之前有学过一些爬虫知识，但是忘得比较快，现在已经都不记得了。索性在这记下笔记好了，也好有个记录，以后复习起来也快。

Python：3.6 / IDE：PyCharm / OS：Win7

爬虫是啥就不用赘述了。

那第一步自然是先做个最简单的页面，把百度主页爬下来吧~

```
#coding=utf-8

import urllib.request;

import urllib.parse;

import urllib.error;

'''#python爬虫第一课，爬取并生成百度首页file = urllib.request.urlopen("http://www.baidu.com");data = file.read();dataline = file.readline();print(dataline);print(data);fhandle = open("D:/baidu.html","wb");fhandle.write(data);fhandle.close();'''
```

比较简单，用fhandle来保存数据。

接下来就是怎么修改自己的头信息勒，毕竟你明着告诉别人自己是爬虫恐怕大部分网站不会让你进去(之后的源码都会忽略import)。

```
'''#Python爬虫第二课，修改header信息，伪装url = "http://blog.csdn.net/weiwei_pig/article/details/51178226";req = urllib.request.Request(url);#头信息可以通过浏览器按F12，观察Network的Header信息得到req.add_header('User-Agent','Mozilla/5.0 (Windows NT 6.1; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/57.0.2987.133 Safari/537.36');data = urllib.request.urlopen(req).read();fhandle = open("D:/test.html","wb");fhandle.write(data);fhandle.close();'''
```

也就是简单的给自己包装了一下~

接下来就是一些简单的设置，比如timeout~

```
'''#Python爬虫第三课，timeout超时设置#99次循环，每次循环试图爬取一次， 每次超时限制为1sfor i in range(1,100):try:file = urllib.request.urlopen("http://yum.iqianyue.com",timeout = 1);data = file.read();print(len(data));except Exception as e:print("Raise Error -->"+str(e));'''
```

再然后就会涉及到用爬虫与网页进行交互了，毕竟光靠静态网页是搜不到什么东西的，这里主要就是涉及到GET和POST两个办法了~下面就是用GET进行百度搜索和用POST的一个小案例，注意GET/POST使用时要多注意上传的表单格式。

```
#Python爬虫第四课，http协议GET请求，通过爬虫进行百度搜索

#中文注释

keywd = '邹屹伟';

url = "http://www.baidu.com/s?wd=";

#中文转码

keywd_code = urllib.request.quote(keywd);

url_all = url + keywd_code

req = urllib.request.Request(url_all);

data = urllib.request.urlopen(req).read();
```

```
fhandle = open("D:/test.html","wb");

fhandle.write(data);

fhandle.close();

"""

#python爬虫第五课，POST请求

url = "http://www.iqianyue.com/mypost/";

postdata = urllib.parse.urlencode({

"name":"ceo@iqianyue.com",

"pass":"aA123456"

}).encode('utf-8') #将数据使用urlencode编码后，使用encode转码为utf-8

req = urllib.request.Request(url,postdata);

req.add_header('User-Agent','Mozilla/5.0 (Windows NT 6.1; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/57.0.2987.133 Safari/537.36');

data = urllib.request.urlopen(req).read();

fhandle = open("D:/test.html","wb");

fhandle.write(data);

fhandle.close();

'''

接下来就讲讲代理服务器，讲讲怎么用Debug log和各种ERROR码了

'''

#python爬虫第六课，代理服务器

#代理来自 yum.iqianyue.com/proxy

def use_proxy(proxy_addr,url):

proxy = urllib.request.ProxyHandler({'http':proxy_addr})

opener = urllib.request.build_opener(proxy, urllib.request.HTTPHandler);

urllib.request.install_opener(opener);

data = urllib.request.urlopen(url).read().decode('utf-8');

return data

proxy_addr = "121.204.165.212:8118";

data = use_proxy(proxy_addr,"http://baidu.com");

print(len(data));

'''

'''
```

#python爬虫第七课，开启Debuglog

httphd = urllib.request.HTTPHandler(debuglevel=1);

httpshd = urllib.request.HTTPSHandler(debuglevel=1);

opener = urllib.request.build_opener(httphd,httpshd);

urllib.request.install_opener(opener);

data = urllib.request.urlopen("http://edu.51cto.com");

'''

#python爬虫第八课，URLError

#200 正常，301 永久重定向，302 临时重定向， 304 请求资源未更新， 400 非法请求， 401 未许可请求，403 禁止访问， 404 未找到页面， 500 服务器内部出错， 501 服务器不支持功能

第一天就到这吧，目前还是比较简单，就简单学了学基本的爬取。关于正则表达式和一些爬虫基本框架像Scrapy这些还都没学。

明天下班早的话复习一下正则表达式。

最后吐槽一下新浪博客，登录半天登录不上去，最后找回密码要我绑定手机，绑定完手机，通过邮件重置密码，发现我并没有输错原密码啊。就是要我绑定个手机么...然后还得想个新密码(原密码不得使用？？？那你也不让我用原密码登录啊)，狗血的是用新密码登录之后，居然提示我短时间登录太频繁，请稍后再试？？？就是不让我登录？行吧，那再见新浪博客。我对贵司的产品经理真是绝望了。

啊，吐槽完爽多了~