# 本科毕业论文


论文题目：基于网络爬虫和数据挖掘技术的新冠疫情可视化分析

姓　　名：曲俊杰　学　　号：16307110125

院　　系：软件学院

专　　业：软件工程

指导教师：徐迎晓　职　　称：高级工程师

单　　位：复旦大学

完成日期：2020　年　5 月 25 日

# 目录

# 摘要

2020 开年，新冠疫情肆虐全球，给人们带来了空前的恐慌，短短三个月内，数以百万计的居民生命安全遭受威胁，亿万人的生活受到不同程度的影响，中国作为 14 亿人口的泱泱大国，充分使用了应对 2003 年非典疫情时的经验，采用地区强制隔离等政策抗击疫情，全国上下齐心协力，共同战疫。中国人民抗击疫情的坚定决心和显著成效给世界各国极大的鼓舞和信心。

项目将从当前世界关于新冠疫情的研究入手，简述该项目的知识背景和实现意义，通过网络爬虫的技术手段收集整理疫情数据，在保证数据准确性的基础上，使用可视化技术对数据进行多维度的直观呈现，并使用 echart 制图技术和 javascript 技术结合搭建每日疫情上报页面，使用机器学习相关技术做疫情发展趋势的简单预测。

此外，项目也对疫情中媒体的作用予以关注，使用相关度判别方法筛选整理出近千篇微信公众号文章，使用中文分词程序结合疫情中出现的具体问题作词频分析，利用 wordcloud 技术生成新冠疫情词图，揭示新冠疫情中媒体视角下的关注热点。

论文将对项目开发具体过程中遇到的问题和解决方案做一个总结，并对项目可供延伸改进的地方做一些猜想和尝试。

**关键词** 新冠疫情，网络爬虫，数据可视化，词频分析，机器学习

# ABSTRACT

Since we stepped into 2020, the sudden outbreak of COVID-19 led to a global pandemic, leaving millions of households' safety at risk, in merely 3 months' time. With unprecedented fear and panic evoked, people's standard of living has been highly degraded as restrictions affected daily life in every aspect. Similar policies in response to SARS in 2003 were adopted immediately, such as implementing travel restrictions within states and regions. As a country with the biggest population of 1.4 billion, the concerted efforts of all Chinese in fighting against this epidemic prompted remarkable results, which purposes as a piece of encouragement and confidence for the whole world.

This project will start with the current research on COVID-19 in the world and briefly describe the knowledge background and the significance. I will collect and organize the epidemic data through Web crawler. On the basis of ensuring the accuracy of the data, some visualization technology is used to show the data more intuitively in multiple dimensions. Also, the echart and JavaScript technology are used to build a daily outbreak report page. As a highlight, this project will try to use the machine learning technology to make a simple forecast of COVID-19.

In addition, the project also paid attention to the role of the media by sorting out nearly 1,000 articles in WeChat platform. I will analyze those articles using the Chinese word segmentation program combined with the wordcloud program which purposes to reveal the focus of attention from the perspective of the media during COVID-19.

The thesis will make a summary of the problems encountered during the implement and their solutions. Then try to make some conjectures and improve the model.

**Keywords** COVID-19, Web crawler, data visualization, word frequency analysis, machine learning

# 第一章 引言

## 1.1 新冠疫情可视化分析的研究背景

2019 年 12 月，新型冠状病毒肺炎（简称新冠肺炎，COVID-19）在中国湖北省武汉市爆发，中国政府迅速组织人力物力救援武汉，防控疫情。2020 年 1 月，世界卫生组织（World Health Organization WHO）将新型冠状病毒疫情认定为“国际关注的突发公共卫生紧急事件”。

新冠病毒爆发以来，各领域工作者积极投身科研工作，发表了大量相关研究论文，辅助实际防疫工作中的决策，为快速有效遏制疫情做出了巨大贡献，也为本项目提供了详实的理论基础。

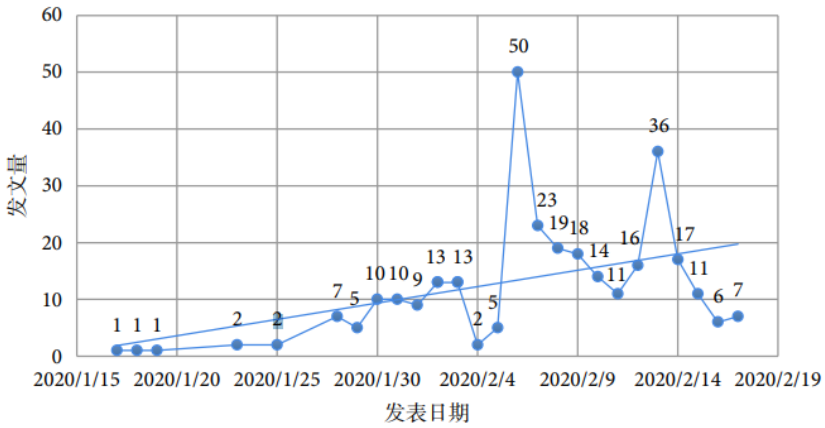


图 1-1 新型冠状病毒相关研究发表数量时间趋势

疫情期间，世界各国疫情相关数据都公开透明，提供了多种可供查看和收集的方法，本次项目也尝试整理收集了不同国家和地区的疫情相关数据供研究参考。使用适当的可视化方法处理这些数据，将使得疫情数据更加清晰直观，更具价值。

媒体是本次疫情中传导信息、增强信心的重要力量，当前关于新冠疫情的相关研究中数据和医疗研究占大多数，本项目对疫情中的媒体文章予以适当关注，分析出民众关注热点，为更好应对类似突发状况累积宝贵经验。

## 1.2 项目结构概述

本项目主要分为如下几个部分：

1. 数据收集部分，项目研究涉及的疫情数据分为两个主要模块，一是世界各国疫情相关人数（包括新增/累计确诊人数、新增/累计治愈人数、新增/累计死亡

人数等）的汇总，通过官方提供的数据接口可以收集取得；另一模块是疫情期间的公众号文章，通过网络爬虫技术进行爬取。

2. 数据清洗部分，疫情相关人数按国家/地区/日期等维度进行分类，形成划分清晰的原始数据；而微信公众号文章则按相关度筛选出近千篇具有代表性的文章作为数据集。
3. 数据可视化部分，数据可视化是本次项目的核心部分，如何清晰直观地描述数据，使用了 Matplotlib 和 Echarts 技术描述人数相关数据，展现其变化趋势并作对比总结；使用中文分词技术和词云技术处理微信公众号文章，作词频分析和热点分析。
4. 疫情预测部分，尝试用机器学习手段预测疫情发展的趋势走向，对本次数据集的构建和使用进行反思和思考，提出一些可行的改进方式。

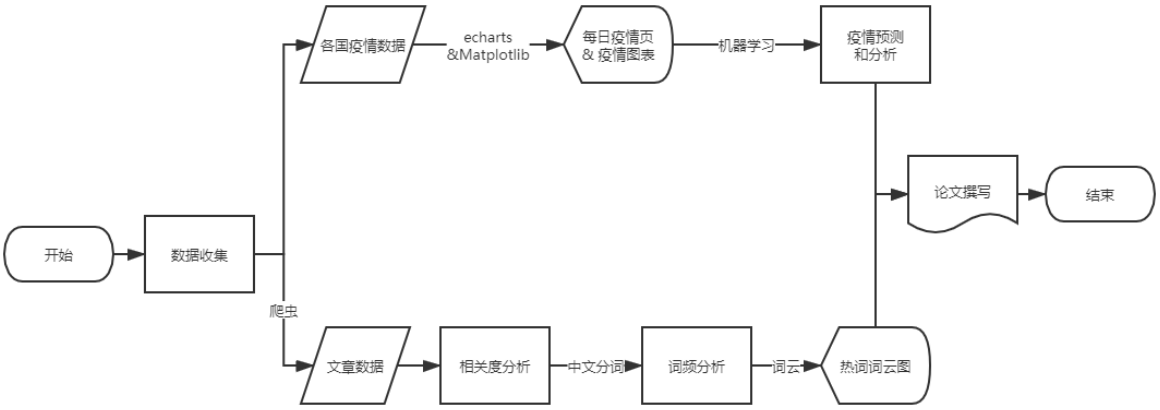


图 1-2 “基于网络爬虫和数据挖掘技术的新冠疫情可视化分析”项目流程图

## 1.3 章节结构介绍

本文将以代码具体实现为基础，按照代码模块划分相关章节，具体说明项目实践过程中遇到的问题和解决方案，并融入对技术的应用和思考，从新冠疫情的第一手数据入手进行可视化探索。

**第一章 引言**

介绍当今新冠疫情的相关研究进展，以及数据的主要来源和论文的框架。

**第二章 项目相关技术介绍**

介绍论文的代码实现中使用到的具体知识。

**第三章 数据收集模块的设计和实现**

介绍利用网络爬虫收集数据中遇到的问题和解决方案。

**第四章 疫情数据的可视化分析**

介绍本项目中疫情数据的可视化手段及其优劣，以及不同可视化手段的技术应用分析、效果呈现。

**第五章 疫情预测的相关尝试**

介绍如何利用收集到的数据进行预测疫情发展趋势的尝试以及一些延伸内容。

**第六章 总结和展望**

本次研究项目的总结。

# 第二章 项目相关技术介绍

## 2.1 网络爬虫 Web spider 技术简述

自互联网问世以来，网络信息资源一直保持着迅猛增长的态势，为了便捷人们方便快速的获取有效信息，网络爬虫应运而生。简单来说，网络爬虫就是一个自动下载网页的自动化脚本，帮助我们在短时间内获取大量网页信息的工具。

根据系统结构和实现技术，网络爬虫可大致分为以下几种类型：通用网络爬虫、主题网络爬虫、增量式网络爬虫和深层网络爬虫。因为我们的目标信息相对明确，可以基于搜索引擎给出的结果进行爬虫，所以采用主题网络爬虫和深层网络爬虫结合的形式进行数据收集。

主题爬虫是按照预先定义的爬行主题，在给定初始 URL 种子集后，根据一定的分析算法，对爬行网页进行主题相关分析，过滤不相关网页，在不断抓取相关网页的过程中，将主题相关链接放入等待队列，直到满足设定条件为止。

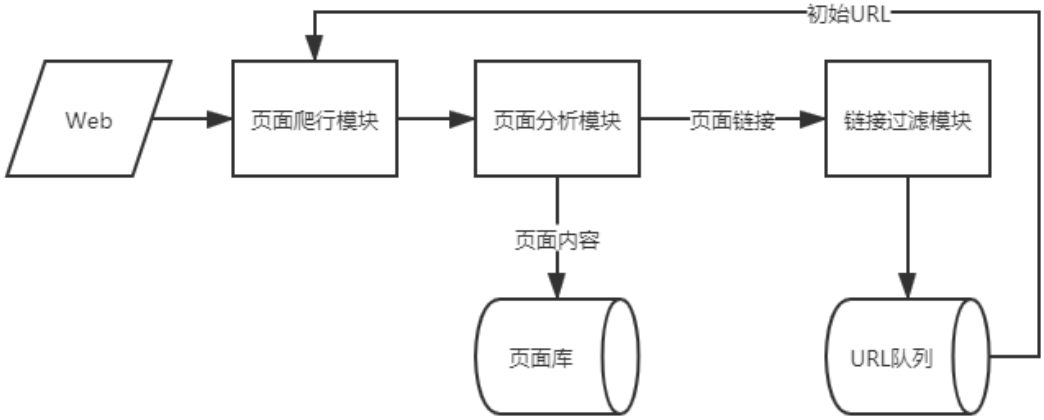


图 2-1 主题网络爬虫体系结构图

关于主题爬虫的爬行策略——针对等待序列的访问顺序的多种处理方式以及主题相关性的判断是主题爬虫中的重点，常分为传统启发式方法、语义分析方法、经验爬行（基于前次爬行结果迭代）方法。如果我们基于搜索引擎的结果进行爬取，则可以将传统启发式方法和经验爬行方法相结合得到理想的结果集，直接利用搜索引擎标签的优先策略（热度、时间顺序等）免去部分设计优先方法的繁杂。

深层网络爬虫作为辅助手段，主要针对那些大部分内容不能通过静态链接获取、隐藏在搜索表单后的页面，这些页面需要用户提交特定关键词才能取得，例如用户注册后可见的页面，或是根据用户喜好有所区别的推荐页面等。

深层网络爬虫的重点是表单/Cookie 的构建填写，针对不同的网站需要不同的应对策略，将在第三章中作具体阐述。

## 2.2 可视化技术简述

我们使用爬虫技术收集来的数据大多数数字构建成的表格，缺乏直观性且数据特征不明显，需要使用可视化手段突显出其数据特征。图数据由于自身的结构特征，可以很好地表示事物之间的关系，便于多方向对比，使用 Matplotlib 库中的画图方法初步处理疫情数据可以形成多种图表，带来直观感受，也为后续的曲线拟合带来便利。

在大数据处理中，好的可视化分析还应该具备一定的交互性，以本次疫情地图为例，疫情地图极大了减少了数据呈现所需要的空间，增强了视觉效果，使用 Echarts 技术和一些 JS 技巧可以帮助我们制作自己的疫情地图。

本项目中还用到了文本可视化的相关知识，文本可视化服务于人们对文本进行分析和理解的基本过程，相当程度上依赖于自然语言处理技术，本项目中主要使用基于词频的文本可视化来探寻疫情发生以来的媒体热点，使用 wordcloud 技术构建词云图。

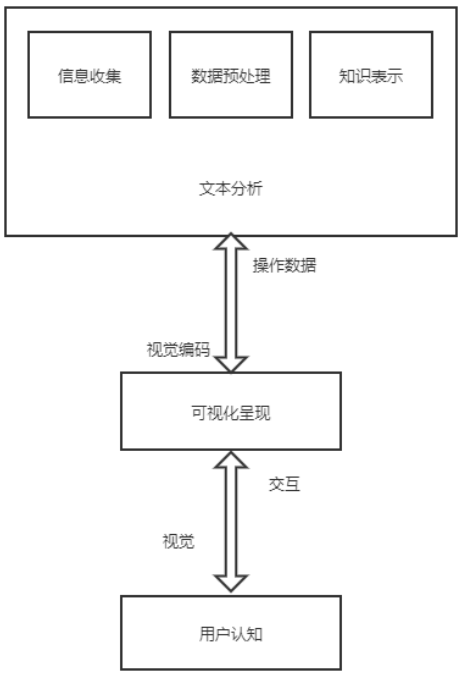


**图 2-2 文本可视化的基本框架**

关于可视化技术部分的应用和思考将在第四章中做详细阐述。

## 2.3 中文分词技术简述

自然语言处理是人工智能的一个重要分支，中文分词是中文自然语言处理中的一项基础性工作，主要难点在于英文中的一个单词就是一个词，而汉语是以单个字为单位进行书写的，词语之间没有类似空格的区分标记，需要通过其他方法进行划分。

目前常见的中文分词方法可以分为三大类：基于字符串匹配的分词方法，基于理解的分词方法，以及基于统计的分词方法。随着大规模语料库的建立，统计机器学习方法不断研究和发展，基于统计的分词方法逐渐成为主流。

项目中采用结巴中文分词的精确模式进行分词，结巴中文分词的优势在于在一般情形下精确度高，支持用户自定义词典（绑定某些较长的词），对于未登录词使用 Viterbi 算法独立成词，由于疫情中出现了很多新词汇，这种对于未登录词的敏感性正是我们所需要的。

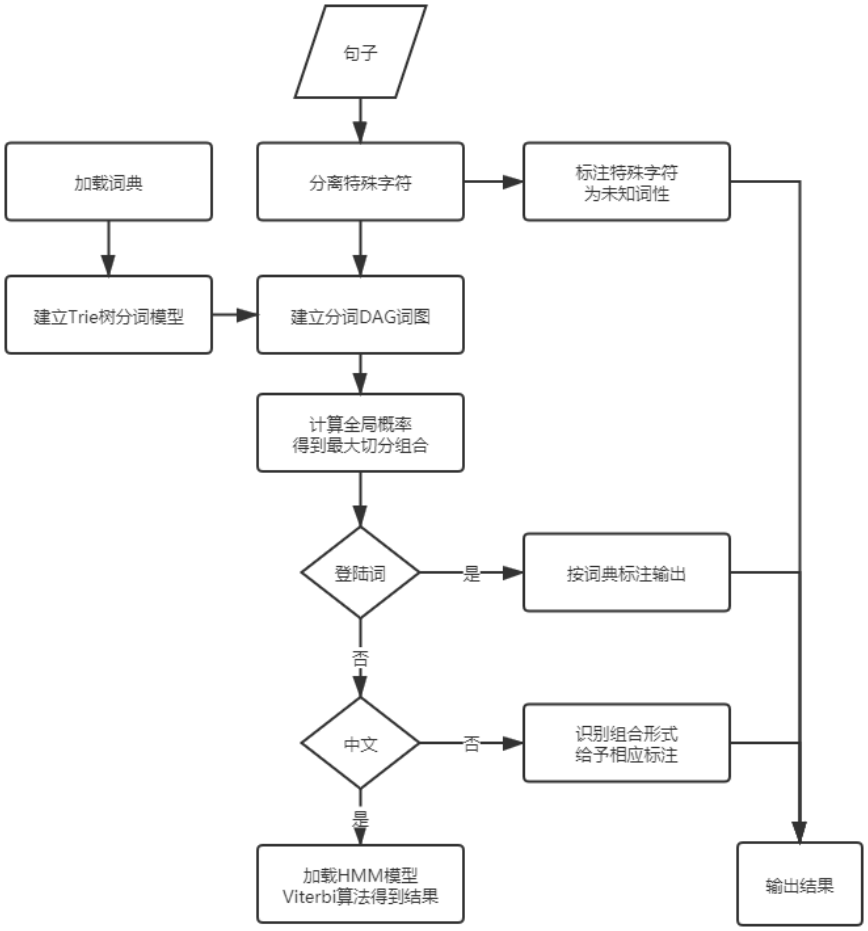


图 2-3 结巴分词流程示意图

## 2.4 机器学习相关技术简述

在项目的最后部分，尝试使用线性回归、贝叶斯曲线拟合、支持向量机（SVM, Support Vector Machines）等机器学习方法进行疫情曲线的拟合用于做疫情趋势的预测。

## 2.5 本章小结

本章介绍了本项目实施中所需要的主要技术及其运作基本原理，包括主题网络爬虫、结巴分词技术、词云可视化技术、Echarts 绘图可视化技术和 SVM 机器学习技术等。

# 第三章 数据收集模块的设计和实现

## 3.1 实际项目中的爬虫策略

### 3.1.1 基于每日上报页面的爬虫策略

疫情人数相关数据，可以在百度、腾讯等疫情每日上报页面中爬取，由于数据都显示在主页面上，编写代码模拟访问该网页遍历相关变量名就可以汇总成数据表格，不需要二级跳转。

```
#获取Json
China = GetHtmlText(Get_China)
City_Count_json = json.loads(China)
# 将json数据中的data字段的数据提取处理
City_Count_json = City_Count_json["data"]
# 将提取出的字符串转换为json数据
City_Count_json = json.loads(City_Count_json)
# 获取每日总信息
lastUpdateTime = City_Count_json["lastUpdateTime"]
chinaTotal_json = City_Count_json["chinaTotal"]
confirmCount = str(chinaTotal_json["confirm"])
suspectCount = str(chinaTotal_json["suspect"])
deadCount = str(chinaTotal_json["dead"])
cure = str(chinaTotal_json["heal"])
```

图 3-1 以中国为例的疫情数据获取爬虫代码

这里使用到 python 的 requests 库模拟浏览器向服务器发送请求获取数据，对 HTML 网页文本直接解析就可以得到目标数据。

### 3.1.2 基于搜索页面的爬虫策略

疫情期间的公众号文章来源不一，需要更为复杂的爬虫策略，我们采用主题网络爬虫和深层网络爬虫相结合的方式进行爬取。

首先，从源 URL 库进行爬虫数据量太大，数据相关性不强，且不同网站的反爬虫策略不同，不利于我们编写通用爬虫代码，为了获取足量的有效数据，本模块将采用基于微信公众平台搜索功能的网络爬虫：若已知主题关键词，可以试用搜狗的微信搜索功能接口；若要收集特定公众号的文章，可以使用公众号内部的搜索链接进行搜索。

通过浏览器的开发者工具，可以观察结果检索页面的网页结构，我们的一级目标是列表中文章 URL，再使用 xpath 获取文章 URL 的详细值（超链接值），在搜索结果的对应项中，我们不难找到相应的链接信息。

```
<a target="_blank" href="/link?url=dn9a_-
gY295K0Rci_xozVXfdMkSQTLW6cwJThYulHEtVjXrGTiVgS8d9Ptk05DKlGpIFNMy0JHdrmFvyZoA3fVqXa8Fpl
pd9D7_pmOxOB2_SWko6RG_KCVvhw5xI37UO0SXCYgC0wKXvxQe2VuT6eXhF3JqsvaDcwk-
WGrJh26xdLV_m1GH3nt26NwWw6s9e5Ol-8iNolMp8R9DzuL1DIVOw-
RpaniPG5AxJYRD84Wwwi0sHSFOubRjegfT2og4rp46-q-
ZgalmE8RsmObDbtQ..&amp;type=2&amp;query=%E7%96%AB%E6%83%85&amp;token=92C354FE3DEEF8E860
65C6E887966A10600D8F9E5ECD08F6" id="sogou_vr_11002601_title_0" uigs="article_title_0">
<em><!--red_beg-->疫情<!--red_end--></em>最新情况通报!</a>
```

图 3-2 搜狗搜索页面的文章 URL 信息

```
https://mp.weixin.qq.com/s?
src=11&timestamp=1590495478&ver=2362&signature=I5fWMcLs3flpUPMsxdDmJYyMrWEavZV7NtGphNuW
JemQIvQSsM*vnkXrgqrnpIXtFFIM-
S7c7W4ctsg3W3FUrkIBIbAXwIvsQfjtQLQByhoBqbTOFqdIcBYgeNUrOlwr&new=1
```

图 3-3 用于访问文章页面的真实 URL

通过对比可以发现两个 URL 不是同一 URL，抓包可以发现中间发生了一些跳转，这里搜狗搜索引擎做了一些反爬虫限制，如果我们补上 https://mp.weixin.qq.com 的域名访问图 3-2 中的 URL，会跳转到访问出错页面，指出这是一个异常访问请求。我们需要用到一些反反爬虫手段解决这个问题。

## 3.2 针对搜狗搜索引擎的反反爬虫策略

为防止站点被恶意访问，大多数站点会使用反爬虫策略侦测那些自动脚本的访问行为。搜狗搜索引擎使用的反爬虫策略主要如下几种：隐藏分页、包装页面 URL、验证用户 Cookie、封禁可疑 IP 地址，本节简要介绍不同策略的反爬虫原理和应对手段。

### 3.2.1 隐藏分页的反制策略

搜狗搜索引擎对未登录的用户限制访问内容量，只允许未登录用户访问检索结果的前十页内容。

由于 requests 库是模拟浏览器访问搜索引擎页面，他继承了你使用浏览器的用户状态（不包括 Cookie），因此我们只需要用个人账号在浏览器上保持登录状态，进行任意搜索即可。

### 3.2.2 包装真实 URL 的反制策略

包装真实页面 URL 是搜索引擎最常见的反爬虫策略，而且以 40-50 天为周期更换不同的包装方法，这种反扒策略成本较低且便于更换，可以有效拦截部分通配的自动爬虫脚本。

我们可以在正常搜索页面中访问目标结果页，使用 Fiddler 抓取数据包观察这种数据结构。请求参数中除了 URL 等常规参数，还包含 k 和 h 参数，这两个参数的构造方式作为 JavaScript 脚本就放置在 html 文件的末端。

```
<script>
    (function(){$("a").on("mousedown click contextmenu",function(){var
    b=Math.floor(100*Math.random())+1,a=this.href.indexOf("url="),c=this.href.indexOf("
    &k=");-1!==a&&-1===c&&
    (a=this.href.substr(a+4+parseInt("21")+b,1),this.href+="&k="+b+"&h="+a)})})();
</script>
```

图 3-4 搜索结果页中 URL 的包装方法

因此我们在爬虫时模拟这种构造方式生成对应的 k、h 值补充在 URL 末尾，通过黏贴到浏览器测试，我们正常访问到了文章页面，则说明构造出的 URL 被正确回应。

```
def get_k_h(url):
    b = int(random.random() * 100) + 1
    a = url.find("url=")
    url = "http://weixin.sogou.com" + url + "&k=" + str(b) + \
          "&h=" + url[a + 4 + 21 + b: a + 4 + 21 + b + 1]
    return url
```

图 3-5 构造真实 URL 的代码

### 3.2.3 验证用户 Cookie 的反制策略

如果不携带任何 Cookie 使用爬虫程序访问目标页面，会进入 antispider 页，被限制访问。

根据抓包可以发现，Cookie 的更新频率和 IP 置信度有关，在没有使用爬虫脚本的很长时间内，Cookies 不会发生变化，一旦单位时间内访问次数达到某个阈值则自动更换 Cookie，继续使用先前的 Cookie 访问则会进入上述验证页面。

Cookie 可以通过访问搜狗网站的其他页面逐步获得构建，且实质是一个有效 Cookie 池，获得有效 Cookie 的方法主要有如下几种：

1.未被检测的可靠 IP 访问搜索页面，通过抓包工具直接截取；

2.触发并手动通过反爬虫验证码页面，抓取全新的 Cookie；

```
1. 得到ABTEST、SNUID、IPLOC、SUID:
   https://weixin.sogou.com/weixin?
type=2&query=%E5%92%B8%E8%9B%8B%E8%B6%85%E4%BA%BA&ie=utf8&s_from=input&_sug_=n&_sug_type_=1
&w=01015002&oq=&ri=1&sourceid=sugg&sut=750912&sst0=1573092594229&lkt=0%2C0%2C0&p=40040108
2. 需要IPLOC、SNUID，得到SUID:
   https://www.sogou.com/sug/css/m3.min.v.7.css
3. 需要ABTEST、IPLOC、SNUID、SUID，得到JSESSIONID:
   https://weixin.sogou.com/websearch/wexinurlenc_sogou_profile.jsp
4. 需要IPLOC、SNUID、SUID，得到SUV
   https://pb.sogou.com/pv.gif
```

**图 3-6 获取 Cookie 的具体方法**

获取到合法 Cookie 值后，便可以根据真实 URL 值构建出请求头，访问目标页面。

### 3.2.4 封禁可疑 IP 地址的反制策略

相较于手动搜索的低频特征，爬虫的效率只会受限于所处终端的处理能力和带宽，服务器很容易将爬虫程序的高频访问定义为脚本行为，所使用的 IP 会被限制访问。使用 time.sleep()方法强制让程序休眠可以降低访问频率，但这样做会牺牲爬虫的效率，增加收集数据所需要的时间，使用 IP 代理池并开多线程访问目标网站可以绕过站点服务器对 IP 地址字段的检测，从而在反制反爬虫策略的同时加快爬取数据的效率。

爬取下来的 IP 代理通过 request 发出 get 请求，若返回码是 200 则说明代理可用，保留在可用代理池中，否则将代理从代理池中剔除。

若想有效提高代理的利用率，可以设置置信度系统，在每次访问失败时扣减置信度分数，挽留那些因为访问异常而被摒弃的可用 IP。

## 3.3 数据的具体内容和编排方法

疫情相关数据主要通过每日上报取得，分为累计/新增两个模块，每个模块分为确诊/重症/治愈/死亡四类数据，每类数据以国家（地区）和日期作为主键形成基准数据，根据项目的不同需要进行排列组合。

而微信文章数据则根据主题不同、时间段不同进行粗略划分，利用标题检索自动删除重复文章，根据阅读量排序剔除阅读量较低的文章，特别的将每日通报的推送文章从中去除，保证推送内容多元化，更具代表性。

微信文章通过结巴中文分词处理，经过词频统计程序生成不同主题、不同时间段的词频分布文件，为文本可视化做准备。

## 3.4 本章小结

数据收集模块使用了主题爬虫技术进行数据的爬取收集，使用一些反反爬虫策略针对搜索引擎的反爬虫机制，最后使用结巴中文分词处理文章得到词频数据，至此，数据集准备就绪，为可视化分析做好了数据基础。

# 第四章 疫情数据的可视化分析

## 4.1 疫情数据每日上报页面的构建

疫情的每日数据牵动着每个人的心，在每一份疫情数据的背后是一个个鲜活的生命，是每一位医疗工作者辛勤的努力，项目中我尝试利用官方给出的数据接口构建一个实时的疫情监测网页，清晰直观地观测疫情变化。

### 4.1.1 疫情实时图表的设计和实现

单项疫情实时数据是时间-人数二维数据，常使用基于位置的方法体现单个数据的差异，常用的有：柱形图/折线图/简单散点图，为表示在连续时间内疫情人数的变化，故采用折线图表示变化曲线，但疫情数据的汇总是类别（累计/新增 确诊/治愈/死亡/重症）-时间-人数三维数据。

由于疫情数据间的联系很紧密，且第三维度是分类字段数据，可以试用分类降维法将三维数据变为二维放在网页中展示，例如新增确诊/新增疑似可用于表示疫情发展的趋势，现有确诊/疑似/重症可用于表示疫情控制现状，全国累计治愈/死亡可以看到疫情治愈/死亡人数比，采用不同颜色多图叠加的方式展示。这样不仅有效节省页面空间，也能有效对比曲线变化趋势

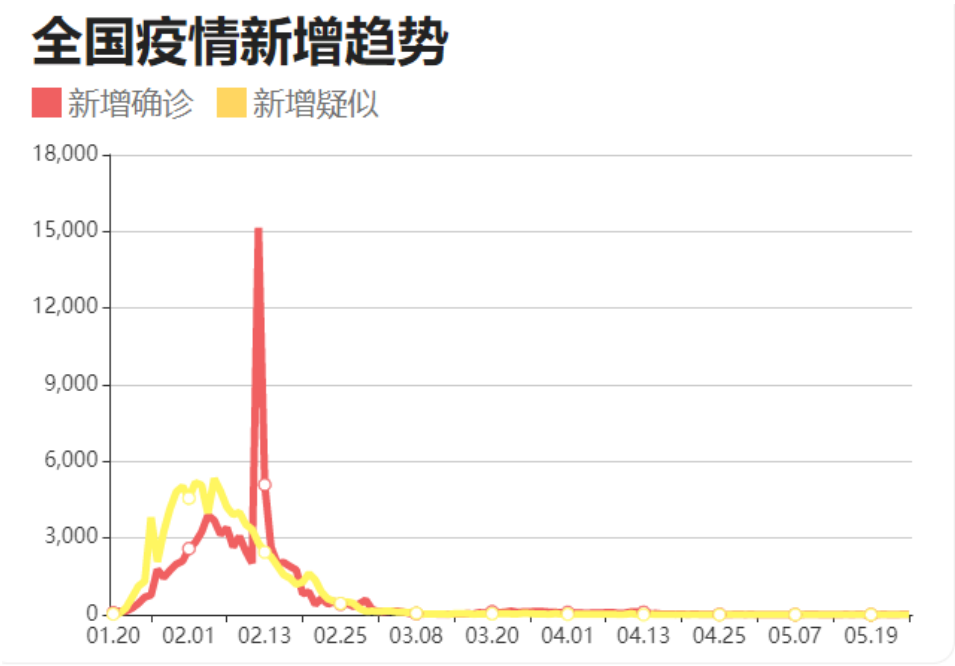


图 4-1 全国疫情新增趋势折线图

在可视化重要的一环交互性上，使用 JavaScript 的图交互技术，设定 trigger 为 axis 使得鼠标在图表上进行移动时可以根据横坐标值显示出当天的疫情数据。

```
let HADChartOption = {
    tooltip: {
        trigger: 'axis',
        formatter: function(params, ticket, callback) {
            var htmlStr = '';
            for (var i = 0,len=params.length;i<len;i++) {
                var param = params[i];
                var name = param.seriesName;
                var data= param.name;
                var value = param.value;
                var color = colorDic[name]; //图例颜色
                if (i === 0) {
                    htmlStr +='<div style="font-size:20px">'+data+'</div>';
                }
                htmlStr += '<div style="display:flex;align-items:center;font-size:20px">';
                //为了保证和原来的效果一样，这里自己实现了一个点的效果
                htmlStr +='<span style="margin-right:5px;display:block;width:20px;background-color:'+color+';"></span>';
                //圆点后面显示的文本
                htmlStr += name + ':' + value + '人';
                htmlStr += '</div>';
            }
            return htmlStr;
        }
```

图 4-2 实现光标移动与疫情图表交互的 JS 代码

对于进入四月以来，疫情得到了有效控制，新增折线趋于平缓，主要增长从内部扩散转为外部输入，且增量变小，同峰值相比数值过小，因此在原先的数据模块中新增境外输入窗口，输入爬取的（一维）境外输入人数数据跟踪外来输入变化。

### 4.1.2 疫情地图的设计和实现

中国的疫情在湖北武汉爆发，湖北省一时间成为新冠病毒的扩散源，中国各区域开始了隔离政策阻断病毒的传播。区域的概念在疫情的可视化分析中就显得尤为重要，在可视化过程中加入地理维度，同时融入具体的视觉表象特征和人的视觉特点将空间的上下文信息展现出来，就会使得数据的可视化效果更好，便于国家针对不同区域的感染状况给予不同程度的人力救助和物资支持。

这种方法在疫情防控的过程中并不少见，比如追踪病患的行动路线寻找可能的感染患者时就会用到基于线的活动轨迹图，据此有效阻止病毒的进一步扩散。


图 4-3 美国一空乘乘客的活动轨迹图

而疫情防控中的病患人数数据则符合地图的区域化特征，地图本身就是区域化特征明显的可视化元素，将地理数据可视化与 GIS（地理信息系统）相结合使用，通过 Echarts 和 D3.js 作为主要绘图工具来将疫情数据和中国各省份地图聚合绘制疫情热力地图。

Echarts 技术是百度前端团队开发的一个 JS 图表库，具有高性能的绘图功能，可以通过相对简单的代码调用绘制中国各省市地图。各省市绑定其疫情数据，并决定其在地图上色块的颜色深浅，体现该城市疫情的严重程度。

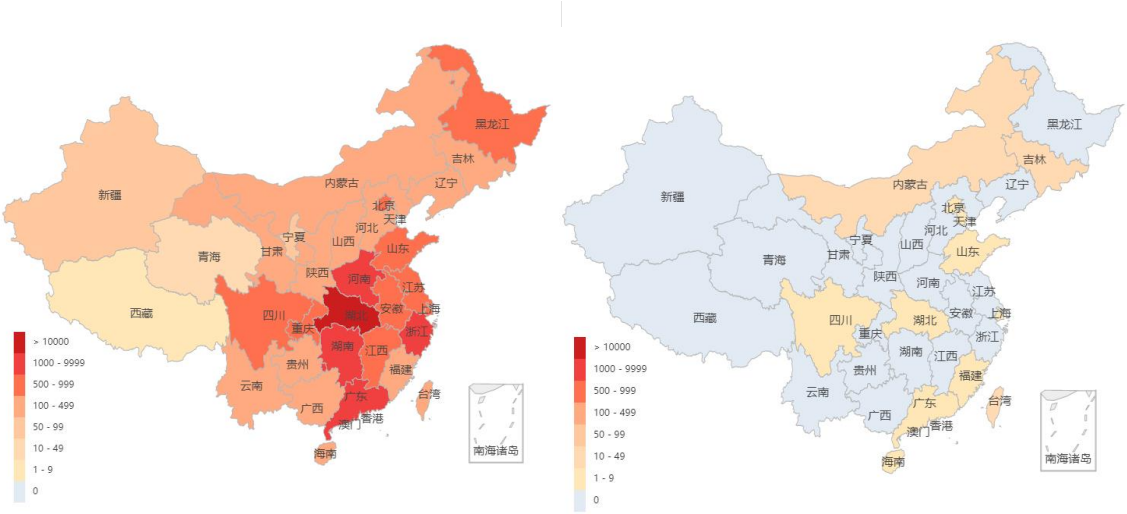


图 4-4 中国累计确诊/现有确诊疫情地图

通过给不同省市的区块绑定点击事件可以进入省市单独的疫情分页，通过同样的经纬定位方式可以实现更为详尽的可交互省市疫情地图。

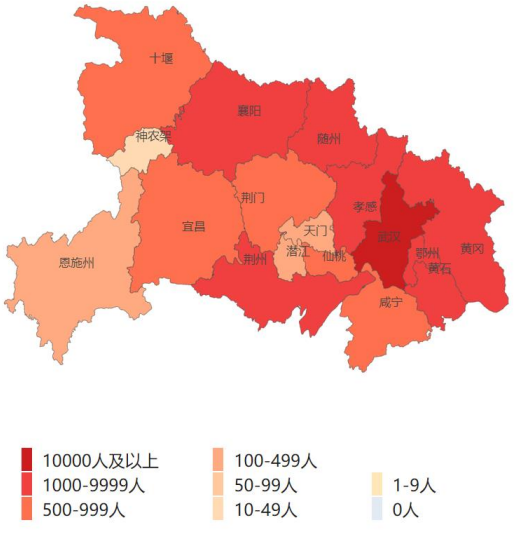


图 4-5 湖北累计确诊疫情地图

```
"properties": {
    "cp": [118.175393, 39.635113],
    "name": "唐山",
    "childNum": 15
```

图 4-6 湖北唐山地图位置绑定 JS 代码

## 4.2 疫情数据可视化中的信息提取

将疫情数据可视化处理后，我们尝试做一些信息处理并与一些已知结论相验证。

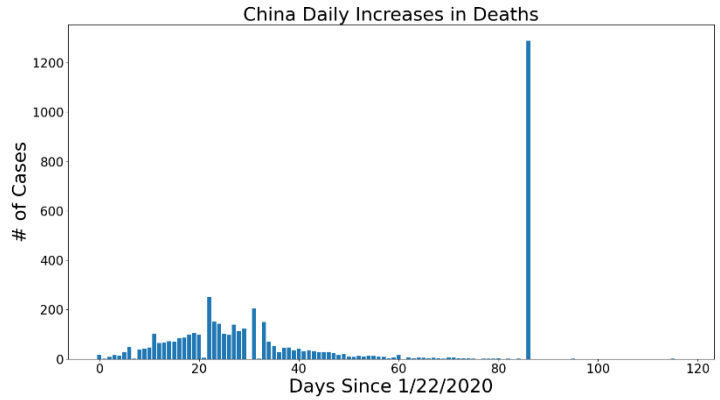


图 4-7 中国每日新增确诊人数

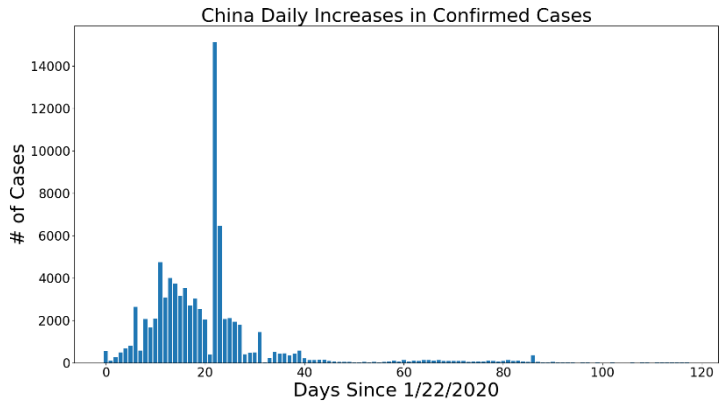


图 4-8 中国每日新增死亡人数

如图 4-7 和图 4-8 中所示，中国每日新增确诊人数和每日新增死亡人数分别经历了一次数据上的跳变，这是武汉先后在 2 月 20 日和 4 月 17 日两次修订疫情相关数据导致的，实际上确诊人数、死亡人数严格意义上不是当日新增，而是由于审定程序更加严谨、统计标准更加严格所产生的数据骤变，数据分析时应尽量减少异常数据带来的影响。

根据《一类潜伏期和染病期均传染的流行病模型》一文中给出的 SEIR 模型：

$$\begin{cases} \dfrac{dS}{dt} = -\beta IS, & \dfrac{dE}{dt} = \beta IS - (\alpha + \gamma_1)E, \\ \dfrac{dI}{dt} = \alpha E - \gamma_2 I, & \dfrac{dR}{dt} = \gamma_1 E + \gamma_2 I. \end{cases}$$

**图** 4-9 SEIR **模型**（Hethcote 2000）

S 代表易感人群数（近似为当地人口数），E 代表感染后处于潜伏期的人群，I 代表感染人群，R 代表已有该疾病免疫力的人群，t 代表第一例感染者出现之后的天数（12 月 1 日可近似作为感染起点），β 代表患者一天感染的人数，α 表示潜伏期转变为患者的速率，则 1/α 表示潜伏期，γ1、γ2 分别代表潜伏期恢复率和患者恢复率（终止传染可能率，即死亡也算在内）。

重点考察新冠肺炎的传染系数 β0，即目标人群缺少相应免疫力的前提下，一个感染者会传播疾病感染人数的平均数，β0>1 时传染病就有发展成地区病的风险，当 β0<1 时，传染病的传染性会逐渐衰弱。β0 越大，控制该疾病的难度就越大。

假设疫情不受控制，持续扩散，那么每日新增人数应当是一个峰值周期出现且不断升高的函数，直到人体的自然免疫建立速度与病毒变异速度达到平衡，每两个峰值之间的距离是受感染者的平均潜伏期，由图中数据疫情的平均潜伏期应当在 14 天左右，根据相关的研究显示，中国境内截至 2 月 13 日新冠病毒的平均病程为 10.5 天，易感人群几乎涵盖所有年龄段，据此来推算制定国家或地区的传染系数 β0。

截止 2020 年 1 月 23 日新冠肺炎的传染系数（主要在湖北境内）高达 5.75，至 2 月 13 日武汉当地的传染系数成功下降到 2.5，武汉以外的传染系数降低至 1.5，远低于武汉，疫情得到了有效遏制。

而反观疫情爆发以来美国、英国等国家的数据，则情况不容乐观，截止本论文书写前，英国的新增确诊人数开始下降，新增死亡人数有所下降，但仍有相当一部分疑似病例；而美国的新增确诊人数仍保持在 β0 水平附近，新增死亡人数有所下降，若不采取任何隔离措施，那么平衡点的到来或将遥遥无期。

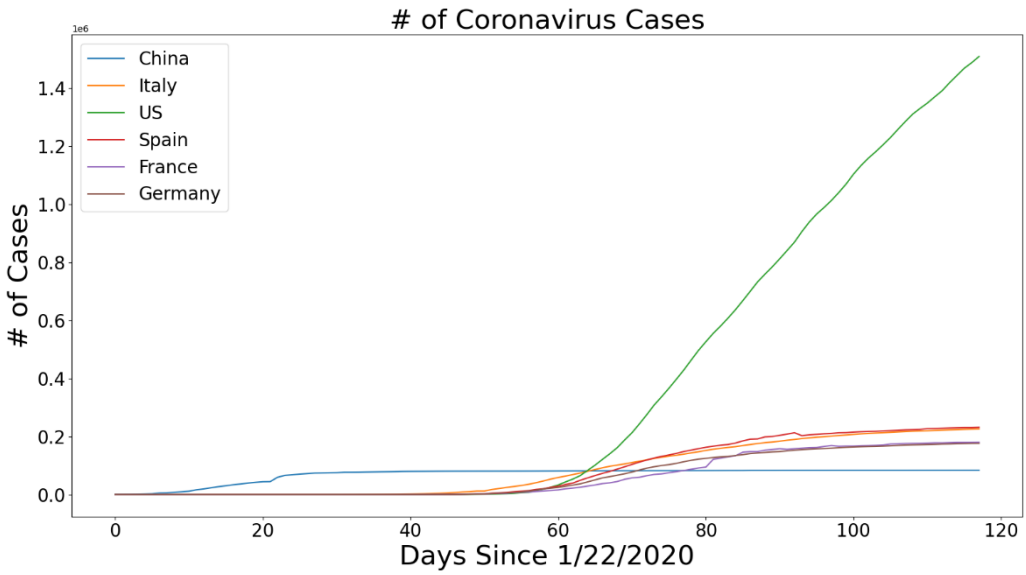


图 4-10 全球新冠肺炎确诊人数曲线图

如图 4-11 所示，美国也成为世界首个确诊新冠肺炎人数超百万的国家，爆发新冠肺炎确诊重灾区也从中国湖北逐渐迁移到美国纽约，湖北目前排在确诊新冠肺炎人数地区的第三位。

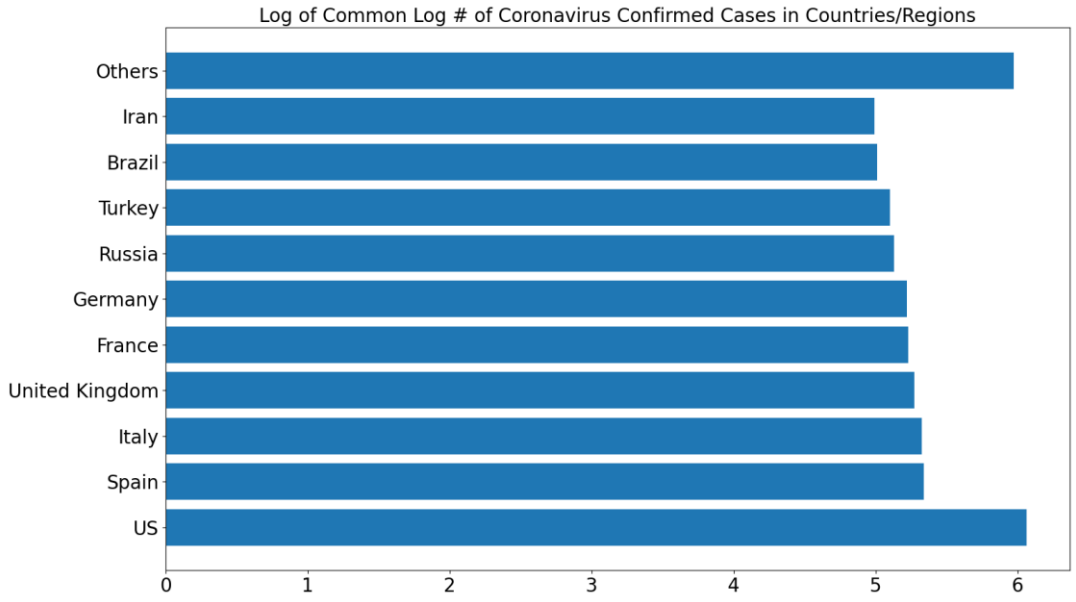


图 4-11 世界各国新冠肺炎确诊人数（取 lg 对数）

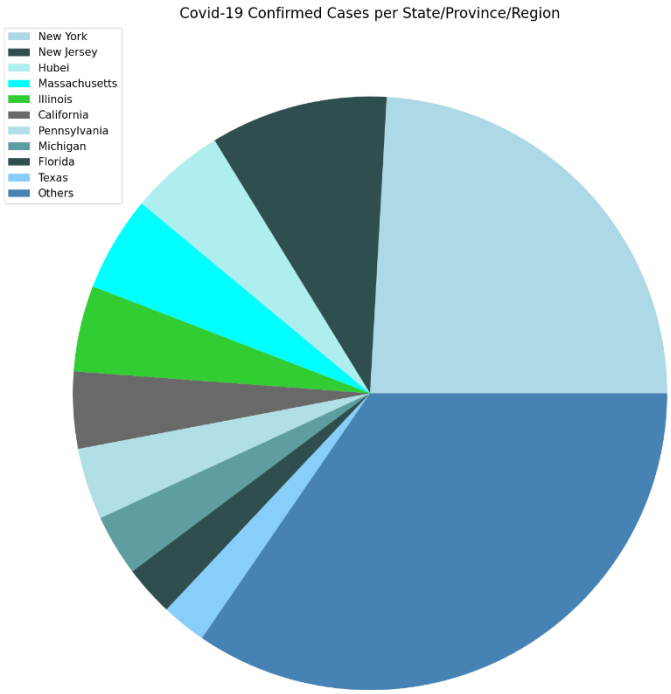


图 4-12 世界各省新冠肺炎确诊人数饼状图

从全球视角上看，自 3 月份中国国内疫情好转之后，世界范围内迎来了第二波疫情高峰，死亡率突破先前的最高值，而康复率出现了跌落反复的情况，目前死亡率还保持在 6%到 7%之间居高不下，但是康复率已经再次突破中位数回到 40%，可以预见在有限的时间内世界上大部分地区将战胜疫情，重返正常的工作生活秩序。

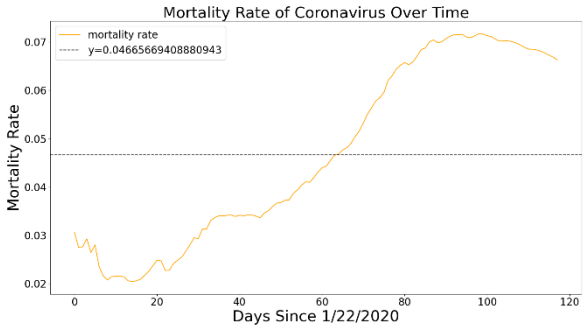


图 4-13 世界范围内的新冠疫情死亡率

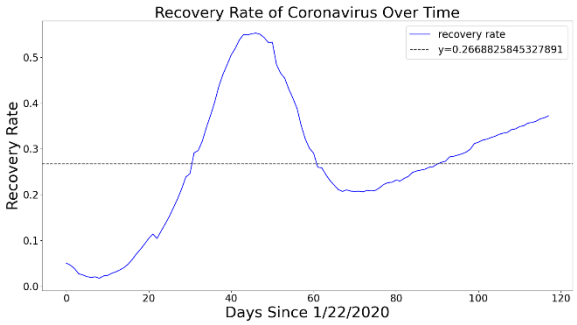


图 4-14 世界范围内的新冠疫情康复率

## 4.3 疫情文章的热点可视化

本项目中关于疫情文本的可视化处理主要指基于文本内容的可视化。

文本可视化的主要内容是如何快速获取文本内容的重点，考察词频和词汇分布两个主要特征。数据主要分为如下两个来源：

1. 爬取微信公众号“新华社”自 1 月 22 日以来的所有文章；

2. 根据步骤 1 中获取的文章进行词频分析，整理出关键词集合进行搜狗微信搜索，爬取前 40 页的内容。

对步骤一中的文章进行初次清洗，因为每日通报的内容已经被包含于疫情数据中，所以将他们从文章集合里去除；通过建立以标题为主键的数据库自动筛除步骤 2 与步骤 1 的重复条目。

采用词云的方式呈现词频分布图，文字间的空隙可以得到充分的利用，可视化结果更加美观。

```
# 生成词云
def create_word_cloud(frequencies, font_path, mask_image):
    mask = plt.imread(mask_image)
    wc = WordCloud(
        font_path=font_path,
        max_words=100,
        width=2000,
        height=1200,
        background_color="white",
        mask=mask,
    )
    word_cloud = wc.generate_from_frequencies(frequencies)
    word_cloud.recolor(color_func=ImageColorGenerator(mask))
    # 写词云图片
    word_cloud.to_file("wordcloud.jpg")
```

图 4-15 **生成词云的部分代码**


图 4-16 **疫情文章词云图**（1.22-4.30）

分时间段整理词云图，1 月热词是“武汉”、“疫情”、“确诊”、“感染”，2 月新晋热词分别是“钟南山”，“医疗队”，“口罩”，“疫苗”，“延迟”，“复工”，“防控”，3 月之后频繁出现了“隔离”，“工作”，“清零”，“美国”等词。由于“疫情”、“感染”等词出现频次较高，我们在生成最后的词云图时适当削减了他们所占的比重，使得更多相对较低频率的热词可以在词云图上呈现出来。

## 4.4 本章小结

本章主要介绍了可视化的具体方法，包括使用各式图表呈现疫情数据，用词云图表现疫情相关文章热词等，展示了自行编写的疫情上报页面，并结合了对当前疫情数据的汇总分析得出一些基础结论。在下一章中，将尝试对数据进行更深一层的分析。

# 第五章 疫情预测的相关尝试

## 5.1 利用新冠疫情数据进行模型建构

人工神经网络进行曲线拟合和数学模型的建构，是传染病预测数学模型研究中的重要一环，根据现有的传染病数据进行模型建构可以为之后的突发传染病提供数据参考和预测支持。

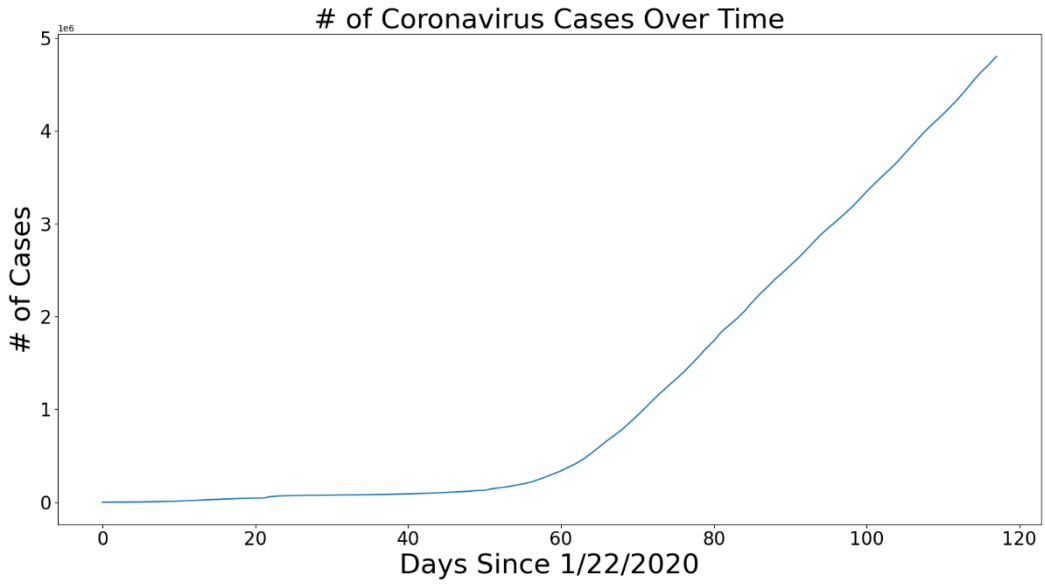


图 5-1 全球新冠肺炎确诊人数图（自 1.22 日以来数据）

编写代码的时间是 4 月初，选择世界范围内的确诊人数作为训练样本，建构出多项式回归、贝叶斯曲线拟合、支持向量机三种不同的线性模型，在撰写论文的五月初做预测拟合的尝试，即采用前 80 天作为训练集，预测未来 40 天的情况。

```
# 寻找参数部分,为SVR寻找最佳参数
c = [0.01, 0.1, 1]
gamma = [0.01, 0.1, 1]
epsilon = [0.01, 0.1, 1]
shrinking = [True, False]
degree = [3, 4, 5]

svm_grid = {'C': c, 'gamma': gamma, 'epsilon': epsilon, 'shrinking': shrinking, 'degree': degree}

svm = SVR(kernel='poly')
svm_search = RandomizedSearchCV(svm, svm_grid, scoring='neg_mean_squared_error',
                                cv=3, return_train_score=True, n_jobs=-1, n_iter=30, verbose=1)
svm_search.fit(X_train_confirmed, y_train_confirmed)

svm_confirmed = svm_search.best_estimator_
```

图 5-2 寻找支持向量机模型参数的相关代码

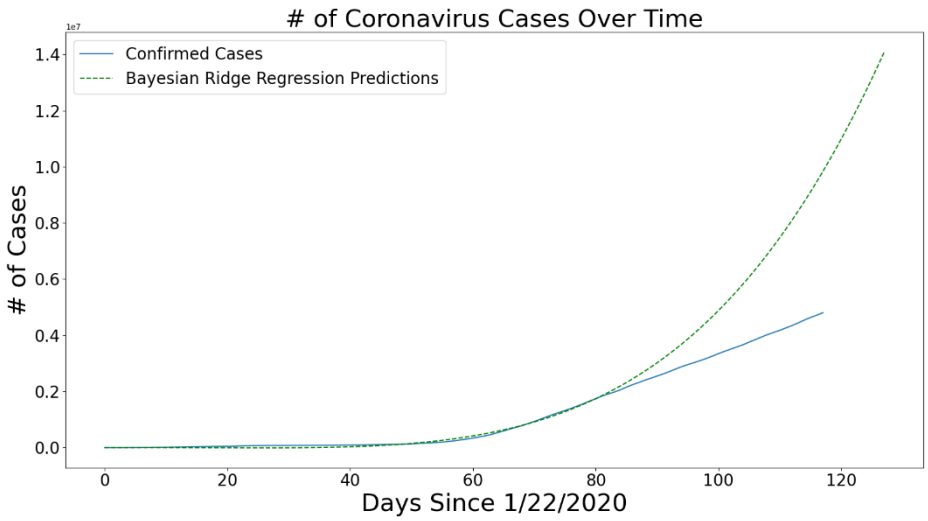


图 5-3 Bayesian 模型拟合全球新冠肺炎确诊人数图

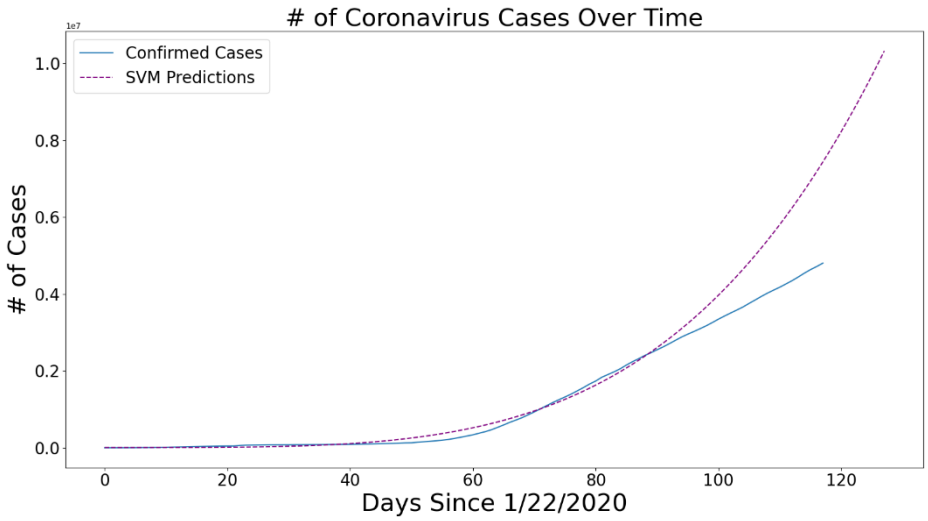


图 5-4 支持向量机模型拟合全球新冠肺炎确诊人数图

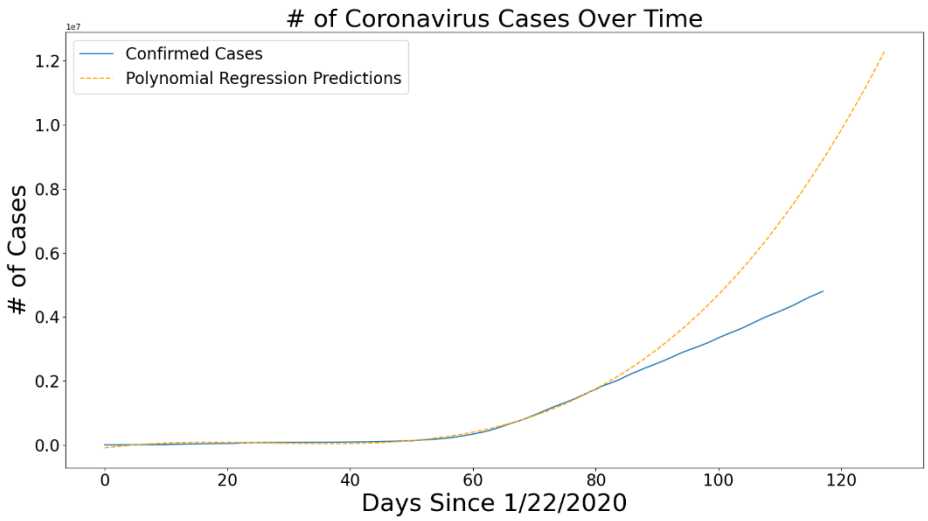


图 5-5 多项式回归模型拟合全球新冠肺炎确诊人数图

三种模型共同特点是在预测后期倾向于将数据放大，都远超出实际值，其中拟合效果较好是支持向量机，平均绝对误差（MAE, Mean Absolute Error）为 $0.84*10^{6}$，均方误差（MSE, Mean Squared Error）是 $1.36*10^{12}$。分析其原因，主要是截止训练集数据收集的日期 4/13/2020 之前，全球新冠疫情正处于爆发期，每日新增的确诊人数正不断升高，曲线拟合过程中如果训练集中的斜率一直处于爬升的状态，二阶导数没有下降，那么之后的曲线预测也会让斜率不断爬升，从而导致误差过大。

根据《传染病疫情早期预警的主要模型》中的理论知识，截止疫情开始（1/22/2020）60 天之前，疫情在全球范围内都处于预警阶段，有相当一部分的感染者体内的病毒处于潜伏期，没有达到确诊人数指数级增长的爆发期，如果在这个阶段及时采用行动轨迹跟踪、区域隔离的手段发现疑似病例，那么病毒是可以被预防和防止扩散的。

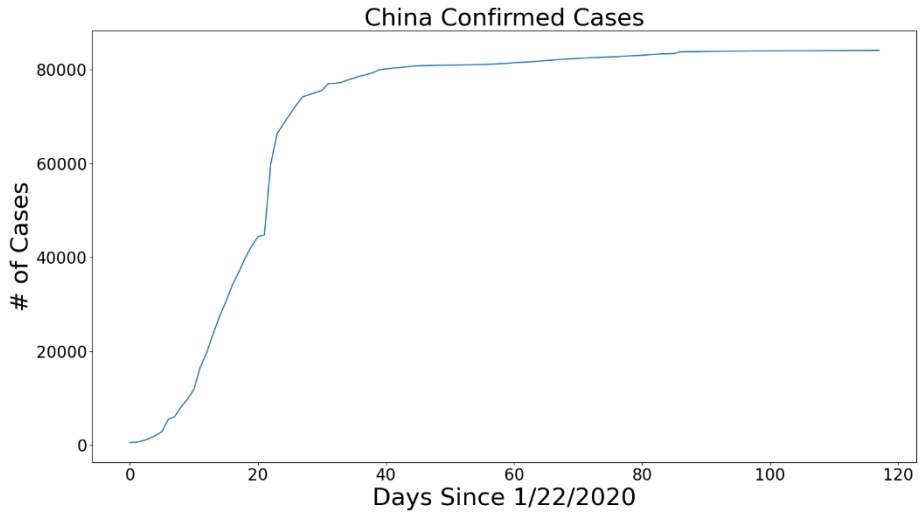


**图 5-6 中国境内新冠肺炎确诊人数图（自 1.22 日以来数据）**

反观中国国内的数据，疫情实际上是在 12 月中下旬于武汉爆发，但 1 月 22 日之前的数据不可考证，所以数据收集是从 1 月 22 日开始，疫情的潜伏期在图上没有明显体现，而是直接迎来爆发期，在接下来的 30 天时间内，确诊人数飙升，经过全国人民的不懈努力，疫情得到了控制，进入平缓期。

## 5.2 利用 SARS 疫情期间数据进行预测的猜想

### 5.2.1 选取 SARS 疫情数据进行预测的理论依据

根据上一节中误差来源的分析，传染病的数据是分段变化的，主要分为：潜伏期、爆发期和平缓期，变化时间节点和诸多因素有关，客观因素有：地域、气候、季

节、病毒自身特性等，主观因素有：医疗条件、防疫策略、物资投入等，如果想要预测一个传染病的传染趋势，最好的方法是考虑过去发生的、客观因素相近的传染病，利用它从爆发到平息全过程的数据进行宏观预测，并根据客观因素的不同进行参数微调，那么我们把目光放在 2003 年同期发生的非典型性肺炎 SARS 病毒上。

这部分的代码和工作主要基于网络资源和已有的针对 SARS 的研究，因此仅作为本次项目的一个猜想补充部分，具体论述说明请见论文“Novel coronavirus 2019-nCoV: early estimation of epidemiological parameters and epidemic predictions”。

### 5.2.2 SARS 数据预测新冠疫情的方法

SARS 的数据集是从 2003 年的 3 月 17 日截止至当年的 5 月 30 日。

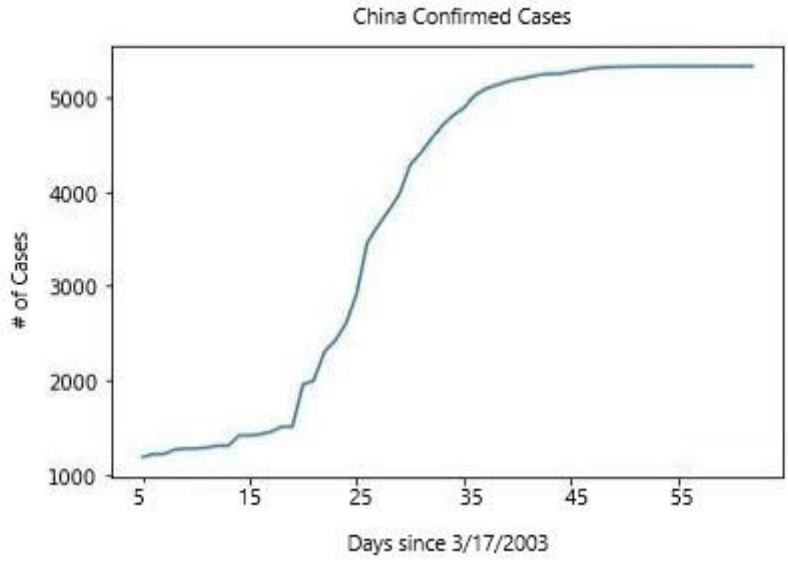


**图 5-7 中国大陆 SARS 非典型性肺炎确诊人数图**（3/17/2003-5/30/2003）*

疫情发展的三个阶段应该有三种数据模型来建模，将所有数据根据斜率变化分为三个阶段，在训练当前阶段数据的模型时，则设置当前阶段数据占比为 0.7，其他两个阶段各为 0.15。衡量模型预测准确度的方法是计算该模型的误差函数 Δ，即当前确诊病例和预测确诊人数的比率。

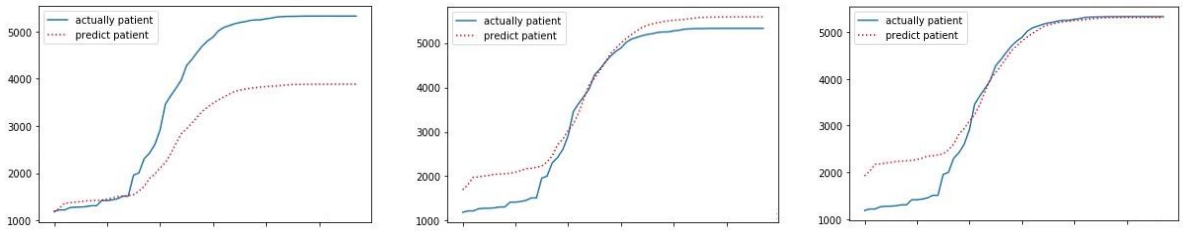


**图 5-8 中国大陆 SARS 非典型性肺炎预测模型（潜伏期-爆发期-平缓期）**

在潜伏期模型的预测中，爆发期和平缓期数据比较小，而在平缓期和爆发期的模型预测中，确诊人数的起点明显高于实际起点。这符合我们对分时期模型的效果预期。我们从潜伏期模型开始拟合新冠疫情数据，当误差函数Δ高于某个值时我们更换为爆发期模型进行拟合，同理拟合进入平缓期的情况，这个两个衔接点的阈值Δk是比较难判断的，在SARS训练集上，潜伏期-爆发期的Δk是0.20，爆发期-平缓期的Δk是0.14。则我们在误差分别达到这两个阈值时更换拟合模型。

### 5.2.3 SARS数据预测新冠疫情的误差分析

在使用该模型预测国内新冠疫情的过程中，爆发期和平缓期都提早到来，且预测确诊人数中位数远低于实际确诊人数中位数，实际误差远高于训练验证误差。

分析误差原因主要有如下几个方面：

1. 非典的严重程度不及新冠病毒肺炎，新冠确诊人数将远高于SARS确诊人数；
2. 新冠疫情的数据采集点在1/22/2020，而疫情的潜伏期发生在这之前，所以数据集不能很好覆盖潜伏期的模型，进而累加整个模型的误差；
3. SARS病毒的转折点发生在20/40天左右，这是根据病毒自身的特性和抗疫措施、医疗条件等人为因素决定的，而即使是对于略过了潜伏期的新冠肺炎病毒来说，20/40天的节点疫情仍处于爆发期，这也就导致模型预见性比较差，不能及时反映出转折点的出现时机。

### 5.2.4 SARS数据预测新冠疫情模型的改进猜想

关于如何提高模型准确性，要点在于时期转折点的判断和确诊人数中位数的大致范围。

首先是病毒的潜伏期长度可以依赖医学实验来判断，例如非典的潜伏期在1-16天左右，最常见的是3-5天，而现在处不得到的新冠肺炎病毒的潜伏期在1-24天，最常见的是2-7天，中位数是3.0天，但病毒变异性强，尚具有变化的可能，而非典疫情图显示的潜伏期为20天，是非典病毒潜伏期最长长度的1.25倍，我们或可据此来预测新冠疫情的潜伏期长度，并倒推回疫情最初发生的时间。

其次是确诊人数的中位数，这个与病毒的感染效率密切相关，传染系数与该中位数成正相关，医疗效率与该中位数成负相关，疫苗研制进度可作为常数项直接有效地避免大规模感染。据此适当调整模型，或许加强模型对不同疾病的适应能力。

## 5.3 本章小结

本章使用机器学习的方法，尝试预测新冠疫情的趋势，并使用最新相关研究的思想利用 SARS 病毒的数据预测新冠疫情，给出该方法的缺陷和改进措施。

# 第六章 总结和展望

## 6.1 全文总结

针对当下时事热点新冠疫情展开研究，利用网络爬虫获取相关的数据，同时利用多种可视化手段展现这些数据，构建了自己的疫情数据网页、词云图和数据图表集合。利用这些数据结合机器学习的方法进行了疫情分析和预测，得出相关结论，并对最后一个 SARS 病毒构建模型预测新冠疫情的环节提出了改进的猜想。该项目使得疫情数据更加清晰直观，并且尝试深挖藏在数据背后的巨大价值，为我国的抗疫事业贡献一份力量。

## 6.2 展望

不管是多大的困难，都终将会过去，我们需要在一次又一次的挫折中汲取宝贵的经验，SARS 病毒、手足口病、新冠肺炎，有多少疾病打击我们，就有多少生命的奇迹在这片土地上发生，如果能建立并不断完善一个展示疫情数据的可视化系统，不仅可以加速疫苗的生产、医疗资源的合理配置，还可以辅助高效的流行病预测模型的开发，最后达到控制传染病传播的效果。在不远的将来，我们一定有理由相信没有任何一种病毒可以让人类陷入困境。让我们一起加油。

# 参考文献

[1]张龙浩,李柏宏,贾鹏,蒲剑,白蓓,李音,朱培嘉,李雷,曾国军,赵欣,董珊珊,刘梦菡,张楠.新型冠状病毒（SARS-CoV-2）全球研究现状分析[J].生物医学工程学杂志,2020,37(02):236-249.

[2]孙立伟,何国辉,吴礼发.网络爬虫技术的研究[J].电脑知识与技术,2010,6(15):4112-4115.

[3]于娟,刘强.主题网络爬虫研究综述[J].计算机工程与科学,2015,37(02):231-237.

[4]唐家渝,刘知远,孙茂松.文本可视化研究综述[J].计算机辅助设计与图形学学报,2013,25(03):273-285.

[5]任磊,杜一,马帅,张小龙,戴国忠.大数据可视分析综述[J].软件学报,2014,25(09):1909-1936.

[6]龙树全,赵正文,唐华.中文分词算法概述[J].电脑知识与技术,2009,5(10):2605-2607.

[7]丁世飞,齐丙娟,谭红艳.支持向量机理论与算法研究综述[J].电子科技大学学报,2011,40(01):2-10.

[8]余豪士,匡芳君.基于 Python 的反反爬虫技术分析与应用[J].智能计算机与应用,2018,8(04):112-115.

[9]易泽顺. 基于 Web 的数据可视化工具设计与实现[D].华中师范大学,2017.

[10]赵序茅,李欣海,聂常虹.基于大数据回溯新冠肺炎的扩散趋势及中国对疫情的控制研究[J].中国科学院院刊,2020,35(03):248-255.

[11]原三领,韩丽涛,马知恩.一类潜伏期和染病期均传染的流行病模型[J].生物数学学报,2001(04):392-398.

[12]宋贽,陶桂洪,张阚.埃博拉病毒病的传播模型及其防控仿真[J].大连工业大学学报,2017,36(03):223-226.

[13]安薇,夏飞,陈敏,杨萍,方莎莎,廖亚玲,许鑫,周琴,李旷宇,张明伟.新型冠状病毒肺炎死亡患者 11 例临床特征分析[J].实用医学杂志,2020,36(09):1125-1130.

[14]王小莉,王全意,栾荣生,曾大军,贺雄.传染病疫情早期预警的主要模型[J].现代预防医学,2008(22):4339-4341.

[15]王丙刚,曲波,郭海强,张蕾,金鑫,李刚,孙高.传染病预测的数学模型研究[J].中国卫生统计,2007(05):536-540.

[16] Jonathan M Read, Jessica RE Bridgen, Derek AT Cummings, Antonia Ho, Chris P Jewell. Novel coronavirus 2019-nCoV: early estimation of epidemiological parameters and epidemic predictions[J]. medRxiv, https://doi.org/10.1101/2020.01.23.20018549

# 致谢

首先谨向尊敬的导师徐迎晓老师致以崇高的敬意和诚挚的谢意。本次论文写作从立题开始，老师就再三叮咛立意不可过深过大，要细于研学，勤于钻研，要站在前人肩膀上提升深度，使用自己的技术耕耘开发，希望这篇文章可以不负所望。

感谢在复旦大学求学期间授予我知识的所有老师，悉心支持、帮助我的辅导员李导，和所有陪伴我四年时光的同学们，这四年有笑有泪，有喜有悲，如此精彩，如此难忘，均是因为有你，谢谢你们。

感谢我的父母和长辈，他们照顾我的饮食起居，养育我成人，学校教我读书，你们教我做人，如果我今后的人生有幸取得零星的成就，那里面都将饱含你们的殷切教诲，远行客的心中涤荡着的不止梦想，还有你们的牵挂。

感谢我的好友邵喆宁同学为本文的英文摘要差缺改错，不间断“催更”式的鼓励支持，谢谢你。

最后，向在本次新冠抗疫事业中无私奉献的每一个工作者致敬！