

信息内容安全实验报告

实验项目名称： 爬虫--爬取某微信公众号的历史原创文章

班级： SC011701

学号： 2017302209

姓名： 胡梦莹

指导教师： 杨黎斌

实验时间： 2020.3.18

一、实验目的与要求

1. 掌握爬虫的基本原理；
2. 通过 Fiddler 抓取手机访问某公众号的流量包，并对其进行分析，最后爬取公众号的原创文章。

二、实验设计思路

1. 用户请求某公众号（Python 爱好者社区）的历史文章；
2. 微信服务器返回数据，Fiddler 进行数据拦截，返回规则；
3. 通过编写 python 脚本模拟请求，抓取数据，并将文章标题以及网址存入 article.txt；
4. 用户得到 txt 文件。

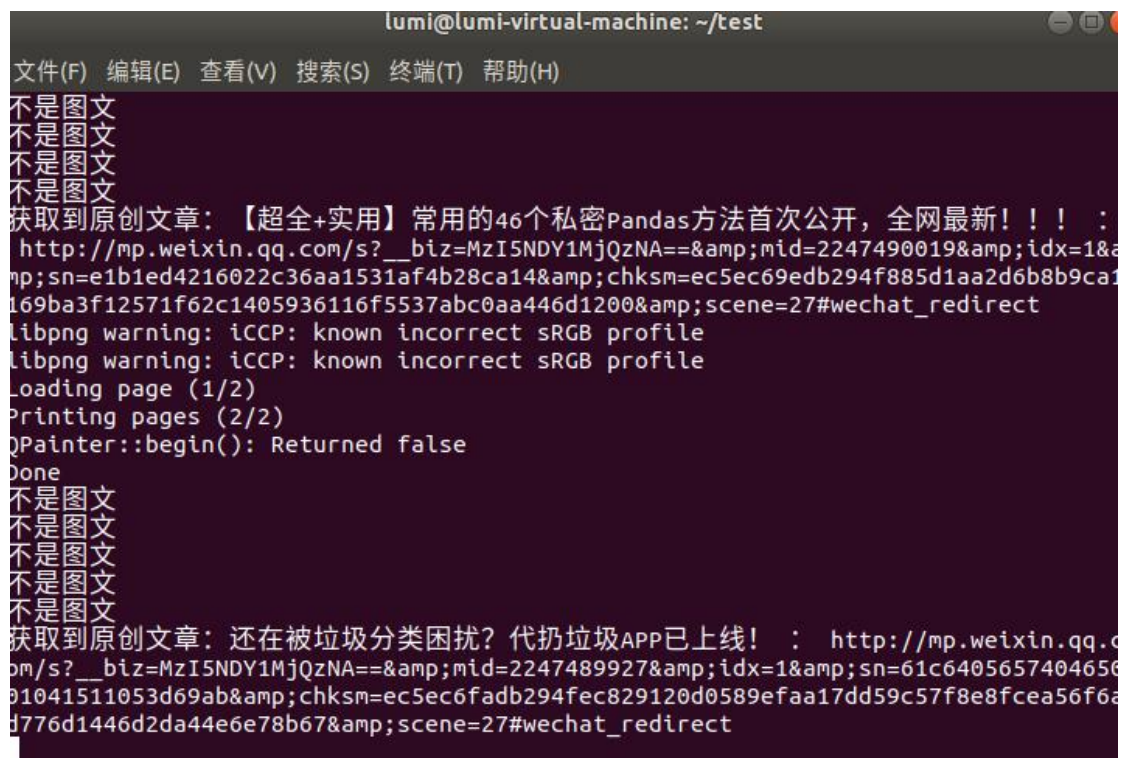
三、实验具体过程

1. 手机和电脑连入同一 WIFI, 打开 Fiddler, 通过手机访问微信某公众号的历史文章；
2. 在 Fiddler 软件中，可以发现哪一接口在请求数据，点开某一具体数据请求，注意到，是否可以加载更多历史文章是由 offset 和 is_ok 两个参数决定的；而服务器端返回的 json 数据的两个参数 next_offset 和 is_continue_ok 控制着 offset 和 is_ok；同时，json 数据的 list 列表中则有历史文章的具体信息，我们期望获取的则只有文章标题以及文章链接两个数据；
3. 区分文章是否原创。通过访问多个公众号可以发现，服务器返回的 json 数据中 copyright_stat=11 代表着文章是原创的。

4. 编写 python 脚本，可以在 Fiddler 中获取 headers 和 cookies 信息，每次发出请求时，只需判断服务器端返回的 is_continue_ok 是否为 1，若为 1，则将上一次的 offset 改变为服务器端返回的 next_offset 即可，并将结果保存至 article.txt 中。

四、实验结果

在 ubuntu18.04, PYTHON3.6.9 中运行结果如下：



```
lumi@lumi-virtual-machine: ~/test
文件(F) 编辑(E) 查看(V) 搜索(S) 终端(T) 帮助(H)
不是图文
不是图文
不是图文
不是图文
获取到原创文章：【超全+实用】常用的46个私密Pandas方法首次公开，全网最新!!! :
http://mp.weixin.qq.com/s?__biz=MzI5NDY1MjQzNA==&mid=2247490019&idx=1&
mp;sn=e1b1ed4216022c36aa1531af4b28ca14&chksm=ec5ec69edb294f885d1aa2d6b8b9ca1
169ba3f12571f62c1405936116f5537abc0aa446d1200&scene=27#wechat_redirect
libpng warning: iCCP: known incorrect sRGB profile
libpng warning: iCCP: known incorrect sRGB profile
Loading page (1/2)
Printing pages (2/2)
QPainter::begin(): Returned false
Done
不是图文
不是图文
不是图文
不是图文
不是图文
获取到原创文章：还在被垃圾分类困扰？代扔垃圾APP已上线！ : http://mp.weixin.qq.c
om/s?__biz=MzI5NDY1MjQzNA==&mid=2247489927&idx=1&sn=61c6405657404656
01041511053d69ab&chksm=ec5ec6fadb294fec829120d0589efaa17dd59c57f8e8fcea56f6a
d776d1446d2da44e6e78b67&scene=27#wechat_redirect
```

article.txt 文件如下：



```
文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)
文章标题：真实的北京IT圈：后厂村姑 vs 后厂村花？ 网址:http://mp.weixin.qq.com/s?__biz=MzI5NDY1MjQzNA==&mid=2
文章标题：真实的上海IT圈：张江男vs漕河泾男 网址:http://mp.weixin.qq.com/s?__biz=MzI5NDY1MjQzNA==&mid=224749
文章标题：为什么你的提问没人解答？ 网址:http://mp.weixin.qq.com/s?__biz=MzI5NDY1MjQzNA==&mid=2247490282&
文章标题：【超全+实用】常用的46个私密Pandas方法首次公开，全网最新!!! 网址:http://mp.weixin.qq.com/s?__biz=MzI5ND
文章标题：还在被垃圾分类困扰？代扔垃圾APP已上线！ 网址:http://mp.weixin.qq.com/s?__biz=MzI5NDY1MjQzNA==&mid=mic
```

五、实验中遇到的问题以及解决办法

1. 问题：编写完脚本后，由于多次运行该脚本，发出请求次数太多，导致 ip 被封，不能访问原网址。如图：



2. 解决方法：使用代理 IP，在相关网站获得免费代理 IP。最后使用代理 IP 访问网站。
3. 改进之处：但是，代理 IP 很容易失效。因此，可以爬取相关网站的大量代理 IP，创建一个 IP 代理池，可以随着时间的推移更新该 IP 代理池，同时检测各 IP 的可用性。