新冠疫情数据爬取程序文档说明

ps：由于数据源自2022年12月17日起不再提供服务，故本程序数据来源于2022年12月11日的存档(即：data目录下的四个json文件)

数据源

- 丁香园：https://ncov.dxy.cn/ncovh5/view/pneumonia(现已停止服务)

程序功能

- 爬取自2020年1月22日的全球各国疫情数据 (生成文件：*"data/corona_virus.json"*)
- 爬取自2020年1月22日的全国各城市疫情数据 (生成文件：*"data/corona_virus_of_china.json"*)
- 爬取今日全球疫情数据 (生成文件：*"data/各国疫情数据.xls"*,*"data/last_day_corona_virus.json"*)
- 爬取今日全国疫情数据 (生成文件：*"data/各省市疫情数据.xls"*,*"data/last_day_corona_virus_of_china.json*")
- 全球疫情可视化
- 全国疫情可视化
- 全球疫情动态发展可视化
- 全国疫情动态发展可视化

依赖库

- **requests**：用于向目标网页发送请求
- **bs4**：用于解析html网页代码
- **re**：用于正则表达式匹配
- **json**：用于处理json文件
- **xlwt**：用于写入excel文件
- **tqdm**：用于显示进度条
- **pyplot**：交互模式，实现动态显示图像
- **pandas**：获取时间数据
- **numpy**：科学计算库
- **matplotlib**：用于图形绘制
- **mpl_toolkits.basemap**：用于绘制地图框架
- **pylab**：用于显示字体
- **cartopy**：用于下载地图数据

执行方法

程序的执行方法可以分为以下四步：**网页获取**，**字符串处理**，**数据处理**，**数据存储**，**数据可视化**

详细说明

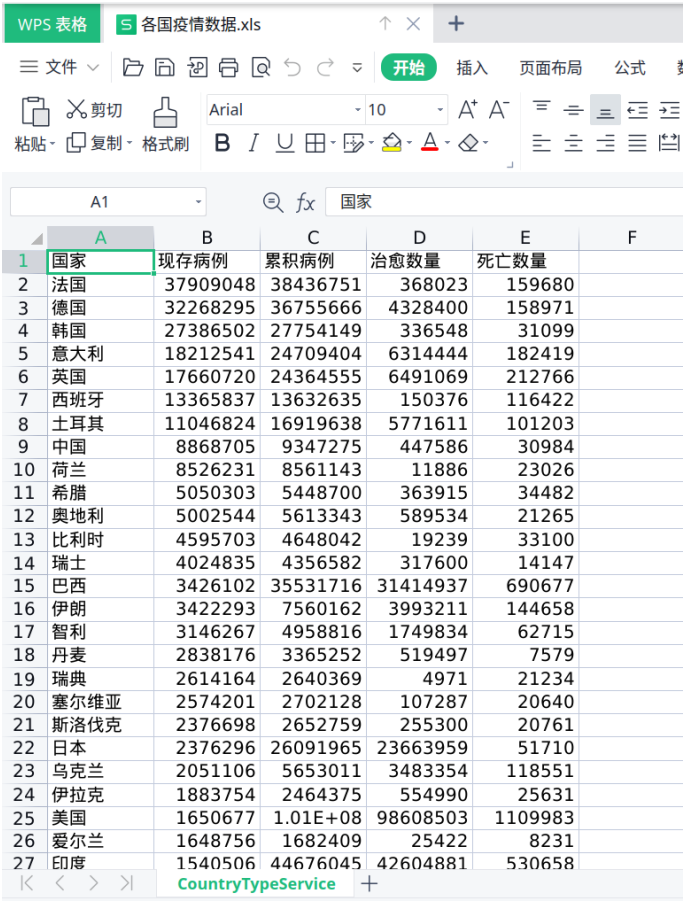
- **网页获取**：使用requests库向目标网页发送请求获取网页源代码，再使用bs4库解析网页源代码
- **字符串处理**：使用json库处理网页获取的json字符串，将其转换为python文件。

- **数据处理**：使用re库进行正则表达式匹配，获取目标数据的列表。
- **数据存储**：采用循环的形式，使用xlwt库将数据存储至excel表格
- **数据可视化**：
    - 从json文件中读取世界现有确诊病例数据
    - 从shapereader库中获取国家地理信息
    - 使用plt库绘制画布与坐标系
    - 使用rgb2hex库根据获取的确诊病例数据和国家地理信息对地图进行填色
    - 在动态图中使用plt库函数绘制条形图，不断获取每一天的数据进行绘制，通过animation实现动态可视化

效果展示

数据存储

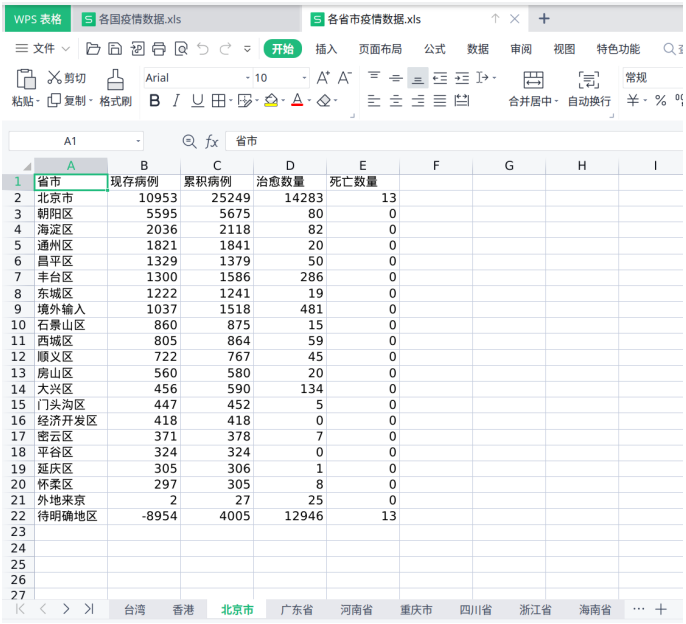
世界各国的疫情数据：



	A	B	C	D	E	F
1	国家	现存病例	累积病例	治愈数量	死亡数量	
2	法国	37909048	38436751	368023	159680	
3	德国	32268295	36755666	4328400	158971	
4	韩国	27386502	27754149	336548	31099	
5	意大利	18212541	24709404	6314444	182419	
6	英国	17660720	24364555	6491069	212766	
7	西班牙	13365837	13632635	150376	116422	
8	土耳其	11046824	16919638	5771611	101203	
9	中国	8868705	9347275	447586	30984	
10	荷兰	8526231	8561143	11886	23026	
11	希腊	5050303	5448700	363915	34482	
12	奥地利	5002544	5613343	589534	21265	
13	比利时	4595703	4648042	19239	33100	
14	瑞士	4024835	4356582	317600	14147	
15	巴西	3426102	35531716	31414937	690677	
16	伊朗	3422293	7560162	3993211	144658	
17	智利	3146267	4958816	1749834	62715	
18	丹麦	2838176	3365252	519497	7579	
19	瑞典	2614164	2640369	4971	21234	
20	塞尔维亚	2574201	2702128	107287	20640	
21	斯洛伐克	2376698	2652759	255300	20761	
22	日本	2376296	26091965	23663959	51710	
23	乌克兰	2051106	5653011	3483354	118551	
24	伊拉克	1883754	2464375	554990	25631	
25	美国	1650677	1.01E+08	98608503	1109983	
26	爱尔兰	1648756	1682409	25422	8231	
27	印度	1540506	44676045	42604881	530658	

图1 世界各国疫情数据

中国各城市的疫情数据：



省市	现存病例	累积病例	治愈数量	死亡数量
北京市	10953	25249	14283	13
朝阳区	5595	5675	80	0
海淀区	2036	2118	82	0
通州区	1821	1841	20	0
昌平区	1329	1379	50	0
丰台区	1300	1586	286	0
东城区	1222	1241	19	0
境外输入	1037	1518	481	0
石景山区	860	875	15	0
西城区	805	864	59	0
顺义区	722	767	45	0
房山区	560	580	20	0
大兴区	456	590	134	0
门头沟区	447	452	5	0
经济开发区	418	418	0	0
密云区	371	378	7	0
平谷区	324	324	0	0
延庆区	305	306	1	0
怀柔区	297	305	8	0
外地来京	2	27	25	0
待明确地区	-8954	4005	12946	13

图2 中国各城市疫情数据

可视化

世界疫情国家top15：

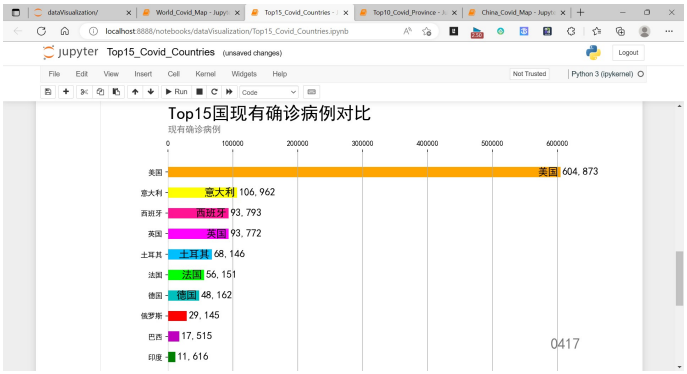


图3 世界疫情国家top15

中国疫情省市top10：

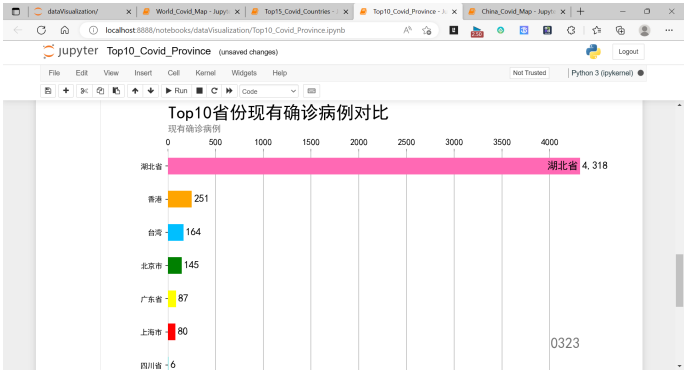


图4 中国疫情省市top10

世界疫情地图：

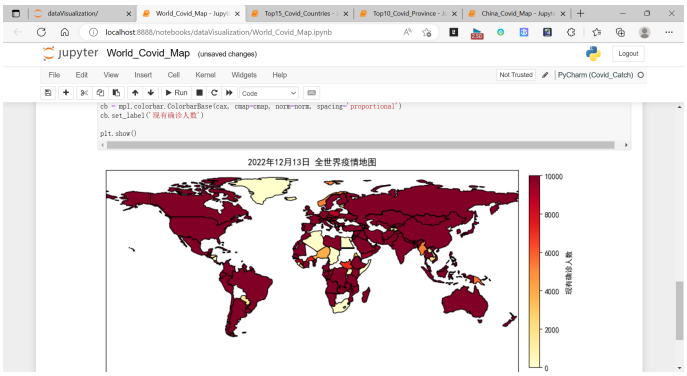


图5 世界疫情地图

中国疫情地图：

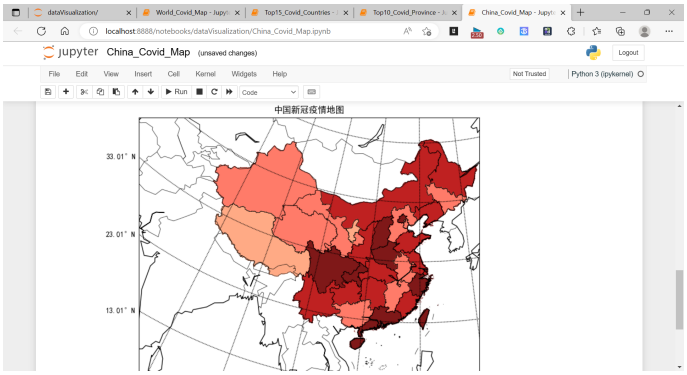


图6 中国疫情数据