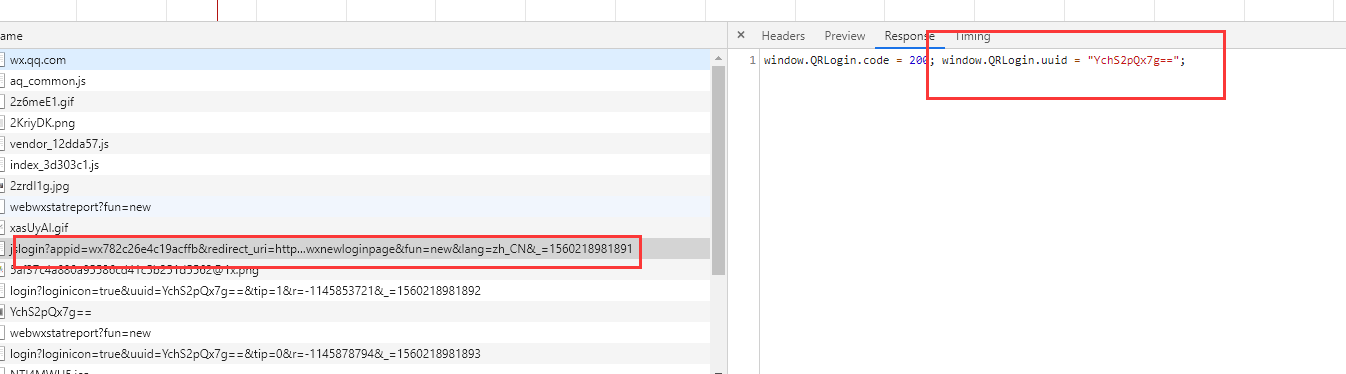
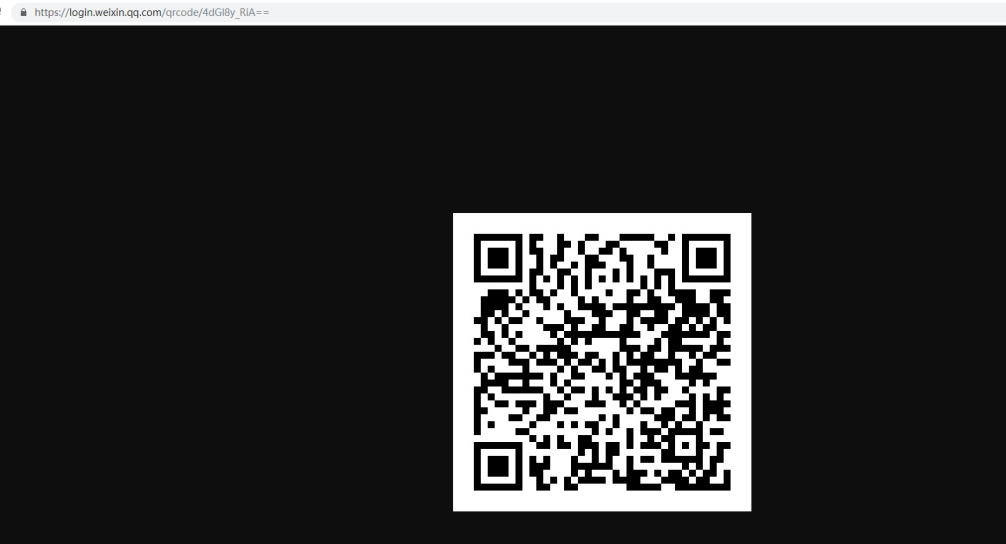
通过分析网页版微信登陆首先获取二维码，发现网页中包含了一个新的url，而这个url就是二维码的来源。https://login.weixin.qq.com/qrcode/YchS2pQx7g==从而发现获取二维码的url就是把最后一部分字符串给去掉，YchS2pQx7g==肯定就是通过请求前面的url后传入的uuid，最终生成二维码图片。

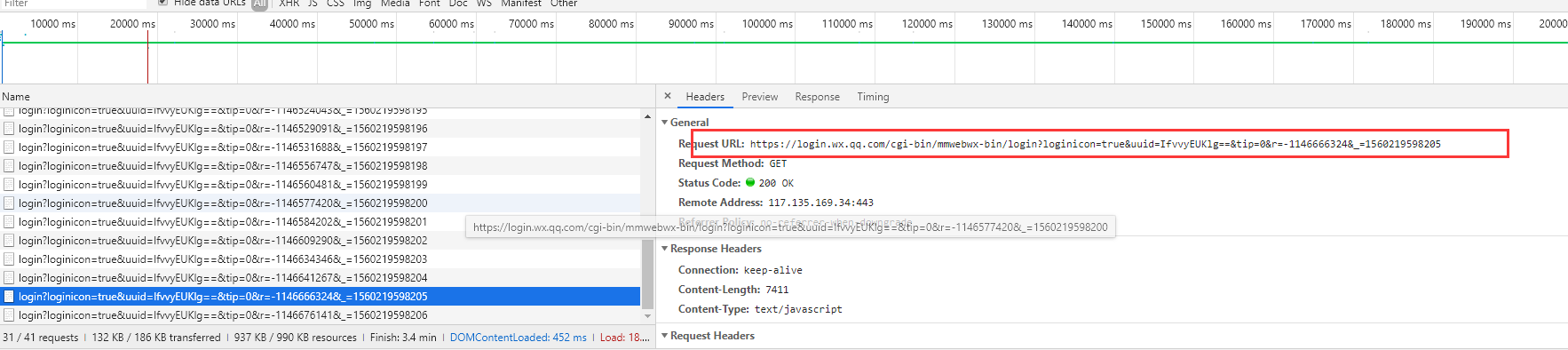


再通过network分析发现这个链接是用来返回uuid的，然后把参数传进去请求这个url并把返回的code跟uuid拿出来





接下来取请求用uuid拼起来的url，并把图片爬取下来用uuid命名存到专门放二维码的文件夹里，并用PIL包里的image打开这个二维码图片，然后用手机端二维码，再观察network里面的变化

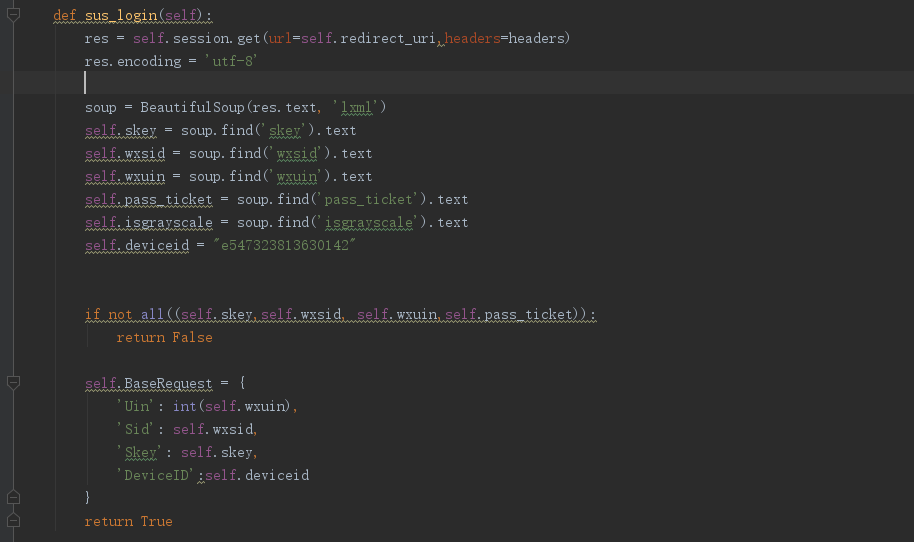


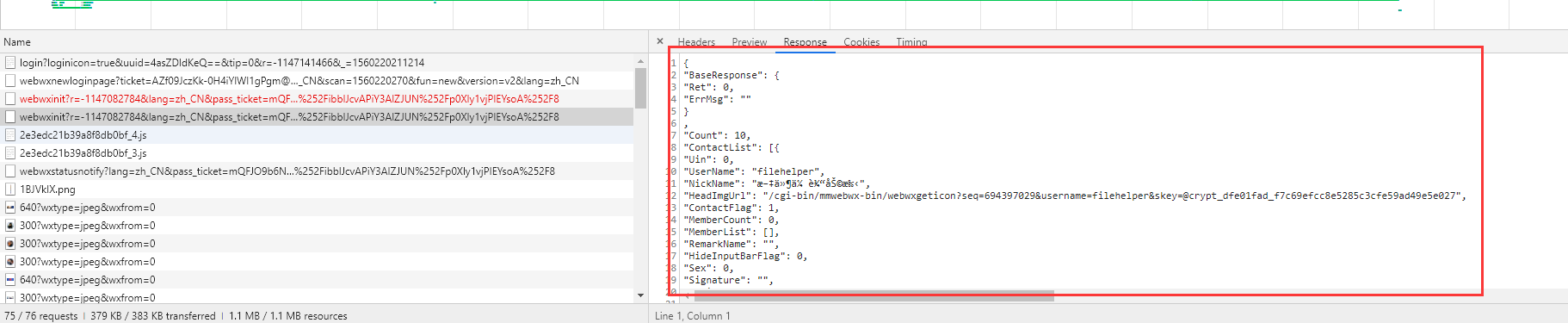
发现如果扫取二维码不登陆的话，会一直有链接在下面更新，也就说这是用来确认登陆的请求，每隔一段时间会请求一次，如果他的返回值是201，就说明手机端已经成功扫描但还没有确认登陆，如果是200，就成功登陆并返回一个uri，这就是登陆成功以后显示信息的uri



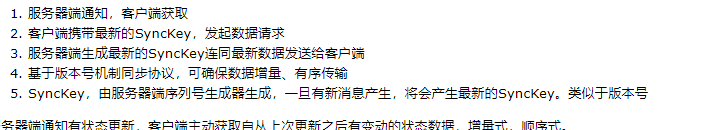


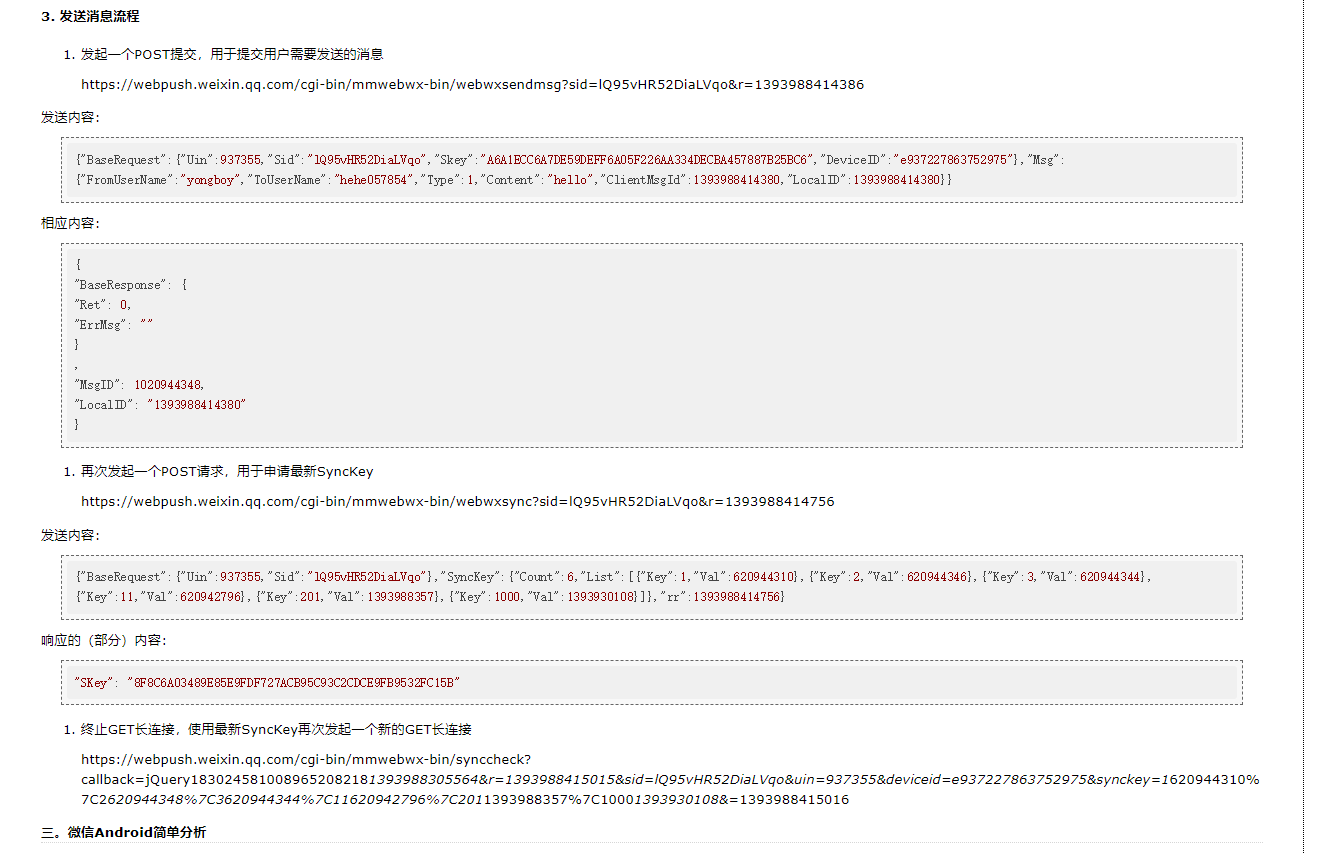
然后发现这个redirect\_uri返回的结果里面有许多属性值也就是键值对，在后面肯定用的到，作为参数，所以先把各个元素用beautifulSoup解析出来存放起来



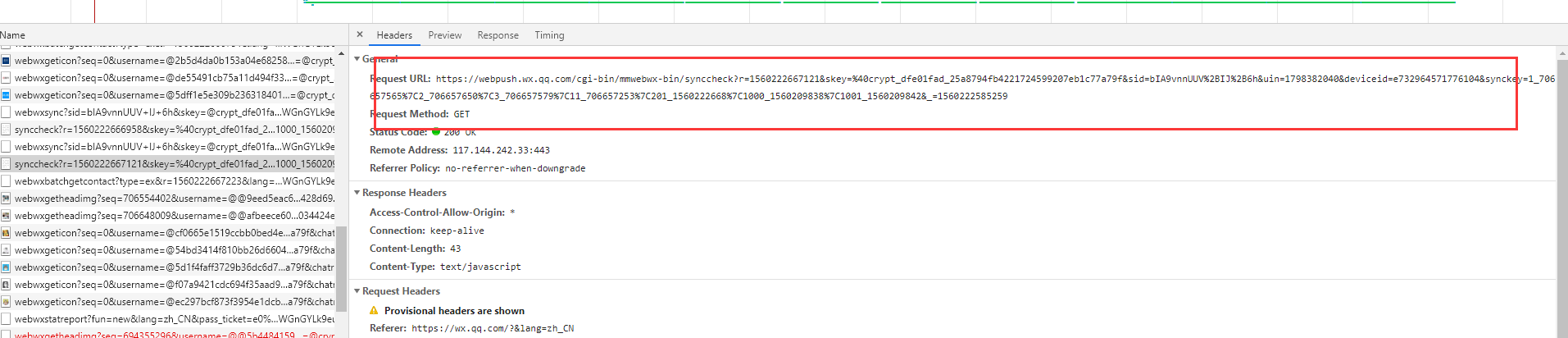


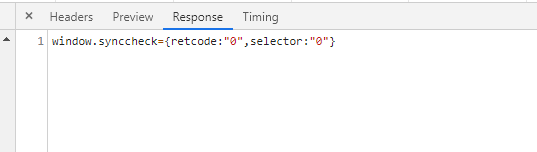
在分析network这个webinit里面就是返回内容的信息，拿到他的请求url [来解析https://wx.qq.com/cgi-bin/mmwebwx-bin/webwxinit?r=-1147082784&lang=zh\_CN&pass\_ticket=mQFJO9b6NHcFRAtrr9TAPMW5ZLJ%252FibblJcvAPiY3AlZJUN%252Fp0Xly1vjPIEYsoA%252F8返回的json串，把](https://wx.qq.com/cgi-bin/mmwebwx-bin/webwxinit?r=-1147082784&lang=zh_CN&pass_ticket=mQFJO9b6NHcFRAtrr9TAPMW5ZLJ%252FibblJcvAPiY3AlZJUN%252Fp0Xly1vjPIEYsoA%252F8来解析返回的json串，把联系人列表，用户)SyncKey取出来

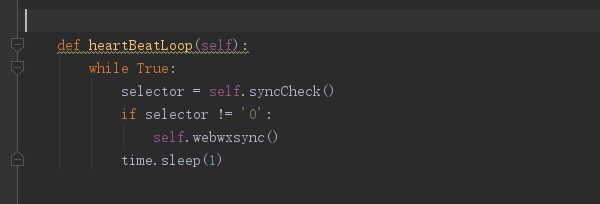




借鉴了一个博客的经验了解到微信是通过一个Synckey协议进行通信的，而且这个协议会更新，接下来就要获取联系人内容，还是对返回的json数据解析，把所有的好友列表保存下来。Synckey是由一个个的键值对组成的，这些key value是消息体的一部分，由于之前已经把Synckey保存了下来，发起GET长连接检测是否存在新的需要同步的数据，会携带上最新SyncKey，这时就把拿到的Synckey连接成串作为参数传给检验Synckey是否最新的url，如果sletcor大于0，则有新的消息需要同步。

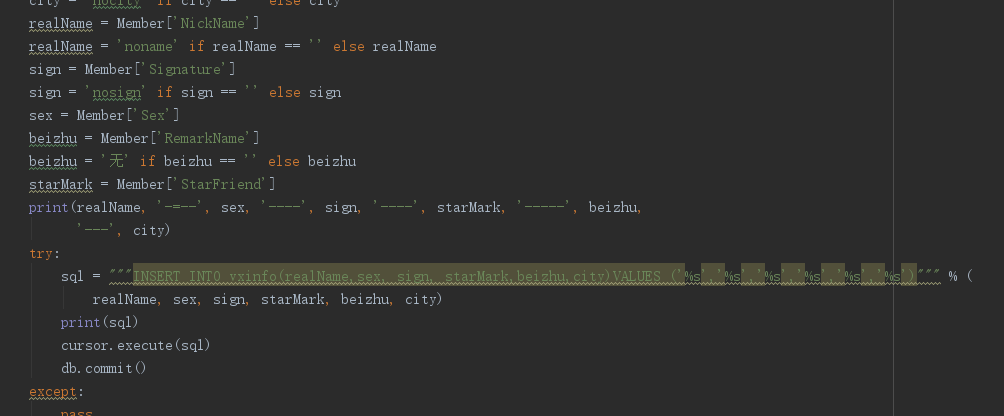






写一个方法用来一直检验是否有新消息需要获取，如果有及时更新Synckey

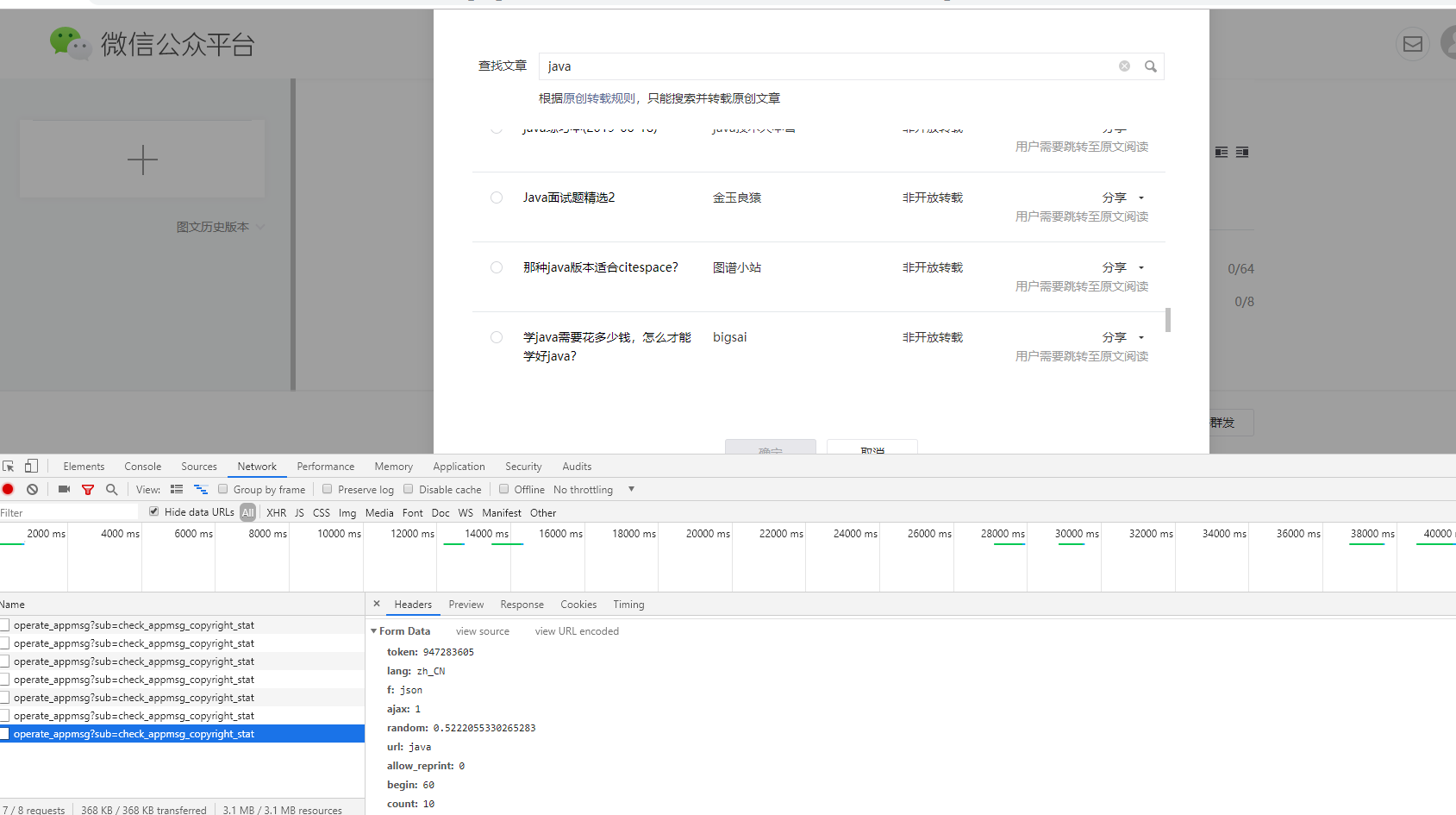
数据处理：对返回的json数据进行解析，通过python字典拿到几个有用的数据并存到数据库里面，包括昵称，备注，个性签名，所在城市，是否是星标朋友等



然后用pyecahrs的geo地理坐标工具，把数据库里面存的好友城市信息，把没有城市跟外国的去掉，展示在中国地图上，他会自动生成一个html文件。

爬取公众号文章：

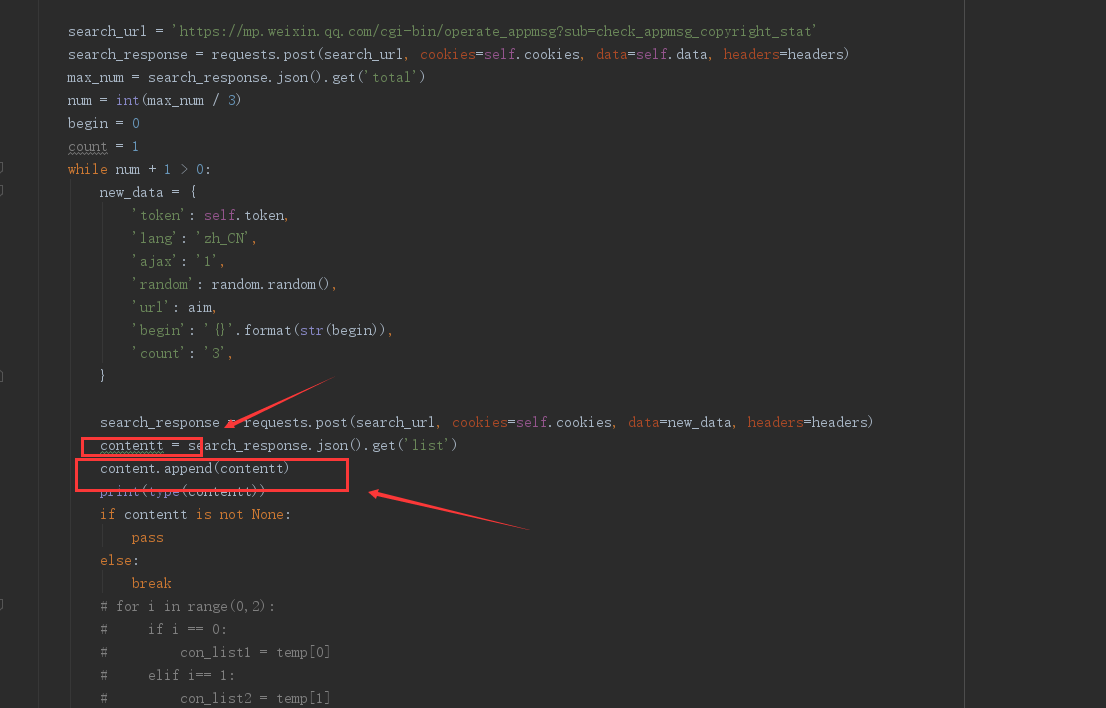
刚开始考虑在网页版微信页面爬取自己订阅的公众号信息，但是network并没有像好友信息那样一次性全部返回过来，通过查阅资料发现有个微信公众平台，可以注册并且发布公众文章，也可以搜索你想要拿到关键词的文章。

。

登录过程用到了selenium自动化测试工具，在Linux上用的是chrmoe的webdriver，通过模拟手动输入账号密码并且点击登陆，拿到跳转的网页的cookie并且保存下来，这样就可以在短时间内（3-4小时）不用重复登陆扫描二维码就可以爬取文章。

接下来就是爬取文章，上面这张图片就是搜索java后出来的所有的java文章，可以发现每往下拉都会重新发送一次请求，请求10个文章，那么可以自己定义header里面的内容一次请求多少个，我自己写的时候是3个，每三个循环一次直到，取到所有的文章。

每次请求一个文章在他的json串中拿到list字段的值，其中的值的类型就是一个list，每次拿几个，意味着最后应该是一个list的list



爬取文章后考虑把他保存成txt文件，但是链接太多，所以就把刚才的list的list写一个算法分成三块，其功能就是把list分成三块，用三个进程来往文件里存内容

