

# SEO - Robots.txt

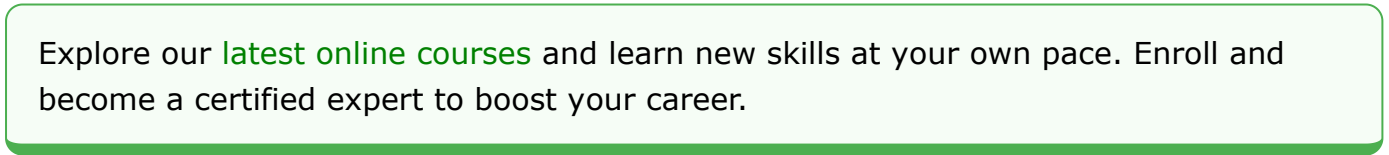
A robots.txt file contains a list of the URLs on the website that search engine spiders may access. This approach doesn't stop websites from being indexed by Google; it is primarily used to control the website from getting overburdened with searches. Use <noindex> to block Google from crawling a website's content or password-protect it to keep it hidden.

## Standard robots.txt file structure

```
User-agent: [user-agent name]  
Allow/Disallow: [URL string not to be crawled]
```

Even though a robot.txt file may consist of numerous lines of user agents and directives (such as disallows, allows, crawl-delays, etc.), these two sections combined are considered an entire robots.txt file.

Here is a real "robots.txt" file illustration



A unique user-agent is used by every search engine to identify itself. Inside the robots.txt file, you can specify particular directions for each. It's possible to use countless user-agents. However, the following few are helpful for SEO –

Platform and Browsers	User-agent Example
Google Chrome for Windows 10	Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/114.0.0.0 Safari/537.36

Mozilla for MS Windows 10	Mozilla/5.0 (Windows NT 10.0; Win64; x64; rv:109.0) Gecko/20100101 Firefox/113.0
Mozilla for macOS	Mozilla/5.0 (Macintosh; Intel Mac OS X 13.4; rv:109.0) Gecko/20100101 Firefox/113.0
Mozilla for Android	Mozilla/5.0 (Android 13; Mobile; rv:109.0) Gecko/113.0 Firefox/113.0
Safari on macOS	Mozilla/5.0 (Macintosh; Intel Mac OS X 13_4) AppleWebKit/605.1.15 (KHTML, like Gecko) Version/16.5 Safari/605.1.15
Microsoft Edge	Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/114.0.0.0 Safari/537.36 Edg/113.0.1774.57

## Note

- Remember that robots.txt treats cases very sensitively for all user-agents.
- To allocate the directives to every user-agent, utilize the Asterisk symbol (\*) wildcard.

**Here is an example of the most popular user agent bots –**

Creator	Bot
Google	Googlebot
Microsoft Bing	Bingbot
Yahoo	Slurp
Google Images	Googlebot-Image
Baidu	Baiduspider
DuckDuckGo	DuckDuckBot

As an illustration, suppose you wished to prevent all bots besides Googlebot from analyzing your website. Below is how you might go about doing it –

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

## Directives

The guidelines you intend the stated user-agents to abide by are called Directives.

## Supported Directives

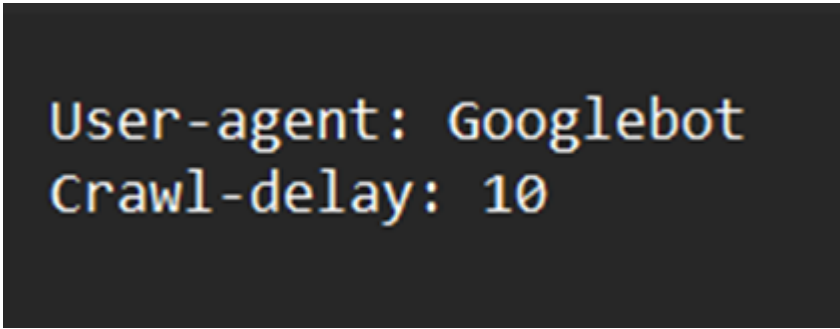
Following are the directives that Google presently recognizes and their applications –

- **Disallow** – This directive is used to prevent search engines from accessing files and webpages that are located along a particular path.
- **Allow** – This directive is used to allow or permit search engines from accessing files and webpages that are located along a particular path.
- **Sitemaps** – To tell the search engines where the sitemap(s) are located, utilize this directive. Sitemaps typically contain the developers of the site intend search engine spiders to scan and index.

## Unsupported Directives

The Google directives listed below are a few that were never officially supported and are not available anymore.

- **Crawl-delay** – The crawl interval of time was formerly specified using this directive. The crawl-delay would be set to 10, for instance, as in the following example, assuming that you intended Googlebot to remain idle for 10 seconds between every crawl action. Bing continues to support this request while Google has stopped.



```
User-agent: Googlebot  
Crawl-delay: 10
```

- **Noindex** – Google never provided any official support for this set of instructions.
- **Nofollow** – Google has never formally backed this directive.

## What is a robots.txt file's largest permitted size?

Approximately 500 kilobytes.

## A robots.txt File Is Required, right?

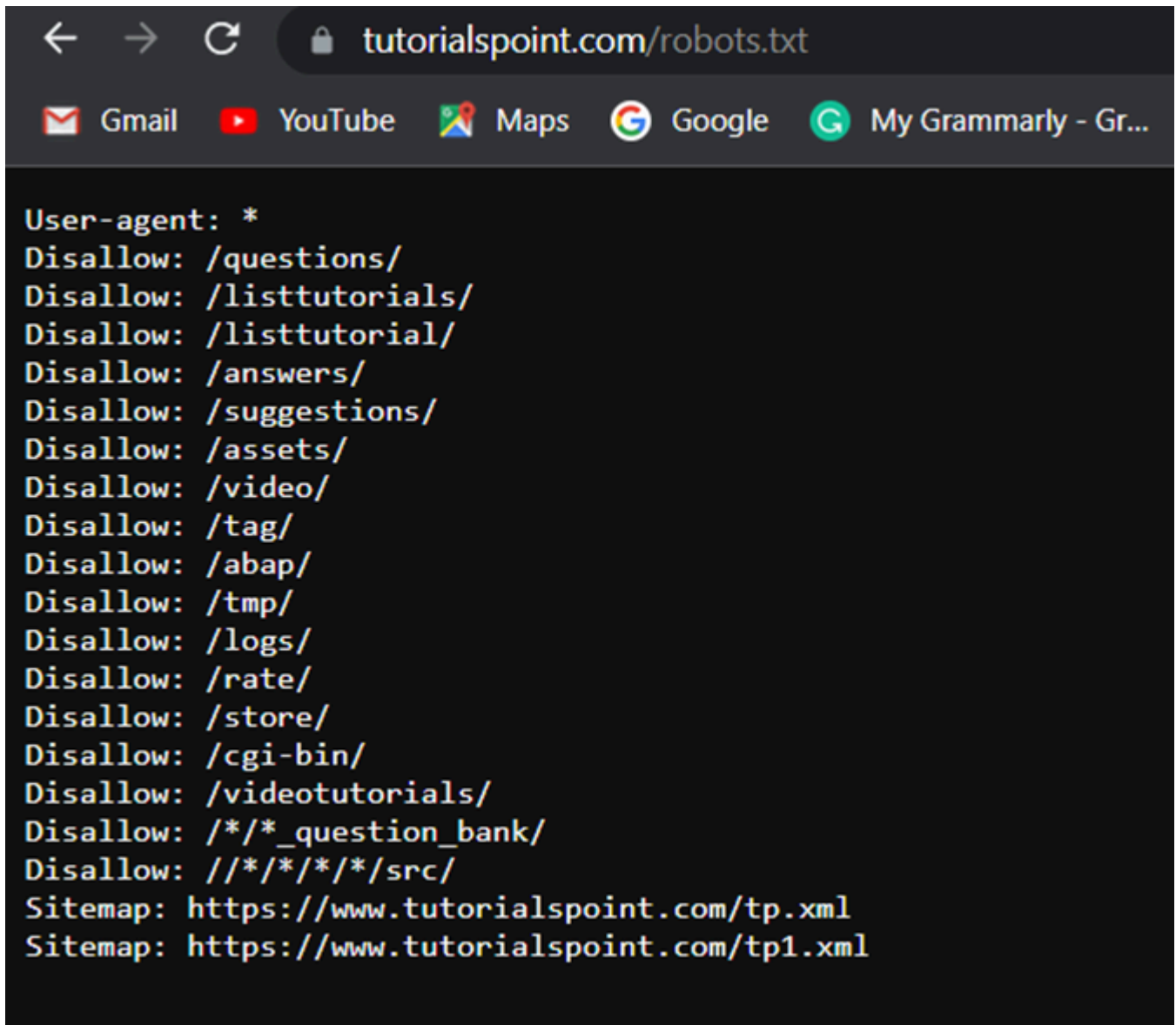
Most websites, particularly those with little traffic, don't essentially need to include a robots.txt file. There is, however, no valid excuse for not having one. With greater authority on what search engine crawlers are permitted to access the website, you can deal with issues like –

- Safeguarding private areas of the website, restricting the crawling of identical data.
- Restricting crawling of internal search results from web pages.
- Protection against server congestion and overload.
- Stop Google from spending the set crawl revenue.
- Prohibiting assets files, videos, and photos from showing up on Google results pages.

Although Google usually fails to index websites with robots.txt restrictions, it is vital to remember that there is no method to ensure removal from search results by using the robots.txt file.

## Methods for Locating The robots.txt File

The website's robots.txt script can be found at "exampledomain.com/robots.txt" if you have already set up one. Enter the URL there in a web browser. You've got a robots.txt file when you view text similar to the following –



## Creating A robots.txt File: Instructions

- A robots.txt file is simple in case you have never created one. Just launch a blank.txt file and start entering instructions. Keep adding to the directives you operate till you've covered all the anticipated fields. Name the file you're saving "robots.txt."
- A robots.txt generator is an additional option. The benefit of employing such a tool is that it reduces syntax errors. This is fortunate, considering a single error may have disastrous SEO effects on your website. The drawback is that there are some restrictions on flexibility.

## Location of the robots.txt file

- The leading directory of the subdomains to which your robots.txt file refers should contain it. The robots.txt file, for instance, must be available at

"tutorialspoint.com/robots.txt" to regulate crawling behavior for "tutorialspoint.com".

- You must be able to view the robots.txt file at "ebooks.domain.com/robots.txt" if you wish to limit crawls to subdomains like "ebooks.domain.com".

## Guidelines for the robots.txt file

### For every directive, start a new line

A distinct line must be created for each directive. Search engine spiders will become confused if it doesn't.

### Directions can be made more accessible by using wildcards

Wildcards (\*) can identify URL sequences and implement them across all user-agents when expressing directives.

### To indicate a URL's end, enter "\$."

To indicate the conclusion of a URL, use the dollar sign "\$". A robots.txt file could resemble something like this if you prefer to stop web crawlers from viewing all the .png files on your website –

```
User-agent: *  
Disallow: /*.png$
```

### Make only one use of each user-agent

Google is okay with it when you utilize a single user-agent repeatedly. However, all the regulations from the different declarations will be combined and followed, thus reducing accuracy and, in some cases, not counting an aspect. Considering that the configuration has fewer complexities, it makes it logical to specify each user-agent just once. Maintaining things organized and straightforward can reduce your risk of serious errors.

### Write comments to inform others about your robots.txt file

Developers—and perhaps even your later self—can understand your robots.txt file easier due to comments. A hash (#) should be used to start a line of comments.

## Be detailed to prevent unintended mistakes

Setting directives without specific guidelines might lead to overlooked errors that can seriously harm your SEO efforts.

```
# This instructs Google not to crawl the website.  
User-agent: Googlebot  
Disallow: /
```

## Issue with Blocks Due to robots.txt

This indicates that there is non-Google-indexed content that has been restricted by robots.txt on your website. Turn off the robots.txt crawler restriction if the data is significant and needs to be crawled and indexed.

## Conclusion

A straightforward but effective file is robots.txt. If used appropriately, it can help your SEO. You'll regret it later on if you use it carelessly.