

# 网络爬虫的“盗亦有道”

WS02

---



嵩天

[www.python123.org](http://www.python123.org)

# The Website is the API ...



## Requests

自动爬取HTML页面  
自动网络请求提交

## robots.txt

网络爬虫排除标准



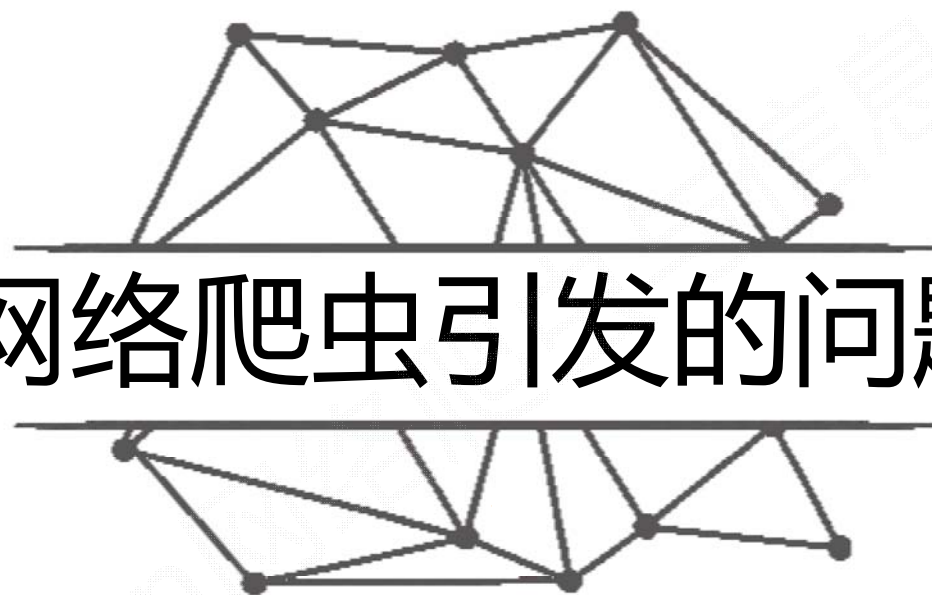
掌握定向网络数据爬取和网页解析的基本能力

# Python网络爬虫与信息提取

**python**  
弹指之间 · 享受创新

04X -Tian

# 网络爬虫引发的问题



# 网络爬虫的尺寸

小规模，数据量小

爬取速度不敏感

Requests库

>90%

爬取网页 玩转网页

中规模，数据规模较大

爬取速度敏感

Scrapy库

爬取网站 爬取系列网站

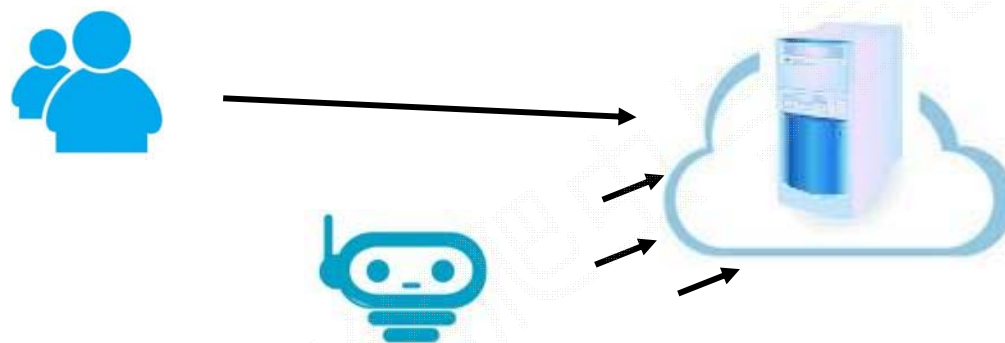
大规模，搜索引擎

爬取速度关键

定制开发

爬取全网

# 网络爬虫的“性能骚扰”



Web服务器默认接收人类访问

受限于编写水平和目的，网络爬虫将会为Web服务器带来巨大的资源开销

# 网络爬虫的法律风险



服务器上的数据有产权归属

网络爬虫获取数据后牟利将带来法律风险

# 网络爬虫的隐私泄露



网络爬虫可能具备突破简单访问控制的能力，获得被保护数据  
从而泄露个人隐私

# 网络爬虫引发的问题

性能骚扰

法律风险

隐私泄露





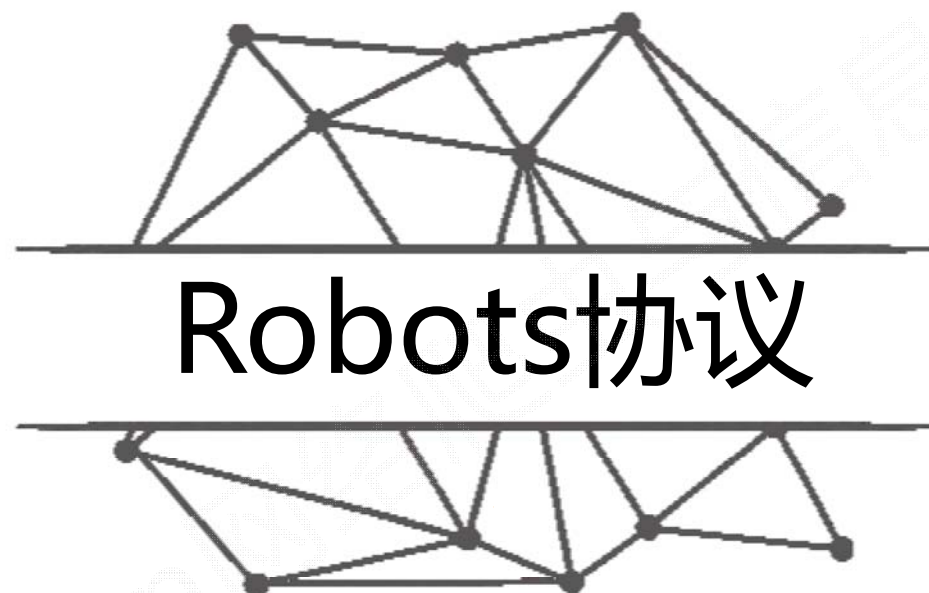
# 网络爬虫的限制

- 来源审查：判断User-Agent进行限制

检查来访HTTP协议头的User-Agent域，只响应浏览器或友好爬虫的访问

- 发布公告：Robots协议

告知所有爬虫网站的爬取策略，要求爬虫遵守



Robots协议

# Robots协议

Robots Exclusion Standard , 网络爬虫排除标准

作用：

网站告知网络爬虫哪些页面可以抓取，哪些不行

形式：

在网站根目录下的robots.txt文件

# 案例：京东的Robots协议

<https://www.jd.com/robots.txt>

```
User-agent: *  
Disallow: /?*  
Disallow: /pop/*.html  
Disallow: /pinpai/*.html?*  
User-agent: EtaoSpider  
Disallow: /  
User-agent: HuihuiSpider  
Disallow: /  
User-agent: GwdangSpider  
Disallow: /  
User-agent: WochachaSpider  
Disallow: /
```

# 注释，\*代表所有，/代表根目录

```
User-agent: *  
Disallow: /
```

Robots协议基本语法

# 案例：真实的Robots协议

<http://www.baidu.com/robots.txt>

<http://news.sina.com.cn/robots.txt>

<http://www.qq.com/robots.txt>

<http://news.qq.com/robots.txt>

<http://www.moe.edu.cn/robots.txt> （无robots协议）



# Robots协议的遵守方式

实际操作中，该如何遵守Robots协议？

# Robots协议的使用

网络爬虫：

自动或人工识别robots.txt，再进行内容爬取

约束性：

Robots协议是建议但非约束性，网络爬虫可以不遵守，但存在法律风险



# 对Robots协议的理解

访问量很小：可以遵守

访问量较大：建议遵守

非商业且偶尔：建议遵守

商业利益：必须遵守

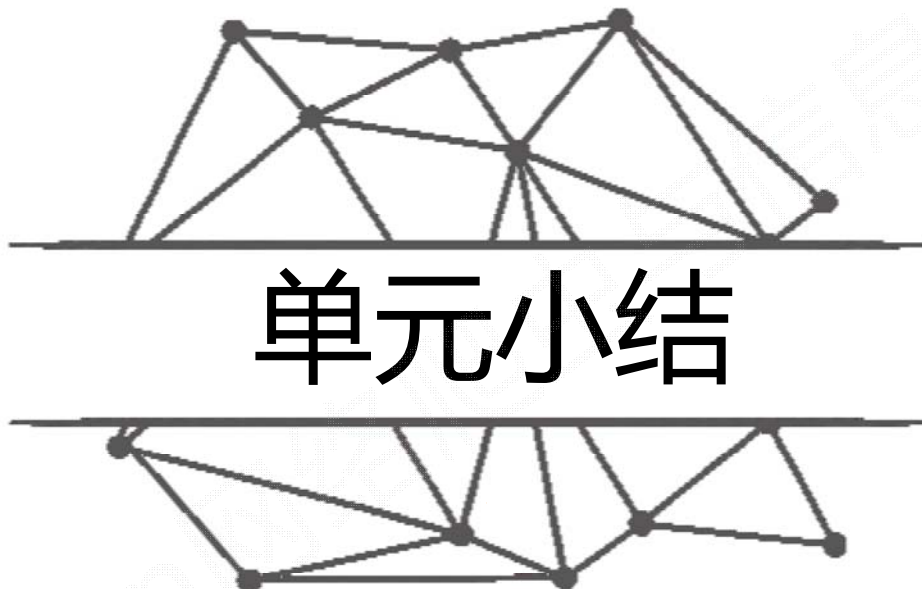
必须遵守

爬取网页 玩转网页

爬取网站 爬取系列网站

爬取全网

**原则**：类人行为可不参考Robots协议



# 单元小结

# 网络爬虫 “盗亦有道”

# 注释，\*代表所有，/代表根目录

User-agent: \*

Disallow: /

Robots协议的使用原则

Robots协议基本语法