Experiment 1

517030910356 Zhihui Xie

September 18, 2018

Contents

1	Exp	eriment Preparation	2
	1.1	Environment	2
	1.2	Purpose	2
	1.3	Principles	2
		1.3.1 What happens after we input the website?	2
		1.3.2 The basic structure of HTML	2
		1.3.3 BeautifulSoup	2
2	Exp	eriment Process	2
	2.1	Exercise 1	2
		2.1.1 Main procedure	2
		2.1.2 Result	3
	2.2	Exercise 2	4
		2.2.1 Main procedure	4
		2.2.2 Result	4
	2.3	Exercise 3	4
		2.3.1 Main procedure	4
		2.3.2 Result	5
3	Exp	eriment Thoughts	6
	3.1	The importance of data processing	6
	3.2	information Retrieval	6

1 Experiment Preparation

1.1 Environment

The experiment is based on Python 2.7 with Linux (running on Ubuntu with the help of VMware workstation).

1.2 Purpose

Given that this experiment are designed for further study of Web Crawler, we are supposed to master the basic structure of HTML, and get started with HTML parser.

From a higher perspective, the experiment aims to push us to try new things, and strengthens the ability of self-study.

1.3 Principles

1.3.1 What happens after we input the website?

As we know, the entire Intenet is organized with tons of small nets, and based on some internet protocols, we can share massive information on the Internet.

When we click the "Access" button, our browser will send the HTML request to the website, which includes the information of Cookie and User-agent. And then the website will respond, returning the HTML text. Finally, the browser will render the page.

1.3.2 The basic structure of HTML

As the name suggests, HTML(HyperText Markup Language) uses different markup tags to organize the page. These tags come in pairs and play different roles. And the content between a pair is called an HTML element, which is usually attached with some attributions.

1.3.3 BeautifulSoup

To deconstruct the HTML page, we use the module BeautifulSoup as the resolver.

The module transforms a complex HTML document into a complex tree of Python objects. There are four kinds of objects we are going to deal with: Tag, NavigableString, BeautifulSoup, and Commment.

2 Experiment Process

2.1 Exercise 1

2.1.1 Main procedure

In this task, I only need to complete the function 'parseURL' and the main, which accepts an URL, and returns a set of all the URLs of hyperlinks.

parseURL I first define the set 'urlset', and convert the URL into an BeautifulSoup object.

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

And then, to filter out some noise, such as 'javascript:;', I use the function 'compile' in the module 're' to define a pattern, which represents the URLs like '//www.baidu.com/duty/' or 'https://www.hao123.com'.

```
pattern = re.compile(('^\w*:*//'))
```

Finally, I search all the tags 'a' in the 'soup' using the function 'findAll', and extract the attribution 'href'.

```
for link in soup.findAll('a', {'href': pattern}):
    url = link.get('href', '')
    url = urljoin(sys.argv[1], url)
    urlset.add(url)
```

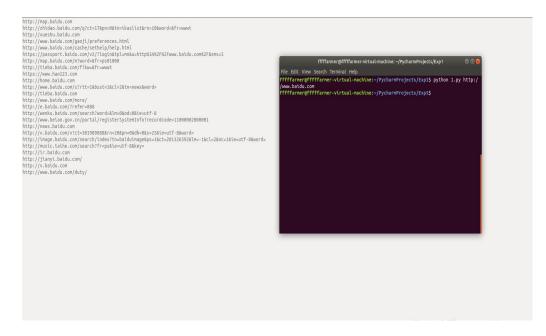
It's worth mentioning that, since the Transfer Protocol is missing, I use the function 'urljoin' to complete it. The parameter is the input URL.

main Here I also use the module 'sys' to get the input URL. And 'urllib' helps me read the URL directly and transforms it into a string.

```
def main():
    url = sys.argv[1]
    content = urllib2.urlopen(url).read()
    print type(content)
    urls = parseURL(content)
    write_outputs(urls, 'res1.txt')
```

2.1.2 Result

We can now directly input the URL in the console and then the result URL will be stored in res1.txt.



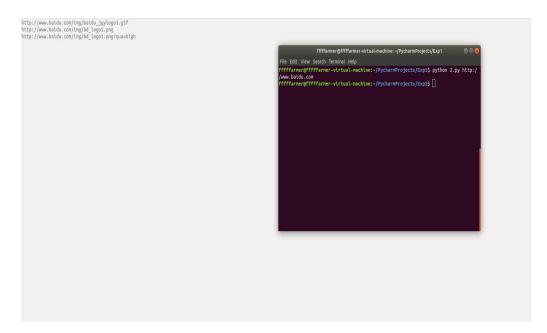
2.2 Exercise 2

2.2.1 Main procedure

Since the procedure in this task is very similar to the last one, I'm not going to discuss it in detail.

2.2.2 Result

We can now directly input the URL in the console and then the all the images will be stored in res2.txt.



2.3 Exercise 3

2.3.1 Main procedure

This time, the task is more tricky. We need to go deeper into the HTML page.

parseQiushibaikePic As usual, I first do some initialization.

```
docs = {}
nextPage = ''
soup = BeautifulSoup(content, features='html.parser')
```

Then, I search for tags whose attribution 'id' begins with 'qiushi_tag'. And I extract the id, as well as the content and the url by using BeautifulSoup method. A BeautifulSoup tag can have different attributions, and we only need to operate just like dealing with a dictionary.

```
for i in soup.findAll('div', {'id': re.compile('^qiushi_tag')}):
    id = i['id'].split('_')[-1]
    id = id.encode('utf-8')

imgurl = urljoin(sys.argv[1], i.img['src'])

content = i.find('div', {'class': 'content'}).span.string
```

```
content = unicode(content)
content = content.encode('utf-8')

tag = {}

tag['content'] = content

tag['imgurl'] = imgurl

docs[id] = tag
```

The search for the nextPage url is similar. To find the url, I first search for the hyperlink tag 'a' whose href starts with '/pic/page/'. And then, I take the url whose class is 'next'.

```
for i in soup.findAll('a', {'href': re.compile('^/pic/page/')}):
    spanClass = i.span['class'][0].encode('utf-8')
    if spanClass == 'next':
        nextPage = urljoin(sys.argv[1], i['href'])
```

main To get access to the website 'Qiushibaike', I set up the 'User-Agent' with the function in the module 'urllib'.

write_outputs I also rewrite the function 'write_outputs' to meet the need of writing out.

```
def write_outputs(info, filename):
    with open(filename, 'w') as f:

    for i in info[0].values():
        f.write(i['content'])
        f.write('\t')
        f.write(i['imgurl'])
        f.write('\n')
```

2.3.2 Result

We can now directly input the URL in the console and then the all the images and the corresponding contents will be stored in res3.txt.

```
不要拦我,我要看看优衣库有没有什么新新闻!!! https://plc.qlushlbalke.com/system/avtnew/2034/20340985/medlum/20170201225809.JPEG
兄弟, 搓衣板上凉吗? https://pic.giushibaike.com/system/avtnew/3100/31001181/medium/2018070710475967.JPEG
今晚公司加班到很贵,公司车送我回去,一路上碰到三个公司得女同事,,我觉得女孩子要互相帮助,在我强烈得要求一起走的时候被他们拒绝了,,,送我回去得周事不遇,开车慢慢带我走过她们男朋友得身边,,,看着她们被披走,,别问我
 为什么不喜欢不诚实得女孩子[[[捂脸] [[捂脸] https://pic.qiushibaike.com/system/avtnew/1244/12442450/medium/20188913175159.jpg
                   https://pic.qiushibaike.com/system/avtnew/3374/33746162/medium/20171030113755.JPEG
                              https://pic.qlushibalke.com/system/avtnew/1429/14298862/medlum/20180728115149.jpg
单身小伙子注意了昂 https://pic.qiushlbalke.com/system/avtnew/3899/38996922/nedtum/20180918152506.jpg
身上母常会是名其妙的青一块,黑一块,自己也没察觉挫到或是碰到,挤压一下,也不疼,百度一下,说是焦虑,严重的可能是血小板减少或是白血病,突然,莫名感到心慌,我是不是快要没了圆[发呆]圆[发呆]
ストラース (Lumbertander) が、一次、日本の日本に日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きな日本の大きなりに対象に対象を与えることに対しまります。
                   https://pic.qiushibaike.com/system/avtnew/3374/33746162/medium/20171030113755.JPEG
                                                                                                                                    File Edit View Search Terminal Help
fffffanner-Artvaul-nachtne:-/PycharnProjects/Exp15 python 3.py https
//www.qtushtbatke.com/ptc/
fffffanner@fffffarner-Virtual-nachtne:-/PycharnProjects/Exp15

值多少钱,等的动手,等价格!!!!! 急 https://pic.qiushibaike.com/system/avtnew/3892/38923654/media
我只是想吃个火锅而已,不要这样吧 https://pic.qiushibaike.com/system/avtnew/551/5512979/medium/201711261
上帝..... https://pic.qlushibalke.com/system/avtnew/3374/33746162/medium/20171030113755.JPEG
4公3母5只??还有没出生的吗? https://pic.qlushibalke.com/system/avtnew/1850/18504853/medium/201401
我哥们借我500说要买东西,他给我发来这个__什么意思 https://pic.qiushibaike.com/system/avtnew/3764/376478
                    https://pic.qiushibaike.com/system/avtnew/3374/33746162/medium/20171030113755.JPEG
没毛病啊,老铁 https://pic.qiushibaike.com/system/avtnew/2652/26520552/medium/20170616170553.JPEG
                   https://pic.qiushibaike.com/system/avtnew/3100/31001181/medium/2018070710475967.JPEG
今天傻逼了,上班路上看见一堆哈罗单车的助力车,就是电动骑得很快的那种,然后上去就是一顿扫码,然后还获得了两块
                                                                                                                                                                                                                                                  现锁不了, 只能去原地
方还车,我内心是崩溃的,妈蛋,这么麻烦老娘骑你干嘛,为了不让系统一直扣我钱,我只得傻逼一样骑到原地方去,路上
                                                                                                                                                                                                                                                  回去了,我他么的又要
走回公司,妹的,迟到了。 https://pic.glushibaike.com/system/avtnew/3930/39303670/medium/20180729204048
                              https://pic.qiushibaike.com/system/avtnew/3855/38557307/medium/2018051807552
 票都买好了,老妈说没有女朋友就别回来给我丢人现眼我该怎么办<u>關</u>[委屈]<u>關</u>[委屈]  [委屈]  https://plc.qlushtbal
```

3 Experiment Thoughts

3.1 The importance of data processing

It is far from completing the task when we finish crawling. There are massive data waiting for processing.

I find it kind of tricky espercially when I deal with the encoding 'utf-8'. I need to do some transformation so that the content can be stored in a right form.

So we really need to pay attention to those raw data.

3.2 information Retrieval

I also find that when we crawl some websites, we will use different methods to search for the information we want. And it's really important to improve our arithmetic espercially when the scale is large.