攻防世界 web2 robots协议

原创

maixinbaogu 于 2021-06-15 10:01:06 发布 157 收藏 1

分类专栏： ctf

本文链接：https://blog.csdn.net/maixinbaogu/article/details/117918482

攻防世界 web基础 robots

robots是网站跟爬虫间的协议，用简单直接的txt格式文本方式告诉对应的爬虫被允许的权限，也就是说robots.txt是搜索引擎中访问网站的时候要查看的第一个文件。当一个搜索蜘蛛访问一个站点时，它会首先检查该站点根目录下是否存在robots.txt，如果存在，搜索机器人就会按照该文件中的内容来确定访问的范围;如果该文件不存在，所有的搜索蜘蛛将能够访问网站上所有没有被口令保护的页面。

如果创建一个纯文本文件robots.txt，在这个文件中声明该网站中不想被robot访问的部分，这样，该网站的部分或全部内容就可以**不被搜索引擎收录了**，或者指定搜索引擎只收录指定的内容。
所以遇到这样的文件首先在根目录中加上robots.txt按访问文件中的内容
下面是VeryCMS里的robots.txt文件:

User-agent: *

该项的值用于描述搜索引擎robot的名字，在"robots.txt"文件中，如果有多条User-agent记录说明有多个robot会受到该协议的限制，对该文件来说，至少要有一条User-agent记录。如果该项的值设为*，则该协议对任何机器人均有效，在"robots.txt"文件中，"User-agent:*"这样的记录只能有一条。

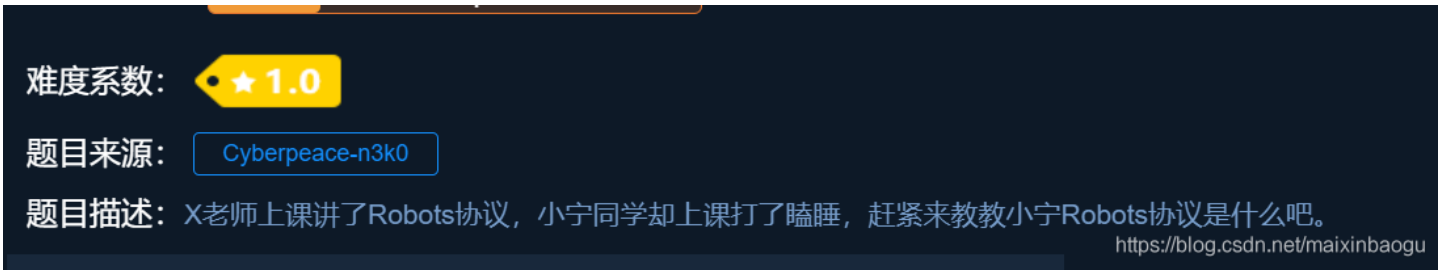
Disallow:

该项的值用于描述不希望被访问到的一个URL，这个URL可以是一条完整的路径，也可以是部分的，任何以Disallow开头的URL均不会被robot访问到。例如"Disallow:/help"对/help.html 和/help/index.html都不允许搜索引擎访问，而"Disallow:/help/“则允许robot访问/help.html，而不能访问/help/index.html。任何一条Disallow记录为空，说明该网站的所有部分都允许被访问，在”/robots.txt"文件中，至少要有一条Disallow记录。如果"/robots.txt"是一个空文件，则对于所有的搜索引擎robot，该网站都是开放的。

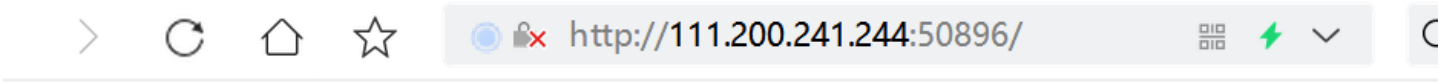
Allow:

该项的值用于描述希望被访问的一组URL，与Disallow项相似，这个值可以是一条完整的路径，也可以是路径的前缀，以Allow项的值开头的URL是允许robot访问的。例如"Allow:/hibaidu"允许robot访问/hibaidu.htm、/hibaiducom.html、/hibaidu/com.html。一个网站的所有URL默认是Allow的，所以Allow通常与Disallow搭配使用，实现允许访问一部分网页同时禁止访问其它所有URL的功能。

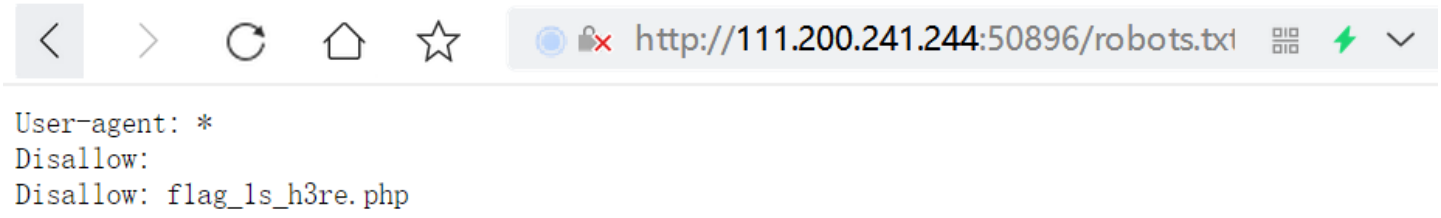
看题



根据题目提示可知道这里他需要用到robots协议，那么先打开网页



为空，说明被隐藏了，遇到这样的文件首先在**根目录中加上robots.txt按访问文件中的内容**



可以发现它禁止了网页访问一个php文件
在根目录输入php文件则可获得flag