 DOI: https://doi.org/10.26549/gcjsygl.v4i11.5295

Analysis of Data on Public Health Based on Python

Minxia Ji　Hongli Shi

Jinling Institute of Technology, Nanjing, Jiangsu, 211169, China

Abstract

With the advent of the era of big data and artificial intelligence, the network and information technology have begun to penetrate into all aspects of human daily life, and the amount of data generated has also shown an exponential growth trend. At the same time, the magnitude of the existing data has far exceeded the scope of the current manpower can handle. In this context, data analysis has become a new research topic in the field of data science. Through such technology, we can obtain data, store data, and analyze data, extract useful and important information from the data.

Keywords

Python; data analysis; Pandas; data visualization; Matplotlib; plotly

基于 Python 对公共卫生进行数据分析

季旻霞　石弘利

金陵科技学院，中国·江苏 南京 211169

摘　要

随着大数据和人工智能时代的到来，网络和信息技术开始渗透到人类日常生活的方方面面，产生的数据量也呈现指数级增长的态势。同时，现有数据的量级已经远远超过了目前人力所能处理的范畴。在此背景下，数据分析成为数据科学领域中一个全新的研究课题。通过这样的技术，我们可以获取数据、存储数据、分析数据，从数据中提炼出有用重要的信息。

关键词

Python；数据分析；Pandas；数据可视化；Matplotlib；plotly

1 引言

2020 年新冠在全球蔓延，不少人因为新冠失去生命与亲人。针对这一情况，我们必须足够重视并采取一定的行为来改善。那我们应如何改善？随着云时代的来临，大数据吸引了越来越多的关注，大数据这一词也越来越多被提及。在这次的新冠疫情中，大数据就发挥的无可替代的作用，分析这些数据内部所蕴含的规律，预测相关运行趋势，得出有效结论，从而预防病毒的肆意蔓延。

2 Python 简介

Python 是一个高层次的结合了解释性、编译性、互动性和面向对象的脚本语言，python 简单开源易上手，拥有丰富强大的库，流程可控，工作高效。

【作者简介】季旻霞，本科学历，金陵科技学院。通讯方式：13306288061；1922506727@qq.com。

3 数据导入

两种方式可以获取到此次疫情数据，第一种是爬虫，爬虫因本身具备的突出优势，被广泛用于数据信息采集中爬虫提取页面源代码 [1]。主要有两种方法：requests 库和 selenium 库。requests 适合提取需要的元素信息直接保存在页面的源代码中，它提取页面快，但不适合用于提取元素不是加载在源代码的项目的情况下。第二种是通过 akshare 库，akshare 是基于 python 的开源数据接口库。这里笔者用的是第一种方法，导入相关部分数据来进行数据分析。笔者使用的是 requests 库，将数据导入之后，由 json 格式转换为方便分析的 dataframe 格式，源数据由于 API 采集的机制，包含大量重复数据，无效数据，缺失数据，所以接下来要对这些数据进行处理 [2]。

4 数据处理

数据处理方面，Python 中的插件 Pandas 功能突出。

 DOI: https://doi.org/10.26549/gcjsygl.v4i11.5295

Pandas 十分方便快捷，主要包括数据输入输出，数据清洗处理，数据挖掘等。Pandas 拥有强大数据清理功能，可以去重复值，通过删除或者填补等处理缺失值和异常值。论文日期不是标准表示，应将它化为年月日的方式，并使用 loc 根据 index 索引来提取月日。并且论文用默认索引不是最佳方案，可以以更新日期作为索引，表一为数据处理后的结果。如下为处理的主要代码：

```
def time_c(timenum):
    timetemp=float(timenum/1000)
    tuptime=time.localtime(timetemp)
     standardtime=time.strftime('%Y-%m-%d %H:%M:%S',tuptime)
    return standardtime
for i in range(len(df)):
    df.iloc[i,16]=time_c(df.iloc[i,16])
for i in range(len(df)):
    df.iloc[i,16]=df.iloc[i,16][5:10]
```

updateTime	continentName	countryName	provinceName	currentConfirmedCount	suspectedCount	curedCount	deadCount
08-05	非洲	摩洛哥	摩洛哥	7081	0	19629	417
08-05	亚洲	乌兹别克斯坦	乌兹别克斯坦	9096	0	18051	167
08-05	北美洲	海地	海地	6605	0	706	165
08-05	非洲	安哥拉	安哥拉	782	0	503	59
08-05	亚洲	孟加拉国	孟加拉国	101657	0	141750	3267
...	...	...	...	...	...	...	...
02-29	欧洲	大不列颠及北爱尔兰联合王国	大不列颠及北爱尔兰联合王国	12	0	8	0
02-29	欧洲	英国（含北爱尔兰）	英国（含北爱尔兰）	12	0	8	0
02-28	欧洲	北爱尔兰	北爱尔兰	1	0	0	0
02-23	亚洲	中国	青海省	0	0	18	0
02-23	亚洲	中国	西藏自治区	0	0	1	0

图 1 数据处理后的结果

5 数据可视化

数据可视化有很多方式，可以通过 excel、matplotlib、seaborn、plotly 以及词云图等方式给我们展示数据的数量以及发展趋势，对可视化结果进行分析，从中提取出有效信息，得出有用结论供人们学习以及商业参考。在论文的疫情数据中，我们可以通过分析得出哪些地区是重灾区，需要引起人们的重视，严格控制人流量预防疫情快速蔓延，提醒相邻地区注意防范等。

在论文中，笔者主要使用了 plotly 方法绘图。Plotly 是一个非常著名且强大的开源数据可视化框架，它通过构建基于浏览器显示的 web 形式的可交互图表来展示信息，可创建多达数十种精美的图表和地图，Matplotlib 存在不够美观、静态性、不易分享等缺点，限制了 Python 在数据可视化中的发展。为了解决这个问题，新型的动态可视化开源模块 Plotly 应运而生，Plotly 具有动态、美观、易用、种类丰富等特性。

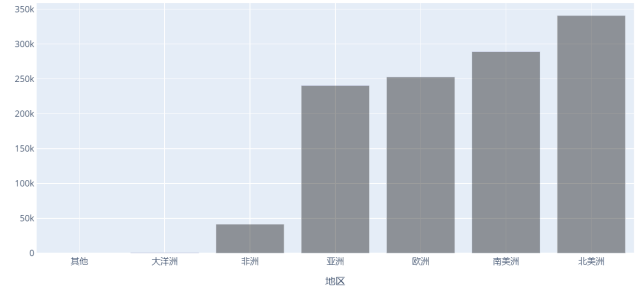

图 2 各区块死亡人数

上图是各区块的死亡人数，使用了 groupby 分组计算出区块死亡人数。从图中可见，北美洲情况最严重，应重点注意。中国相关部门应严格把控输入输出来控制疫情在区块间的传播。

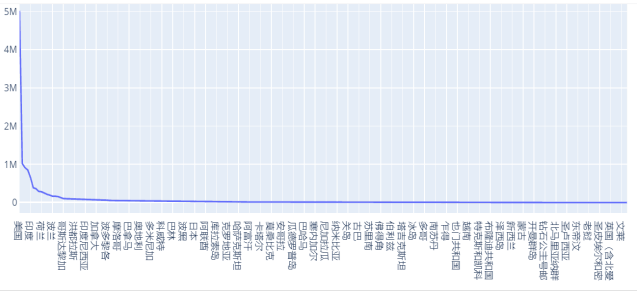

图 3 各地区的确诊人数

上图是使用 plotly 绘制的折线图，折线图易于展现趋势，上图清晰可见美国疫情的严重性。

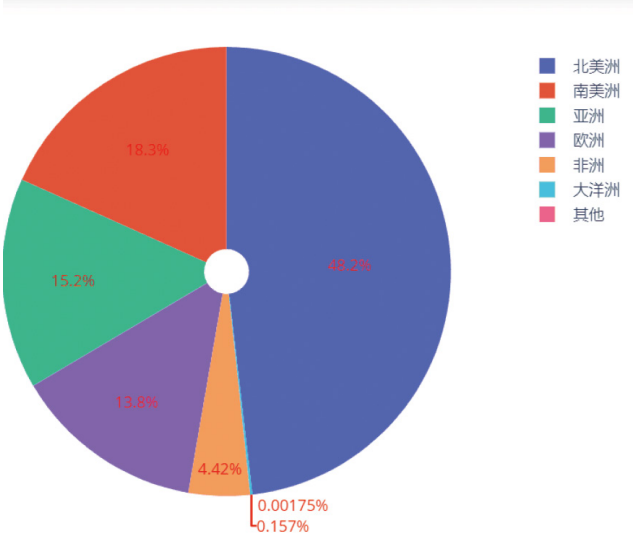

图 4 08-06 各州确诊情况

 DOI: https://doi.org/10.26549/gcjsygl.v4i11.5295

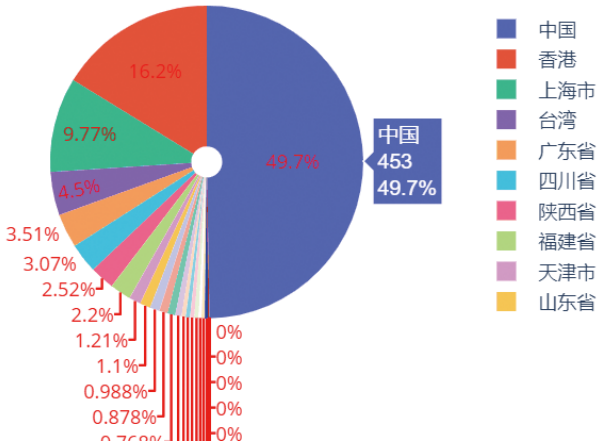

图 5 08-06 中国确诊人数

可视化核心代码：

```
df7=df12[['provinceName','currentConfirmedCount']]
df7=df7['currentConfirmedCount'].sort_values(ascending=False).head(10)
trace=[go.Pie(labels=df7.index.tolist(),values=df7.values.tolist(),\
        hole=0.1,textfont=dict(size=12,color='red'))]
layout=go.Layout(title='八月六号中国确诊人数较多地区')
fig=go.Figure(data=trace,layout=layout)
pyplot(fig)
```

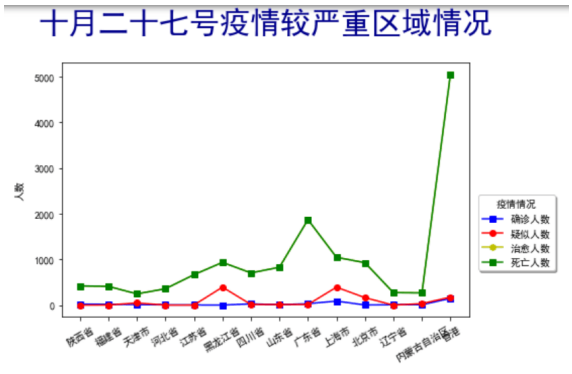

图 6 08-05 疫情较严重区域情况

笔者使用了 matplotlib 绘制折线图来呈现疫情情况描绘在八月五号重要疫情较严重地区疫情发展趋势，先通过数据筛选，再使用可视化绘图 [3]。下图可以明显发现香港地区疫情较为严重。居住香港地区人应注意减少外出，严格佩戴口罩，相邻地区也要引起注意和重视。其他地区人应尽量避免出入疫情严重区。其中，主要代码为：

```
fig=plt.figure(figsize=(7,5))
fig.add_axes([0.2,0.15,0.8,0.7])
plt.plot(df11.iloc[:,2],df11.iloc[:,3],'bs--',\
        df11.iloc[:,2],df11.iloc[:,4],'ro--',\
        df11.iloc[:,2],df11.iloc[:,5],'yh--',
        df11.iloc[:,2],df11.iloc[:,5],'rs--')
plt.legend([' 确诊人数 ',' 疑似人数 ',' 治愈人数 ',' 死亡人数 '],bbox_to_anchor=(1.01,0.5),ncol=1,title=' 疫情情况 ',shadow=True,loc=0)
```

图 7 08-06 疫情词云图

词云图可以有效划重点，高亮出重要信息呈现出来。论文可以显示出中国疫情较严重地区。

6 结语

随着网络信息化的飞速发展，人们面对越来越多纷繁复杂的数据时，需要分析处理，需要利用数据可视化后的结果指导和解决各种学习工作中的问题。论文利用 python 语言作为编程基础，利用 pandas 进行数据分析以及 matplotlib 和 plotly 来进行数据可视化，由此来提取出有效信息 [3]。

参考文献

[1] 宋永生 , 黄蓉美 , 王军 . 基于 Python 的数据分析与可视化平台研究 [J]. 现代信息科技 ,2019(11):143-145.

[2] Allen B Downey. 像科学家 - 样思考 Python[M]. 北京 : 人民邮电出版社 ,2013.

[3] Ivan Idris. Python. 数据分析基础教程 [M]. 北京 : 人民邮电出版社 ,2014.