用 Python 抓取公号文章保存成 PDF

原创 極光 Python技术 1周前

今天为大家介绍如何将自己喜欢的公众号的历史文章转成 PDF 保存到本地。前几天还有朋友再问，能不能帮把某某公众号的文章下载下来，因为他很喜欢这个号的文章，但由于微信上查看历史文章不能排序，一些较早期的文章翻很长时间才能找到，而且往往没有一次看不了几篇，下次还得再重头翻，想想就很痛苦。

抓取的思路

目前我在网上找了找，看到实现的方式大概分为以下三种：

1. 通过手机和电脑相连，利用 Fiddler 抓包获取请求和返回报文，然后通过报文模拟请求实现批量下载。
2. 通过搜狗浏览器或者用 `wechatsogou` 这个 Python 模块，去搜索公号后，实现批量下载。
3. 通过公众号平台，这个需要你能登陆到公众号平台即可，剩下就比较简单。

整体来看最后一种方式是最简单的，接下来将以第三种方式为例，为大家介绍如何达到批量下载的目的。

获取 Cookie

首先我们登陆到公众号平台，登陆成功后会跳转到公众号管理首页，如下图：

然后我们在当前页面打开浏览器开发者工具，刷新下页面，在网络里就能看到各种请求，在这里我们点开一个请求 url，然后就能看到下图网络请求信息，里面包含请求的 Cookie 信息。

接下来我们需要把 Cookie 信息复制下来转换成 Json 格式串保存到文本文件里，以供后面请求链接时使用。这里需要写一段 Python 代码进行处理，新建文件 `gen_cookies.py` 写入代码如下：

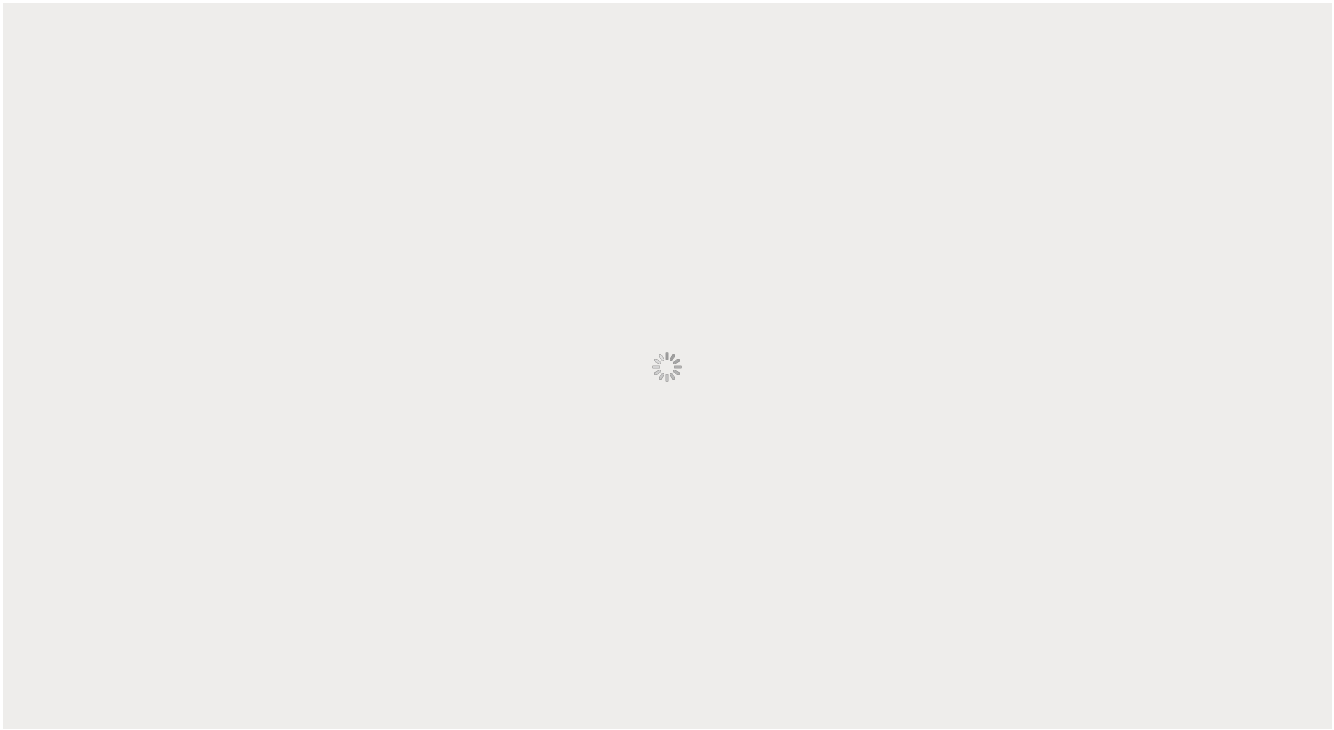
```
# gen_cookies.py

import json

# 从浏览器中复制出来的 Cookie 字符串
cookie_str = "pgv_pvid=9551991123; pac_uid=89sdjfklas; XWINDEXGREY=0; pgv_pvi=89273492834; tvfe_boss
……中间部分省略 \
 EXIV96Zg=sNOaZlBxE37T1tqbsOL/qzHBtiHUNZSxr6TMqpb8Z9k="

cookie = {}
# 遍历 cookie 信息
for cookies in cookie_str.split("; "):
    cookie_item = cookies.split("=")
    cookie[cookie_item[0]] = cookie_item[1]
# 将cookies写入到本地文件
with open('cookie.txt', "w") as file:
    #  写入文件
    file.write(json.dumps(cookie))

```

好了，将 Cookie 写入文件后，接下来就来说下在哪里可以找到某公号的文章链接。

获取文章链接

在公号管理平台首页点击左侧素材管理菜单，进入素材管理页面，然后点击右侧的新建图文素材按钮，如下图：

进入新建图文素材页面，然后点击这里的超链接：



在编辑超链接的弹出框里，点击选择其他公众号的连接：

在这里我们就能通过搜索，输入关键字搜索我们想要找到公众号，比如在这里我们搜索 "Python 技术"，就能看到如下搜索结果：

然后点击第一个 Python 技术的公众号，在这里我们就能看到这个公众号历史发布过的所有文章：

我们看到这里文章每页只显示五篇，一共分了31页，现在我们再打开自带的开发者工具，然后在列表下面点下一页的按钮，在网络中会看到向服务发送了一个请求，我们分析下这个请求的参数。

通过请求参数，我们大概可以分析出参数的意义，`begin` 是从第几篇文章开始，`count` 是一次查出几篇，`fakeId` 对应这个公号的唯一 Id，`token` 是通过 Cookie 信息来获取的。好了，知道这些我们就可以用 Python 写段代码去遍历请求，新建文件 `gzh_download.py`,代码如下：

```
# gzh_download.py
# 引入模块
import requests
import json
import re
import random
import time
```

```
import pdfkit

# 打开 cookie.txt
with open("cookie.txt", "r") as file:
    cookie = file.read()
cookies = json.loads(cookie)
url = "https://mp.weixin.qq.com"
#请求公号平台
response = requests.get(url, cookies=cookies)
# 从url中获取token
token = re.findall(r'token=(\d+)', str(response.url))[0]
# 设置请求访问头信息
headers = {
    "Referer": "https://mp.weixin.qq.com/cgi-bin/appmsg?t=media/appmsg_edit_v2&action=edit&isNew=1&t
    "Host": "mp.weixin.qq.com",
    "User-Agent": "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_12_6) AppleWebKit/537.36 (KHTML, like G
}

# 循环遍历前10页的文章
for j in range(1, 10, 1):
    begin = (j-1)*5
    # 请求当前页获取文章列表
    requestUrl = "https://mp.weixin.qq.com/cgi-bin/appmsg?action=list_ex&begin="+str(begin)+"&count=
    search_response = requests.get(requestUrl, cookies=cookies, headers=headers)
    # 获取到返回列表 Json 信息
    re_text = search_response.json()
    list = re_text.get("app_msg_list")
    # 遍历当前页的文章列表
    for i in list:
        # 将文章链接转换 pdf 下载到当前目录
        pdfkit.from_url(i["link"], i["title"] + ".pdf")
    # 过快请求可能会被微信问候，这里进行10秒等待
    time.sleep(10)


```

好了，就上面这点代码就够了，这里在将 URL 转成 PDF 时使用的是 `pdfkit` 的模块，使用这个需要先安装 `wkhtmltopdf` 这个工具，官网地址在文末给出，支持多操作系统，自己下载安装即可，这里就不再赘述。

安装完后，还需要再执行 `pip3 install pdfkit` 命令安装这个模块。安装好了，现在来执行下 `python gzh_download.py` 命令启动程序看下效果怎么样。

看来是成功了，这个工具还是很强大的。

总结

本文为大家介绍了如何通过分析公众号平台的功能，找到可以访问到某个公众号所有文章的链接，从而可以批量下载某公众号所有文章，并转为 PDF 格式保存到本地的目的。这里通过 Python 写了少量代码就实现文章的抓取和转换的工作，如果有兴趣你也可以试试。

参考

https://wkhtmltopdf.org/downloads.html