python空气污染指数JS爬虫代码文档

I11714008 经济学院 经济学类 吴峰宇

代码截止至2018年12月17日13:25分运行无误

目标网站：https://www.aqistudy.cn/historydata 中各个城市2014-2018年污染物数据

输出方式：整理成面板数据格式输出至excel

网站数据生成类型：由Js动态加载

爬取流程：通过selenium库调用无头浏览器phantomjs获得网页全部源代码（包括由JS加载的数据）；后使用BeautifulSoup对网页进行解析，将所需数据提取出来形成列表；利用pandas库整理成DataFrame格式后输出

循环方式：循环生成每个城市的月份信息，加入原始url链接形成目标网页链接。

缺陷：非分布式，爬取效率较低；

调用phantomjs获取全部网页信息，占用资源较大

只能对单个城市不同月份进行循环，无法对城市进行循环（地址中代表城市的信息为二进制形式，无法自动生成）

会产生一个warnning,但是不影响使用