

# 100行代码爬取公众号所有文章

原创 卡比兽matt 卡比兽matt的笔记 2019-08-09



遇到了难得一遇的宝藏级公众号  
生怕主人哪天抽风删掉文章或被微信block  
今天就带你把文章都爬下来在硬盘慢慢撸

- 领域：网络爬虫
- 关键词：抓包、公众号
- 序号：002
- 日期：2019年8月9日

## /01/ 项目简介

利用requests库爬取公众号文章，再用pdfkit库保存为图文pdf文档

## /02/ 项目思路

1. 通过Fiddler抓包，找出公众号文章链接请求和包含内容。（笔者尝试用ios手机端微信抓包，但Fiddler无法抓取信息，后改用PC端微信）
2. 获取公众号文章url，获取其html后用pdfkit保存为pdf文档即可。

## /03/ 项目过程遇到的问题

**问题1: Fiddler无法对iOS手机端公众号文章抓包**

解决：解决改用PC端微信抓包；

**问题2: pdfkit在获取了文章url后使用pdfkit.from\_url()保存pdf文档，图片缺失**

解决：改成获取html，使用pdfkit.from\_string()保存文档

**问题3: requests获取response时，提示InsecureRequestWarning**

解决：requests是给予urllib编写的，urllib官方强制验证https的安全证书，解决办法是禁用urllib3。即添加代码requests.packages.urllib3.disable\_warnings()

**问题4: 爬取公众号文章时，提示各种错误，keyerror、IOError、internet error...**

解决：直接采用try..except处理，多跑几遍就可以获取到全部历史文章

**问题5: pdfkit保存的文档中，有部分图片分辨率失真**

未解决！原因已经找到：因为公众号中的图片有些是用cmyk格式，如果转为RPG就可以解决！但我不会！大神赐教吧！

**/04/ 项目代码**

先打开Fiddler，再打开你要爬取的公众号，并进入历史文章界面，如图（Fiddler的设置我不再讲啦，翻我上期的“撩妹神技-爬取抖音视频【无水印】把喜欢的小姐姐藏在硬盘里”看看吧）：



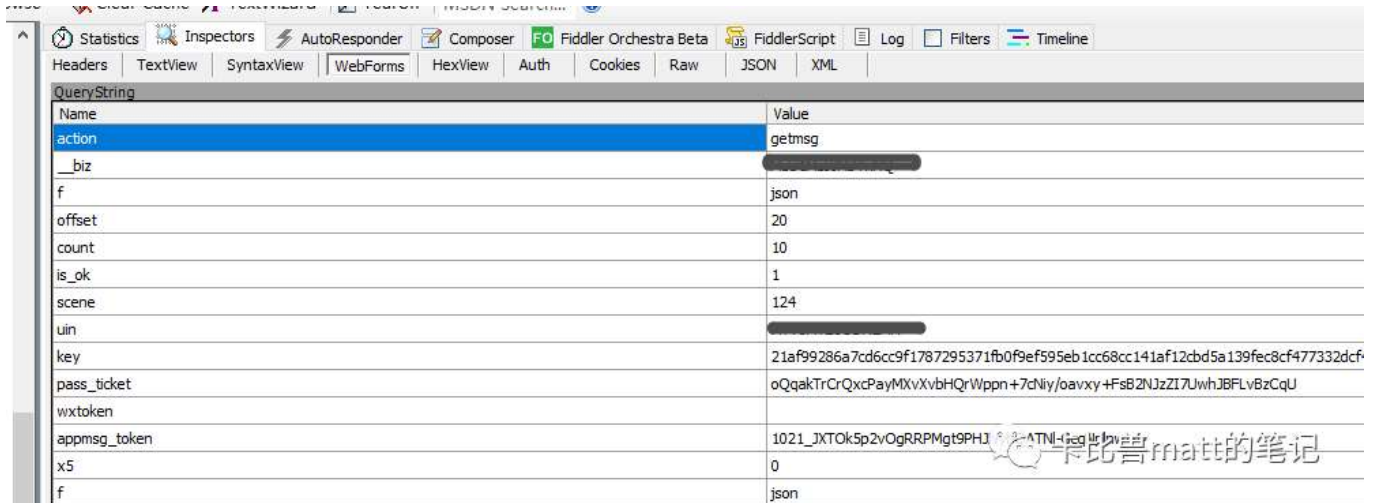
向下滚动页面，loading几页。再回到Fiddler，找到URL是带有/mp/profile\_ext的，就是我们需要获取的内容啦！

#	Result	Protocol	Host	URL	Body	Caching	Content-Type	Process
64	200	HTTP	mmbiz.qpic.cn	/mmbiz_jpg/5689LcmCoXdElpxBLX5zx...	66,577	max-age=250000...	image/jpeg	wechat:12672
65	200	HTTP	mmbiz.qpic.cn	/mmbiz_jpg/5689LcmCoXee1mSgQ6M...	78,670	max-age=250000...	image/jpeg	wechat:12672
12	200	HTTPS	mp.weixin.qq.com	/cgi-bin/spellingcheck?	64	no-cache, must-r...	application/...	chrome:10600
13	200	HTTPS	mp.weixin.qq.com	/mp/jsmonitor?idkey=59475_4_1&t=0...	153	no-cache, must-r...	application/...	chrome:10600
14	200	HTTPS	mp.weixin.qq.com	/mp/jsmonitor?idkey=69271_78_1&t=...	153	no-cache, must-r...	application/...	chrome:10600
16	200	HTTPS	mp.weixin.qq.com	/mp/jsmonitor?idkey=65080_31_1&t=...	153	no-cache, must-r...	application/...	chrome:10600
17	200	HTTPS	mp.weixin.qq.com	/mp/jsmonitor?idkey=65080_118_1	153	no-cache, must-r...	application/...	chrome:10600
18	200	HTTPS	mp.weixin.qq.com	/cgi-bin/filetransfer?action=upload_m...	248	no-cache, must-r...	application/...	chrome:10600
19	200	HTTPS	mp.weixin.qq.com	/mp/jsmonitor?idkey=69271_78_1&t=...	153	no-cache, must-r...	application/...	chrome:10600
21	200	HTTPS	mp.weixin.qq.com	/mp/jsmonitor?idkey=65080_31_1&t=...	153	no-cache, must-r...	application/...	chrome:10600
23	200	HTTPS	mp.weixin.qq.com	/mp/jsmonitor?idkey=69271_78_1&t=...	153	no-cache, must-r...	application/...	chrome:10600
24	200	HTTPS	mp.weixin.qq.com	/mp/jsmonitor?idkey=65080_31_1&t=...	153	no-cache, must-r...	application/...	chrome:10600
27	302	HTTPS	mp.weixin.qq.com	/mp/getmasssendmsg?_biz=MzU1Mz...	0			wechat:12672
28	200	HTTPS	mp.weixin.qq.com	/mp/profile_ext?action=home&_biz=...	10,196		text/html; c...	wechat:12672
30	200	HTTPS	mp.weixin.qq.com	/mp/profile_ext?action=urlcheck&uin=...	36	no-cache, must-r...	application/...	wechat:12672
31	200	HTTPS	mp.weixin.qq.com	/mp/readtemplate?t=pages/video_ad...	3,113		text/html; c...	wechat:12672
44	200	HTTPS	mp.weixin.qq.com	/mp/profile_ext?action=getmsg&_biz=...	2,970	no-cache, must-r...	application/...	wechat:12672
55	200	HTTPS	mp.weixin.qq.com	/mp/profile_ext?action=getmsg&_biz=...	3,024	no-cache, must-r...	application/...	wechat:12672
66	200	HTTPS	mp.weixin.qq.com	/mp/jsmonitor?idkey=59475_4_1&t=0...	153	no-cache, must-r...	application/...	chrome:10600
67	200	HTTPS	mp.weixin.qq.com	/mp/jsmonitor?idkey=59475_4_1&t=0...	153	no-cache, must-r...	application/...	chrome:10600
68	200	HTTPS	mp.weixin.qq.com	/cgi-bin/spellingcheck?	64	no-cache, must-r...	application/...	chrome:10600

再看看Fiddler右边界面，点击inspectors==>Headers会看到User-Agent，Cookies等信息。我们在代码中只需要Headers带User-Agent，Cookies为空即可，其他

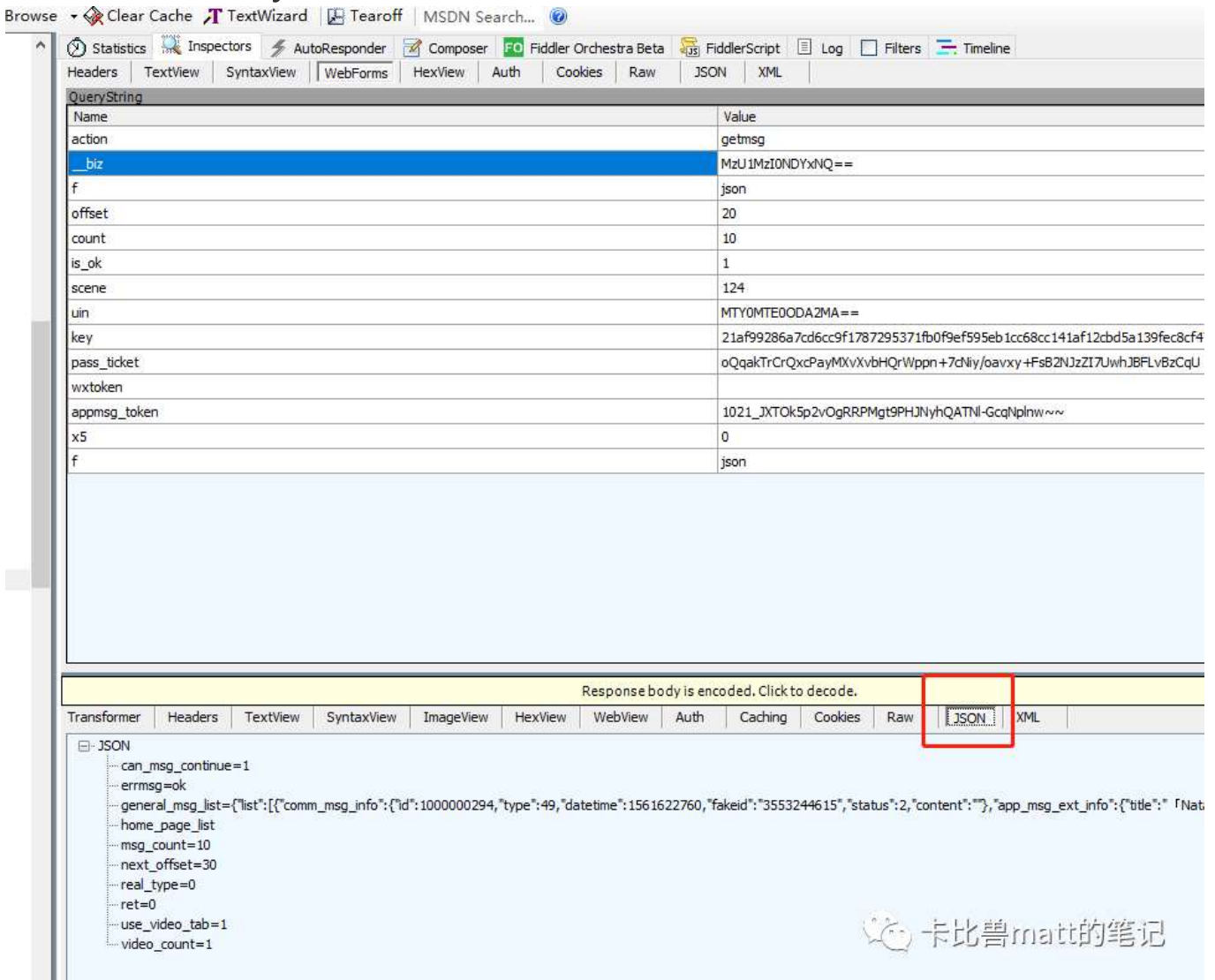
信息都不需要。

我们再点WebForms看看：



\_\_biz是公众号的代码，固定的；f的value是json代表返回json文档；uin表示你的微信号，也是固定的；最重要的是key，大约半小时就失效，所以在爬取代码中，key值都要新鲜。

再点击下方看看json带有什么东西：



can\_msg\_continue=1表示可以继续往下刷新文章列表；general\_msg\_list带有当前刷新页中包含的几篇文章的全部内容，正是我们想要的东东！next\_offset代表下次刷



新的步长。

下面上完整代码：

```

1  #-*- coding:UTF-8 -*-
2  import json
3  import time
4  import pdfkit
5  import os
6  import wechatsogou
7  import re
8  import requests
9
10 base_url = 'https:去掉这里的中文//mp.weixin.qq.com/mp/去掉这里的中文profile_去掉
11
12 # 这些信息不能拿我的，要用你自己的ok
13 headers = {
14     # 'Connection': 'keep-alive',
15     # 'Accept': '*/*',
16     'User-Agent': '复制你fiddler里的',
17     # 'Accept-Encoding': 'gzip, deflate',
18     # 'X-Requested-With': 'XMLHttpRequest'
19     # 'key': '21af99286a7cd6cc292140f938e28243313af8fec95ff68725125bba4a64242l
20 }
21
22 cookies = {
23     # 'devicetype': 'Windows10',
24     # 'lang': 'zh_CN',
25     # 'pass_ticket': 'mf2myiNN1c0GYhagys4+LwBZd0Vj63v0e70ZqUdKxh9esxBoVW8R1ir
26     # 'version': '62060844',
27     # 'wap_sid2': 'CJzdx44GEnA2Q1ZsS2NacmNBZWpwN1dTc3RIUFBURjR1emxua1E1U1Mwcc
28     # 'wxtokenkey': '777',
29     # 'wxuin': '1641148060'
30
31 }
32
33 # 初始化API。captcha_break_time为验证码输入错误的重试次数，默认为1
34 ws_api = wechatsogou.WechatSogouAPI(captcha_break_time=3)
35
36 def get_params(offset):
37     params = {

```

```

38         'action': 'getmsg',
39         '__biz': '这里写你要爬取的公众号的代码',
40         'f': 'json',
41         'offset': '{}'.format(offset),
42         'count': '10',
43         'is_ok': '1',
44         'scene': '126',
45         'uin': '这里写你的微信号代码',
46         'key': '这个key复制fiddler里抓到的',
47         # 'pass_ticket': 'mf2myiNN1c0GYhagys4+lwBZd0Vj63voe70ZqUdKxh9esxBoVWk',
48         # 'appmsg_token': '1021_AKKMhy%2B%2FHmCaW4J47-5sLRtmBiYAjYC4XSiyEg~~',
49         # 'x5': '0',
50     }
51
52     return params
53
54 # 保存文件时，去除名字中的非法字符
55 def validateTitle(title):
56     rstr = r'[\\/:*?"<>|\r\n]+'
57     new_title = re.sub(rstr, "_", title) # 替换为下划线
58     return new_title
59
60 # 保存文件时，加上文章发布日期，方便管理文章
61 def datetime_toString(dt):
62     return time.strftime('%Y-%m-%d', time.localtime(dt))
63
64 def get_list_data(offset):
65     requests.packages.urllib3.disable_warnings() # 禁用urllib
66     res = requests.get(base_url, headers=headers, params=get_params(offset),
67                         data=json.loads(res.text))
68
69     can_msg_continue = data['can_msg_continue']
70     next_offset = data['next_offset']
71
72     general_msg_list = data['general_msg_list']
73     list_data = json.loads(general_msg_list)['list']
74
75     # pdfkit的设置内容
76     pdf_options = {
77         'encoding': "UTF-8",

```

```

78     'quiet': '' # 写上这句保存pdf时就不会显示一大串保存进度过程输出
79 }
80 # config = pdfkit.configuration(wkhtmltopdf='C:/Program Files/wkhtmltopdf;
81 for data in list_data:
82     date = data["comm_msg_info"]["datetime"]
83     try:
84         if data['comm_msg_info']['type']==49 and data['app_msg_ext_info']:
85
86             msg_info = data['app_msg_ext_info']
87             title = str(datetime_toString(date)) + '_' + str(validateTitl
88             content_url = msg_info['content_url']
89             print(content_url)
90             # wechatsogou根据文章url对html进行处理, 使图片显示
91             content_info = ws_api.get_article_content(content_url)
92             # 得到html代码
93             html_code = content_info['content_html']
94             # 自己定义存储路径
95             file = 'wechat_article/{0}.pdf'.format(title)
96             if not os.path.exists(file):
97                 pdfkit.from_string(html_code, file,options=pdf_options)
98                 print('获取到原创文章: %s : %s' % (title, content_url))
99             elif data['comm_msg_info']['type']==1:
100                 content = data['comm_msg_info']['content']
101                 title = str(datetime_toString(date)) + '_' + str(data['comm_r
102                 file = 'wechat_article/{0}.pdf'.format(title)
103                 if not os.path.exists(file):
104                     pdfkit.from_string(content, file, options=pdf_options)
105                     print('获取到原创文段: %s ' % (title))
106
107             except Exception as e:
108                 print(e)
109
110         if can_msg_continue == 1:
111             time.sleep(1)
112             get_list_data(next_offset)
113
114
115 if __name__ == '__main__':
116     get_list_data(0)

```

以上代码在跑的过程，可能会遇到网络问题，没关系，多跑几次应该就能抓全。注意在跑代码时把fiddler关闭，因为它占用了一些端口。

我跑出来的结果：



保存了以上pdf就可以慢慢撸了！！ 感谢各位，拜拜啦您呐！

文章已于2019-08-09修改