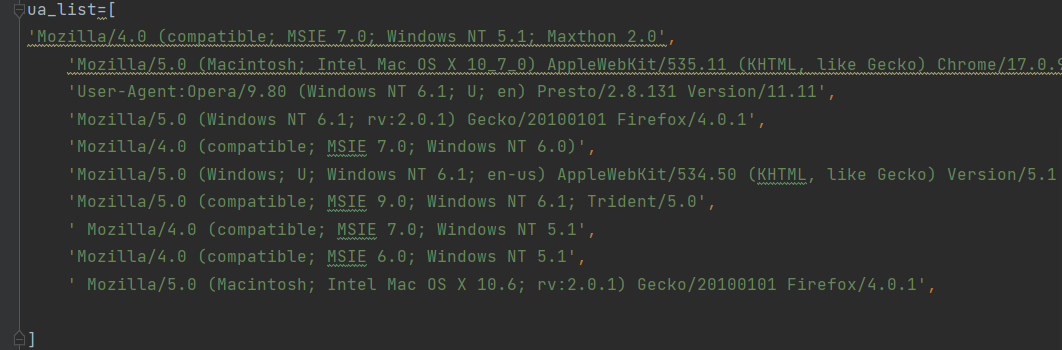
python爬虫开发记录（爬疫情数据）

本文的目的是来记录我们小组三人的开发记录，我们三人所编写的程序是爬取相关疫情 的数据，生成相应的表格形式的数据，最后进行可视化的操作。生成的疫情数据是可利用资源，供我们其他相关研究使用。以下将要开始讨论我们在开发过程中学习的记录，以及遇到的难点的记录。

1. 开发过程中的学习的记录：
2. 了解了网页的一般构成：网页一般由三部分组成，分别是 HTML（超文本标记语言）、CSS（层叠样式表）和 JavaScript（简称“JS”动态脚本语言）。
3. 获取网页http信息并进行编解码操作：认识了第一个爬虫库 urllib，以及urlopen(),Request()方法的使用。



1. 用户代理池：即访问网页的主机型号类型等，利用用户代理池是一种反爬技术。



1. 正则表达式的学习：在使用 Python 编写爬虫的过程中，re 模块通常做为一种解析方法来使用。通过审查网页元素来获取网页的大体结构，然后使用解析模块来提取你想要的网页信息，最终实现数据的抓取。



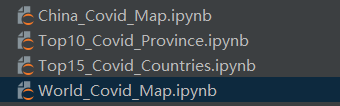


1. 绘制生成表格：

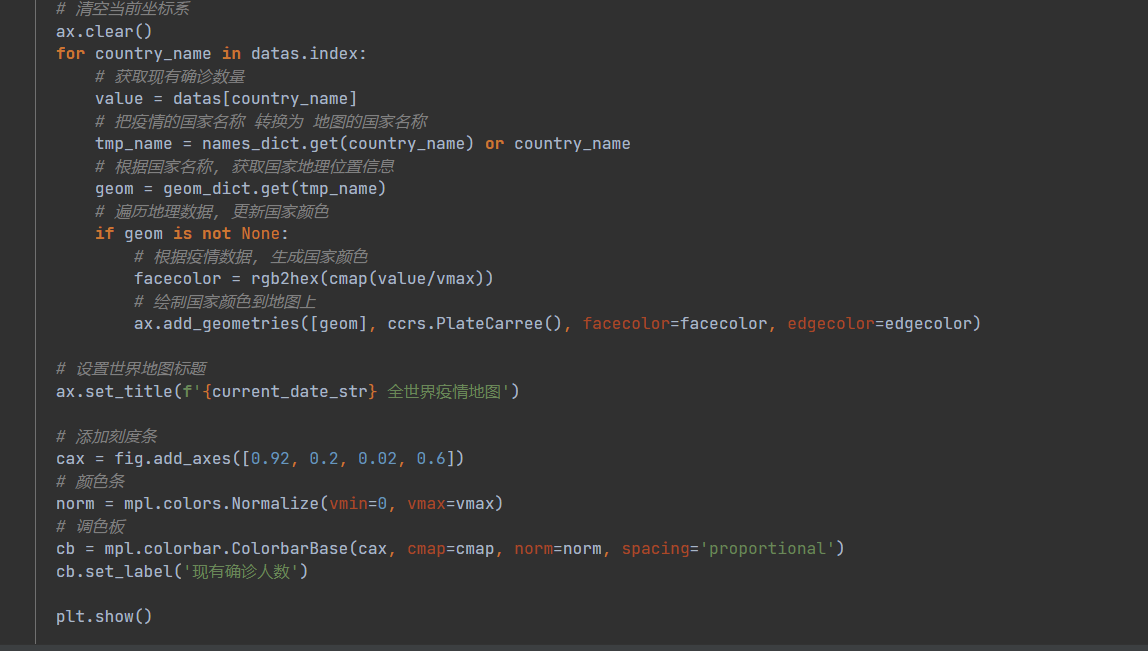


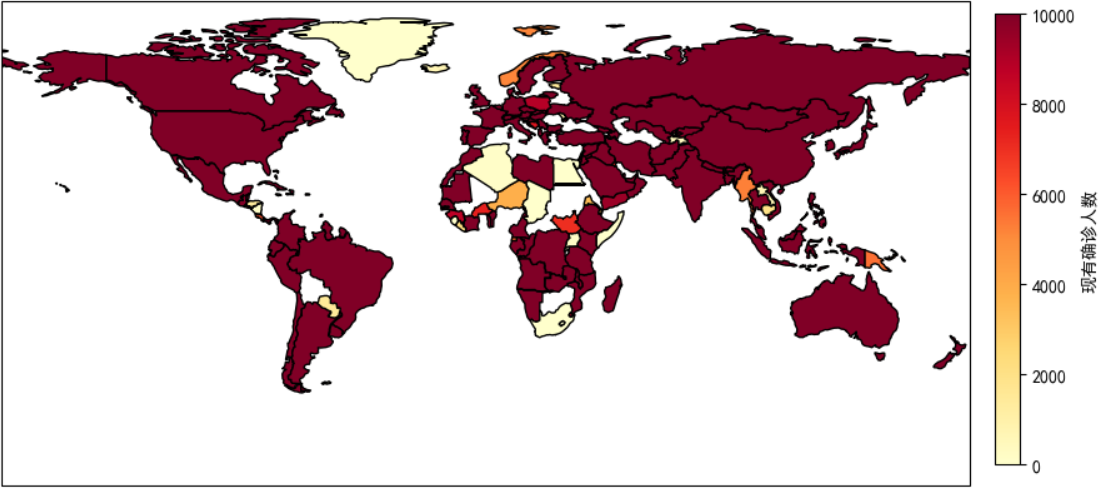
可视化操作：

具体可视化的几个文件如下：



这是选取世界疫情地图的一部分实现代码：主要是通过获取先前爬取来的数据之后，将疫情数据数据中的国家名字转换成我们可视化地图上的国家名字，并根据疫情数据来生成相应的国家颜色。





1. 开发过程中的难点记录：
2. 下载问题：在编写python爬虫程序的时候，通常需要使用一些第三方库文件，有些库文件和python环境不匹配导致运行不成功。
3. 环境问题：在编写python程序的时候通常需要调试好python，不然导致程序运行总出错或者干脆不运行。
4. 代码问题：basemap的参数数字不符合要，通过在网上寻找解决方案，找到数字进行替换成功解决问题。
5. 其他问题：爬取的网页源代码需要转换成Python文件，且需要多次确认数据类型。

5.过程问题：可视化的建立部分