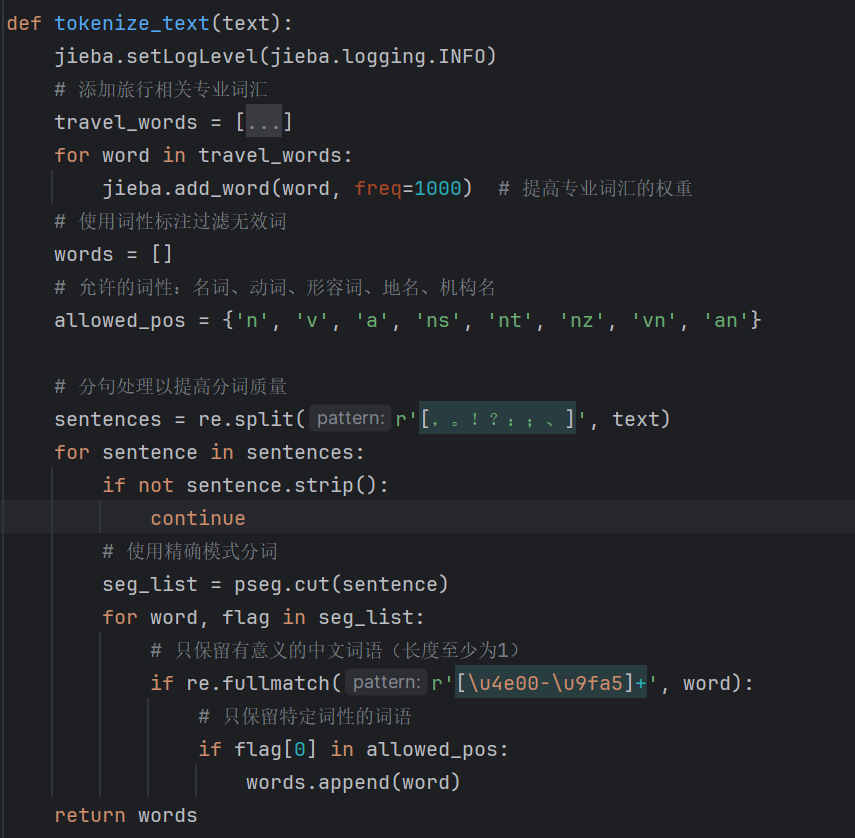
##### 3.1 数据清洗​

使用 Python 的re库对文本数据进行清洗，去除其中的特殊字符、标点符号以及空白字符，使得文本更加规范。具体代码如下：



##### 3.2 分词处理​

使用jieba库对清洗后的文本进行分词，将连续的文本分割成一个个独立的词语，便于后续的分析。代码如下：​

​ 

##### 3.3 去除停用词​

为了提高分析的准确性，去除文本中如 “的”“了”“在” 等无实际意义的停用词。从网上下载中文停用词表chinese\_stopwords.txt，设置了三级停用词库：读取并去除数据集中的停用词。

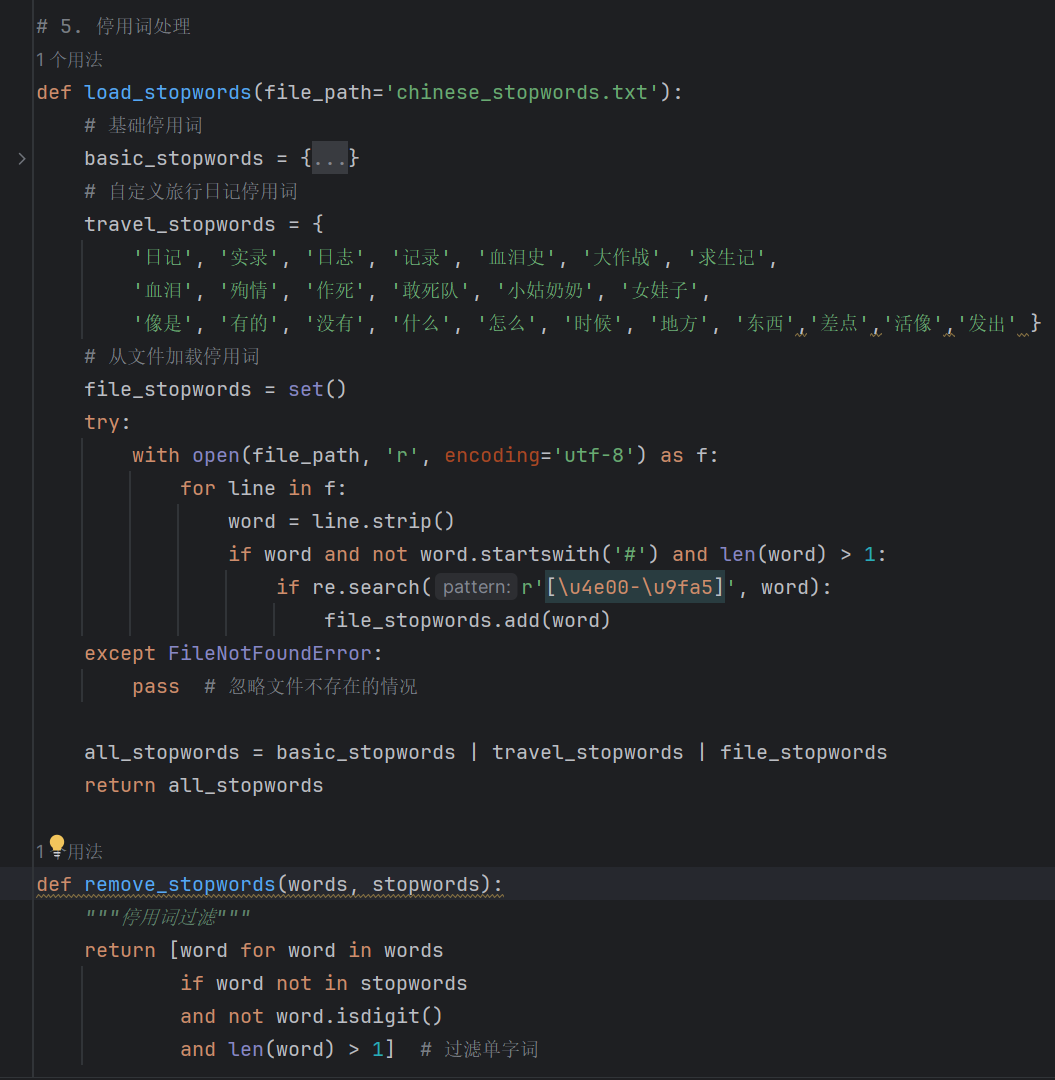
**三级停用词库**：

（1）基础停用词（的/了/是）

（2）旅行场景停用词（日记/记录/差点）

（3）文件停用词（从**chinese\_stopwords.txt**加载）

具体代码如下：



**数据处理后效果**：



#### 四、数据展示​

##### 4.1 词云展示​

使用wordcloud库生成词云图，直观展示数据集中出现频率较高的词语。同时，使用matplotlib库对词云图进行美化和展示。

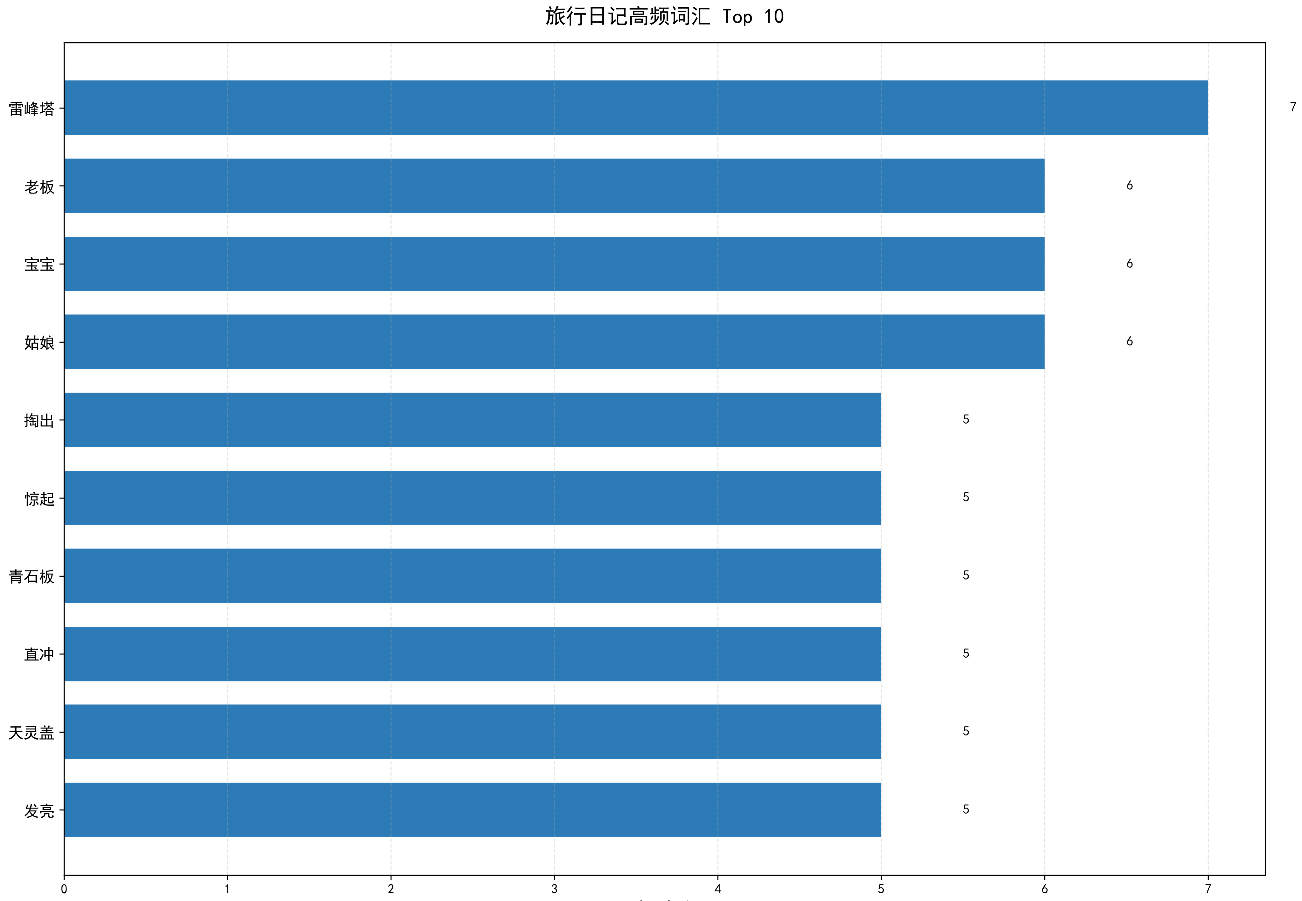


通过词云图可以清晰地看到，“雷峰塔”，“鼓浪屿”，“客栈” 等词语出现频率较高，这也符合旅行日记的主题内容。

##### 4.2 词语频率统计图表​

使用Counter对预处理后的词语列表进行计数，计算词语出现频率，按词频从高到低排序

可视化：用matplotlib生成横向柱状图，展示出现频率前10的词语。



#### 五、总结与展望​

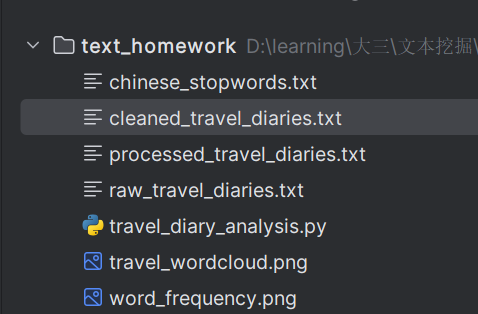
本次作业完成了个人文本数据集的收集与预处理工作，通过手动收集旅行日记数据，运用 Python 相关库进行清洗、分词、去除停用词等预处理操作，并通过词云图和柱状图对数据进行了展示分析。完成从数据收集→清洗→分析→可视化的完整流程。通过此次实践，加深了对数据收集与预处理流程的理解，掌握了相关技术的应用。​

在未来的学习和研究中，可以进一步探索更多的数据收集方式，丰富数据集的来源和类型。同时，尝试运用更复杂的文本分析技术，如情感分析、主题模型等，深入挖掘文本数据背后的潜在信息，为文本信息挖掘领域的研究提供更多有价值的参考。​

#### 参考文献​

1. Scapy 官方网站. <https://scapy.net/>, Accessed 2021-03-15​
2. jieba 分词库官方文档. <https://github.com/fxsjy/jieba>​
3. wordcloud 库官方文档. <https://amueller.github.io/word_cloud/index.html>​
4. matplotlib 库官方文档. <https://matplotlib.org/>​

**代码清单：**



原始数据文件raw\_travel\_diaries.txt

清洗后的数据文件 cleaned\_travel\_diaries

保存处理后的文本（processed\_travel\_diaries.txt）

停用词表chinese\_stopwords.txt

Python数据预处理运行代码travel\_diary\_analysis.py

生成词云图（travel\_wordcloud.png）

生成词频统计图（word\_frequency.png）