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

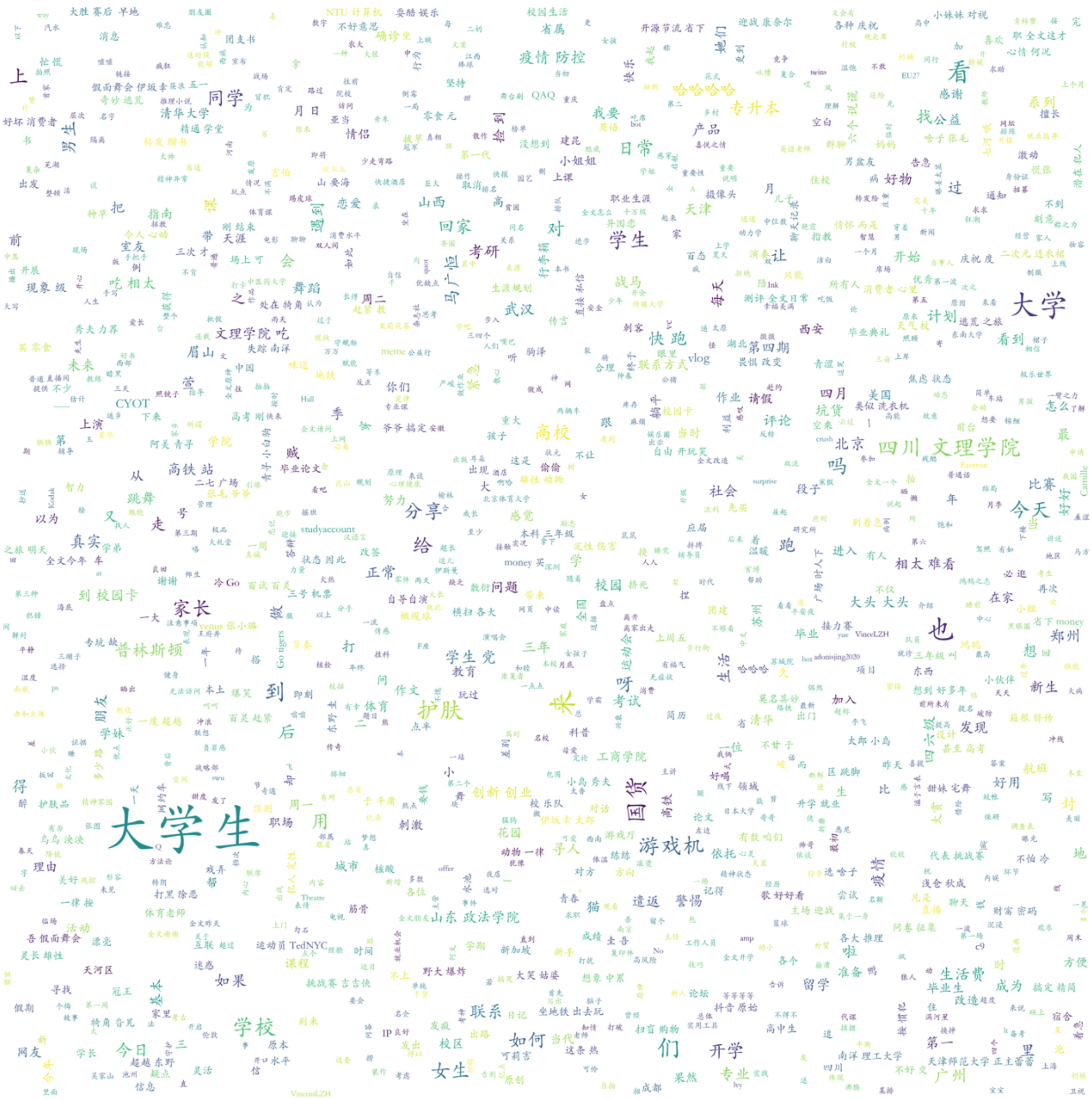
```
import pandas as pd
import jieba
from wordcloud import WordCloud
from stopwords import stop_words
import os
def chinese_jieba(text):
    wordlist_jieba = jieba.cut(text)
    space_wordlist = " ".join(wordlist_jieba)
    return space_wordlist
def ciyun():
    topic_list = [f for f in os.listdir('../work1/') 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

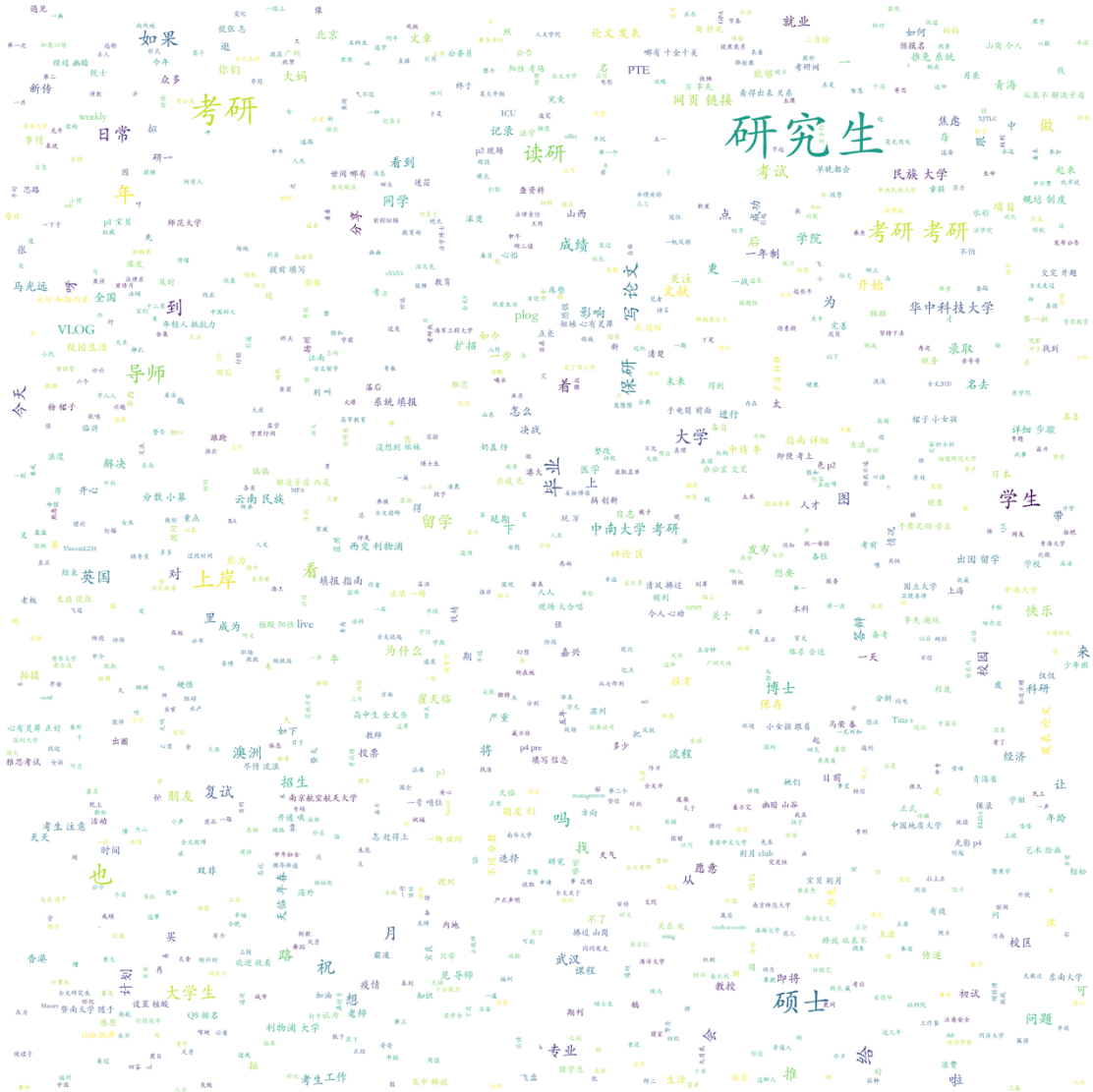
```python
# -*- coding: utf-8 -*-
# @Author  : zqbai
# @Time    : 2023/4/3 下午4:14
# @Function:

stop_words = {'微博', '视频', '搜', '真的', '什么', '的', '了', '在', '是', '我', '你', '他', '她', '它', '和', '与', '去', '吧','请', '都','啊', '知道', '用'
              '或', '这', '那', '个', '些', '很', '多', '觉得', '出', '自己', '结果', '最近', '不', '一个', '能', '谁', '呢', '要','有', '没', '好', '这么', '有'
              '关于', '人', '全文', '就是', '说', '我们', '大家', '被', '所以', '就', '一些', '一定', '一点', '一直', '一样',
              '一次', '一起', '一些', '一份', '一下', '一些', '两个', '三个', '几个', '这些', '那些', '这个', '那个', '所有',
              '所有的', '任何', '任何人', '任何事', '任何时候', '任何方式', '任何情况', '别人', '另外', '其他', '其他人', '其他事', '其他时候',
              '其他方式', '其他情况', '自己的', '有些', '有时候', '有着', '只有', '只是', '只有', '就是说', '好的', '很多', '许多',
              '大部分', '绝大部分', '少数', '少部分', '一切', '每个', '每一个', '每件事', '每一时刻', '任何地方', '任何时候', '任何原因',
               '每个人', '每一件事', '每种', '每种情况', '什么样', '那样', '这样', '那样的', '这样的', '那个', '这个', '那种', '这种',
              '哪个', '哪一个', '哪种', '那一种', '这一种', '哪些', '什么的', '那些的', '这些的', '一切的', '全部的', '每个人的',
              '每件事的', '任何人的', '任何事的', '任何情况的', '有关', '当然', '之前', '之后', '之内', '之间', '之外', '可以', '可能',
              '不可能', '实现', '实际上', '现在', '同时', '但是', '然而', '虽然', '尽管', '无论如何', '无论', '而且', '此外',
              '即使', '因为', '所以', '由于', '要求', '需要', '推荐', '请求', '建议', '希望', '尽可能', '最好', '通常', '一般',
              '有时', '很少', '偶尔', '有时候', '可能会', '有些', '一些', '有点', '好像', '似乎', '好了', '不错', '不行', '不要', '不必', '不能', '不想',
              '不用', '不必要', '没有', '不是', '不然', '不如', '不会','可以', '可能', '应该', '应当', '一定', '必须', '需要', '要求', '要是', '要不', '要不然', '比如', '例如', '像是', '只是', '还是', '还有',
              '以及', '尤其','并且', '但是', '然后', '之后', '接着', '同时', '而且', '除了', '以外', '依然', '依旧', '仍然', '仍旧', '尽管', '虽然', '但', '还', '这样',
              '这时', '这里','那样', '那么', '那时', '那里', '只有', '就要', '才能', '才会', '越来越', '已经', '正在', '还有', '就像', '时候', '自然', '应该', '为了',
              '对于', '同时','一直', '经常', '常常', '大概', '不断', '不时', '不曾', '从未', '偶尔', '突然', '马上', '立即', '立刻', '随时', '时刻', '每次', '总是',
              '往往', '通常', '一般','常常', '总的来说', '总的说来', '总的来讲', '绝对', '完全', '彻底', '全部', '从不', '从未', '决不', '毫不', '一点', '有点', '稍微', '有些',
              '比较', '更加', '最为','非常', '极其', '异常', '相当', '特别', '十分', '细心', '特意', '其实', '事实上', '确实', '实际上', '当然', '无论', '不管', '不论', '任何',
              '只要', '就算', '即使','这个', '它们', '他们', '自己', '的话', '什么的', '等等', '之类', '之类的', '啥的', '什么之类', '别的', '或者', '或是'}
```

3. 运行结果：

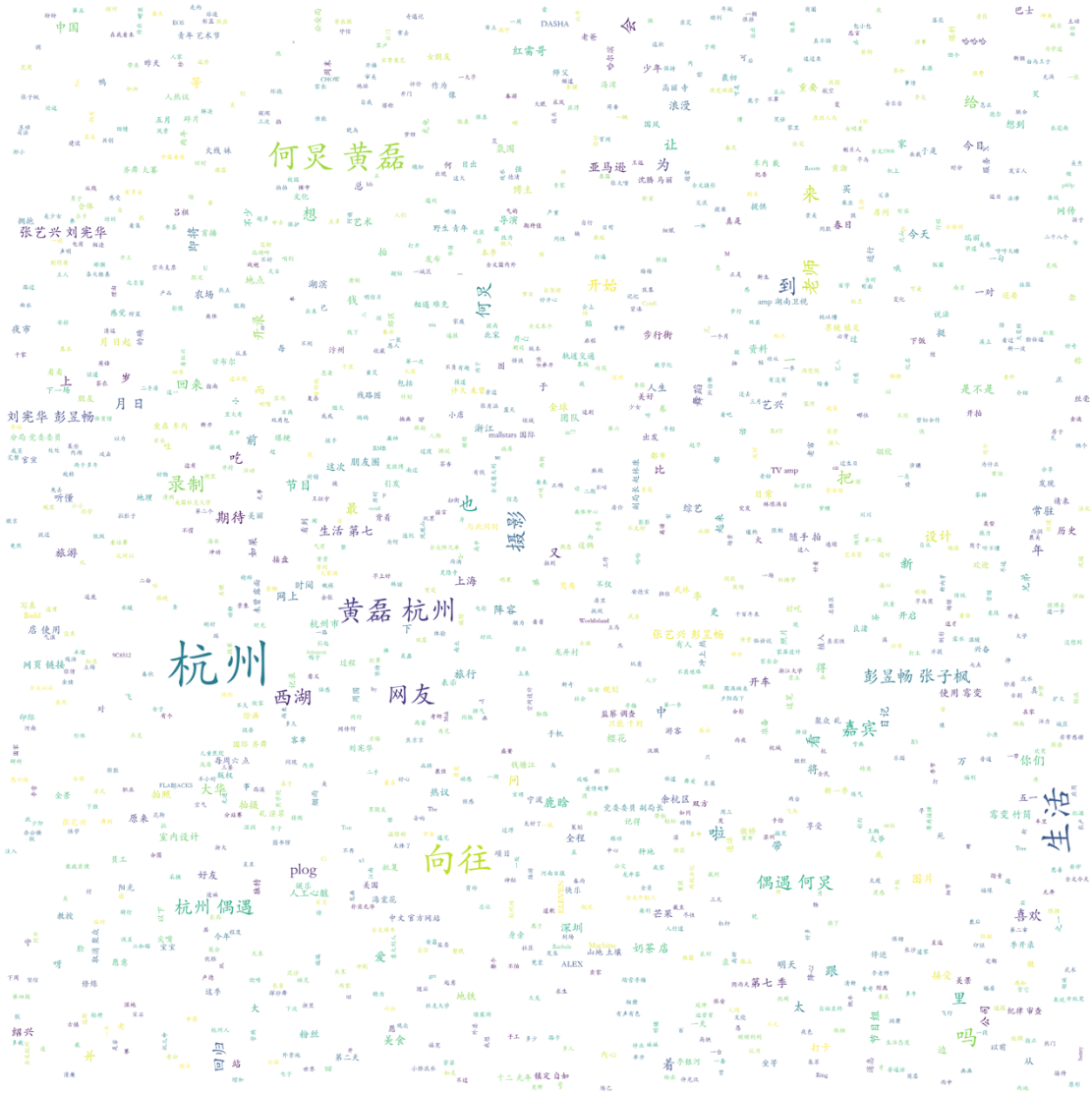
（1）“大学生”主题词云图：



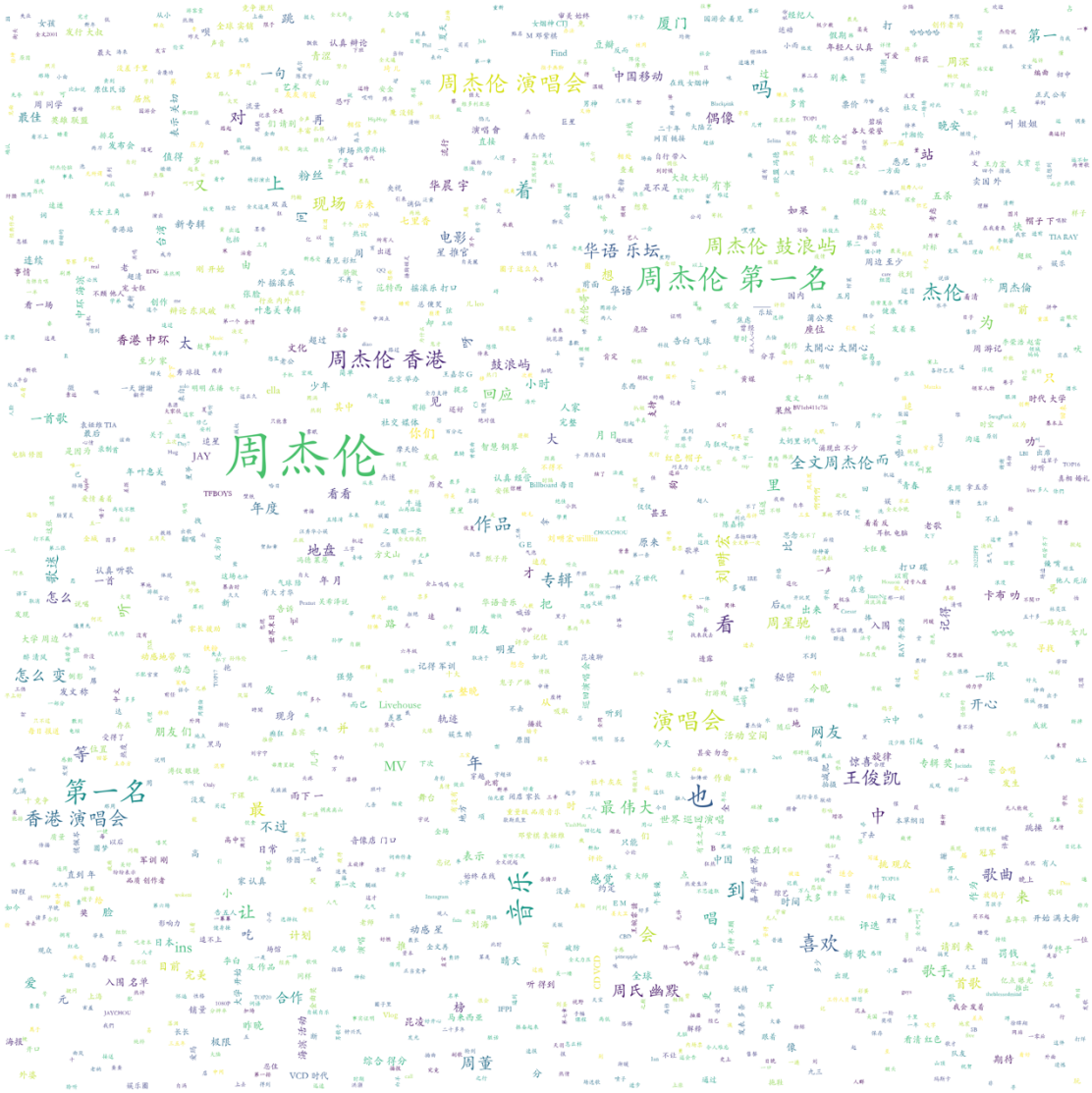
（2）“研究生”主题词云图：


研究生
考研
日常
读研
考试
学生
大学
导师
上岸
英国
留学
保研
写论文
硕士
博士
考研 考研
华中科技大学
民族 大学
学院
开始
为
点
后
一年制
更
关注
文献
论文
plog
影响
成绩
山西
同学
分享
看到
年
张
马光远
全国
到
VLOG
校园生活
今天
解决
分数 小幕
云南 民族
对
看
能力
里
成为
live
为什么
澳洲
招生
朋友
复试
考生 注意
时间
也
双非
买
香港
计划
大学生
设置 核磁
QS 排名
想
老师
利物浦 大学
考生工作
抱
祝
路
月
疫情
内地
只带
见 导师
武汉
课程
专业
留学生
飞盘
会
给
推
啦
生活
期刊
教授
即将
初试
问题
可
东南大学
校区
课
光影 pt
艺术 绘画
轻松
让
年龄
经济
科研
快乐
来
回
校园
带
学生
日本
中南大学
出国 留学
图
人才
指南 详细
申请 季
大学
进行
太
名去
录取
开题
规培 制度
项目
焦虑
做
中
青海
网页 链接
PTE
名
论文 发表
就业
如何
推免 系统
二月份
山岗 令人
如果
北京
文章
众多
你们
大妈
今年
新传
weekly
日常
事情
研一
年
同学
p1 室友
师范大学
VLOG
校园生活
导师
今天
全国
到
年轻人 抓紧力
解决
分数 小幕
云南 民族
英国
对
上岸
看
能力
里
成为
live
为什么
澳洲
招生
朋友
复试
考生 注意
时间
也
双非
买
香港
计划
大学生
设置 核磁
QS 排名
想
老师
利物浦 大学
考生工作
抱
祝
路
月
疫情
内地
只带
见 导师
武汉
课程
专业
留学生
飞盘
会
给
推
啦
生活
期刊
教授
即将
初试
问题
可
东南大学
校区
课
光影 pt
艺术 绘画
轻松
让
年龄
经济
科研
快乐
来
回
校园
带
学生
日本
中南大学
出国 留学
图
人才
指南 详细
申请 季
大学
进行
太
名去
录取
开题
规培 制度
项目
焦虑
做
中
青海
网页 链接
PTE
名
论文 发表
就业
如何
推免 系统
二月份
山岗 令人
如果
北京
文章
众多
你们
大妈
今年
新传
weekly
日常
事情
研一
年
同学
p1 室友
师范大学

（3）“杭州”主题词云图：


杭州
黄磊 杭州
何炅 黄磊
向往
生活
网友
西湖
摄影
偶遇 何炅
杭州 偶遇
彭昱畅 张子枫
嘉宾
录制
老师
到
来
开始
何炅
也
设计
plog
吗
张艺兴 刘宪华
刘宪华 彭昱畅
期待
节目
生活 第七
阵容
喜欢
会
让
为
给
旅游
旅行
上海
深圳
中国
红雷哥
亚马逊
大华
手机
热议
全程
第七季
张艺兴 彭昱畅
切茶店

（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。