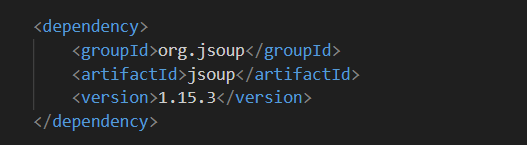
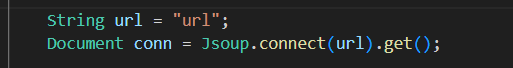
本篇文章用于记录自己的爬虫使用经验：

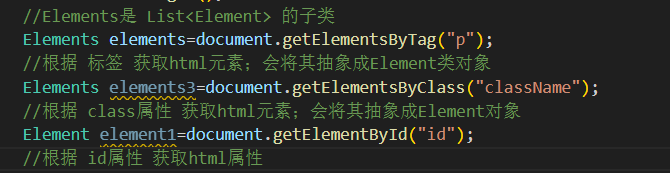


通过在maven中导入jsoup依赖，可以进行简单的爬虫操作

**根据url发送http请求获取对应的html页面，封装成Document类对象**：



**常用的提取元素的方式**：



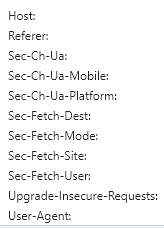
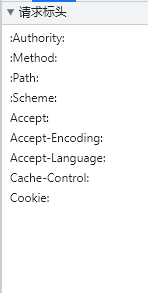
**提取元素中的内容**：一般用于过滤信息

**但是实际情况是，许多网站都会有点“反爬虫”措施，或者需要你登录才能看到想爬取的信息，这种情况该怎么办呢**？

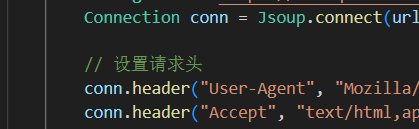
为了解决这个问题，我们需要知道上面的爬虫发出的http请求和我们平常访问网站时有什么不同：

首先，你得能正常访问需要爬取的网站，之后利用f12查看正常访问时的http请求

之后找到左边这样的http请求头

但是呢，我们上面发送的爬虫请求，是没有带请求头的http请求

**因此我们只需要在爬虫请求中添加这些请求头，就能骗过目标服务器，具体操作**为：

在conn对象调用get方法前，设置请求头即可

此外，通过请求头也可以设置cookie，JWT等信息，因此也能通过这种方式完成**登录**。