服务器对爬虫的限制：

- 来源审查：判断User-Agent进行限制

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

- 对服务器的维护者会有一定的技术要求

- 发布公告：Robots协议

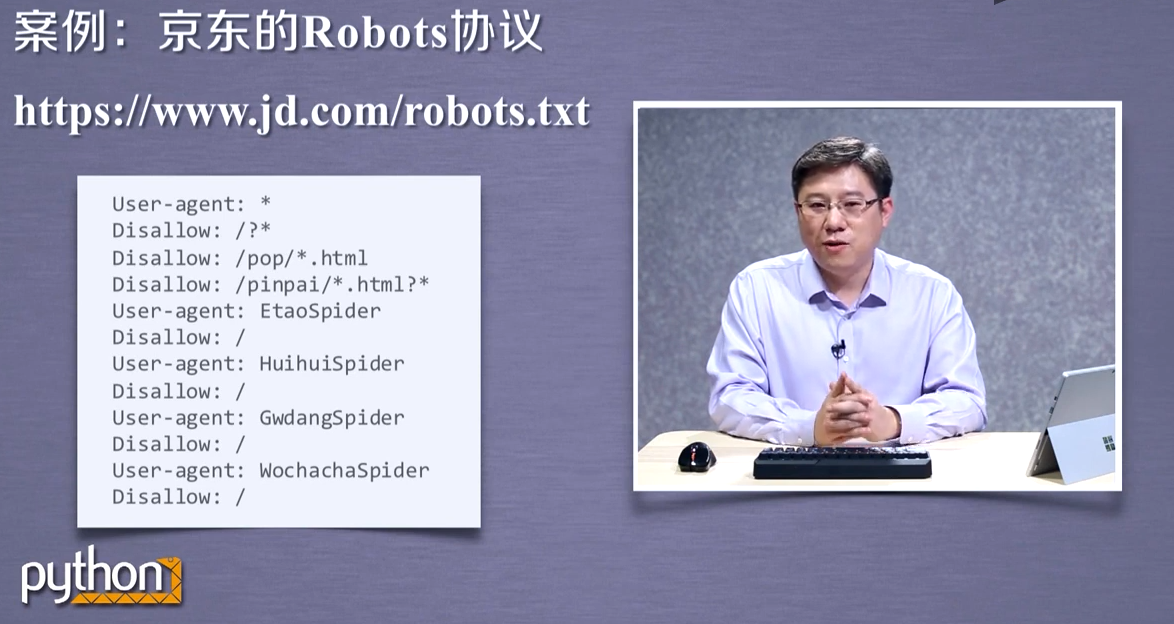
- 告知所有爬虫网站的爬取策略，要求爬虫遵守

Robots协议：

Robots Exclusion Standard网络爬虫排除标准

作用：网络告知网络爬虫那些页面可以抓取，那些不行

形式：在网络根目录下的robots.txt文件（文件中写了哪些目录是可以爬取，哪些不行）



Robots基本语法：

# \* 代表所以, / 代表根目录

User-agent: \*

Disallow: /

其他案例：

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

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

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

<http://www.moe.edu.con/robots.txt> （中国教育局网站，无robots协议，代表任何内容可以被任何爬虫自由爬取）

注意：Robots协议一定是放在网站的根目录下

马来西亚教育局网站的robots协议 XD

User-agent: \*

Allow: /\*.js\*

Allow: /\*.css\*

Allow: /\*.png\*

Allow: /\*.jpg\*

Allow: /\*.gif\*

Disallow: /administrator/

Disallow: /bin/

Disallow: /cache/

Disallow: /cli/

Disallow: /includes/

Disallow: /installation/

Disallow: /language/

Disallow: /layouts/

Disallow: /libraries/

Disallow: /logs/

Disallow: /tmp/