最近在做爬虫相关的工作，通过爬虫获取一些公开信息。今天与大家一起尝试使用模拟浏览器的方式爬取职场社交平台—脉脉网站上面的信息。

1.selenium安装与配置

Selenium本来是一个用于自动化测试的框架，但是能够完全模拟浏览器的行为，因此在爬虫中也可以发挥很大的作用。对于一些有很严格的反爬措施的网站，比如验证码、动态加载请求，都可以将Selenium作为爬虫的选择。

Python提供了爬虫的第三方库selenium，在使用之前，先要安装必要的工具。

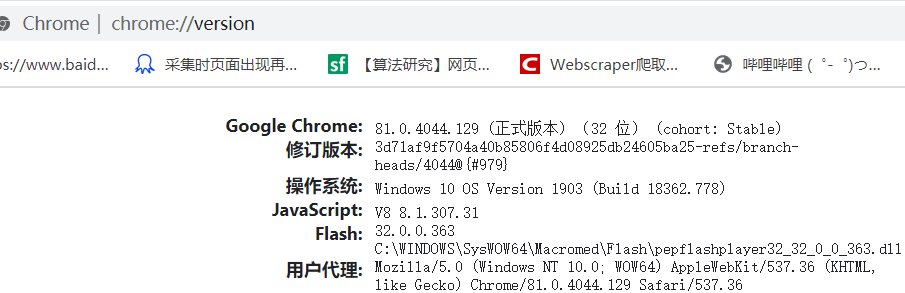
1.1 selenium安装

pip install selenium

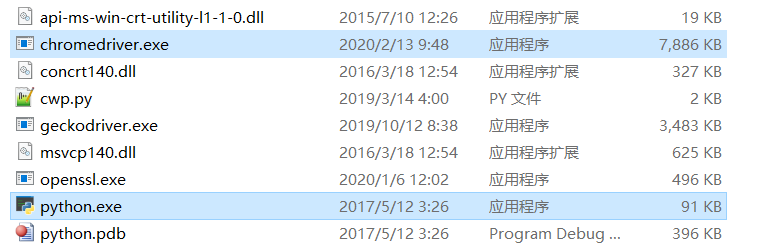
1.2 web driver安装

Selenium支持模拟目前所有的主流浏览器，包括chrome、firefox等，不同的浏览器需要安装不同的web driver。以chrome driver为例：

chrome drive的版本号必须与PC内安装的浏览器版本号一致，打开<chrome://version/>，



发现版本号为81.0.4044.129，打开<https://npm.taobao.org/mirrors/chromedriver/>，选择与浏览器版本号一致的chrome drive。找到81.0.4044.69（前三位一致就行，81.0.4044.69和81.0.4044.20选择更高的前者）。解压后，放入python的安装路径下。



1.3测试

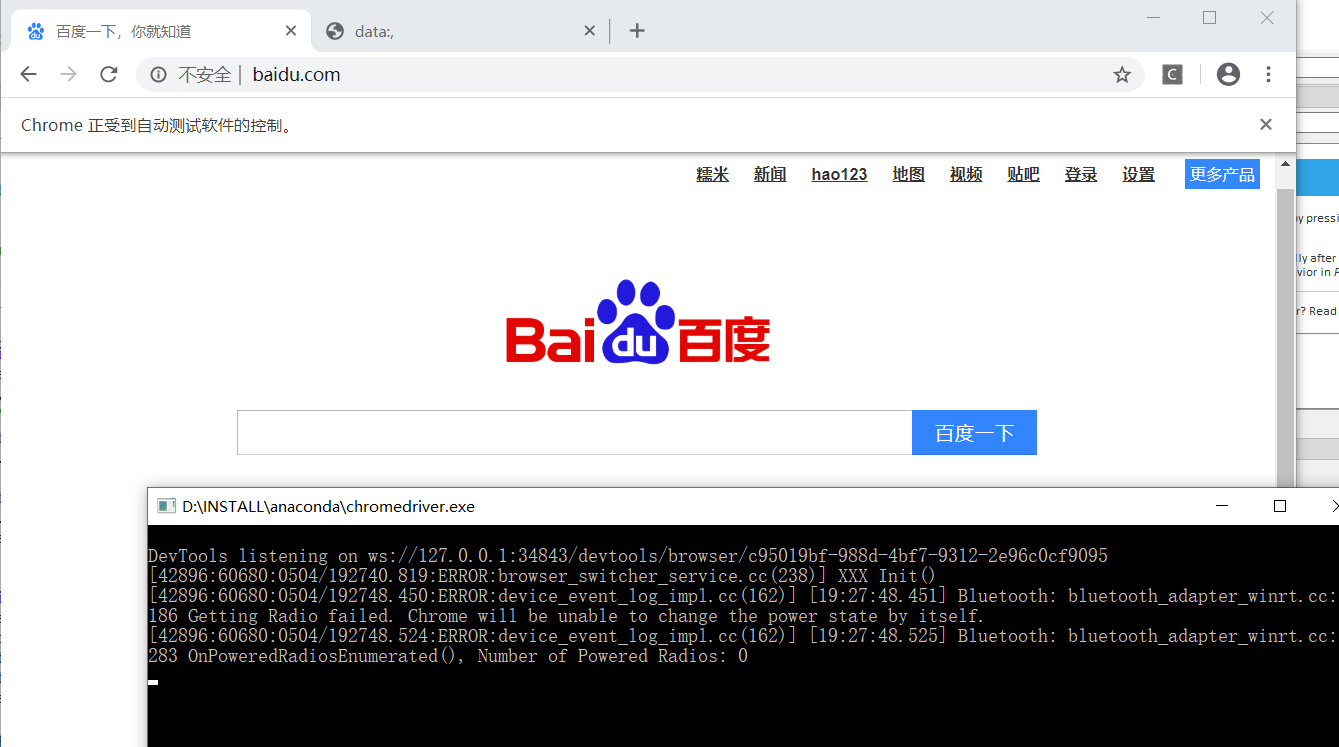
上代码：

from selenium import webdriver

browser = webdriver.Chrome()

browser.get('http://www.baidu.com/')

若能自动打开下面的界面，表示安装成功。



2.开始“白嫖”（开始爬虫）

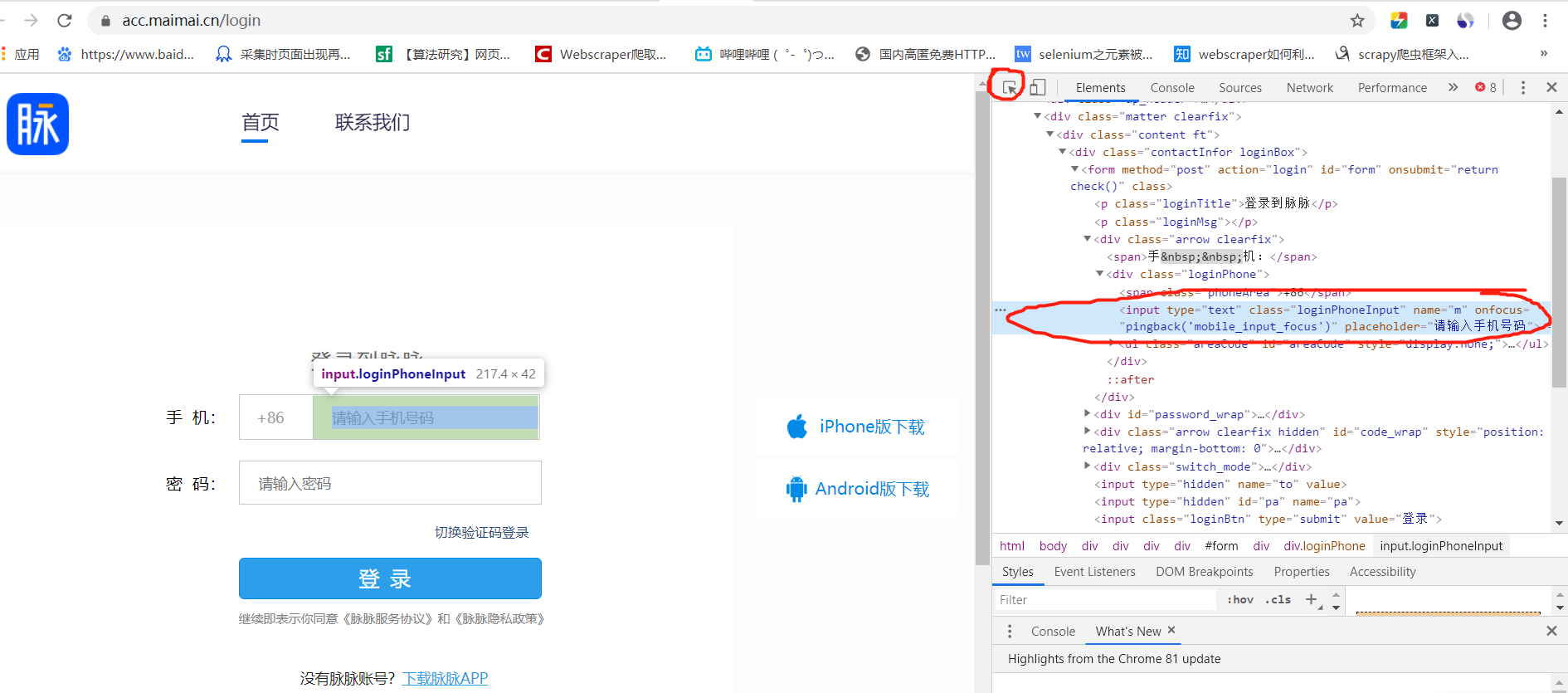
前面说了，selenium是模拟人使用浏览器的一系列操作，因此我们先要看看人是怎么浏览脉脉这个网站的。

2.1登录

打开官网<https://acc.maimai.cn/login>，先要登录。正常的流程是分别输入手机号码和密码，再点击“登录”按钮。同样，使用selenium也需要告诉浏览器输入账号密码以及点击登录，我们在输入账号密码之前可以使用鼠标定位到输入框，那么使用模拟浏览器怎么才能定位到输入框呢？答案是Xpath。

XPath即为XML路径语言（XML Path Language），它是一种用来确定XML文档中某部分位置的语言，使用XPath可以在网页中快速定位想要获取的元素，比如文字、按钮、输入框等（具体教程移步<https://www.runoob.com/xpath/xpath-tutorial.html>）。

举一个栗子，在登录脉脉时先要输入账号，账号的输入框可以使用xpath定位。右击网页—检查，使用“inspect”点击某一元素（以手机账号输入框为例）可以快速在源码中定位到对应位置，右击选择Copy—Copy Xpath即可获取输入框的xpath路径，便于后续的浏览器定位。采用同样的方法去定位密码输入框和“登录”按钮。



在定位到三者的xpath路径后需要告诉浏览器相应的操作，比如在输入框输入账号密码、点击按钮。直接上代码：

browser.find\_elements\_by\_xpath('//\*[@id="form"]/div[1]/div/input')[0].send\_keys(“账号”) browser.find\_elements\_by\_xpath('//\*[@id="login\_pw"]')[0].send\_keys(“密码”) browser.find\_elements\_by\_xpath('//\*[@id="form"]/input[3]')[0].click()

'//\*[@id="form"]/div[1]/div/input'、'//\*[@id="login\_pw"]'、'//\*[@id="form"]/input[3]'分别为账号输入框、密码输入框以及登录按钮的xpath路径。输入框使用方法send\_keys()输入账号密码，按钮使用方法click()实现点击。

2.2爬取个人信息（敏感信息，未经授权，请勿转载！）

在登录进去后，如果想搜索某个人，直接在搜索框输入搜索信息即可，也可以在url直接指定搜索信息，比如<https://maimai.cn/web/search_center?type=feed&query=金融壹账通&highlight=true>表示搜索“金融壹账通”，本次demo也使用方法二展示。该网站提供了json接口：https://maimai.cn/search/contacts?count=10&page=1&query=金融壹账通&dist=0&searchTokens=&highlight=true&jsononly=1&pc=1，可以快速解析数据。

当点开某个用户的主页时，同样可以上述的方式快速定位工作经历、教育经历等等。





2.3应用

你可以试一下“金融壹账通 张杰”或者“金融壹账通 吴满芳”亦或是其他everything。