第一次实验报告

1 实验概览

本实验分为两部分。第一部分首先了解了浏览器访问网址的具体过程，随后学习了HTML 语言，它是一种用设计标签的方式设计网页的超文本标记语言，由标签、元素组成，标签还可拥有一些属性。一般包括<head>和<body>这两个标签，分别表示网页的头部和正文。还了解了 HTML 的常见标签。

实验的第二部分，是使用 BeautifulSoup 处理网页 HTML。通过 BS4 库中的各种方法获得特定标签的内容，在这次试验中常用的是 findAll()方法，简单介绍几个常见参数：

1. name 参数可以查找所有名字为‘name’的 tag，字符串对象会被自动忽略。
2. attrs 参数搜索每个 tag 的属性。
3. text 参数可以搜搜文档中的字符串内容，与 name 参数的可选值一样, text 参数接受字符串,正则表达式，列表, True。还可以与其它参数混合使用来过滤 tag，会找到.string 方法与 text 参数值相符的 tag。

最后，对正则表达式进行了简要介绍，它可以帮助我们更精确地匹配出网页中想要的内容。

2 实验环境

本实验基于 Docker 中的 sjtucmic/ee208 镜像，以及 Python 3.10。

使用到的库有 BeautifulSoup、Regular Expression。

3 练习 1

3.1 解决思路

练习 1 需要我们对给定的网页内容，返回网页中所有超链接的 URL（不包括图片地址），并将结果打印至文件 res1.txt 中。在助教提供的 example1.py 参考文件中，已经写好了获取网站 HTML 和输出的函数，我们只需要补充 parseURL 函数，它以 content（储存着网站的 HTML 代码）作为参数，并返回一个集合 urlset（储存着网站所有 URL）。

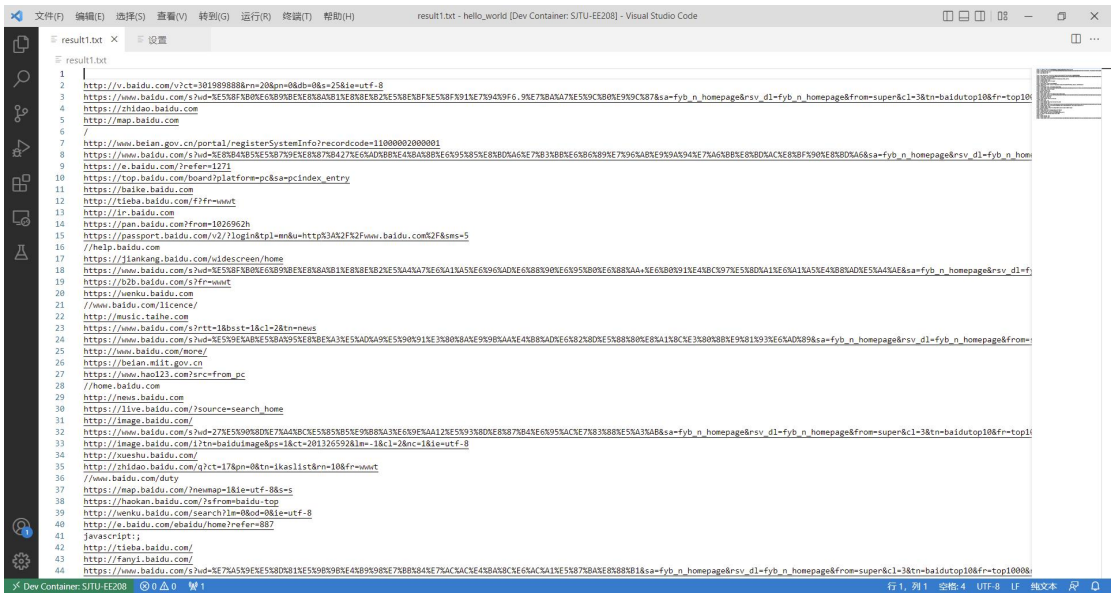
因为链接地址只需要考虑形如<a href=”...”>的形式，所以只需要用 content 生成一个 BeautifulSoup 的对象，再用 findAll 方法找到其中所有标签为’a’的元素，再用 get 方法获得每一个标签的 href 属性的值，将每一个值添加到 urlset 集合中即可。

3.2 代码及运行结果

```
    urlset = set()

    soup = BeautifulSoup(content, "html.parser")
    for i in soup.findAll('a'):
        urlset.add(i.get('href',''))

    return urlset
```



3.3 分析与思考

1. 第 13 行代码 soup = BeautifulSoup(content, "html.parser")中，如果这个构造函数不加第二个参数，那么 python3 会报错。给出的原因是，如果不加第二个参数，那么这段代码在另一个系统或虚拟环境中运行时，可能会调用其他的 parser，使得到的结果不同，因此 python 建议我们加上这个参数，指定 BeautifulSoup 的解析器为"html.parser"。

2. 为什么得到的结果中，总有一行是空白的呢？在练习 2 和练习 3 中并未出现同样的情况。在查阅一些资料后，我推断原因应该是，网站 HTML 中有一个空的 href，它表示不需要跳转，相当于访问当前 URL，点击会刷新页面。

4 练习 2

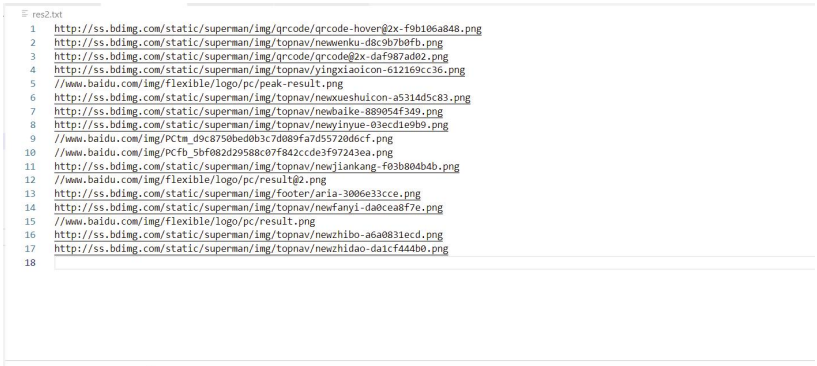
4.1 解决思路

这道题与上一道思路类似，我们只需要补充 parseURL 函数，它以 content（储存着网站的 HTML 代码）作为参数，并返回一个集合 imgset（储存着网站所有图片的地址）。

由于图片地址只需要考虑形如<img src= “..” >这样的形式，我们只需先用 findAll()方法搜索’img’标签，再循环每一项提取其中’src’属性的值，将它们加入 imgset 即可。

4.2 代码及运行结果

```
10 def parseURL(content):
11     imgset = set()
12 
13     soup = BeautifulSoup(content, "html.parser")
14     for i in soup.findAll('img'):
15         imgset.add(i.get('src',''))
16 
17     return imgset
```



5 练习 3

5.1 解决思路

这道题需要将知乎日报网页中的图片地址，相应文本，对应的超链接网址以规定格式打印至 res3.txt 中。我们只需要补充 parseZhihuDaily 函数，该函数有两个参数，一个是 content（储存着网站的 HTML 代码），另一个是 url（用于超链接网址转化为绝对地址），并返回一个列表 zhihulist（包含所有文章信息）。

我们首先观察一下该网页的 HTML 代码。在浏览器中打开开发者工具后，可以看到所有图片都有一个标签为<a>，属性'class' = 'link-button'的父节点，该标签中还包含了图片对应的'href'属性。同时，还有<img>和两个子节点，分别包含了图片地址和相应文本，如下图：

因此我们每次先创建一个空列表 list1，然后搜索到所有标签为<a>，属性'class' = 'link-button'的元素，采用正则表达式来匹配搜索。之后获取这个元素<img>标签的'src'属性内容，标签的元素内容，最后获取元素的'href'属性的值，也就是图片对应的超链接网址。这里需要注意，直接得到的超链接是相对地址，需要用 urllib.parse.urljoin(url, linkpage) 将当前页面的 URL 添加上，将相对地址改成绝对地址。最后，将上述三个值依次添加到 list1 中，并将 list1 添加到 zhihulist 中就可以了。

具体代码如下图所示：

```
def parseZhihuDaily(content, url):
    zhihulist = list()

    soup = BeautifulSoup(content, "html.parser")
    for i in soup.findAll('a',{"class":"link-button"}):
        list1 = list()
        list1.append(i.img.get('src', ''))
        list1.append(i.span.text)
        link = i.get('href', '')
        linkpage = urllib.parse.urljoin(url, link)
        list1.append(linkpage)
        zhihulist.append(list1)

    return zhihulist
```

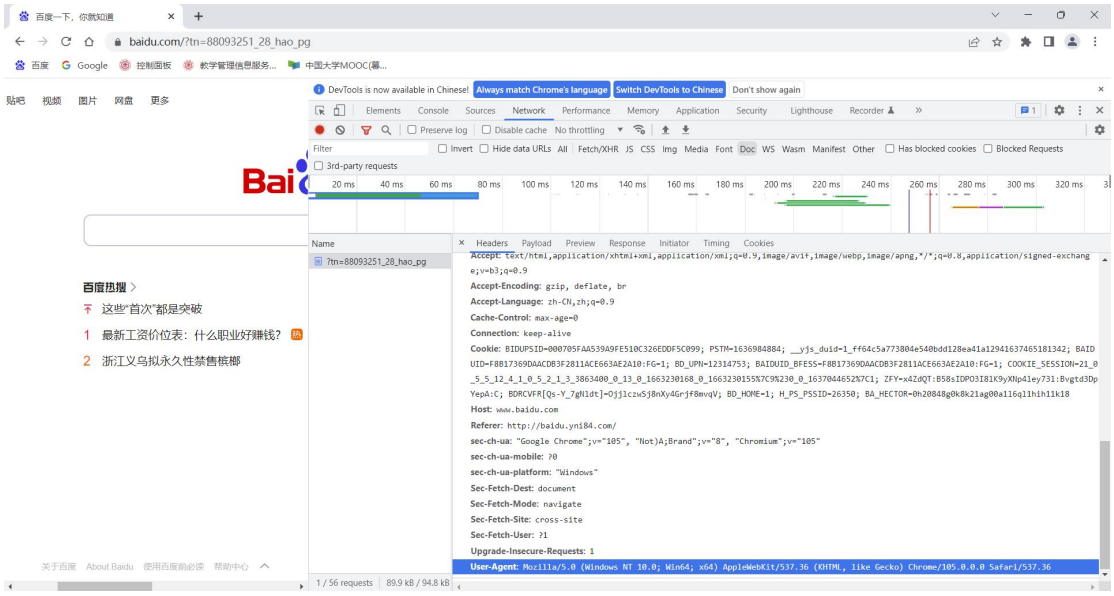
5.2 添加报头

在编写爬虫的过程中，有些网站会设置反爬机制，对于非浏览器的访问拒绝响应，来防止别人恶意爬取信息；或者短时间的频繁爬取会触发网站的反爬机制，导致 ip 被封无法访问网页。在阅读了助教提示的 CSDN 的链接后，我选择了利用 add_header()函数给对象添加报头的方式来解决这个问题。

原理是，如果我们设置一些 Headers 信息，让爬虫模拟成浏览器去访问这些网站，网站会把我们当成正常访问的用户。对来访者身份的判定一般基于 Headers 里的 User 值，User-Agent 会告诉网站服务器，访问者是通过什么工具来请求的，如果是爬虫请求，一般会拒绝，如果是用户浏览器，就会应答。

具体实现：首先使用 urllib.request.Request(url)创建一个 Request 对象并赋予 req，再利用 add_header()函数给对象添加报头。这个函数有两个参数，第一个是对象名，第二个是对象值。

那么如何获取一个网页的 User-Agent 值呢？我们可以任意打开一个网页，然后按 F12 打开开发者工具，切换到 Network 标签页，选择访问的页面地址，再往下拖动到底部，就会看到“User-Agent”标签的一串信息，将其复制下来即可。如下图所示：



5.3 代码实现

```
37  def main():
38      url = "http://daily.zhihu.com/"
39      req = urllib.request.Request(url)
40      req.add_header('U (module) request lla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KH
41      content = urllib.request.urlopen(req).read()
42      zhihus = parseZhihuDaily(content, url)
43      write_outputs(zhihus, "result3.txt")
44
```

5.4 运行结果

result3.txt

```
https://picx.zhimg.com/v2-ee8737b97d93e8edd32356494c03cc45.jpg?source=8673f162 有哪些冒通人很少听说，但在生活中很重要的化学元素？ http://daily.zhihu.com/story/9753144
https://pic1.zhimg.com/v2-3dd8c5fa5bca0a5a018e058f649c28e2.jpg?source=8673f162 青梅煮酒为什么要配青梅，并且要煮酒？ http://daily.zhihu.com/story/9753140
https://pic1.zhimg.com/v2-5924cbf975d91d60e2de20a64523b0a9.jpg?source=8673f162 如果宇宙中真的存在经真文明，我们有胜算吗？ http://daily.zhihu.com/story/9753136
https://pica.zhimg.com/v2-f9645f66ad9f440b3cd4cc9176ede7e0.jpg?source=8673f162 南朝还有荆扬之争吗，荆扬之争什么时候退出历史的？ http://daily.zhihu.com/story/9753132
https://pic1.zhimg.com/v2-2cb9fd3e51d4536fe5118095ad328dbb.jpg?source=8673f162 人脑能承载多少知识？ http://daily.zhihu.com/story/9753124
https://pica.zhimg.com/v2-3dd8c5fa5bca0a5a018e058f649c28e2.jpg?source=8673f162 瞎扯 · 如何正确地吐槽 http://daily.zhihu.com/story/9753145
https://pica.zhimg.com/v2-b781a2a3dbf10fa278ae37a622d9bb4b.jpg?source=8673f162 小事 · 这个陌生的城市是不是喜欢找啊？ http://daily.zhihu.com/story/9753068
https://pic2.zhimg.com/v2-3e7382fb6f54ddcf8513a3c8d4e9d5eb.jpg?source=8673f162 《水浒传》中林冲真的窝囊吗？ http://daily.zhihu.com/story/9753108
https://pic2.zhimg.com/v2-9708104aee896e2595878d61b64acefe5.jpg?source=8673f162 人们为什么对负面信息更感兴趣？ http://daily.zhihu.com/story/9753099
https://pic3.zhimg.com/v2-57b6e4cca2ac81b4f74dc6701e37e417.jpg?source=8673f162 电脑的发展以后会脱掉键件哪个部位？ http://daily.zhihu.com/story/9753089
https://pic1.zhimg.com/v2-cfb2b4fb04bdf8f15190815cdc41b9c.jpg?source=8673f162 教练是怎么指导比自己强的运动员的？ http://daily.zhihu.com/story/9753080
https://pic1.zhimg.com/v2-f9d0f1a4653225aeb0cd2a5c6b640429.jpg?source=8673f162 把汉语言文学学到极致会有怎样意想不到的收获？ http://daily.zhihu.com/story/9753073
https://pic2.zhimg.com/v2-fab9c8b85e6e77d52fb656527d61d451.jpg?source=8673f162 小事 · 「钱不够打电话，千万不要省吃的」 http://daily.zhihu.com/story/9753059
https://picx.zhimg.com/v2-ee412acafe753be5491179ec8d624910.jpg?source=8673f162 蚊子最喜欢什么血型（ABO血型）？ http://daily.zhihu.com/story/9753051
https://pic1.zhimg.com/v2-6cf90626d4667b5d5982bebec4170c8f.jpg?source=8673f162 临床医学五年制学生如何顺成为法医？ http://daily.zhihu.com/story/9753044
https://pic1.zhimg.com/v2-6f663448b113e619a263420357781025.jpg?source=8673f162 为什么乒乓球运动员都长得很好看？ http://daily.zhihu.com/story/9753043
https://pic1.zhimg.com/v2-5a67c9847466d480bb2b6ec92786b3f6.jpg?source=8673f162 人类向太空中发射过哪些意想不到的东西？ http://daily.zhihu.com/story/9753041
https://picx.zhimg.com/v2-172ec4c0e9c27917500f6bdda5f8bcea.jpg?source=8673f162 从生物学上来看，人类有哪些设定不适合永生？ http://daily.zhihu.com/story/9753038
https://pic2.zhimg.com/v2-9c3f5d01087cf5470e76acaee7d076dc.jpg?source=8673f162 为什么《数码宝贝》只有第一部那么火？ http://daily.zhihu.com/story/9753006
https://picx.zhimg.com/v2-e24853e7ec6df9a863339a8285b1afd9.jpg?source=8673f162 如果给你一百万，你会用来做什么？怎么用才有意义？ http://daily.zhihu.com/story/9753029
https://pica.zhimg.com/v2-cb6022f2caa63b567287820990ceecaf.jpg?source=8673f162 炫富真的很爽吗？ http://daily.zhihu.com/story/9753026
https://picx.zhimg.com/v2-51e1e4f301af4151d9116afdc3eb531f.jpg?source=8673f162 无线鼠标安全吗？ http://daily.zhihu.com/story/9753022
https://pic1.zhimg.com/v2-71a47b50d48262da0763a390da65da79.jpg?source=8673f162 为什么其他生物就不能进化成「人」呢？ http://daily.zhihu.com/story/9753016
https://pic2.zhimg.com/v2-47932c99a3060179d9c547e107ec17f7.jpg?source=8673f162 瞎扯 · 如何正确地吐槽 http://daily.zhihu.com/story/9753036
https://pic1.zhimg.com/v2-3d385ec1f253c0ef228937a3ccbab34f.jpg?source=8673f162 很多动物都能通过气味识别幼崽，人类可以吗？ http://daily.zhihu.com/story/9752951
https://pic1.zhimg.com/v2-a34281d348e94e5307d5ae03375a70d6.jpg?source=8673f162 有哪些隐含编匾或细节、情节的名画？ http://daily.zhihu.com/story/9753003
https://pic2.zhimg.com/v2-23abec2296e70acf6ca2e1211f3654a4.jpg?source=8673f162 骨头是生的碰还是熟的键？ http://daily.zhihu.com/story/9753000
https://pic1.zhimg.com/v2-1f3d3b17ec70cd6471ed23230be6ae3e.jpg?source=8673f162 为什么有人说开车时，最能看出一个人的福商和潜力？ http://daily.zhihu.com/story/9752984
https://pica.zhimg.com/v2-768140debb9582aa54bf9b852c57fa33.jpg?source=8673f162 欧洲中世纪的饮食，平民饮食状况是怎样的？ http://daily.zhihu.com/story/9752968
https://pic1.zhimg.com/v2-01abbf13dd9cc4d0b8817326580fd524.jpg?source=8673f162 瞎扯 · 如何正确地吐槽 http://daily.zhihu.com/story/9752976
```

6 拓展思考

思考：爬取到的 href 链接有哪几种形式？

根据练习 1 的运行结果，href 的链接形式有如下 5 种：

1. http://.../：代表未加密的超链接。HTTP 协议以明文方式发送内容，不提供任何方式的数据加密

2. https://.../：代表加密的超链接。HTTPS 经由 HTTP 进行通信，但利用 SSL/TLS 来加密数据包，保护交换数据的隐私与完整性。

3. //.../：代表未指定链接的加密形式，根据相对协议进行 url 转换。它会判断当前的页面协议是 http 还是 https，比如当前页使用的是 https 协议，那么转换后的 url 就是 https://...

4. /.../：代表根据服务器根目录进行 url 转换的链接。

5. javascript:;：代表的是执行一段 Javascript 代码。如果是 href='javascript:;'，那么这个代码是空的，所以什么也不执行。