# 第4 天正则爬去海量图片

## 一、获取数据

参考数据源： 斗图：https://www.doutula.com/photo/list/

### 1. 打开浏览器的调试模式

首先，在打开网页后右键，点击 `检查`


### 2. 查找图片地址


1. 首先，点击调试窗口的左上角的箭头
2. 之后，将鼠标点击某一张图片上，在下方的红框中就会看到图片的链接地址，实际是 `img` 标签的一个属性: `data-original` 的值

# 3. 获取图片

双击 `data-original` 属性，之后用鼠标选定地址，再右键


> http://tva2.sinaimg.cn/bmiddle/9150e4e5gy1gct8nwvd5yj205i05bjr9.jpg

# 4. 开始下载

复制好地址后就可以使用 `requests` 模的的 `get()` 方法来请求这个数据了

图片、视频等非普通问题的文件，返回的数据都是二进制的，所以要保存到本地就需要使用文件的 `wb` 的方式写。

```
import requests,io

# io 模块中的 open 函数和 #open 函数使用方法一样
 io.open 适用 python2 和 python3 ，比较通用

url = 'http://tva2.sinaimg.cn/bmiddle/9150e4e5gy1gct8nwvd5yj205i05bjr9.jpg'

# 方法的源码： def get(url, params=None, **kwargs):

# 获取响应对象
res = requests.get(url=url)

file_name = "反应" + '.jpg'

# 二进制方式打开文件，不需要传递字符编码参数:encoding
with  io.open(file_name, 'wb') as bf:
    bf.write(res.content)
```


# 二、批量下载



## 1. 思路分析

上面的方式只是想让大家对爬虫爬取图片数据的流程有个简单的认识 可以看出来步骤非常的繁琐，且效率很低

实际上，图片的地址都是通过代码获取到的 思路就是把所有需要下载的图片地址放到一个列表中，之后循环这个列表一个一个的下载，或者分批下载。

## 2 获取整个页面内容

首先我们需要把整个页面的内容获取到，我们知道获取到的网页数据实际上是普通文本，之前我们也操作过。

那我们只需要利用字符串的方法或者利用正则的方式，把需要的地址找出来并放到一个列表中就可以了。

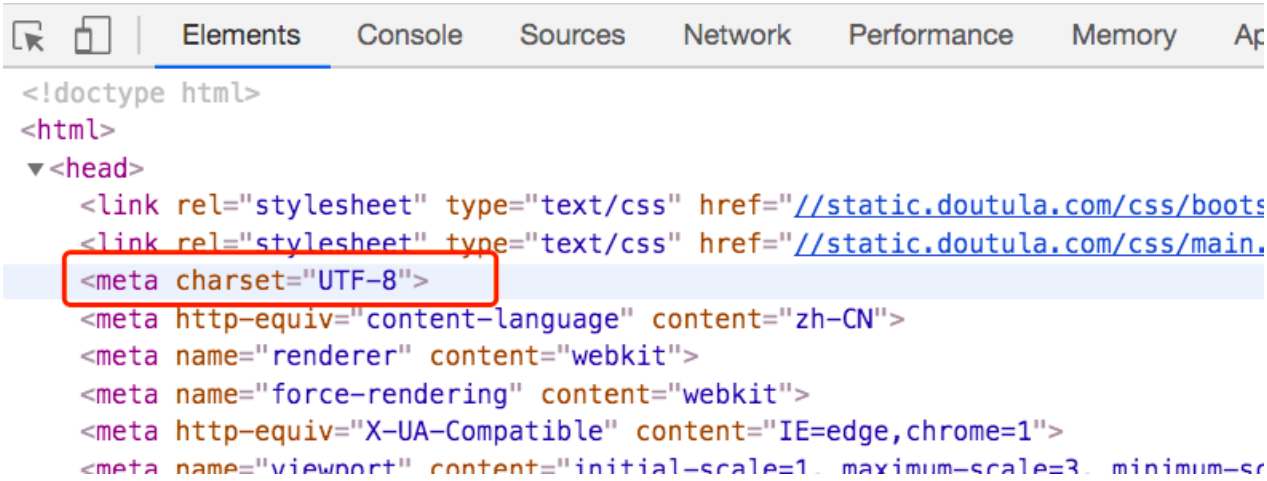
参考数据源： 斗图：https://www.doutula.com/photo/list/


```python
import requests

# 发送请求
res = requests.get('https://www.doutula.com/photo/list/')

# 获取到普通文本内容，需要指定字符编码
# 字符编码一般从页面中的 meta 标签中获取，请看下面的图片
html = str(res.content,encoding='utf-8')

# 打印出内容后，在内容中搜索关键字 data-original
print(html)
```




# 3. 清理数据

根据我们找到 `img` 标签在整个页面上的特点清理数据，获取到图片的地址

```
13  for line in html.splitlines():
14      if 'data-original=' in line:
15          print(line)
16
17          # 可以利用 break 可以只看一行，以便于观察下一步如何处理
18          break
```

```
$ /Users/yanshunjun/.virtualenvs/QF-Online2/bin/python3 "/Users/yanshunjun/Mygitlab/devops/打开你的任督二脉/web_server/test2.py"
<img referrerpolicy="no-referrer" src="//static.doutula.com/img/loader.gif?33" data-original="http://tva4.sinaimg.cn/bmiddle/9150e4e5gy1gct8m7o4xqj206o06ojs5.jpg" alt="看来你不懂珍惜生命啊" class="img-responsive lazy image_dta" data-backup="http://img.doutula.com/production/uploads/image/2020/03/14/20200314182907_tSFzyM.png">
(QF-Online2)
$
```

# 4. 保存数据