Lab1　HTML-Parser

孙济宸　　学号：520030910016　　班级：F2003003

2021.9

1　实验概览

爬取 https://www.baidu.com 和 https://daily.zhihu.com 并且用 beautifulsoup 库对 html 文件进行解析，并对其中的各种 tag（如超链接、图片 url、问题标题等）进行查找、筛选并将结果存储在文件中。

2　实验环境

- Docker
- beautifulsoup (bs4)
- urllib
- re

3　练习题的解决思路

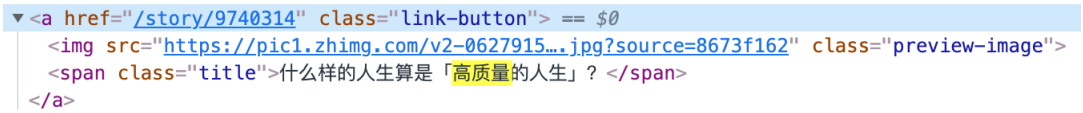
3.1　问题 1

问题要求返回网页中所有超链接的 URL（不包括图片地址），只要用 soup.findAll() 方法匹配 <a href = “...” >，再用 tag.get() 方法获取 href 中内容即可。

3.2　问题 2

思路同问题一，将匹配的标签变为 <img src = ”...”>,get() 获取 src 中内容即可。

3.3 问题 3



通过观察知乎日报的 html 文件，可以发现日报的每个条目是以 <a href = ”/story/[日报 id]”> 的格式开始。所以首先用 soup.findAll() 方法定位所有这个格式的标签；接着对于每个该种标签，

1. href 中内容即为日报文章 url（相对路径）
2. 子标签 0 为 <img src = ”...”>，类似问题 2 获取预览图 url
3. 子标签 1 为 [文章标题]？ 通过 tag.string 方法获取文章标题。

最后将获取到的网页 url、图片 url 及 title 存进 list。

4 代码运行结果

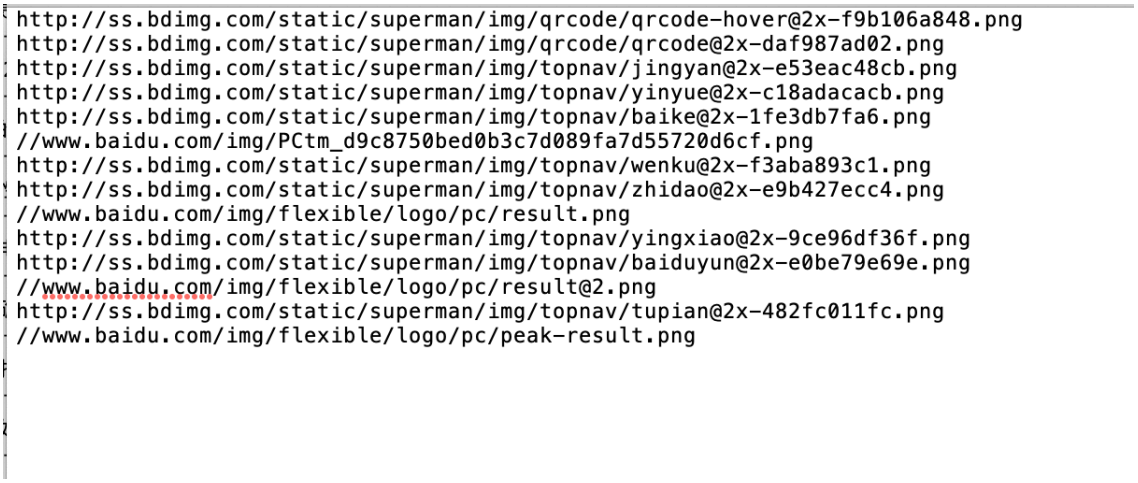
4.1 问题 1 结果

完整结果见 txt 文件

```
https://www.baidu.com/s?rtt=1&bsst=1&cl=2&tn=news
http://news.baidu.com
http://image.baidu.com
http://www.baidu.com/more/
http://wenku.baidu.com/search?lm=0&od=0&ie=utf-8
https://pan.baidu.com
http://tieba.baidu.com/f?fr=wwwt
https://haokan.baidu.com/?sfrom=baidu-top
//www.baidu.com/licence/
https://www.baidu.com/s?
cl=3&tn=baidutop10&fr=top1000&wd=3%E5%90%8D%E8%88%AA%E5%A4%A9%E5%91%98%E7%A(
_n_homepage&hisfilter=1
/
http://image.baidu.com/i?tn=baiduimage&ps=1&ct=201326592&lm=-1&cl=2&nc=1&ie=
http://music.taihe.com
https://beian.miit.gov.cn
//home.baidu.com
http://ir.baidu.com
https://www.baidu.com/s?
cl=3&tn=baidutop10&fr=top1000&wd=%E5%9C%B0%E9%9C%87%E8%87%B4%E6%B3%B8%E5%B7%
yb_n_homepage&hisfilter=1
http://tieba.baidu.com
https://www.baidu.com/s?
```

图 1: res1.txt

4.2 问题 2 结果



```
http://ss.bdimg.com/static/superman/img/qrcode/qrcode-hover@2x-f9b106a848.png
http://ss.bdimg.com/static/superman/img/qrcode/qrcode@2x-daf987ad02.png
http://ss.bdimg.com/static/superman/img/topnav/jingyan@2x-e53eac48cb.png
http://ss.bdimg.com/static/superman/img/topnav/yinyue@2x-c18adacacb.png
http://ss.bdimg.com/static/superman/img/topnav/baike@2x-1fe3db7fa6.png
//www.baidu.com/img/PCtm_d9c8750bed0b3c7d089fa7d55720d6cf.png
http://ss.bdimg.com/static/superman/img/topnav/wenku@2x-f3aba893c1.png
http://ss.bdimg.com/static/superman/img/topnav/zhidao@2x-e9b427ecc4.png
//www.baidu.com/img/flexible/logo/pc/result.png
http://ss.bdimg.com/static/superman/img/topnav/yingxiao@2x-9ce96df36f.png
http://ss.bdimg.com/static/superman/img/topnav/baiduyun@2x-e0be79e69e.png
//www.baidu.com/img/flexible/logo/pc/result@2.png
http://ss.bdimg.com/static/superman/img/topnav/tupian@2x-482fc011fc.png
//www.baidu.com/img/flexible/logo/pc/peak-result.png
```

图 2: res2.txt

4.3 问题 3 结果

```
https://pic1.zhimg.com/v2-06279159e0ea4fa7a3dc146d8703070a.jpg?source=8673f162  什么样的人生算是「高质量的人生」？
https://pica.zhimg.com/v2-a93839fc4fef75ad5b70adb5d4f57a0c.jpg?source=8673f162  一文看尽 2021 苹果秋季新品发布会
https://pic3.zhimg.com/v2-9880ae74b62ccd42609c7e11ce70b580.jpg?source=8673f162  有什么逆境翻盘的故事?        http:
https://pica.zhimg.com/v2-298e722668c226abbbe2bc15ab8c99f5.jpg?source=8673f162  为什么说打破惯性是一件需要勇气和力量的

https://pic2.zhimg.com/v2-b6793dd931c92c25df979c008eae32cb.jpg?source=8673f162  为什么现在国产电视剧里似乎穷人越来越少

https://pic2.zhimg.com/v2-19f2eafe4f64b5b274e83f430a34ac94.jpg?source=8673f162  瞎扯 · 如何正确地吐槽        http:
https://pic2.zhimg.com/v2-a97a8520d555e346a38020e25a9d4203.jpg?source=8673f162  DC 电影宇宙在扎克施耐德离开之后，整体
daily.zhihu.com//story/9740283
https://pic3.zhimg.com/v2-bb8b8344db9fa0fd9fc5df1768524b0e.jpg?source=8673f162  骑手谜云：法律如何打开外卖平台用工的

https://pic2.zhimg.com/v2-4779892ebdf5403108650f16a90f5993.jpg?source=8673f162  如何制作干花?      http://daily.
https://pic3.zhimg.com/v2-ee5a9289ee2f7519ecba6ff203523715.jpg?source=8673f162  有哪些优秀的美术馆设计案例?   http:
https://pic2.zhimg.com/v2-ec8154c7ae3b70302d7f21300971d161.jpg?source=8673f162  从科学的角度来看，护肤党从什么时候开始
9740281
https://pic3.zhimg.com/v2-15a04b59331448b09018ceba16757853.jpg?source=8673f162  瞎扯 · 如何正确地吐槽        http:
https://pic1.zhimg.com/v2-e7706bd539fd70bd2246c44343047387.jpg?source=8673f162  科学与艺术的关系是什么?      http:
https://pic2.zhimg.com/v2-0a4b53a4818e62678fa5918175b4d723.jpg?source=8673f162  创意工作有多难？总会发现自己的想法已经

https://pic2.zhimg.com/v2-47f55a1db1ed5d4bec5353785fbd41aa.jpg?source=8673f162  人在猝死前有哪些先兆吗?      http:
```

图 3: res3.txt

5 分析与思考

- beautifulsoup 解析 html 文件得到的是 bs4.BeautifulSoup 类型对象; 解析方法如果不指定 features = "html.parser"，会弹 warning。网上资料显示这是默认的解析方法，还

可以选择其他的解析方法。

- soup.findAll() 方法得到一个存有许多标签（tag）的 list。这些 tag 常常是多层嵌套的，可以用 parent,contents 等多种方法访问。
- **拓展思考：** 问题 1 中的 href url 有绝对路径、相对路径和 javascript。如果需要进一步处理可能需要分类
- tag.find() 方法会从这个标签的所有**子标签**中查找指定内容，并不能查找该标签同层（**包括自己**）的内容！所以问题 3 中用 i.contents[0].find(...) 会什么也找不到。这里被坑了好久
- 知乎日报文章的标题**并没有**使用 <title> 标签，而是用了 标签，所以不能直接用 tag.title，因此我选择使用 tag.string 获取标题内容。