Hay ocasiones en las que queremos que los distintos buscadores no muestren en sus resultados algunas de nuestras páginas, porque tenemos información personal, por ejemplo. Para ello, vamos a ver como evitar que sus robots indexen estas páginas.
Solo contamos con un requisito, y es tener acceso a la carpeta raíz de nuestro servidor, conviene consultarlo antes con nuestro proveedor de servicios. Al final del artículo damos una solución para aquellos usuarios que no tengan este acceso.
Creamos un archivo llamado "robots.txt". En este archivo pondremos las páginas que no queremos que sean indexadas por estos robots; estos leerán este archivo lo primero, y luego indexarán las páginas restantes.

**La estructura del archivo será la siguiente:**

```
User-agent: *
Disallow: /tmp/
Disallow: /archivo.html
...
```

La primera línea la utilizamos para indicar a que buscadores nos referimos, si ponemos un asterisco, nos referimos a todos, para referirnos solo a Google, por ejemplo, pondríamos:

```
User-agent: Google
```

En las siguientes líneas escribiremos los directorios o archivos que permitimos que sean indexados o no. Vamos a ver algunos ejemplos y posibilidades:

### Que todos los robots puedan indexar todo lo que tengamos en el servidor, esto sería equivalente a no tener el archivo robots.txt:

```
User-agent: *
Disallow:
```

### Que ningún robot indexe ninguno de los contenidos:

```
User-agent: *
Disallow: /
```

### Permitir indexar todos los contenidos solo al robot de Google, y ningún contenido a los demás:

```
User-agent: Google
Disallow:
User-agent: *
Disallow: /
```

### Excluir dos directorios completos, escribiríamos una línea por cada directorio:

```
User-agent: *
Disallow: /files/
Disallow: /private/
```

## Para excluir ciertos archivos, de la misma forma que antes, escribiríamos una línea por cada archivo:

```
User-agent: *
Disallow: /private/data.html
Disallow: /private/personal.html
```

## ¿Qué hago si no tengo acceso a la carpeta raíz de mi servidor?

En este caso tenemos la posibilidad de añadir en cada página que no queremos que sea indexada la siguiente etiqueta, la cual añadiremos dentro del <head> :

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

De la misma manera que hemos indicado antes, si queremos evitar solo al robot de Google, la etiqueta sería la siguiente:

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

*Debemos tener en cuenta que esto no garantiza que nadie llegue a estas páginas. Puede haber robots con otros fines, spam o similares, que no hagan caso de estos archivos. Por lo que si tenemos información confidencial, deberíamos protegerla siempre con contraseña.*