**作业2:词云**

姓名：白植权 学号：221050020

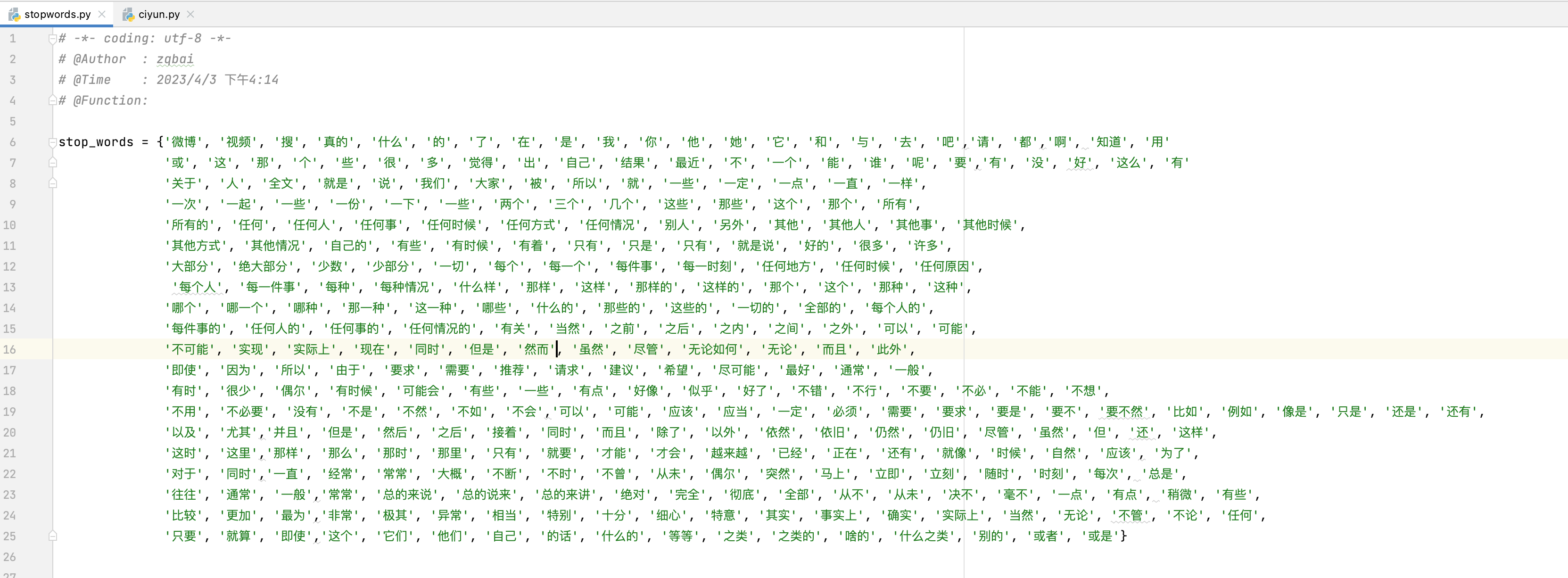
1.作业简介

本作业读取作业1中网络爬虫程序爬取并存储在CSV文件中的数据，对这些数据进行分词处理并生成词云图 。其中词云图（Word Cloud）是一种常见的可视化工具，用于显示文本数据。在词云图中，每个单词的大小表示其在文本中出现的频率。出现频率较高的词语以较大的字体显示，反之则以较小的字体显示，这使得用户可以一眼看出文本中的主要主题和关键词。

2.Python代码

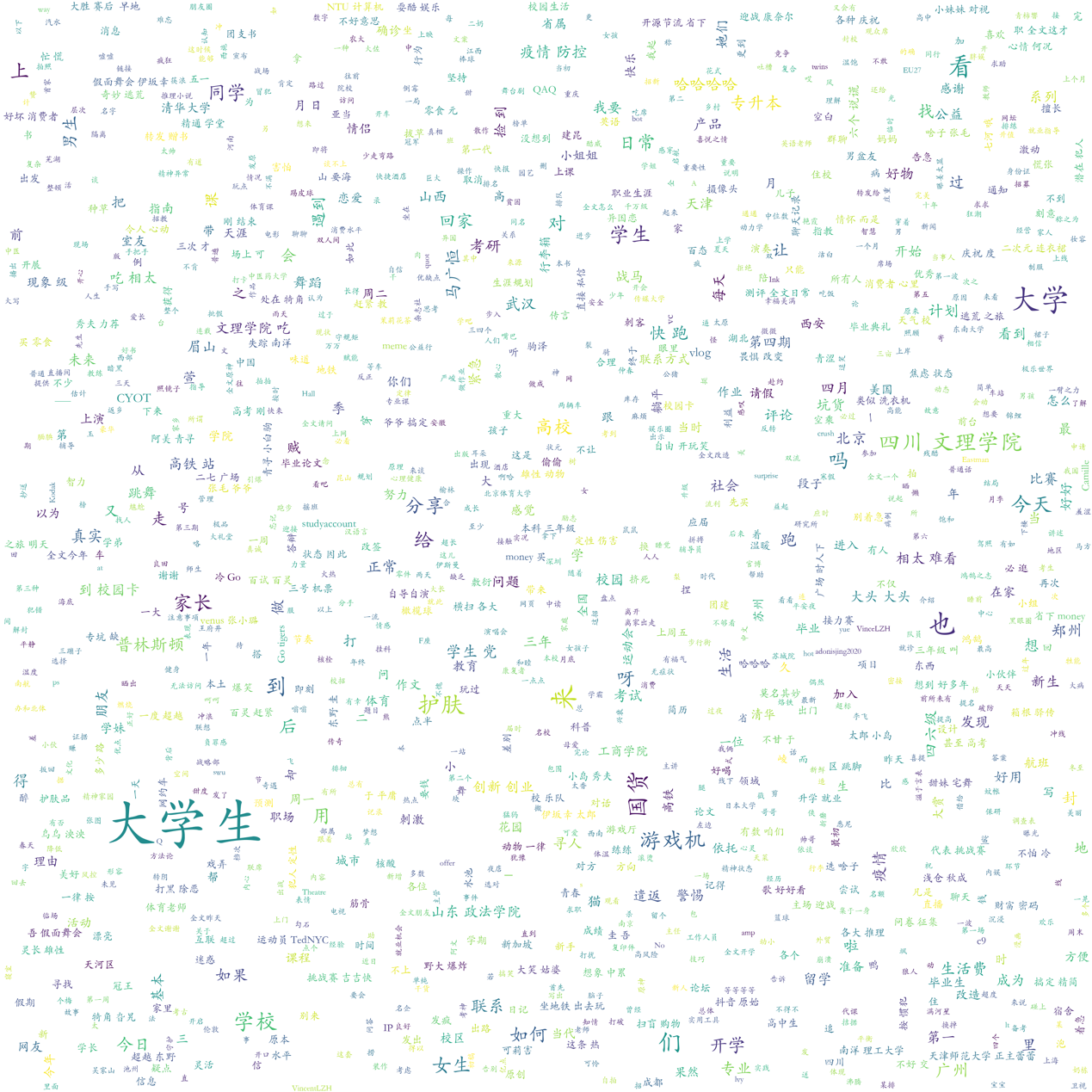
|  |
| --- |
| import pandas as pd import jieba from wordcloud import WordCloud from stopwords import stop\_words import os def chinese\_jieba(text):  wordlist\_jieba = jieba.cut(text)  space\_wordlist = " ".join(wordlist\_jieba)  return space\_wordlist def ciyun():  topic\_list = [f for f in os.listdir('../work1/') if f.endswith('.csv')]  for topic in topic\_list:  df = pd.read\_csv('../work1/' + topic)  comment\_list = df['comment'].values.tolist()  text = ""  for jj in range(len(comment\_list)):  text = text + chinese\_jieba(comment\_list[jj])  print(text)  *# 调用包PIL中的open方法，读取图片文件，通过numpy中的array方法生成数组* wordcloud = WordCloud(width=3000, height=3000,  font\_path="font.ttf", *# 字体文件* background\_color="white", *# 设置背景颜色* max\_font\_size=150, *# 设置字体最大值* max\_words=2000, *# 设置最大显示的字数* stopwords=stop\_words, *# 设置停用词，停用词则不再词云图中表示* ).generate(text)  image = wordcloud.to\_image()  wordcloud.to\_file(topic.split('.')[0] + '-ciyun.png') *# 导出文件* image.show() if \_\_name\_\_ == "\_\_main\_\_":  ciyun() |

存放停用词文件stopwords.py

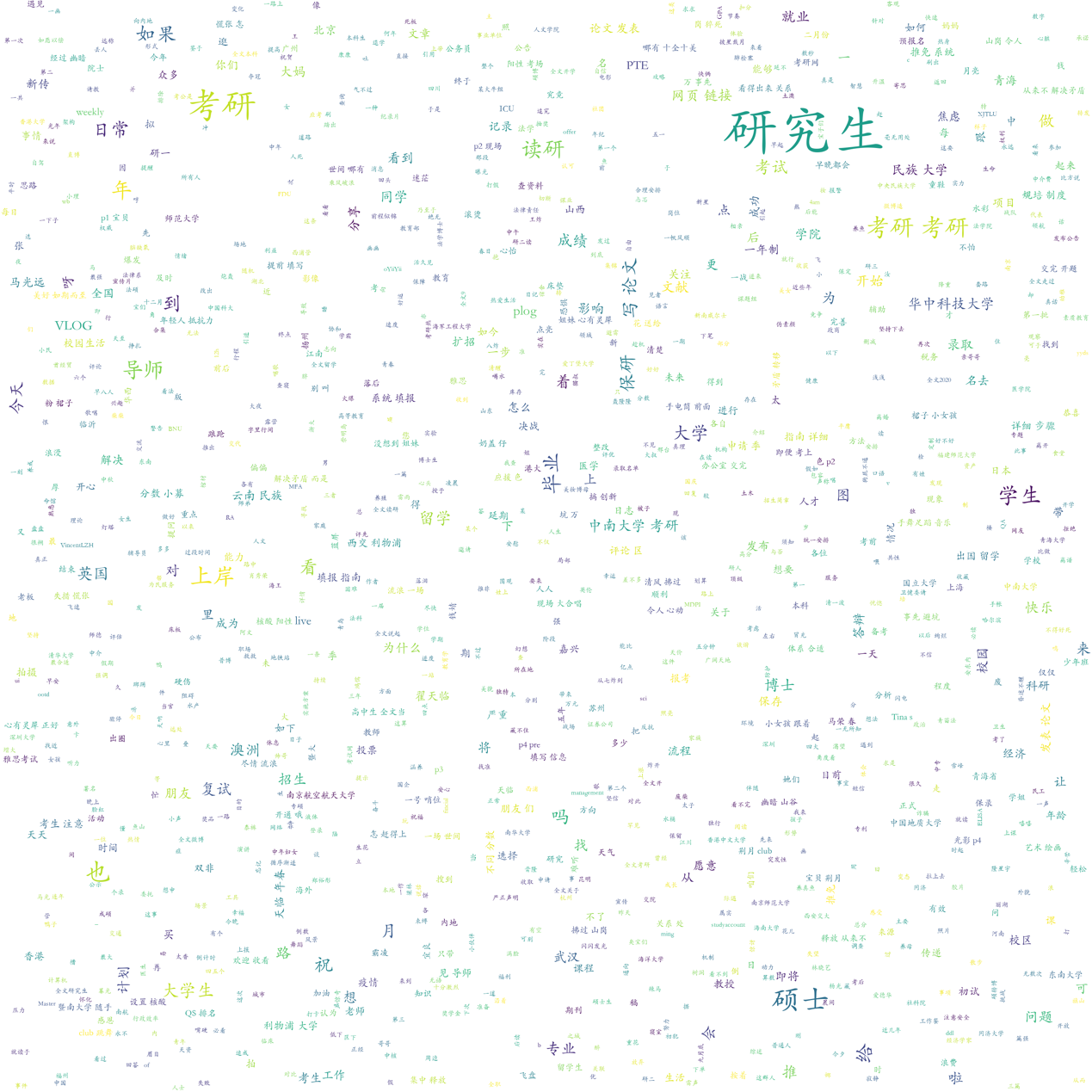


3.运行结果：

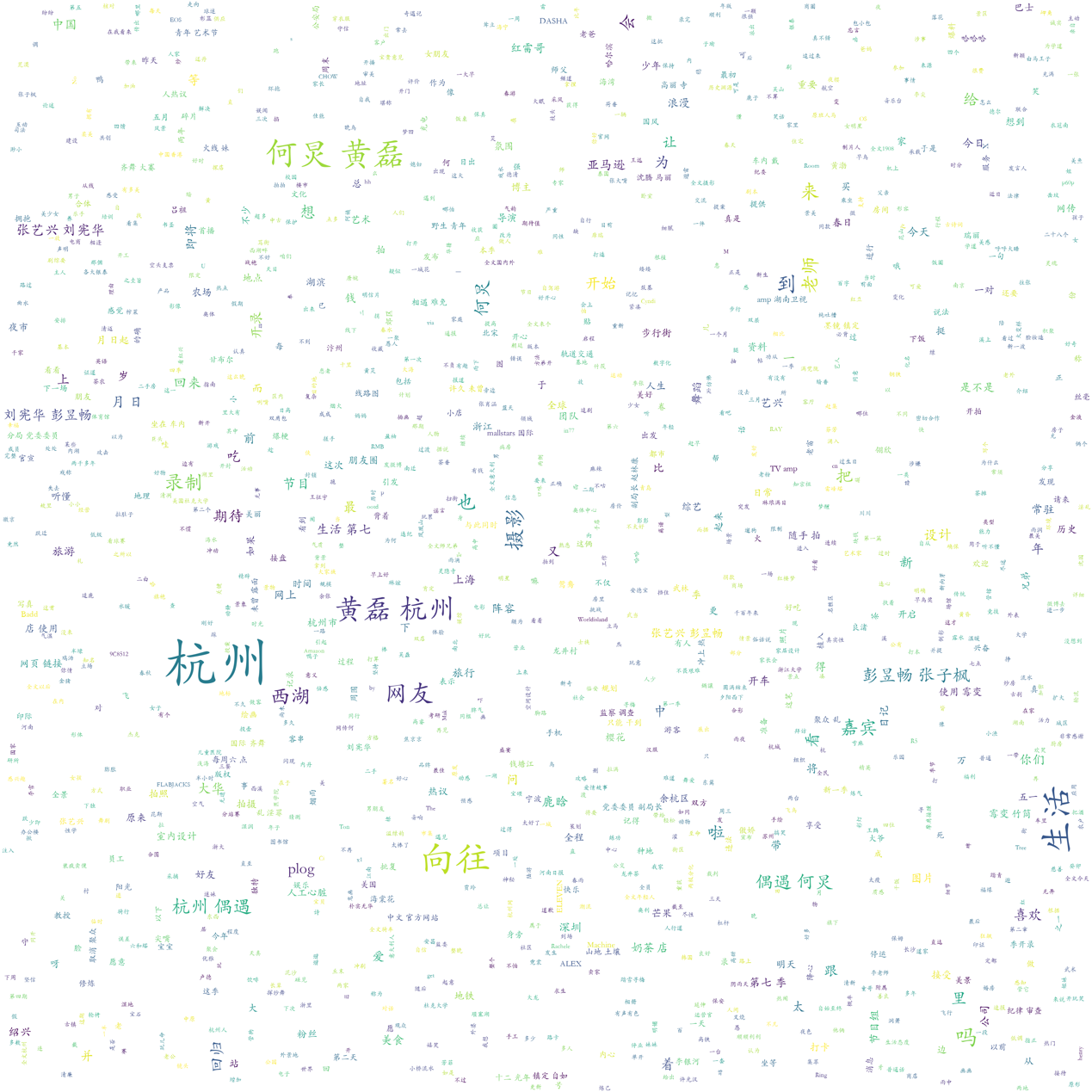
**（1）“大学生”主题词云图：**

****

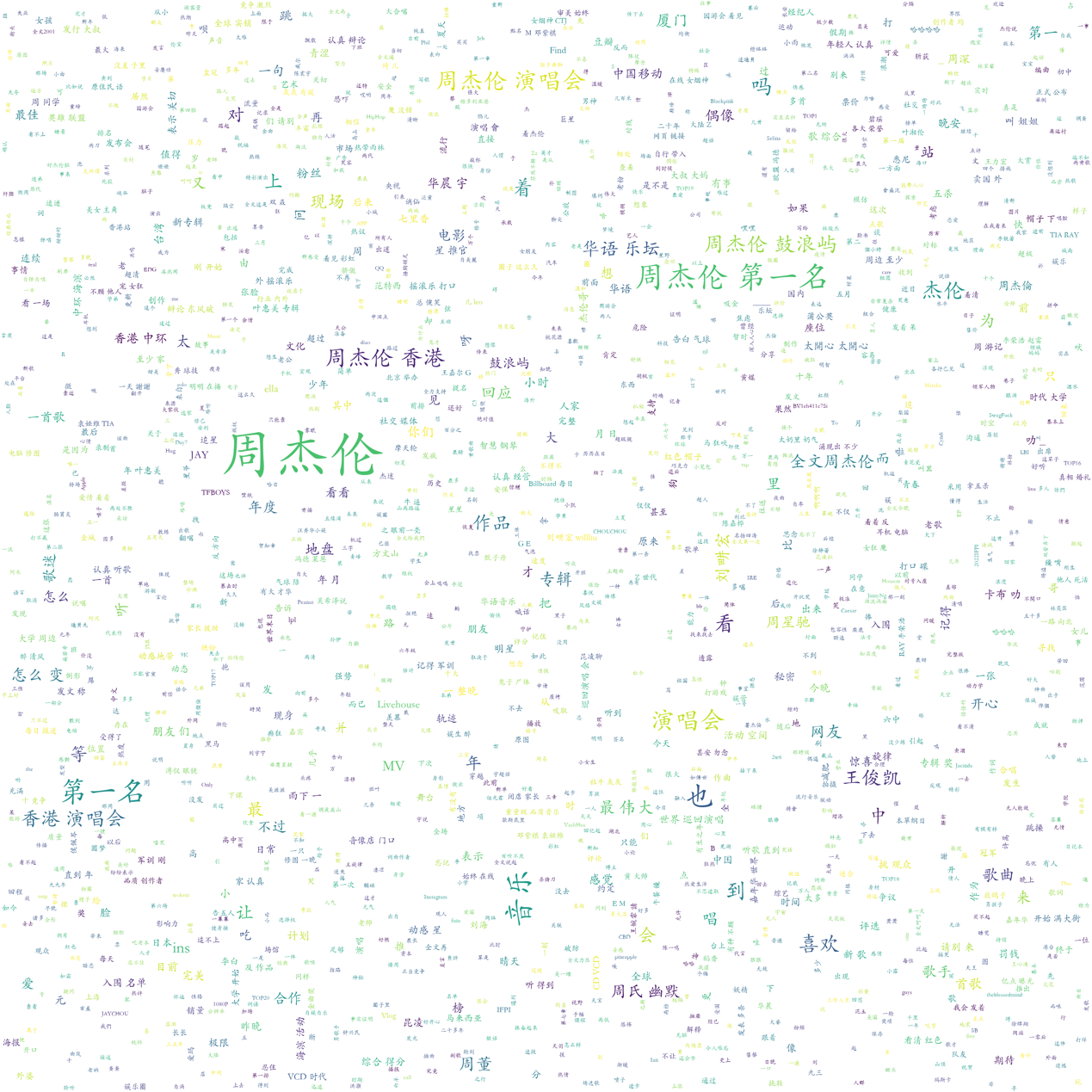
**（2）“研究生”主题词云图：**

****

**（3）“杭州”主题词云图：**

****

**（4）“周杰伦”主题词云图：**

****

4.代码简介

其中chinese\_jieba函数使用jieba库进行中文分词。接收一段文本作为输入，利用jieba的cut方法对文本进行分词，并将分词结果用空格连接成一个字符串返回。

ciyun函数则调用chinese\_jieba函数并执行以下步骤：

首先，遍历指定目录下的所有CSV文件。对于每个CSV文件，读取数据并提取评论列。再对提取的微博内容进行分词处理，然后将所有评论的分词结果连接成一个大的字符串。使用WordCloud库创建一个词云对象。词云对象的参数包括画布的宽度和高度、字体路径、背景颜色、最大字体大小、最大显示单词数以及停用词列表。这里的停用词来自'stopwords'模块的'stop\_words'变量，这些词在生成词云图时将被忽略（该模块中具体的stop\_words变量情况在2中有展示）。最后调用WordCloud对象的generate方法生成词云图，然后保存到文件并显示。