爬虫步骤

1、 确定需求

2、 寻找需求

3、 发送请求

4、 解析数据

5、 存储数据

请求方式

GET 和 POST

- GET是默认的HTTP请求方法，用于直接输入网址的方式去访问网页
- POST方法主是向Web服务器提交表单数据，通常表单提交时采用POST方法
- GET把请求参数包含在URL中，POST通过请求体传递参数
- GET相对POST不安全，参数直接暴露在URL上，用来传递敏感信息

Requests

安装

```
pip install requests
```

发送get请求

```
# 百度
import requests
url = "https://www.baidu.com/"
response = requests.get(url)
response.encoding='utf-8'

print (response.text)
print (response.content)
print (response.content.decode('utf-8'))
print (response.headers)
print (response.status_code)
print (response.url)
print(response.request.headers)
```

添加请求头

```
# 西祠代理
import  requests

url = 'https://www.xicidaili.com/nn/'
headers = {
    'User-Agent': 'Mozilla/5.0 (Windows NT 6.1; Win64; x64)
AppleWebKit/537.36 (KHTML, like Gecko) Chrome/72.0.3626.109
Safari/537.36'
}
resp = requests.get(url, headers=headers)
with open('xici.html', 'wb') as f:
    f.write(resp.content)
```

发送post请求

参数通过请求体提交

- requests.post()

```python
# 百度翻译
import requests
import json

url = 'https://fanyi.baidu.com/sug'
data = {
    'kw':'w'
}
resp = requests.post(url, data=data)
res = json.loads(resp.text)
print(res)
```

添加cookie

```python
# 微博 / 京东
url = 'https://weibo.com'
headers = {
    'User-Agent': '',
    'Cookie': ''
}

resp = requests.get(url, headers=headers)
print(resp.status_code)
with open ('weibo.html', 'wb')as f:
    f.write(resp.content)
```

人人影视登录

```
url = 'http://www.zmz2019.com/User/Login/ajaxLogin'

httphead={
    'User-Agent':'Mozilla/5.0 (Windows NT 6.3; WOW64)
AppleWebKit/537.36 (KHTML, like Gecko) Chrome/50.0.2661.94
Safari/537.36',
}

index_url = 'http://www.zmz2019.com/User/user'
data = {
    'account':'',
    'password':'',
    'remember':'0',
    'url_back':index_url,
}
resp = requests.post(url, data=data, headers=httphead)
res = json.loads(resp.text)
```

requests.session()

在同一个Session实例发出的所有请求都保持同一个cookies, 而requests模块每次会自动处理cookies，这样就很方便地处理登录时的cookies问题

```
url = 'http://www.zmz2019.com/User/Login/ajaxLogin'

httphead={
    'User-Agent':'Mozilla/5.0 (Windows NT 6.3; WOW64)
AppleWebKit/537.36 (KHTML, like Gecko) Chrome/50.0.2661.94
Safari/537.36',
}

index_url = 'http://www.zmz2019.com/User/user'
data = {
    'account':'',
    'password':'',
    'remember':'0',
    'url_back':index_url,
}
# ###### 1用的session对象
s = requests.session()
resp = s.post(url, data=data, headers=httphead)
# resp.encoding = 'utf-8'
res = json.loads(resp.text)
print(resp.text)
print(res)

"""{'status': 1, 'info': '登录成功! ', 'data': {'url_back':
'http://www.zmz2019.com/User/user'}}
"""
###### 2 用的是同一个session对象
resp_index = s.get(index_url, headers=httphead)
```