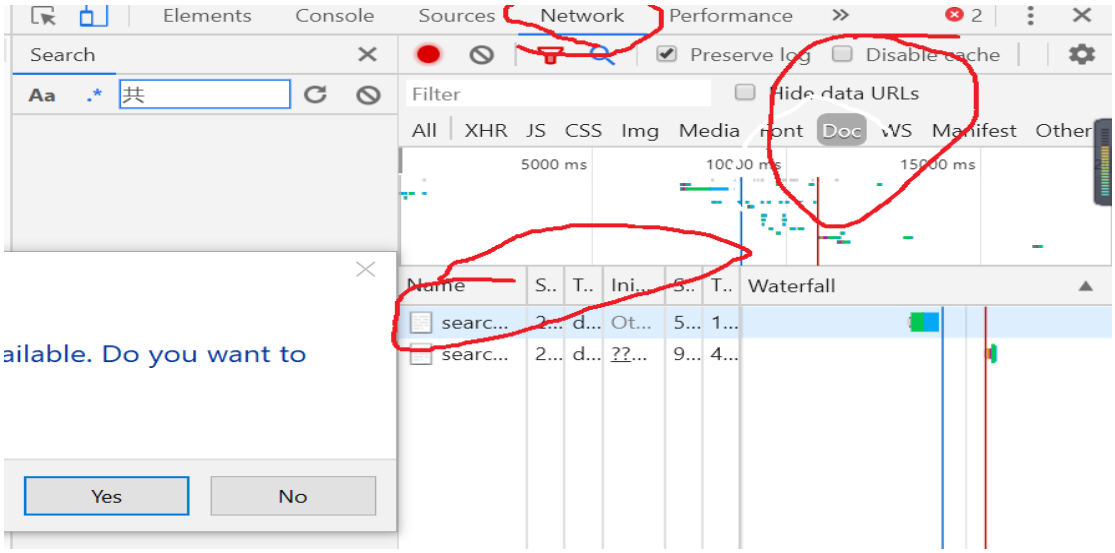
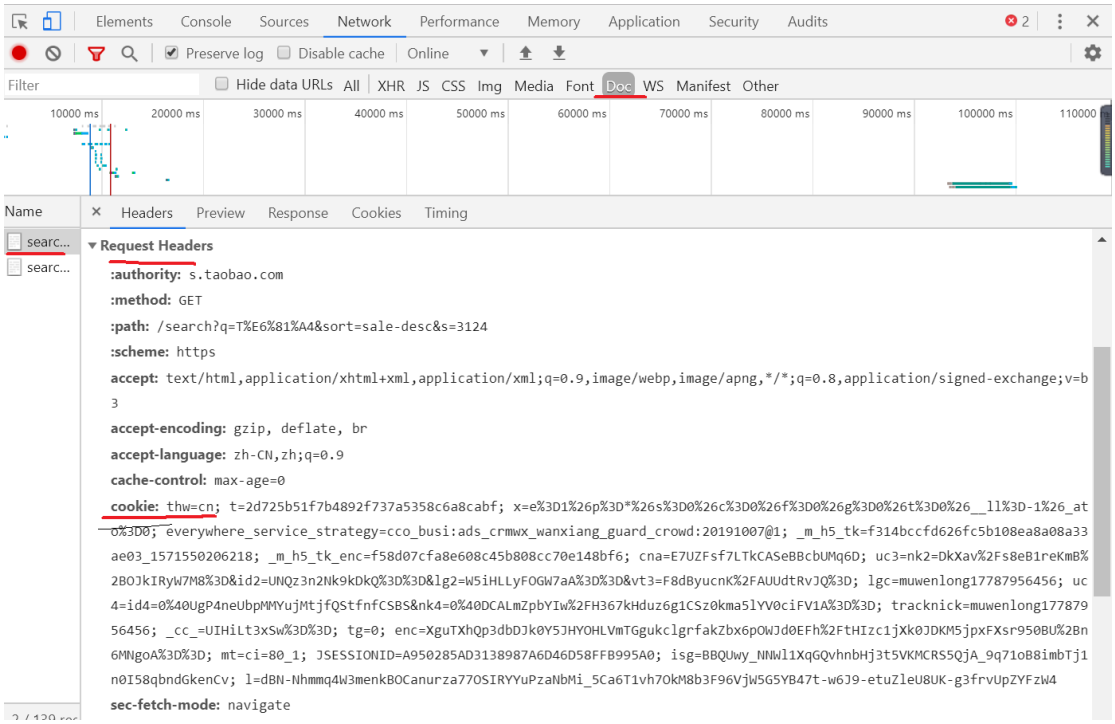
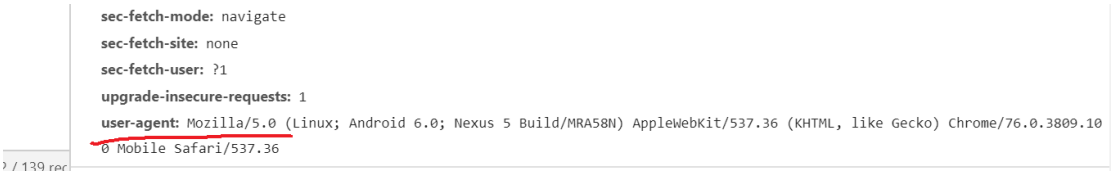
文件执行顺序如下：
catch.py
（connect.py）
deal_data.py
analyze.py

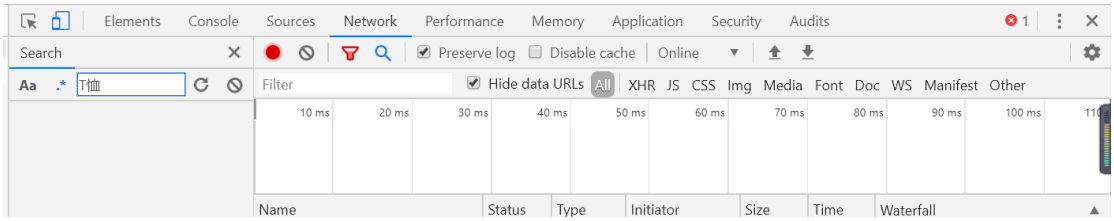
catch.py:
对于淘宝数据爬取过程中，第一步是进行模拟登录，第二步是进行 json 分析，第三步把数据存到 excel。
模拟登录：采用 requests 的 header 进行登录（cookie，user-agent）获取 header 形式如下：







然后通过 chrome 的自带的对照功能，或者下面的 search 功能进行查询：



在爬取内容之前，要先查看网页源代码，知道它的内容是直接摆在网页内部的，或者用脚本文件（json）。
不要先右键 检查————element，不然很可能白费功夫。!!!!!!!!

通过正则表达式提取 json 文件：正则表达式是自己的软肋!
解析 json 文件：

网站： *http://www.bejson.com/*

工具：VS code，文件后缀名 json
域名：
url="https://s.taobao.com/search?q=%s&sort=sale-desc&s=%d"%(keyword,pagenum)
关键字，页数*44（淘宝是按商品数确定页数标签，不是 1，2，3 顺序排列，*44 是说一夜页之内有 44 个商品）

本程序难点在于，对于初学者，要先分析网页结构，再选择爬取方式，少走弯路

（connect.py）
这个程序的存在就是我整个程序不够完备的最好说明，taobao 网存在反扒的机制，爬的过多或者爬的过快，都会被限制的。要么让你重新登录，要么让你滑动滑块验证，两者随便一者就会让你的爬虫终止运行。用 request 的 header 不是长久之计。
该程序就是说，如果你爬虫爬到一半中止，你可以再重新开启一个爬虫（把 for 循环里的参数改一下，知道它是在那一页终止的，就从哪里开始）。然后这个程序把你爬取的几个 excel 连接到一起。

deal_data.py
本程序主要是通过 for 循环，对字符串加以处理，变得规范化，没太多难点

analyze.py
本程序主要对数据加以处理，完成数据可视化工作，第一项处理是原来学习过的词云处理，

第二项是我用新发现的工具 pyecharts 进行处理，pyecharts 是中国人写的一个模块，有全中文版文档 https://pyecharts.org/#/zh-cn/intro 强烈推荐!!!!
比 matplotlib 简单上很多很多，还易于理解，但有一点比不了的是 matplotlib 和 pandas 是配套使用的，更方便。用 pyecharts 也不难，但应该把 dataframe 转化为列表（list），用 pyecharts 保存的是 html 形式，若需保存为图片，还需进行其他环境配置，pyecharts 是一个宝藏，还有好多好多东西需要我挖掘，这次我就运用 pyecharts 完成了之前认为很麻烦的地理热图的制作，很酷。而且 matplotlib 要显示中文标签的话需要很多其他设置，略显繁琐，pyecharts 就不用啦，本身就是中国人写的。

整个程序的编译并不是十分高深，但是通过这两百来行的代码，自己学会运用很多新的方式，新的工具，不完美的地方还有很多，需再接再厉！

https://github.com/onroadmuwl/clawer_2

2019/10/22

牟文龙