爬虫概述

什么是爬虫?

不知道各位是否遇到过这样的需求. 就是我们总是希望能够保存互联网上的一些重要的数据信息为己所用.

比如,

- 在浏览到一些优秀的让人血脉喷张的图片时. 总想保存起来留为日后做桌面上的壁纸
- 在浏览到一些重要的数据时(各行各业), 希望保留下来日后为自己进行各种销售行为增光添彩
- 在浏览到一些奇奇怪怪的劲爆视频时, 希望保存在硬盘里供日后慢慢品鉴
- 在浏览到一些十分优秀的歌声曲目时, 希望保存下来供我们在烦闷的生活中增添一份精彩

那么恭喜你. 本课程将十分的适合于你. 因为爬虫就是通过编写程序来爬取互联网上的优秀资源(图片, 音频, 视频, 数据)

爬虫和Python

爬虫一定要用Python么? 非也~ 用Java也行, C也可以. 请各位记住, 编程语言只是工具. 抓到数据是你的目的. 用什么工具去达到你的目的都是可以的. 和吃饭一样, 可以用叉子也可以用筷子, 最终的结果都是你能吃到饭. 那为什么大多数人喜欢用Python呢? 答案: 因为Python写爬虫简单. 不理解? 问: 为什么吃米饭不用刀叉? 用筷子? 因为简单! 好用!

而Python是众多编程语言中, 小白上手最快, 语法最简单. 更重要的是, 这货有非常多的关于爬虫能用到的第三方支持库. 说直白点儿. 就是你用筷子吃饭, 我还附送你一个佣人. 帮你吃! 这样吃的是不是更爽了. 更容易了~

爬虫合法么?

首先, 爬虫在法律上是不被禁止的. 也就是说法律是允许爬虫存在的. 但是, 爬虫也具有违法风险的. 就像菜刀一样, 法律是允许菜刀的存在的. 但是你要是用来砍人, 那对不起. 没人惯着你. 就像王欣说过的, 技术是无罪的. 主要看你用它来干嘛. 比方说有些人就利用爬虫+一些黑客技术每秒钟对着bilibili撸上十万八千次. 那这个肯定是不被允许的.

爬虫分为善意的爬虫和恶意的爬虫

- 善意的爬虫, 不破坏被爬取的网站的资源(正常访问, 一般频率不高, 不窃取用户隐私)
- 恶意的爬虫, 影响网站的正常运营(抢票, 秒杀, 疯狂solo网站资源造成网站宕机)

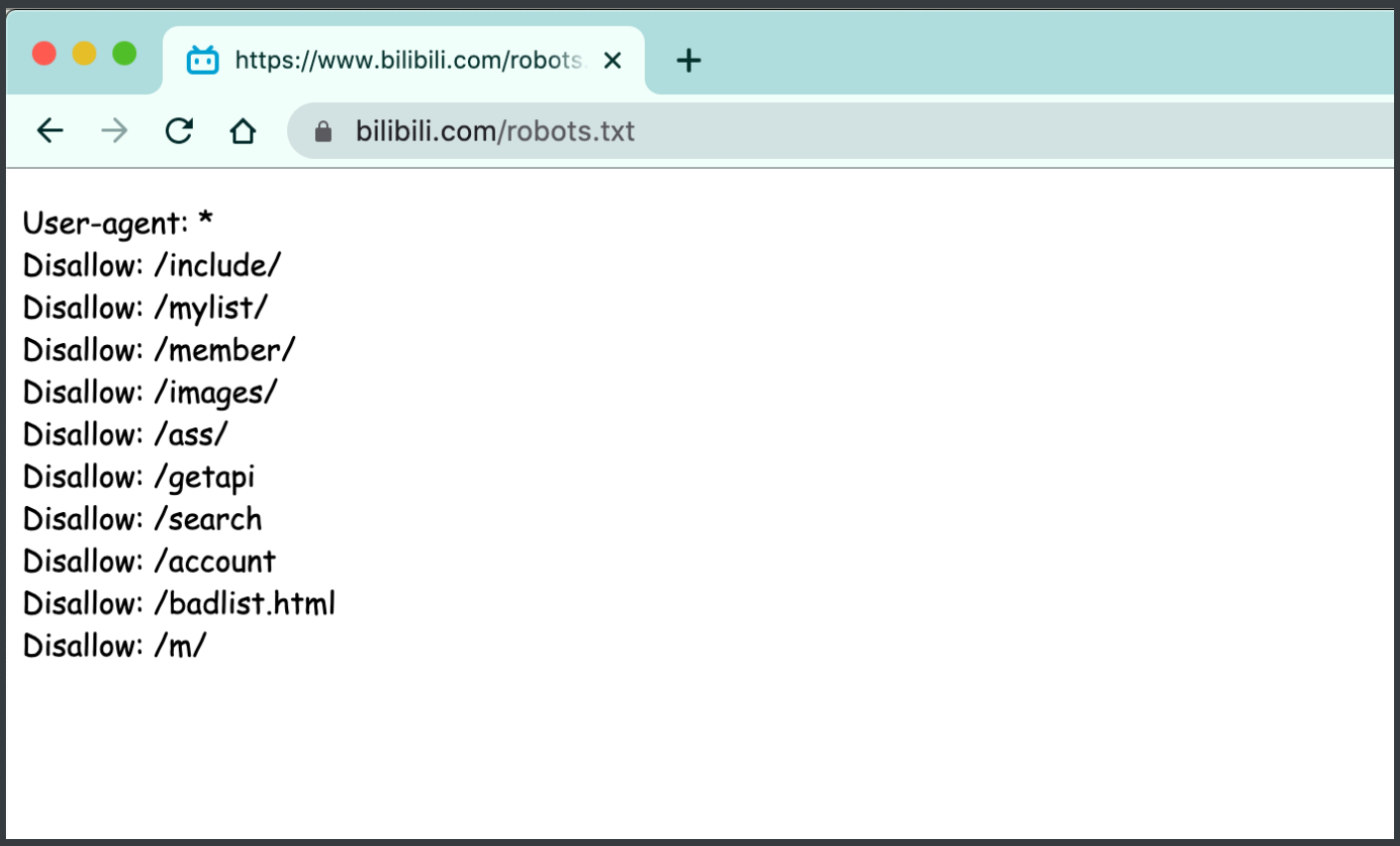
综上, 为了避免进🍊 我们还是要安分守己. 时常优化自己的爬虫程序避免干扰到网站的正常运行. 并且在使用爬取到的数据时,发现涉及到用户隐私和商业机密等敏感内容时, 一定要及时终止爬取和传播

爬虫的矛与盾

反爬机制 门户网站，可以通过制定相应的策略或者技术手段，防止爬虫程序进行网站数据的爬取。

反反爬策略 爬虫程序可以通过制定相关的策略或者技术手段，破解门户网站中具备的反爬机制，从而可以获取门户网站中相关的数据。

robots.txt协议： 君子协议。规定了网站中哪些数据可以被爬虫爬取哪些数据不可以被爬取。


https://www.bilibili.com/robots
bilibili.com/robots.txt
User-agent: *
Disallow: /include/
Disallow: /mylist/
Disallow: /member/
Disallow: /images/
Disallow: /ass/
Disallow: /getapi
Disallow: /search
Disallow: /account
Disallow: /badlist.html
Disallow: /m/