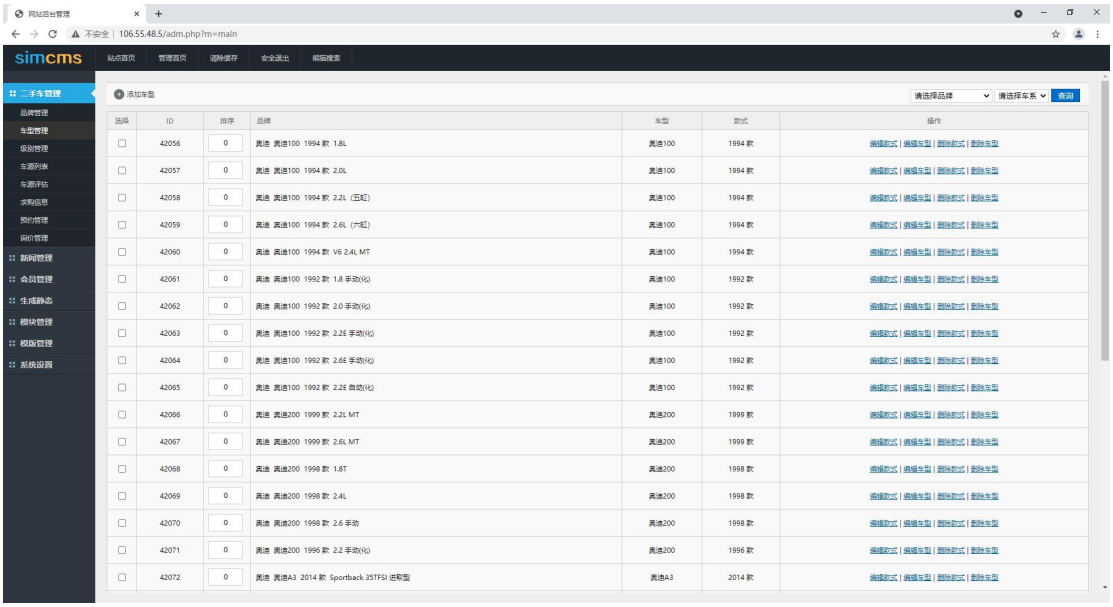
阿乐数据采集器 V2.0 使用说明书

1.1 软件简介

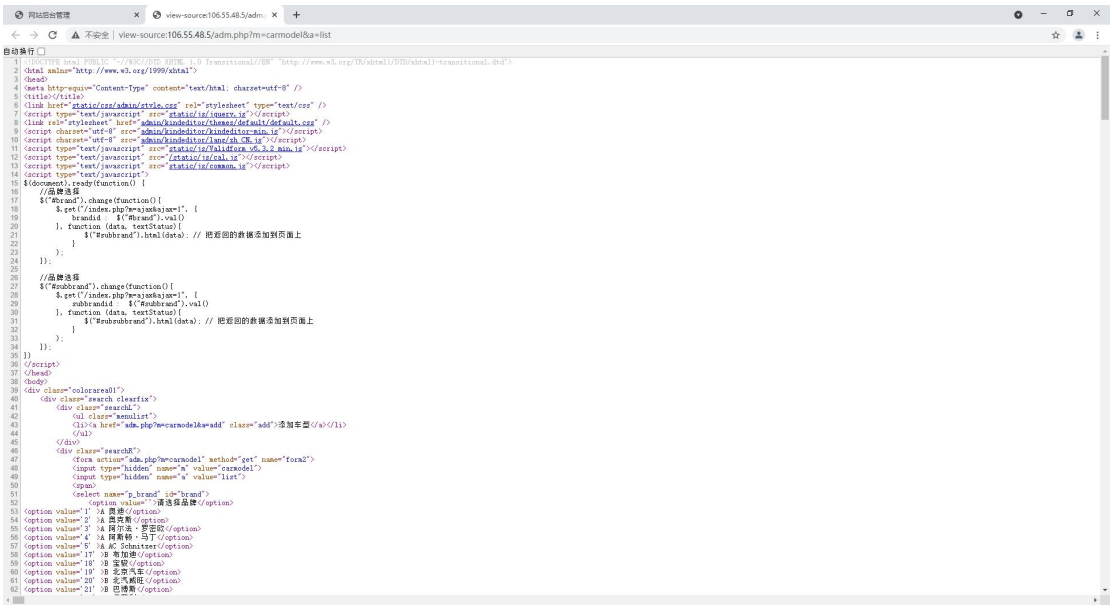
快速爬取 url 有规律变化的网站，保存静态的 html 文件。如果是后台管理系统或者网站中包含表格，可以自动将表格保存至 csv 文件。同时也可以爬取前端渲染类网站的接口，将 json 格式的数据，转化保存至 csv 文件。

1.2 使用说明

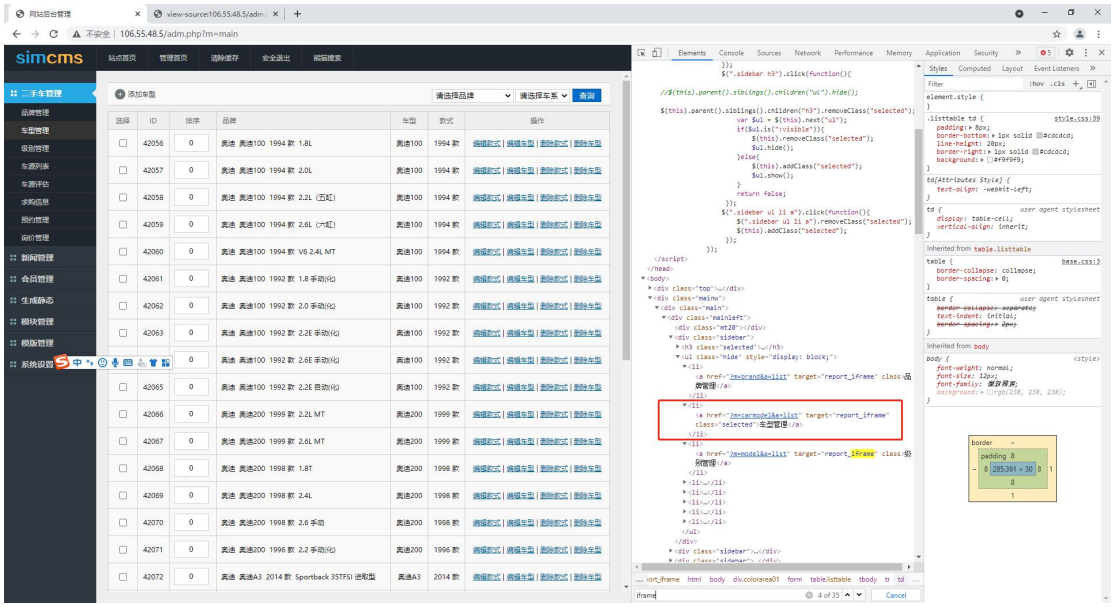
一、用浏览器打开需要爬取的网站，打开需要采集的模块。如下图



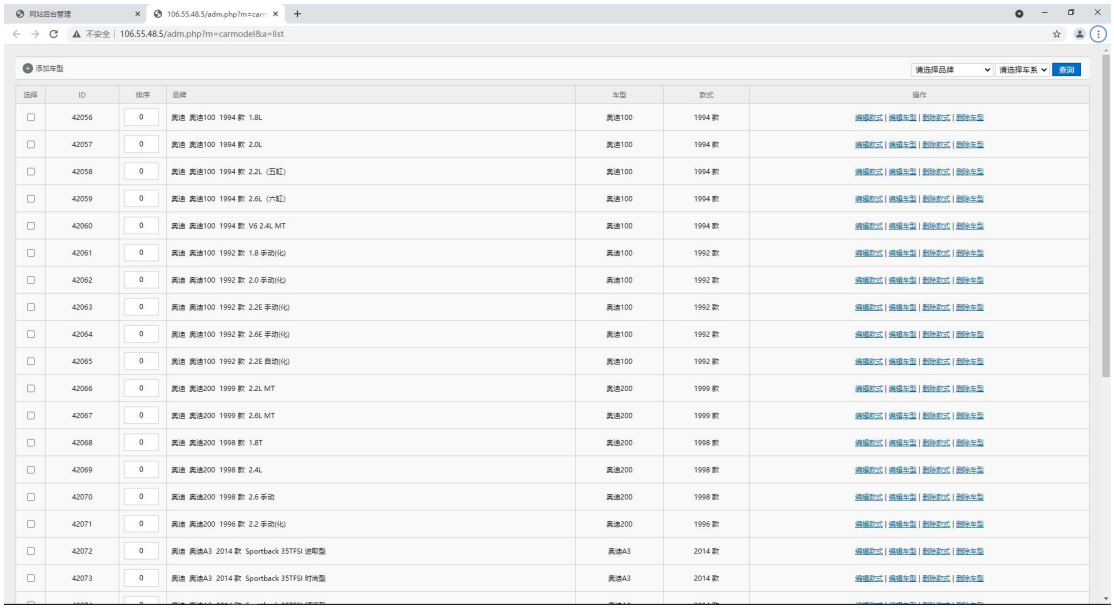
二、如果有 iframe 框架，需要去掉框架，即打开真实的链接。这里有两种方法，第一种可以通过右键表格区域，查看框架源代码，这里以软件内置的绿色版谷歌为例，如下图



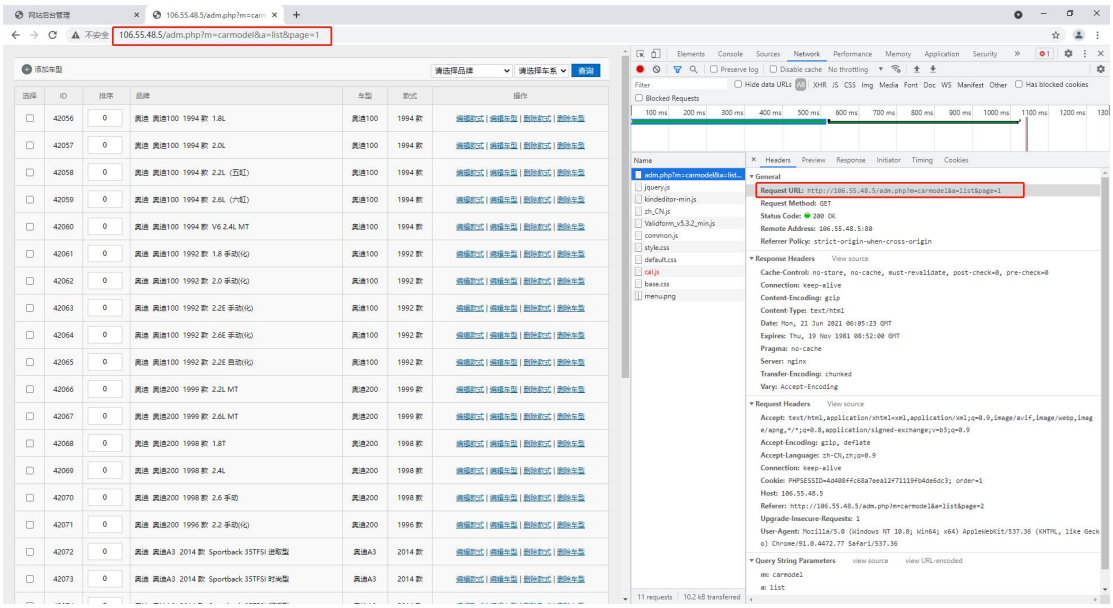
去掉地址栏中的 view-source:即可打开真实的链接。如果是火狐浏览器则更方便，右键点击表格区域，点击此框架，再点击新建标签页打开框架。另一种方式是通过 F12 去查看网页源代码，搜索 iframe，如下图所示

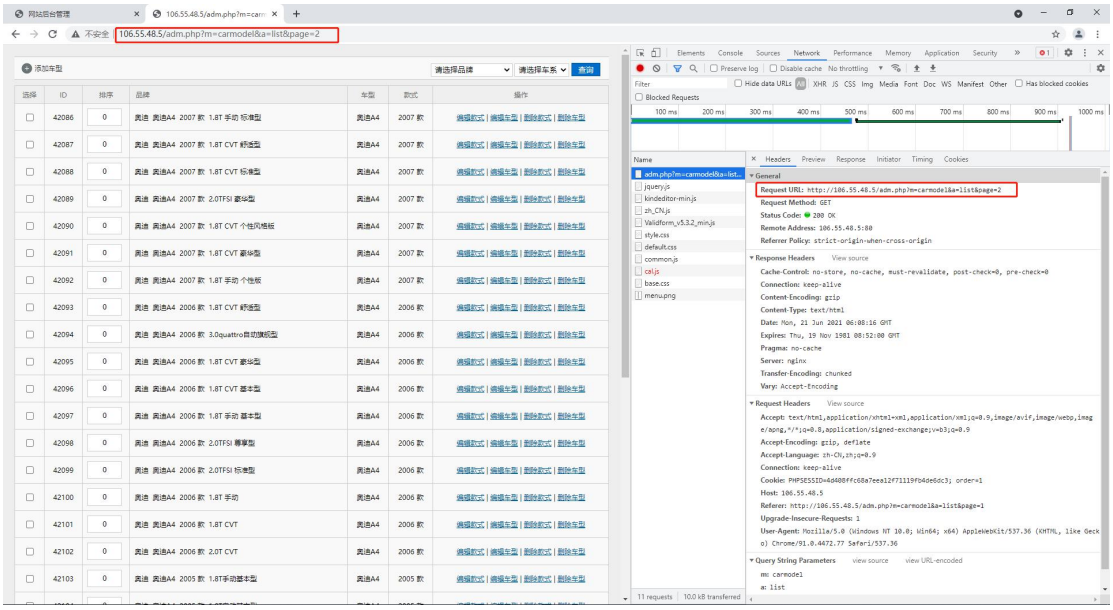


右键点击找到的链接，选择 Open in new tab（在新的标签页中打开），打开后如下图



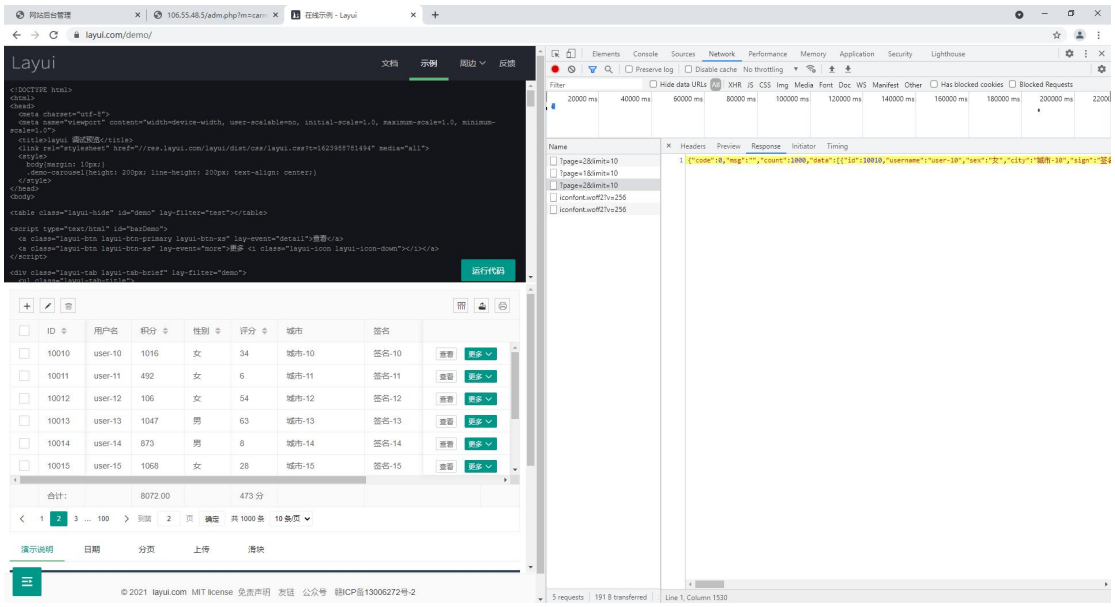
三、在新打开的标签页中，右键点击检查或者按 F12，接着点击 Network。开始前可以先点击 clear 清除之前的一些内容。我们通过点击翻页按钮，来观察 Network 中请求的变化，如图

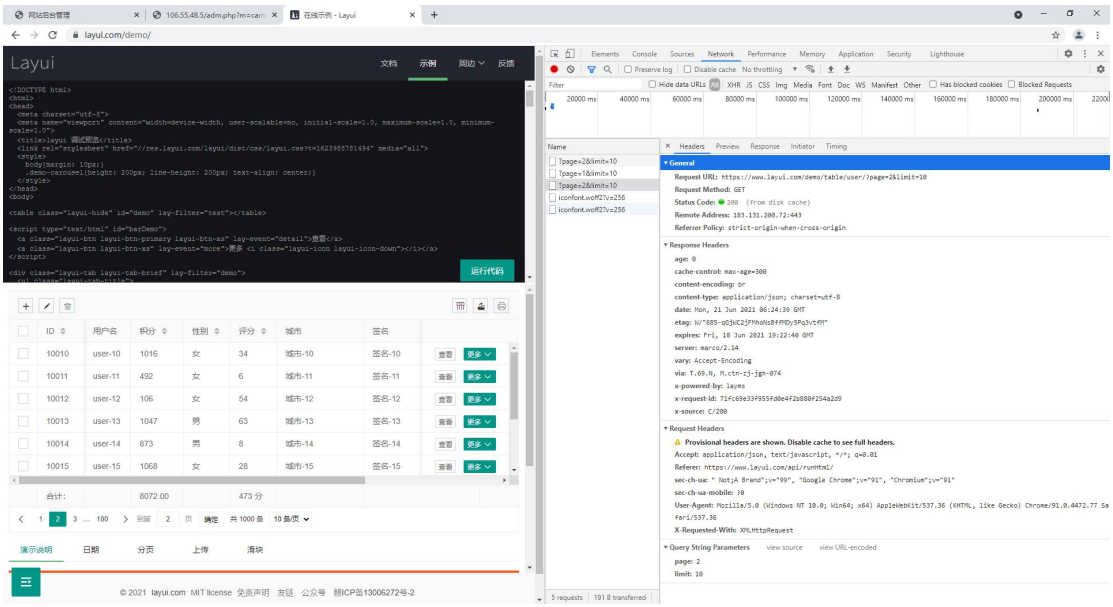




这里不断变化的是 page 后面的数字，打开阿乐数据采集器 V2.0.exe，将 http://106.55.48.5/adm.php?m=carmodel&a=list&page=1 中&前面的部分填入 url 的框中，将 page 填入 variate（变量）中，在第三项 page（页数）中填入需要爬取的页数。如下图

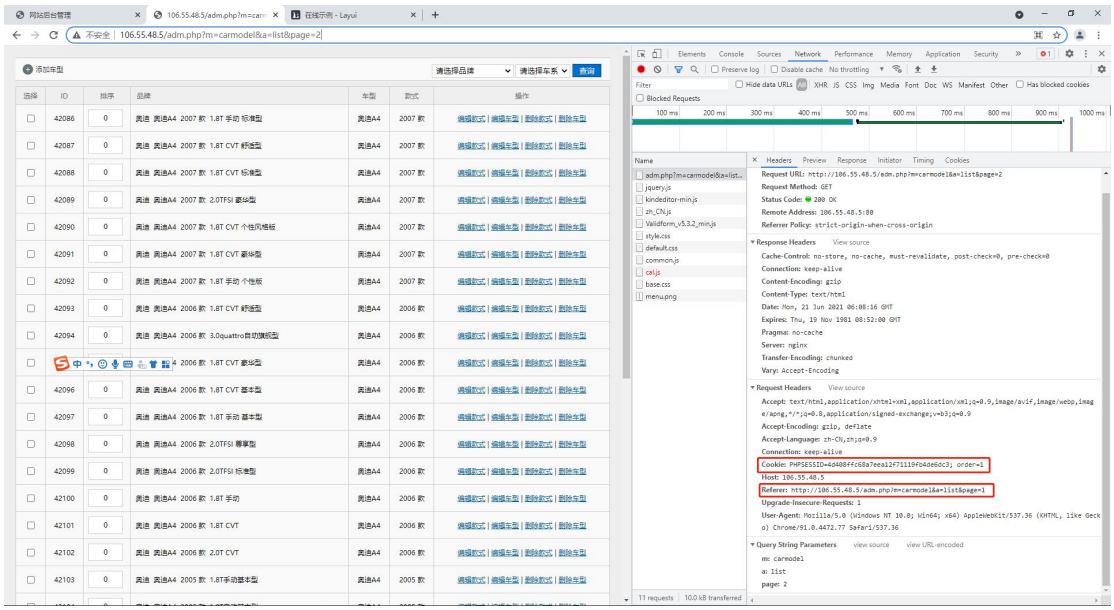
如果是前端渲染类的网站，地址栏中 url 是不变的，在 Network 中可以看到翻页时，请求的 url 也是有规律的变化，服务器返回来的是 json 格式的数据，如下图





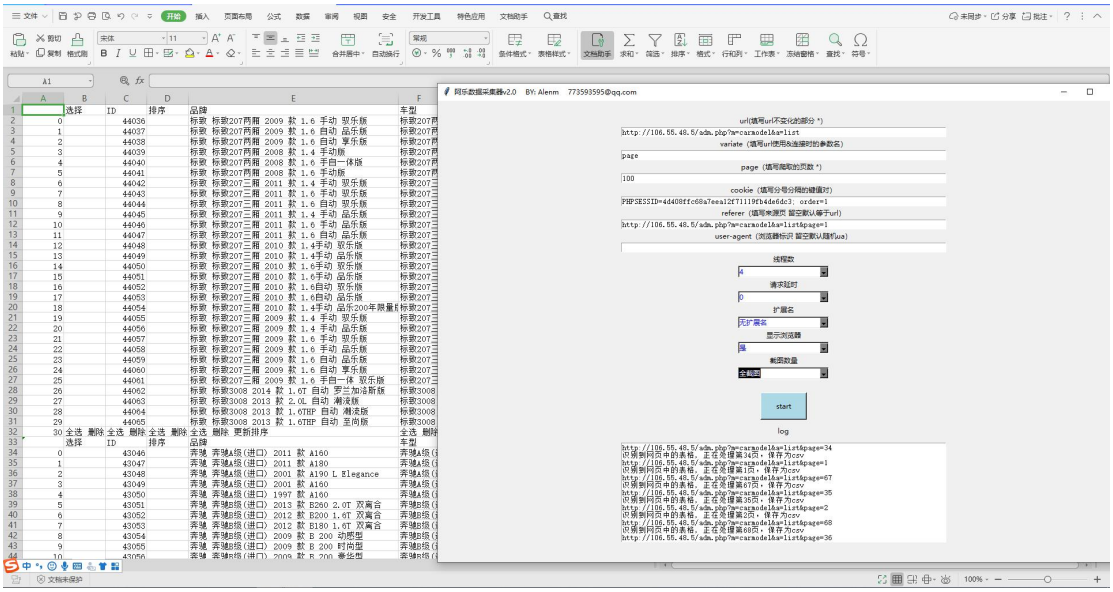
这类的网站，填写方式和前面是一样的，去掉中间变化的部分填入 url，参数填到 variate 中，如图

四、其他内容的填写。如果是网站后台，必须要填写 cookie。有些网站必须要填写 referer。这些内容也都在 Network 中可以找到，如下图填写



五、根据网络、网站性能调整线程数等设置。勾选截图会调用绿色版谷歌浏览器截图，会影响爬取速度。如果网站有.html 或者.jsp 等扩展名，则需要勾选相应的扩展名。

1.3 效果截图及性能



我自己测试的网站，PHP 做的 cms，里面有表格。服务器配置是腾讯云轻量级服务器，2 核 2g 内存，带宽 5m。网络是电信 300m。

Speed：

单线程，爬取 800 页，处理表格，不截图，用时 130s

单线程，爬取 800 页，处理表格，全截图，用时 280s

四线程，爬取 800 页，处理表格，不截图，用时 40s

四线程，爬取 800 页，处理表格，全截图，用时 108s

八线程，爬取 800 页，处理表格，不截图，用时 36s

八线程，爬取 800 页，处理表格，全截图，用时 57s

2.1 其他正在开发的功能

一、浏览器输入网址，自动提取 url，cookie，referer 等。

二、爬取的日志，以及 html 文件的哈希值。

三、post 请求网站的爬取。

四、动态加载网站的爬取和保存。

五、多个参数变化的网站爬取。

六、服务器端，分布式爬取。

七、页数选择，从中间的页数开始提取。

八、代理，加入 ip 代理池。

九、文件锁，爬取网站时，打开 csv 文件用只读，不会使爬虫停止。

十、UI 太丑了，研究一下 qt5 重新做一下。

2.2 项目地址

GitHub：https://github.com/zmzmon/CJ

Gitee：https://gitee.com/saa1028/cj