

TASK 2 – DISCUSSION

How many ways can you find to prevent your website from being crawled or indexed by search engines?

1. 'robots.txt'

The 'robots.txt' file contains directives to search engine crawlers on what is allowed to be crawled and indexed in the website, and is placed at the root of the website directory. The directives in 'robots.txt' are generally respected by search engines, but can be ignored by some bots as well.

2. 'noindex' metatag

The `<meta name="robots" content="noindex,nofollow">` metatag is inserted into the `<head>` section of a page's HTML. This is a directive that tells search engine crawlers not to index a web page. Additionally, the "nofollow" metatag instructs search engines to not follow the links on the web page.

3. X-Robots-Tag HTTP header

An X-Robots-Tag is added to the HTTP response header of a given URL, and has the same effect as a 'noindex' metatag, but has additional options to specify various conditions for various search engines.

4. Password protection

A .htaccess file can be set up for a directory of the website to enable password protection for that directory of web pages. Password protected web pages will not be accessible to search engine crawlers.

5. Complex javascript loading of web pages

If the webpage requires the execution of complex Javascript code, most search engines will not execute it and thus will not be able to read and index the webpages.