4行Python代码获取所在城市天气预报

链接：**https://zhuanlan.zhihu.com/p/26369491**

使用Python获取天气预报，想想是件很简单的事情。无非是发送一个HTTP请求，再解析请求返回的结果。当你真的使用Python程序去获取天气预报以后，会发现，有不少坑在等着你。这里简单记录一下我遇到的坑，供大家参考。

## 如何获取

### 传统方式

使用Python获取天气有两种不同的方法，一种是像平时爬虫一样，获取天气预报网站的HTML页面，再使用**XPath或BeautifulSoup解析HTML页面**的内容。**这是比较传统的爬虫方式**。

### 通过API获取

**此外，还有另外一种比较合适的方法——通过天气预报网站提供的API。通过API，直接获取结构化的数据，省去了解析HTML的烦恼**。

## 使用API

搜索"天气预报 API"这两个关键字，会有很多相关的内容，例如，这个

**https://wwwzhihu.com/question/20575288**

答案下就列出了不少提供API访问天气预报的网站。

然而，大部分都已经不可用了。部分可用的需要收费或者需要注册，都比较麻烦。有没有比较省事的方案呢？找来找去，我找到了中国天气网的API。无需注册直接可用，返回json格式的数据，无需使用BeautifulSoup或XPath解析，非常的方便。赞！

例如，可以直接访问下面的地址，在浏览器中查看中国天气网返回的json数据：

**http://www.weather.com.cn/data/sk/101020100.html**

有了API处理起来就很简单了，直**接使用Python世界最知名的requests访问API即可**。

## 使用requests库

安装requests：

pip install requests

检查安装是否成功：

python -c "import requests"

使用ipython测试：

In [1]: import requests

In [2]: r = requests.get('http://www.weather.com.cn/data/sk/101020100.html')

In [3]: r.status\_code

Out[3]: 200

In [4]: r.content

Out[4]: '{"weatherinfo":{"city":"\xe4\xb8\x8a\xe6\xb5\xb7","cityid":"101020100","temp":"15","WD":"\xe4\xb8\x9c\xe9\xa3\x8e","WS":"1\xe7\xba\xa7","SD":"50%","WSE":"1","time":"17:08","isRadar":"1","Radar":"JC\_RADAR\_AZ9210\_JB","njd":"\xe6\x9a\x82\xe6\x97\xa0\xe5\xae\x9e\xe5\x86\xb5","qy":"1020","rain":"0"}}'

In [5]: r.json()

Out[5]: {u'weatherinfo': {u'Radar': u'JC\_RADAR\_AZ9210\_JB',

u'SD': u'50%',

u'WD': u'\xe4\xb8\x9c\xe9\xa3\x8e',

u'WS': u'1\xe7\xba\xa7',

u'WSE': u'1',

u'city': u'\xe4\xb8\x8a\xe6\xb5\xb7',

u'cityid': u'101020100',

u'isRadar': u'1',

u'njd': u'\xe6\x9a\x82\xe6\x97\xa0\xe5\xae\x9e\xe5\x86\xb5',

u'qy': u'1020',

u'rain': u'0',

u'temp': u'15',

u'time': u'17:08'}}

## resquests库中的json方法

requests库包含一个名为json的方法，当请求的地址返回的是json格式的数据时，直接使用该方法访问即可，无需使用标准库的json库。

解决乱码

如果大家刚才在浏览器中打开了我给的地址，会发现，输出结果是乱码的。如下所示：

我们可以在ipython中，查看数据编码：

In [6]: r.encoding

Out[6]: 'ISO-8859-1'

我们知道，乱码是因为解码的字符集与编码的字符集不一样，所以才会有乱码。那么，我们怎么知道数据的编码字符集呢？这个时候就靠猜了。众所周知，utf-8因为各种优点，是使用最广泛的字符集编码，因此，我们可以尝试使用utf-8进行解码。如下所示：

In [7]: r.json()['weatherinfo']['city']

Out[7]: u'\xe4\xb8\x8a\xe6\xb5\xb7'

In [8]: '\xe4\xb8\x8a\xe6\xb5\xb7'.decode('utf-8')

Out[8]: u'\u4e0a\u6d77'

In [9]: print '\xe4\xb8\x8a\xe6\xb5\xb7'.decode('utf-8')

上海

可以看到，使用utf-8解码以后，可以正确的显示数据。也就是说，中国天气网返回给我们的数据，应该是utf-8格式的。那么，为什么会乱码呢？这可能是中国天气网的工程师水平不行，也可能是故意不想让我们使用，谁知道呢。

我们已经知道了正确的编码，接下来，只要将相应的数据，使用utf-8格式解码即可。requests库本身提供了这样的功能，如下所示：

**In [10]: r.encoding = 'utf-8'**

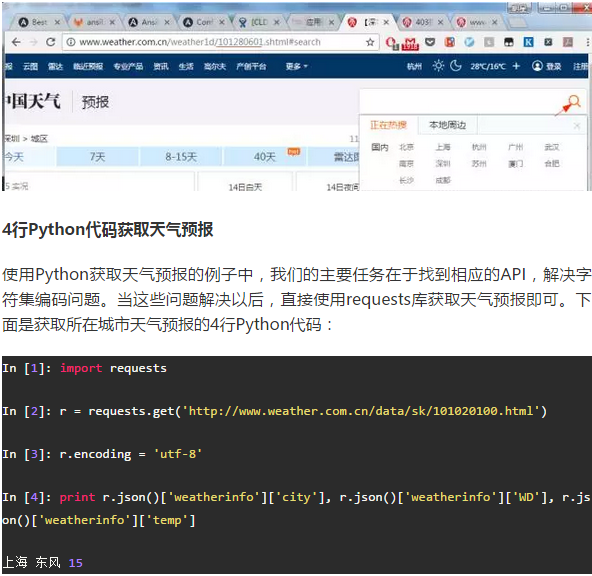
**In [11]: print r.json()['weatherinfo']['city']**

**上海**

## 获取不同城市的天气预报

前面的例子，获取的是上海的天气预报。如果想要使用中国天气网的API，获取其他城市的天气预报呢？中国天气网并没有提供相应的接口，我们只能自己想办法。

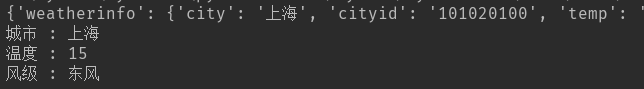
在我们测试的URL中，101020100是城市的代码，我们只需要找到其他城市的代码，将101020100替换成相应的代码即可。**查找方法是，在中国天气网的首页，搜索城市的名称，地址栏中会显示相应城市的代码**。如下所示：



**示例：**

import requests  
res = requests.get('http://www.weather.com.cn/data/sk/101020100.html')  
res.encoding = 'utf-8'  
'''  
利用json库的loads方法，可以将json字符串转化为字典  
import json  
resJson2 = json.loads(res.text)  
'''  
'''  
requests库中的response对象有一个json方法，  
直接将获取的json字符串转换成dict类型  
前提：获取的文本必须是json格式，否则报错  
'''  
resJson = res.json()  
print(resJson)  
weatherinfo = resJson['weatherinfo']  
print('城市 : '+weatherinfo['city'])  
print('温度 : '+weatherinfo['temp'])  
print('风级 : '+weatherinfo['WD'])

结果：



**写了一个简单的获取城市Id的爬虫示例：**

import requests  
from lxml import etree  
import re  
respon = requests.get('http://www.weather.com.cn/weather1d/101010100.shtml#dingzhi\_first')  
respon.encoding='utf-8'  
# print(respon.text)  
with open('./html.txt','w',encoding='utf-8') as f:  
 f.write(respon.text);  
 f.close();  
sele = etree.HTML(respon.text)  
aList = sele.xpath('//div[starts-with(@class,"w\_city city\_guonei")]/dl/dd/a[@href and @target="\_blank"]')  
cityDict = {}  
for a in aList:  
 cityName = str(a.xpath('text()')[0])  
 cityID = re.findall(r'\d{6,}',str(a.xpath('@href')[0]))[0]  
 cityDict[cityName]=cityID  
print(cityDict)  
for name in cityDict:  
 print(name + ' : '+cityDict[name])

**结果如下：**

