前面的例子显示了一个 robots.txt 文件，这个文件允许机器人 Slurp 和 Webcrawler 访问除了 private 子目录下那些文件之外所有的文件。这个文件还会阻止所有其他机器人访问那个站点上的任何内容。

我们来看看 User-Agent、Disallow 和 Allow 行。

1. User-Agent行

每个机器人记录都以一个或多个下列形式的 User-Agent 行开始：

```
User-Agent: <robot-name>
```

或


```
User-Agent: *
```

在机器人 HTTP GET 请求的 User-Agent 首部中发送（由机器人实现者选择的）机器人名。

机器人处理 robots.txt 文件时，它所遵循的记录必须符合下列规则之一：

- 第一个 <robot-name> 是机器人名的大小写无关的子字符串；
- 第一个 <robot-name> 为“*”。

如果机器人无法找到与其名字相匹配的 User-Agent 行，而且也无法找到通配的 User-Agent:* 行，就是没有记录与之匹配，访问不受限。

由于机器人名是与大小写无关的子字符串进行匹配，所以要小心不要匹配错了。比如，User-Agent:bot 就与名为 Bot、Robot、Bottom-Feeder、Spambot 和 Dont-Bother-Me 的所有机器人相匹配。

2. Disallow和Allow行

Disallow 和 Allow 行紧跟在机器人排斥记录的 User-Agent 行之后。用来说明显式禁止或显式允许特定机器人使用哪些 URL 路径。

机器人必须将期望访问的 URL 按序与排斥记录中所有的 Disallow 和 Allow 规则进行匹配。使用找到的第一个匹配项。如果没有找到匹配项，就说明允许使用这个 URL。[21]

注 21：总是应该允许访问 robots.txt 的 URL，它一定不能出现在 Allow/Disallow 规则中。