python与深度学习基础第一次作业

PB20000010母子跃

2023/05/05

实验内容

利用python爬取博客上单一作者的所有文章保存到本地docx文件中

看上博客上一个作者的文章，想一次性下载到一个word文件中，输入作者的个人主页网址， 导入到本地的docx文档中，通过word的“导航窗格”快速定位单篇文章。具体的实现如下：

1. 先获取到所有文章的标题、发表日期、链接
2. 通过链接获取文章的内容
3. 将文章标题作为“1级”，发表日期和内容作为正文写入word文件
4. 保存wrod文件 下面就按照以上步骤进行操作。

实验思路

本代码分为两个文件： 第一个文件用于实现包括爬取、导入等功能操作。 第二个文件用于实现图形化界面展示，包括输入和输出的展示。 第二个文件为主文件，通过subprocess库来python调用来执行第一个python文件。 爬取使用了request库和bs4中的beautifulsoup库中的函数，保存到本地docx用了python的docx库。 图形化界面展示使用了tkinter库

文件一：爬取并保存

先进入到目标博客的主页，点击“博文目录”，这样就在网址栏看到“blog.sina.com.cn/s/articlelist_5119330124_0_1.html” 。

再点击下一页，可以看到网址末尾的“1”变成了“2”。这样我们就知道所有页对应的网址了（尾号从1到5）。

先挑第一页的网址，定位我们需要的信息，以便后续批量爬取。 在博文的标题和发表日期上分别点右键，选择“检查”. 可见博文标题和博文链接都位于 `class="atc_title"` 下面，发表时间位于 `class="atc_tm SG_txtc"` 下面。因此使用 `soup.select('.atc_title')` 就可以获取当前网页的所有博文的链接和标题；使用 `soup.select('.atc_tm')` 可获取所有博文的发表日期。

In [2]:

```
import requests
from bs4 import BeautifulSoup
#测试数据：http://blog.sina.com.cn/s/articlelist_5119330124_0_1.html
url=input("请输入博客作者的网页：")
```

```
请输入博客作者的网页：https://blog.sina.com.cn/s/articlelist_1223240311_0_1.html
```

In [3]:

```
wb_data = requests.get(url)
soup = BeautifulSoup(wb_data.content)
```

In [4]:

```
#获取当页所有文章的标题和链接
soup.select('.atc_title')
```

Out[4]:

```
[<span class="atc_title">
 <a href="//blog.sina.com.cn/s/blog_48e92a770102z5i2.html" target="_blank" t
itle="">《中国人在德国》连载265《德国是…</a></span>,
 <span class="atc_title">
 <a href="//blog.sina.com.cn/s/blog_48e92a770102z5i0.html" target="_blank" t
itle="">《中国人在德国》连载264《在德国…</a></span>,
 <span class="atc_title">
 <a href="//blog.sina.com.cn/s/blog_48e92a770102z5hz.html" target="_blank" t
itle="">《中国人在德国》连载263《在德国…</a></span>,
 <span class="atc_title">
 <a href="//blog.sina.com.cn/s/blog_48e92a770102z5hv.html" target="_blank" t
itle="">《中国人在德国》连载262《德国，…</a></span>,
 <span class="atc_title">
 <a href="//blog.sina.com.cn/s/blog_48e92a770102z5hu.html" target="_blank" t
itle="">《中国人在德国》连载261《在教授…</a></span>,
 <span class="atc_title">
 <a href="//blog.sina.com.cn/s/blog_48e92a770102z5ht.html" target="_blank" t
```

In [5]:

```
#获取当页所有文章的发表时间
soup.select('.atc_tm')
```

Out[5]:

```
[<span class="atc_tm SG_txtc">2023-04-19 08:01</span>,
 <span class="atc_tm SG_txtc">2023-04-17 08:03</span>,
 <span class="atc_tm SG_txtc">2023-04-14 09:26</span>,
 <span class="atc_tm SG_txtc">2023-04-12 08:16</span>,
 <span class="atc_tm SG_txtc">2023-04-10 07:52</span>,
 <span class="atc_tm SG_txtc">2023-04-07 08:29</span>,
 <span class="atc_tm SG_txtc">2023-04-05 08:48</span>,
 <span class="atc_tm SG_txtc">2023-04-03 08:22</span>,
 <span class="atc_tm SG_txtc">2023-03-31 08:34</span>,
 <span class="atc_tm SG_txtc">2023-03-29 08:25</span>,
 <span class="atc_tm SG_txtc">2023-03-27 08:38</span>,
 <span class="atc_tm SG_txtc">2023-03-24 08:46</span>,
 <span class="atc_tm SG_txtc">2023-03-22 08:32</span>,
 <span class="atc_tm SG_txtc">2023-03-20 08:31</span>,
 <span class="atc_tm SG_txtc">2023-03-17 09:35</span>,
 <span class="atc_tm SG_txtc">2023-03-15 08:34</span>,
 <span class="atc_tm SG_txtc">2023-03-10 09:17</span>,
```

如上获取的文章标题及链接信息是存在一个大列表中的。现在以第一个元素为例从中提取出链接和标题信息。观察发现链接位于 a 标签里的 href 里面，于是使用 select 方法选中 a 标签，可以看到结果是一个新的列表（如下）。

In [6]:

```
soup.select('.atc_title')[0].select('a')
```

Out[6]:

```
[<a href="//blog.sina.com.cn/s/blog_48e92a770102z5i2.html" target="_blank" title
="">《中国人在德国》连载265《德国是…</a>]
```

然后再从这个新列表中提取出链接和标题。使用 get("href") 方法获得链接；使用 text 方法获得标题。

In [7]:

```
soup.select('.atc_title')[0].select('a')[0].get("href")
```

Out[7]:

```
'//blog.sina.com.cn/s/blog_48e92a770102z5i2.html'
```

In [8]:

```
soup.select('.atc_title')[0].select('a')[0].text
```

Out[8]:

```
'《中国人在德国》连载265《德国是…'
```

发表时间的获取就简单很多了，直接用 text 方法即可。

In [9]:

```
soup.select('.atc_tm')[0].text
```

Out[9]:

```
'2023-04-19 08:01'
```

单页的信息搞定，然后就可以批量处理了。使用 for 循环遍历所有页，然后逐个提取。因为我们已知作者的文章共有5页，所以直接使用 range(1,6) 。将最终的信息存入字典 all_links 。其中，“标题”作为键，文章链接和发表时间作为值。通过 len(all_links) 查看获取的文章链接数，一共211篇文章。

In [10]:

```
#获取所有博客文章的链接
import requests
from bs4 import BeautifulSoup

all_links = {}
for i in range(1,6):
    wb_data = requests.get(url)
    soup = BeautifulSoup(wb_data.content)
    links = soup.select('.atc_title')
    times = soup.select('.atc_tm')
    for i in range(len(links)):
        http_link = links[i].select('a')[0].get('href')
        c="https:"
        link=c+http_link
        title = links[i].text.strip()
        time = times[i].text
        all_links[title] = [link, time]
```

In [11]:

```
len(all_links)
```

Out[11]:

50

In [12]:

```
all_links
```

Out[12]:

{'《中国人在德国》连载265《德国是…': ['https://blog.sina.com.cn/s/blog_48e92a770102z5i2.html', '2023-04-19 08:01'], '《中国人在德国》连载264《在德国…': ['https://blog.sina.com.cn/s/blog_48e92a770102z5i0.html', '2023-04-17 08:03'], '《中国人在德国》连载263《在德国…': ['https://blog.sina.com.cn/s/blog_48e92a770102z5hz.html', '2023-04-14 09:26'], '《中国人在德国》连载262《德国，…': ['https://blog.sina.com.cn/s/blog_48e92a770102z5hv.html', '2023-04-12 08:16'], '《中国人在德国》连载261《在教授…': ['https://blog.sina.com.cn/s/blog_48e92a770102z5hu.html', '2023-04-10 07:52'], '《中国人在德国》连载260《德国污…': ['https://blog.sina.com.cn/s/blog_48e92a770102z5ht.html',

拿到所有文章链接后，先取一个来测试一下如何获取页面的文字。在文字上点右键，选择“检查”，可见其内容位于 class=articalContent　newfont_family 里面，因此使用 soup.select(".articalContent.newfont_family") 就可以获取到（注意articalContent和newfont_family之

间的空格要用"."代替）。将其存入 article 变量，显示一下，可以看到这是一个大列表，其中的文本就是我

In [13]:

```
#获取单篇文章中的文字
url = 'https://blog.sina.com.cn/s/blog_48e92a770102z5hi.html'
wb_data = requests.get(url)
soup = BeautifulSoup(wb_data.content)
article = soup.select(".articalContent.newfont_family")
article
```

Out[13]:

[<div class="articalContent newfont_family" id="sina_keyword_ad_area2">
<p style="margin-top:0cm;margin-right:0cm;margin-bottom:3.15pt;margin-left: 0cm;line-height:13.15pt;background:white">
《中国人在德国》
雅兰·著 <wbr/> <wbr> <wbr> <wbr> <wbr> <wbr> <wbr> <wbr> <wbr> <wbr> <wbr> <wbr> <wbr> <wbr> <wbr> <wbr> <wbr> <wbr> <wbr> <wbr/></wbr></wbr></wbr></wbr></wbr></wbr></wbr></wbr></wbr></wbr></wbr></wbr></wbr></wbr></wbr></wbr></wbr></wbr></p>
<p style="margin-top:0cm;margin-right:0cm;margin-bottom:3.15pt;margin-left: 0cm;line-height:13.15pt;background:white">
享有著权严禁刊载转载侵权必究！ <wbr/> <wbr> <wbr> <wbr> <wbr> <wbr> <wbr> <wbr> <wbr> <wbr> <wbr> <wbr> <wbr> <wbr>
<wbr> <wbr/></wbr></wbr></wbr></wbr></wbr></wbr></wbr></wbr></wbr></wbr></wbr></wbr></wbr></wbr></p>
<p style="margin-top:0cm;margin-right:0cm;margin-bottom:3.15pt;margin-left: 0cm;line-height:13.15pt;background:white">

《谦宇》篇</p>
<p style="margin-top:0cm;margin-right:0cm;margin-bottom:3.15pt;margin-left: 0cm;line-height:13.15pt;background:white">

</p>
<p> <wbr/>
 <wbr> <wbr> <wbr>
<wbr> <wbr> <wbr>
<wbr> <wbr> <wbr>
<wbr> <wbr> <wbr>
<wbr> <wbr> <wbr>
<wbr> <wbr> <wbr>
<wbr> <wbr> <wbr>
<wbr> <wbr>
<wbr/></wbr></wbr></wbr></wbr></wbr></wbr></wbr></wbr></wbr></wbr></wbr></wbr></wbr></wbr></wbr></wbr></wbr></wbr></wbr></wbr></wbr></wbr></wbr>微信公众号订阅：直接搜索“大千德国”</p>
<p>
《中国人在德国》珍藏版</p>
<p><b>
已被：中国国家图书馆、北京海淀区图书馆、北京大学图书馆、清华大学图书馆、海南大

学图书馆、南京图书馆、金陵图书馆、世界华商基金会、国际中医养生大会理事会、新媒体国际合作组织等海内外国家图书馆和机构收藏！

<wbr/>

……
011/
读硕士的这一年半，很辛苦
 <wbr/>

我印象比较深的，是学语言的时候。老师跟我们说，你们现在觉得学德语很难，很痛苦，当你们通过语言考试，进入大学后，你们会发现，这只是冰山一角。

我们大部分人，花了近一年的时间，才拿到DSH语言证书，属于德语的C1级别。听起来级别挺高，其实是对于德国人来说，这只是中学水平。

我们这些国内过来的大学生，受了九年的教育，天天上学读书，参加了全国难度最高的考试，千军万马过了那个独木桥，进入一本大学，进行学习。

那个年代的大学生，含金量还是很高的。但在德国人眼里，国内的学业不被认可，德语水平，也很吃力。某种程度上，我们等于是初中生，直接去上了大学。

可以说，还是有落差感的。

我在FH的专业是Photogrammetry and
Geoinformatics（摄影测量和地理信息），跟信息工程相近。

读硕士的这一年半，很辛苦。

不能再像读语言的时候，为了赚钱而逃课，我找了一个只在周六周日的兼职。这样，就不会耽误平时的学业。周一到周五，我可以全身心地投入学习。周六周日的两天，可以安心地打工。一周七天，全部都是占满的。

当时，除了每隔两三天，跟女朋友（现在的老婆）视频一下，聊聊天，或者给父母打电话，已经完全没有其他时间。每天，我都在不停地运转。</p>
<p style="text-indent:21.0pt">
也可以说很规律。生活中只有五件事：做饭，吃饭，睡觉，上课，打工。打一天工后，回来会很累的。</p>
<p style="text-indent:21.0pt">
毕业设计时，就显得很紧张。每天要做这么多事情，时间被压缩的特别厉害。所有的环节，都必须完成。当时这样，感觉时间，也过得非常快。整个脑子里，似乎没有其他东西。除了睡觉可以休息，其他时间，都在忙碌。</p>
<p style="text-indent:21.0pt">
<wbr> <wbr> <wbr>
<wbr> <wbr> <wbr>
<wbr> <wbr> <wbr>
<wbr> <wbr> <wbr>
<wbr> <wbr> <wbr>
<wbr> <wbr> <wbr>
<wbr> <wbr> <wbr>
<wbr> <wbr>
<wbr> <wbr><b style="text-align: center; text-indent: 21pt;">2006<b style="text-align: center; text-indent: 21pt;">年9月9日，国际田联田径大奖赛男子110米栏决赛中，</wbr></wbr></wbr></wbr></wbr></wbr></wbr></wbr></wbr></wbr></wbr></wbr></wbr></wbr></wbr></wbr></wbr></wbr></wbr></wbr></wbr></wbr></wbr></wbr></wbr></p>
<p style="text-indent:21.0pt"><b style="text-align: center; text-indent: 21pt;">
<wbr> <wbr> <wbr>
<wbr> <wbr> <wbr>
<wbr> 刘翔以12秒93夺得冠军，并打破赛会纪录。摄于斯图加特戴姆勒体育场</wbr></wbr></wbr></wbr></wbr></wbr></wbr></p>
<p style="text-indent:21.0pt">那时，打工的地方，搞笑的说法是：世界500强之一，汉堡王。</p>
<p style="text-indent:21.0pt">
那里，一般是女生做Kasse（收银），男生做后厨。我们是有调班的。有时候，我上早班，有时候上晚班。我的工作，就是准备食材、包汉堡和炸鸡块等。</p>
<p style="text-indent:21.0pt">
在食品店打工的好处就是，每天累了，能够免费吃到汉堡。有时候，也能带一些出来，给舍友。尽管店里有声明，不准往外带汉堡，但所有的员工，在走的时候，都会带上几个汉堡。因为汉堡只要是超过一个小时，就不允许再卖了，必须处理掉。扔掉，也是浪费，还不如带给同学吃。</p>
<p style="text-indent:21.0pt"><span style="mso-bidi-font-size: 10.5pt;font-family:宋体;mso-fareast-font-family:宋体;mso-fareast-theme-font:minor-fareast; colo

r:#0000CC">
那家汉堡王里的员工，有很多是波黑内战那段时间，从巴尔干地区来的难民，经理和小头都是这么来的。

他们对我们这些打工的学生，都挺客气。有时闲聊，会和我们说：“现在，你们也就是短时间，来这里打工，会辛苦一点。等到毕业了，就会有好工作的。但我们可能一辈子，都只能待在这个店工作了”。

有个黑人同事，人非常好，他原来是美军在德驻军。退役后，娶了个德国老婆，就一直留在德国了。后来，我终止工作，与他告别时，他还送给我一本英文版的斯诺的《红星照耀中国》。这本书，我一直没时间看。

我在水果厂打工的时候，也会带一些水果，给同学。不仅是我带东西给同学，同学有时也会带东西给我。同学之间，都是有时在这个厂，有时在那个厂。大家彼此之间，都会互相给。

因为斯图加特是打工城，有很多人，都是通过打工，来交朋友的

在汉堡王的工资待遇还行，每月400欧不到。听起来不高，一年下来，也能挣4000多欧。不仅学业不会受影响，生活费也有了保障，还能买上回国的机票。

暑假里的时间，还可以打另外一份工。就是去工厂的流水线，集中工作一个月。这一个月，可以有两三千欧得收入。再加上平时打工的收入，一年可以挣到六千多欧。这些足够支撑我在德国的生活。

光打工也不行，我的首要任务，还是学习。
FH的学习，比较累。一年半的学制虽然短，但强度很高。课程多，作业多，上机操作等实际训练也多。一个学期，有七八门课。要想按时毕业，每个学期，都必须通过这些课。我们当时的要求是，如果一门课，三次考不过，学生必须退学。

当时的班上，有个中国同学，打工打得太狠了。导致他有一门课，三次考试，都没过，退学了。后来，他转到德国的另外一个大学，把学业学完了。

我们都觉得，他很可惜。

那时，我们已经学到第二学期，都快毕业了。我们一般的补考，努力一下，基本上都是可

以过的。他已经处于补考阶段，可还在打工，这样，学业肯定是被耽误了。

<wbr/>

无论如何，到德国来，还是要以学习为主。打工什么的，平时无所谓，考试期间，好歹把重心转移一下。

……

书中有关话题与范畴，逐入：人文，社会，文化，哲学，历史，教育，医疗等；

具有很高的阅读价值，参考价值，史料价值，收藏价值……

 <wbr/>

对比、沉淀、融合、提升、超越，是《中国人在德国》的核心创作理念与思想。以此，全面有力的推动中国社会发展进程，真正做到：铸造大国，屹立东方，傲然世界！
 <wbr>
<wbr> <wbr> <wbr>
<wbr> <wbr> <wbr>
<wbr> <wbr> <wbr>
<wbr> <wbr> <wbr>
<wbr> <wbr/></wbr></wbr></wbr></wbr></wbr></wbr></wbr></wbr></wbr></wbr></wbr></wbr></wbr></wbr></wbr>
 <wbr>
<wbr> <wbr> <wbr>
<wbr> <wbr> <wbr>
<wbr> <wbr> <wbr>
<wbr> <wbr> <wbr>
<wbr> <wbr> <wbr>
<wbr> <wbr> <wbr>
<wbr> <wbr>
<wbr>扫码进《中国人在德国》读

者群　<wbr>　<wbr/></wbr></wbr></wbr></wbr></wbr></wbr></wbr></wbr></wbr></wbr></wbr></wbr></wbr></wbr></wbr></wbr></wbr></wbr></wbr></wbr></wbr></wbr></wbr></wbr></wbr></wbr></p>

<p style="margin:0cm;margin-bottom:.0001pt;line-height:15.75pt;background:white">
………………………………</p>
<p style="margin:0cm;margin-bottom:.0001pt;text-indent:.4pt;line-height:15.75pt; background:white">
　<wbr/></p>
<p style="margin:0cm;margin-bottom:.0001pt;text-indent:.4pt;line-height:15.75pt; background:white">
内容简介：</p>
<p style="margin:0cm;margin-bottom:.0001pt;text-indent:21.0pt;line-height:15.75pt; background:white">
　<wbr/></p>
<p style="margin:0cm;margin-bottom:.0001pt;text-indent:21.0pt;line-height:15.75pt; background:white">
这里的人物影像，不是直观谫浅。而是迄今为止，唯一集中呈示在德华人的一本专著。</p>
<p style="margin:0cm;margin-bottom:.0001pt;text-indent:21.0pt;line-height:15.75pt; background:white">
书中文字，具有穿透力。通过阅读，能够让读者感受到，他们心域的宽度，广度，深度，以及每一寸气息里，饱含的温度。</p>
<p style="margin:0cm;margin-bottom:.0001pt;text-indent:21.0pt;line-height:15.75pt; background:white">
他们是不同的。个人成长，教育背景，出国原因，融入状态，东西方文化的碰撞与冲突；他们又都是相同的，他们都是华人在德国的一个缩影。</p>
<p style="margin:0cm;margin-bottom:.0001pt;text-indent:21.0pt;line-height:15.75pt; background:white">
他们每个人的内心，都是一个广袤的未知世界。也正是由此，他们构铸了在德华人的大千万象。</p>
<p style="margin:0cm;margin-bottom:.0001pt;text-indent:21.0pt;line-height:15.75pt; background:white">
就国家与民族而言，他们是弱小的，但在生命的进程中
，在不屈的闯拓中，他们又都是强大而无憾的！</p>
<p style="margin:0cm;margin-bottom:.0001pt;text-indent:21.0pt;line-height:15.75pt; background:white">
他们每个人的声音，都是中国的心跳。他们每个人的步履，都是中国的深脉。虽然他们都在德国，但他们都是有着一颗中国心！</p>
<p style="margin:0cm;margin-bottom:.0001pt;text-indent:21.0pt;line-height:15.75pt; background:white">
文字，是他们的灵魂，是无声的音符。</p>
<p style="margin:0cm;margin-bottom:.0001pt;text-indent:21.0pt;line-height:15.75pt; background:white">
不同阶层的阅读者，从中，都能找到自己想要的答案…</p>
<p style="margin:0cm;margin-bottom:.0001pt;text-indent:21.0pt;line-height:15.75pt; background:white">
　<wbr/></p>
<p style="margin:0cm;margin-bottom:.0001pt;line-height:15.75pt;background:white">
…………………………………………………………………………</p>
<p style="margin:0cm;margin-bottom:.0001pt;line-height:15.75pt;background:white">

 <wbr/>

书中主要内容：

 <wbr/>

着力刻画人物的成长与蜕变。历尽万象艰辛后，在异国他乡，最终形成了独立个体。更深层的，是想把在德国的华人历练后的生存状态，用文字呈现给海内外的中国读者……

所有的美景，不是天然而成。凤凰涅槃，烈焰燃烧后的重生，由内而外，都是极致的景象。

“他们当中，有餐厅老板，自由职业者，科学家，机构职员（联邦政府公职），公司雇员，射击协会会长，律师，工程师，赛车教练，留学生，年轻夫妻等，他们是不同的。个人成长，教育背景，出国原因，融入状态，东西方文化的碰撞与冲突；他们又都是相同的，他们都是华人在德国的一个缩影。他们是浓墨，他们是重彩……他们每个人的内心，都是一个广袤的未知世界。也正是由此，他们构铸了在德华人的大千万象……”

 <wbr/>

“所有的一切，抵达到他们，即是切面，也是交错融合的焦点。”
这是序言《他们是浓墨，他们是重彩》中的文字。

 <wbr/>

在构思《中国人在德国》时，我是想尽可能，让每个人物，都能鲜活饱满起来。这只是外在的阅读感受……随着文字的流淌和加深，再让读者，去触探每个人物的内心。思想。及灵魂……再扩充到大我的家国情怀！根系处，仍是人物（被采访者）和读者难以割舍的羁绊与绕缠：一个小我的人，到底该怎样去生存，去融合，去追索等方面的思考……另一方面，作者也想表达：一个人，又该怎样活着，才能真正拥有人生的价值和意义！？

 <wbr/>

整本书，处处都是看点。章章都有收获！

 <wbr/>

……

 <wbr/><span style="font

-size:11.0pt;color:red">《中国人在德国》腰封文案：

 <wbr/>

一、《中国人在德国》

灵魂，都已打开了！只在等你，静心阅读……

 <wbr/> <wbr/>

亮点阅读：

 <wbr/>

《中国人在德国》，不仅在国内会产生影响（尤其是青年学子），在德国的华人界也会产生影响…因为：迄今为止，无论是国内作家还是华人作家，都还没有写出一本《中国人在德国》的书。或者是类似的书…就像凌鼎年老师说的，写出欧洲的，也没有…《中国人在德国》是第一本。

 <wbr/>

读者，通过此书的阅读，也会产生更强的民族凝聚力量。

华人，他们每个人，不是一个人。他们是一群人。一个群体。他们是一种符号和象征…在德国，他们是中国人…他们是中国。他们的血管里，流淌着中国血液…他们站在德国的大地上。他们的身影，毅力着。从没倒下过…他们的身后，有强大的中国，在做支撑！所有文字，都是表象…物质，不仅有表象。真正有价值的，都是潜藏在内核…

 <wbr/>

书中人物，除一人之外，其他人，都是以留学生身份，去往德国。

现在的中国，是留学大国。每年，去国外留学的学生人数，都在递增。德国因为免学费，所以，去德国留学，是很多国内中层阶级孩子的向往。

 <wbr/>

学生通过《中国人在德国》，可以了解到，留学德国后，所要面临和将来需要解决的各种问题。也可谓：此书，即是一本留学指南针<span style="font-size:

10.5pt;color:#0000CC">…对于想去德国旅居和定居的中国人来说，此书，也可谓是一把解剖德国各阶层的手术刀。因为书中人物，已经渗透至德国社会各阶层，他们的亲临感受，让此类读者，通过文字阅读，即能做到恰当的选择与规避。以便让自己，能够在德国，求得最好的生存方式。

学子睿才，追求与思考并行。书中，尽显东西方文化与教育的碰撞，互通，共融。

通过阅读，人文，社会，文化，哲学，历史，教育，医疗等领域的读者，都会寻索到自己想要的答案……

 <wbr/>

地球，已趋国际化…《中国人在德国》方方面面的生活、经历，所见所闻，所思所想，话题广泛，可令读者增广见闻。

对于国内的各阶层阅读者，此书，既是了解德国的一个界面，也能让阅读者，通过文字本身，能够更深层的了解到：德国独有的文化特色和社会风情。通过阅读，神思上，会产生无尽遐想……

……

作者简介

雅兰，著名作家，社会学家，旅德作家。以思想敏锐，文字敏锐，视角独特著称。游离于言论、散文、小说、诗歌等诸多文体。在公开发行的刊物及相关网页开设各版《雅兰专栏》。著书《中国很高兴》、《性殇》、《从压抑到泛滥》、《在我离开你之前》。

因创作微型小说，而引发的风暴漩流，被中国文坛列为《新世纪以来的文学事件》之一。被《北京文学》杂志评为“2004年中国十大文化热点之一”。轰动中国文化界。与此同时，引发海内外海量媒体的关注与跟进报道。并在2018年，入选“江苏省改革开放40年最具影响的40篇微型小说”。

 <wbr/>……

雅兰信箱:

<wbr/>

直接使用 text 方法就能提取出来。“\xa0”是个什么鬼？明显不是我们要的，百度了一下，说是什么不间断空格符，管他呢，直接使用 replace("\xa0", "") 删掉，这下就美丽了。“\n”是换行，就不要删了，保持原格式比较好。

In [14]:

```
article[0].text
```

Out[14]:

```
'\n\n《中国人在德国》\n雅兰·著\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\n\n享有著权严禁刊载转载侵权必究！\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\n\xa0 \xa0\n\n\n《谦宇》篇\n\n\n\xa0\xa0 \xa0 \xa0\n\xa0 \xa0 \xa0\n\xa0 \xa0 \xa0\n\xa0 \xa0 \xa0\n\xa0 \xa0 \xa0\n\xa0 \xa0 \xa0\n\xa0 \xa0\n\xa0微信公众号订阅：直接搜索“大千德国” \n\n《中国人在德国》珍藏版\n\n已被：中国国家图书馆、北京海淀区图书馆、北京大学图书馆、清华大学图书馆、海南大学图书馆、南京图书馆、金陵图书馆、世界华商基金会、国际中医养生大会理事会、新媒体国际合作组织等海内外国家图书馆和机构收藏！\n\n\xa0\n\n…………\n011/ \n读硕士的这一年半，很辛苦\n\xa0\n\n我印象比较深的，是学语言的时候。老师跟我们说，你们现在觉得学德语很难，很痛苦，当你们通过语言考试，进入大学后，你们会发现，这只是冰山一角。\n\n我们大部分人，花了近一年的时间，才拿到DSH语言证书，属于德语的C1级别。听起来级别挺高，其实是对于德国人来说，这只是中学水平。\n\n我们这些国内过来的大学生，受了九年的教育，天天上学读书，参加了全国难度最高的考试，千军万马过了那个独木桥，进入一本大学，进行学习。\n\n那个年代的大学生，含金量还是很高的。但在德国人眼里，国内的学业不被认可，德语水平，也很吃力。某种程度上，我们等于是初中生，直接去上了大学。\n\n可以说，还是有落差感的。\n\n我在FH的专业是
```

In [15]:

```
article[0].text.replace("\xa0","")
```

Out[15]:

```
'\n\n《中国人在德国》\n雅兰·著\n\n享有著权严禁刊载转载侵权必究！\n \n\n\n《谦宇》篇\n\n\n \n \n \n \n \n \n \n微信公众号订阅：直接搜索“大千德国”\n\n《中国人在德国》珍藏版\n\n已被：中国国家图书馆、北京海淀区图书馆、北京大学图书馆、清华大学图书馆、海南大学图书馆、南京图书馆、金陵图书馆、世界华商基金会、国际中医养生大会理事会、新媒体国际合作组织等海内外国家图书馆和机构收藏！\n\n\n\n…………\n011/ \n读硕士的这一年半，很辛苦\n\n\n我印象比较深的，是学语言的时候。老师跟我们说，你们现在觉得学德语很难，很痛苦，当你们通过语言考试，进入大学后，你们会发现，这只是冰山一角。\n\n我们大部分人，花了近一年的时间，才拿到DSH语言证书，属于德语的C1级别。听起来级别挺高，其实是对于德国人来说，这只是中学水平。\n\n我们这些国内过来的大学生，受了九年的教育，天天上学读书，参加了全国难度最高的考试，千军万马过了那个独木桥，进入一本大学，进行学习。\n\n那个年代的大学生，含金量还是很高的。但在德国人眼里，国内的学业不被认可，德语水平，也很吃力。某种程度上，我们等于是初中生，直接去上了大学。\n\n可以说，还是有落差感的。\n\n我在FH的专业是Photogrammetry and\nGeoinformatics（摄影测量和地理信息），跟信息工程相近。\n\n读硕士的这一年半，很辛苦。\n\n不能再像读语言的时候，为了赚钱而逃课，我找了一个只在周六周日的兼职。这样，就不会耽误平时的学业。周一到周五，我可以全身心地投入学习。周六周日的两
```

```
<p style="margin:0cm;margin-bottom:.0001pt;text-indent:21.0pt;line-height:15.75
pt; background:white">
<b><span style="font-size:9.0pt;font-family:">shuiyanqinglian@126.com</span></b></p>
<p style="margin:0cm;margin-bottom:.0001pt;text-indent:21.0pt;line-height:15.75
pt; background:white">
<b><span style="font-size:9.0pt;font-family:"><br/></span></b></p>
<p style="margin:0cm;margin-bottom:.0001pt;text-indent:21.0pt;line-height:15.75
pt; background:white">
<b><span style="font-size:9.0pt;font-family:"><br/></span></b></p>
<p style="margin:0cm;margin-bottom:.0001pt;text-indent:21.0pt;line-height:15.75
pt; background:white">
<b><span style="font-size:9.0pt;font-family:"><br/></span></b></p>
<p style="margin-top:0cm;margin-right:0cm;margin-bottom:3.15pt;margin-left: 0cm;line-height:13.15pt;background:white">
<br/></p>
<p style="margin-top:0cm;margin-right:0cm;margin-bottom:3.15pt;margin-left: 0cm;line-height:13.15pt;background:white">
<br/></p>
<p style="margin-top:0cm;margin-right:0cm;margin-bottom:3.15pt;margin-left: 0cm;line-height:13.15pt;background:white">
<br/></p>
<p style="margin-top:0cm;margin-right:0cm;margin-bottom:3.15pt;margin-left: 0cm;line-height:13.15pt;background:white">
<br/></p>
<p style="margin-top:0cm;margin-right:0cm;margin-bottom:3.15pt;margin-left: 0cm;line-height:13.15pt;background:white">
<br/></p>
<p style="margin-top:0cm;margin-right:0cm;margin-bottom:3.15pt;margin-left: 0cm;line-height:13.15pt;background:white">
<br/></p>
<p style="margin-top:0cm;margin-right:0cm;margin-bottom:3.15pt;margin-left: 0cm;line-height:13.15pt;background:white">
<br/></p>
<p style="margin-top:0cm;margin-right:0cm;margin-bottom:3.15pt;margin-left: 0cm;line-height:13.15pt;background:white">
<br/></p>
</div>
```

In [3]:

```
#获取单篇文章中的图片链接
wb_data = requests.get(url)
soup = BeautifulSoup(wb_data.content)
img_link = soup.select(".articalContent.newfont_family")[0].find_all("img")[0].get("real_src")
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
~\AppData\Local\Temp\ipykernel_13644\1716739296.py in <module>
      1 #获取单篇文章中的图片链接
----> 2 wb_data = requests.get(url)
      3 soup = BeautifulSoup(wb_data.content)
      4 img_link = soup.select(".articalContent.newfont_family")[0].find_a
ll("img")[0].get("real_src")

NameError: name 'requests' is not defined
```

In [21]:

```
soup.select(".articalContent.newfont_family")[0].find_all("img")[0].get("real_src")
```

Out[21]:

```
'https://s15.sinaimg.cn/mw690/001kMAGrzy840EQ9EPD19&690'
```

In [22]:

```
#图片下载函数
def downloadImg(img_url, file_path):
    req = requests.get(url=img_url)
    with open(file_path, 'wb') as f:
        f.write(req.content)
downloadImg(url,'1.jpg')
```

以上理顺，就可以大刀阔斧地开干了。定义一个函数 to_word ，一个参数，就是上面获取到的数据字典 all_links 。设定好 header ，假装是浏览器在访问。然后新建一个word文档，设置全局字体为宋体。因为有些文章被加密，无法访问并获取内容，所以最终获取到的文章数不一定等于链接数。于是增加一个初始值为0的计数器，用于记录写入word文档中的文章数，以便心中有数。然后遍历所有文章的标题，将标题按照“1级”写入word文档，这样才能在“导航窗格”看到文章目录，方便后续选取阅读。日期和内容都作为段落写入。有些文章被加密，获取不到内容，此时 article 变量为空，所以加个if语句判断，以免程序崩溃。每写入一篇文章，计数器自动加1，然后通过 print 输出信息。最后保存文件，366页，35万字的博客就到手了，结果是美丽的！从此阅读博客文章轻松多了。

In [23]:

```
#写入标题，内容到word文件
import docx
from docx.oxml.ns import qn #用于应用中文字体

def to_word(all_links):
    header = {"User-Agent":"Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.36 (KHTML, like
    doc=docx.Document() #新建word文档
    doc.styles['Normal'].font.name=u'宋体'
    doc.styles['Normal']._element.rPr.rFonts.set(qn('w:eastAsia'), u'宋体')

    counter = 0 #计数器，用于记录写入word的文章数
    for title in all_links.keys():
        doc.add_heading(title,1)
        date = all_links[title][1][:10]#只取日期，不要时间
        doc.add_paragraph(date)
        wb_data = requests.get(all_links[title][0],headers = header)
        soup = BeautifulSoup(wb_data.content)
        article = soup.select(".articalContent.newfont_family")
        #有些文章被加密，获取不到内容，此时article为空，所以加个if语句判断
        if article:
            text = article[0].text.replace("\xa0","")
            doc.add_paragraph(text)
            print(f"写入文章 {title} 。")
            counter += 1
    print(f"共写入 {counter} 篇文章。")
    doc.save("新浪微博文章.docx")

to_word(all_links)
print('保存完成')
```

写入文章 《中国人在德国》连载265《德国是… 。
写入文章 《中国人在德国》连载264《在德国… 。
写入文章 《中国人在德国》连载263《在德国… 。
写入文章 《中国人在德国》连载262《德国，… 。
写入文章 《中国人在德国》连载261《在教授… 。
写入文章 《中国人在德国》连载260《德国污… 。
写入文章 《中国人在德国》连载259《骗我的… 。
写入文章 《中国人在德国》连载258《在德国… 。
写入文章 《中国人在德国》连载257《我是一… 。
写入文章 《中国人在德国》连载256《德国不… 。
写入文章 《中国人在德国》连载255《我们的… 。
写入文章 《中国人在德国》连载254《我更看… 。
写入文章 《中国人在德国》连载253《餐厅楼… 。
写入文章 《中国人在德国》连载252《读硕士… 。
写入文章 《中国人在德国》连载251《坚持到… 。
写入文章 《中国人在德国》连载250《德国对… 。
写入文章 《中国人在德国》连载249《中德最… 。
写入文章 《中国人在德国》连载248《南北朝… 。
写入文章 《中国人在德国》连载247《我是这… 。
写入文章 《中国人在德国》连载246《我读了… 。
写入文章 《中国人在德国》连载245《我在德… 。
写入文章 《中国人在德国》连载244《斯特加… 。
写入文章 《中国人在德国》连载243《DSH不是… 。
写入文章 《中国人在德国》连载242《德国免… 。
写入文章 《中国人在德国》连载241《德国人… 。
写入文章 《中国人在德国》连载240《德国人… 。
写入文章 《中国人在德国》连载239《就从德… 。
写入文章 《中国人在德国》连载238《国内的… 。
写入文章 《中国人在德国》连载237《人，竟… 。
写入文章 《中国人在德国》连载236《都可以… 。
写入文章 《中国人在德国》连载235《在德国… 。
写入文章 《中国人在德国》连载234《从2010… 。
写入文章 《中国人在德国》连载233《在心里… 。
写入文章 《中国人在德国》连载232《我所谓… 。
写入文章 《中国人在德国》连载231《没有最… 。
写入文章 《中国人在德国》连载230《我是这… 。
写入文章 《中国人在德国》连载229《只有我… 。
写入文章 《中国人在德国》连载228《帮一次… 。
写入文章 《中国人在德国》连载227《语言班… 。
写入文章 《中国人在德国》连载226《中德之… 。
写入文章 《中国人在德国》连载225《因为德… 。
写入文章 《中国人在德国》连载224《在德国… 。
写入文章 《中国人在德国》连载223《本来，… 。
写入文章 《中国人在德国》连载222《现在，… 。
写入文章 《中国人在德国》连载221《我会更… 。
写入文章 《中国人在德国》连载220《移民对… 。
写入文章 《中国人在德国》连载219《我从不… 。
写入文章 《中国人在德国》连载218《德国改… 。
写入文章 《中国人在德国》连载217《你究竟… 。
写入文章 《中国人在德国》连载216《丧尽天… 。
共写入 50 篇文章。

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
~\AppData\Local\Temp\ipykernel_11400\4021679727.py in <module>
     27 
     28 to_word(all_links)
---> 29 print(保存完成)

NameError: name '保存完成' is not defined
```

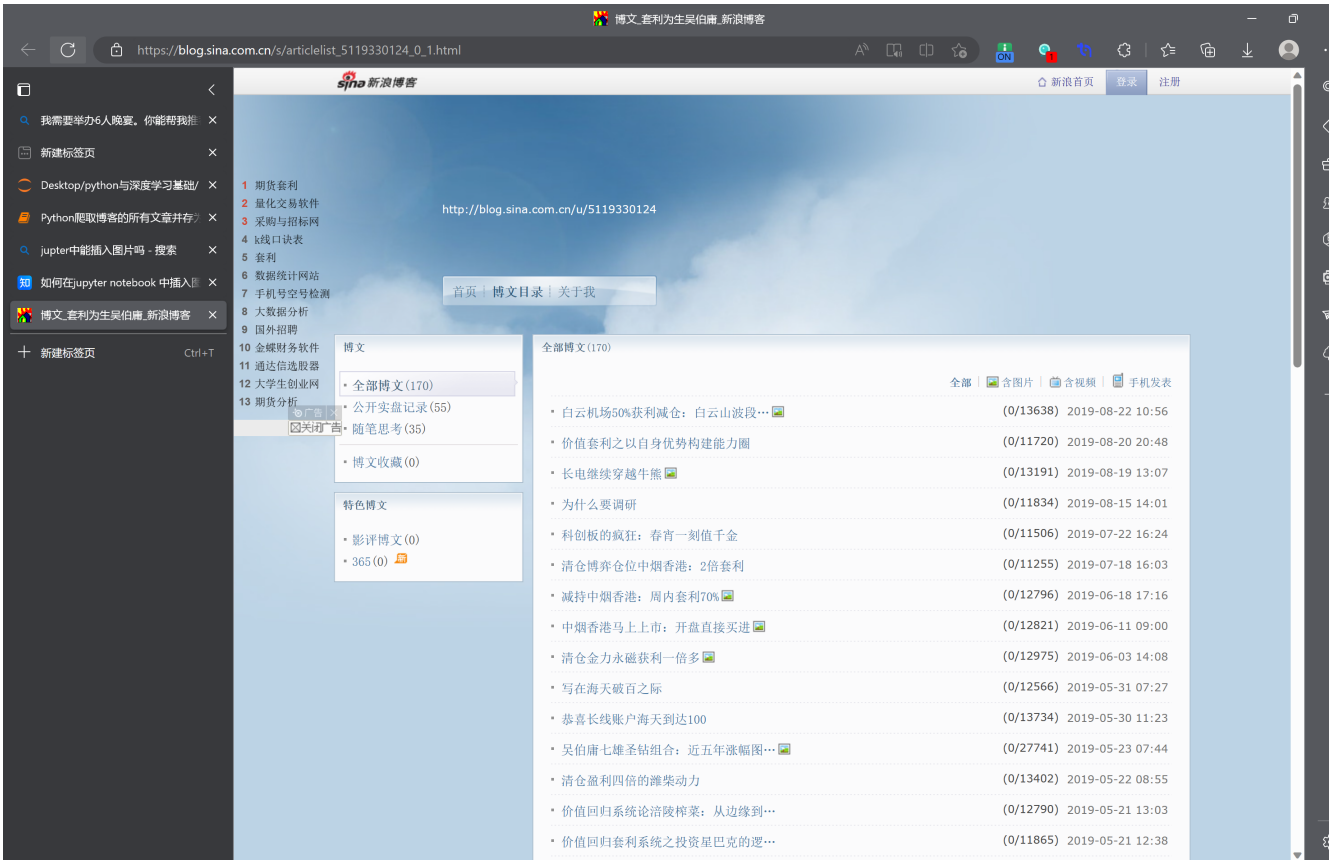
文件二：主函数，负责图形化展示、输入输出

文件一中已经较为完整地实现了希望达到的主要功能，本文件的内容主要是为了更好地在图形化的界面内完成输入和输出，而非仅仅在命令框中进行输入输出

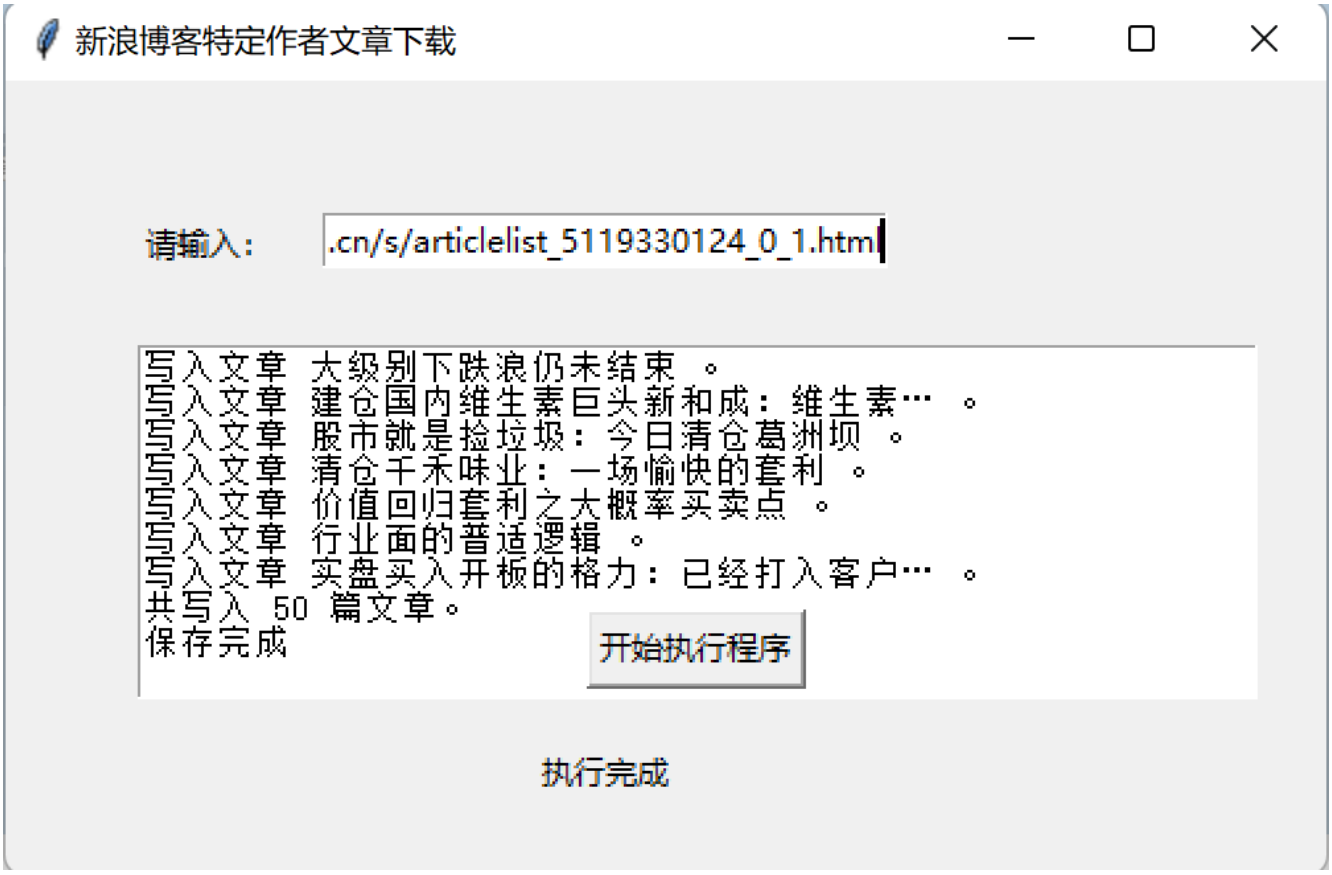
In [ ]:

```python
import subprocess
from tkinter import *

def execute_program():
    # 从文本框获取用户输入
    user_input = input_entry.get()

    # 将用户输入发送到另一个程序
    process = subprocess.Popen(['python', 'C:/Users/ziyuemu/Desktop/pyshenduxuexilab01/lab01pyth
    output, _ = process.communicate(user_input.encode())

    # 在文本框中显示输出
    output_text.delete('1.0', END)
    output_text.insert(END, output.decode())

    output_label.config(text="执行完成")

top = Tk()
top.title("新浪博客特定作者文章下载")
top.geometry('500x300')

# 创建标签和按钮
input_label = Label(top, text="请输入：")
input_label.place(x=50, y=50)
input_entry = Entry(top, width=30)
input_entry.place(x=120, y=50)
output_text = Text(top, width=60, height=10)
output_text.place(x=50, y=100)
output_label = Label(top, text="点击开始执行")
output_label.place(x=200, y=250)
execute_button = Button(top, text="开始执行程序", command=execute_program)
execute_button.place(x=220, y=200)

top.mainloop()
```

在“请输入”的框图中，输入希望爬取的博主的全部博文主页



将该网址输入后，点击“开始执行程序”，则此时开始执行程序，相关过程中的输出仍然输出在vscode中， 但结果的输出将展示在图形化的框图中

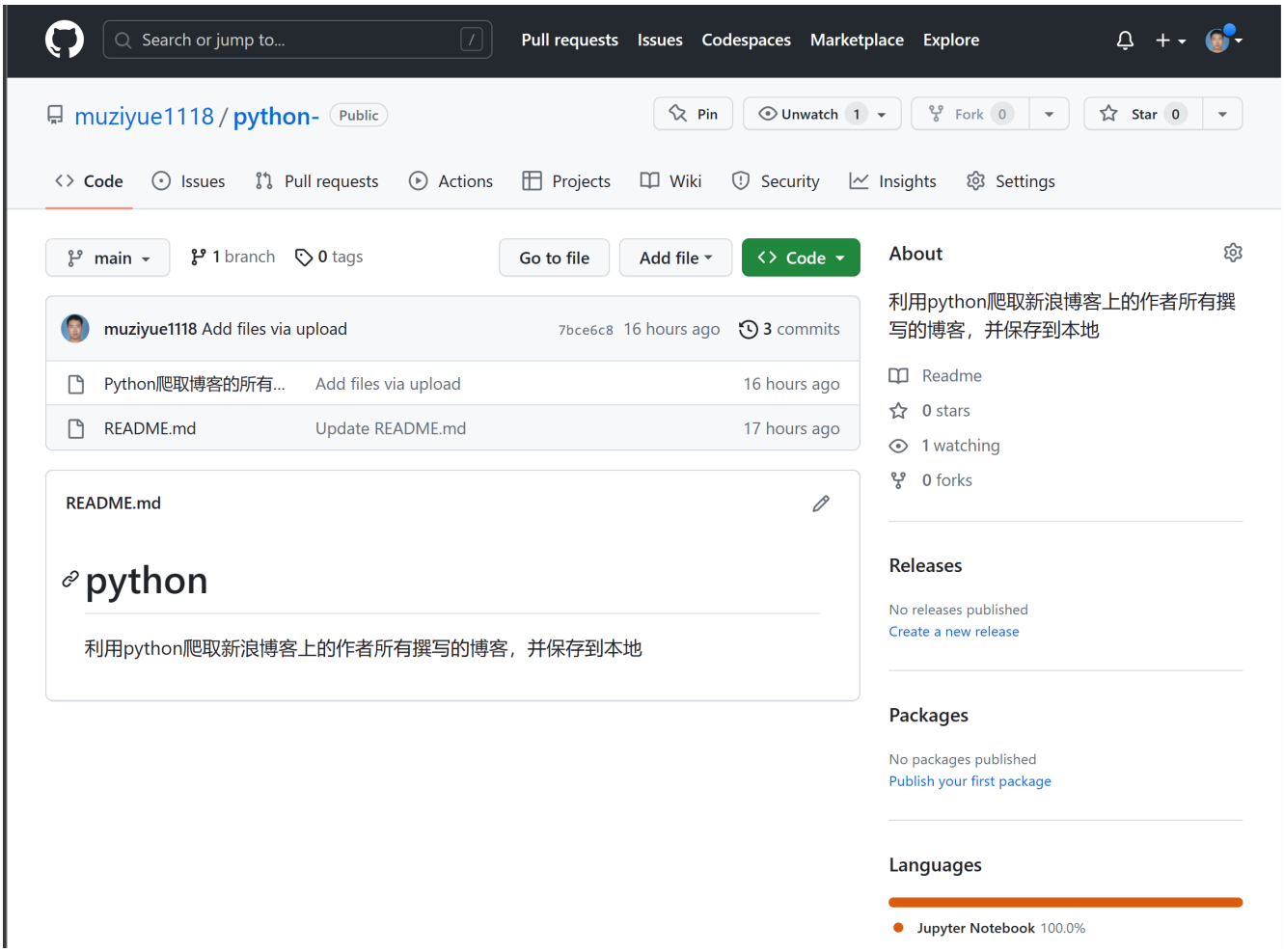


此时我们打开文件夹中，即可发现文章已经存储在本地中

点击进入发现文章已经按照我们的要求完成

实验源代码

本次实验的所有代码、中间文件均保存于本人的github仓库中，网址为：
https://github.com/muziyue1118/python- (https://github.com/muziyue1118/python-)



实验总结

本次实验是我第一次使用python完成一个较为独立、自主的项目，虽然之前张越一老师已经讲授了足够充分、详实的python基础语法、常用库等基本内容，但当我刚刚拿到本次的大作业时，仍然头脑发空，不知道如何下手。

开始时，本打算从推荐的选题中选择爬取各个国家GDP和化学元素来撰写代码，但在实践的过程中发现自己没有能力处理化学元素中“如果希望获取元素的进一步信息，需要点击大图，但点击出现大图的xpath路径是一样的”这个问题；然后尝试了爬取各个国家的GDP，但这个数据的存储方式也不是之前在课堂上讲授的典型类型，而是采用表格筛选的形式，故也放弃了。

后采用了本次的一个选题，一方面是因为本项目内容在网络上类似的内容之前做的人比较多，故有现成的代码可供借鉴，另一方面是本人曾经是一个比较喜爱使用博客的用户，但随着时代变迁，现在已经不怎么使用它了，故本次借此机会也是对自己童年的一个缅怀，当然，最重要的原因还是博客的爬取比较简单，数据基本没有受过包装，可以直接爬取。

然而，在实验具体进行中还是出现了许多问题，其中最致命的问题是我无法将使用ipynb上写成的代码转换为py文件在VScode上运行，起初报错的原因是无法找到docx包中oxml库，但import docx这一步却显示为正确。根据网上的提示，我反复删除docx库，下载、更新python-docx库了四遍，仍然无法解决问题。后来才发现是自己VScode中的python路径配置没有选用anaconda的python路径，而是选用了之前在电脑中下载的python，导致VScode在执行的过程中无法找到我的库中的包。

除此之外，还有从网页上初步爬取的网页内容的调整也比较繁琐，需要把一些乱七八糟多余的东西去除，不过相对来讲网络上的教程比较多，可以完成。

不过还有一个没解决的问题，在窗口中点击开始执行后，窗口就会卡住（虽然程序已经开始执行），且会被Windows判定为未响应，我目前不知道该如何解决该问题。

万事开头难，非常感谢老师布置这样的一个作业，之前学习python主要都是纸上谈兵，本次实际操演之后才

In [ ]: