2. 响应码

很多 Web 站点都没有 robots.txt 资源，但机器人并不知道这一点。它必须尝试着从每个站点上获取 robots.txt 资源。机器人会根据对 robots.txt 检索的结果采取不同的行动。

- 如果服务器以一个成功状态（HTTP 状态码 2XX）为响应，机器人就必须对内容进行解析，并使用排斥规则从那个站点上获取内容。
- 如果服务器响应说明资源并不存在（HTTP 状态码 404），机器人就可以认为服务器没有激活任何排斥规则，对此站点的访问不受 robots.txt 的限制。
- 如果服务器响应说明有访问限制（HTTP 状态码 401 或 403），机器人就应该认为对此站点的访问是完全受限的。
- 如果请求尝试的结果是临时故障（HTTP 状态码 503），机器人就应该推迟对此站点的访问，直到可以获取该资源为止。
- 如果服务器响应说明是重定向（HTTP 状态码 3XX），机器人就应该跟着重定向，直到找到资源为止。

231

9.4.3 robots.txt文件的格式

robots.txt 文件采用了非常简单的，面向行的语法。robots.txt 文件中有三种类型的行：空行、注释行和规则行。规则行看起来就像 HTTP 首部（<Field>:<value>）一样，用于模式匹配。比如：

```
# this robots.txt file allows Slurp & Webcrawler to crawl
# the public parts of our site, but no other robots...

User-Agent: slurp
User-Agent: webcrawler
Disallow: /private

User-Agent: *
Disallow:
```

robots.txt 文件中的行可以从逻辑上划分成“记录”。每条记录都为一组特定的机器人描述了一组排斥规则。通过这种方式，可以为不同的机器人使用不同的排斥规则。

每条记录中都包含了一组规则行，由一个空行或文件结束符终止。记录以一个或多个 User-Agent 行开始，说明哪些机器人会受此记录的影响，后面跟着一些 Disallow 和 Allow 行，用来说明这些机器人可以访问哪些 URL。[20]

注 20：出于实际应用的原因，机器人软件应该很强壮，可以灵活地使用行结束符。应该支持 CR、LF 和 CRLF。