使用 Requests 实现一个简单网页爬虫

第一节我们简单介绍了爬虫的基本原理，理解原理可以帮助我们更好的实现代码。Python 提供了非常多工具去实现 HTTP 请求，但第三方开源库提供的功能更丰富，你无需从 socket 通信开始写，比如使用Pyton内建模块 urllib 请求一个 URL 代码示例如下：

```python
import ssl

from urllib.request import Request
from urllib.request import urlopen

context = ssl._create_unverified_context()

# HTTP 请求
request =
Request(url="https://foofish.net/pip.html",
                method="GET",
                headers={"Host":
"foofish.net"},
                data=None)

# HTTP 响应
response = urlopen(request, context=context)
headers = response.info()  # 响应头
content = response.read()  # 响应体
code = response.getcode()  # 状态码
```

发起请求前首先要构建请求对象 Request，指定 url 地址、请求方法、请求头，这里的请求体 data 为空，因为你不需要提交数据给服务器，所以你也可以不指定。urlopen 函数会自动与目标服务器建立连接，发送 HTTP 请求，该函数的返回值是一个响应对象 Response，里面有响应头信息，响应体，状态码之类的属性。

但是，Python 提供的这个内建模块过于低级，需要写很多代码，使用简单爬虫可以考虑 Requests，Requests 在GitHub 有近30k的 Star，是一个很Pythonic的框架。先来简单熟悉一下这个框架的使用方式

安装 requests

```
pip install requests
```

GET 请求

```
>>> r = requests.get("https://httpbin.org/ip")
>>> r
<Response [200]>  # 响应对象
>>> r.status_code   # 响应状态码
200


>>> r.content   # 响应内容
'{\n  "origin": "183.237.232.123"\n}\n'
```

POST 请求

```
>>> r = requests.post('http://httpbin.org/post',
data = {'key':'value'})
```

自定义请求头

这个经常会用到，服务器反爬虫机制会判断客户端请求头中的User-Agent是否来源于真实浏览器，所以，我们使用Requests经常会指定UA伪装成浏览器发起请求

```
>>> url = 'https://httpbin.org/headers'
>>> headers = {'user-agent': 'Mozilla/5.0'}
>>> r = requests.get(url, headers=headers)
```

参数传递

很多时候URL后面会有一串很长的参数，为了提高可读性，requests支持将参数抽离出来作为方法的参数（params）传递过去，而无需附在 URL 后面，例如请求 url http://bin.org/get?key=val ，可使用

```
>>> url = "http://httpbin.org/get"
>>> r = requests.get(url, params={"key":"val"})
>>> r.url
u'http://httpbin.org/get?key=val'
```

指定Cookie

Cookie 是web浏览器登录网站的凭证，虽然 Cookie 也是请求头的一部分，我们可以从中剥离出来，使用 Cookie 参数指定

```
>>> s =
requests.get('http://httpbin.org/cookies',
cookies={'from-my': 'browser'})
>>> s.text
u'{\n  "cookies": {\n    "from-my": "browser"\n
}\n}\n'
```

设置超时

当发起一个请求遇到服务器响应非常缓慢而你又不希望等待太久时，可以指定 timeout 来设置请求超时时间，单位是秒，超过该时间还没有连接服务器成功时，请求将强行终止。

```
r = requests.get('https://google.com', timeout=5)
```

设置代理

一段时间内发送的请求太多容易被服务器判定为爬虫，所以很多时候我们使用代理IP来伪装客户端的真实IP。

```
import requests

proxies = {
    'http': 'http://127.0.0.1:1080',
    'https': 'http://127.0.0.1:1080',
}

r =
requests.get('http://www.kuaidaili.com/free/',
proxies=proxies, timeout=2)
```

Session

如果想和服务器一直保持登录（会话）状态，而不必每次都指定 cookies，那么可以使用 session，Session 提供的API和 requests 是一样的。
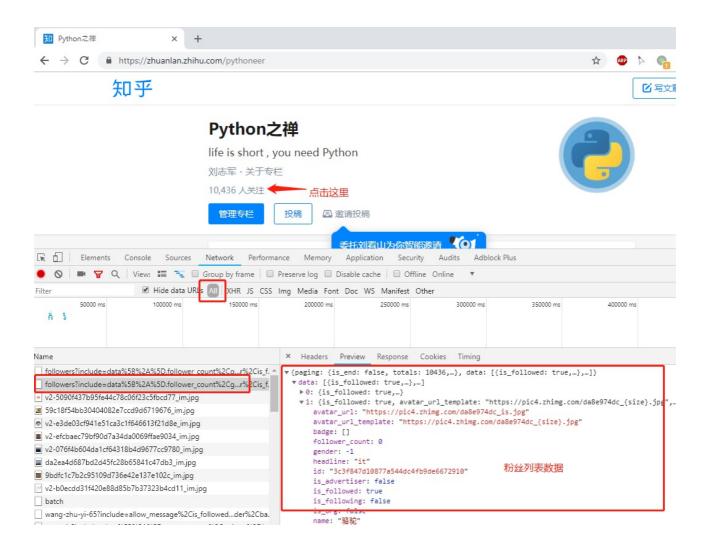
```
import requests

s = requests.Session()
s.cookies =
requests.utils.cookiejar_from_dict({"a": "c"})
r = s.get('http://httpbin.org/cookies')
print(r.text)
# '{"cookies": {"a": "c"}}'

r = s.get('http://httpbin.org/cookies')
print(r.text)
# '{"cookies": {"a": "c"}}'
```
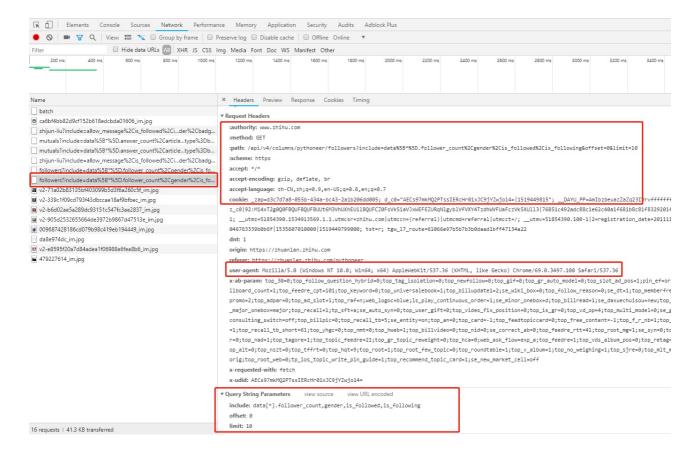
小试牛刀

现在我们使用Requests完成一个爬取知乎专栏用户关注列表的简单
爬虫为例，找到任意一个专栏，打开它的关注列表
(https://zhuanlan.zhihu.com/pythoneer)。用 Chrome 找到获
取粉丝列表的请求地
址：https://www.zhihu.com/api/v4/columns/pythoneer/follo
include=data%5B%2A%5D.follower_count%2Cgender%2Cis_foll
(https://www.zhihu.com/api/v4/columns/pythoneer/follower
include=data%5B%2A%5D.follower_count%2Cgender%2Cis_foll
我是怎么找到的？就是逐个点击左侧的请求，观察右边是否有数据出
现，那些以 .jpg，js，css 结尾的静态资源可直接忽略。

现在我们用 Requests 模拟浏览器发送请求给服务器，写程序前，我们要先分析出这个请求是怎么构成的，请求URL是什么？请求头有哪些？查询参数有哪些？只有清楚了这些，你才好动手写代码，掌握分析方法很重要，否则一头雾水。

回到前面那个URL，我们发现这个URL是获取粉丝列表的接口，然后再来详细分析一下这个请求是怎么构成的。

- 请求URL：
  https://www.zhihu.com/api/v4/columns/pythoneer/follo
- 请求方法：GET
- user-agent: Mozilla/5.0 (Windows NT 10.0; Win64; x64)
  AppleWebKit/537.36 (KHTML, like Gecko)
  Chrome/69.0.3497.100 Safari/537.36
- 查询参数：

  - include:
    data[*].follower_count,gender,is_followed,is_followi
  - offset: 0
  - limit: 10

利用这些请求数据我们就可以用requests这个库来构建一个请求，通过Python代码来抓取这些数据。

```
import requests
```

```python
class SimpleCrawler:

    def crawl(self, params=None):
        # 必须指定UA，否则知乎服务器会判定请求不合法

        url = 
"https://www.zhihu.com/api/v4/columns/pythoneer/f
ollowers"
        # 查询参数
        params = {"limit": 20,
                  "offset": 0,
                  "include": 
"data[*].follower_count, gender, is_followed,
is_following"}

        headers = {
            "authority": "www.zhihu.com",
            "user-agent": "Mozilla/5.0 (Windows
NT 10.0; Win64; x64) "
                          "AppleWebKit/537.36
(KHTML, like Gecko) Chrome/63.0.3239.84
Safari/537.36",
        }
        response = requests.get(url,
headers=headers, params=params)
        print("请求URL: ", response.url)
        # 你可以先将返回的响应数据打印出来，拷贝到
http://www.kjson.com/jsoneditor/ 分析其结构。
        print("返回数据: ", response.text)

        # 解析返回的数据
```

```python
        for follower in
response.json().get("data"):
            print(follower)


if __name__ == '__main__':
    SimpleCrawler().crawl()
```

输出：

请求URL：
https://www.zhihu.com/api/v4/columns/pythoneer/followers?
limit=20&offset=0&include=data%5B%2A%5D.follower_count%2C+gender%2C+is_followed%2C+is_following
返回数据：{'paging': {'is_end': False, 'totals': 10436, 'previous': 'http://www.zhihu.com/api/v4/columns/pythoneer/followers? include=data%5B%2A%5D.follower_count%2C+gender%2C+is_followed%2C+is_following&limit=20&offset=0', 'is_start': True, 'next': 'http://www.zhihu.com/api/v4/columns/pythoneer/followers? include=data%5B%2A%5D.follower_count%2C+gender%2C+is_followed%2C+is_following&limit=20&offset=20'}, 'data': [{'is_followed': False, 'avatar_url_template': 'https://pic4.zhimg.com/v2-71a02b83135bf403099b5d3f6a260c5f_{size}.jpg', 'user_type': 'people', 'is_following': False, 'headline': '欢迎关注公众号：「乔尔事务所」 → 城市 | 地理 | 规划 | 经济', 'url_token': 'joe1324', 'id':

'130894ff204e72b96719a09b5a6565bc', 'type':
'people', 'name': '乔尔爱冒险', 'url':
'http://www.zhihu.com/api/v4/people/130894ff204e7
2b96719a09b5a6565bc', 'gender': 1,
'is_advertiser': False, 'avatar_url':
'https://pic4.zhimg.com/v2-
71a02b83135bf403099b5d3f6a260c5f_is.jpg',
'is_org': False, 'follower_count': 4, 'badge':
[]}, {'is_followed': False,
'avatar_url_template':
'https://pic3.zhimg.com/v2-
339c1f09cd793f43dbccae18af9bfbec_{size}.jpg',
'user_type': 'people', 'is_following': False,
'headline': '', 'url_token': 'hei-tai-yang-91',
'id': 'bbeba0903e14a32d19b27ffa0016ac55', 'type':
'people', 'name': 'JOJO', 'url':
'http://www.zhihu.com/api/v4/people/bbeba0903e14a
32d19b27ffa0016ac55', 'gender': -1,
'is_advertiser': False, 'avatar_url':
'https://pic3.zhimg.com/v2-
339c1f09cd793f43dbccae18af9bfbec_is.jpg',
'is_org': False, 'follower_count': 4, 'badge':
[]}, {'is_followed': False,
'avatar_url_template':
'https://pic3.zhimg.com/v2-
b6d02ae5a289dc93151c547fc3ae2837_{size}.jpg',
'user_type': 'people', 'is_following': False,
'headline': '2333333333', 'url_token': 'wo-zai-
xue-xi-ni-87', 'id':
'696dcf2ff66b23ae0ff39474f9f48c5e', 'type':
'people', 'name': '我在学习呢', 'url':
'http://www.zhihu.com/api/v4/people/696dcf2ff66b2
3ae0ff39474f9f48c5e', 'gender': -1,

'is_advertiser': False, 'avatar_url': 'https://pic3.zhimg.com/v2-b6d02ae5a289dc93151c547fc3ae2837_is.jpg', 'is_org': False, 'follower_count': 0, 'badge': []}, {'is_followed': False, 'avatar_url_template': 'https://pic2.zhimg.com/v2-905d2532655664de3972b9867d47513e_{size}.jpg', 'user_type': 'people', 'is_following': False, 'headline': '逻辑至上', 'url_token': 'wang-zhu-yi-65', 'id': '0580427bffd2815ee8e8007fbb2419b1', 'type': 'people', 'name': '王朮屹', 'url': 'http://www.zhihu.com/api/v4/people/0580427bffd2815ee8e8007fbb2419b1', 'gender': 1, 'is_advertiser': False, 'avatar_url': 'https://pic2.zhimg.com/v2-905d2532655664de3972b9867d47513e_is.jpg', 'is_org': False, 'follower_count': 17, 'badge': []}, {'is_followed': False, 'avatar_url_template': 'https://pic1.zhimg.com/009687428186cd079b98c419eb194449_{size}.jpg', 'user_type': 'people', 'is_following': False, 'headline': '笨鸟先飞', 'url_token': 'wei-sen-ken', 'id': '7d938d55321c34eee0cde88359e827ff', 'type': 'people', 'name': '魏森肯', 'url': 'http://www.zhihu.com/api/v4/people/7d938d55321c34eee0cde88359e827ff', 'gender': 1, 'is_advertiser': False, 'avatar_url': 'https://pic1.zhimg.com/009687428186cd079b98c419eb194449_is.jpg', 'is_org': False, 'follower_count': 0, 'badge': []}, {'is_followed': False, 'avatar_url_template':

'https://pic4.zhimg.com/da8e974dc_{size}.jpg',
'user_type': 'people', 'is_following': False,
'headline': 'holo', 'url_token': 'zhang-feng-
wei', 'id': 'c6ba181c10593256fa71d9ecb8dcdb90',
'type': 'people', 'name': 'goingon', 'url':
'http://www.zhihu.com/api/v4/people/c6ba181c10593
256fa71d9ecb8dcdb90', 'gender': 1,
'is_advertiser': False, 'avatar_url':
'https://pic4.zhimg.com/da8e974dc_is.jpg',
'is_org': False, 'follower_count': 9, 'badge':
[]}, {'is_followed': False,
'avatar_url_template':
'https://pic2.zhimg.com/v2-
e8595f20a7d84adea1f06988e6fea8b6_{size}.jpg',
'user_type': 'people', 'is_following': False,
'headline': '在读大学生', 'url_token': 'fu-teng-
53', 'id': 'b7ef951f5b89373ed369a07ea5cb7553',
'type': 'people', 'name': 'fuPatrick', 'url':
'http://www.zhihu.com/api/v4/people/b7ef951f5b8937
3ed369a07ea5cb7553', 'gender': -1,
'is_advertiser': False, 'avatar_url':
'https://pic2.zhimg.com/v2-
e8595f20a7d84adea1f06988e6fea8b6_is.jpg',
'is_org': False, 'follower_count': 1, 'badge':
[]}, {'is_followed': False,
'avatar_url_template':
'https://pic4.zhimg.com/da8e974dc_{size}.jpg',
'user_type': 'people', 'is_following': False,
'url_token': 'hu-chen-92-42-74', 'id':
'f88db2870e4e9cdbe589b5846fccd70c', 'type':
'people', 'name': '胡沈', 'url':
'http://www.zhihu.com/api/v4/people/f88db2870e4e9
cdbe589b5846fccd70c', 'gender': -1,

'is_advertiser': False, 'avatar_url':
'https://pic4.zhimg.com/da8e974dc_is.jpg',
'is_org': False, 'follower_count': 0, 'badge':
[]}, {'is_followed': False,
'avatar_url_template':
'https://pic4.zhimg.com/da8e974dc_{size}.jpg',
'user_type': 'people', 'is_following': False,
'headline': '', 'url_token': 'zhi-bei-bu-nan',
'id': '63220b9e8cadb266b236ed0e1d76d1ed', 'type':
'people', 'name': '只北不南', 'url':
'http://www.zhihu.com/api/v4/people/63220b9e8cadb
266b236ed0e1d76d1ed', 'gender': -1,
'is_advertiser': False, 'avatar_url':
'https://pic4.zhimg.com/da8e974dc_is.jpg',
'is_org': False, 'follower_count': 0, 'badge':
[]}, {'is_followed': False,
'avatar_url_template':
'https://pic4.zhimg.com/479227614_{size}.jpg',
'user_type': 'people', 'is_following': False,
'headline': '道路不止一条 终点不必相同 人生各自精彩',
'url_token': 'xi-zi-31-56', 'id':
'8e3014e0c8e4bf4cd73f34398ee21379', 'type':
'people', 'name': '西子', 'url':
'http://www.zhihu.com/api/v4/people/8e3014e0c8e4b
f4cd73f34398ee21379', 'gender': 1,
'is_advertiser': False, 'avatar_url':
'https://pic4.zhimg.com/479227614_is.jpg',
'is_org': False, 'follower_count': 1, 'badge':
[]}, {'is_followed': False,
'avatar_url_template':
'https://pic1.zhimg.com/eff9fca94_{size}.jpg',
'user_type': 'people', 'is_following': False,
'headline': 'I know nothing.', 'url_token':

'42195', 'id':
'ce5e19c125f8570bee02cdb414f6e233', 'type':
'people', 'name': 'Gladius', 'url':
'http://www.zhihu.com/api/v4/people/ce5e19c125f85
70bee02cdb414f6e233', 'gender': 1,
'is_advertiser': False, 'avatar_url':
'https://pic1.zhimg.com/eff9fca94_is.jpg',
'is_org': False, 'follower_count': 74, 'badge':
[]}, {'is_followed': False,
'avatar_url_template':
'https://pic4.zhimg.com/da8e974dc_{size}.jpg',
'user_type': 'people', 'is_following': False,
'url_token': 'willbillion', 'id':
'0ee103930c20c21d314a998a371abe82', 'type':
'people', 'name': 'willbillion', 'url':
'http://www.zhihu.com/api/v4/people/0ee103930c20c
21d314a998a371abe82', 'gender': -1,
'is_advertiser': False, 'avatar_url':
'https://pic4.zhimg.com/da8e974dc_is.jpg',
'is_org': False, 'follower_count': 0, 'badge':
[]}, {'is_followed': False,
'avatar_url_template':
'https://pic4.zhimg.com/da8e974dc_{size}.jpg',
'user_type': 'people', 'is_following': False,
'headline': '', 'url_token': 'yang-guang-xiao-
tong', 'id': '1637b7761fe90da99519d4878a4a991f',
'type': 'people', 'name': '阳光小桐', 'url':
'http://www.zhihu.com/api/v4/people/1637b7761fe90
da99519d4878a4a991f', 'gender': 0,
'is_advertiser': False, 'avatar_url':
'https://pic4.zhimg.com/da8e974dc_is.jpg',
'is_org': False, 'follower_count': 1, 'badge':
[]}, {'is_followed': False,

'avatar_url_template': 'https://pic4.zhimg.com/da8e974dc_{size}.jpg', 'user_type': 'people', 'is_following': False, 'headline': 'Access', 'url_token': 'wang-hua-18-79', 'id': 'd0a5fd983546174636af438651c13018', 'type': 'people', 'name': '王华', 'url': 'http://www.zhihu.com/api/v4/people/d0a5fd983546174636af438651c13018', 'gender': -1, 'is_advertiser': False, 'avatar_url': 'https://pic4.zhimg.com/da8e974dc_is.jpg', 'is_org': False, 'follower_count': 0, 'badge': []}, {'is_followed': False, 'avatar_url_template': 'https://pic3.zhimg.com/5fa6a8e5aa7207fcbfd917dc8ce50d50_{size}.jpg', 'user_type': 'people', 'is_following': False, 'headline': '', 'url_token': 'oriver', 'id': 'cd9bcc01b125999b10d3f19e872601f2', 'type': 'people', 'name': '瑞木', 'url': 'http://www.zhihu.com/api/v4/people/cd9bcc01b125999b10d3f19e872601f2', 'gender': 1, 'is_advertiser': False, 'avatar_url': 'https://pic3.zhimg.com/5fa6a8e5aa7207fcbfd917dc8ce50d50_is.jpg', 'is_org': False, 'follower_count': 28, 'badge': []}, {'is_followed': False, 'avatar_url_template': 'https://pic4.zhimg.com/v2-613200caa6f358b464f77943dd2f3e58_{size}.jpg', 'user_type': 'people', 'is_following': False, 'headline': '略。', 'url_token': 'yuan-yi-ming-75', 'id': '7fa634179ecab220235bfb92b7fb3774', 'type': 'people', 'name': '明明就是我', 'url': 'http://www.zhihu.com/api/v4/people/7fa634179ecab

220235bfb92b7fb3774', 'gender': 1, 'is_advertiser': False, 'avatar_url': 'https://pic4.zhimg.com/v2-613200caa6f358b464f77943dd2f3e58_is.jpg', 'is_org': False, 'follower_count': 10, 'badge': []}, {'is_followed': False, 'avatar_url_template': 'https://pic4.zhimg.com/v2-ae335a25be4fd6c5bc3c051bb609e2bf_{size}.jpg', 'user_type': 'people', 'is_following': False, 'headline': '纯属无聊，打发时间', 'url_token': 'abandon-65-76', 'id': 'e98a2e7d53e048dfc8d8a05e60b361fd', 'type': 'people', 'name': 'Abandon', 'url': 'http://www.zhihu.com/api/v4/people/e98a2e7d53e048dfc8d8a05e60b361fd', 'gender': -1, 'is_advertiser': False, 'avatar_url': 'https://pic4.zhimg.com/v2-ae335a25be4fd6c5bc3c051bb609e2bf_is.jpg', 'is_org': False, 'follower_count': 0, 'badge': []}, {'is_followed': False, 'avatar_url_template': 'https://pic3.zhimg.com/v2-050a58566ea8d633236ea4cfddf8fa82_{size}.jpg', 'user_type': 'people', 'is_following': False, 'headline': '这个人很懒什么都没留下', 'url_token': '10-07-zhihu-er', 'id': '2652bced43a18f45849f4f8ea760ef05', 'type': 'people', 'name': '10点睡7点起', 'url': 'http://www.zhihu.com/api/v4/people/2652bced43a18f45849f4f8ea760ef05', 'gender': 1, 'is_advertiser': False, 'avatar_url': 'https://pic3.zhimg.com/v2-

050a58566ea8d633236ea4cfddf8fa82_is.jpg',
'is_org': False, 'follower_count': 15, 'badge':
[]}, {'is_followed': False,
'avatar_url_template':
'https://pic4.zhimg.com/da8e974dc_{size}.jpg',
'user_type': 'people', 'is_following': False,
'headline': '', 'url_token': 'zjsroom', 'id':
'50b846d7a5988ac7bbc74b2a4f651c7f', 'type':
'people', 'name': 'zjsroom', 'url':
'http://www.zhihu.com/api/v4/people/50b846d7a5988
ac7bbc74b2a4f651c7f', 'gender': -1,
'is_advertiser': False, 'avatar_url':
'https://pic4.zhimg.com/da8e974dc_is.jpg',
'is_org': False, 'follower_count': 0, 'badge':
[]}, {'is_followed': False,
'avatar_url_template':
'https://pic4.zhimg.com/v2-
ce57364925dc28389b7daed195abe1b6_{size}.jpg',
'user_type': 'people', 'is_following': False,
'headline': '程序猿', 'url_token': 'yi-di-jin-
guang', 'id': 'e7d124798d91d28e983330ec2b72d311',
'type': 'people', 'name': '一地金光', 'url':
'http://www.zhihu.com/api/v4/people/e7d124798d91d
28e983330ec2b72d311', 'gender': -1,
'is_advertiser': False, 'avatar_url':
'https://pic4.zhimg.com/v2-
ce57364925dc28389b7daed195abe1b6_is.jpg',
'is_org': False, 'follower_count': 0, 'badge':
[]}]}
{'is_followed': False, 'avatar_url_template':
'https://pic4.zhimg.com/v2-
71a02b83135bf403099b5d3f6a260c5f_{size}.jpg',
'user_type': 'people', 'is_following': False,

'headline': '欢迎关注公众号：「乔尔事务所」 → 城市 ｜ 地理 ｜ 规划 ｜ 经济', 'url_token': 'joe1324', 'id': '130894ff204e72b96719a09b5a6565bc', 'type': 'people', 'name': '乔尔爱冒险', 'url': 'http://www.zhihu.com/api/v4/people/130894ff204e72b96719a09b5a6565bc', 'gender': 1, 'is_advertiser': False, 'avatar_url': 'https://pic4.zhimg.com/v2-71a02b83135bf403099b5d3f6a260c5f_is.jpg', 'is_org': False, 'follower_count': 4, 'badge': []}
{'is_followed': False, 'avatar_url_template': 'https://pic3.zhimg.com/v2-339c1f09cd793f43dbccae18af9bfbec_{size}.jpg', 'user_type': 'people', 'is_following': False, 'headline': '', 'url_token': 'hei-tai-yang-91', 'id': 'bbeba0903e14a32d19b27ffa0016ac55', 'type': 'people', 'name': 'JOJO', 'url': 'http://www.zhihu.com/api/v4/people/bbeba0903e14a32d19b27ffa0016ac55', 'gender': -1, 'is_advertiser': False, 'avatar_url': 'https://pic3.zhimg.com/v2-339c1f09cd793f43dbccae18af9bfbec_is.jpg', 'is_org': False, 'follower_count': 4, 'badge': []}
{'is_followed': False, 'avatar_url_template': 'https://pic3.zhimg.com/v2-b6d02ae5a289dc93151c547fc3ae2837_{size}.jpg', 'user_type': 'people', 'is_following': False, 'headline': '2333333333', 'url_token': 'wo-zai-xue-xi-ni-87', 'id': '696dcf2ff66b23ae0ff39474f9f48c5e', 'type': 'people', 'name': '我在学习呢', 'url':

```
'http://www.zhihu.com/api/v4/people/696dcf2ff66b2
3ae0ff39474f9f48c5e', 'gender': -1,
'is_advertiser': False, 'avatar_url':
'https://pic3.zhimg.com/v2-
b6d02ae5a289dc93151c547fc3ae2837_is.jpg',
'is_org': False, 'follower_count': 0, 'badge':
[]}
...
}

Process finished with exit code 0
```

作业

上面的示例只抓到了20条粉丝数据，如何将该专栏的所有粉丝爬下来并保存到文件中。

小结

这就是一个最简单的基于 Requests 的单线程知乎专栏粉丝列表的爬虫，requests 非常灵活，请求头、请求参数、Cookie 信息都可以直接指定在请求方法中，返回值 response 如果是 json 格式可以直接调用json()方法返回 python 对象。关于 Requests 的更多使用方法可以参考官方文档：http://docs.python-requests.org/en/master/ (http://docs.python-requests.org/en/master/)