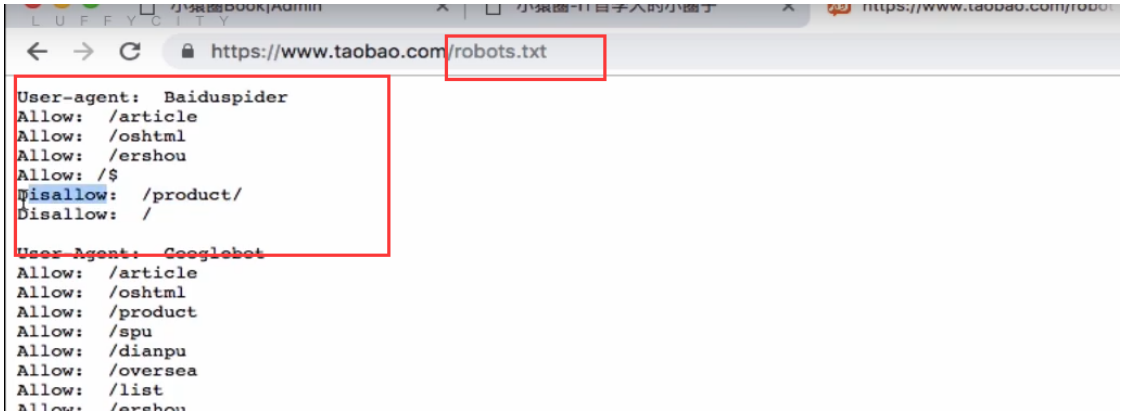
6/18

笔记本: reptile_draft
创建时间: 2021/6/18 13:35 **更新时间:** 2021/6/18 13:56
作者: 134exetj717

爬虫在使用场景中的分类:

- 通用爬虫:
  - 提取系统重要组成部分，抓取的是一整张页面数据。
- 聚焦爬虫:
  - 是建立在通用爬虫的基础之上，抓取的是页面中特定的局部内容。
- 增量式爬虫:
  - 检测网站中数据更新的情况，只会抓取网站中最新更新出来的数据。

robots 协议:
君子协议。规定了网站中哪些内容可以被爬取，哪些不可以。



htpp 协议:

- 概念:就是服务器和客户端进行数据交互的一种形式。

常用请求头信息:

- user-agent:请求载体的身体标识
- connection:请求完毕后，是断开连接还是保持连接

常用响应头信息:

- Content-Type:服务器响应回客户端的数据类型

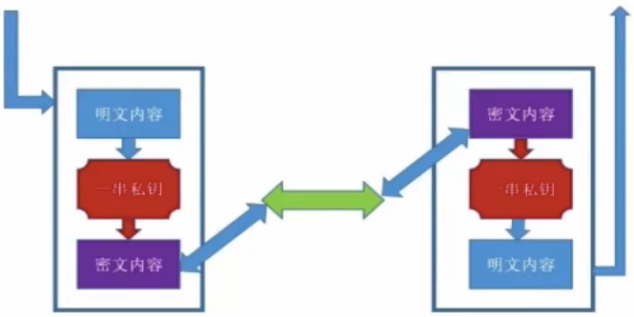
https协议:

- 安全的超文本传输协议(数据是否加密)

数据加密方式:

- 对称密钥加密
- 非对称密钥加密
- 证书密钥加密

公钥加密原理图



这里的A加密勇的的密钥是可以是A的私钥也可以是B的公钥
如果加密时用的是A的私钥，B解密就需要用A的公钥，反之亦然

- 非对称秘钥加密：“非对称加密”使用的时候有两把锁，一把叫做“私有密钥”，一把是“公开密钥”，使用非对象加密的加密方式的时候，服务器首先告诉客户端按照自己给定的公开密钥进行加密处理，客户端按照公开密钥加密以后，服务器接受到信息再通过自己的私有密钥进行解密，这样做的好处就是解密的钥匙根本就不会进行传输，因此也就避免了被挟持的风险。就算公开密钥被窃听者拿到了，它也很难进行解密，因为解密过程是对离散对数求值，这可不是轻而易举就能做到的事。以下是非对称加密的原理图：

  - 但是非对称秘钥加密技术也存在如下缺点：
    - 第一个是：如何保证接收端向发送端发出公开秘钥的时候，发送端确保收到的是预先要发送的，而不会被挟持。只要是发送密钥，就有可能有被挟持的风险。
    - 第二个是：非对称加密的方式效率比较低，它处理起来更为复杂，通信过程中使用就有一定的效率问题而影响通信速度

- **证书秘钥加密：**在上面我们讲了非对称加密的缺点，其中第一个就是公钥很可能存在被挟持的情况，无法保证客户端收到的公开密钥就是服务器发行的公开密钥。此时就引出了公开密钥证书机制。数字证书认证机构是客户端与服务器都可信赖的第三方机构。证书的具体传播过程如下：
    - 服务器的开发者携带公开密钥，向数字证书认证机构提出公开密钥的申请，数字证书认证机构在认清申请者的身份，审核通过以后，会对开发者申请的公开密钥做数字签名，然后分配这个已签名的公开密钥，并将密钥放在证书里面，绑定在一起
    - 服务器将这份数字证书发送给客户端，因为客户端也认可证书机构，客户端可以通过数字证书中的数字签名来验证公钥的真伪，来确保服务器传过来的公开密钥是真实的。一般情况下，证书的数字签名是很难被伪造的，这取决于认证机构的公信力。一旦确认信息无误之后，客户端就会通过公钥对报文进行加密发送，服务器接收到以后用自己的私钥进行解密。