**《中国疫情现状分析》**

**课程考核报告**

[1、引言 2](#_Toc92532443)

[爬虫 2](#_Toc92532444)

[爬虫准备工作 3](#_Toc92532445)

[2、课程设计目的 3](#_Toc92532446)

[课程设计内容 3](#_Toc92532447)

[背景知识 4](#_Toc92532448)

[3、设计步骤与方法 5](#_Toc92532449)

[步骤1，数据的爬取 requests 5](#_Toc92532450)

[步骤2，数据可视化 6](#_Toc92532451)

[4、设计结果及分析 10](#_Toc92532452)

[5、问题及心得体会 10](#_Toc92532453)

[6、参考文献 11](#_Toc92532454)

# 1、引言

## 爬虫

网络爬虫（又被称为网页蜘蛛，网络机器人）就是模拟浏览器发送网络请求，接收请求响应，一种按照一定的规则，自动地抓取互联网信息的程序。  
原则上,只要是浏览器(客户端)能做的事情，爬虫都能够做。

为什么我们要使用爬虫

互联网大数据时代，给予我们的是生活的便利以及海量数据爆炸式的出现在网络中。

过去，我们通过书籍、报纸、电视、广播或许信息，这些信息数量有限，且是经过一定的筛选，信息相对而言比较有效，但是缺点则是信息面太过于狭窄了。不对称的信息传导，以致于我们视野受限，无法了解到更多的信息和知识。

互联网大数据时代，我们突然间，信息获取自由了，我们得到了海量的信息，但是大多数都是无效的垃圾信息。

例如新浪微博，一天产生数亿条的状态更新，而在百度搜索引擎中，随意搜一条——减肥100,000,000条信息。

在如此海量的信息碎片中，我们如何获取对自己有用的信息呢？

答案是筛选！

通过某项技术将相关的内容收集起来，在分析删选才能得到我们真正需要的信息。

这个信息收集分析整合的工作，可应用的范畴非常的广泛，无论是生活服务、出行旅行、金融投资、各类制造业的产品市场需求等等……都能够借助这个技术获取更精准有效的信息加以利用。

网络爬虫技术，虽说有个诡异的名字，让能第一反应是那种软软的蠕动的生物，但它却是一个可以在虚拟世界里，无往不前的利器。

## 爬虫准备工作

我们平时都说Python爬虫，其实这里可能有个误解，爬虫并不是Python独有的，可以做爬虫的语言有很多例如：PHP,JAVA,C#,C++,Python，选择Python做爬虫是因为Python相对来说比较简单，而且功能比较齐全。

# 2、课程设计目的

在目前山西疫情的爆发，牵动着所有人的心，对整个社会的经济发展都造成了严重的影响，所以我们想通过网络爬虫去分析疫情目前的状态

## 

## 课程设计内容

本次课设主要由两部分组成：

1. 首先是网站数据的爬取，是通过新浪疫情的数据统计，根据他们提供出来的API去做疫情分析
2. 然后是使用pandas库对这些数据进行分析出来，整理出国内每个城市的现状，以及疫情对人们的影响

## 背景知识

经过一个学期python课程的学习，我对python语言也有了一定程度的掌握，特别是python丰富的第三方库，可以使得我更加容易的完成对该网络疫情信息的批量爬取与分析。

# 3、设计步骤与方法

实验环境：

Windows 10

pycharm 2021.2

python 3.9

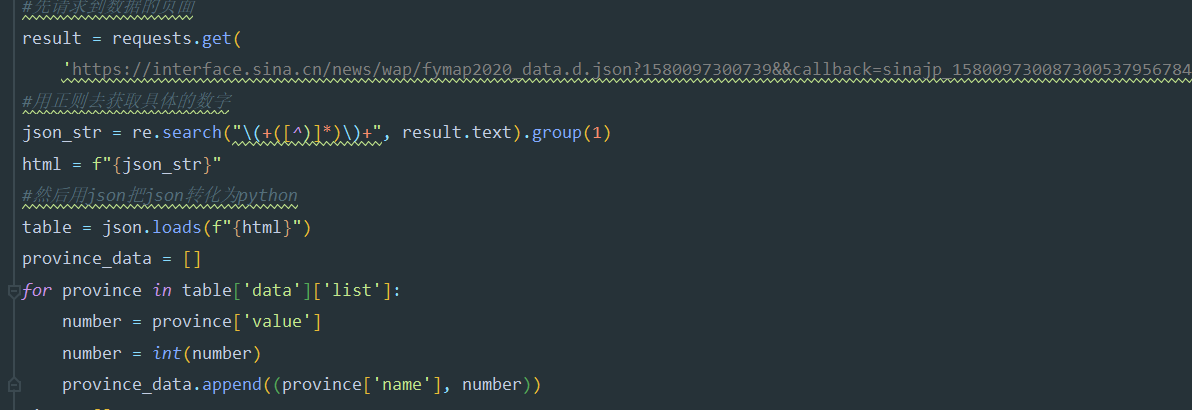
## 步骤1，数据的爬取 requests

我们打开chromd 谷歌浏览器的开发人员工具查看该信息在网站时获取到的请求信息



去读取它对应的API信息

然后再把获取到的json信息

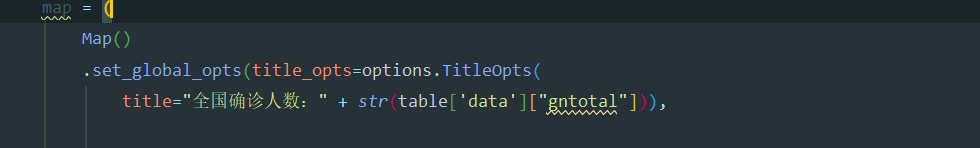


转化为python格式，然后再对其进行定位

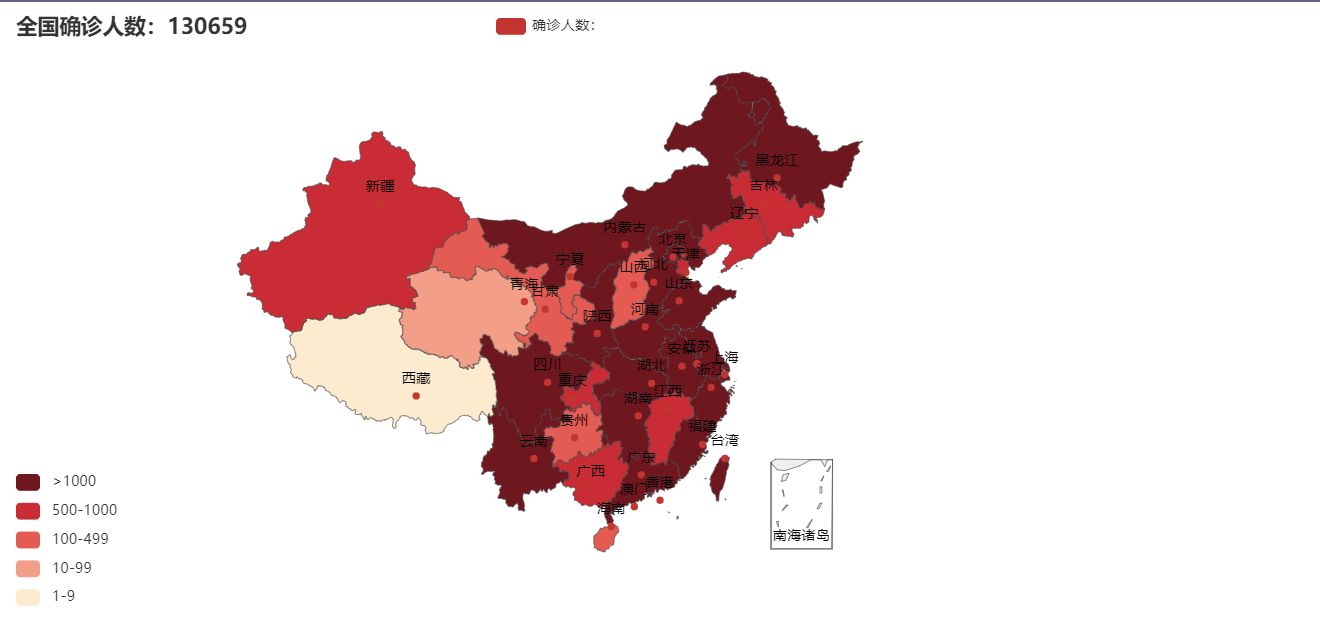
因为里面json都是已经处理好的数据，所以我们只要用json定位正确，那就可以获取我们需要的数据了，无需再去处理

## 步骤2，数据可视化

获取全国疫情累加确诊人数如下

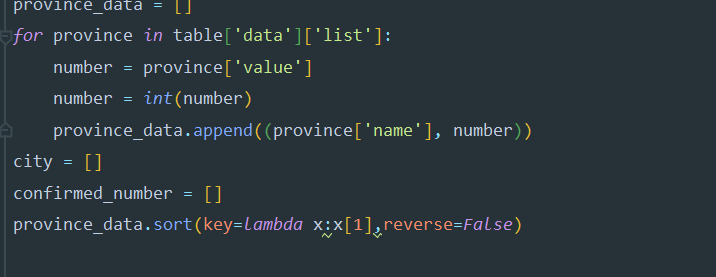


然后用地图的形式，更加清晰的表现出来



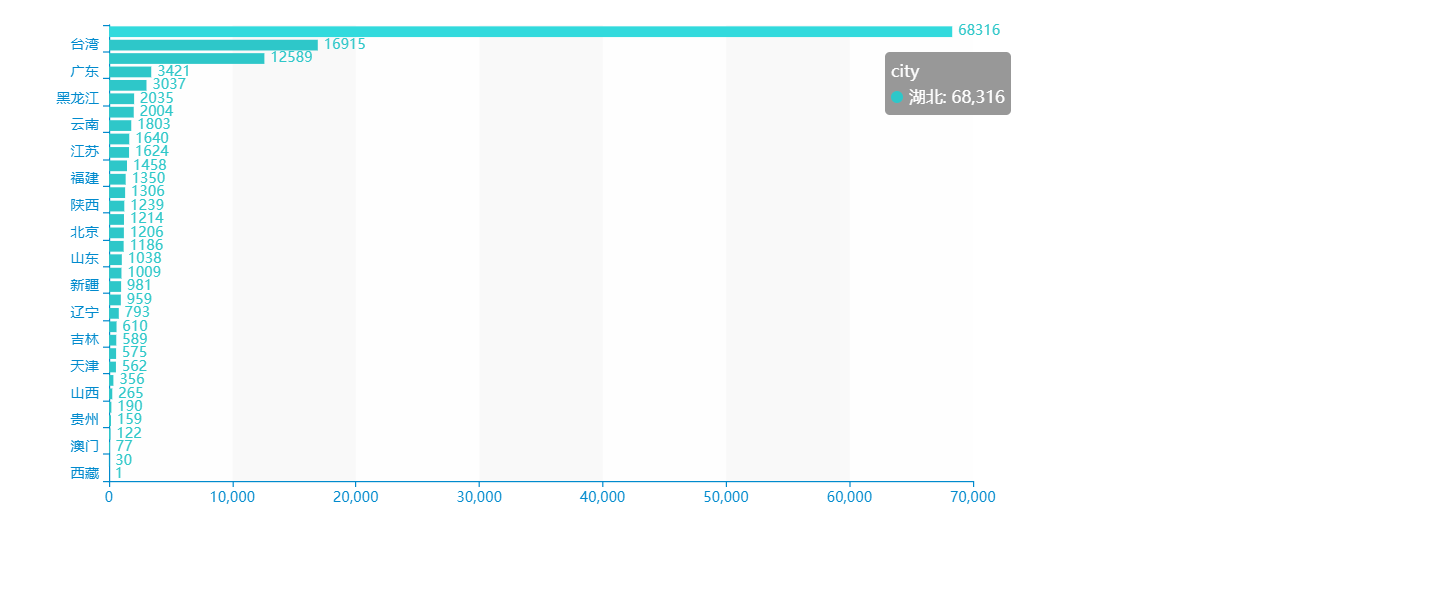
这个是查看排名的

先用列表把数据保存下来

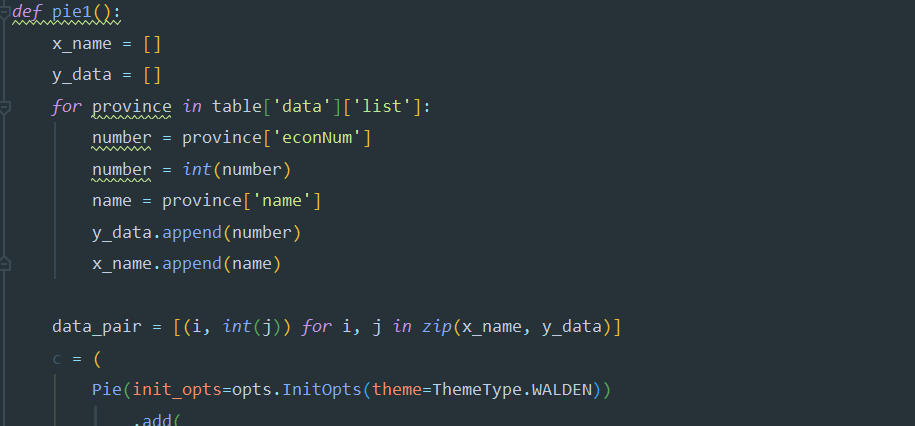


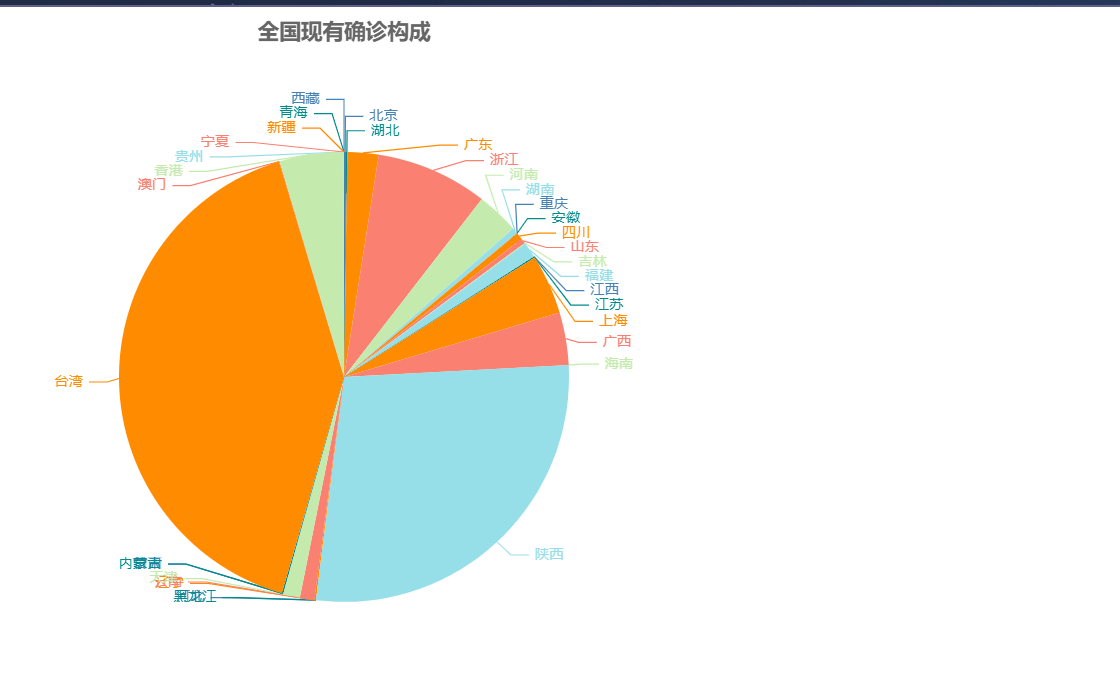
然后再让列表从大到小

再用条形图的方式展示出来



然后查看最近的发生的疫情病例





通过现有案例，发现台湾和陕西目前疫情是最严重的

# 

# 4、设计结果及分析

通过上面的数据集可以清晰的认知关于中国疫情的现状，各个省份疫情的数量，以及占比，哪有城市相对安全等这些非常重要的信息，也清晰的认知自己的能力不足，以及和别人之间的差异，还是有多虚心学习，争取在毕业的时候能找到一份自己满意的工作

# 5、问题及心得体会

本次课程完成的过程中，遇到了许多问题，大部分问题都通过自己的debug或者网络教程解决。

关于中国疫情爬取的教程网上也有很多，不过以较为老旧，经过了解，大多数都不能用了，只能自己重新去学习，去爬取对应的数据用于做数据可视化，数据分析。

本次的课程设计巩固和加深了我对python爬取信息知识的理解，提高了综合运用所学知识的能力，增强了根据课程需要选学参考资料，查阅手册、图表和文献资料的自学能力。

同时，在本次课程中，我通过实际项目，更加熟悉python的使用，并且体会到它的强大之处。首先它易于扩展，在pypi商店和GitHub上，有许多免费开源的包提供我们直接使用，这大大减少了我们的工作量，并且方便我们实行现水平下难以实行的功能。其次，它的数据结构类型很丰富，列表，字典，dataframe等数据结构大大方便了我们的使用，不过，在数据处理中，也体会到python的运行速度逊于C、C++的方面，所以也学会了通过优化算法减少时间复杂度来给程序提速。

# 6、参考文献

[Python 爬虫介绍](https://www.runoob.com/w3cnote/python-spider-intro.html)

[pyecharts可视化库使用说明](https://gallery.pyecharts.org/#/Funnel/funnel\_sort\_ascending)