requests库

1. requests库的get()用法

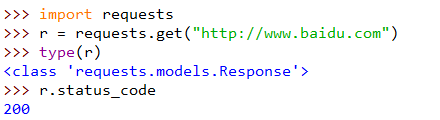
R = requests.get(url)

右边: 构造一个向服务器请求资源的**Requests**对象

左边：返回一个包含服务器资源的**Response**对象

**Response**对象包含爬虫返回的内容

1.1 **Response**对象

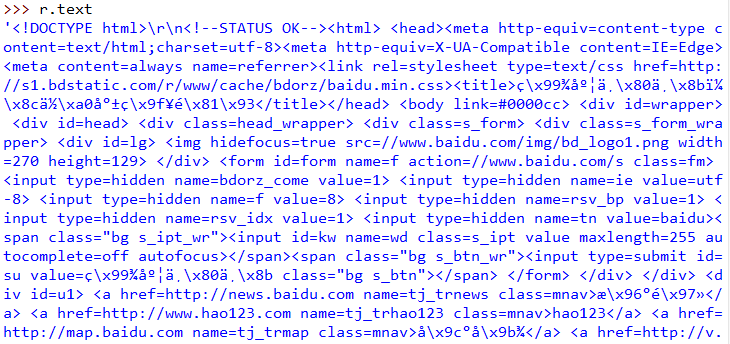


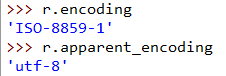
这里的r即为Response对象（参见type(r)）

status\_code为Response对象的一个属性，表示HTTP请求的返回状态，200表示连接成功，404表示失败

1.1.1 **Response**对象的属性

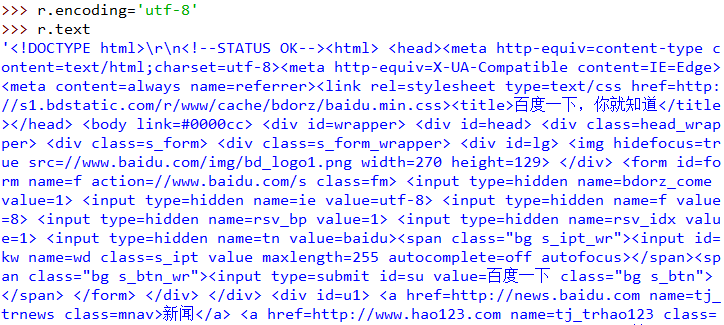
|  |  |
| --- | --- |
| 属性 | 说明 |
| r.status\_code | HTTP请求的返回状态，200表示连接成功，404表示失败 |
| r.text | HTTP相应内容的字符串形式，即url对应的页面内容 |
| r.encoding | 从HTTP header中猜测的响应内容编码方式 |
| r.apparent\_coding | 从内容中分析出的响应内容编码方式 |
| r.content | HTTP响应内容的二进制形式 |





注意：r.encoding只是从header中猜测响应内容的编码方式，r.text根据r.encoding显示网页内容，则此时可能出现不了中文

而r.apparent\_coding是根据网页内容分析出编码方式，所以将r.encoding改为’utf-8’后，可以显示出中文



1. 爬取网页的通用代码框架

|  |
| --- |
| **import** requests  **def get\_HTML\_Text(**url**):  try:** r **=** requests.get**(**url,timeout**=**30**)** r.raise\_for\_status**()** r.encoding **=** r.apparent\_encoding  **return** r.text  **except:  return "产生异常"  if** \_\_name\_\_**=="\_\_main\_\_":** url **= "http://www.baidu.com"** print**(**get\_HTML\_Text**(**url**))** |

HTTP协议

HTTP ：Hpertext Transfer Protocol ,超文本传输协议

HTTP是一个基于“请求与响应”模式的、无状态的应用层协议，HTTP协议采用URL作为定位网络资源的标识

URL：URL是通过HTTP协议存取资源的Internet路径，一个URL对应一个数据资源

HTTP协议对资源的操作

|  |  |
| --- | --- |
| 方法 | 说明 |
| GET | 请求获取URL位置的资源 |
| HEAD | 请求获取URL位置资源的相映消息报告，即获得该资源的头部信息 |
| POST | 请求向URL位置资源后附加新的数据 |
| PUT | 请求向URL位置存储一个资源，覆盖原URL位置的资源 |
| PATCH | 请求局部更新URL位置的资源，即改变该处资源的部分内容 |
| DELETE | 请求删除URL位置存储的资源 |

HTTP协议与Requests库

|  |  |  |
| --- | --- | --- |
| HTTP协议方法 | Requests库方法 | 功能一致性 |
| GET | requests.get() | 一致 |
| HEAD | requests.head() | 一致 |
| POST | requests.post() | 一致 |
| PUT | requests.put() | 一致 |
| PATCH | requests.patch() | 一致 |
| DELETE | requests.delete() | 一致 |

2.requests库的head()用法

|  |
| --- |
| **import** requests  r **=** requests.head**("http://www.baidu.com")** print**(**r.headers**)**  {'Server': 'bfe/1.0.8.18', 'Date': 'Thu, 23 Mar 2017 14:16:38 GMT', 'Content-Type': 'text/html', 'Last-Modified': 'Mon, 13 Jun 2016 02:50:45 GMT', 'Connection': 'Keep-Alive', 'Cache-Control': 'private, no-cache, no-store, proxy-revalidate, no-transform', 'Pragma': 'no-cache', 'Content-Encoding': 'gzip'}  Process finished with exit code 0 |

|  |
| --- |
| **import** requests  r **=** requests.head**("http://www.baidu.com")** print**(**r.text**)**  Process finished with exit code 0 |

注意只用head方法，r.text什么都没有

|  |
| --- |
| **import** requests  r **=** requests.head**("http://www.baidu.com")** print**(**r.encoding,r.apparent\_encoding**)**  ISO-8859-1 None  Process finished with exit code 0 |

注意只用head方法，r.apparent\_encoding为None

1. requests库的post()用法

????