lab1 实验报告

黄霄童 520030910229

练习1

- 问题： 给定任意网页内容，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，每一行为一个链接地址。
- 函数介绍

> 完整代码请参照 parseHREF.py 文件

1. parseURL(content)函数
    - 功能：解析content内容, 返回集合包含所给内容里的链接地址
    - 关键代码介绍
        1. 将内容交给BeautifulSoup模块解析，得到soup变量

```
soup = BeautifulSoup(content, 'html.parser')
```

        2. 截取形如<a href=.....>的内容，存在两种方式

        方法一： 使用find_all函数，返回元素名为a，含有href属性的tag

```
a_label=soup.find_all('a',attrs={'href':True})
```

        方法二： 使用筛选器select()函数

```
a_label=soup.select('a[href]')
```

        3. 遍历每一个tag, 取出href的链接

```
for a in a_label:
    urlSet.add(a.get('href'))
```

2. write_outputs(urls, filename)函数
    - 功能：将urls的内容写入文件名为filename的文件中，一行一个超链接。
3. main()

- 功能：主函数。

伪代码

1. 使用urllib读取链接获取html内容。
2. 调用parseURL()函数解析内容并返回content中的链接。
3. 调用write_outputs()函数写入'res1.txt'文件中。

练习2

- 问题： 给定任意网页内容，返回网页中所有图片地址，并将结果打印至文件res2.txt中，每一行为一个图片地址。
- 函数介绍

> 完整代码请参照 parseIMG.py 文件

同练习1，只需要将a和href位置替换为img和src即可。

练习3

- 问题：给定知乎日报的url，返回网页中的图片和相应文本，以及每个图片对应的超链接网址。并将图片地址，相应文本，超链接网址以下述格式打印至res3.txt中，每一行对应一个图片地址，相应文本和超链接网址，格式为：图片地址 \t 相应文本 \t 超链接网址。
- 函数介绍

> 完整代码请参照 parse.py 文件

分析知乎日报的html源码可以了解到，每一个话题的地址，图片和相应文本都放在 ... 中，其中话题地址在<a href=...>中，相应文本在中,图片在<img src=...>中。

示例：

```
<div class="box">
    <a href="/story/9740505" class="link-button">
     <img src="https://pic2.zhimg.com/v2-
7192ff0dadc26d671b2273041cfbd112.jpg?source=8673f162" class="preview-image">
     <span class="title">什么是「有效」的关心？ </span>
    </a>
 </div>
```

关键操作：

```
boxes=soup.findAll('div',{'class':'box'})
for box in boxes:
      linkpage=urllib.parse.urljoin(url,box.a['href'])
      img=box.img['src']
      text=box.get_text()
      zhihulist.append([img,text,linkpage])
```

```
伪代码：
boxes=筛选div属性下class为box的代码
对于boxes每一段代码box：
  linkpage=url+box中href的相对地址
  img=box中src后的地址
  text=box中的文字(get_text()函数)
  将[linkpage,img,text]存入zhihulist
```

其他函数类似于练习1，2

思考

问题： 你爬取到的href链接有哪几种形式?

三种：

1. 调用Js代码实现功能

```
<a href="javascript:;">
```

2. 实现超链接

   分为相对地址和绝对地址，绝对地址指向另一个站点，相对地址指向站点内某一个文件。

```
<a href="http://www.baidu.com">
```

3. "#"

   产生锚点的作用，跳转到指定的页面

```
<a href="#">
```

4. 打开文件

```
<a href="css.css">
```