1.http:(1)当用户在地址输入了网址 发送网络请求的过程是什么

(2)http的请求方式

get请求

(1)比较便捷

缺点:不安全:明文

参数的长度有限制

post请求

(1)比较安全

(2)数据整体没有限制

(3)上传文件

put(不完全的)

delete(删除一些信息)

head(请求头)

发送网络请求(需要带一定的数据给服务器不带数据也可以)

请求头里面requestheader

返回数据:response

(1)Accept:文本的格式

(2)Accept-Encoding:编码格式

(3)Connection:长链接 短链接

(4)Cookie:验证用的

(5)Host:域名

(6)Referer:标志从哪个页面跳转过来的

(7)User-Agent:浏览器和用户的信息

2.爬虫入门:使用代码模拟用户 批量的发送网络请求 批量的获取数据

(1)爬虫的价值:

1.买卖数据(高端的领域价格特别贵)

2.数据分析:出分析报告

3.流量

4.指数阿里指数,百度指数

(3)合法性:灰色产业

政府没有法律规定爬虫是违法的,也没有法律规定爬虫是合法的

公司概念:公司让你爬数据库(窃取商业机密)责任在公司

(4)爬虫可以爬取所有东西?(不是)爬虫只能爬取用户能访问到的数据

爱奇艺的视频(vip非vip)

1.普通用户 只能看非vip 爬取非vip的的视频

2.vip 爬取vip的视频

3.普通用户想要爬取vip视频(黑客)

爬虫的分类:(1)通用爬虫

1.使用搜索引擎:百度 谷歌 360 雅虎 搜狗

优势:开放性 速度快

劣势:目标不明确

返回内容:基本上%90是用户不需要的

不清楚用户的需求在哪里

(2)聚焦爬虫(学习)

1.目标明确

2.对用户的需求非常精准

3.返回的内容很固定

增量式:翻页:从第一页请求到最后一页

Deep 深度爬虫:静态数据:html css

动态数据:js代码,加密的js

robots:是否允许其他爬虫(通用爬虫)爬取某些内容

聚焦爬虫不遵守robots

爬虫和反扒做斗争:资源对等 胜利的永远是爬虫

爬虫的工作原理:

1.缺人你抓取目标的url是哪一个(找)

2.使用python代码发送请求获取数据(java Go)

3.解析获取到的数据(精确数据)

(1)找到新的目标(url)回到第一步(自动化)

4.数据持久化

python3(原生提供的模块):urlib.rquest:

(1)urlopen :

1.返回response对象

2.response.read()

3.bytes.decode("utf-8")

(2)get:传参

1.汉字报错 :解释器ascii没有汉字,url汉字转码

(3)post

(4)handle处理器的自定义

(5)urlError

python(原生提供的):urlib2

接下来将的知识点:

5.request(第三方)

6.数据解析:xpath bs4

7.数据存储