COESIONE ITALIA 21-27 PIEMONTE

Cofinanziato dall'Unione europea

REGIONE PIEMONTE

Web Developer

HTML, CSS e Strumenti di Digital Marketing (SEO, SEM, SEA)

Docente: Shadi Lahham

Intervento realizzato da

ITS
TECNOLOGIE
DELL'INFORMAZIONE E
DELLA COMUNICAZIONE

Robots

Technical seo

Shadi Lahham - Web development

Robots.txt

Robots

robots.txt is a simple text file that is essential for controlling how search engines discover and index a website's content, providing instructions to web crawlers about which parts they can access and index

the robots.txt file is typically placed in the root directory of a website, at the same level as the main HTML files and folders that make up the website's content

it's important to note that robots.txt is a guideline, not a rule, and search engines may not always follow its instructions

Robots - syntax

robots.txt uses a straightforward syntax to specify rules for user-agents to:

- allow or disallow access to specific parts of a website
- control the crawling behavior of specific bots
- prevent indexing of certain pages or directories

the robots.txt file is composed of one or more groups, each specifying a user-agent, such as googlebot and bingbot, and the directories or pages that are disallowed or allowed for that specific user-agent

Robots - common directives

User-agent
identifies the web crawler to which the rules apply, where * represents all crawlers, and specific crawler names can target specific bots

Disallow
prevents web crawlers from accessing specific directories or files, acting as the primary method for blocking crawlers from certain parts of a site

Allow
overrides a Disallow directive for specific URLs, allowing access to specific directories or files within a restricted section

Sitemap
provides the location of the XML sitemap, helping search engines understand the site's structure and locate all pages

Robots - rare directives

Crawl-delay
controls the rate at which a crawler requests pages from the server, helping to reduce server load

Host
specifies the preferred domain if the site is available under multiple domains, useful for canonicalization purposes, though not all crawlers respect this directive

Clean-param
instructs search engines to ignore certain URL parameters when crawling a site, helping to reduce duplicate content

Simple example

```
User-agent: *
Disallow: /admin/
Disallow: /private/
Allow: /public/

User-agent: Googlebot
Allow: /blog/

Sitemap: https://www.example.com/sitemap.xml
```

More complex example

```
User-agent: *
Disallow: /
Allow: /$
Allow: /public/
Allow: /products/*.html$

User-agent: Googlebot
Disallow: /no-google/
Allow: /

User-agent: Googlebot-Image
Disallow: /private-images/

User-agent: Googlebot, Bingbot, Yandex
Disallow: /admin/

Crawl-delay: 10
Clean-param: sessionid /product/

Host: www.example.com
Sitemap: https://www.example.com/sitemap.xml
```

Example explained

Allow: /
allows crawling of all pages on the website, including all directories and subdirectories, and overrides any Disallow directives for the specified user-agent

Allow: /$
specifically allows crawling of the root page (homepage) of the website, using the $ symbol as a regular expression to match the end of the URL, applying only to the URL that ends immediately after the domain name (e.g., https://example.com/)

Disallow: / and Allow: /$
this combination blocks all pages except the homepage, where $ represents the end of the URL, ensuring that / matches only the root URL

Allow: /products/*.html$
allows crawling of all HTML files in the /products/ directory. The * acts as a wildcard, and $ ensures it matches only URLs ending with .html

Example explained

`Crawl-delay: 10`
`sets a 10-second delay between page requests for all bots, helping to reduce server load`

`Clean-param: sessionid /product/`
`instructs crawlers to ignore the sessionid parameter within the /product/ directory, reducing duplicate content caused by dynamic URLs`

`Host`
`specifies the preferred domain version for canonicalization purposes, guiding crawlers to prioritize one version over others`

User-agent strings

search engines frequently update their user-agent strings, malicious actors often disguise or forge their identities to evade detection, and major search engines also use multiple official crawlers

the most prominent user-agent tokens

- googlebot
- bingbot
- yahoo! slurp
- baiduspider
- yandexbot
- duckduckgobot

Google crawlers
Bing crawlers

Optimization

prioritize important pages
ensure that these pages contain the most valuable content, making them accessible to search engines

block unnecessary content
prevent indexing of low-quality or duplicate content, which negatively impact site rankings

optimize for mobile
consider mobile-specific robots.txt rules so that mobile content is correctly indexed and accessible

test and monitor
regularly test the robots.txt file to ensure it works and maintains site visibility in search engines

optimization techniques

- might want to exclude pages with no SEO value such as admin and login pages, etc.
- allow access to essential resources such as CSS, JavaScript, and images
- avoid blocking important sections of the site unintentionally

Robots meta tag

Robots meta tag

the robots meta tag is an HTML tag placed within the <head> section of an individual page to provide specific instructions to search engine crawlers about that page

```html
<meta name="robots" content="index, follow">

<!-- other combinations -->
<meta name="robots" content="noindex, follow">
<meta name="robots" content="index, nofollow">
<meta name="robots" content="noindex, nofollow">
<meta name="robots" content="noarchive">
```

index: allows the page to be indexed by search engines
noindex: prevents the page from being indexed
follow: allows search engines to follow links on the page
nofollow: prevents search engines from following links on the page
noarchive: prevents search engines from caching the page

Robots.txt vs robots meta tag

use robots.txt for site-wide rules and directory-level control and the robots meta tag for page-specific control and exceptions since **Disallow** in robots.txt prevents crawling while the robots meta tag **noindex** prevents indexing

noindex
should be used to help search engines understand the page's content and relationships with other pages, but prevent the page from appearing in search results

situations where noindex could be used

- internal pages or pages no SEO value, such as admin, login, shopping cart pages
- duplicate content
- low-quality content which lowers the site's SEO value
- temporary content such as pages created for a temporary event

nofollow Attribute

nofollow Attribute

the nofollow attribute can be added to anchor tags <a> to indicate that a search engine should not follow the link or pass SEO value

```
<a href="https://example.com" rel="nofollow">Visit Example</a>
```

When to use nofollow

- external links to untrusted or user-generated content (UGC) such as forums or user comments
- paid or sponsored links
- links to low-quality pages or to pages that are irrelevant to a website's content

recommended: read about link equity (juice) and nofollow, sponsored and UGC links for more depth

Best practices

- combine robots.txt and meta tags for maximum control
- ensure consistency between robots.txt file and meta tags
- regularly review and update robots.txt file
- use the nofollow attribute wisely to manage link equity
- test robots.txt file regularly using reliable tools

robots.txt is a **guideline**, not a strict rule - some crawlers may ignore it

Try

Write your own robots.txt file with different rules

Test and validate it with the following tools, try both

Robots.txt validator and testing tool

Robots.txt testing & validator tool

References

Robots.txt

Yoast: ultimate guide to robots.txt

Woorank: robots.txt - a beginner's guide

Robots Meta Tag

Meta robots tag & x-robots-tag explained

What are robots meta tags

Robots meta tags specifications

References

Nofollow

nofollow, sponsored and ugc links: what you need to know

Woorank: understanding link juice in seo

Tools

Robots.txt validator and testing tool

Robots.txt testing & validator tool

Crawlers

Google crawlers

Bing crawlers