

武汉大学

2020 年暑期 python 编程学习 课程设计报告

项目名称：利用 python 进行网页数据抓取

所在院系：电子信息学院

所在专业：电子信息类

项目成员：郭晏银

项目指导：杨剑锋老师

2020 年 6 月 30 日

目录

第一章 项目分析.....	3
第二章 项目方案.....	5
第三章 项目成果.....	9
第四章 感想及参考文献.....	21

第一章 项目分析

1.1 问题描述

湖北省人民检察院网站 (<http://www.hbjc.gov.cn/>), 界面如下。现需要开发 android 版本的手机客户端程序。由于该门户网站安全要求很高, 不允许其网站数据库开放任何接口。因此手机端的软件的数据库就没有了来源。为此, 需要设计一个爬虫程序, 可以实时监控门户网站上的任何数据, 并可以抓取、存入本地数据库 (手机端软件使用)。

网站包括了很多栏目, 数据类型主要有文字、图片、视频等。请你利用 python 编程, 将门户网站的“权威发布”、“工作报告”这 2 个栏目的内容进行抓取, 并存入本地。

1.2 流程分析

(一)、我发现该网站的“权威发布”这个栏目由案件信息和指导案例组成, “工作报告”栏目由年度报告、半年报告和专题报告组成。因此, 我总共要爬取五个部分的内容。



(二)、“案件信息”部分

点击“更多”后发现该部分总共有 18 页, 共计 268 个小网页, 我的任务便是爬取这 268 个网页内容, 且该部分有些只有文字, 而有一些则包含图片, 都要爬取出来。



（三）、“指导案例” 部分

点击“更多”后发现该部分总共有 2 页，且该部分只有文字，我只需将文字部分全部爬取出来即可。

指导案例

当前位置: 首页>>权威发布>>指导案例

• 最高人民法院第十七批指导性案例

[2020-04-08]

• 最高人民法院第十六批指导性案例

[2020-03-06]

• 最高人民法院发布第十五批指导性案例

[2019-09-26]

• 最高人民法院第十四批指导性案例

[2019-05-23]

• 最高人民法院第十三批指导性案例

[2018-12-25]

• 最高人民法院第十二批指导性案例

[2018-12-20]

• 最高人民法院第十一批指导性案例

[2018-11-19]

• 最高人民法院第十批指导性案例

[2018-07-15]

• 最高人民法院第九批指导性案例

[2017-10-26]

• 最高人民法院第八批指导性案例

[2017-04-28]

• 最高人民法院第七批指导性案例

[2017-04-28]

• 最高人民法院第六批指导性案例

[2017-04-28]

• 最高人民法院第五批指导性案例

[2014-09-17]

• 最高人民法院惩治危害食品安全犯罪指导性案例（第四批指导性案例）

[2014-09-17]

• 最高人民法院第三批指导性案例

[2014-09-17]

每页15条，共2页

上一页0102下一页

指导案例

当前位置: 首页>>权威发布>>指导案例

最高人民法院第二批指导性案例

时间: 2014-09-17 来源: 最高人民法院 阅读量:

关于印发第二批指导性案例的通知

各省、自治区、直辖市人民检察院，军事检察院，新疆生产建设兵团人民检察院

经2012年10月31日最高人民法院第十一届检察委员会第八十一次会议审议通过，现附最高人民法院、陈某某等滥用职权案、罗某某等滥用职权案、胡某某等滥用职权案不移交刑事强制措施和某某玩忽职守、徇私枉法、受贿案等五个案例印发你们，供参考。

最高人民法院

2012年11月15日

崔某某环境监管失职案

(检第4号)

【关键词】

法职罪主体国有事业单位工作人员环境监管失职罪

【要旨】

实践中，一些国有企业、企业和事业单位经合法授权从事具体的管理市场经济和社会生活的工作，具有一定管理公共事务和社会事务的职能，这些实际行使国家行政管理职权的公司、企业和事业单位工作人员，符合渎职罪主体要求对其实施渎职行为构成犯罪的，应当依照刑法关于渎职罪的规定追究刑事责任。

【相关立法】

（四）、“年度报告” 部分

点击“更多”后发现该部分总共有 2 页，该部分有些只有文字，而有一些则包含图片，我要将他们全都爬取出来，整体来看与（二）相似。

指导案例

当前位置: 首页>>权威发布>>指导案例

• 最高人民法院第十七批指导性案例

[2020-04-08]

• 最高人民法院第十六批指导性案例

[2020-03-06]

• 最高人民法院发布第十五批指导性案例

[2019-09-26]

• 最高人民法院第十四批指导性案例

[2019-05-23]

• 最高人民法院第十三批指导性案例

[2018-12-25]

• 最高人民法院第十二批指导性案例

[2018-12-20]

• 最高人民法院第十一批指导性案例

[2018-11-19]

• 最高人民法院第十批指导性案例

[2018-07-15]

• 最高人民法院第九批指导性案例

[2017-10-26]

• 最高人民法院第八批指导性案例

[2017-04-28]

• 最高人民法院第七批指导性案例

[2017-04-28]

• 最高人民法院第六批指导性案例

[2017-04-28]

• 最高人民法院第五批指导性案例

[2014-09-17]

• 最高人民法院惩治危害食品安全犯罪指导性案例（第四批指导性案例）

[2014-09-17]

• 最高人民法院第三批指导性案例

[2014-09-17]

每页15条，共2页

上一页0102下一页

（五）、“半年报告” 部分

点击“更多”后发现该部分总共只有 1 页，该部分有些有文字，而有一些则包含图片，我要将他们全都爬取出来，整体来看与（二）也相相似。

半年报告

当前位置: 首页>>工作报告>>半年报告

• 王晋检察长向省人大常委会作半年检察工作报告

[2018-07-23]

• 省检察院向省人大常委会作半年工作报告

[2018-07-23]

• 依法履职尽责 服务保障大局

王晋检察长向省人大常委会作半年检察工作报告

[2017-07-24]

• 依法履职尽责 服务保障大局

王晋检察长向省人大常委会作半年工作报告

[2016-07-26]

• 湖北省人民检察院关于2014年上半年工作情况的报告

[2014-09-12]

每页15条，共1页

（六）、“专题报告”部分

点击“更多”后发现该部分总共只有 1 页，该部分有些有文字，而有一些则包含图片，我要将他们都爬取出来，整体来看与（二）也相相似。



第二章 项目方案

2.1 实施步骤

6 月 27 日，确定语言为 python 语言，查询与 python 爬虫相关的例子和知识点，先进行较系统性的学习。学习后发现有多中方法可以爬取网页，如 BeautifulSoup 或 Xpath 和 lxml 等等，综合考量后，我决定采取 Xpath 和 lxml 来采取数据。

6 月 28 日，开始进行 python 编程爬取数据，成功将权威发布的案件信息的文本以及图片分别以 txt 和 jpg 形式保存到指定文件夹下。

6 月 29 日，仿照 6 月 28 日的方法，成功将权威发布的指导案例以及工作报告的年度报告、半年报告和专题报告的文本以及图片分别以 txt 和 jpg 形式保存到指定文件夹下。

6 月 30 日，整理并优化代码，添加注释，回顾流程，编写课程设计报告。

2.2 具体细节

（一）、五个模块分别对应四个 urllist 和 dizhilib，其中 urllist 存储的是五大部分的每一页的网址，其中 dizhilib 存储的是每个地址里面所有的小地

址。

```
6 #urllist中存储四大内容总的网页网址, dizhilib存储内容里面每个小网址
7 urllist1_1=['http://www.hbjc.gov.cn/qwfb/ajxx/index.shtml']
8 dizhilib1_1=[]
9 urllist1_2=['http://www.hbjc.gov.cn/qwfb/zdal/', 'http://www.hbjc.gov.cn/qwfb/zdal/index_1.shtml']
10 dizhilib1_2=[]
11 urllist1_3=['http://www.hbjc.gov.cn/gzbg/ndbg/index.shtml', 'http://www.hbjc.gov.cn/gzbg/ndbg/index_1.shtml']
12 dizhilib1_3=[]
13 urllist1_4=['http://www.hbjc.gov.cn/gzbg/bnbg/']
14 dizhilib1_4=[]
15 urllist1_5=['http://www.hbjc.gov.cn/gzbg/ztb/']
16 dizhilib1_5=[]
17
```

(二)、分别通过下面的函数得到每个地址内的文本和图片。

```
def gethtml(url): #通过url得到html
    res = requests.get(url)
    html = etree.HTML(res.content)
    return html
def getdizhi(html): #得到大的内容里面的每个小地址
    return html.xpath("//div[@class='conblock_listpage']/ul/li/a/@href")
def gettitle1_1(html): #得到标题方法1
    return html.xpath("//h1[@class='arttitle']/text()")
def gettitle1_2(html): #得到标题方法2
    return html.xpath("//div[@class='detail_tit']/text()")
def gettext1_1(html): #得到文本方法1
    return html.xpath("//div[@class='article']/text()")
def gettext1_2(html): #得到文本方法2
    return html.xpath("//div[@class='detail_con']/text()")
def getimg(html): #得到图片
    return html.xpath("//div[@class='article']/img/@oldsrc")
```

(三)、最后就是输出图片和文本到对应的文件夹

```
#输出图片
if deepimg==[]:
    pass
else:
    for img in deepimg:
        s = img[2:8]
        k = img[17]
        img = 'http://www.hbjc.gov.cn/ajxx/{0}/{1}'.format(s, img)
        try:
            urllib.request.urlretrieve(img, r'D:\text\权威发布\案件信息\{0}-{1}.jpg'.format(deeptitle, k))
        except:
            pass
#输出文本
with open(r'D:\text\权威发布\案件信息\{0}.txt'.format(deeptitle), 'w', encoding="utf-8") as file:
    for texts in deeptext:
        if (texts[0] == '\n' and (texts[-1] == ' ' or texts[-1] == '\n')):
            pass
        else:
            file.write(texts + '\n')
print('权威发布的案件信息内容爬取完毕!')
print()
```

2.3 遇到的困难及解决方案

(一)、翻页问题, 权威发布的案件信息部分有 18 页, 要提取每一页内的网址。
解决方法: 通过观察, 发现网址之间存在规律, 从而得到每一页的网址。

```
for i in range(1,18):
    m='http://www.hbjc.gov.cn/qwfb/ajxx/index_'+str(i)+'.shtml'
    urlist1_1.append(m)
```

(二)、前后网址类型不同问题，权威发布的指导案例部分前后的网址类型不一样。

```
▼<div class="span830">
  ▼<div class="conblock articlepage">
    ▶<div class="h">...</div>
    ▼<div class="article">
      ...
      <h1 class="arttitle">最高人民法院第六批指导性案例</h1> == $0
      ▶<p class="artinfo">...</p>
      ▶<div class="readbox">...</div>
      ▶<div class="arttext">...</div>
      <p class="editor">作者：</p>

▼<div class="container bgfff">
  ▶<div class="BreadcrumNav">...</div>
  ▼<div class="detail_con">
    <div class="detail_tit">第十七批指导性案例</div> == $0
    ▶<div class="detail_extend">...</div>
    <div class="clear"></div>
```

解决方法：找到不同网址的分界处，对前后两种不同的网址采取不同的数据提取方法以及数据加工处理的方法，从而将他们写入到对应的文件夹之中

```
for deepurl in dizhilist1_2[0][:9]:
    deephtml = gethtml(deepurl)
    deeptitle = gettitle1_2(deephtml)[0]
    deeptext = gettext1_2(deephtml)
```

```
for deepurl in dizhilist1_2[0][9:]:
    printtext1_2(deepurl)
for deepurl in dizhilist1_2[1]:
    printtext1_2(deepurl)
```

(三)、文本写入异常，无法将提取的数据写进对应的文件夹之中。

```
UnicodeDecodeError: 'gbk' codec can't decode byte 0x88 in position 7: illegal multibyte sequence
```

解决方法：将它写成‘utf-8’格式即可，在 open 函数后面加一个参数 encoding=‘utf-8’ 即可。

```
with open(r'D:\text\权威发布\案件信息\{0}.txt'.format(deeptitle), 'w', encoding="utf-8") as file:
```


(四)、直接输出的文本空行特别多，且格式很混乱。

日前，湖北省武汉园林绿化建设发展有限公司原党委书记、董事长吴涛（副厅级）涉嫌受贿等一案，由武汉市人民检察院向武汉市中级人民法院提起公诉。

检察机关在审查起诉阶段，依法告知了被告人吴涛享有的诉讼权利，并依法讯问了被告人，听取了辩护人的意见。武汉市人民检察院起诉指控：被告人吴涛在任武汉园林绿化建设发展有限公司党委

作者：

上一篇新闻：

解决方法：按照一定方法加工数据后在输出，下图为之后的输出结果。

访问量：



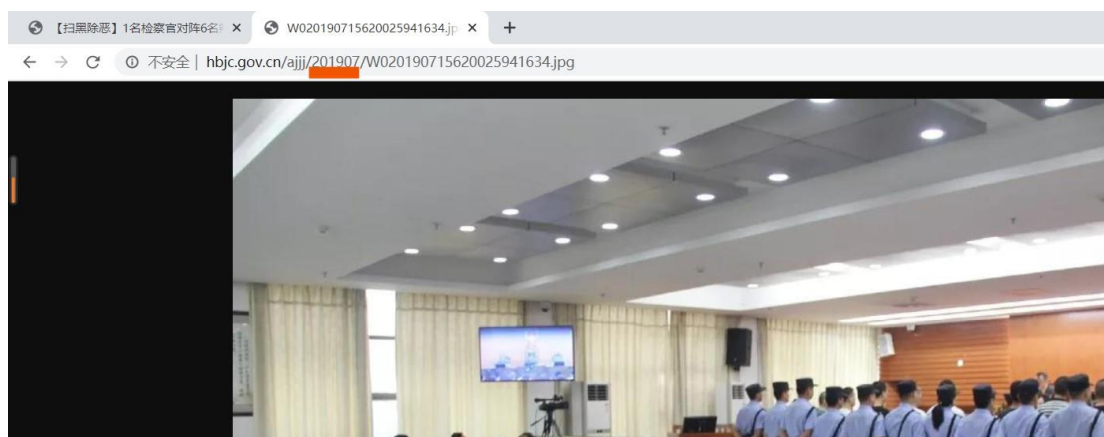
日前，湖北省武汉园林绿化建设发展有限公司原党委书记、董事长吴涛
检察机关在审查起诉阶段，依法告知了被告人吴涛享有的诉讼权利，
作者：

下一篇新闻：

仙桃市人民检察院依法对兰红林涉嫌受贿案提起公诉

(五)、图片网址问题，无法直接从中提取到图片的网址，可以看到图片的网址那里还有一个时间戳，就无法像之前一样简单地得到图片网址。

```
<p align="center">  
***  
 == $0
```



解决方案：提取到的图片网址，如./W20190715620025941634，可以看到第 2 到第 7 位正好是对应的时间戳。

```
for img in deepimg:
    s = img[2:8]
    k = img[17]
    img = 'http://www.hbjc.gov.cn/jcyw/{0}/{1}'.format(s, img)
    try:
        urllib.request.urlretrieve(img, r'D:\text\工作报告\年度报告\{0}\{1}.jpg'.format(deeptitle, k))
    except:
        pass
```

第三章 项目成果

3.1 代码展示

```
from lxml import etree
import requests
import urllib.request
import re

#urllist 中存储四大内容总的网页网址，dizhilist 存储内容里面每个小网址
urllist1_1=['http://www.hbjc.gov.cn/qwfb/ajxx/index.shtml']
dizhilist1_1=[]
urllist1_2=['http://www.hbjc.gov.cn/qwfb/zdal/', 'http://www.hbjc.gov.cn/qwfb/zdal/index_1.shtml']
dizhilist1_2=[]
urllist2_1=['http://www.hbjc.gov.cn/gzbg/ndbg/index.shtml', 'http://www.hbjc.gov.cn/gzbg/ndbg/index_1.shtml']
dizhilist2_1=[]
urllist2_2=['http://www.hbjc.gov.cn/gzbg/bnbg/']
dizhilist2_2=[]
urllist2_3=['http://www.hbjc.gov.cn/gzbg/ztbg/']
dizhilist2_3=[]

def gethtml(url):    #通过url 得到html
    res = requests.get(url)
    html = etree.HTML(res.content)
    return html

def getdizhi(html):    #得到大的内容里面的每个小地址
    return html.xpath("//div[@class='conblock listpage']/ul/li/a/@href")

def gettitle1_1(html):    #得到标题方法1
```

```

        return html.xpath("//h1[@class='arttitle']/text()")
def gettitle1_2(html):    #得到标题方法 2
    return html.xpath("//div[@class='detail_tit']/text()")
def gettext1_1(html):    #得到文本方法 1
    return html.xpath("//div[@class='article']//text()")
def gettext1_2(html):    #得到文本方法 22
    return html.xpath("//div[@class='detail_con']//text()")
def getimg(html):        #得到图片
    return html.xpath("//div[@class='article']//img/@oldsrc")
def printtext1_2(deepurl):    #指导案例的一部分文本输出
    deepurl = 'http://www.hbjc.gov.cn/qwfb/zdal/' + deepurl
    deephtml = gethtml(deepurl)
    deeptitle = gettitle1_1(deephtml)[0]
    deeptext = gettext1_1(deephtml)
    # 输出文本
    with open(r'D:\text\权威发布\指导案例\{0}.txt'.format(deeptitle),
            'w', encoding="utf-8") as file:
        for texts in deeptext:
            if (texts[0] == '\n' and (texts[-1] == ' ' or texts[-1] ==
            '\n')):
                pass
            else:
                file.write(texts + '\n')

"""
#权威发布的案件信息的内容提取
"""
#获得 18 页内所有案例（268 个）的网址
print('开始爬取权威发布的案件信息的内容')
print('-----请稍后-----')
for i in range(1,18):
    m='http://www.hbjc.gov.cn/qwfb/ajxx/index_'+str(i)+'.shtml'
    urllist1_1.append(m)
for url in urllist1_1:
    html = gethtml(url)
    dizhilist1_1.append(getdizhi(html))
#将每个网址里面对应的内容以 TXT 格式输出到对应文件夹中,并将 jpg 图片也输出到文件夹
中
for url in dizhilist1_1:
    for deepurl in url:
        deepurl='http://www.hbjc.gov.cn/qwfb/ajxx/'+deepurl
        deephtml=gethtml(deepurl)
        deeptitle=gettitle1_1(deephtml)[0]
        deepimg=getimg(deephtml)

```

```

    deeptext=gettext1_1(deephtml)
    #输出图片
    if deepimg==[]:
        pass
    else:
        for img in deepimg:
            s = img[2:8]
            k = img[17]
            img = 'http://www.hbjc.gov.cn/ajjj/{0}/{1}'.format(s,
img)

            try:
                urllib.request.urlretrieve(img, r'D:\text\权威发布\案
件信息\{0}{1}.jpg'.format(deeptitle, k))
            except :
                pass
        #输出文本
        with open(r'D:\text\权威发布\案件信息\{0}.txt'.format(deeptitle),
'w', encoding="utf-8") as file:
            for texts in deeptext:
                if (texts[0] == '\n' and (texts[-1] == ' ' or texts[-1]
== '\n')):
                    pass
                else:
                    file.write(texts + '\n')
print('权威发布的案件信息内容爬取完毕')
print()

"""
#权威发布的指导案例的内容提取
"""
print('开始爬取权威发布的指导案例的内容')
print('-----请稍后-----')
#获得2 页内所有指导案例的网址
for url in urllist1_2:
    html = gethtml(url)
    dizhilist1_2.append(getdizhi(html))
#将每个网址里面对应的内容以 TXT 格式输出到对应文件夹中
for deepurl in dizhilist1_2[0][:9]:
    deephtml = gethtml(deepurl)
    deeptitle = gettitle1_2(deephtml)[0]
    deeptext = gettext1_2(deephtml)
    # 输出文本
    with open(r'D:\text\权威发布\指导案例\{0}.txt'.format(deeptitle),
'w', encoding="utf-8") as file:

```

```

        for texts in deeptext:
            if (texts[0] == '\r'):
                if (texts[-1] == '\n'):
                    file.write(texts)
                else:
                    pass
            else:
                file.write(texts)
    for deepurl in dizhilist1_2[0][9:]:
        printtext1_2(deepurl)
    for deepurl in dizhilist1_2[1]:
        printtext1_2(deepurl)
    print('权威发布的指导案例的内容爬取完毕')
    print()

    """
    #工作报告的年度报告的内容提取
    """

    print('开始爬取工作报告的年度报告的内容')
    print('-----请稍后-----')
    #获得2 页内所有年度报告的网址
    for url in urllist2_1:
        html = gethtml(url)
        dizhilist2_1.append(getdizhi(html))
    for i in range(0,18):
        if i==0:

            dizhilist2_1[0][0]='http://www.hbjc.gov.cn/'+str(dizhilist2_1[0][0])
            else:
                s=0 if i<15 else 1
                dizhilist2_1[s][i-
s*15]='http://www.hbjc.gov.cn/gzbg/ndbg/'+str(dizhilist2_1[s][i-
s*15])
    #将每个网址里面对应的内容以 TXT 格式输出到对应文件夹中, 图片以 jpg 格式输出到对应文
    件夹
    for url in dizhilist2_1:
        for deepurl in url:
            deephtml=gethtml(deepurl)
            deeptitle=gettitle1_1(deephtml)[0]
            deepimg = getimg(deephtml)
            deeptext = gettext1_1(deephtml)
            # 输出图片
            if deepimg == []:
                pass

```

```

else:
    for img in deepimg:
        s = img[2:8]
        k = img[17]
        img = 'http://www.hbjc.gov.cn/jcyw/{0}/{1}'.format(s,
img)

        try:
            urllib.request.urlretrieve(img, r'D:\text\工作报告\年
度报告\{0}\{1}.jpg'.format(deeptitle, k))
        except:
            pass

    # 输出文本
    with open(r'D:\text\工作报告\年度报告\{0}.txt'.format(deeptitle),
'w', encoding="utf-8") as file:
        for texts in deeptext:
            file.write(texts)
print('工作报告的年度报告的内容爬取完毕')
print()

"""
#工作报告的半年报告的内容提取
"""

print('开始爬取年度报告的半年报告的内容')
print('-----请稍后-----')
#获得所有半年报告的网址
for url in urllist2_2:
    html = gethtml(url)
    dizhilist2_2.append(getdizhi(html))
#将每个网址里面对应的内容以 TXT 格式输出到对应文件夹中
for i in range(0,5):
    if(i==4):
        dizhilist2_2[0][i] = 'http://www.hbjc.gov.cn/gzbg/bnbg/' +
str(dizhilist2_2[0][i])
    else:
        dizhilist2_2[0][i] = 'http://www.hbjc.gov.cn/' +
str(dizhilist2_2[0][i])
for url in dizhilist2_2:
    for deepurl in url:
        deephtml = gethtml(deepurl)
        deeptitle = gettitle1_1(deephtml)[0]
        deepimg = getimg(deephtml)
        deeptext = gettext1_1(deephtml)
        # 输出图片
        if deepimg == []:

```

```

        pass
    else:
        for img in deepimg:
            s = img[2:8]
            k = img[17]
            img = 'http://www.hbjc.gov.cn/jcyw/{0}/{1}'.format(s,
img)

            try:
                urllib.request.urlretrieve(img, r'D:\text\工作报告\半
年度报告\{0}\1}.jpg'.format(deeptitle, k))
            except:
                pass

        # 输出文本
        with open(r'D:\text\工作报告\半年报告\{0}.txt'.format(deeptitle),
'w', encoding="utf-8") as file:
            for texts in deeptext:
                if (texts[0] == '\r'):
                    if (texts[-1] == '\n'):
                        file.write(texts)

                    else:
                        pass

                else:
                    file.write(texts)
print('年度报告的半年报告的内容爬取完毕')
print()

"""
#工作报告的专题报告的内容提取
"""

print('开始爬取年度报告的专题报告的内容')
print('-----请稍后-----')
#获得所有专题报告的网址
for url in urllist2_3:
    html = gethtml(url)
    dizhilist2_3.append(getdizhi(html))
#将每个网址里面对应的内容以 TXT 格式输出到对应文件夹中
for i in range(0,4):
    if(i==3):
        dizhilist2_3[0][i] = 'http://www.hbjc.gov.cn/gzbg/ztbg/' +
str(dizhilist2_3[0][i])
    else:
        dizhilist2_3[0][i] = 'http://www.hbjc.gov.cn/' +
str(dizhilist2_3[0][i])
for url in dizhilist2_3:

```



```

for deepurl in url:
    deephtml = gethtml(deepurl)
    deeptitle = gettitle1_1(deephtml)[0]
    deepimg = getimg(deephtml)
    deeptext = gettext1_1(deephtml)
    # 输出图片
    if deepimg == []:
        pass
    else:
        for img in deepimg:
            s = img[2:8]
            k = img[17]
            img = 'http://www.hbjc.gov.cn/jcyw/{0}/{1}'.format(s,
img)

            try:
                urllib.request.urlretrieve(img, r'D:\text\工作报告\专
题报告\{0}\{1}.jpg'.format(deeptitle, k))
            except:
                pass
        # 输出文本
        with open(r'D:\text\工作报告\专题报告\{0}.txt'.format(deeptitle),
'w', encoding="utf-8") as file:
            for texts in deeptext:
                if (texts[0] == '\r'):
                    if (texts[-1] == '\n'):
                        file.write(texts)
                    else:
                        pass
                else:
                    file.write(texts)
print('年度报告的专题报告的内容爬取完毕')
print()

```

3.2 实现情况

代码运行后，会出现如下界面，实现人机之间的交互。如果只爬取五个部分的文本，只需要 10-20 秒左右的时间，但如果要爬取网址内的图片，下载的速度就会变慢，则需要 5-7 分钟左右的时间。

```
D:\pacong\env\Scripts\python.exe D:/pacong/realpacong.py
```

```
开始爬取权威发布的案件信息的内容
```

```
-----请稍后-----
```

```
权威发布的案件信息内容爬取完毕
```

```
开始爬取权威发布的指导案例的内容
```

```
-----请稍后-----
```

```
权威发布的指导案例的内容爬取完毕
```

```
开始爬取工作报告的年度报告的内容
```

```
-----请稍后-----
```

```
工作报告的年度报告的内容爬取完毕
```

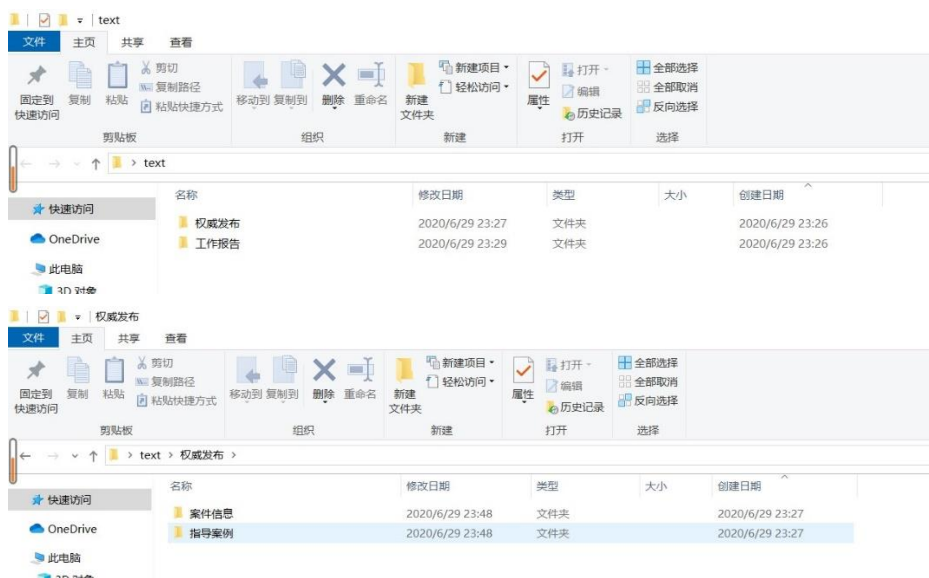
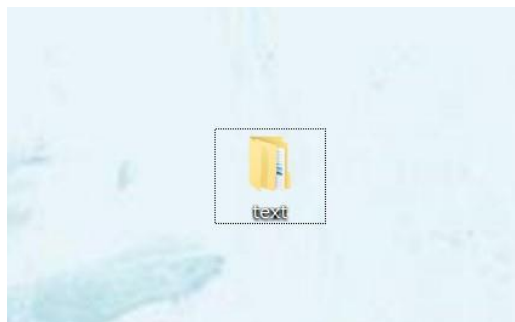
```
开始爬取年度报告的半年报告的内容
```

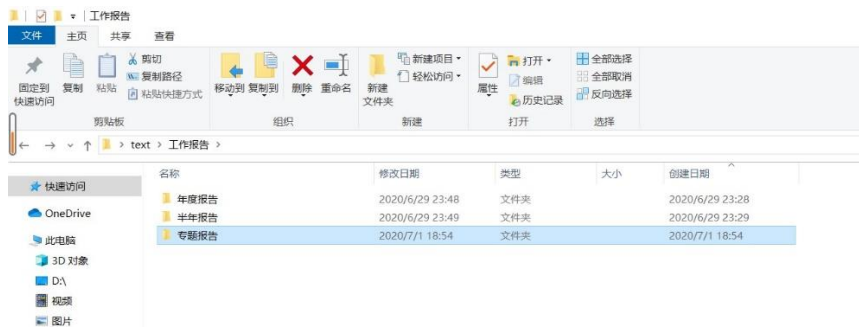
```
-----请稍后-----
```

```
年度报告的半年报告的内容爬取完毕
```

```
Process finished with exit code 0
```

数据存储在对应该文件夹之内，文件夹名为 text，其中有权威发布和工作报告两部分，而权威报告中又有案件信息和指导案例两部分，工作报告中又有年度报告、半年报告和专题报告三部分。





首先是权威发布的案件信息部分的展示，可以看到一共有 334 个项目，其中有 268 个文本文件以及 66 个 jpg 图片。



【扫黑除恶】1名检察官对阵6名辩护人，武六市检察院提起公诉的陈智等11人恶势力犯罪团伙案公开审理

时间：2019-07-15 来源： 访问量：



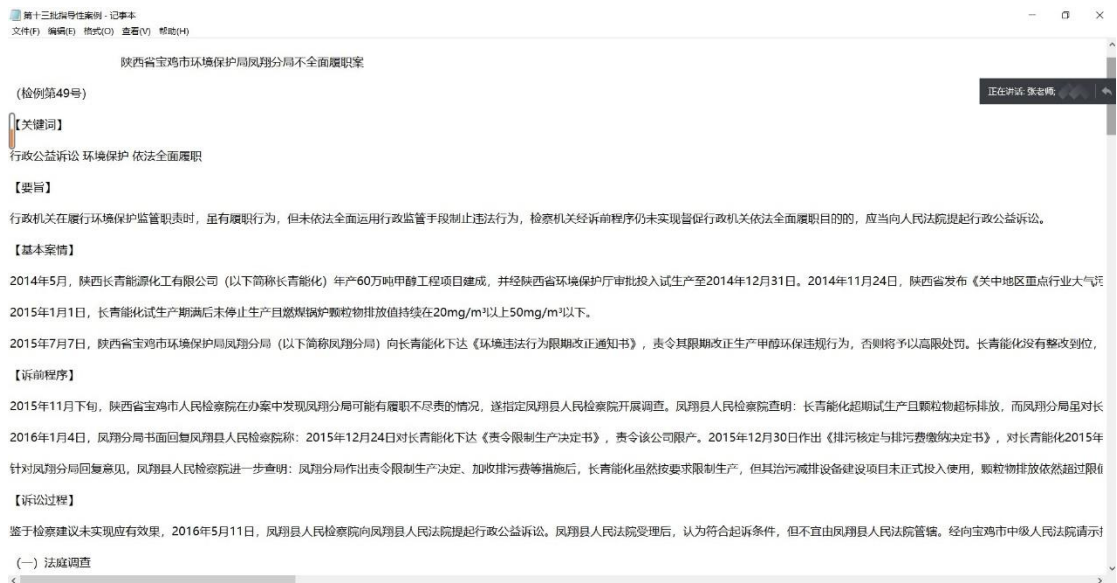
庭审现场

7月12日，由武六市人民检察院依法提起公诉的陈智等11名被告人涉嫌非法采矿罪，寻衅滋事罪，掩饰、隐瞒犯罪所得

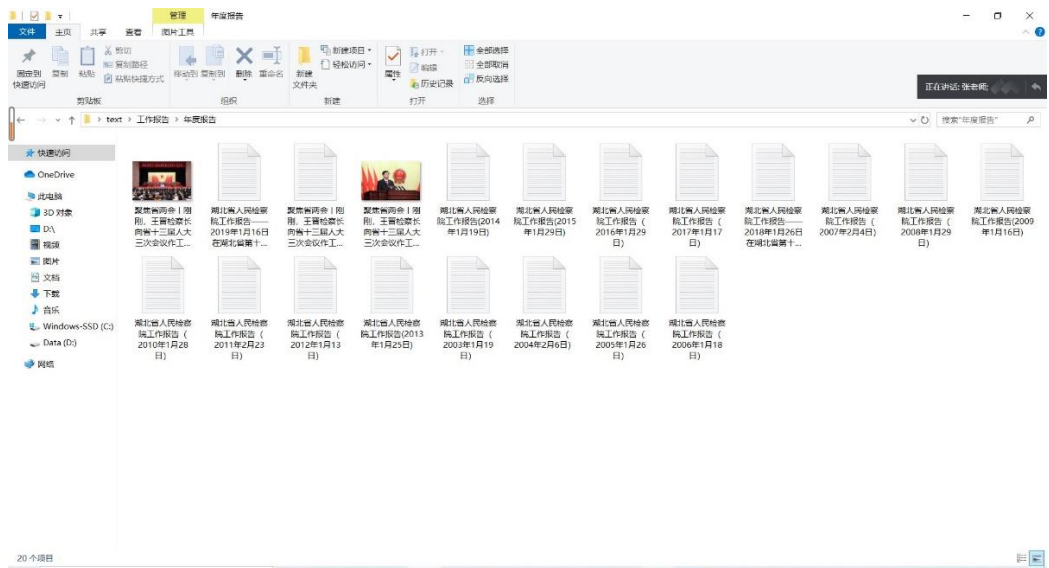
然后是权威发布的指导案例部分的展示，该部分共有 17 个 txt 文件，没有 jpg。

File Explorer window showing a directory named "指导案例" (Guiding Cases). The directory contains 17 text files (txt) related to guiding cases, all dated 2020/6/30 19:36. The files are listed in a table with columns: 名称 (Name), 修改日期 (Modified Date), 类型 (Type), 大小 (Size), and 创建日期 (Created Date).

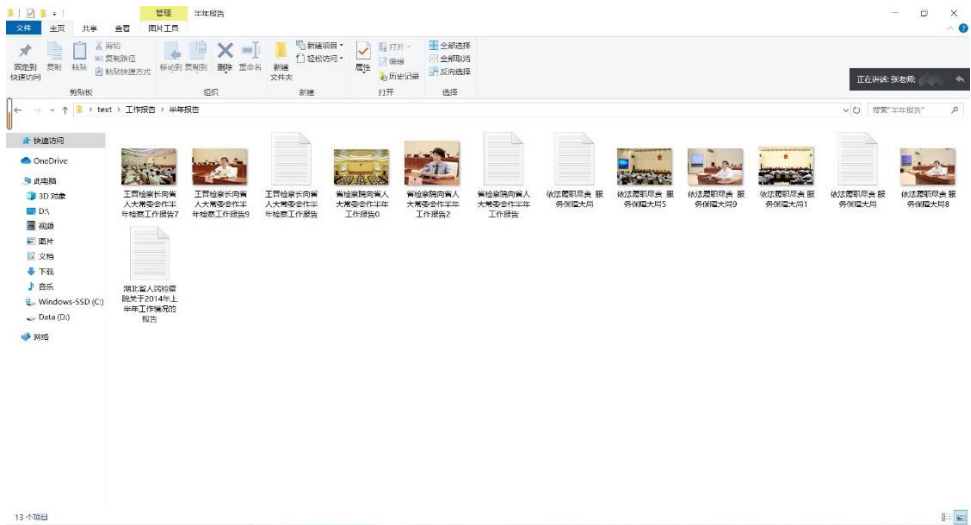
名称	修改日期	类型	大小	创建日期
第十七批指导性案例	2020/6/30 19:36	文本文档	38 KB	2020/6/29 23:48
第十六批指导性案例	2020/6/30 19:36	文本文档	48 KB	2020/6/29 23:48
第十三批指导性案例	2020/6/30 19:36	文本文档	24 KB	2020/6/29 23:48
第十四批指导性案例	2020/6/30 19:36	文本文档	37 KB	2020/6/29 23:48
最高人民检察院发布第十五批指导性案例...	2020/6/30 19:36	文本文档	54 KB	2020/6/29 23:48
第十二批指导性案例	2020/6/30 19:36	文本文档	28 KB	2020/6/29 23:48
第十批指导性案例	2020/6/30 19:36	文本文档	35 KB	2020/6/29 23:48
第十一批指导性案例	2020/6/30 19:36	文本文档	27 KB	2020/6/29 23:48
第九批指导性案例	2020/6/30 19:36	文本文档	31 KB	2020/6/29 23:48
最高人民检察院惩治危害食品安全犯罪指...	2020/6/30 19:36	文本文档	11 KB	2020/6/29 23:48
最高人民检察院第八批指导性案例	2020/6/30 19:36	文本文档	58 KB	2020/6/29 23:48
最高人民检察院第二批指导性案例	2020/6/30 19:36	文本文档	29 KB	2020/6/29 23:48
最高人民检察院第六批指导性案例	2020/6/30 19:36	文本文档	18 KB	2020/6/29 23:48
最高人民检察院第七批指导性案例	2020/6/30 19:36	文本文档	45 KB	2020/6/29 23:48
最高人民检察院第三批指导性案例	2020/6/30 19:36	文本文档	11 KB	2020/6/29 23:48
最高人民检察院第五批指导性案例	2020/6/30 19:36	文本文档	7 KB	2020/6/29 23:48
最高人民检察院第一批指导性案例	2020/6/30 19:36	文本文档	19 KB	2020/6/29 23:48



然后是工作报告的年度报告部分的展示，共有 20 个项目，包括 18 个 txt 文件以及 2 张 jpg 图片。



然后是工作报告的半年报告部分的展示，有 13 个项目，5 个 txt 文件和 8 张 jpg 图片。



“湖北省人民检察院关于 2014 年上半年工作情况的报告” 刘事本
文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)
湖北省人民检察院关于 2014 年上半年工作情况的报告

时间: 2014-09-12
来源: 湖北省人民检察院
访问量:
湖北省人民检察院
关于 2014 年上半年工作情况的报告

—— 2014 年 7 月 28 日在湖北省第十二届人民代表大会
常务委员会第十次会议上

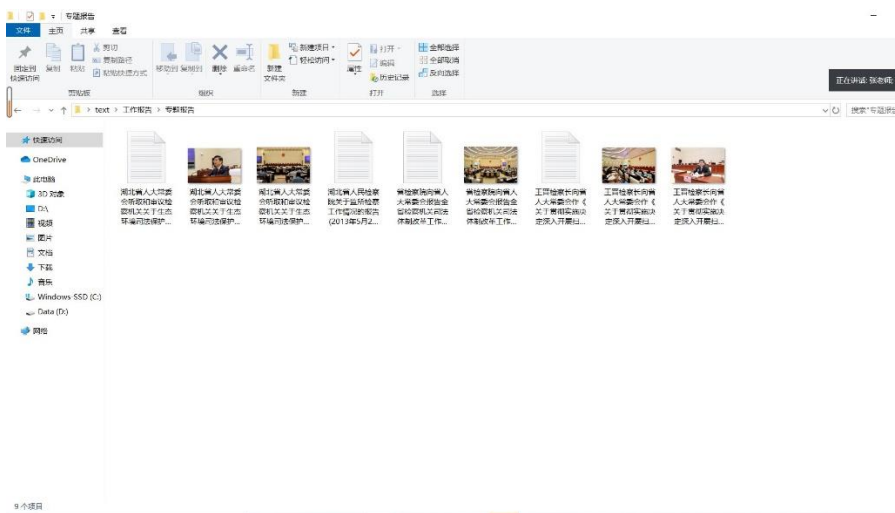
湖北省人民检察院检察长 敬人刀

主任、各位副主任、秘书长、各位委员:

现在, 我代表省人民检察院向会议报告上半年工作, 请予审议。

今年以来, 省人民检察院在省委、最高人民检察院正确领导下, 在省人大及其常委会有力监督下, 深入贯彻落实党的十八大精神和“三个走在前列”要求, 牢牢把握三项主要任务和五项职责, 按照省十二届
坚持稳中求进、充分履职, 着力为经济社会发展提供司法保障。
坚持在大局中谋划、推进检察工作, 继续全面深化改革充分履职, 为经济社会发展提供有力司法保障。
着力服务全面深化改革, 主动适应经济、政治、文化、社会和生态文明建设体制改革要求, 自觉当好改革的促进者、守护者。今年 1 至 6 月 (时间下同), 起诉非法集众、制假售假等破坏市场经济秩序犯罪 1014 人, 查办
扎实开展三个专项工作, 针对代表反映的企业内部人员侵占资产资金、“告状难”、国家工作人员向企业“吃拿卡要”、涉企案件裁判不公等问题, 突出抓好惩治涉及企业的违法犯罪专项工作, 促进企业健康发展。
正确把握法律政策界限和办案方式方法, 贯彻省委“护、告、不、敢”和建设性司法司法要求, 落实代表相关意见, 开展“公”与“私”界限问题专题调研, 正确区分改革发展、招商引资、实行优惠政策近
二、坚持严格执法、强化监督, 着力促进业务工作健康发展
依法严格履行宪法和法律赋予的各项职责, 维护社会大局稳定, 促进社会公平正义, 保障人民安居乐业。
全力维护社会和谐稳定。把维护政治稳定放在首位, 把人民群众对平安的期待作为努力方向, 积极刑事犯罪案件 13987 人, 起诉 19400 人, 积极参与严厉打击暴力恐怖活动专项行动, 严惩“法轮功”、“全能神”
积极查办和预防职务犯罪。认真贯彻落实中央部署的反腐败体制机制改革和制度建设, 严格执行中央决策部署, 坚持有案必查、有腐必惩, 立案侦查各类职务犯罪 1945 人, 同比上升 11%; 查办县处级以上干部职务犯罪 169 人 (含行刑)
全面加强诉讼活动的法律监督。针对人民群众反映强烈的执法不严司法不公问题, 强化对诉讼活动的法律监督, 努力让人民群众在每一个司法案件中都能感受到公平正义。加强刑事诉讼监督, 对应当立案而不立案初
三、坚持深化改革、积极创新, 着力提升检察工作活力
认真贯彻落实中央、省委和最高人民检察院部署, 结合湖北实际, 扎实推进司法改革、检察改革和工作机制创新。
全力抓好司法体制改革试点准备工作。作为司法体制改革 6 个试点省份之一, 省检察院在中央、省委、最高人民检察院领导下, 高度重视改革试点工作, 专门成立司法改革领导小组, 把握正确方向, 先后 8 次召开专
认真落实最高人民检察院部署的四项改革任务。深入推进涉法涉诉信访改革, 完善受理、审查机制, 实行信访分离、释法说理、息诉维稳和司法救助, 妥善处置涉法涉诉信访 6617 件。深化检务公开,
深入推进三项工作机制创新。推进诉讼监督工作制度化、规范化、程序化、体系化, 省检察院成立专班, 对 2013 年全省 3 万余件诉讼监督案件进行全面核查, 针对存在的问题, 推行量化立案及处理标准, 规范办案质
四、坚持强基固本、提升公信, 着力促进自身建设
以改革创新精神加强自身建设, 为严格公正廉洁司法奠定坚实基础。
狠抓过硬检察队伍建设。深入贯彻“五个过硬”要求, 全面加强检察队伍建设。把学习贯彻习近平总书记系列重要讲话精神落实到行动上, 组织全省检察机关领导干部专题培训, 进一步坚定理想信念, 确保正确政

最后是工作报告的专题报告部分的展示，有 13 个项目，5 个 txt 文件和 8 张 jpg 图片。



王晋检察长向省人大常委会作《关于贯彻实施决定深入开展扫黑除恶专项斗争情况的报告》

时间: 2019-11-26

来源:

访问量:

图为湖北省第十三届人民代表大会常务委员第十二次会议现场

本网讯 (记者 蔡欣 摄影 刘化梅) 11月26日,湖北省人民检察院党组书记、检察长王晋在湖北省第十三届人民代表大会常务委员第十二次会议上作《关于贯彻实施决定深入开展扫黑除恶专项斗争情况的报告》

图为湖北省检察院检察长王晋向湖北省十二届人大常委会第十二次会议作《关于贯彻实施决定深入开展扫黑除恶专项斗争情况的报告》

报告显示,扫黑除恶专项斗争开展以来,全省检察机关认真学习领会习近平总书记系列重要讲话和指示批示精神,深入贯彻中央和省委、高检院决策部署,全面落实省人大常委会《关于深入开展扫黑除恶专项斗争王晋检察长在报告中从持续加大办案力度,保持高压态势;持续强化法律监督,促进严格执法、公正司法;持续深化“打伞断财”,铲除滋生土壤;持续参与综合治理,助力源头防范;持续加强组织保障,夯实持续加大办案力度,保持高压态势

湖北省检察机关按照《决定》“切实打好扫黑除恶专项斗争主动攻坚战整体仗”的要求,坚持以办案为中心,对黑恶势力犯罪出重拳、下重手,自专项斗争开展以来,共批捕涉黑涉恶犯罪嫌疑人2625件6349人,起诉1061通过提前介入和跟踪审查等方式深挖彻查,共监督立案黑恶势力犯罪案件104件185人;追捕涉黑涉恶犯罪嫌疑人1295人、追诉393人。充分发挥领导干部示范带头作用,131名检察长、副检察长以主办检察官、持续强化法律监督,促进严格执法、公正司法

按照《决定》“确保全省扫黑除恶专项斗争在法治轨道上运行”的要求,坚守客观公正立场,切实把好案件事实关、证据关、程序关和法律适用关,做到“不是涉黑涉恶的,一个不凑数;是涉黑涉恶的,一个不放过;专项斗争以来,全省检察机关共提出补充完善证据等工作建议8900余条,增(追)加犯罪事实1472笔,纠正侦查活动违法46件次;共对2563名黑恶案件犯罪嫌疑人开展讯问合法性核查,排除非法证据1177份持续深化“打伞断财”,铲除滋生土壤

按照《决定》“坚持依法严惩、综合治理、标本兼治”的要求,在惩治黑恶势力犯罪的同时,深挖其背后的“关系网”“保护伞”,对其非法聚集的财物依法追缴,不给黑恶势力卷土重来、死灰复燃的机会。全省检察机关向纪检监察机关移送“保护伞”线索774条;依法办理“保护伞”案件,受理“保护伞”等案件310件,立案侦查司法人员涉嫌徇私枉法等“保护伞”案件12件13人;及时引导侦查机关准确、全面

按照《决定》“大力加强基层基础建设,健全防止黑恶势力滋生的长效机制”的要求,把参与综合治理作为办理黑恶势力犯罪案件的硬任务,增强整治的辐射力。全省检察机关向有关部门提出关于长江生态保护、规范房地产中介市场和金融市场秩序等各类检察建议317份;选派837名检察官担任中小学法治副校长,防止未成年人被黑恶势力拉拢利用;开展多层次、多辉持续加强组织保障,夯实工作基础

按照《决定》“压实扫黑除恶专项斗争的政治责任和法律责任”的要求,把扫黑除恶专项斗争作为“一把手”工程,在组织领导、机制建设、能力水平等方面下功夫,确保扫黑除恶专项斗争有序有效推进。全省检察机关组建了1600余名专项斗争队伍,共投入专项经费5581万元;加强实战练兵,举行“全省公诉人与律师电视论辩赛”;加强高端人才培养,遴选32名业务骨干组建全省检察机关扫黑除恶专项斗争专家人才王晋检察长表示,下一步,全省检察机关将以学习贯彻十九届四中全会精神为统领,认真落实会议的审议意见,持续发力、攻坚克难,推动扫黑除恶专项斗争不断取得新突破,为湖北落实促进中部地区崛起战略

部署,本次会议上,王晋检察长还作了《关于湖北省第十三届人民代表大会第二次会议代表建议、批评和意见办理情况的报告》。2019年,省检察院共承办代表建议7件,内咨涉及司法办案、诉讼监督、司法公

第四章 感想及参考文献

4.1 心得感想

在本次 python 爬虫爬取网址内容这个项目完成的过程中,我学到了很多东 西。前几天的基础知识虽然很多,但由于是听讲的方式,并没有真正上手,所有 感触并不是很深。即使每天都有作业,但作业比较简单,并不能让我深刻领略 python 的魅力。后来中科大的研究生对 python 项目的介绍真正让我开了眼界, 说实话,我听得不是很懂,很多东西都是云里雾里,但毫无疑问的是,我感受到了 python 的强大,我对它有了浓厚的兴趣。后来,我和许敬一同学合作制作 web 网页更是让我看到了 python 的实用性,它利用很多库和模块,使得编程变得很 简单很容易上手,着实让我兴奋不已。

在做本次这个小项目时,我一开始是非常茫然地,虽然中科大的研究生以及 曹鑫老师都已经给我们介绍过了 python 爬虫的基础知识,但当我自己亲手来做 时还是很茫然的。不过我很快调整了心态,查看了网页,心中就有了基础的框架, 后来我又上网查询了资料确定了大致的方案,这时候我就比较信心十足了。但是, 真正在编程的时候,我又遇到了各种各样的困难,我在本报告书的第二部分也有 提到,但实际上,我真正遇到的困难远不止这些,我还遇到了一些比如格式、语 法、基础知识等的小困难,完成这个项目实在不容易。不过,我并没有被困难打 倒,我一步一步逐渐地解决了所有的困难,完成了这个项目。在这个过程之中, 我亲身经历了代码调试的过程,经历了 python 编程的过程,真切感受到了 python 的魅力。

在未来的日子里,希望我能运用好 python 这个工具,也希望我能将在学习 python 过 程中的感悟融入到生活的点滴之中,给我们未来以启示。

4.2 参考文献

- [1]CSDN python 爬虫基础及实例——代码经过实测 https://blog.csdn.net/q_q_39481696/article/details/82493493
- [2]CSDN python 爬虫——第一个小爬虫（经典）修改版 https://blog.csdn.net/weixin_42247452/article/details/82317065
- [3]CSDN 第一个 Python 爬虫 https://blog.csdn.net/sunon_/article/details/90634253