VISIBILIDAD WEB UNAM

ARCHIVO ROBOTS

DGTIC

MAYO 2020

Archivo robots

Un archivo robots.txt proporciona a los buscadores, la información sobre cuáles páginas y archivos pueden ser o no solicitados de un sitio web. Se utiliza principalmente para evitar la sobrecarga de solicitudes al sitio web en cuestión.

¿Para qué sirve el robots.txt?

El archivo robots.txt sirve para gestionar el tráfico de los rastreadores a un sitio web; se utiliza también en menor medida para indicar cuáles páginas, directorios y archivos pueden ser solicitados por los rastreadores. No se utiliza para evitar que Google u otros buscadores no indexen alguna página, para esto último se recomienda utilizar la etiqueta noindex.

¿Cómo se crea un archivo robots.txt y dónde se ubica?

Un archivo robots.txt se crea a partir de una archivo de texto sin formato y se debe de ubicar en la raíz de nuestro directorio de archivos, es decir, en donde se encuentra la página principal o index.html, como https://www.unam.mx/robots.txt. Si el sitemap se ubicara en un directorio distinto a la raíz, cabe la posibilidad de que las indicaciones no apliquen a las páginas fuera de este.

Una vez creado el archivo de texto sin formato, se tiene que nombrar como robots.txt e ir agrupando las indicaciones. En este archivo se deben incluir uno o varios grupos, permitiendo o no, el acceso de un rastreador o todos, a un archivo o directorio del sitio web.

A continuación podemos observar el ejemplo más básico de un archivo robots.txt:

```
# Grupo 1

User-agent: Googlebot

Disallow: /unam-recursos/

# Grupo 2

User-agent: *

Allow: /

Sitemap: http://www.tic.unam.mx/sitemap.xml
```

En el Grupo 1 estamos indicando que el rastreador de Google no tenga acceso al directorio unam-recursos. En el Grupo 2 estamos indicando que todos los rastreadores puedan acceder a todos los directorios del sitio web. Por último se indica la ubicación del Sitemap XML del sitio en cuestión. En el siguiente enlace se puede consultar el Estándar de exclusión de robots para más ejemplos de acceso o bloqueo de directorios mediante robots.txt.

Ejemplo de archivo robots.txt en un navegador.

También se puede utilizar la metaetiqueta “robots” para indicarle a los buscadores cómo se debe indexar y mostrar a los usuarios una página en particular.

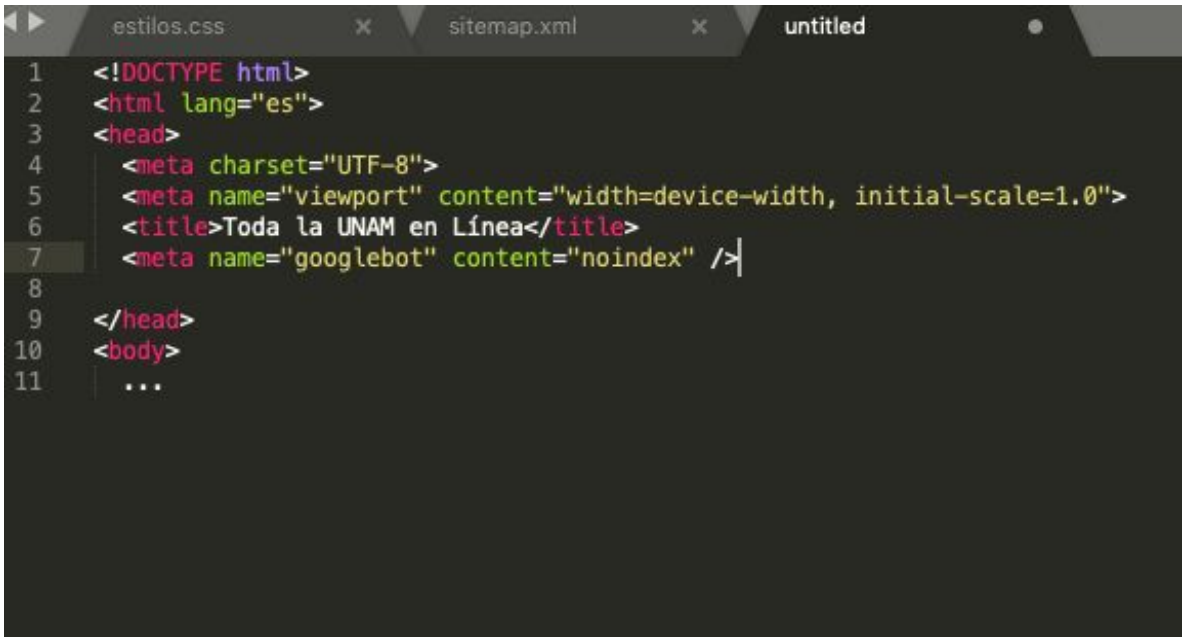
Se trata en realidad del valor “robots” o el nombre del robot en específico, que pertenece al atributo “name” de una metaetiqueta. Esta metaetiqueta robots se ubica dentro del <head> de la página en particular, como en los siguientes ejemplos:

```
<head>
<meta name="robots" content="noindex" />
</head>
```

```
<head>
<meta name="googlebot" content="noindex" />
</head>
```

En el primer ejemplo estamos indicando que ningún robot o buscador indexe la página.

En el segundo ejemplo le estamos indicando al robot de Google que no indexe la página a la que se le ha insertado la metaetiqueta. Se pueden utilizar varias metaetiquetas a la vez en la misma página.



Documento HTML con la metaetiqueta robots.txt.

Como se puede apreciar el valor “robots” cambiará si solo queremos modificar el comportamiento de algún robot en particular. En el siguiente enlace encontraremos una lista de los robots de Google.

El valor del atributo “content” cambiará de acuerdo al comportamiento que se desea para el robot que se especificó en el atributo “name”. En el siguiente enlace encontraremos una lista de los atributos que modificarán el comportamiento de los robots en una página en particular.

Consideraciones para crear un archivo robots.txt

- Es necesario tener acceso a la raíz del dominio para poder crear un archivo robots.txt.
- Es importante señalar que las instrucciones de los robots.txt son solamente indicaciones o recomendaciones. Es decir, aunque los rastreadores suelen hacerles caso, no necesariamente las deben cumplir.
- El archivo robots.txt proporciona información de índole público, por lo que para proteger información sensible es conveniente utilizar otros mecanismos, como la protección con contraseñas seguras en determinados directorios y el servidor.

Existe una herramienta de Google, denominada probador de robots.txt, que sirve para comprobar el funcionamiento.

Créditos

Coordinación del proyecto

Dra. Marcela Peñaloza Báez

Directora de Colaboración y Vinculación

Mtro. Juan Manuel Castillejos Reyes

Subdirector de Visibilidad Web

Lic. Irene G. Sánchez García

Enlace Institucional de Iniciativas de Visibilidad

Equipo de trabajo

L.C.G. Miguel Ángel Islas Flores

Autor del documento

L.I. Cristhian Eder Alavez Barrita

MATIE. Juan Manuel Castillejos Reyes

Revisores del documento