**运维工程师**

**《爬虫基础》**

**制作人：杜正斌**

**18981693537@163.com**

**时间：2021年05月**

# 目录

[目录 1](#_Toc477465994)

[版权声明： 2](#_Toc477465995)

[文档信息： 2](#_Toc477465996)

[文档格式： 2](#_Toc477465997)

# 版权声明：

**文章的所有内容属于作者的个人资料，只允许作者给属的个人进行使用，如需进行传播与它人，请联系作者，获取传播知晓！如发现私自传播，违者取消后续文档获取资格，请看文档的人员严格遵守！**

# 文档信息：

**文档作者：杜正斌**

**文档版本：Version 1.0**

**修改记录：2021-05-16**

**软件版本：Python-3.7.2**

# 文档格式：

**[绿色背景] 知识重点**

**[黄色背景] 错误警告**

**[红色背景] 注意事项**

**[灰色底纹] 命令操作**

**[蓝色底纹] 命令格式**

**命令格式**

**执行命令**

# 1、URI和URL

URI：全称为Uniform Resource Identifier，即统一资源标志符。

URI一般由三部组成：1、访问资源的命名机制。2、有效资源的主机名。3、资源自身的名称，由路径表示。

URL：全称为Universal Resource Locator，即统一资源定位符。

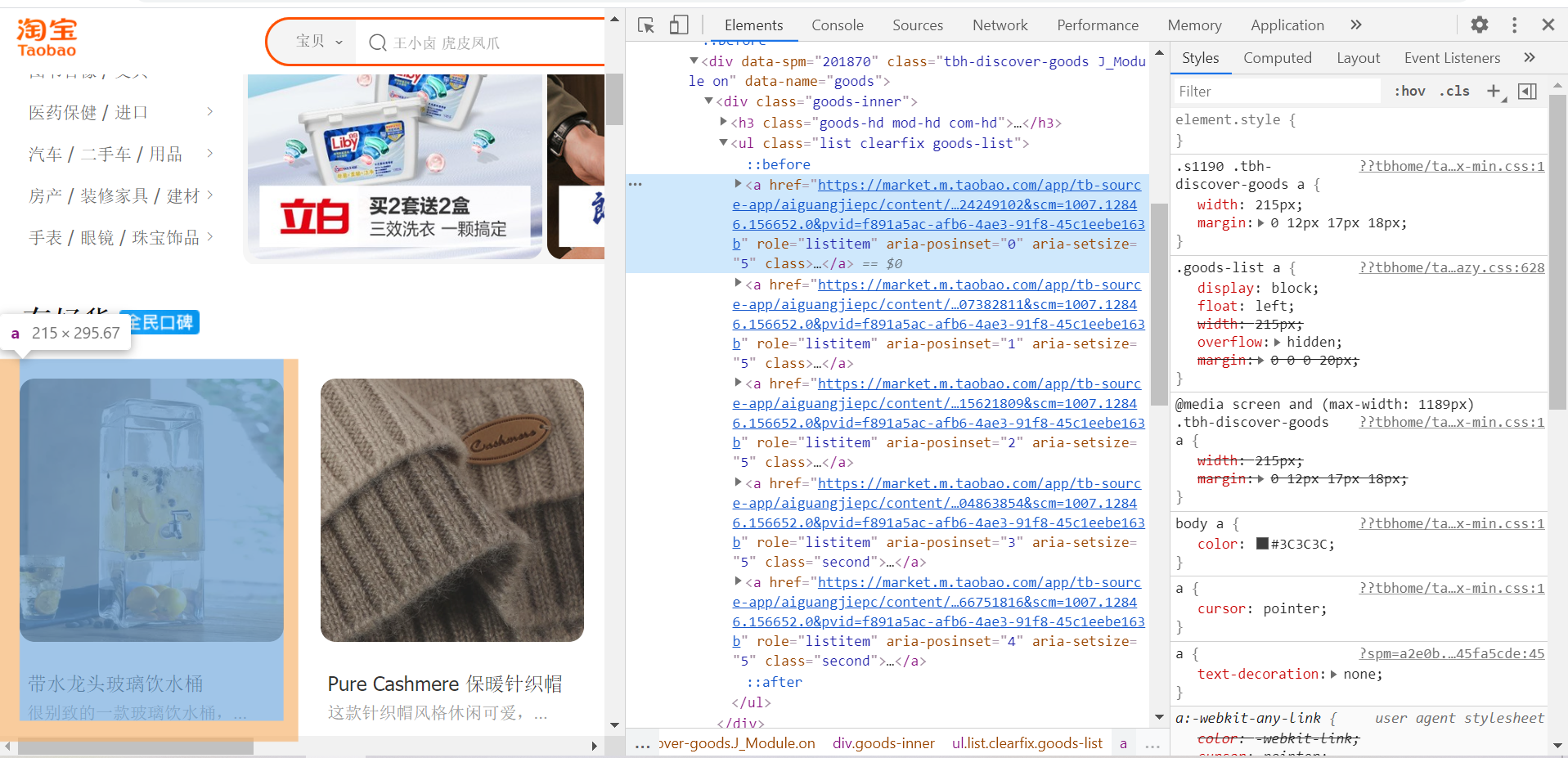
URL一般由三部组成：1、协议（或称为服务方式）。2、存有该资源的主机IP地址（有时也包含端口号）。3、主机资源的具体地址，如：目录和文件等。

URL是URI的子集，每个URL都是URI，但不是每个URI都是URL。URI还不包括一个子类叫做URN，它的全称为Universal Resource Name，即统一资源名称。URN只命名资源而不指定如何定位资源。

# 2、超文本

什么是超文本：我们在浏览器里看到的网页就是超文本解析而成的，其网页源代码是一系列HTML代码，里面包含一系列标签。浏览器解析这些标签后，便形成我们平时看到的网页，而网页的源代码HTML就可以称作为（超文本）

如何查看网页源代码：打开浏览器🡪右击任意地方并选择“检查”或者（直接按F12），即可打开浏览器开发工具，这时在Elements选项卡即可看到当前网页的源代码，这些代码都是超文本。



# 3、HTTP和HTTPS