# Python爬虫之解析robots.txt文件

网络爬虫以获取网站的数据为目的，它带来了一些问题：

1. 网络爬虫不断的访问服务器，增加了服务器的性能负担。
2. 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。
3. 网络爬虫可以较为轻松的突破安全防控能力低的网站，从而非法获取被保护数据，比如个人隐私。

通常，网站的根目录都会防止robots.txt文件，告诉爬虫程序哪些页面数据可以爬去，哪些不能。

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。因此，从规避法律风险的角度出发，建议遵守。

## robots.txt文件示例

下面是京东的robots.txt文件。

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: /

User-agent: \* 表示任何浏览器代理。

Disallow: /?\* 表示任何带?的路径都不允许爬取。

/pop/\*.html 表示pop路径下的任何html文件不允许爬去。

User-agent: EtaoSpider

Disallow: /

上面两句则表示用户代理为EtaoSpider的爬虫不允许爬去网站任何数据。

## 利用robotparser解析robots.txt文件

[**https://www.jianshu.com/robots.txt**](https://www.jianshu.com/robots.txt)

# See http://www.robotstxt.org/wc/norobots.html for documentation on how to use the robots.txt file

#

# To ban all spiders from the entire site uncomment the next two lines:

User-agent: \*

Disallow: /search

Disallow: /convos/

Disallow: /notes/

Disallow: /admin/

Disallow: /adm/

Disallow: /p/0826cf4692f9

Disallow: /p/d8b31d20a867

Disallow: /collections/\*/recommended\_authors

Disallow: /trial/\*

Disallow: /keyword\_notes

Disallow: /stats-2017/\*

User-agent: trendkite-akashic-crawler

Request-rate: 1/2 # load 1 page per 2 seconds

Crawl-delay: 60

User-agent: YisouSpider

Request-rate: 1/10 # load 1 page per 10 seconds

Crawl-delay: 60

User-agent: Cliqzbot

Disallow: /

User-agent: Googlebot

Request-rate: 2/1 # load 2 page per 1 seconds

Crawl-delay: 10

[**http://www.wave12.com/robots.txt**](http://www.wave12.com/robots.txt)

User-agent: \*

Disallow: /bigfish/news/

表明除了/bigfish/news/路径下的网页，其他都可以爬取。

**import** urllib.robotparser  
  
rp = urllib.robotparser.RobotFileParser()  
rp.set\_url(**'http://www.wave12.com/robots.txt'**)  
rp.read()  
user\_agent = **'wsCrawler'**url = **'http://www.wave12.com/bigfish/doc/doc\_list'**ret = rp.can\_fetch(user\_agent, url)  
print(ret)  
  
url = **'http://www.wave12.com/bigfish/news/newsList?cateId=1'**ret = rp.can\_fetch(user\_agent, url)  
print(ret)

## 爬虫程序开发建议

* 爬取互联网公开数据
* 尽量放慢你的速度
* 尽量遵循robots协议
* 不要用于商业用途
* 不要公布爬虫程序与数据