

# 爬虫简单入门

- 爬虫简单入门
- 爬搜狗
- 爬知乎
- 爬百度
- ...

## 案例1

### 1. 下载爬虫框架:requests

```
pip install requests
```

### 2. 使用request爬网页流程

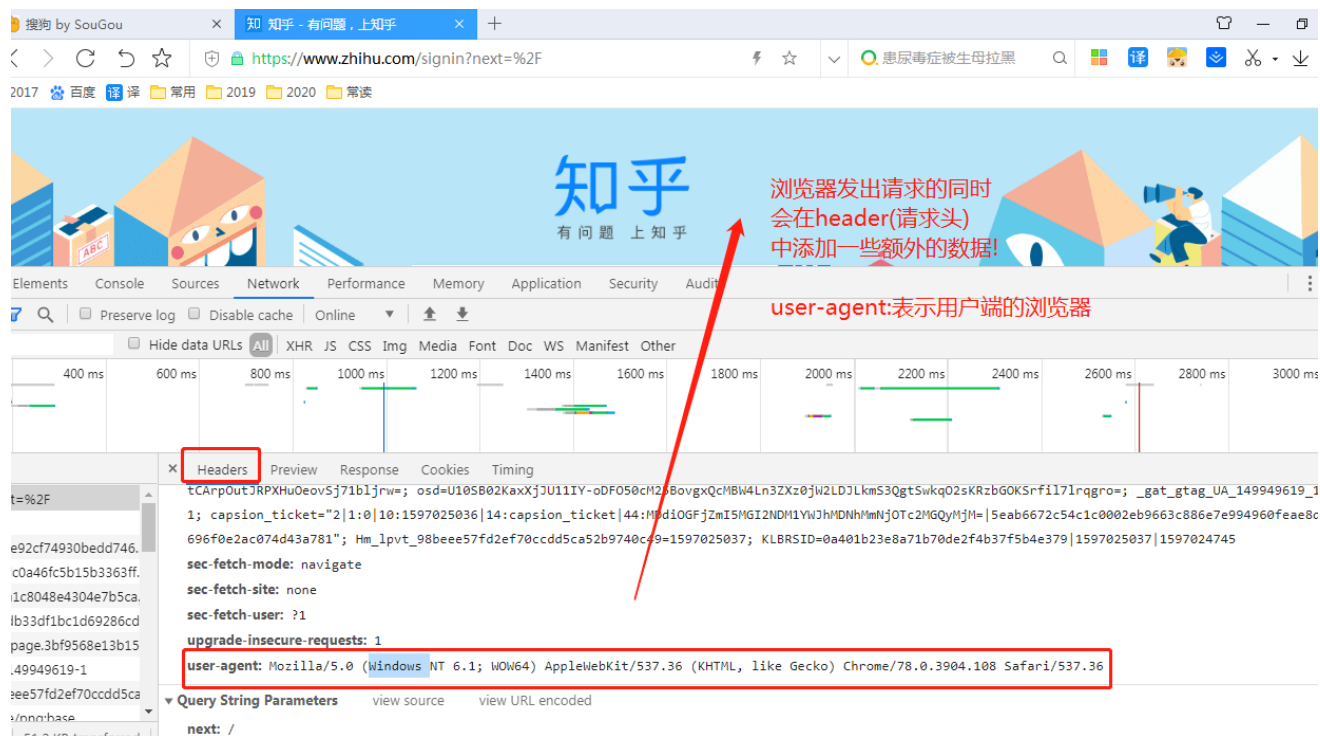
```
# 爬搜狗(用程序模拟浏览器访问搜狗)
# 1. 定义url
# 2. 发出请求
# 3. 获取响应回来的页面数据
# 4. 存储
```

```
'''
爬搜狗
1. 定义url https://www.sougou.com
2. 发起请求
3. 获取相应的页面数据
4. 存储
'''
import requests

# 1. url (CSDN加反扒了!)
# url = 'http://www.sougou.com'
url = 'https://blog.csdn.net/qq_30242609/article/details/53788228#commentBox'
# 2. 发送请求
res = requests.get(url=url)
# 3. 获取页面数据
res_text = res.text
print(res_text)
# 4. 存储(保存到数据,txt.csv.表格...)
with open('博文.html',mode='w',encoding='utf8') as f:
    f.write(res_text)
```

## 案例2:爬知乎

- 浏览器正常访问能得到页面. 因为请求中有user-agent变量.标识我的请求是通过浏览器发出的!



user-agent: Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/78.0.3904.108 Safari/537.36

- 通过requests模块写代码访问网站,被拒绝返回空页面或错误页面

因为:被检测到不是通过浏览器访问的!

```
案例2 x
D:\py_env\flask\Scripts\python.exe E:/workplace/pycharm/爬!
<html>
<head><title>400 Bad Request</title></head>
<body bgcolor="white">
<center><h1>400 Bad Request</h1></center>
<hr><center>openresty</center>
</body>
</html>
```

Terminal Python Console ▶ 4: Run ≡ 6: TODO

```
'''
知乎
https://www.zhihu.com
'''
import requests
# 1.url
#url = 'https://www.zhihu.com'
url = 'https://blog.csdn.net/qg_30242609/article/details/53788228#commentBox'
# 指定请求头,记性UA伪装(伪装为浏览器)
headers = {
    'user-agent':'Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.36 (KHTML, like
    Gecko) Chrome/78.0.3904.108 Safari/537.36'
}
#2. 请求(知乎加反扒,返回就错误页面或空页面)
response = requests.get(url,headers=headers)
# 3. 获取返回文本
#print(response.text)
with open('博文2.html',mode='w',encoding='utf8') as f:
    f.write(response.text)
```

### 案例3:爬百度翻译

Ajax请求, 后台返回的接口数据

Request URL: `https://fanyi.baidu.com/sug`

Request Method: `POST`

Status Code: `200`

Remote Address: `220.195.22.202:443`

Referrer Policy: `no-referrer-when-downgrade`

Form Data

`kw: big`

表单需要传递的参数kw:值

errno: 0, data: [{k: "big", v: "adj. (体积、程度、数量等)大的, 巨大的; 年龄较大的; 重大的; 严重的;"}]

data: [{k: "big", v: "adj. (体积、程度、数量等)大的, 巨大的; 年龄较大的; 重大的; 严重的;"}]

big", v: "adj. (体积、程度、数量等)大的, 巨大的; 年龄较大的; 重大的; 严重的;"}]

1: {k: "bigbang", v: "2006年YG Entertainment推出的重点新人组合。bigbang为宇宙大爆炸"}]

2: {k: "biggest", v: "adv. 大大; 给人印象深地; adj. (体积、程度、数量等)大的, 巨大的;"}]

3: {k: "bigger", v: "adv. 大大; 给人印象深地; adj. (体积、程度、数量等)大的, 巨大的;"}]

4: {k: "big bang", v: "n. (某些科学家提出的关于宇宙起源的)创世大爆炸;"}

errno: 0

```
'''
爬百度翻译
'''

import requests
# 1. 地址
url = 'https://fanyi.baidu.com/sug'
# 2. 发请求加请求头
h = {
    'user-agent': 'Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/78.0.3904.108 Safari/537.36'
}
# 请求加参数
print('万能英译汉:')
kw = input('请输入一个单词:')
mydata = {'kw': kw}
```

```
# get方法表示get请求, post表示post请求
```

```
response = requests.post(url, headers=h, data=mydata) # url参数, headers请求头, data参数
```

```
print(response.json()) #获取json数据
```

## 爬取百度产品列表

- `https://www.baidu.com/more/`
- 熟悉 requests 的相关方法

## 作业

---

- 爬小狗照片

[https://image.baidu.com/search/index?tn=baiduimage&ipn=r&ct=201326592&cl=2&lm=-1&st=-1&fm=result&fr=&sf=1&fmq=1597028621766\\_R&pv=&ic=&nc=1&z=&hd=&latest=@right=&se=1&showtab=0&fb=0&width=&height=&face=0&istype=2&ie=utf-8&sid=&word=小狗](https://image.baidu.com/search/index?tn=baiduimage&ipn=r&ct=201326592&cl=2&lm=-1&st=-1&fm=result&fr=&sf=1&fmq=1597028621766_R&pv=&ic=&nc=1&z=&hd=&latest=@right=&se=1&showtab=0&fb=0&width=&height=&face=0&istype=2&ie=utf-8&sid=&word=小狗)