

Report Date

Report Date	06/03/2020	Name	Yiming Zhang
-------------	------------	------	--------------

Period covered by this report

Start Date	29/02/2020	End Date	06/03/2020
------------	------------	----------	------------

一、本周研究内容

研究内容	<p>本周的研究内容主要为：</p> <ol style="list-style-type: none"> 1. 研究上周会议讨论提出的新问题，即在数据源上增加 2 个国内医学杂志以及选定了在百度舆情中拟使用的 10 个关键词； 2. 针对自己上周发现的新问题（反爬虫机制）进行了研究； 3. 继续研究计划书中研究内容(一)的内容，完善了爬虫部分代码中的解析部分； 4. 对爬虫中的并行加速 (multiprocessing) 部分进行了研究；
------	--

二、项目实施当前状态

项目进度实施情况	由于上周会议增加了一部分新的内容，研究进度比原计划推迟了半周，目前研究计划中的 研究内容(一) 已经完成了大部分，预计下周完成所有 研究内容（一） 的内容并可以初步开展 研究内容（二） 中有关数据库的内容。
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。

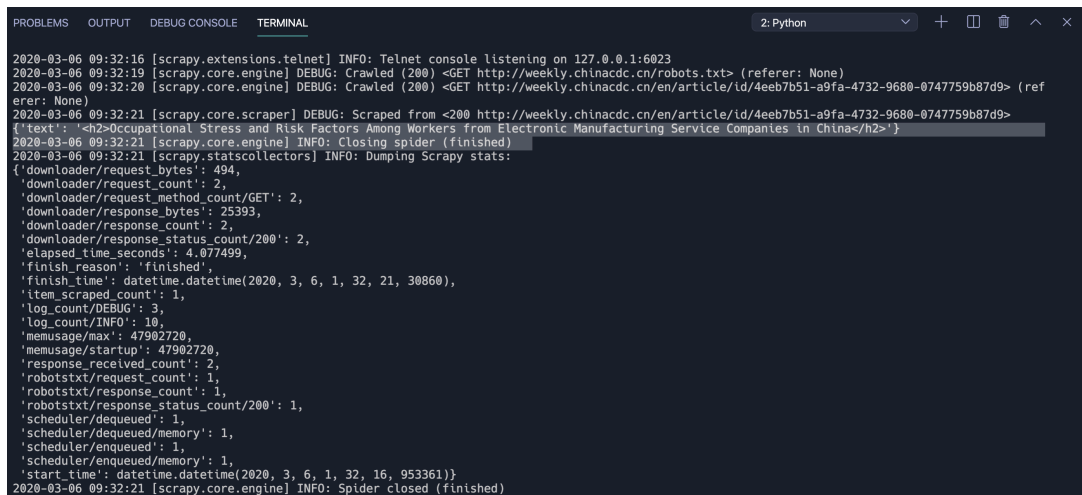
三、本周成果

<p>本周研究的主要成果分为以下四点。</p> <ol style="list-style-type: none"> 1. 对上周会议提出的新问题进行了补充： <ol style="list-style-type: none"> 1.1：研究所使用的数据源增加了两个国内医疗期刊杂志数据源(Data Source)： <ul style="list-style-type: none"> 中华传染病杂志: http://www.zhcrbzz.com/ 解放军医学杂志: http://www.jfjyxxzz.org.cn/WKE/WebPublication/index.aspx?mid=jfjy 1.2：百度舆情拟使用的 10 个关键词：非典，新型冠状病毒，武汉肺炎，肺炎，流行病，病毒，传染病，瘟疫，不明疾病，不明肺炎 2. 对反爬虫技术进行了研究，使用 user-agent 方法绕过反爬虫。具体来说，使用 scrapy 进行爬取时，默认的 user-agent header 内容是"Scrapy/{version}(+http://scrapy.org)"，但是这个内容往往会被很多网站阻拦。因此，要将参数里的 header 参数进行修改，如下图所示：
--

```
def start_requests(self):
    headers= {'User-Agent': 'Mozilla/5.0 (X11; Linux x86_64; rv:48.0) Gecko/20100101 Firefox/48.0'}
    for url in self.start_urls:
        yield Request(url, headers=headers)
```

图 1： scrapy 使用 user-agent 方法绕过反爬虫

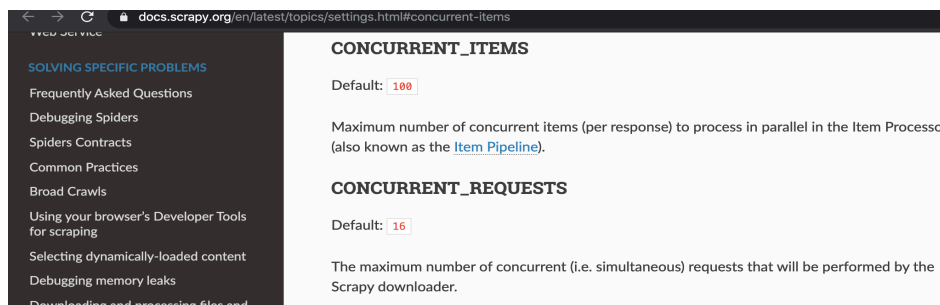
3. 对爬虫代码中的解析部分进行了完善，解析可以使用 XPath 解析和 CSS 解析。如下图所示，在中国疾控中心周报爬取到的 URL 中，用 CSS 解析方法获取到了文章的标题。



```
2020-03-06 09:32:16 [scrapy.extensions.telnet] INFO: Telnet console listening on 127.0.0.1:6023
2020-03-06 09:32:19 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://weekly.chinacdc.cn/robots.txt> (referer: None)
2020-03-06 09:32:20 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://weekly.chinacdc.cn/en/article/id/4eeb7b51-a9fa-4732-9680-0747759b87d9> (ref
erer: None)
2020-03-06 09:32:21 [scrapy.core.scrapy] DEBUG: Scraped from <200 http://weekly.chinacdc.cn/en/article/id/4eeb7b51-a9fa-4732-9680-0747759b87d9>
{'text': '<h2>Occupational Stress and Risk Factors Among Workers from Electronic Manufacturing Service Companies in China</h2>'}
2020-03-06 09:32:21 [scrapy.core.engine] INFO: Closing spider (finished)
2020-03-06 09:32:21 [scrapy.statscollectors] INFO: Dumping Scrapy stats:
{
  'download/request_bytes': 494,
  'download/request_count': 2,
  'download/request_method_count/GET': 2,
  'download/response_bytes': 25393,
  'download/response_count': 2,
  'download/response_status_count/200': 2,
  'elapsed_time_seconds': 4.077499,
  'finish_reason': 'finished',
  'finish_time': datetime.datetime(2020, 3, 6, 1, 32, 21, 30860),
  'item_scraped_count': 1,
  'log_count/DEBUG': 3,
  'log_count/INFO': 10,
  'memusage/max': 47902720,
  'memusage/startup': 47902720,
  'response_received_count': 2,
  'robotstxt/request_count': 1,
  'robotstxt/response_count': 1,
  'robotstxt/response_status_count/200': 1,
  'scheduler/dequeued': 1,
  'scheduler/dequeued/memory': 1,
  'scheduler/enqueued': 1,
  'scheduler/enqueued/memory': 1,
  'start time': datetime.datetime(2020, 3, 6, 1, 32, 16, 95361)}
2020-03-06 09:32:21 [scrapy.core.engine] INFO: Spider closed (finished)
```

图 2： CSS 解析方法 terminal 运行截图

4. 在爬虫代码中用了并行计算 (multiprocessing)进行了研究。研究发现，scrapy 爬虫爬取的速度是非常快的，因此是不建议使用并行计算的。如果想增加并发计算量，可以调整 scrapy 中爬虫部分的有关参数，如调大 CONCURRENT_ITEMS 值等。CONCURRENT_ITEMS 为每个响应的在项目处理器中的并行处理的最大并行项目数。



CONCURRENT_ITEMS
Default: 100
Maximum number of concurrent items (per response) to process in parallel in the Item Processor (also known as the [Item Pipeline](#)).

CONCURRENT_REQUESTS
Default: 16
The maximum number of concurrent (i.e. simultaneous) requests that will be performed by the Scrapy downloader.

图 3： scrapy 文档有关并发部分的截图

相关文档链接: <https://docs.scrapy.org/en/latest/topics/settings.html#concurrent-items>

四、上周问题解决情况

1. 上周提出的爬取网站公开披露信息是否涉及违法，已经在上周会议上解决，爬取网站公开信息用于科研是合法行为。
2. 上周会议提出的增加国内医疗期刊杂志数据源以及 10 个百度舆情搜索关键词问题已经在本周解决。

五、项目当前可能出现的问题

具体问题描述	无具体问题
解决方案	

六、下周计划

下周计划为：

1. 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。本周这部分内容没有全部完成，下周拟将目前所有数据源的网站都进行爬取并解析。