链滴

Python 爬虫系列（二）基本库的使用

作者：luofeng0603
原文链接：https://ld246.com/article/1679314427712
来源网站：链滴

urllib的使用

urlib是python内置的请求库，不需要额外安装

urllib包含如下4个模块：

- request：最基本的http请求模块，模拟请求发送
- error：异常处理模块
- parse：工具模块
- robotparser：用来识别网站的robots.txt，用的很少

urlopen发送请求

使用urlopen来发送请求

一个小例子：

```
import urllib.request

response = urllib.request.urlopen('https://www.python.org')
print(response.read().decode('utf-8'))
```

response 是一个HTTPRequest对象，主要包含read、readinto、getheader、getheaders、fileno方法，以及msg、version、status、reason、debuglevel、closed等属性。

urlopen的API：

```
urllib.request.urlopen(url, data=None,[timeout,]*,cafile=None,capath=None,cadeFault=False,
ontext=None)
```

用法就不写了，写几个小例子吧：

```
import urllib.request
import urllib.parse

# data参数需要变成字节码
data = bytes(urllib.parse.urlencode({'nanme': 'zhangsan'}), encoding='utf-8')
response = urllib.request.urlopen('https://www.httpbin.org/post', data = data)
print(response.read().decode('utf-8'))

# timeout
response = urllib.request.urlopen('https://www.httpbin.org/get', timeout=0.1)
print(response.read())
```

Request发送请求

复杂请求，urlopen无法满足，比如需要附加Headers，这时候需要可以用更强大的Request

```
# Request的基本使用
request = urllib.request.Request('https://www.python.org')
response = urllib.request.urlopen(request)
print(response.read().decode('utf-8'))

# Request多参数
url = 'https://www.httpbin.org/post'
headers = {
    'User-Agent': 'Mozilla/4.0 (compatible; MSIE 5.5; Windows NT)',
    'Host': 'www.httpbin.org'
}
dict = {'name': 'zhangsan'}
data = bytes(urllib.parse.urlencode(dict), encoding='utf-8')
req = urllib.request.Request(url=url, data=data, headers=headers, method='POST')
response = urllib.request.urlopen(req)
print(response.read().decode('utf-8'))
```