lab1 实验报告

班级：　　姓名：　　学号：

一、实验概览

本实验旨在了解 HTML 的结构，学习用 BeautifulSoup 整理 Python 模拟抓取的 HTML，并获取需要的信息。练习 1 中，需获取网页中所有超链接的 URL；练习 2 中，需获取网页中所有图片的地址；练习 3 中，需获取知乎日报网页中，每一条消息对应的图片地址、文本、网址。

二、实验环境

Docker + VS Code Python3.8.5　　Dev Container: SJTU-EE208

三、解决思路

1、 练习 1

Example1 中已完成模拟浏览器抓取 HTML 网页，以及将 urlSet 中的内容写入文件的操作，所以我只需要写 parseURL(content)函数内部的代码。

先用 BeautifulSoup 转化 content。根据链接地址只需要考虑形如<a href="...">的形式，所以我先用 findAll 找到所有带有属性<a>的元素，再去获取它们的 href 属性。如果存在 href 属性，就将链接地址存入 urlset；不存在 href 属性的忽略即可。

```
def parseURL(content):
    urlset = set()
    ########################
    # write your code here
    soup = BeautifulSoup(content, features="html.parser")
    for i in soup.findAll("a"):
        url = i.get("href", "")
        if url:
            urlset.add(url)
    ########################
    return urlset
```

2、 练习 2

练习 2 与练习 1 相似。只需要根据图片地址考虑形如<img src="…">这样的形式，去找到所有 tag<img>，再去获取其中的 src 属性。（一般网页有 tag<img>的都会有 src，否则图片无法显示，但是保险起见，仍然忽略了没有 src 的 tag<img>）

```
def parseIMG(content):
    imgset = set()
    ########################
    # write your code here
    soup = BeautifulSoup(content, features="html.parser")
    for i in soup.findAll("img"):
        src = i.get("src", "")
        if src:
            imgset.add(src)
    ########################
    return imgset
```

3、 练习 3

蚊子最喜欢什么血型（ABO血型）？

```
<a href="/story/9753051" class="link-button"> == $0
    <img src="https://picx.zhimg.com/v2-ee412ac….jpg?source=8673
    f162" class="preview-image">
    <span class="title">蚊子最喜欢什么血型（ABO血型）？ </span>
</a>
```

观察 http://daily.zhihu.com/的 HTML 发现，所有需要获取信息的消息（如左上图），它们的 HTML 格式都一致（如右上图），且有且仅有这些消息有 class=”link-button”的属性。所以我先用 findAll 找到所有符合属性 class=”link-button”的元素，并获取该元素的 href，即 linkpage。再得到该元素的子节点 tag<img>和 tag，来获取 src 属性和 title 的内容。然后将[src, title, linkpage]加入 zhuhulist 即可。

参考课件所附链接，使用 add_header()添加报头。一些网页具有反爬虫机制，添加报头将代码包装成一个浏览器，从而突破这些机制。

```
def parseZhihuDaily(content, url):
    zhihulist = list()
    ##########################
    # write your code here
    soup = BeautifulSoup(content, features="html.parser")
    for i in soup.findAll('a', {"class": "link-button"}):
        linkpage = urllib.parse.urljoin(url, i.get("href"))
        content = i.contents
        src = content[0].get("src")
        title = content[1].string
        zhihu = [src, title, linkpage]
        zhihulist.append(zhihu)
    ##########################

    return zhihulist
```

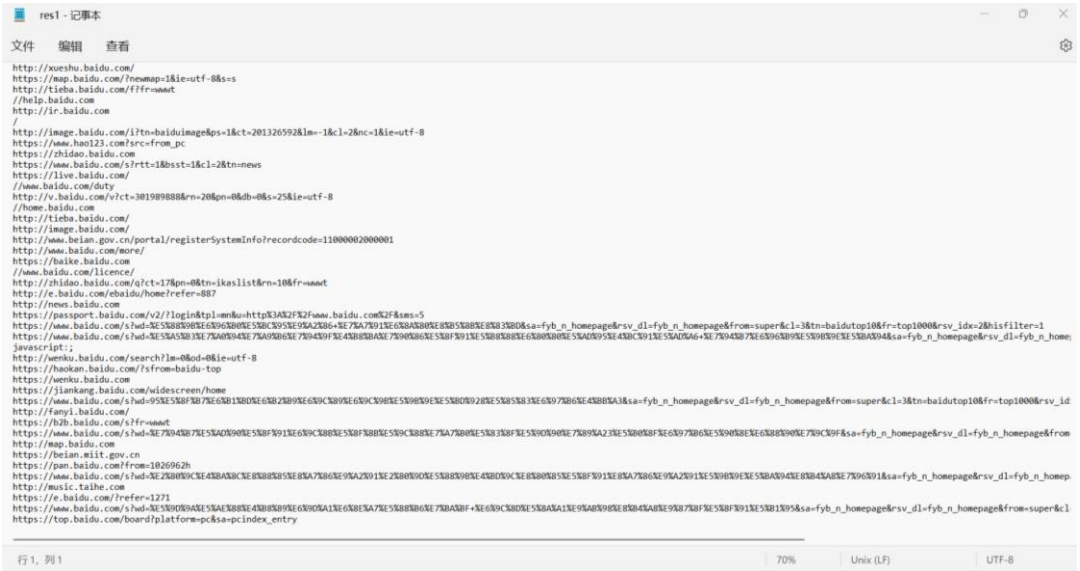
```
def main():
    url = "http://daily.zhihu.com/"
    req = urllib.request.Request(url)
    agent = "Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/99.0.4844.84 Safari/537.36"
    req.add_header("User-Agent", agent)
    content = urllib.request.urlopen(req).read()
    zhihus = parseZhihuDaily(content, url)
    write_outputs(zhihus, "res3.txt")
```

四、代码运行结果

因相关网页会实时更新，以下运行结果仅为该时刻的结果。

1、 练习 1

以下为 2022/9/20 18：55 运行 exercise1 的结果，因屏幕大小，并未截取完全，详细请见文件 res1.txt



2、 练习 2

以下为 2022/9/20 18:58 运行 exercise2 的结果，详细请见文件 res2.txt

```
https://dss0.bdstatic.com/5aV1bjqh_Q23odCf/static/superman/img/topnav/newjiankang-f03b804b4b.png
https://dss0.bdstatic.com/5aV1bjqh_Q23odCf/static/superman/img/topnav/newzhibo-a6a0831ecd.png
http://ss.bdimg.com/static/superman/img/qrcode/qrcode-hover@2x-f9b106a848.png
https://dss0.bdstatic.com/5aV1bjqh_Q23odCf/static/superman/img/topnav/newwenku-d8c9b7b0fb.png
https://dss0.bdstatic.com/5aV1bjqh_Q23odCf/static/superman/img/topnav/newbaike-889054f349.png
http://ss.bdimg.com/static/superman/img/footer/aria-3006e33cce.png
//www.baidu.com/img/flexible/logo/pc/peak-result.png
http://ss.bdimg.com/static/superman/img/qrcode/qrcode@2x-daf987ad02.png
//www.baidu.com/img/flexible/logo/pc/result.png
https://dss0.bdstatic.com/5aV1bjqh_Q23odCf/static/superman/img/topnav/newxueshuicon-a5314d5c83.png
https://dss0.bdstatic.com/5aV1bjqh_Q23odCf/static/superman/img/topnav/yingxiaoicon-612169cc36.png
https://dss0.bdstatic.com/5aV1bjqh_Q23odCf/static/superman/img/topnav/newfanyi-da0cea8f7e.png
https://dss0.bdstatic.com/5aV1bjqh_Q23odCf/static/superman/img/topnav/newyinyue-03ecd1e9b9.png
//www.baidu.com/img/PCtm_d9c8750bed0b3c7d089fa7d55720d6cf.png
https://dss0.bdstatic.com/5aV1bjqh_Q23odCf/static/superman/img/topnav/newzhidao-da1cf444b0.png
//www.baidu.com/img/PCfb_5bf082d29588c07f842ccde3f97243ea.png
//www.baidu.com/img/flexible/logo/pc/result@2.png
```

3、 练习 3

以下为 2022/9/22 16:45 运行 exercise3 的结果，详细请见文件 res3.txt

```
https://pic1.zhimg.com/v2-d611e914bf52462cafc59c3c3acc6e7f.jpg?source=8673f162  教练是怎么指导比自己强的运动员的?        http://daily.zhihu.com/story/9753237
https://pic4.zhimg.com/v2-1ba1fa50383a100f2301a15e4f95067a.jpg?source=8673f162  这篇文章带你玩转 AI 绘画 http://daily.zhihu.com/story/9753272
https://pic4.zhimg.com/v2-9c5e37d076233a796254595790 4ab249.jpg?source=8673f162  不吃晚饭为什么对减肥无用?        http://daily.zhihu.com/story/9753266
https://picx.zhimg.com/v2-3080c17ac5150d71d94126e055a82160.jpg?source=8673f162  最让你震撼的游戏细节有哪些?        http://daily.zhihu.com/story/9753254
https://picx.zhimg.com/v2-0ee7a9e91708f4b72709290a2c2c7c4d.jpg?source=8673f162  瞧扯 · 如何正确地吐槽     http://daily.zhihu.com/story/9753257
https://picx.zhimg.com/v2-a1f08b9ac171aa3ce8ac44484ffe2d2e.jpg?source=8673f162  为什么人无法忍受 40°C 的环境温度，在 40°C 的水里却很滋润? http://daily.zhihu.com/story/9753183
https://pic3.zhimg.com/v2-f77bb378afc7524bebed1dc223939378.jpg?source=8673f162  你见过哪些错误或者不太恰当的翻译?        http://daily.zhihu.com/story/9753209
https://pica.zhimg.com/v2-335fa4c21786b70a3448120ba73b1f7d.jpg?source=8673f162  围棋中隐藏的定律?        http://daily.zhihu.com/story/9753199
https://picx.zhimg.com/v2-db8908006e745cc3651fbca15594404c.jpg?source=8673f162  时间是什么，其它生物怎么感知时间? http://daily.zhihu.com/story/9753198
https://pic1.zhimg.com/v2-be14d2aa8b6ffec63ca7071ff0e57e40.jpg?source=8673f162  意识到底是如何产生的? 能通过技术手段进行读取吗?    http://daily.zhihu.com/story/9753191
https://picx.zhimg.com/v2-86d93cf047f0c3b1cd1de8717db43b40.jpg?source=8673f162  瞧扯 · 如何正确地吐槽     http://daily.zhihu.com/story/9753216
https://picx.zhimg.com/v2-72888cd3944ad45cf7fe9046a8f8f043.jpg?source=8673f162  人静止捉着物体不做功，那为什么会感到累?    http://daily.zhihu.com/story/9753154
https://pica.zhimg.com/v2-3615706646562696ded4fd300fe6a8be.jpg?source=8673f162  什么叫「好」的词曲，「好」的编曲，「好」的混音，「好」的音色...如何定义「好」?      http://daily.zhihu.com/story/9753175
https://pic1.zhimg.com/v2-34b8e5d518d610c24cdcd285ea2e48ca.jpg?source=8673f162  你相册中最舍不得删的一张「饭照」是哪张? 吃的什么食物，有什么故事?    http://daily.zhihu.com/story/9753174
https://picx.zhimg.com/v2-9517da6d8b2a4a267bff659d0b2e98f7.jpg?source=8673f162  外行对你熟知的领域有哪些误解?      http://daily.zhihu.com/story/9753168
https://pic1.zhimg.com/v2-9dfdeaa6c7c1eace9d350a0523ac42af.jpg?source=8673f162  有什么和「生命都会死亡」一样的基础规则，可以帮助我们对生命意义进行思考?      http://daily.zhihu.com/story/9753164
https://pica.zhimg.com/v2-eaf29f39043d1c1f80b506263f14375c.jpg?source=8673f162  瞧扯 · 如何正确地吐槽     http://daily.zhihu.com/story/9753177
https://pica.zhimg.com/v2-ee8737b97d93e8edd32356494c03cc45.jpg?source=8673f162  有哪些普通人很少听说，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/v2-5924cbf979d91d60e2de20a64523b0a9.jpg?source=8673f162  如果宇宙中真的存在修真文明，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  瞧扯 · 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我啊?      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/v2-970810aee896e2595878d61b64acefe5.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硬件哪个部位? http://daily.zhihu.com/story/9753089
https://picx.zhimg.com/v2-cfb2b4fb04bdf8f151990815cdc41b9c.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
```

五、分析与思考

1、 拓展思考

我爬取到的 href 链接有以下几种形式

(1) 以“http://”开始，以“/”结束

例如：http://xueshu.baidu.com/

(2) 以“http://”开始，末尾无“/”

例如：http://map.baidu.com

(3) 以“https://”开始，以“/”结束

例如：https://live.baidu.com/

(4) 以“https://”开始，末尾无“/”

例如：https://baike.baidu.com

(5) 以“//”开始，末尾无“/”

例如：//help.baidu.com

(6) 以“//”开始，以“/”结束

例如：//www.baidu.com/licence/

(7) javascript:;

(8) /

经过资料查找，我对以上不同形式的理解：

http 和 https 的区别：http 不对数据进行加密。而 https 为了安全，以 SSL 为基础，对数据进行加密，比 http 更适合用来传输敏感信息。

末尾有无“/”的区别：没有“/”的 URL 会先认为是一个文件，如果找不到该

路径对应的文件，才会认为该路径可能是是目录而去查找，也就是说最多会进行两次查找。而有“/”会直接认为这个URL是目录，只会进行一次查找。

以“//”开始的含义：不指定使用http协议或https协议，而是使用与当前使用相同的协议。

href=” javascript:;”的含义：这个URL表示要执行一个javascript脚本，而直接用“;”结束，表明不进行任何操作。

href=”/”的含义：回到根目录

2、其他发现

（1） 回车和
的区别

回车\n会让HTML在代码整理时换行，但在网页上不会换行，即对网页显示没有任何影响。
为真实显示在网页上的换行。

（2） 空格和

在HTML中无论写多少个空格，在网页中只会显示一个。而 的数量，会真实地显示在网页中。

（3） 图片地址source

`https://pica.zhimg.com/v2-3615706646562696ded4fd300fe6a8be.jpg`

`https://pica.zhimg.com/v2-3615706646562696ded4fd300fe6a8be.jpg?source=8673f162`

在完成练习3的过程中，我发现类似以上两个图片地址，指向的是同一张图片，且随意改变“source=”后的数字，仍指向此图。经向助教请教，了解到“source=”后的数字代表访问来源，即是从哪一个网页，去访问这张图片的。