电类工程导论(C 类)实验 1 报告

贾萧松 516030910548

一、 实验概述

本次实验中，主要内容为通过学习 html 基本知识，python 库 beautifulsoup 的使用以及正则表达式的基本知识来实现对网页信息的获取。

二、 实验环境

Ubuntu14.04+python2.7+beautifulsoup4

三、 实验内容

1. 实验 1

给定任意网页内容，返回网页中所有链接地址（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。

首先调用 urlopen 函数读取指定 url 并用 read 函数获取代码

```
content = urllib2.urlopen(url).read()
```

然后创建 beautifulsoup 对象，创建储存结果的 set

```
    urlset = set()
    soup = BeautifulSoup(content)
```

接着使用 findAll 函数得到所有标签名为'a'的标签并获得其'href'的内容

```
for i in soup.findAll('a'):
        tmp = str(i.get('href',' '))
```

此时从得到的结果看获取的内容有部分不是链接，有部分链接不规范（例如："//www.baidu.com"），在此我给所有链接加入了"http://"开头

```
if(tmp != ' '): #make all url begin with http://
        if(tmp[:2] == "//" ):
                tmp = "http:"+tmp
        elif(tmp[:1] == '/' and len(tmp)>=3):
                tmp = "http:/"+tmp
        elif(tmp[0] == 'w' and len(tmp)>=3):
                tmp = "http://" + tmp
        if(tmp[0] == 'h'):
                res += (tmp + "\n");
                urlset.add(tmp)
```

最后把结果加入 set 并写入文件

```
                                urlset.add(tmp)
                f.write(res)
        f.close();
        return urlset
```

2. 实验 2

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。

代码主体与实验 1 相同，只是获取的是标签名为"img"的"src"属性

```
for i in soup.findAll('img'):
        tmp = str(i.get('src',' '))
```

3.实验 3

给定糗事百科有图有真相任意一页内容，返回网页中图片和相应文本，以及下一页的网址，并将图片地址与相应文本以下述格式打印至文件 res3.txt 中，每一行对应一个图片地址与相应文本，格式为 ：图片地址\t 相应文本

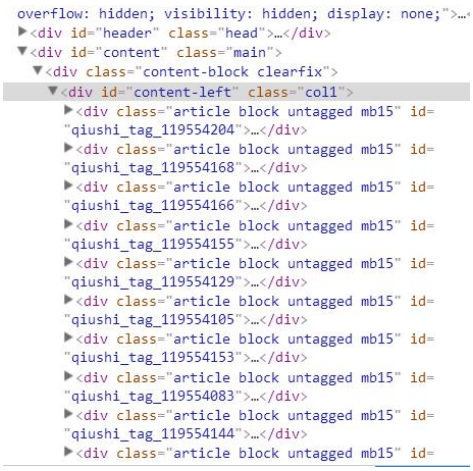
首先，获取糗事百科的访问权限

```
req = urllib2.Request(url, None, {'User-agent':'Custom User Agent'})
```

然后读取，创建 beautifulsoup 对象

```
content = urllib2.urlopen(req).read()
soup = BeautifulSoup(content)
```

通过观察网页 html 代码，发现我们要找的内容都在 id 为'qiushi_tag_数字'的 div 标签下



具体的，图片的地址在每个 qiushi_tag 下 class 为'thumb'的 div 标签下，只需获取其 a 标签的"href"内容

而文字内容在这里

根据以上，写出下面这段代码

```
for qiu_pic in soup.findAll('div', {'id':re.compile('^qiushi_tag_\d+')}):
    qiushi_tag = str(qiu_pic.get('id',' ')[11:])
    sentence = str(qiu_pic.find('div',{'class':'content'}).span.text)
    pic =  'http:'+str(qiu_pic.find('div',{"class":"thumb"}).find('a').find('img').get('src',' '))
    docs[qiushi_tag] = {"content":sentence, "imgurl":pic}
```

用正则匹配找到 qiushi_tag,并获取数字部分

注意到文字部分可能有多行，直接使用 get 函数不是很方便，于是经过阅读文档，我找到了 text 成员。这个成员忽略了所有标签，只保留文本，也就是我们要找的内容。

接下来，获取下一页的 url，观察到在这里

```
▼<li>
  ▼<a href="/pic/page/2?s=5018770" rel="nofollow">
     <!--<a href="/pic/page/2/" rel="nofollow">-->
     <span class="next">
     下一页
     </span>
   </a>
  </li>
 </ul>
</div>
```

如果直接一层一层找到这里是繁琐的，于是我采用找到 span 标签,class 属性为”next”的标签，再找它的 parent 的方法，很简单的实现了要求。

由于获取的是相对地址，所以使用到了 urljoin 函数将相对地址转换为绝对地址

```
url = "https://www.qiushibaike.com/pic"
nextpage = urlparse.urljoin(url , soup.find('span', {'class':'next'}).parent.get('href', ' '))
```

四、 问题与解决

在本次实验中，我遇到的最大的问题就是在实验 3 中文的输入输出，这也是使用 python2.7 经常会遭遇到的问题。

经过网上的查询和学习，我总结了如下几点，并应用这些知识解决了这个问题

1. 编码分为 ascii(单字节), unicode(双字节), utf8(变长字节)，而在 python 中，建议程序过程中统一使用 unicode 编码，保存文件和读取文件时使用 utf8

2. python 默认使用 ascii 编码去解释源文件。如果源文件中出现了非 ASCII 码字符，不在开头声明 encoding 会报错。我们可以声明为 utf8，告诉解释器用 utf8 去读取文件代码，这个时候源文件有中文也不会报错。

3. 在编程过程中推荐设置相应的默认编码为 utf8；读文件拿到 str 类型：str -> decode('utf8') -> unicode；程序处理：用 unicode；写文件：unicode -> encode('utf8') -> str，用 str 类型写入文件

根据以上知识，写出下面的代码

```
# coding = utf8
import sys, os
reload(sys)
sys.setdefaultencoding('utf-8')
```

```
http://pic.qiushibaike.com/system/pictures/11955/119554345/medium/app119554345.jpg  闲来无事信手之作，不喜勿喷。从经典口号“天王盖地虎，宝塔镇河妖”这个年代过来的糗百老人扶我上高楼！
http://pic.qiushibaike.com/system/pictures/11955/119554362/medium/app119554362.jpg  自动厨房，不错。
http://pic.qiushibaike.com/system/pictures/11955/119554351/medium/app119554351.jpg  今天刚看到的新闻，泡妞就泡妞吧，拿人家手机就成犯罪了
http://pic.qiushibaike.com/system/pictures/11955/119554275/medium/app119554275.jpg  谁知道医生写的什么？？在线急等
```

写中文到文件

除此之外，即使你设置了默认的编码为”utf-8”，此时 print string 没问题，但如果 print 字典（dic）和元组（tuple）还是会编码错误（因为内置的 print 函数的一些问题）。于是我就直接去掉转义字符\,然后打印出来，问题得到解决。

```
print str(parseQiushibaikePic(content)).decode('string_escape')
```

去掉转义符后 print

```
jia@jia-VirtualBox:~/Desktop/JIa/hw1$ python 3.py http://www.qiushibaike.com/pic
{'119554220': {'content': '\xe4\xba\x94\xe6\x98\x9f\xe5\xa5\xbd\xe8\xaf\x84\xef
\xbc\x81', 'imgurl': 'http://pic.qiushibaike.com/system/pictures/11955/119554220
/medium/app119554220.jpg'}, '119554254': {'content': '\xe4\xb8\xad\xe7\xa7\x8b\x
```

未设置之前，直接 print docs

```
jia@jia-VirtualBox:~/Desktop/JIa/hw1$ python 3.py http://www.qiushibaike.com/pic
{'119554220': {'content': '五星好评！', 'imgurl': 'http://pic.qiushibaike.com/s
ystem/pictures/11955/119554220/medium/app119554220.jpg'}, '119554254': {'content
': '中秋节，我告诉孩他爸整点实在的，结果……', 'imgurl': 'http://pic.qiushibaike.c
om/system/pictures/11955/119554254/medium/app119554254.jpg'}, '119554345': {'con
tent': '闲来无事信手之作，不喜勿喷。从经典口号“天王盖地虎，宝塔镇河妖”这个年代过
```

设置之后，print docs

五、 总结

通过本次实验，我学习到了：ubuntu 的基本使用方法，html 的入门知识，正则表达式的入门知识以及用 python 获取网页内容的方法和一些注意点；与此同时，在调试程序，debug 的过程中，我也感觉到自己编程的能力得到了提升。

最后，感谢老师和助教们的耐心教导和辛勤付出！

贾萧松 516030910548

2017.9.17