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

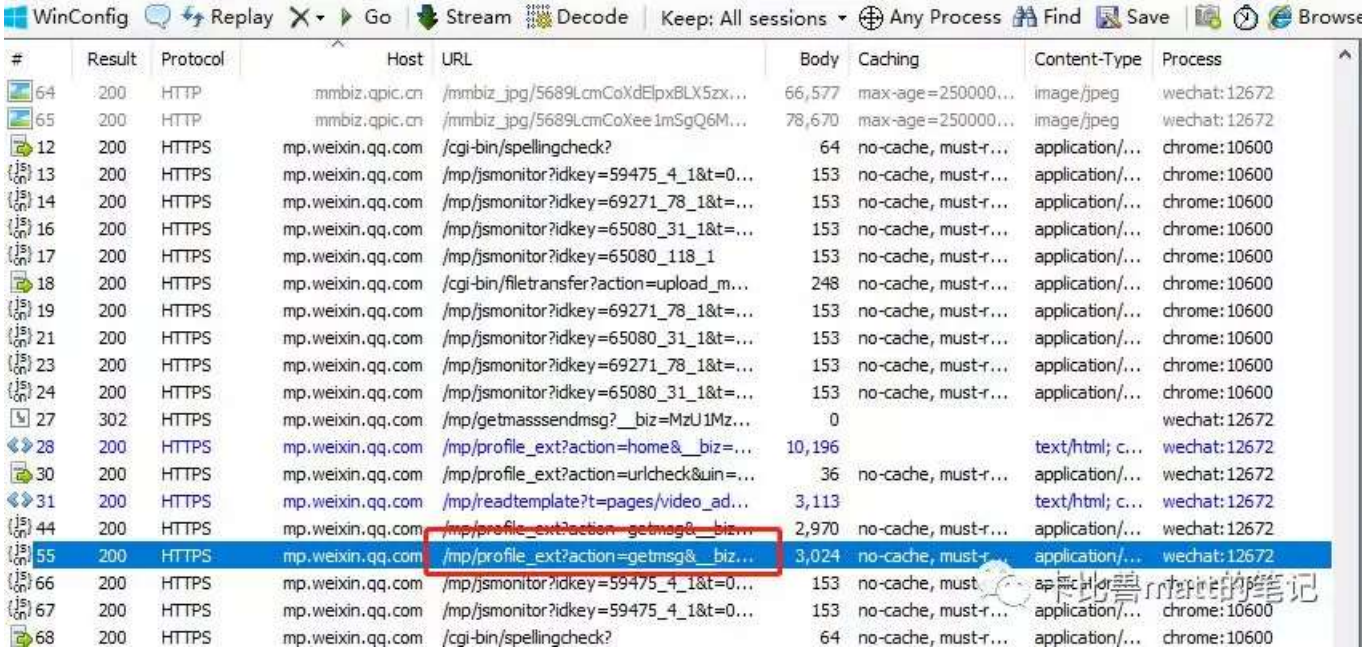
问题5：pdfkit保存的文档中，有部分图片分辨率失真

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）：

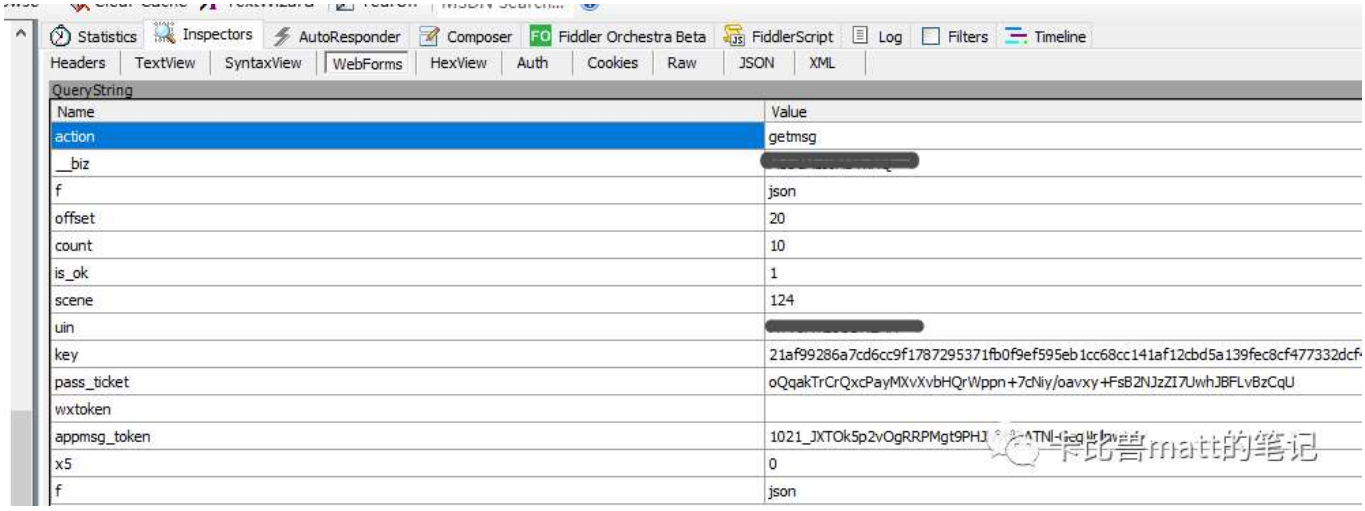
向下滚动页面，loading几页。再回到Fiddler，找到URL是带有/mp/profile_ext的，就是我们需要获取的内容啦！



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Heders带User-Agent，Cookies为空即可，其他

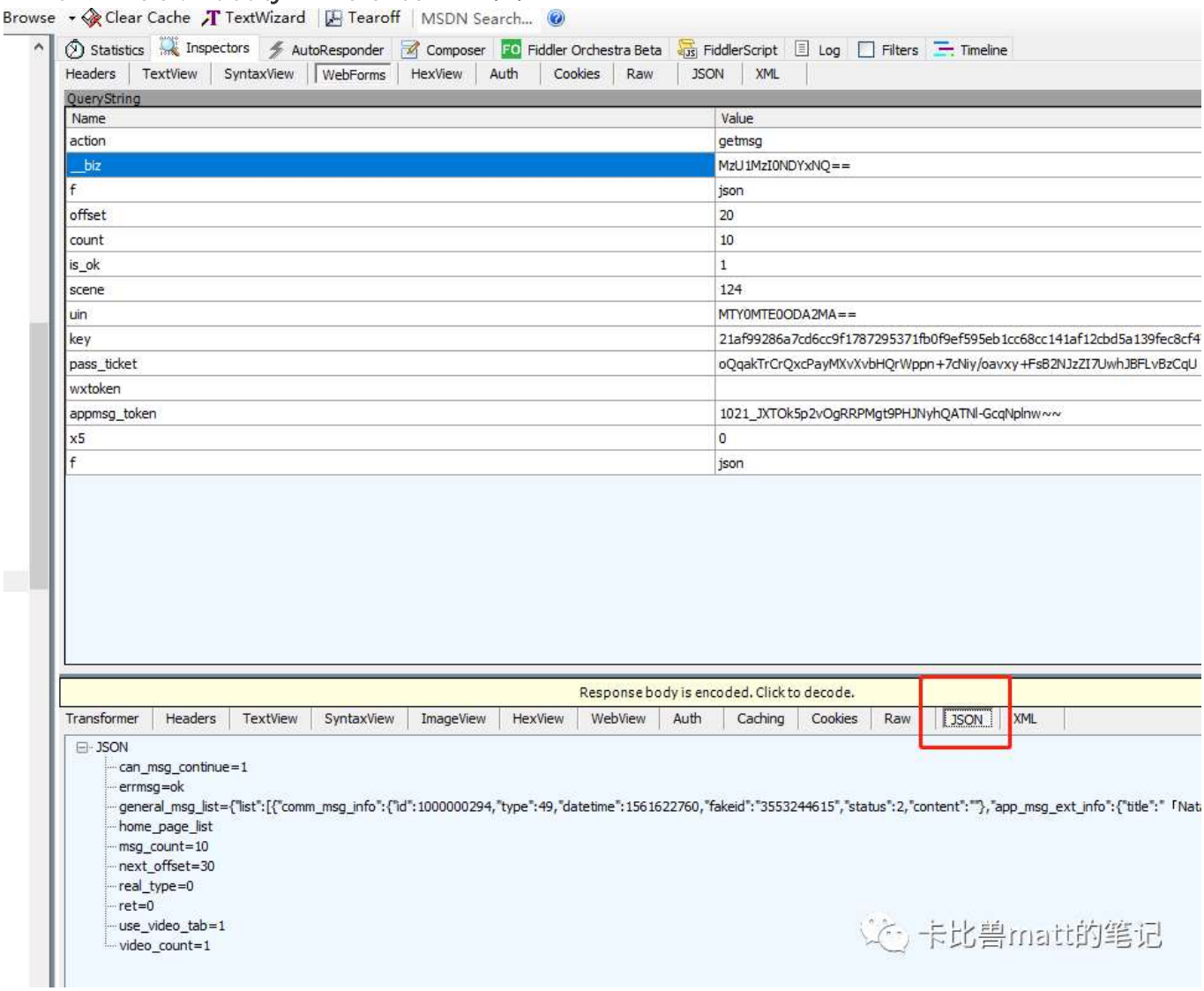
信息都不需要。

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：

```
#-*- coding:UTF-8 -*-
import json
import time
import pdfkit
import os
import wechatsogou
import re
import requests

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

# 这些信息不能拿我的，要用你自己的ok
headers = {
    # 'Connection': 'keep-alive',
    # 'Accept': '*/*',
    'User-Agent': '复制你fiddler里的',
    # 'Accept-Encoding': 'gzip, deflate',
    # 'X-Requested-With': 'XMLHttpRequest'
    # 'key':'21af99286a7cd6cc292140f938e28243313af8fec95ff68725125bba4a64242l
}

cookies = {
    # 'devicetype': 'Windows10',
    # 'lang': 'zh_CN',
    # 'pass_ticket': 'mf2myiNN1c0GYhagys4+lwBZd0Vj63voe7OZqUdKxh9esxBoVW8R1ir
    # 'version': '62060844',
    # 'wap_sid2': 'CJzdx44GEnA2Q1ZsS2NacmNBZWpwN1dTc3RIUFBURjR1emxua1E1U1Mwc(
    # 'wxtokenkey':'777',
    # 'wxuin': '1641148060'

}

# 初始化API。captcha_break_time为验证码输入错误的重试次数，默认为1
ws_api = wechatsogou.WechatSogouAPI(captcha_break_time=3)

def get_params(offset):
    params = {
```

```
        'action': 'getmsg',
        '__biz': '这里写你要爬取的公众号的代码',
        'f': 'json',
        'offset': '{}'.format(offset),
        'count': '10',
        'is_ok': '1',
        'scene': '126',
        'uin': '这里写你的微信号代码',
        'key': '这个key复制fiddler里抓到的',
        # 'pass_ticket': 'mf2myiNN1c0GYhagys4+lwBZd0Vj63voe7OZqUdKxh9esxBoVW8
        # 'appmsg_token': '1021_AKKMhy%2B%2FHmCaW4J47-5slRtmBiYAjYC4XSiyEg~~
        # 'x5': '0',
    }

    return params

# 保存文件时，去除名字中的非法字符
def validateTitle(title):
    rstr = r'[\\/:*?"<>|\r\n]+'
    new_title = re.sub(rstr, "_", title)  # 替换为下划线
    return new_title

# 保存文件时，加上文章发布日期，方便管理文章
def datetime_toString(dt):
    return time.strftime('%Y-%m-%d',time.localtime(dt))

def get_list_data(offset):
    requests.packages.urllib3.disable_warnings() # 禁用urllib
    res = requests.get(base_url, headers=headers, params=get_params(offset),
    data = json.loads(res.text)

    can_msg_continue = data['can_msg_continue']
    next_offset = data['next_offset']

    general_msg_list = data['general_msg_list']
    list_data = json.loads(general_msg_list)['list']

    # pdfkit的设置内容
    pdf_options = {
        'encoding': "UTF-8",
```

```
        '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
    }
    # config = pdfkit.configuration(wkhtmltopdf='C:/Program Files/wkhtmltopdf
    for data in list_data:
        date = data["comm_msg_info"]["datetime"]
        try:
            if data['comm_msg_info']['type']==49 and data['app_msg_ext_info']

                msg_info = data['app_msg_ext_info']
                title = str(datetime_toString(date)) + '_' + str(validateTit
                content_url = msg_info['content_url']
                print(content_url)
                # wechatsogou根据文章url对html进行处理，使图片显示
                content_info = ws_api.get_article_content(content_url)
                # 得到html代码
                html_code = content_info['content_html']
                # 自己定义存储路径
                file = 'wechat_article/{}.pdf'.format(title)
                if not os.path.exists(file):
                    pdfkit.from_string(html_code, file,options=pdf_options)
                    print('获取到原创文章：%s ： %s' % (title, content_url))
            elif data['comm_msg_info']['type']==1:
                content = data['comm_msg_info']['content']
                title = str(datetime_toString(date)) + '_' + str(data['comm_
                file = 'wechat_article/{}.pdf'.format(title)
                if not os.path.exists(file):
                    pdfkit.from_string(content, file, options=pdf_options)
                    print('获取到原创文段：%s ' % (title))

        except Exception as e:
            print(e)

    if can_msg_continue == 1:
        time.sleep(1)
        get_list_data(next_offset)


if __name__ == '__main__':
    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