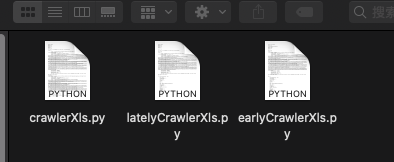
**一、目标**

收取2015年科学计量学文献特征：作者h指数、主题词、国家、机构、学科、早两年被引次数、早两年被引次数、早两年被引次数、早两年被引次数、早两年被引次数、近两年被引次数、近两年被引学科数、近两年来源出版物数、近两年被引机构数、近两年被引国家数

Ps：H指数需要请求的网页较多，等待时间较长，错误率较大，且可能因为频繁进入网页需要多次换header，收取时可先注释掉，最后可以考虑最后单独手工收取。

**二、爬虫文件简介**

爬虫文件：



1.crawlerXls.py：爬取作者h指数、主题词、国家、机构、学科

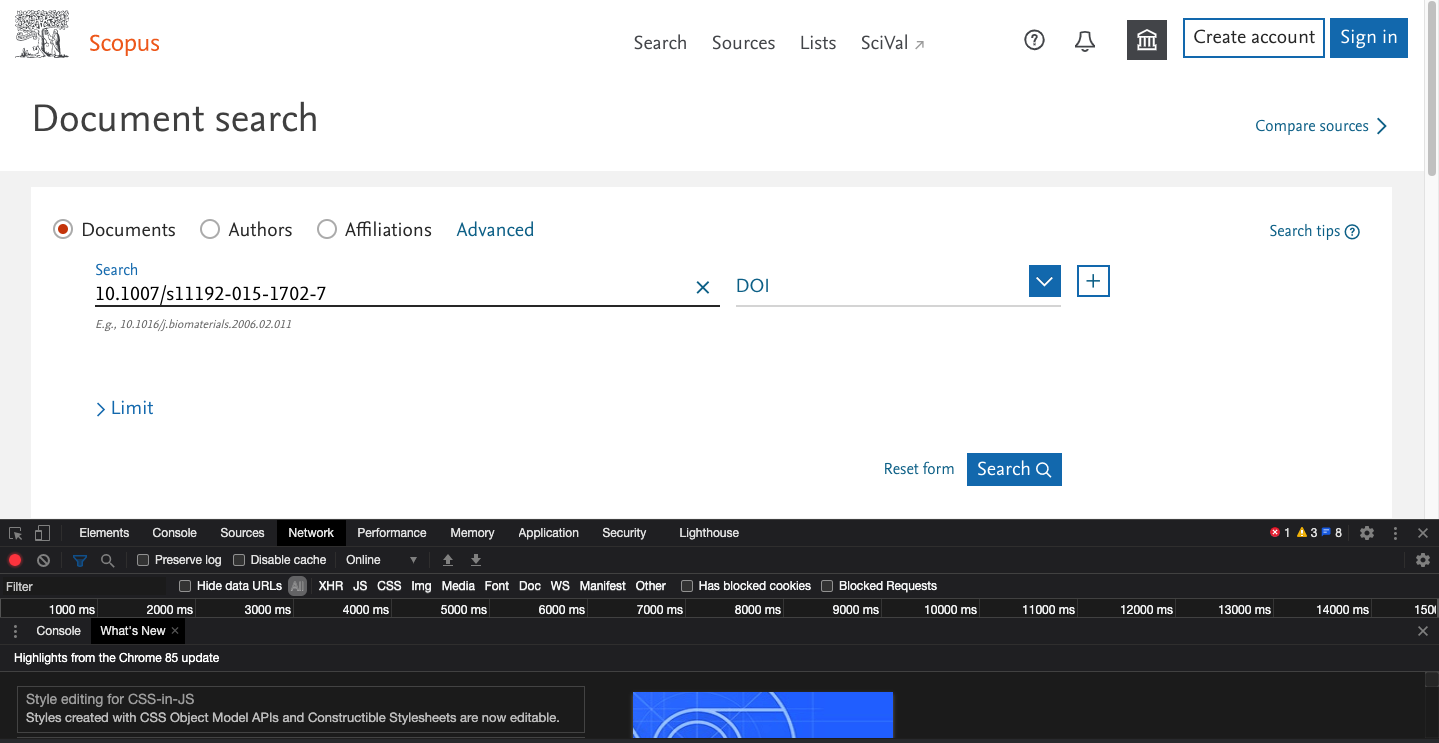
2.earlyCrawlerXls.py：早两年被引次数、早两年被引次数、早两年被引次数、早两年被引次数、早两年被引次数

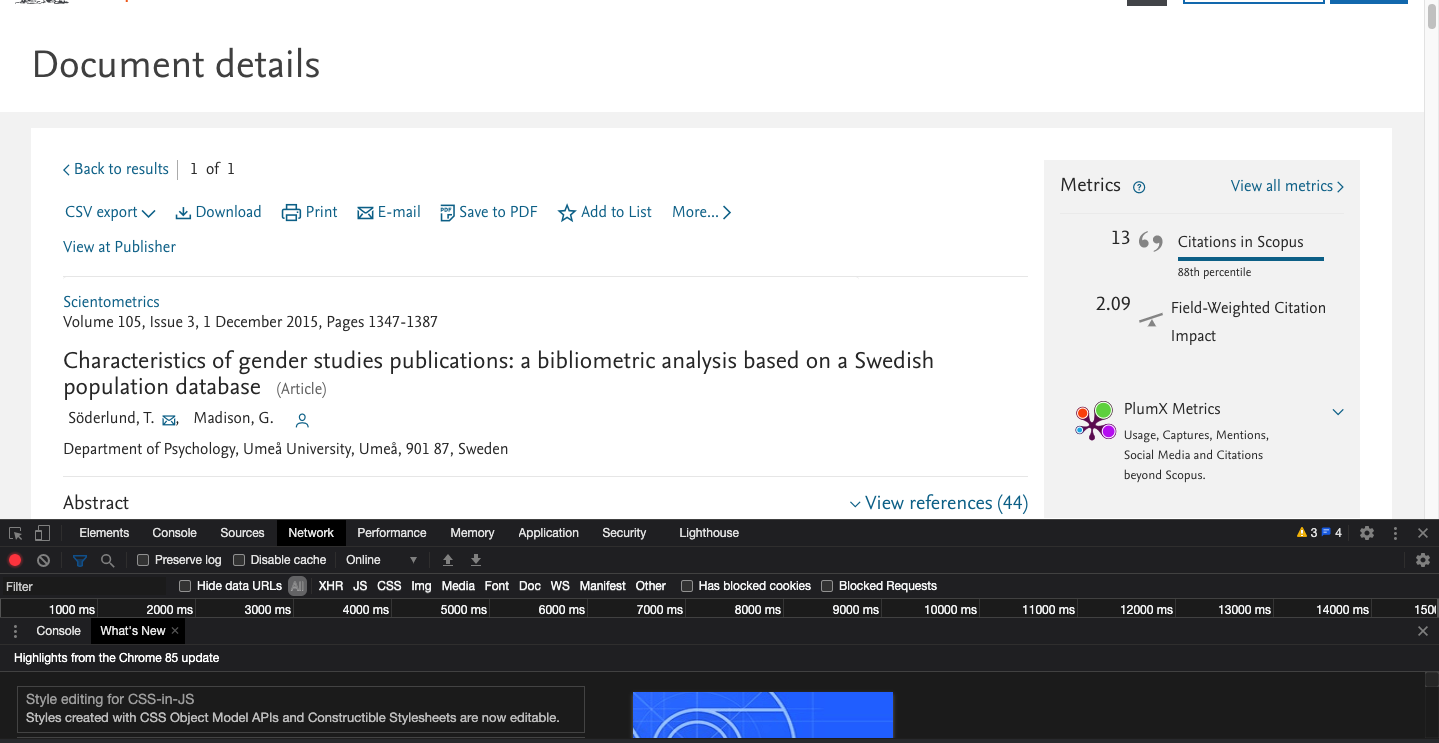
3.latelyCrawlerXls.py：近两年被引次数、近两年被引学科数、近两年来源出版物数、近两年被引机构数、近两年被引国家数

**三、爬虫使用说明**

无论用哪一个文件，都需要修改的是以下几处代码：

1. 更新header和headerDetail中的‘cookie’。且一旦网页进不去，就需要更新‘cookie’。

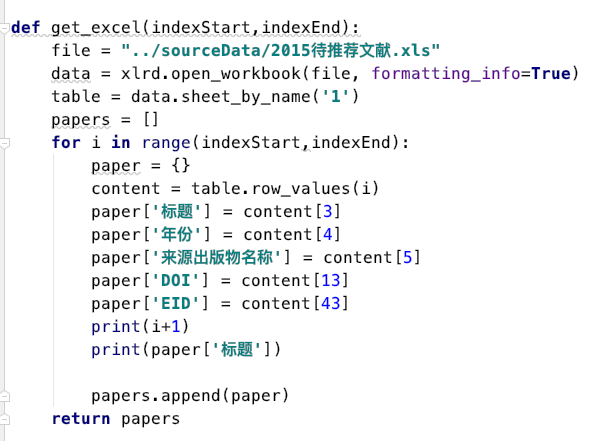
更新header，直接在scopus主页刷新获取；

更新headerDetail，在任意文献的详情页刷新获取；

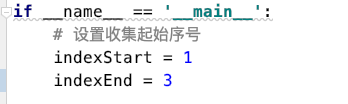
换header 的cookie的详情见录屏视频，headerDetail换法想同。



2.将get\_excel方法中file = "../sourceData/2015待推荐文献.xls"，文件的位置改为自己电脑上的地址。

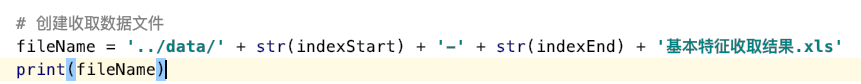


3.按需修改开始序号indexStart和结束序号indexEnd。



4.将fileName中'../data/'，改为自己想要存数据的路径。

注意：+ str(indexStart) + '-' + str(indexEnd) + '基本特征收取结果.xls' 后面这串不变



5.以下为收取结果，最后记得将特征复制粘贴汇总。

