爬虫入门

I. 爬虫应用领域

A. 企业的资源与价格竞争

B. 个人用途的商品最低价

C. 爬取的网站如果提供全面的API接口，则优先考虑使用

II. 爬虫的合法性

A. 如果爬取的数据用户个人使用的话，则不会存在任何相关法律问题

B. 如果爬取的数据用户商用或爬取原创数据等，通常都是会有版权限制的

III 网站背景调研

A. 检查robots.txt文件

遵从该文件可以有效的防止爬虫被封禁 (robotparser)

B. 检查sitemap文件

能够最快的获取网站的网站地图 (re)

C. 估算网站页面数量

考虑是否使用多线程/多进程/分布式下载 (site:)

D. 识别网站使用技术

判断网站是静态还是动态网站，用于分析爬取难度和方法 (builtwith)

E. 判断网站的所有者

查找所有者是否发布该网站的相关可用信息 (python-whois)

IV. 爬虫基本要点

A. 请求模块设计

a. 下载重试

b. 下载限速

c. 解析reobts文件

d. 设置用户代理

e. 网页下载缓存

防止网页重复下载

f. 避免下载无限递归

设置下载深度限制

B. 下载网页方式

a. 网站提供了API接口

b. 网站提供的站点地图

c. 通过遍历网页ID号

d. 通过追踪网页链接

C. 使用爬虫框架

a. PySpider

b. Scrapy

V. 爬取思维总结

A. 通过分析需要爬取网站的页面数量级来决定适用何种下载方式和组织架构

B. 针对不同的下载方式使用多种方案爬取来防止网站封禁，并且要不断优化爬取性能和效率