本节知识点概述

爬取知乎关注者信息

爬虫的一般思路

抓取网页

解析网页

储存数据

实战环节

分析知乎

动手写爬虫

爬虫的一般思路

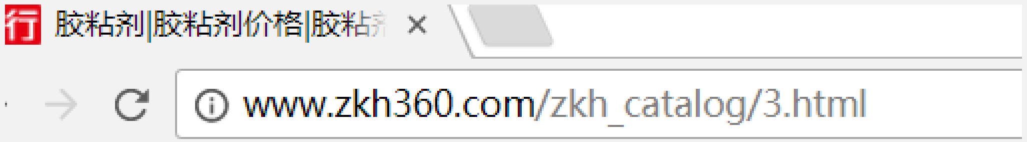
1 抓取网页、分析请求。

2 解析网页、寻找数据。

3 储存数据、多页处理。

分析具体网页

http://www.zkh360.com/zkh_catalog/3.html



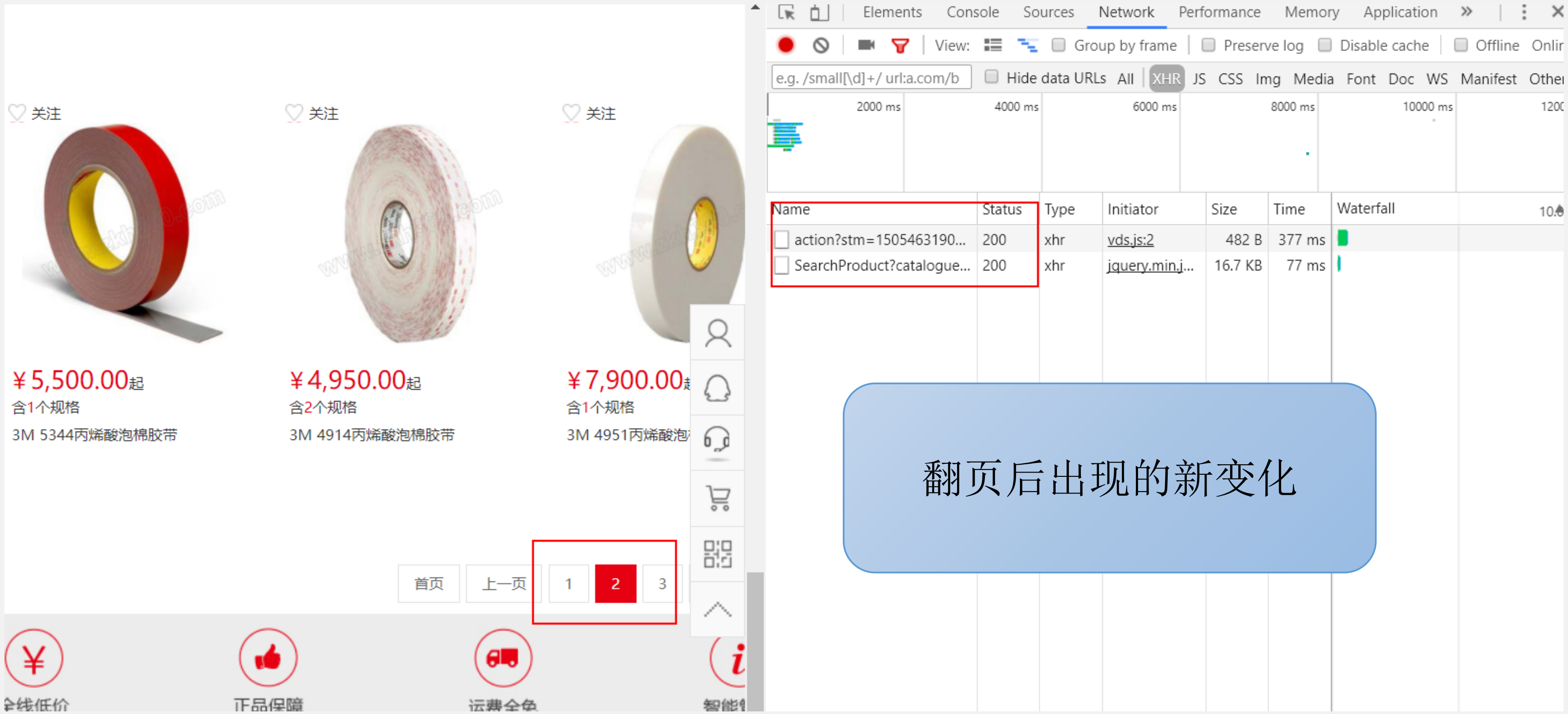
胶粘剂|胶粘剂价格|胶粘剂

www.zkh360.com/zkh_catalog/3.html

翻页后URL不变
如何寻找请求？

哪些网站也是这样？

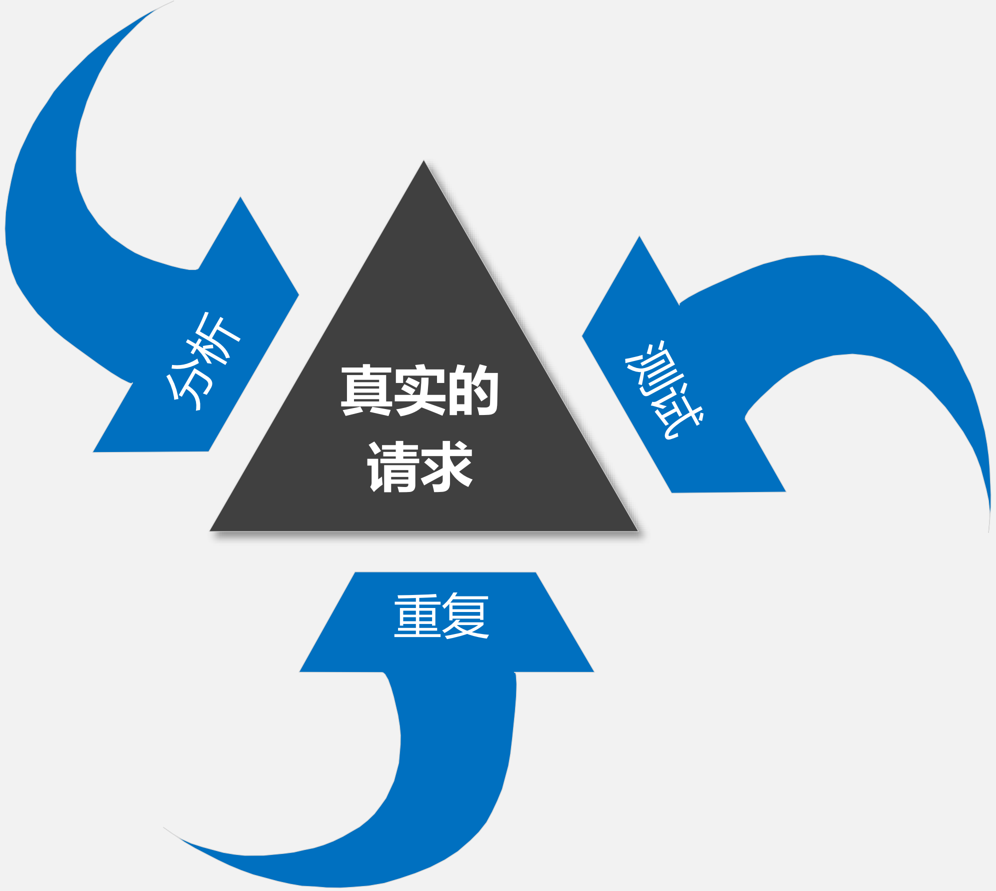
使用Chrome



Name	Status	Type	Initiator	Size	Time	Waterfall
action?stm=1505463190...	200	xhr	vds.js:2	482 B	377 ms	
SearchProduct?catalogue...	200	xhr	jquery.min.j...	16.7 KB	77 ms	

翻页后出现的新变化

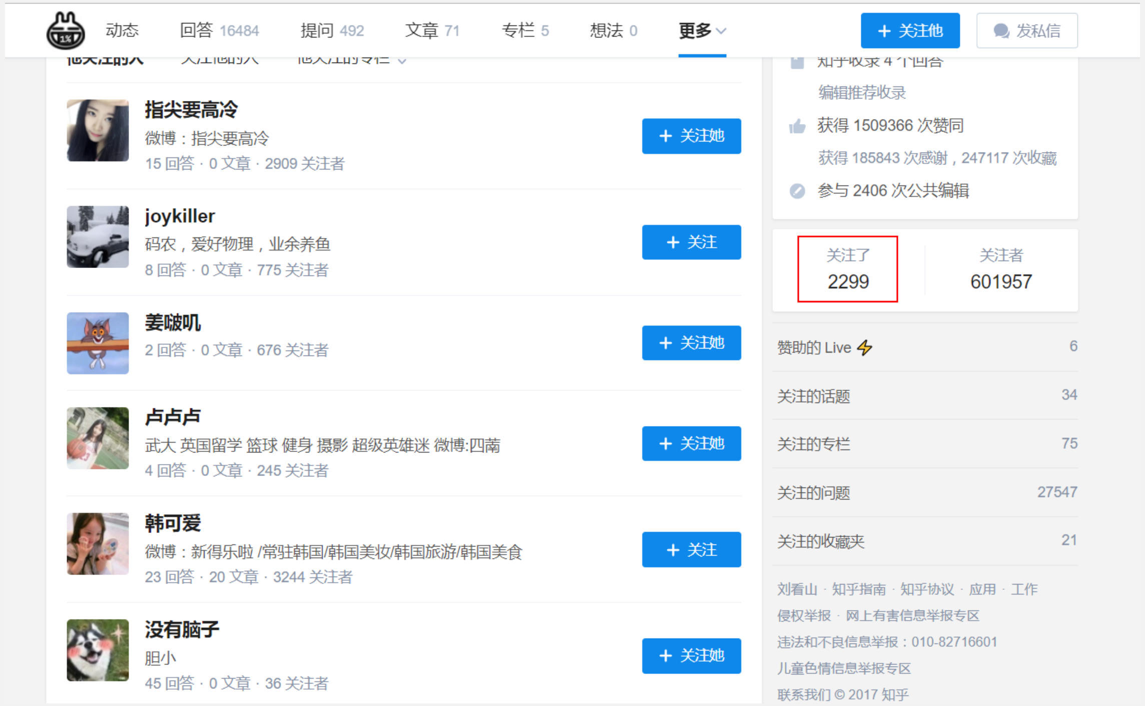
如何寻找真实的请求



尝试寻找
淘宝的请求

实战环节

https://www.zhihu.com/people/excited-vczh/following



指尖要高冷
微博：指尖要高冷
15 回答 · 0 文章 · 2909 关注者
＋ 关注她

joykiller
码农，爱好物理，业余养鱼
8 回答 · 0 文章 · 775 关注者
＋ 关注

姜啵叽
2 回答 · 0 文章 · 676 关注者
＋ 关注她

卢卢卢
武大 英国留学 篮球 健身 摄影 超级英雄迷 微博:四菡
4 回答 · 0 文章 · 245 关注者
＋ 关注她

韩可爱
微博：新得乐啦 /常驻韩国/韩国美妆/韩国旅游/韩国美食
23 回答 · 20 文章 · 3244 关注者
＋ 关注

没有脑子
胆小
45 回答 · 0 文章 · 36 关注者
＋ 关注她

知乎收录 4 个回答
编辑推荐收录
获得 1509366 次赞同
获得 185843 次感谢，247117 次收藏
参与 2406 次公共编辑

关注了	关注者
2299	601957

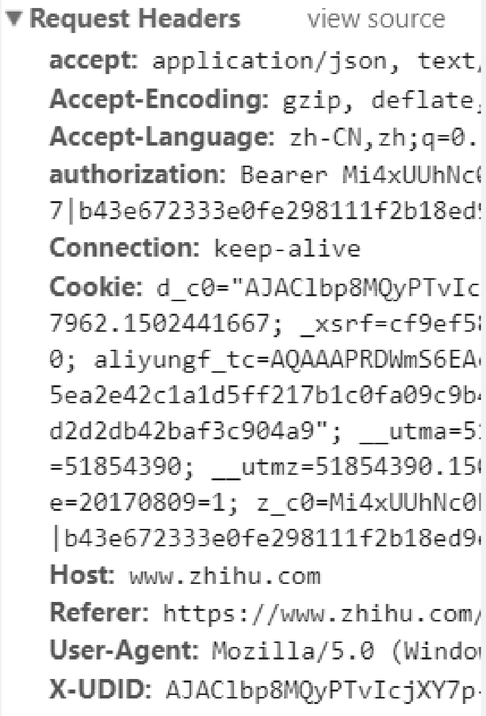
赞助的 Live ⚡	6
关注的话题	34
关注的专栏	75
关注的问题	27547
关注的收藏夹	21

刘看山 · 知乎指南 · 知乎协议 · 应用 · 工作
侵权举报 · 网上有害信息举报专区
违法和不良信息举报：010-82716601
儿童色情信息举报专区
联系我们 © 2017 知乎

寻找请求

动手写爬虫

知乎请求头信息



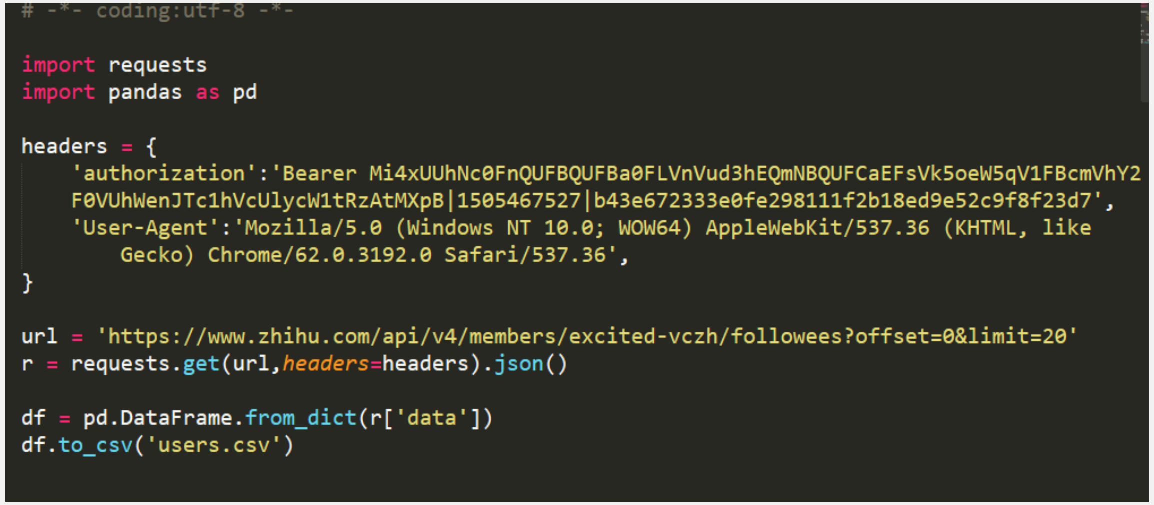
```
▼ Request Headers      view source
    accept: application/json, text,
    Accept-Encoding: gzip, deflate,
    Accept-Language: zh-CN,zh;q=0.
    authorization: Bearer Mi4xUUhNc0
    7|b43e672333e0fe298111f2b18ed
    Connection: keep-alive
    Cookie: d_c0="AJAClbp8MQyPTvIc
    7962.1502441667; _xsrf=cf9ef5
    0; aliyungf_tc=AQAAAPRDWmS6EA
    5ea2e42c1a1d5ff217b1c0fa09c9b
    d2d2db42baf3c904a9"; __utma=5
    =51854390; __utmz=51854390.15
    e=20170809=1; z_c0=Mi4xUUhNc0
    |b43e672333e0fe298111f2b18ed9
    Host: www.zhihu.com
    Referer: https://www.zhihu.com/
    User-Agent: Mozilla/5.0 (Windo
    X-UDID: AJAClbp8MQyPTvIcjXY7p-
```

找不同

常见的请求头

```
Accept:
Accept-Encoding:
Accept-Language:
Connection:
Cookie:
Host:
Referer:
User-Agent:
```

保存单网页信息



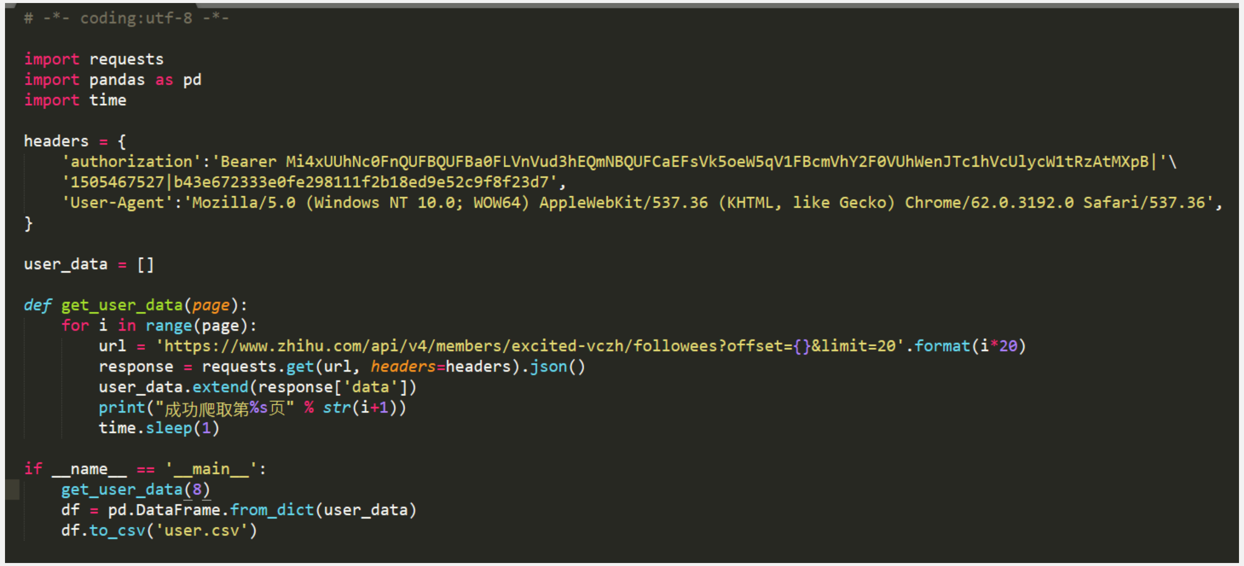
```
# -*- coding:utf-8 -*-

import requests
import pandas as pd

headers = {
    'authorization':'Bearer Mi4xUUhNc0FnQUFBQUFBa0FLVnVud3hEQmNBQUFCaEFsVk5oeW5qV1FBcmVhY2
    F0VUhWenJTc1hVcUlycW1tRzAtMXpB|1505467527|b43e672333e0fe298111f2b18ed9e52c9f8f23d7',
    'User-Agent':'Mozilla/5.0 (Windows NT 10.0; WOW64) AppleWebKit/537.36 (KHTML, like
        Gecko) Chrome/62.0.3192.0 Safari/537.36',
}

url = 'https://www.zhihu.com/api/v4/members/excited-vczh/followees?offset=0&limit=20'
r = requests.get(url,headers=headers).json()

df = pd.DataFrame.from_dict(r['data'])
df.to_csv('users.csv')
```

设置请求头

获得数据

保存数据

保存多页



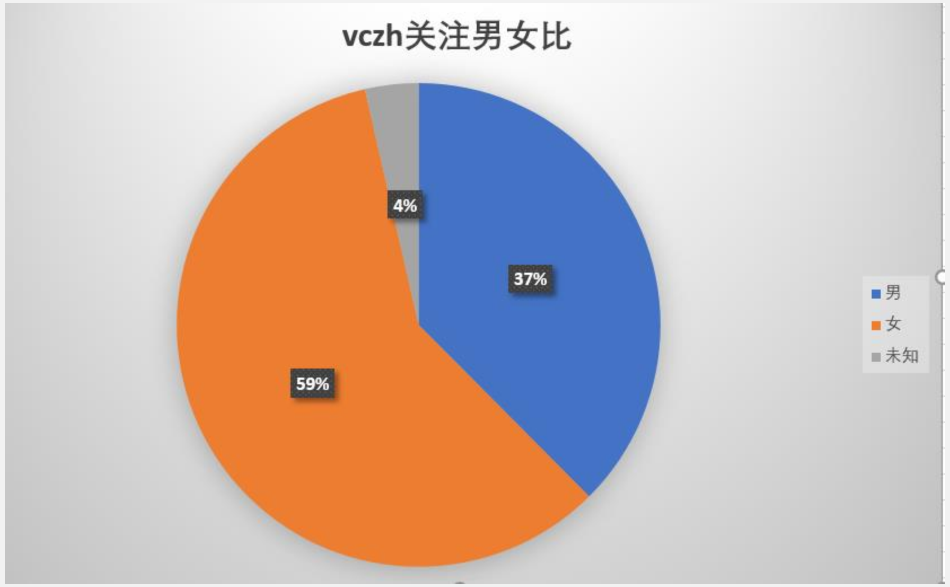
```
# -*- coding:utf-8 -*-

import requests
import pandas as pd
import time

headers = {
    'authorization':'Bearer Mi4xUUhNc0FnQUFBQUFBa0FLVnVud3hEQmNBQUFCaEFsVk5oeW5qV1FBcmVhY2F0VUhWenJTc1hVcUlycW1tRzAtMXpB|'\
    '1505467527|b43e672333e0fe298111f2b18ed9e52c9f8f23d7',
    'User-Agent':'Mozilla/5.0 (Windows NT 10.0; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/62.0.3192.0 Safari/537.36',
}

user_data = []

def get_user_data(page):
    for i in range(page):
        url = 'https://www.zhihu.com/api/v4/members/excited-vczh/followees?offset={}&limit=20'.format(i*20)
        response = requests.get(url, headers=headers).json()
        user_data.extend(response['data'])
        print("成功爬取第%s页" % str(i+1))
        time.sleep(1)

if __name__ == '__main__':
    get_user_data(8)
    df = pd.DataFrame.from_dict(user_data)
    df.to_csv('user.csv')
```

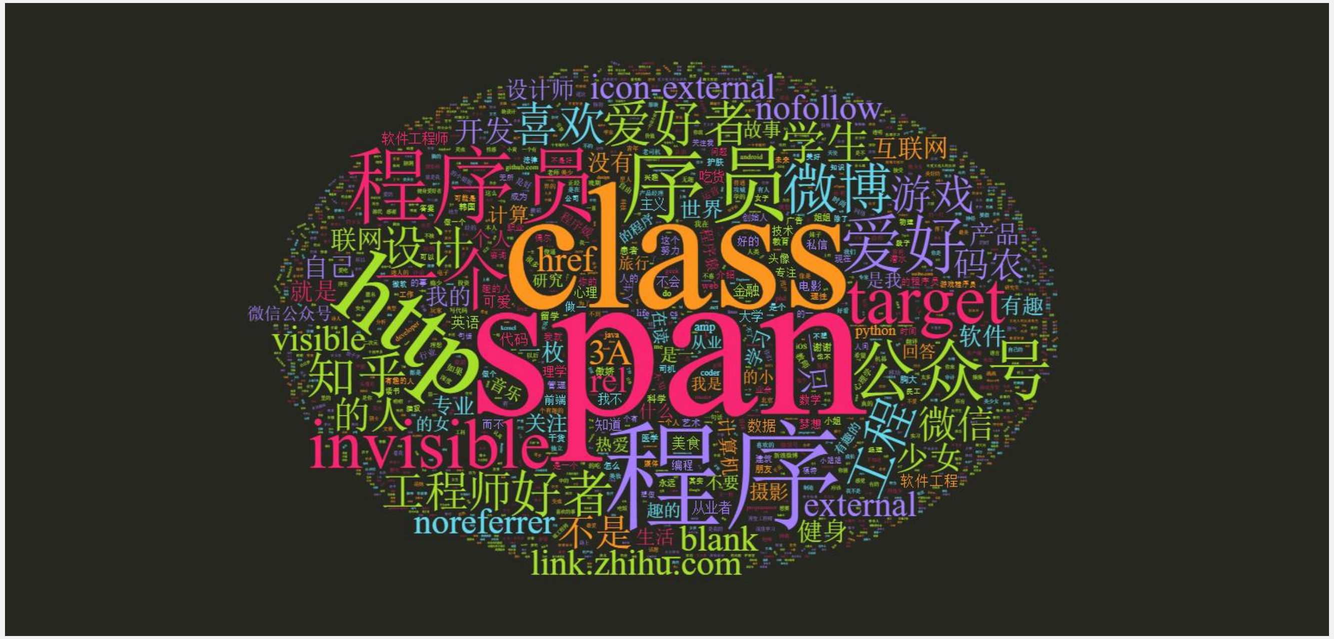
翻页处理

暂停爬虫

调用函数

简单数据可视化


vczh关注男女比
4%
37%
59%
■ 男
■ 女
■ 未知


span
class
http
程序员
程序
invisible
target
公众号
微博
爱好
喜欢爱好者
设计
知乎
工程师好者
工程
noreferrer
link.zhihu.com
blank
external
icon-external
nofollow
visible
href
rel
3A
学生
游戏
码农
微信
少女
互联网
设计师
开发
产品
有趣
软件
就是
自己
联网
微信公众号
的人
专业
关注
摄影
健身
不是
生活
数据
美食
python

完成作业

- 查看淘宝的翻页
- 将vczh改为其他人
- 尝试爬取知乎粉丝信息

DC学院 class.pkbigdata.com | 造数

爬虫工程师

更多数据科学课程，上DC学院： class.pkbigdata.com

关注 DataCastle

关注造数