

Report Date

Report Date	28/02/2020	Name	Yiming Zhang
-------------	------------	------	--------------

Period covered by this report

Start Date	21/02/2020	End Date	28/02/2020
------------	------------	----------	------------

一、本周研究内容

研究内容	本周的研究内容主要为研究计划中的 研究内容(一) ，即“通过网络爬虫等技术对传染病(新冠肺炎，H1N1, SARS)等关键信息进行大数据搜集。爬虫技术拟使用 python 来实现数据搜集。研究所使用的大数据获取渠道包括:世界卫生组织(WHO), 国家疾病预防控制中心，省级、市级的疾病预防控制中心，专业的医疗期刊杂志等，预计可能共需要研究 数百万篇相关文章来满足训练深度学习模型的需要。”
------	---

二、项目实施当前状态

项目进度实施情况	研究计划中的 研究内容(一) 已经完成一半，预计下周完成所有 研究内容(一) 的内容。
项目整体进度完成情况	研究计划中的研究内容总共有四部分，预计 第一个月 内完成研究内容(一)和研究内容(二)，即对大数据进行爬取，数据库的建立以及数据清洗和数据集成。

三、本周成果

本周成果	<p>本周研究的主要成果分为以下三点。</p> <p>1. 对研究所使用的数据来源进行确定，目前确定的数据来源(Data Source)为：</p> <ul style="list-style-type: none"> 世界卫生组织 (WHO) : https://www.who.int/ 中华人民共和国国家卫生健康委员会 (NHC) : http://www.nhc.gov.cn/wjw/xwdt/list.shtml 中国疾病预防控制中心 (CCDC) : http://weekly.chinacdc.cn/en/zcustom/currentVolume/1 中国香港政府新闻网 : https://www.news.gov.hk/eng/categories/health/index.html 新英格兰医学杂志 (NEJM) : https://www.nejm.org/ 柳叶刀 (Lancet) : https://www.thelancet.com/ 美国医学会杂志 (JAMA) : http://jama.ama-assn.org/ 英国医学期刊 (BMJ) : http://www.bmj.com/
------	--

	<p>2. 对研究所使用的网络爬虫技术进行确定，尝试了 Requests 库, BeautifulSoup 库和 Scrapy 框架，目前决定使用 Scrapy 框架进行爬虫的开发 (https://scrapy.org/)。</p> <p>3. 使用 Scrapy 框架写了爬虫代码 (后台运行截图参见 Appendix 图一)，已经成功爬取香港政府新闻网上的新闻内容。</p>
--	--

四、上周问题解决情况（2020.02.28 更新）

<p>2020 年 2 月 28 日下午 2 点微信电话会议对问题讨论结果：</p> <ol style="list-style-type: none"> 1. 按照当前计划继续开展项目。 2. 爬取网站公开披露信息属于合法行为，不涉及数据隐私等问题。 3. 补充了百度舆情使用费用 (一年 30000 元) 以及解析不同爬取后的网站数据，需要修改相应的代码，因此工作量有所增加。 <p>2020 年 2 月 28 日晚上 8 点半微信电话会议讨论结果：</p> <ol style="list-style-type: none"> 4. 在数据来源上，增加 2 个国内医学杂志。 5. 在百度舆情的使用上，详细列出 10 个关键词。
--

五、项目当前可能出现的问题

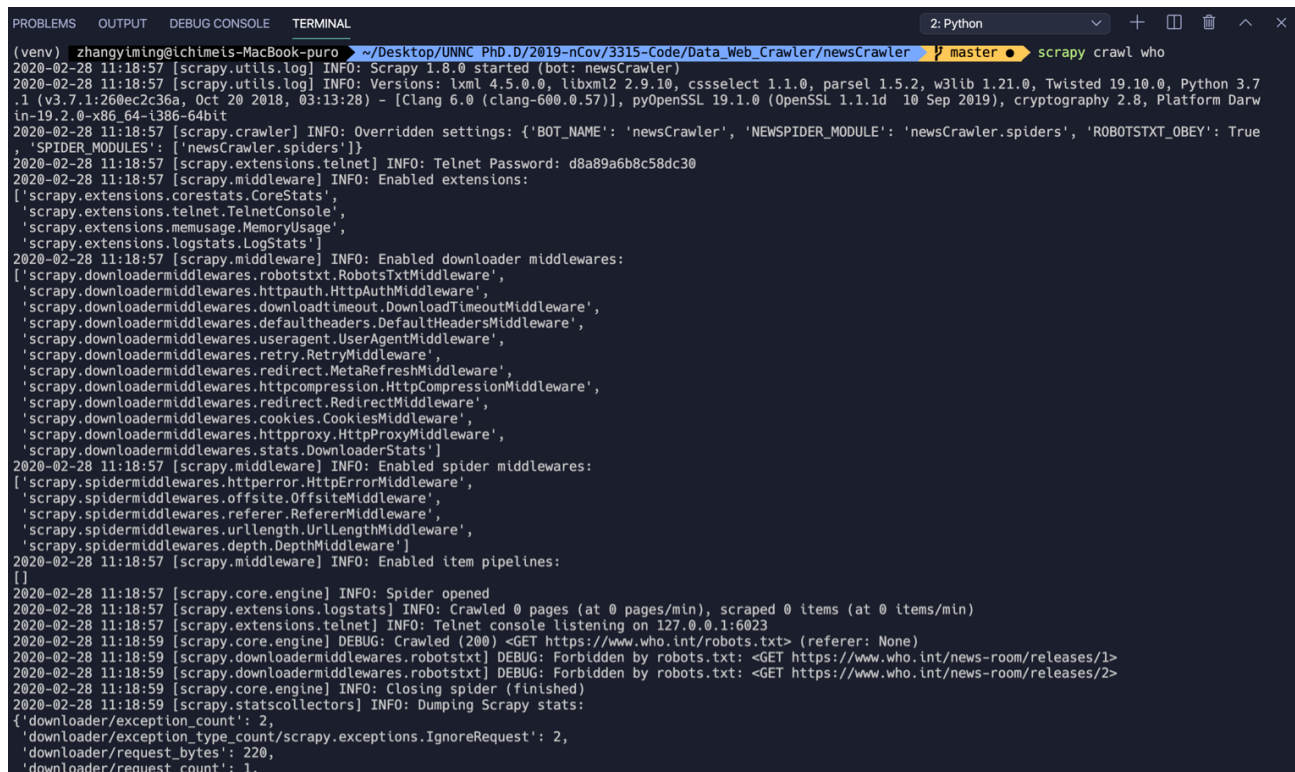
具体问题描述	<p>本周发现的主要问题：</p> <ol style="list-style-type: none"> 1. 有些网站有反爬虫的保护机制，无法成功爬取网页链接（Appendix 图二所示） 2. 在利用爬虫技术去爬取医学杂志信息，以及爬取国家卫健委公开信息等用于科研目的，是否需要向杂志及卫健委等部门争得同意，在法律方面是否完全合法，以及是否需要签订相关数据保密协议。
--------	---

<p>解决方案</p>	<ol style="list-style-type: none"> 1. 研究一下如何能通过反爬虫机制。 2. 看是否有相关网站提供 API 接口，提供数据服务，如 Appendix 图四的百度舆情 API（https://cloud.baidu.com/product/byapi.html），该服务覆盖超过 8000 万家网站，提供自定义数据源的订制数据服务，可考虑通过订阅该类 API，丰富数据来源。百度舆情 SaaS 的数据提供订阅 API 能够覆盖百度搜索、百度贴吧、微博、微信、新闻、论坛、博客等全网数据源；日采集数据源达 1 亿条以上，并且支持自定义数据源。 <p>本项目可提供关键词如 “SARS”，“肺炎” 或 “咳嗽” 等与传染病有关的关键词，通过百度舆情 SaaS 来获取来自新闻、百度搜索等大量有效数据。在收费标准方面，百度舆情 SaaS 是根据关键词数量进行收费，最低关键词购买数量为 10 词，最低购买年限为 1 年，因此最低使用费用为 1 年 30000 元人民币（参考 Appendix 图 6）。</p> <ol style="list-style-type: none"> 3. 个人认为爬取网站的公开信息是合法行为，但是有些网站有反爬虫机制，因此为了安全起见，是否应该在爬取对方的网站前，给相关方发邮件，先征得同意，再进行爬取文件。
--------------------	--

六、下周计划

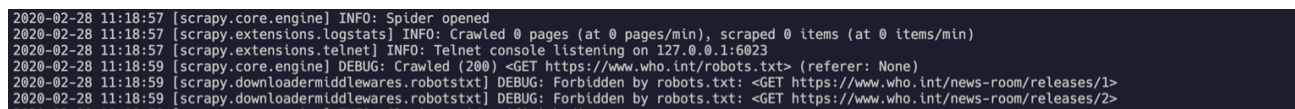
<p>下周计划为：</p> <ol style="list-style-type: none"> 1. 解决反爬虫的问题，优化目前的爬虫部分代码，考虑完善爬虫代码的解析页面功能，并考虑增加并行加速 (multiprocessing)功能，以更快地爬取网页内容。 2. 下周拟以目前的代码为基础，并用优化后的代码对目前所有能够爬取的数据来源网站进行内容爬取（如 Appendix 图三所示，对每个目标网站进行爬取并解析）。网站爬取后的解析过程，需要根据网站前端代码如 div, class 里的信息等去逐步解析文章标题，文章内容，文章发布时间等具体信息，因此需要针对每个不同的网站，都需要对代码进行不同程度的修改和调整从而抓取到所需信息，因此这方面的工作量比预期有所增加。

Appendix



```
(venv) zhangyiming@ichimeis-MacBook-puro: ~/Desktop/UNNC_PhD.D/2019-nCov/3315-Code/Data_Web_Crawler/newsCrawler $ scrapy crawl who
2020-02-28 11:18:57 [scrapy.utils.log] INFO: Scrapy 1.8.0 started (bot: newsCrawler)
2020-02-28 11:18:57 [scrapy.utils.log] INFO: Versions: lxml 4.5.0.0, libxml2 2.9.10, cssselect 1.1.0, parsel 1.5.2, w3lib 1.21.0, Twisted 19.10.0, Python 3.7.1 (v3.7.1:260ec2c36a, Oct 20 2018, 03:13:28) - [Clang 6.0 (clang-600.0.57)], pyOpenSSL 19.1.0 (OpenSSL 1.1.1d 10 Sep 2019), cryptography 2.8, Platform Darwin-19.2.0-x86_64-i386-64bit
2020-02-28 11:18:57 [scrapy.crawler] INFO: Overridden settings: {'BOT_NAME': 'newsCrawler', 'NEWSPIDER_MODULE': 'newsCrawler.spiders', 'ROBOTSTXT_OBEY': True, 'SPIDER_MODULES': ['newsCrawler.spiders']}
2020-02-28 11:18:57 [scrapy.extensions.telnet] INFO: Telnet Password: d8a89a6b8c58dc30
2020-02-28 11:18:57 [scrapy.middleware] INFO: Enabled extensions:
['scrapy.extensions.corestats.CoreStats',
 'scrapy.extensions.telnet.TelnetConsole',
 'scrapy.extensions.memusage.MemoryUsage',
 'scrapy.extensions.logstats.LogStats']
2020-02-28 11:18:57 [scrapy.middleware] INFO: Enabled downloader middlewares:
['scrapy.downloadermiddlewares.robotstxt.RobotsTxtMiddleware',
 'scrapy.downloadermiddlewares.httpauth.HttpAuthMiddleware',
 'scrapy.downloadermiddlewares.downloadtimeout.DownloadTimeoutMiddleware',
 'scrapy.downloadermiddlewares.defaultheaders.DefaultHeadersMiddleware',
 'scrapy.downloadermiddlewares.useragent.UserAgentMiddleware',
 'scrapy.downloadermiddlewares.retry.RetryMiddleware',
 'scrapy.downloadermiddlewares.redirect.MetaRefreshMiddleware',
 'scrapy.downloadermiddlewares.httpcompression.HttpCompressionMiddleware',
 'scrapy.downloadermiddlewares.redirect.RedirectMiddleware',
 'scrapy.downloadermiddlewares.cookies.CookiesMiddleware',
 'scrapy.downloadermiddlewares.httpproxy.HttpProxyMiddleware',
 'scrapy.downloadermiddlewares.stats.DownloaderStats']
2020-02-28 11:18:57 [scrapy.middleware] INFO: Enabled spider middlewares:
['scrapy.spidermiddlewares.httperror.HttpErrorMiddleware',
 'scrapy.spidermiddlewares.offsite.OffsiteMiddleware',
 'scrapy.spidermiddlewares.referer.RefererMiddleware',
 'scrapy.spidermiddlewares.urllength.UrlLengthMiddleware',
 'scrapy.spidermiddlewares.depth.DepthMiddleware']
2020-02-28 11:18:57 [scrapy.middleware] INFO: Enabled item pipelines:
[]
2020-02-28 11:18:57 [scrapy.core.engine] INFO: Spider opened
2020-02-28 11:18:57 [scrapy.extensions.logstats] INFO: Crawled 0 pages (at 0 pages/min), scraped 0 items (at 0 items/min)
2020-02-28 11:18:57 [scrapy.extensions.telnet] INFO: Telnet console listening on 127.0.0.1:6023
2020-02-28 11:18:59 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.who.int/robots.txt> (referer: None)
2020-02-28 11:18:59 [scrapy.downloadermiddlewares.robotstxt] DEBUG: Forbidden by robots.txt: <GET https://www.who.int/news-room/releases/1>
2020-02-28 11:18:59 [scrapy.downloadermiddlewares.robotstxt] DEBUG: Forbidden by robots.txt: <GET https://www.who.int/news-room/releases/2>
2020-02-28 11:18:59 [scrapy.core.engine] INFO: Closing spider (finished)
2020-02-28 11:18:59 [scrapy.statscollectors] INFO: Dumping Scrapy stats:
{'downloader/exception_count': 2,
 'downloader/exception_type_count/scrapy.exceptions.IgnoreRequest': 2,
 'downloader/request_bytes': 220,
 'downloader/request_count': 1,
```

图 1： scrapy 后台爬取过程截图



```
2020-02-28 11:18:57 [scrapy.core.engine] INFO: Spider opened
2020-02-28 11:18:57 [scrapy.extensions.logstats] INFO: Crawled 0 pages (at 0 pages/min), scraped 0 items (at 0 items/min)
2020-02-28 11:18:57 [scrapy.extensions.telnet] INFO: Telnet console listening on 127.0.0.1:6023
2020-02-28 11:18:59 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.who.int/robots.txt> (referer: None)
2020-02-28 11:18:59 [scrapy.downloadermiddlewares.robotstxt] DEBUG: Forbidden by robots.txt: <GET https://www.who.int/news-room/releases/1>
2020-02-28 11:18:59 [scrapy.downloadermiddlewares.robotstxt] DEBUG: Forbidden by robots.txt: <GET https://www.who.int/news-room/releases/2>
```

图 2： scrapy 爬取时发现被禁止访问(forbidden)截图

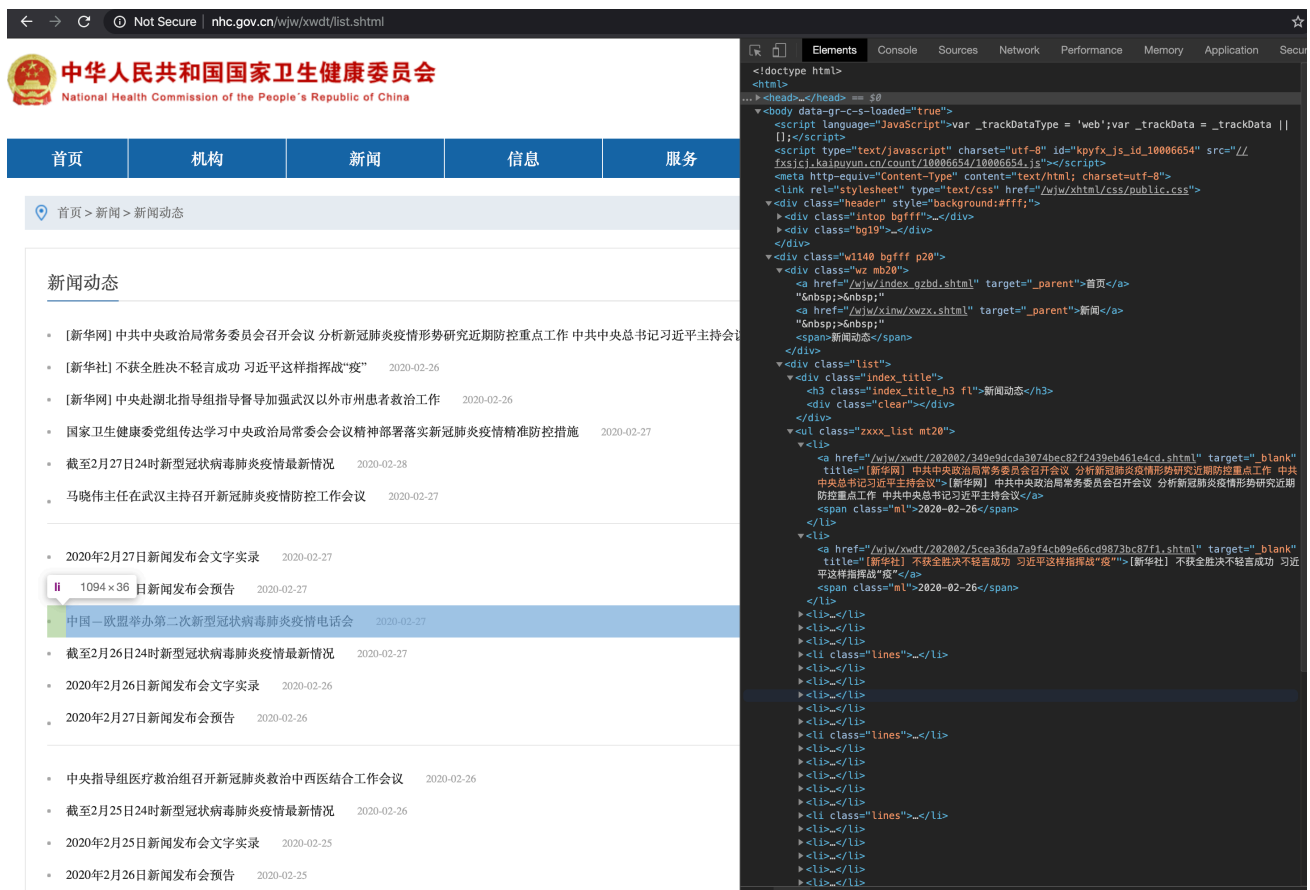


图 3：目标网站代码审查截图



图 4：百度舆情 SaaS
(<https://cloud.baidu.com/product/byapi.html>)

产品功能

舆情文本分析

包含对文本的基本分析，如文本摘要提取、相似文章聚合、舆情情感分析等，通过设置接口中的参数可返回对应的结果。

文本观点聚类

数据 对用户观点进行分类，掌握一批用户发言中用户持有的不同观点类别及各自占比，提升用户评论管理效率，更好地把握传播态势及用户声音。

订阅API

事件脉络梳理

帮助舆情服务商、企业开发者梳理舆情传播时间和传播节点，并按新闻发布时间进行排序，辅助客户快速了解事件发展脉络。

传播路径识别

通过对传播路径的分析帮助客户获得所关心的舆情专题中微博数据的传播关系，数据以嵌套JSON的形式提供。

地域

地域热点汇总

风向标

API

为用户提供以数亿网民的搜索行为作为数据基础、以关键词为统计对象的搜索关键词排行榜。用户可在API中携带地理位置信息（如：国家、省份、城市）、时间信息（时间起点、时间终点）、热词数量，即可实时获取指定地理位置在指定时间段内的百度搜索词排行榜。

舆情平台SAAS

舆情SAAS平台

拥有强大的舆情监控、传播分析、相关搜索词分析、受众画像、事件挖掘、情感提炼等功能，并提供客户定制化的舆情预警通知、舆情分析简报、对比分析报告等辅助模块，依托百度丰富的数据优势与人工智能能力，提供稳定的舆情采集、分析及展示服务。

图 5：百度舆情 SaaS 产品功能截图

数据订阅API

数据订阅API:

☒ 购买数据订阅API

☐ 暂不需要

计费方式:

调用时长和购买关键词数，独立阶梯定价

计费价格:

关键词数量	计费价格
[10词 ~ 50词]	¥3000/词/年
(50词 ~ 200词]	¥2400/词/年
(200词 ~ 500词]	¥2200/词/年
(500词 ~ 999词]	¥2000/词/年
999词以上	¥1800/词/年

关键词数:

10

购买时长:

一年

所选配置

清空配置

配置:

数据订阅API: 10关键词*1年

配置费用

¥ 30,000.00

下一步

温馨提示: 您还没有实名认证, 请立即去认证

图 6：百度舆情 SaaS 产品收费情况截图