Report Date

Report Date	20/03/2020	**Name**	Yiming Zhang

Period covered by this report

Start Date	14/03/2020	**End Date**	20/03/2020

一、本周研究内容

研究内容	本周的研究内容主要为： **1.** 创建了 GitHub 仓库，并上传并更新了当前最新的爬虫部分代码（https://github.com/yiming95/3315_Project）。 **2.** 研究并探讨了 3315 项目中关于大数据的部分，阅读了几篇关于自然语言处理、大数据等与传染病预警的相关文献。 **3.** 继续爬虫部分的代码工作：利用 scrapy 爬虫框架对新闻网站进行了爬取、解析并储存为 JSON 格式。

二、项目实施当前状态

项目进度实施情况	目前正在进行研究计划中的**研究内容(一)**爬虫部分。
项目整体进度完成情况	预计于下周完成对爬虫进行批量爬取，并将爬取的数据存储到 MongoDB 数据库中。

三、本周成果

本周研究的主要成果主要针对爬虫部分。

1. 对国家卫健委 CCDC 周报的新闻内容 HTML 网页进行分析，如图一所示，发现新闻的文本信息主要在 abs-con 的 div 里：

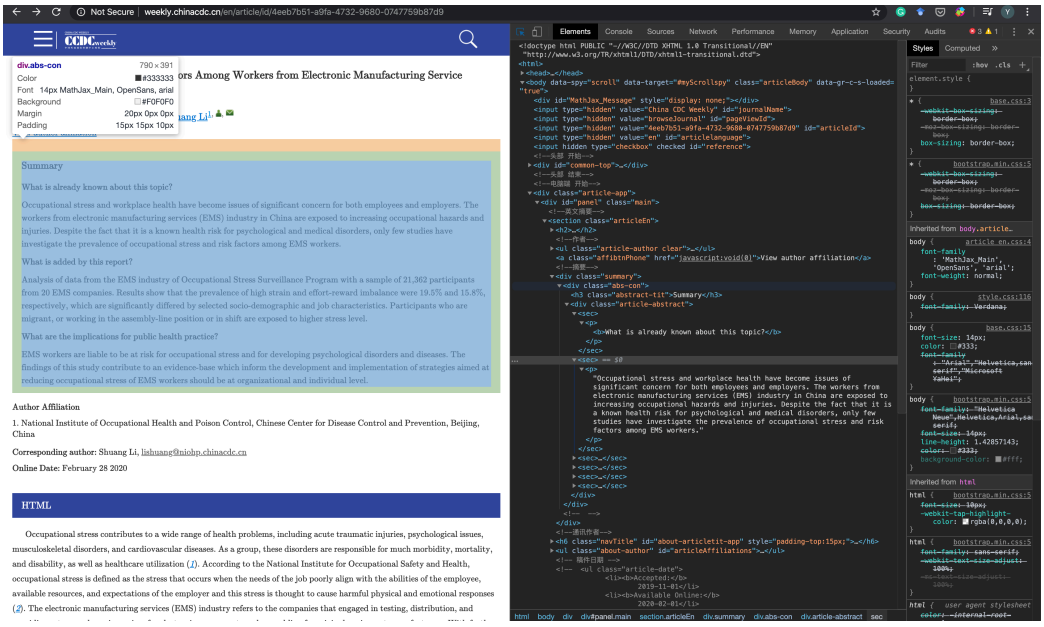

图 1： CCDC 网站审查元素内容

2. 对新闻网页中的内容进行了提取，爬虫部分的代码以及解析的代码如图二所示：

```
newsCrawler > newsCrawler > spiders > ccdc_spider.py > ...
import scrapy
from bs4 import BeautifulSoup

#!  中国疾控中心网站ccdc 爬虫部分核心代码

#Todo: 从主网页作为起始URL,然后对每个超链接里的下一层的网页进行爬取


class cdcSpider(scrapy.Spider):
    # * 爬虫名字为: ccdc
    name = "ccdc"

    start_urls = [
        #中国疾控中心一篇英文报道的网址URL作为爬虫的起始URL
        'http://weekly.chinacdc.cn/en/article/id/4eeb7b51-a9fa-4732-9680-0747759b87d9'
    ]

    # * 从爬取到的网页数据中提取数据
    def parse(self, response):

    #    yield {
    #        'text': response.css('div.article-app h2').get()
    #    }
        response = BeautifulSoup(response.body)

        #! 提取新闻中的Abstract内容, 该内容在网页 abs-con class下
        # for news in response.select('.abs-con'):
        #     print(news.select('.article-abstract')[0].text)
        print(response.select('.article-abstract')[0].text)
```

图 2： CCDC 爬虫部分核心代码

3. 运行爬虫后，成功获取到了新闻的内容，结果如图三所示：

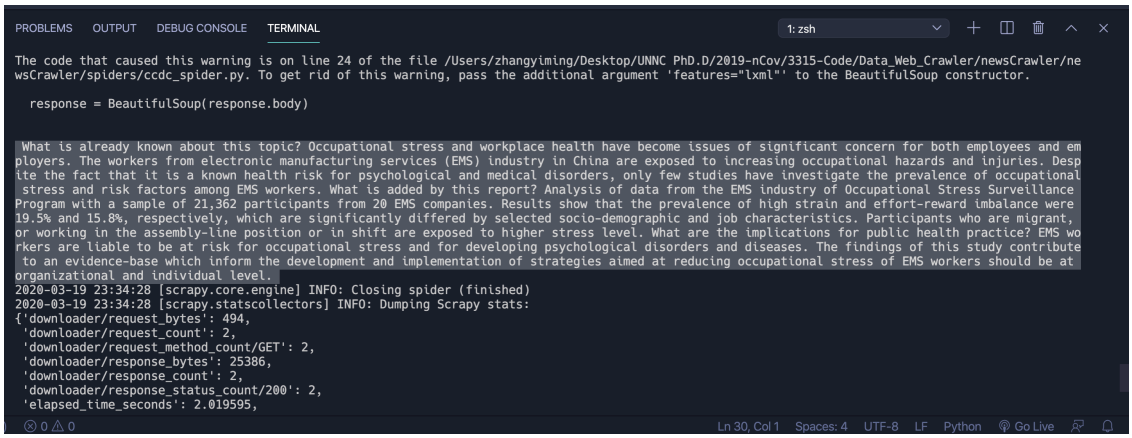

图 3： 成功爬取出新闻内容

4. 在爬虫代码中的 item 进行了定义，并将爬取后的内容转化为 JSON 格式存储在本地，如图 4 所示，存储为 ‘ccdc.json’ 。

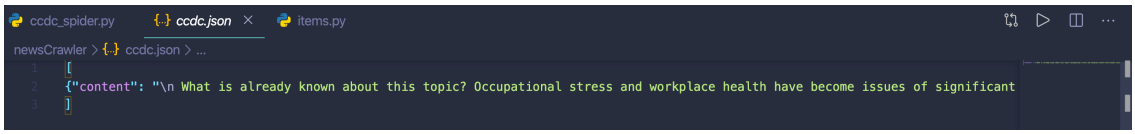

图 4：将爬取后的文件存储为了 JSON 格式

四、上周问题解决情况

上周无具体问题

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. 继续爬虫部分的内容，目前主要是针对新闻页面的爬取，下周尝试从主网页作为起始 URL, 然后对每个超链接里的下一层的网页进行爬取以实现批量爬取。

2. 目前主要是针对新闻类的网站爬取，下周尝试对期刊类的网站进行爬取。