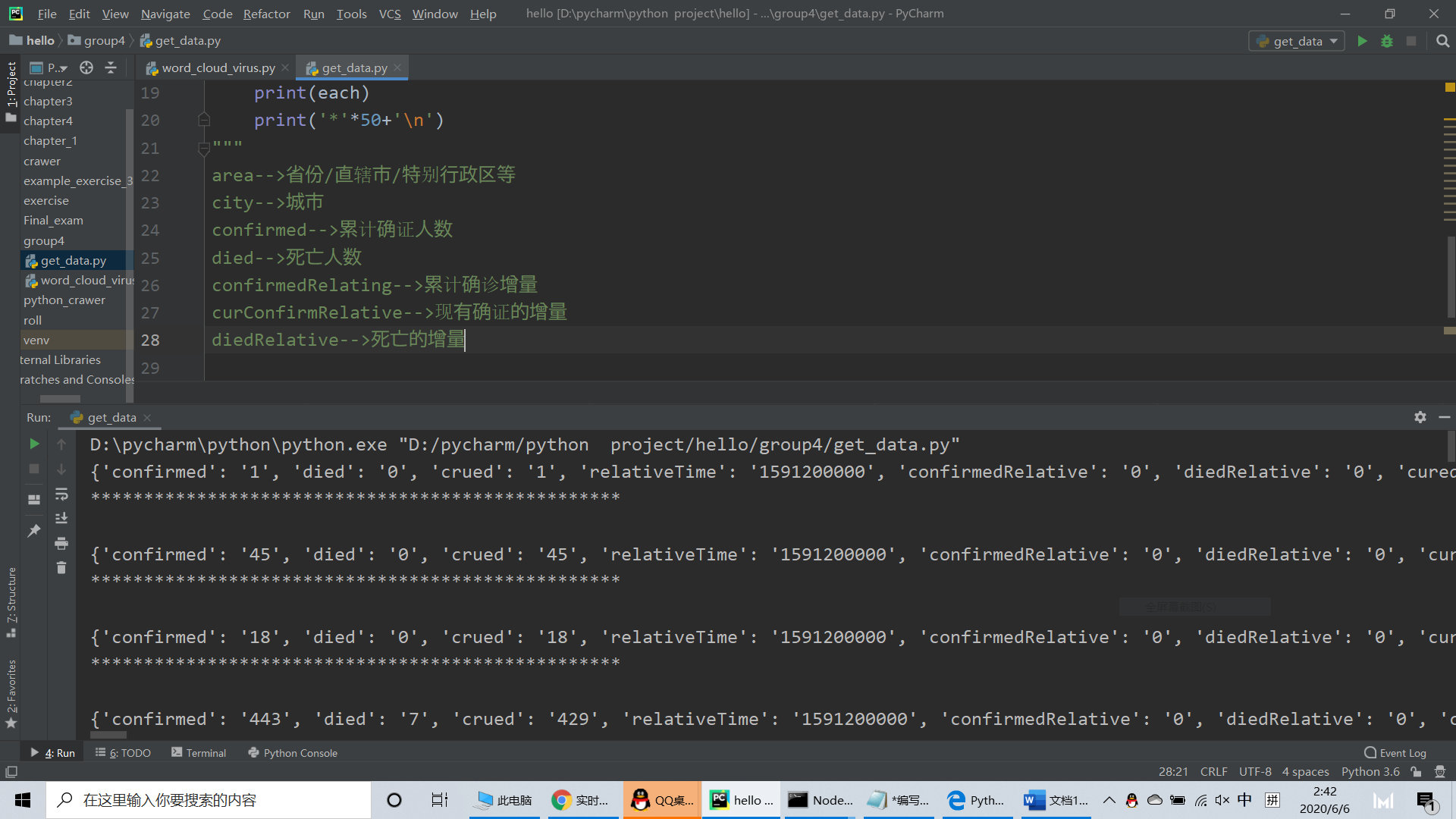
第一步: 首先打开网站地址：<https://voice.baidu.com/act/newpneumonia/newpneumonia>

查看源代码分析json文件

分析头部文件进行分析

通过导入json模块通过loads进行解码从而对爬取的数据进行一次遍历

第一次爬取的数据图如图所示:



通过对上图进行分析可以进行第一次爬虫总结然后对爬取的数据与实际的比对发现完全

正确

总结如下: """

area-->省份/直辖市/特别行政区等

city-->城市

confirmed-->累计确证人数

died-->死亡人数

confirmedRelating-->累计确诊增量

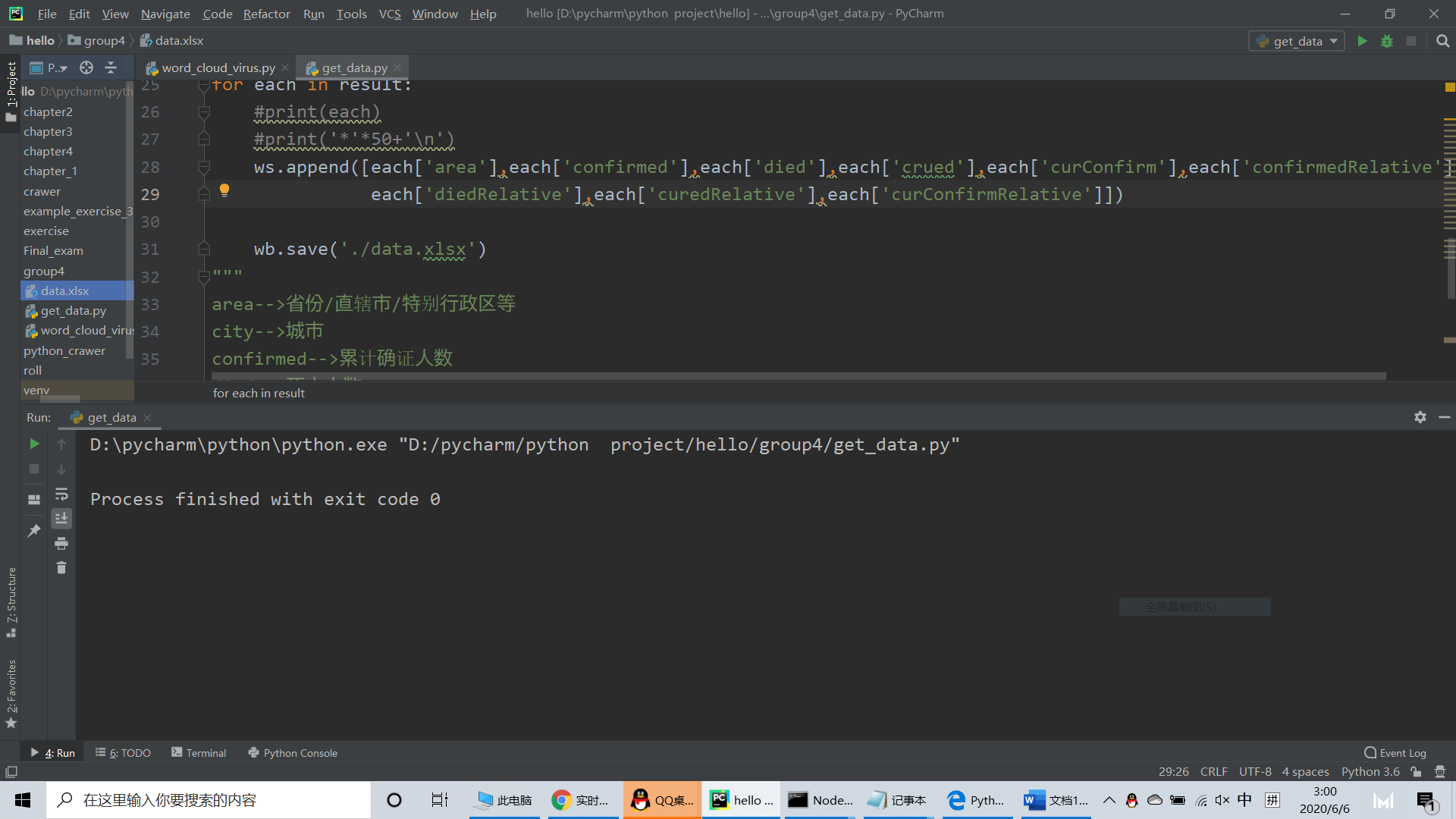
curConfirmRelative-->现有确证的增量

diedRelative-->死亡的增量

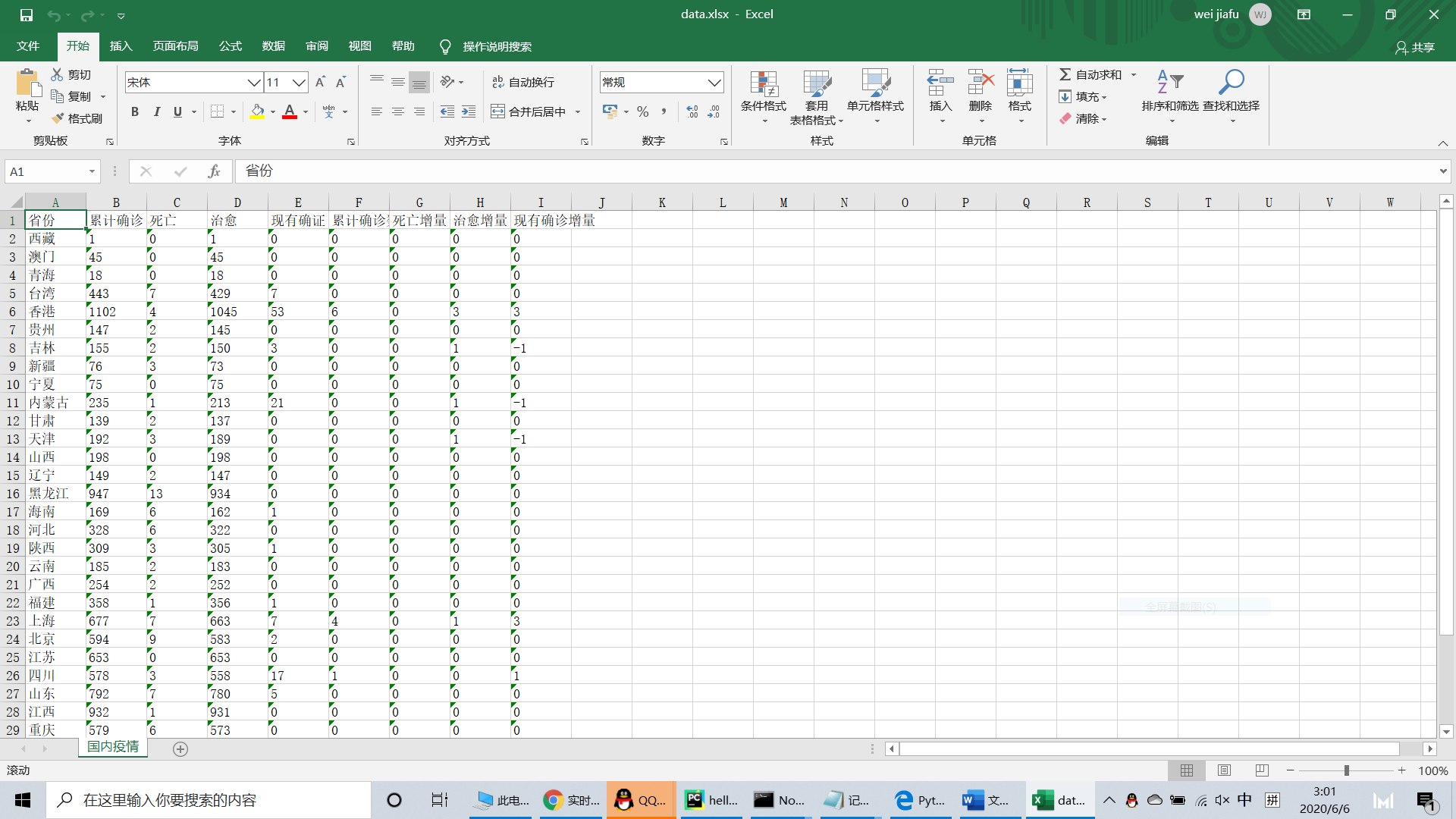
"""

D第二步：通过导入openpyxl模块将爬取到的数据写入excel表中

截图如下:



EXCEL表实现如图所示:此处为国内疫情数据的爬取



总结get\_data.py主要主要实现的是对疫情网站信息的爬取和利用openpyxl模块对爬取数据的保存

1.1通过对源码的分析找到了疫情信息的标签名称从而对进行遍历和解码从而爬取

核心代码为:

for each in result:

#print(each)

#print('\*'\*50+'\n')

#为了让代码简洁要对代码进行一次赋值

temp\_list=[each['area'],each['confirmed'],each['died'],each['crued'],each['curConfirm'],each['confirmedRelative'],

each['diedRelative'],each['curedRelative'],each['curConfirmRelative']]

ws.append(temp\_list)

#为了让爬取的数据的空值变为0进行一次遍历

for i in range(len(temp\_list)):

if temp\_list[i]=='':

temp\_list[i]='0'

ws.append(temp\_list)

1.2调用excel从而对爬取内容的核心代码为

Openpyxl模块的 #创建工作簿

wb=openpyxl.Workbook()

#创建工作表

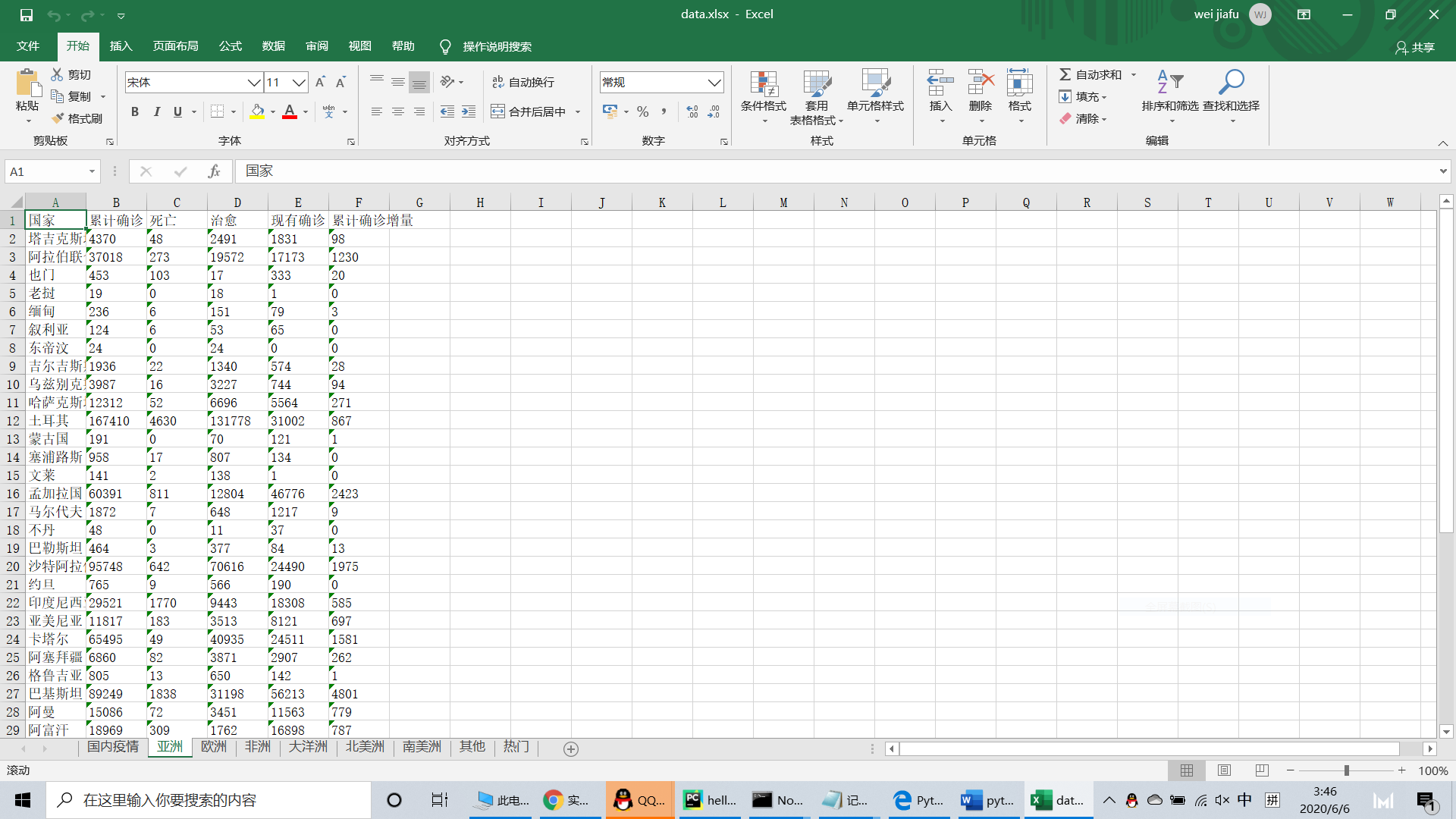
ws=wb.active

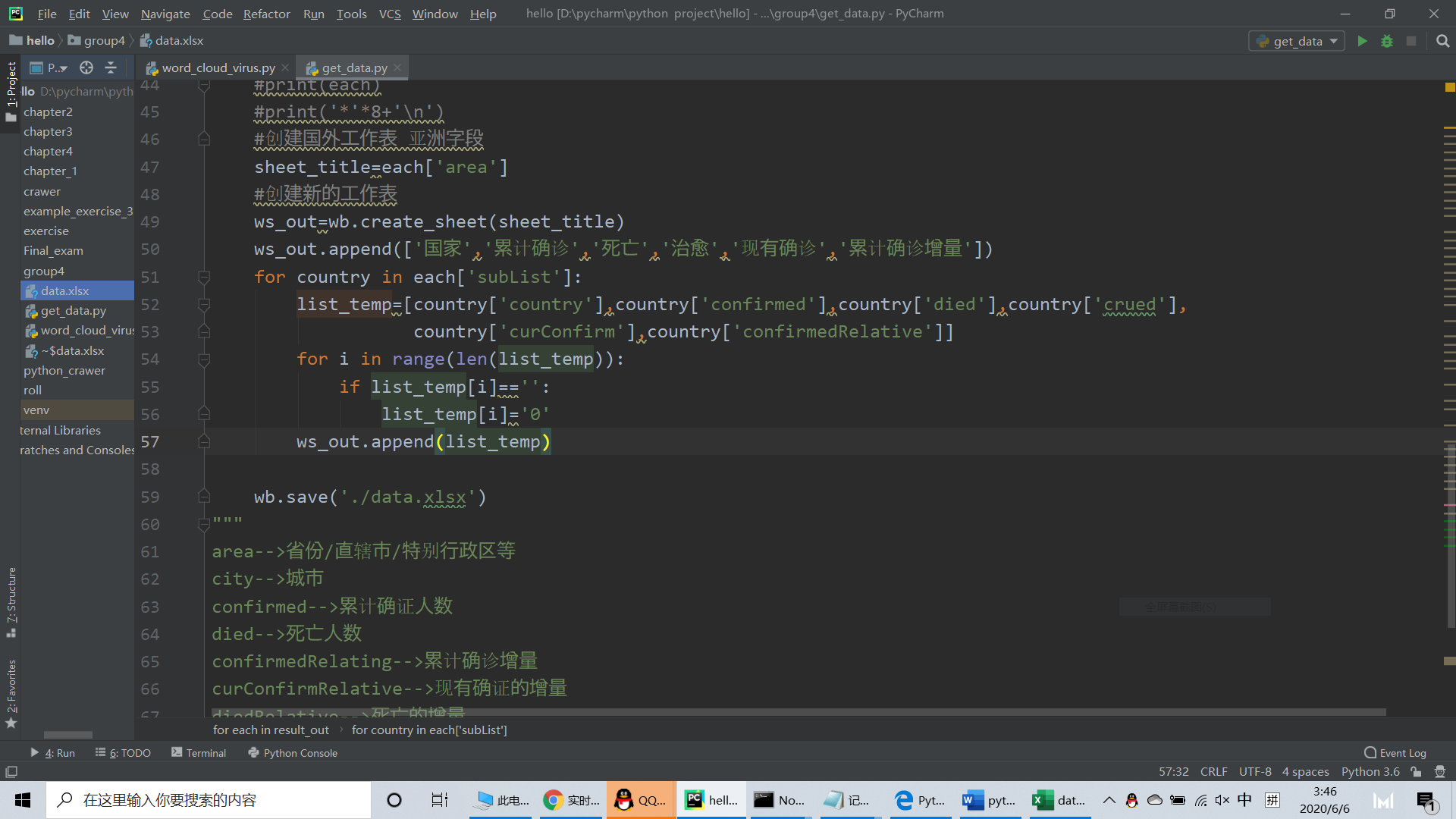
1.3还有国内疫情的标签为caseList 国外的为globalLIst

通过区分从而进行列表的遍历从而爬取到国外的数据到excel

表中

如图所示:





这样实现了基础对疫情网站的基本爬虫

二开始利用爬下来的数据进行写一个词云分析图

2.1新建词云文件为word\_cloud\_virus.py

首先要导入openpyxl模块进行读取excel表中的内容

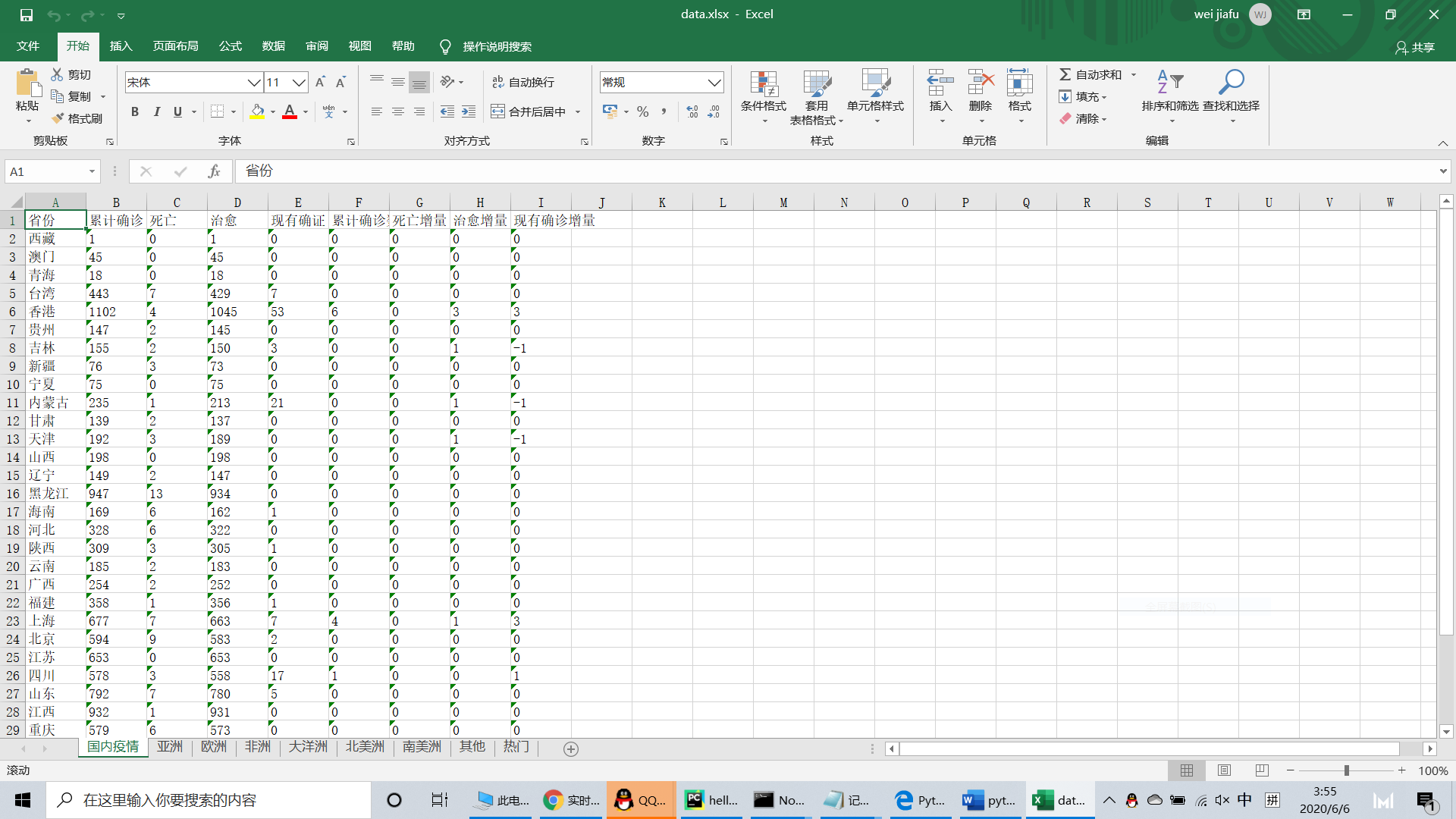
然后导入wordcloud(词云模块)进行词云分析图的绘制

2.2首先读取数据代码为：

#读取数据

wb=openpyxl.load\_workbook('data.xlsx')

然后打开excel表进行结构分析



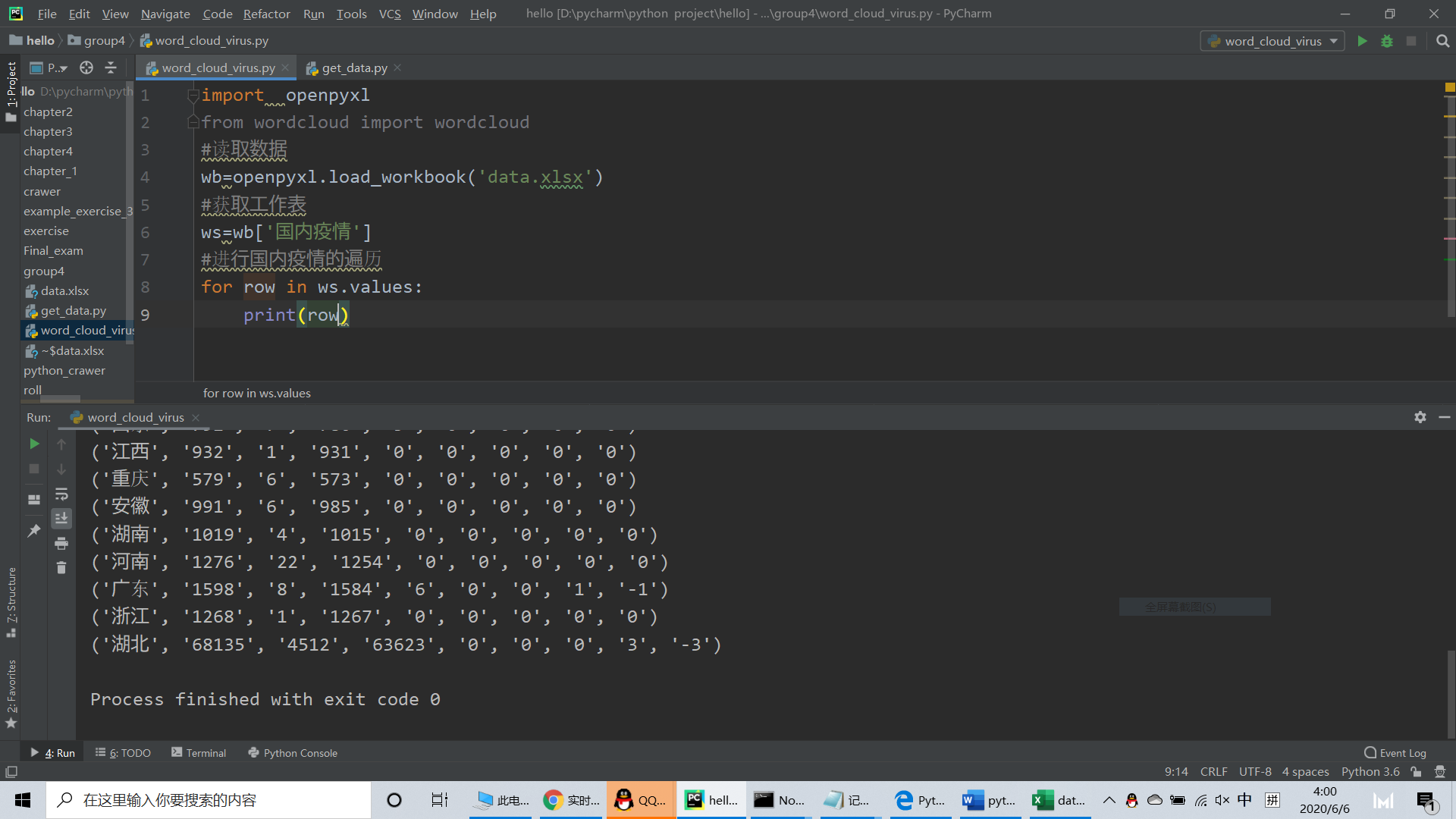
首先分析国内疫情首先我们小组的分析是根据疫情的严重程度从而进行绘制相应的疫情词云分析图

国内的第一列是省份的名称 第二列为累计确诊人数由此可得根据确诊人数的多少从而判断出各省的疫情严重程度

首先通过写代码

进行国内疫情遍历

截图如图所示:



然后第一列不需要通过判断删除

1.7通过字典分析代码如下：

import openpyxl

from wordcloud import WordCloud

#读取数据

wb=openpyxl.load\_workbook('data.xlsx')

#获取工作表

ws=wb['国内疫情']

#创建字典

frequency={}

for row in ws.values:

if row[0]=='省份':

pass

else:

#通过字典将省份那一列去除

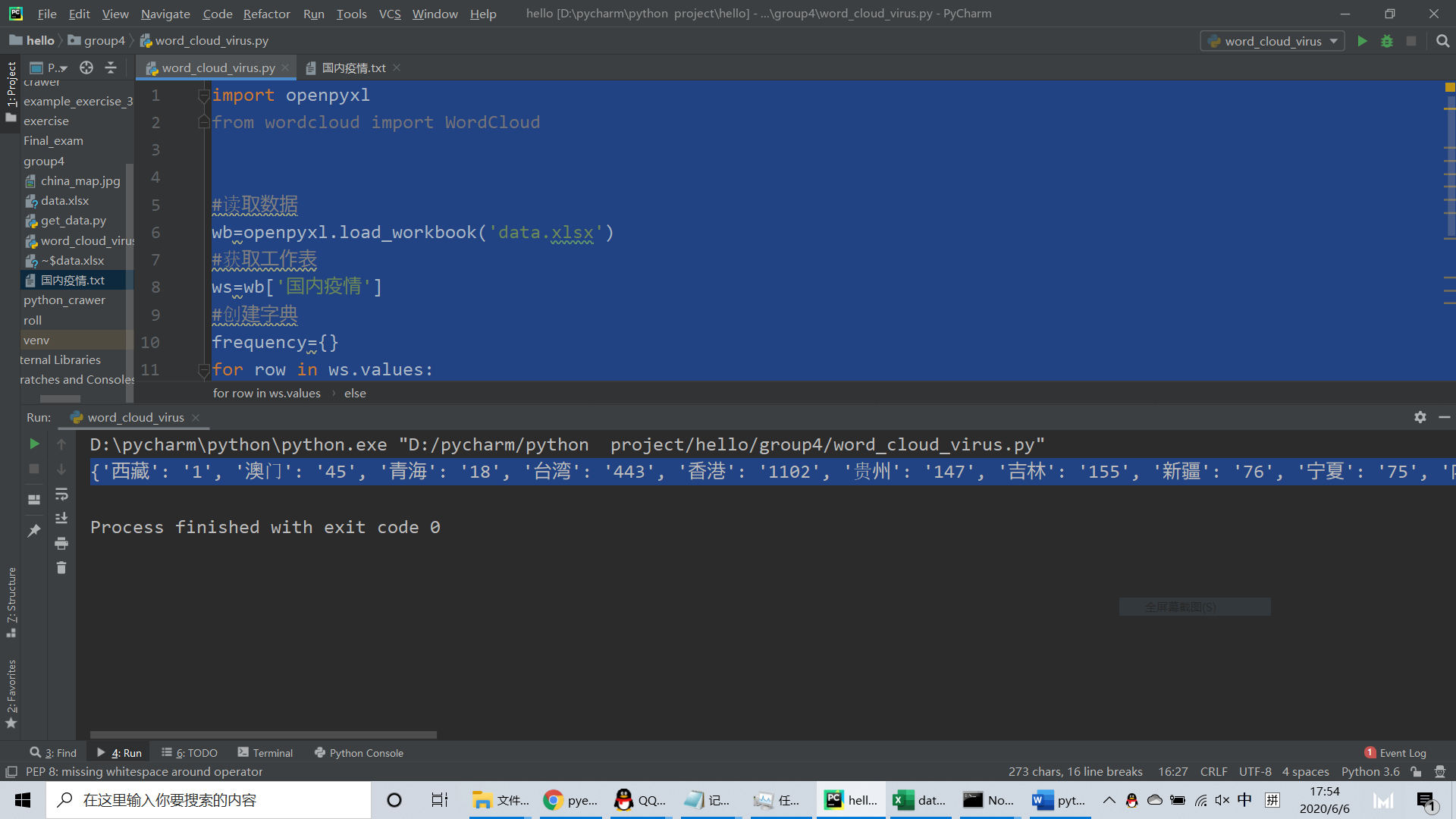
frequency[row[0]]= row[1]

print(frequency)

将各省的疫情人数从excel表读取出来通过字典将没有的列省份删除

将数据保存在国内疫情.txt文档中

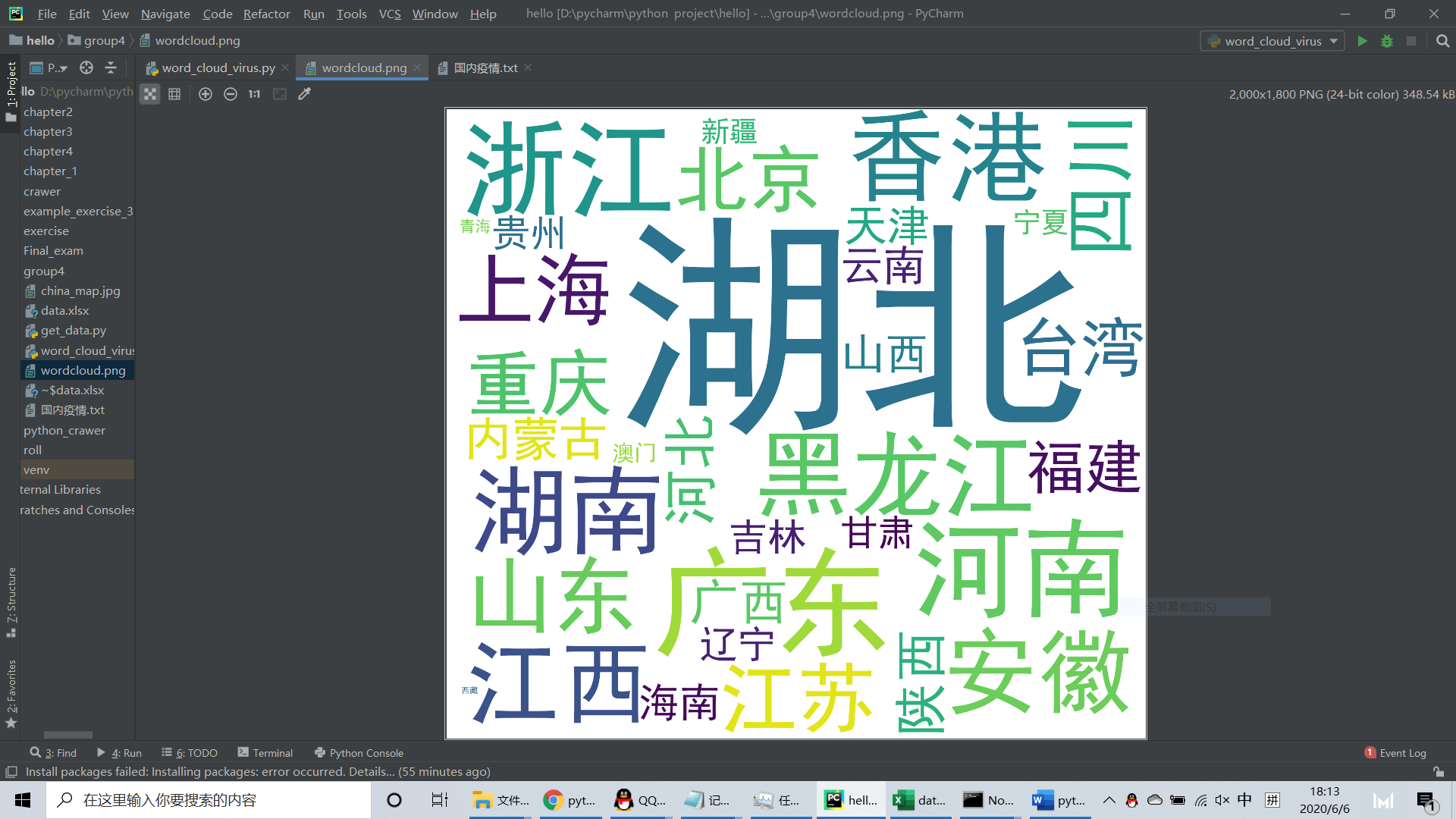
效果图如图所示:



首先通过字典对各省的数据大小进行排序然后用词云模块进行分析，得出国内疫情的分析图

图中的字体越大表示越严重反之（注意要进行编码这里的编码格式用的是windows自用的字体 ）

如图所示:

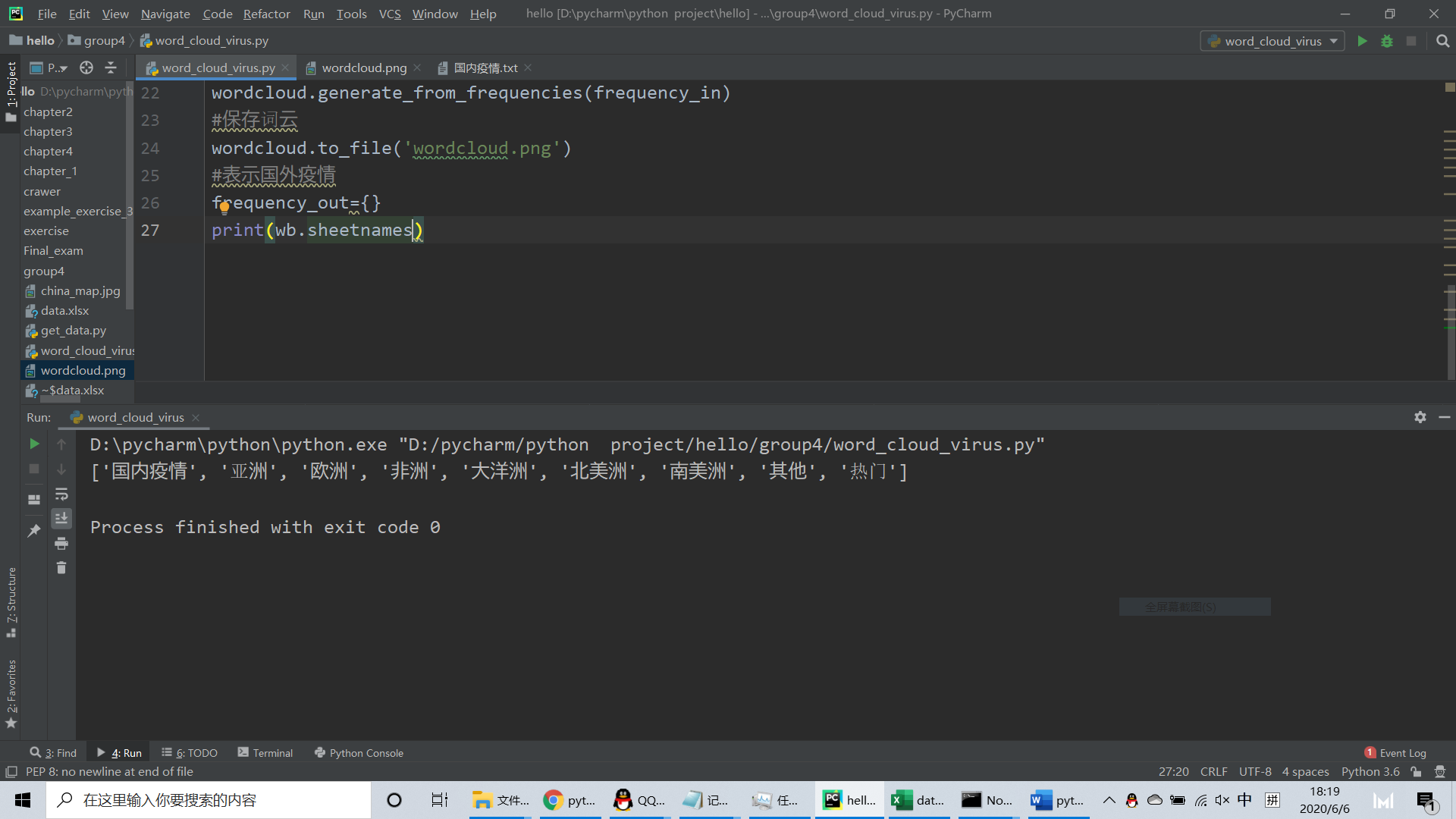


注意这里是一个没有背景图片的词云分析

这就完成了对国内疫情的绘制

然后分析国外疫情

通过国外的分析国外的是以洲来划分，从而用列表进行遍历来获取excel文件



列表核心代码如下:

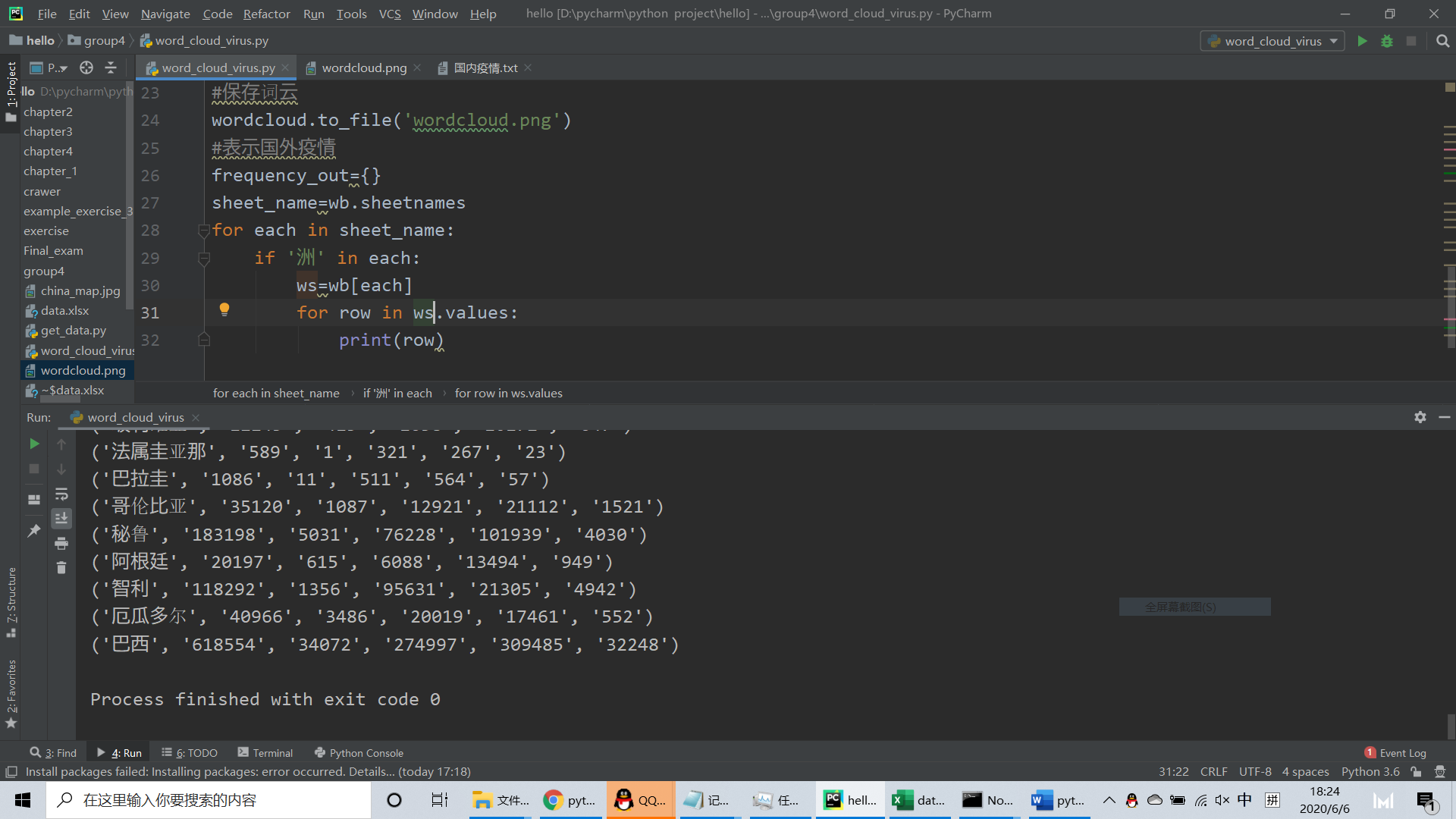
for each in sheet\_name:

if '洲' in each:

ws=wb[each]

for row in wb.values:

print(row)



通过字典的遍历将国家这个不需要的列删除

代码如下:

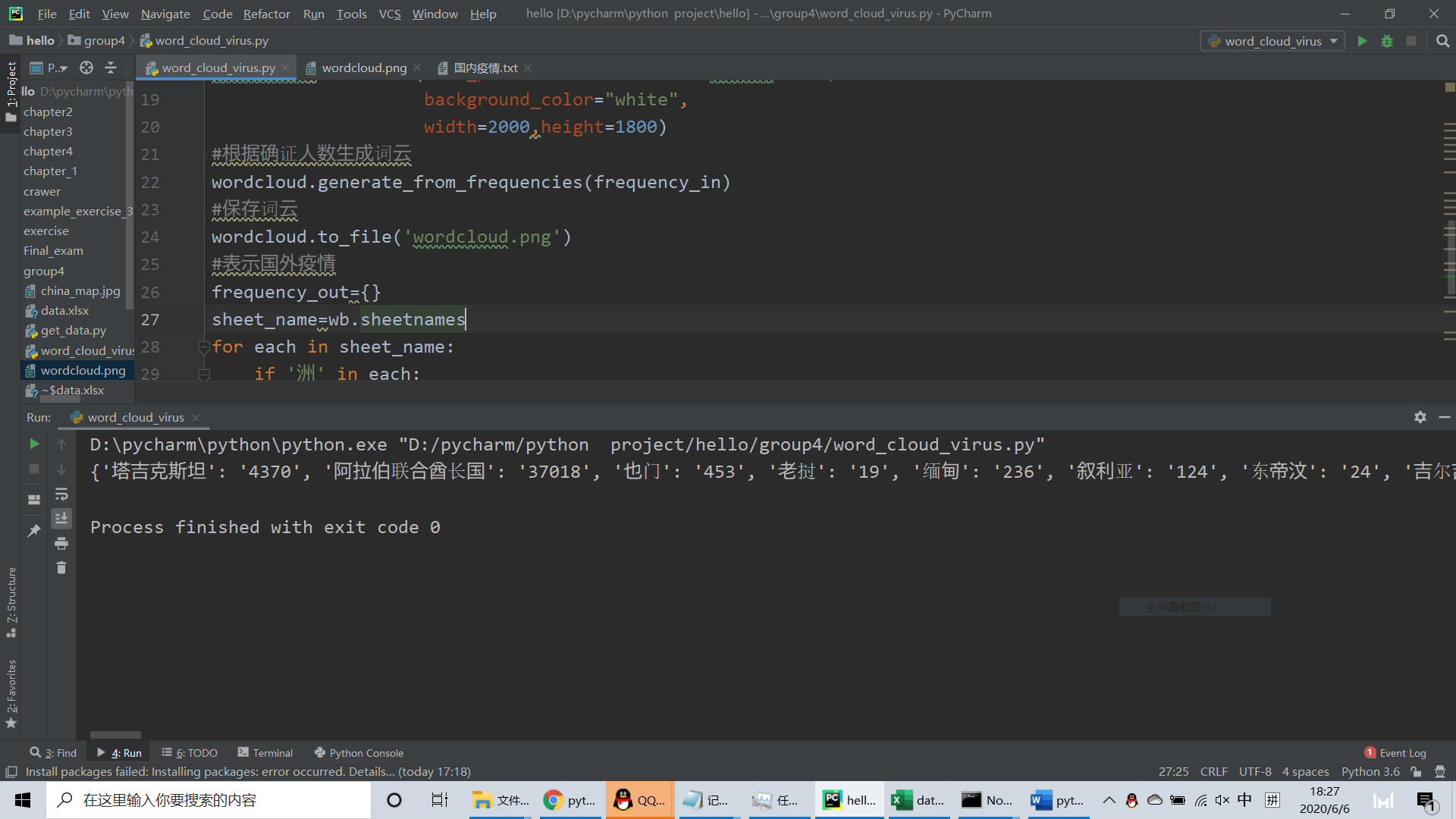
for row in ws.values:

if row[0]=='国家':

pass

else:

frequency\_out[row[0]]=row[1]

最后得出世界疫情词云图:



看出美国字体最大所以最严重、