第一章：

http ://www.robotstxt .org

1、 robots.txt可以让爬虫了解爬取该网站时存在哪些限制。

2、robots.txts

关于robots.txt协议的更多信息

http : // www. robotstxt . org

#section 1

#robots.txt文件禁止用户代理为BadCrawler的爬虫爬取该网 站,不过这种写法可能无法起 到应有的作用，因为恶意爬虫根本不会遵从robots. txt 的要求

User-agent:BadCrawler

Disallow:

#sect ion 2 规定,无论使用哪种用户代理,都应该在两次下载请求之间给出5秒的抓取延迟,我们需要遵从该建议以避免服务器过载

#section 2

User-agent: \*

Crawl-delay:5

Disallow: trap /trap链接,用于封禁那些爬取了不允许链接的恶意爬虫

#sect ion 3定义了一个Sitemap文件

#section 3

Sitemap: <http://example.webscraping.com/sitemap.xml>

3、检查网站地图

获取网站地图标准的定义：

<https://www.sitemaps.org/protocol.html>

4、估算网站大小

目标网站的大小会影响我们如何进行爬取。

通过Goo gle搜索的s ite关键词过滤域名结果,从而获取网站大小

例如：搜索site: example.webscraping. com/view

5、识别网站所用技术

有一个十分有用的工具可以检查网站构建的技术类型----builtwith模块