

Robots.txt SEO: The 1994 Rule That Still Matters

Category: Technical Blog

October 17, 2026

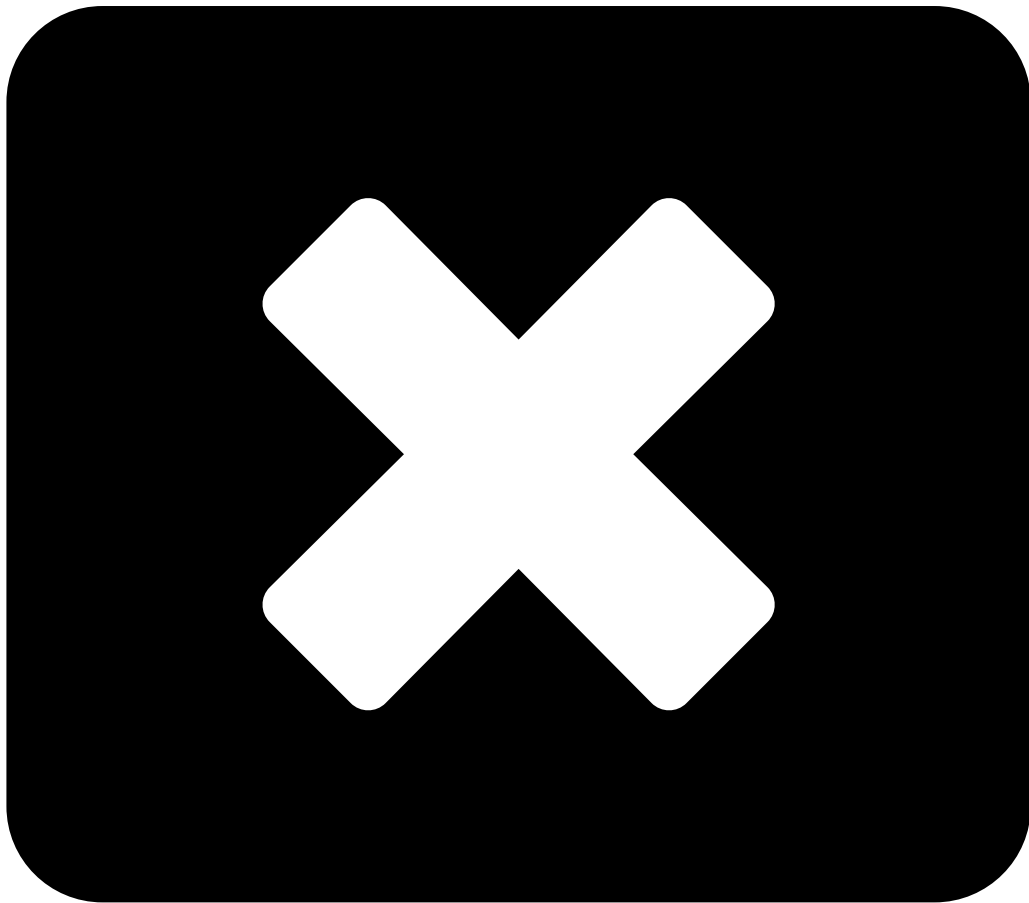


- [Services](#)
- [Projects](#)
- [Blogs](#)
- [About](#)
- [Contact](#)
- [Languages](#)
 - [NL](#)
 - [BE](#)

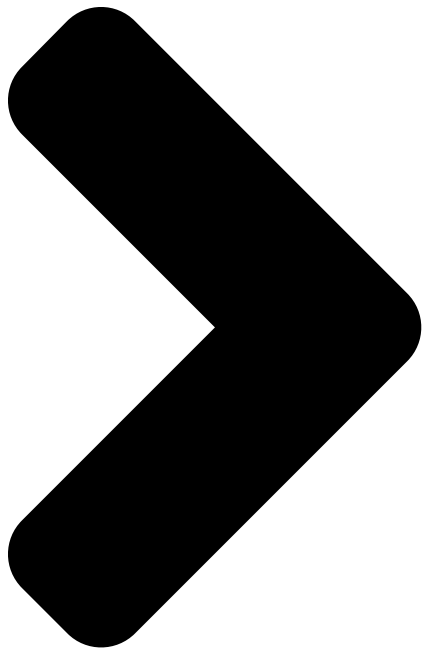
Hamburger Toggle Menu



[Free Quote](#)

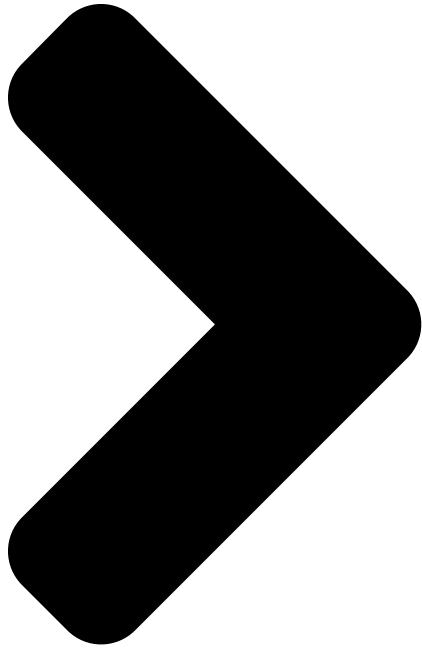


.



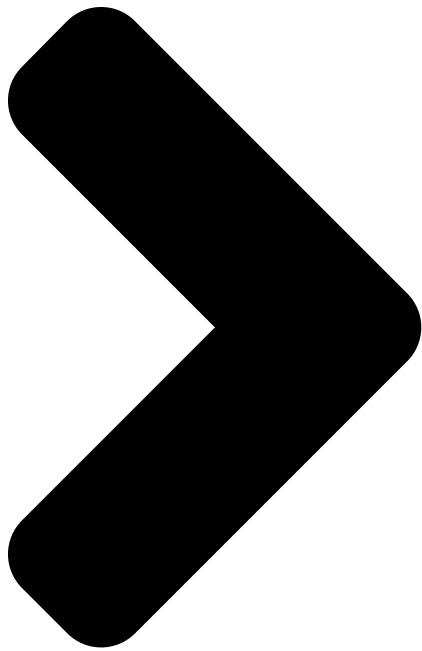
About us

.



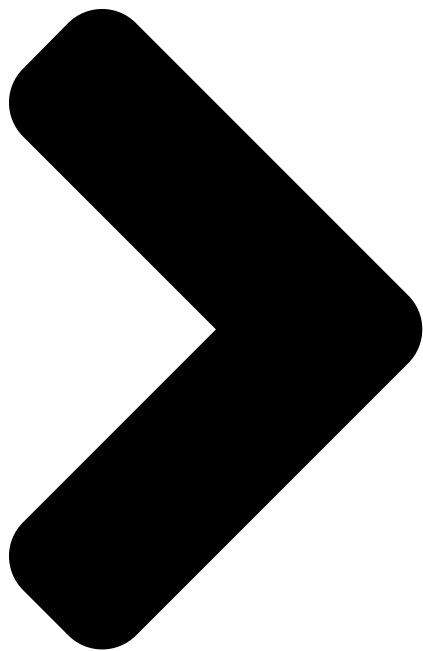
Request for detailed website quote

.



Courses University TSI

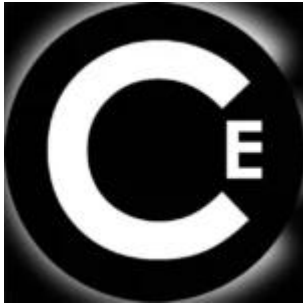
.



Login Team TSI Digital Solution

OUR PARTNERS



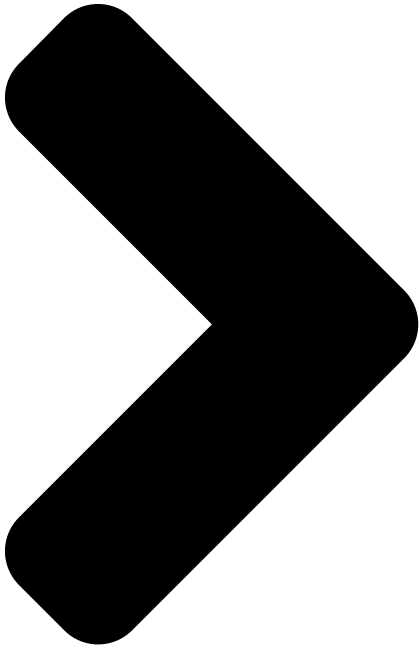


Trending Topic

Publicity

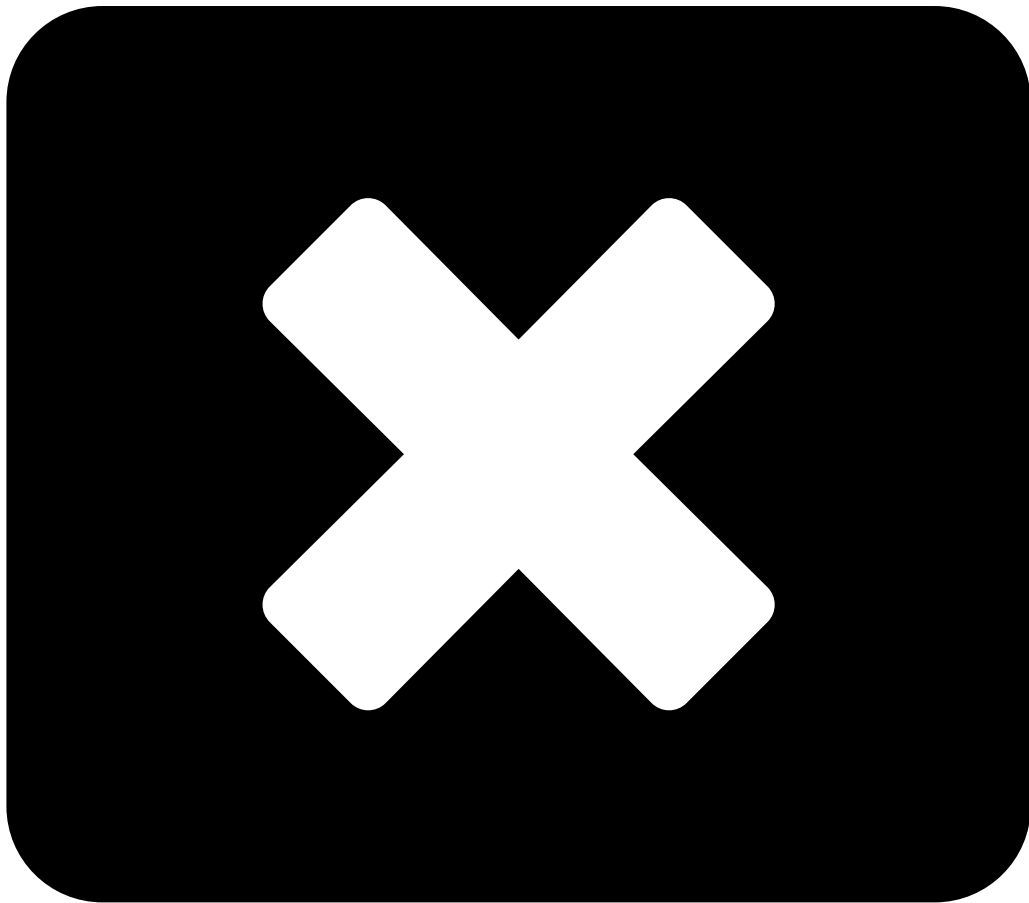
Reach out



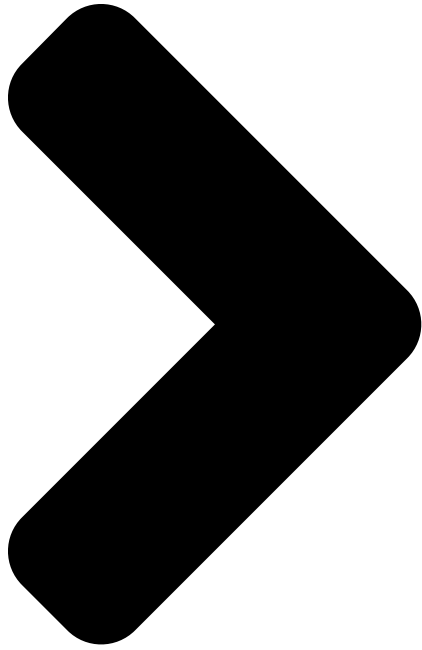


Some projects

Edit Template

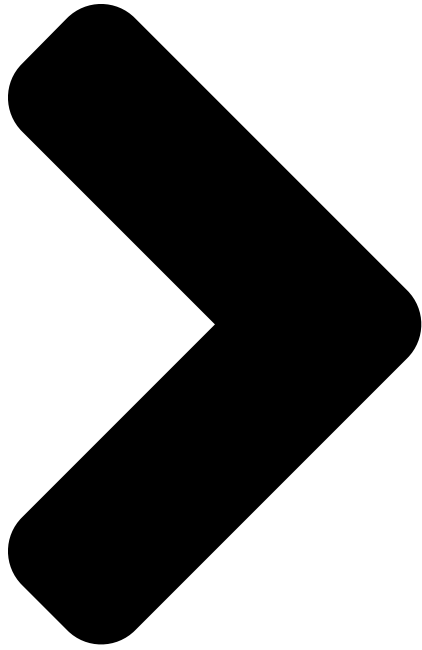


.



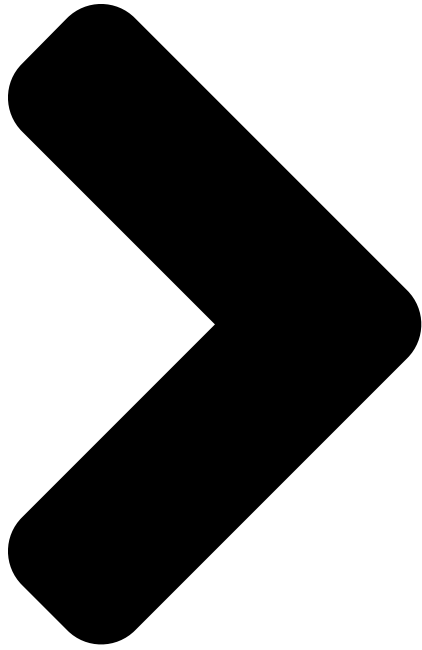
Services

-



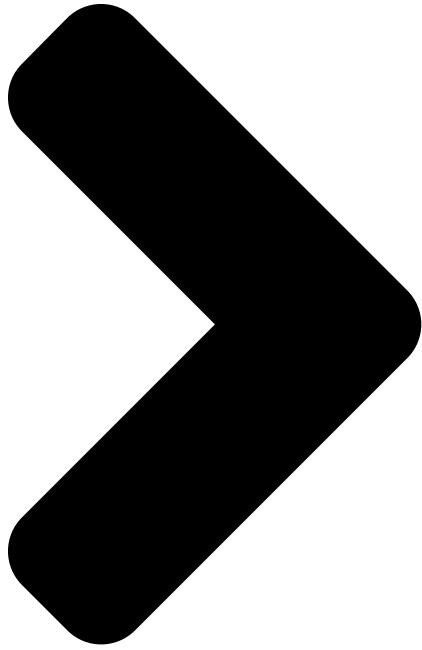
Web Design & Development

.



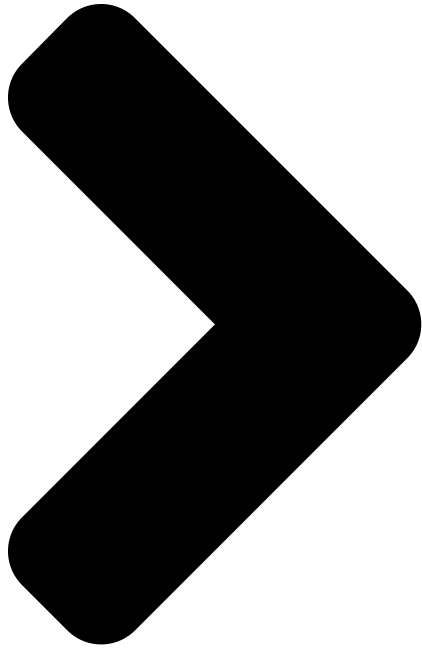
Hosting





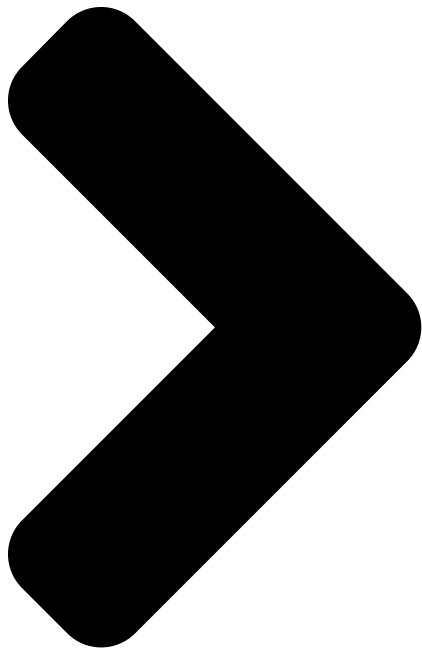
SEO/Ads/GMB/Speedup/Technical

.



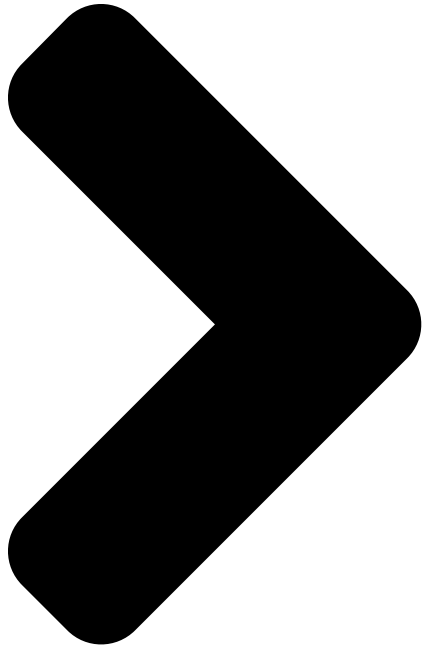
App Design & Development

•



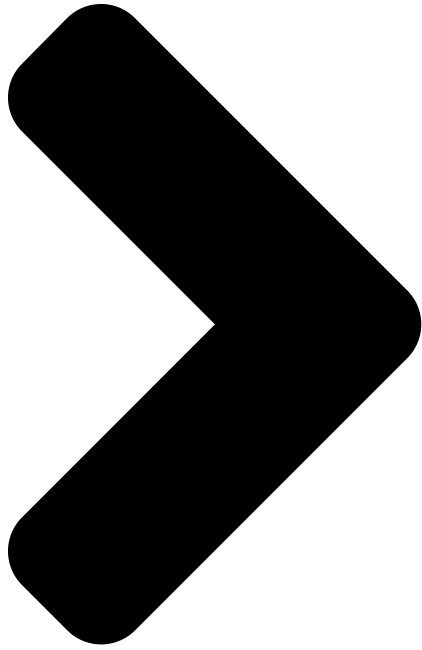
AI Agents





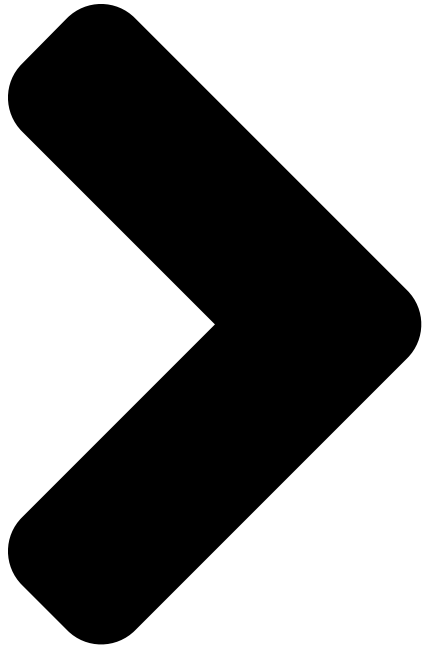
e-Commerce





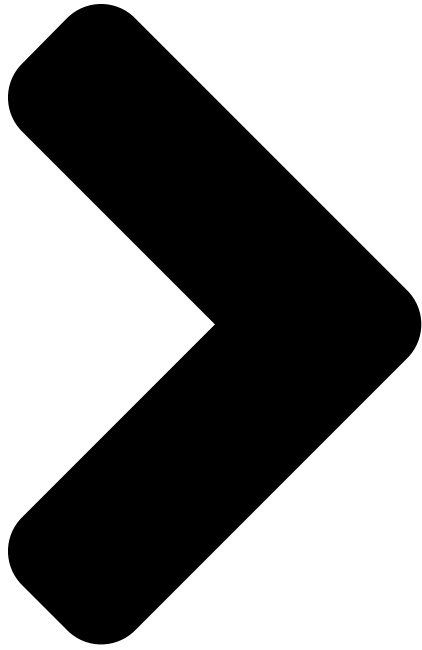
Branding

.



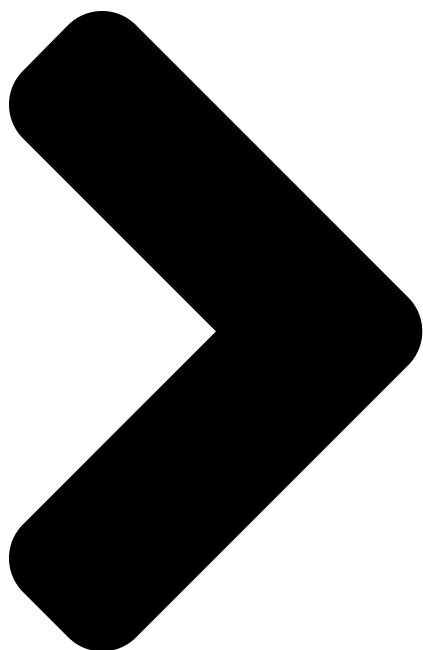
Social Media & Content Creation

.



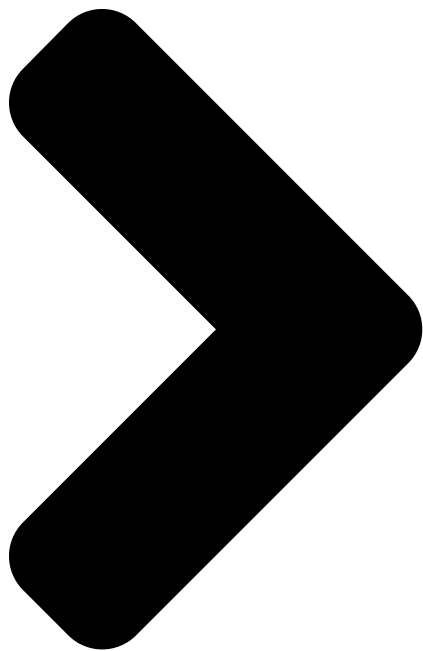
Graphic Design

.



Copywriting & Translations

.



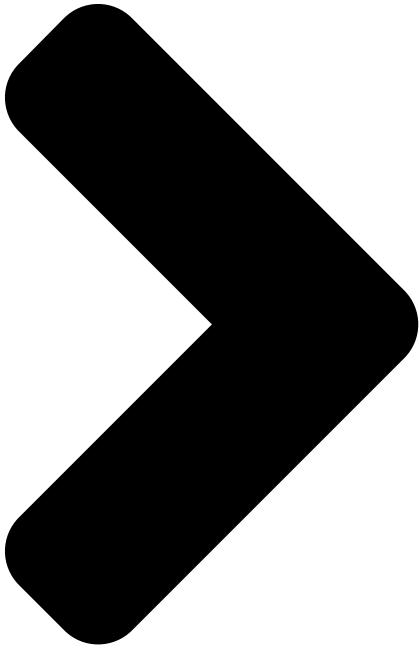
[Photo- & Videography](#)

[Trending Topic](#)

[Publicity](#)

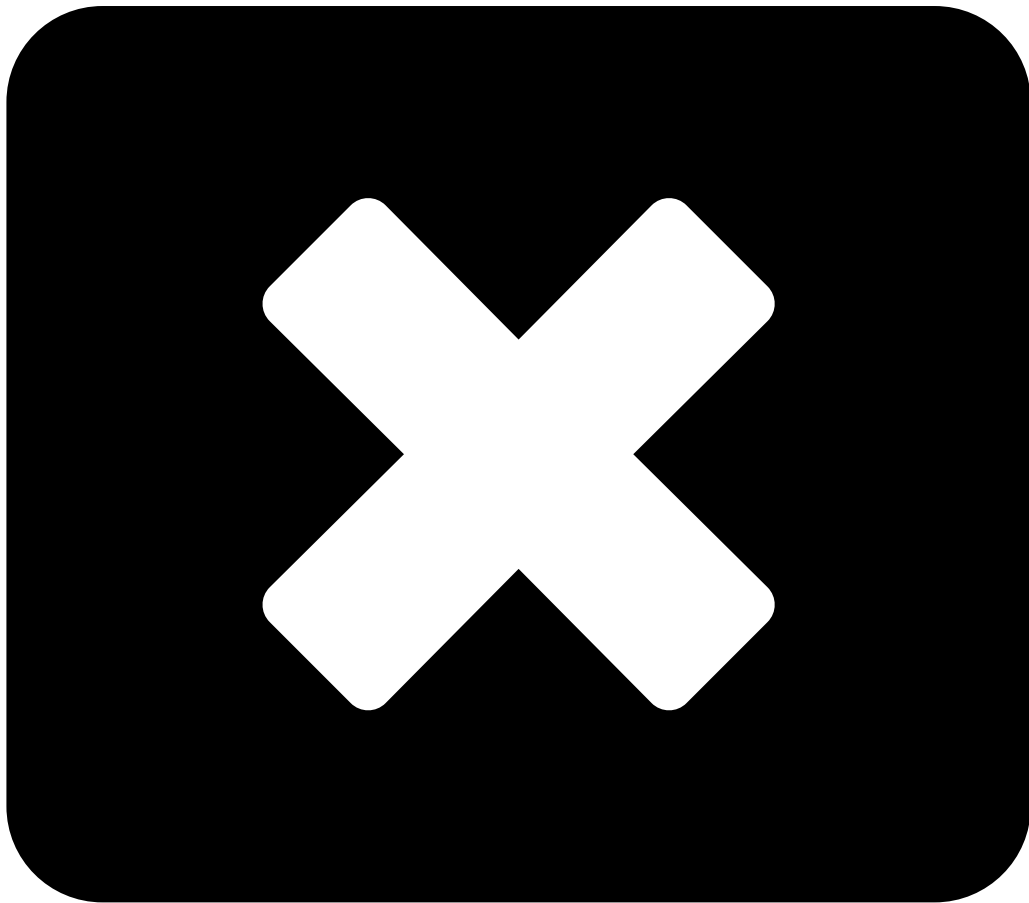
[Calculate a quote
online yourself](#)



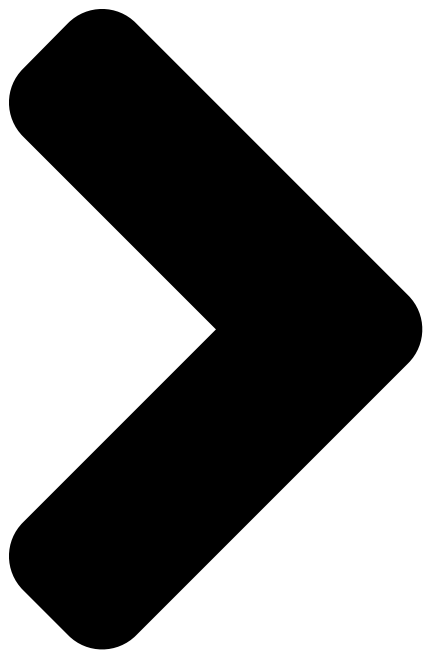


The way we work

Edit Template



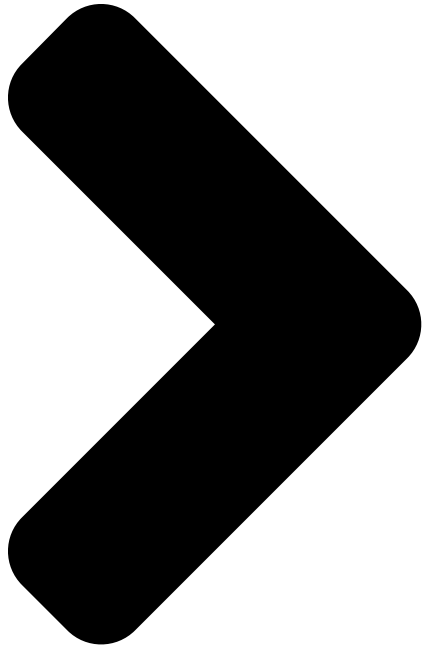
.



Realized projects

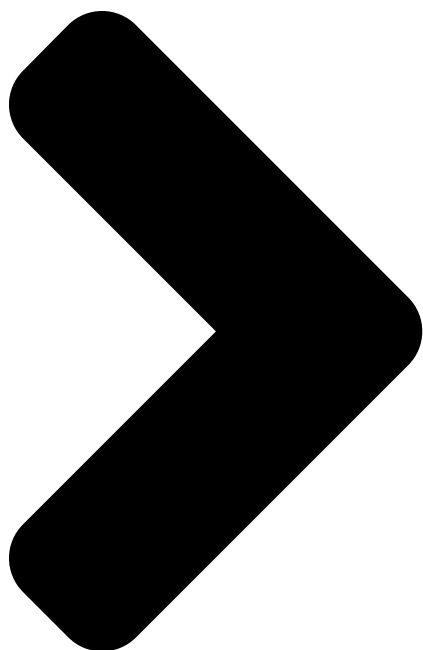
Calculators for your project costs

.



Request for detailed website quote

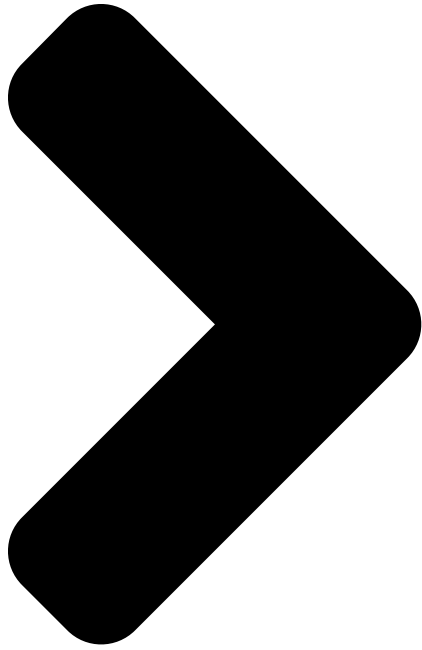
.



Overview of quote calculators

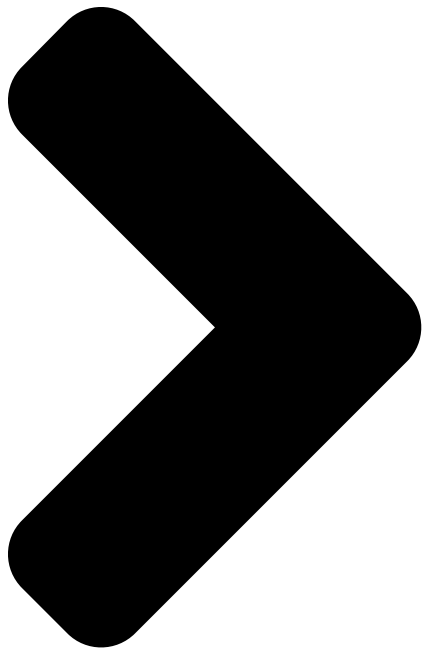
Useful information

-



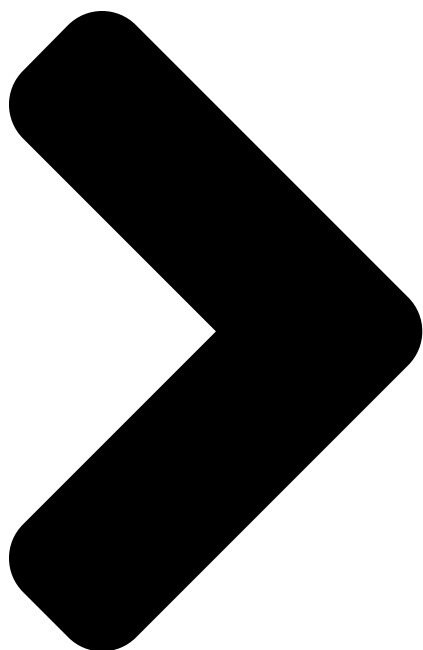
Jobs @ TSI Digital Solution

.



Hosting at €75/year

.



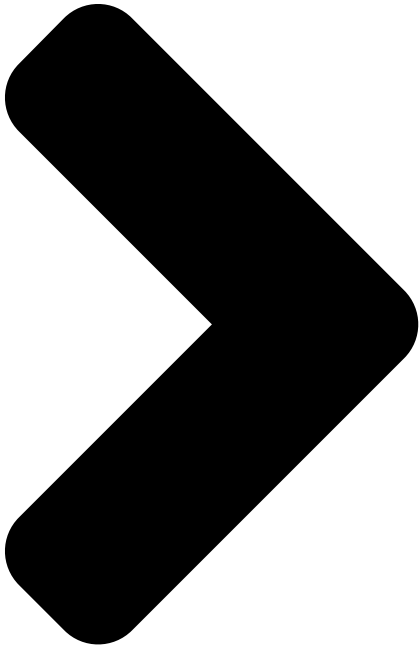
TSI Free Tools

Trending Topic

Publicity

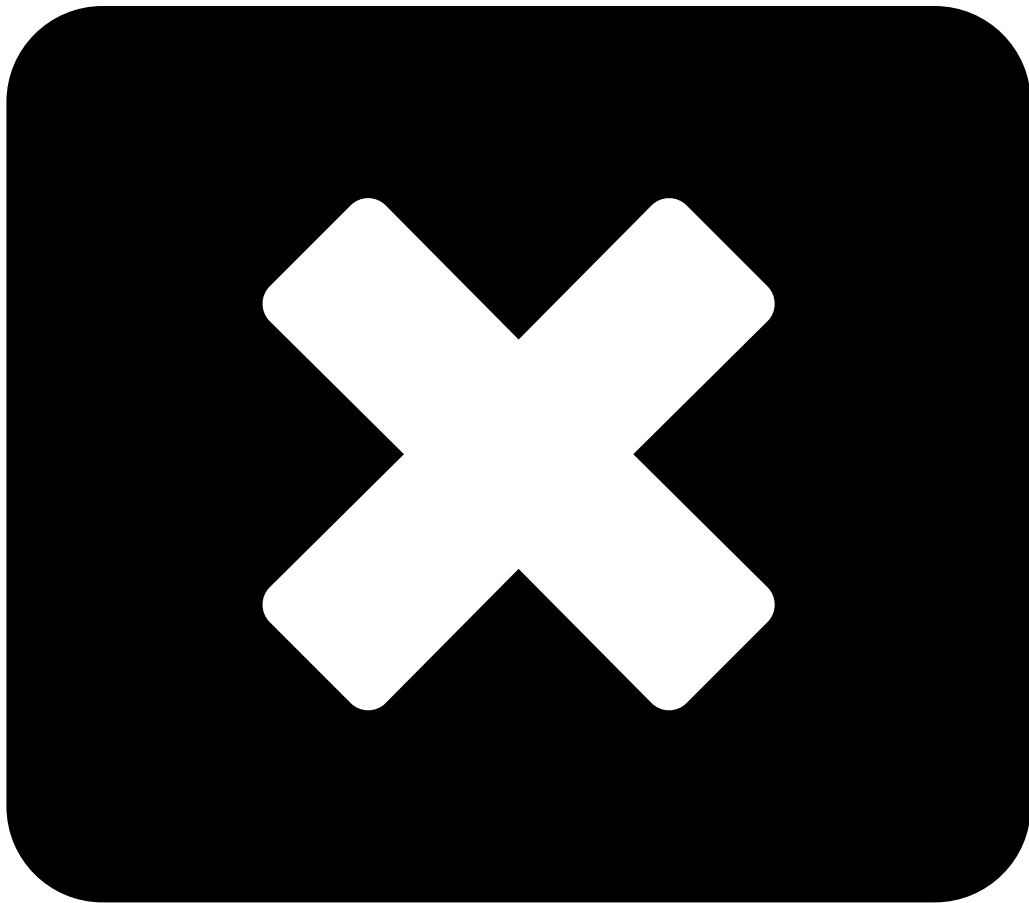
Interesting stuff to read



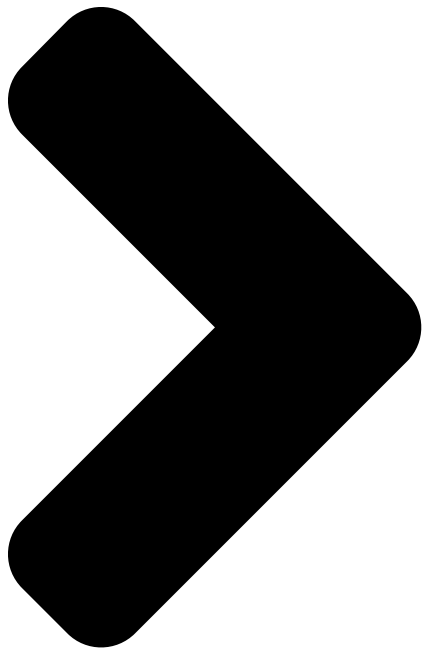


[See More](#)

[Edit Template](#)

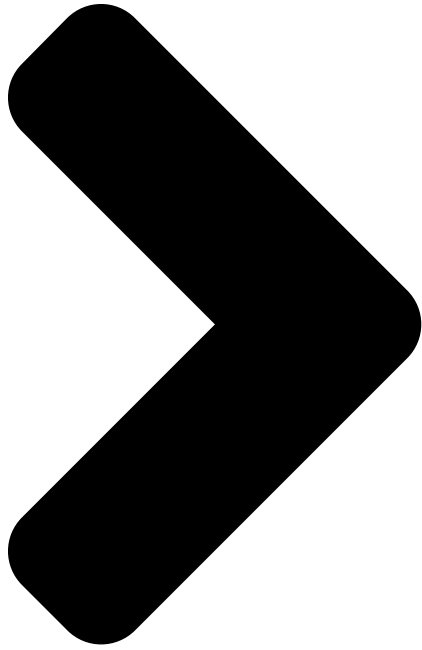


.



Blogs

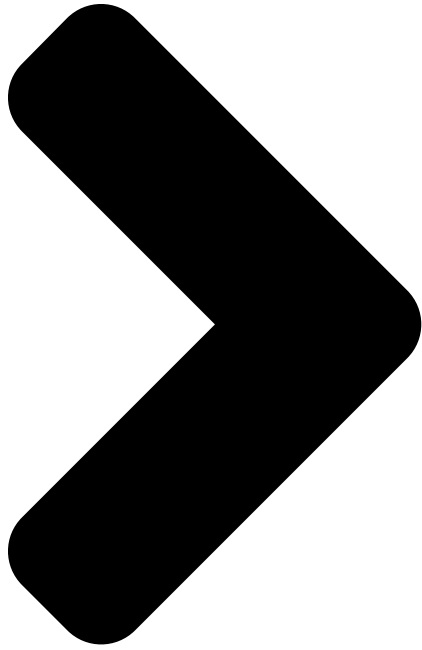
.



Request for detailed website quote

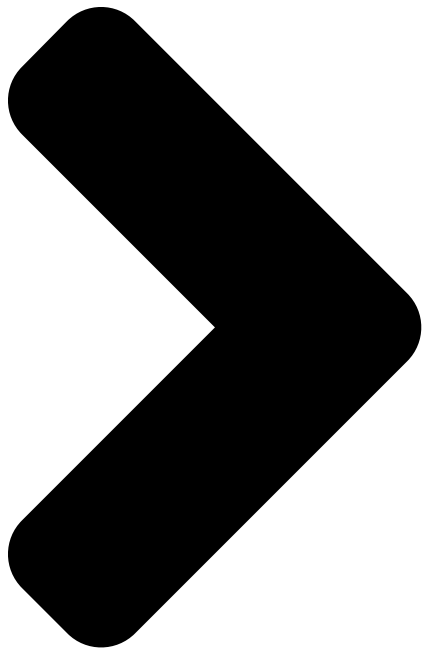
Some weekly posted topics

.



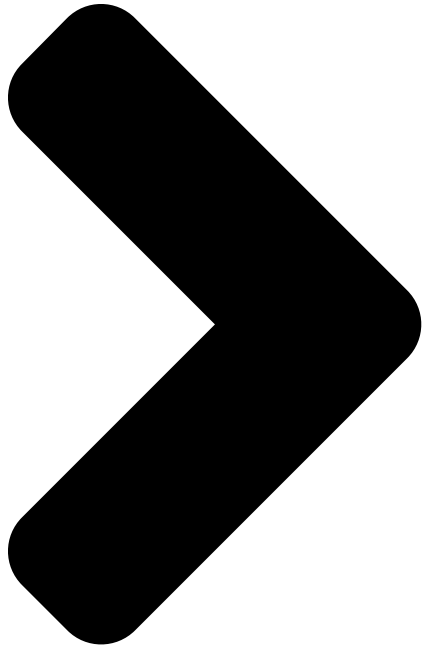
Cheap Website vs. High Performance

.



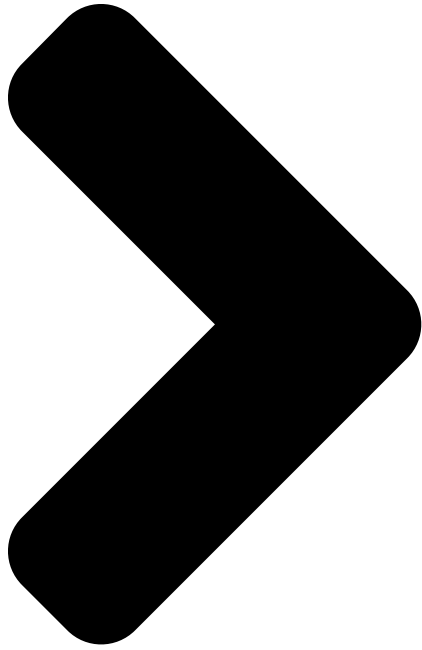
What Is an AI Agent?

.



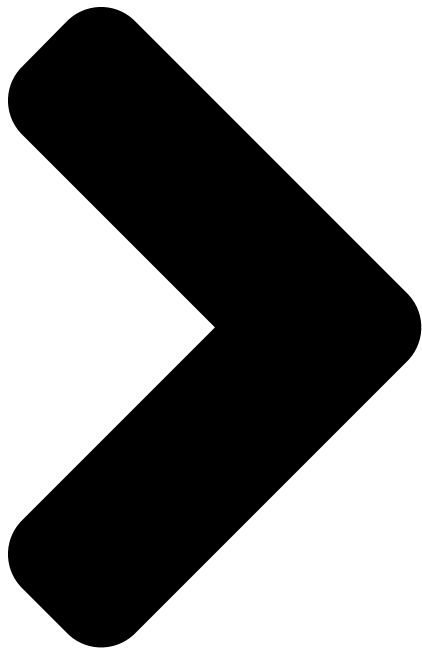
Client Communication in Digital Marketing

-



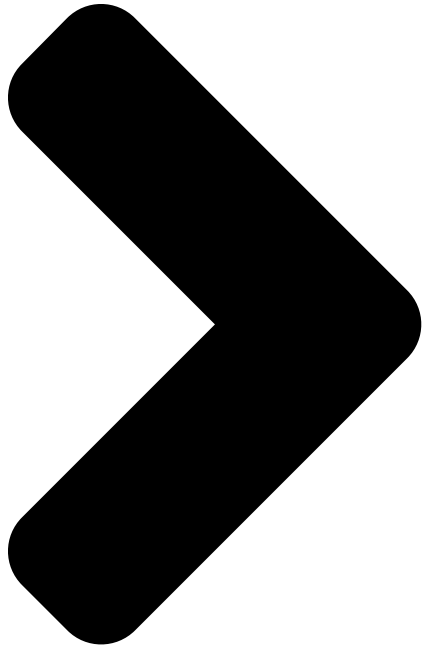
Human Made Websites vs AI Websites

.



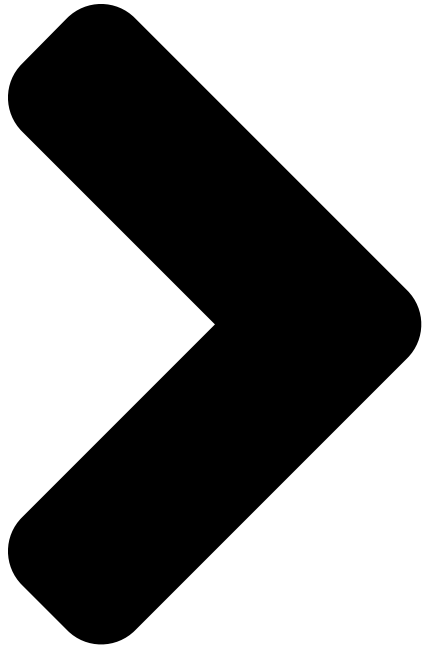
Your Digital Marketing Ecosystem Explained

.



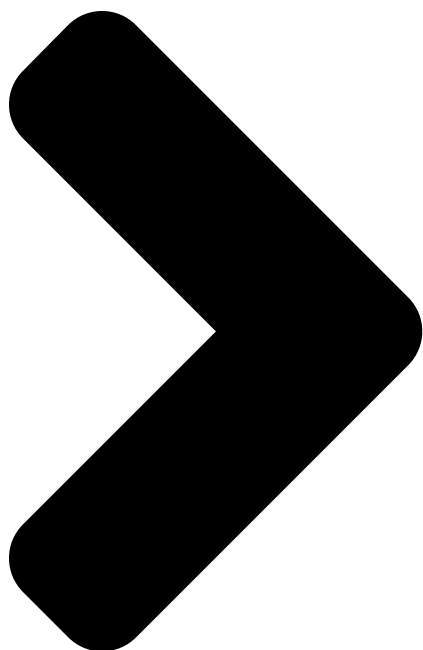
Mastering Answer Engine Optimization (AEO)

.



Beware the Template Trap

-

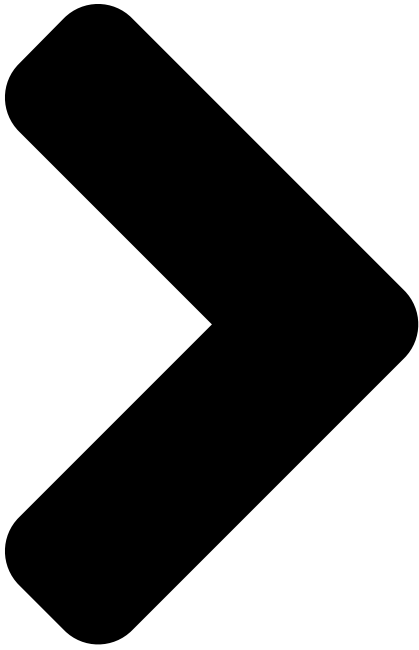


NWOW: Results Matter, Not Presence

Trending Topic

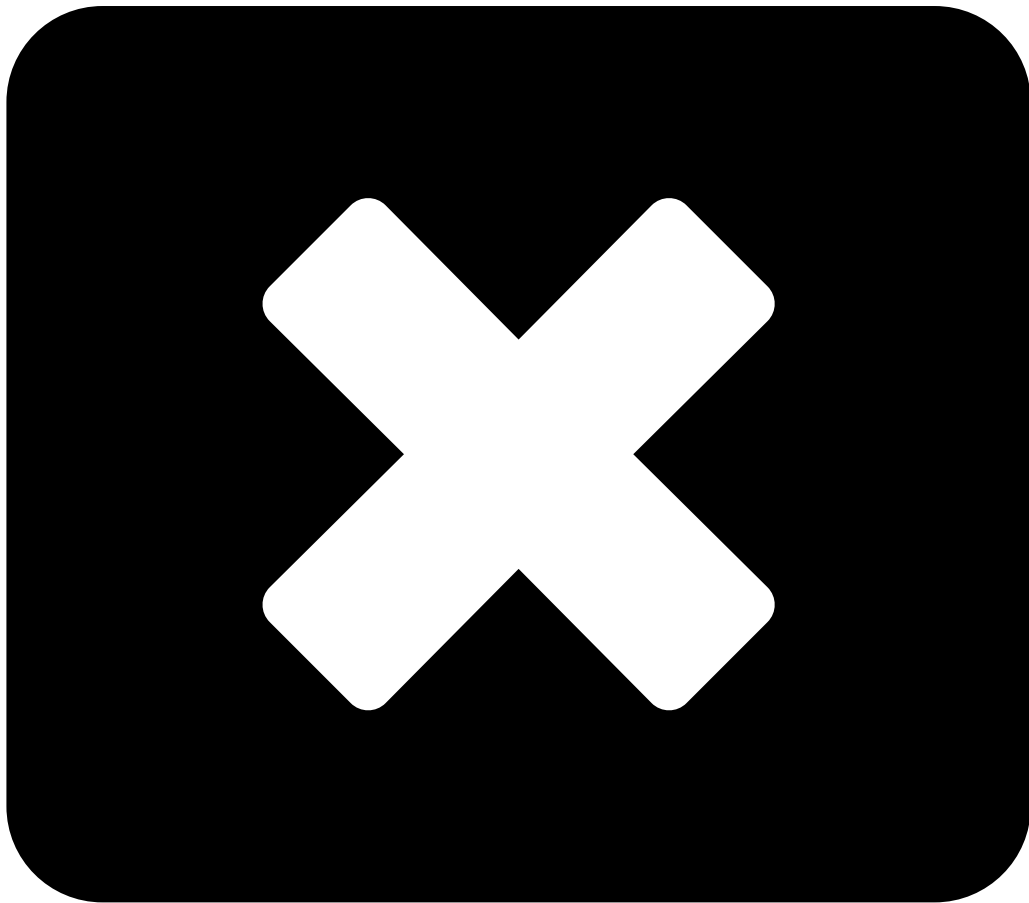
Publicity



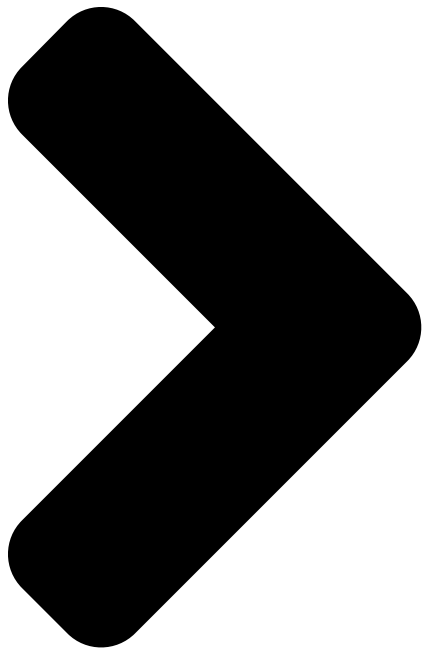


Who we are

Edit Template

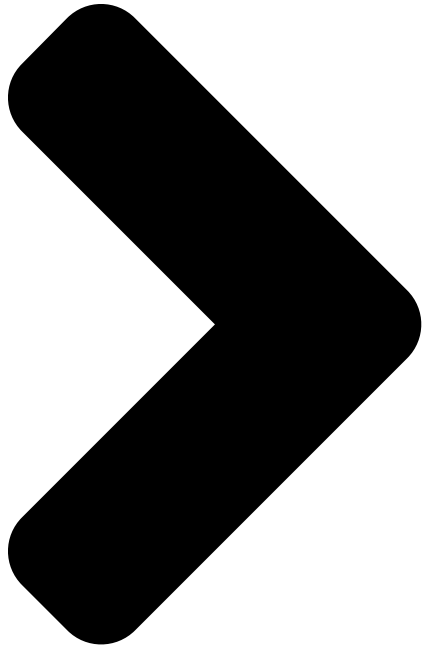


.



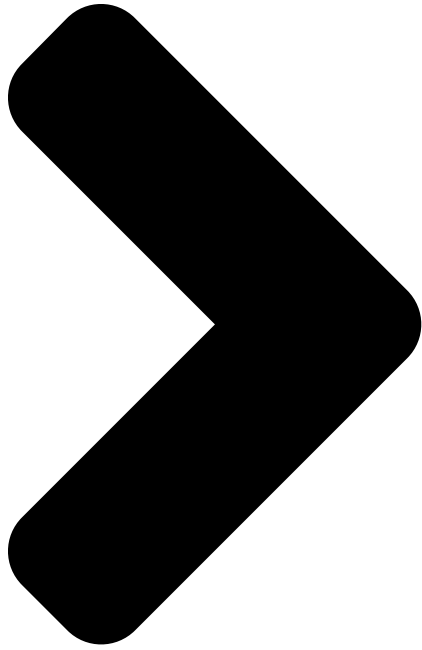
Contact us

.



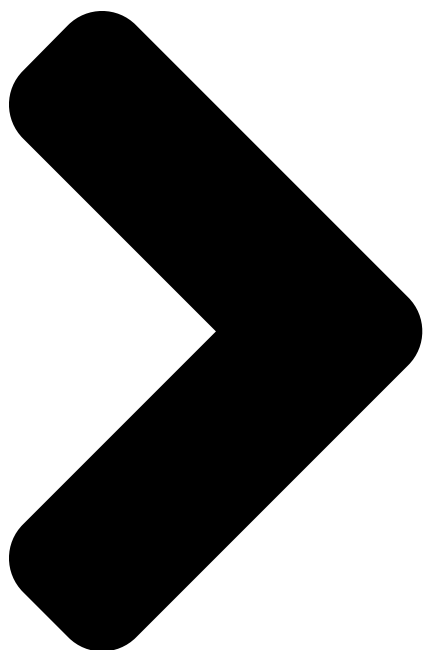
Request for detailed website quote

.



Jobs @ TSI Digital Solution

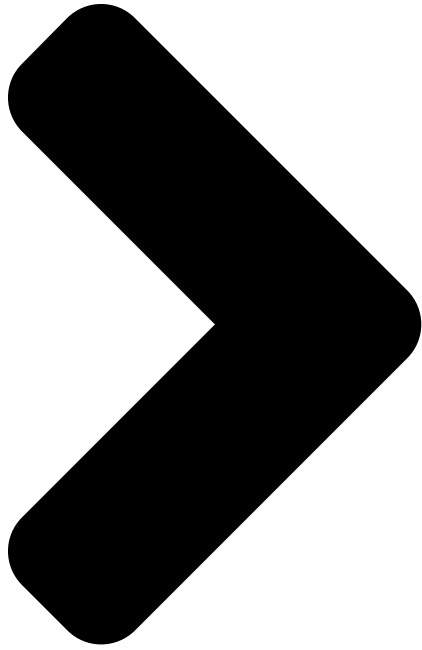
.



Affiliate program

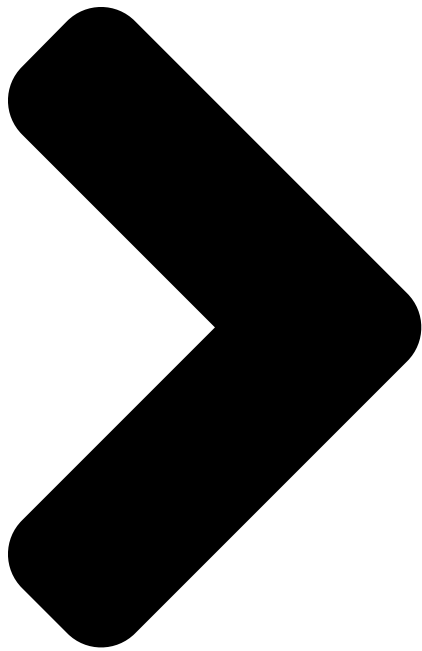
Download for free

.



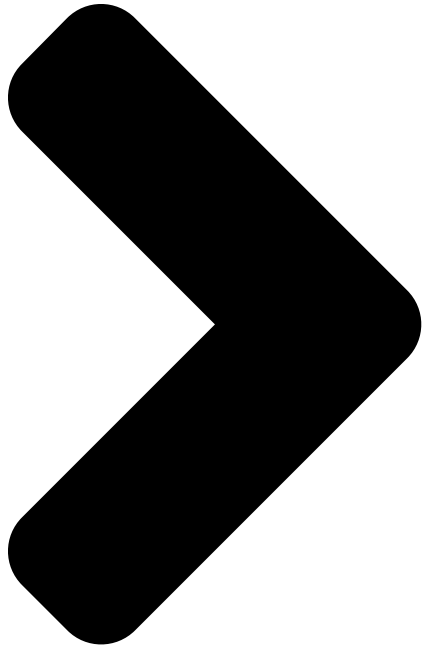
TSI Free Tools

.



TSI Spam Blocker

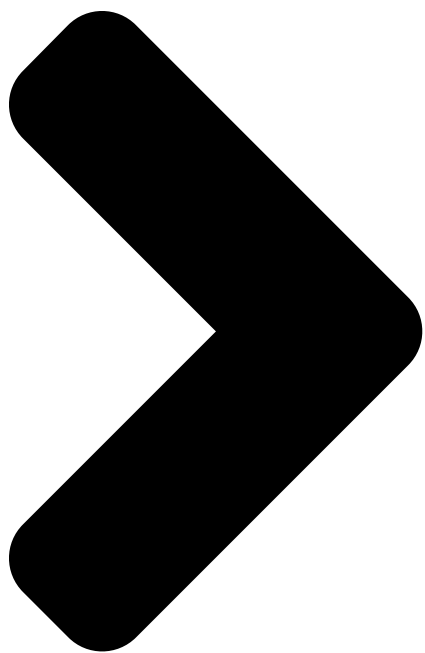
.



TSI Downloadable Webp Converter

GSAP

.

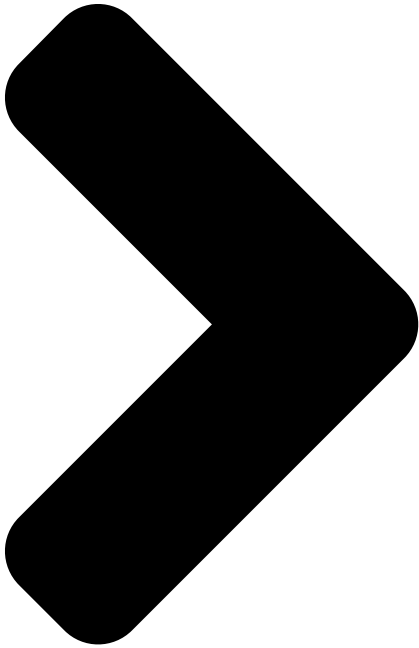


Demo site GSAP animation

Trending Topic

Publicity





[See More](#)

[Edit Template](#)

[Edit Template](#)

Robots.txt SEO: The 1994 Rule That Still Matters

[Get An Online Quote](#)

Robots.txt SEO



Best Practices for Crawlability and Indexing

Imagine launching a redesigned website, investing in content, and waiting for Google to discover your best pages—only to find that a single line of

configuration is preventing search engines from crawling them. It sounds like a small technical mistake. But if that rule affects your service pages, product categories, or high-performing articles, the consequences can reach far beyond your website settings.

This is why **robots.txt remains an important part of technical SEO and index hygiene**. The file is simple, but misunderstanding what it does can create expensive problems. Its story also goes back further than many marketers realize. The protocol originated in 1994, long before modern SEO, ecommerce platforms, and AI-powered search.

Here's how robots.txt works, what its limitations are, and how to use it without accidentally restricting the pages that matter most to your business.

What Is Robots.txt in SEO?

A robots.txt file is a plain-text file that tells compliant web crawlers which URLs they may or may not access on a particular website. It is generally located at the root of a website's origin. For example:
<https://www.example.com/robots.txt>

Search engines consult this file to understand the site's crawling preferences. Website owners can use it to restrict access to selected URL patterns, manage unnecessary crawler requests, and guide bots away from areas that do not need to be crawled.

However, one distinction matters more than anything else:

Robots.txt controls crawling; it is not a reliable way to prevent a page from appearing in search results.

Google explains that a blocked URL can still be indexed if it is discovered through links or other signals. If your goal is to remove a page from Google Search, you need an appropriate indexing or access-control method instead.

Official guidance: [Google Search Central's robots.txt documentation](#).

The History of Robots.txt: A Rule Born in 1994

The web has changed dramatically since the early days of search engines. Yet one of its fundamental crawling conventions traces back to a problem that website owners faced in the early 1990s. As automated crawlers explored websites, they could request pages that owners did not want them to visit, or generate more requests than a server could comfortably handle.

In February 1994, internet pioneer Martijn Koster proposed the Robots Exclusion Protocol as a simple way for website administrators to communicate crawling preferences to automated clients. The idea was straightforward: a website could publish rules that cooperating crawlers were expected to respect.

For years, this convention was widely used without a formal Internet Engineering Task Force standard. That changed in September 2022, when the IETF published [RFC 9307: Robots Exclusion Standard](#), formalizing and extending the protocol. That means a practice originating in 1994 received its formal standards-track specification 28 years later.

Why does this history matter to modern SEO? Because robots.txt was designed to communicate crawling preferences, not to serve as a security system, a deindexing tool, or a shortcut to higher rankings.

Understanding its original purpose helps prevent many of the mistakes that still occur today.

How Does Robots.txt Work?

When a crawler visits a website, it can request the site's robots.txt file and check the rules relevant to its crawler identity. The file uses directives such as User-agent and Disallow. A Sitemap line can also identify an XML sitemap.

Here is a simple example:

```
User-agent: *
```

```
Disallow: /internal-search/
```

```
Disallow: /account/
```

```
Sitemap:
```

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

In this example:

- User-agent: * applies the following rules to all crawlers that honor the wildcard group.
- Disallow: /internal-search/ asks compliant crawlers not to crawl URLs matching that path pattern.
- Disallow: /account/ asks them not to crawl URLs under that path.
- Sitemap: identifies the location of the XML sitemap.

The example is illustrative. Before applying similar rules to a live website, verify that the paths really are appropriate to restrict.

A robots.txt file is not automatically required to contain restrictive rules. If your website has no reason to limit crawling, a carefully configured file with no disallow rules may be appropriate.

Robots.txt syntax: Common directives

Directive	Purpose
User-agent	Identifies the crawler or crawler group addressed by the rules
Disallow	Specifies URL paths that the targeted crawler should not crawl
Allow	Allows crawling of a path that would otherwise be restricted by a broader rule, where supported
Sitemap	Identifies an XML sitemap location

Crawlers can interpret some syntax differently, so check the relevant search engine's documentation before relying on complex patterns. Google provides further details in its [robots.txt rules and interpretation guide](#).

One important limitation: robots.txt is a request to cooperating crawlers, not an access-control mechanism. A malicious bot may ignore it. Private dashboards, customer records, payment pages, and confidential files must be protected through authentication and server-side access controls.

Robots.txt vs. Noindex: What Is the Difference?

This is one of the most important distinctions in technical SEO. Robots.txt and noindex solve different problems. Choosing the wrong one can leave unwanted URLs in search results or prevent search engines from reading the instructions needed to remove a page.

Feature	Robots.txt	noindex
Main purpose	Restrict crawling of selected URL paths	Tell supporting search engines not to index a page
Can Google still discover a blocked URL?	Yes	Yes, but a crawlable page carrying noindex should eventually be excluded from Google Search
Can Google read the page's noindex tag if crawling is blocked?	No, not if it cannot access the page	The page must remain crawlable for Google to read the directive
Suitable for securing confidential content?	No	No
Typical use	Managing crawling of selected URL patterns	Keeping a crawlable page out of search results

Suppose you have an old campaign landing page that should no longer appear in Google Search but still needs to be accessible to users with a direct link.

A suitable option may be to leave the page crawlable and add a supported noindex directive. Blocking it in robots.txt at the same time can prevent Google from seeing that instruction. For genuinely private information, use authentication or remove public access rather than relying on either robots.txt or noindex.

Google's documentation explains the details: [Block search indexing with noindex](#).

Why Robots.txt Matters for Index Hygiene

Index hygiene is the process of helping search engines discover, crawl, and index the URLs that deserve to appear in search results while reducing unnecessary or problematic URLs.

Robots.txt is one part of that process, alongside canonical tags, internal links, XML sitemaps, redirects, HTTP status codes, and page-level indexing directives. Consider an ecommerce website with thousands of product-filter combinations. Filters for size, colour, sorting, and price may create large numbers of URLs with overlapping content.

If search engines spend substantial time crawling low-value URL variations, crawling resources may be diverted from other pages. Carefully chosen robots.txt rules can sometimes reduce unnecessary crawling—but they must be based on the site's URL structure and indexing goals.

A few practical principles follow:

- Keep valuable product, category, service, and editorial pages crawlable.
- Review filter and parameter URLs before restricting them.
- Ensure important pages can be discovered through internal links.
- Keep XML sitemaps focused on URLs you want search engines to consider for indexing.
- Use canonical tags and indexing directives for their intended purposes rather than expecting robots.txt to replace them.

Remember that robots.txt does not automatically consolidate duplicate URLs or tell Google which version should be canonical. Those are separate technical SEO decisions.

Real-World Technical SEO Case Study: Ahrefs Blocks Two Ranking Pages

The risks become clearer when we look at a documented experiment rather than relying only on theoretical examples.

In an article published by Ahrefs, technical SEO specialist Patrick Stox described what happened when he deliberately blocked two existing, ranking blog articles from being crawled by Google.

The setup

On January 30, 2023, Stox added these rules to Ahrefs' robots.txt file:

```
Disallow: /blog/top-bing-searches/
```

```
Disallow: /blog/top-youtube-searches/
```

The two affected articles were “Top Bing Searches” and “Top YouTube Searches.” The experiment was intended to explore how restricting access to

existing content might affect search performance.

What happened?

In his published account, Stox reported several observations:

- Both pages lost some organic traffic, although the overall effect was smaller than he had expected.
- The pages lost their featured snippets while blocked.
- Some individual keyword positions declined, while others improved.
- The search results showed a message indicating that page information was unavailable instead of the usual custom snippet.
- Clicks appeared to decline more noticeably than impressions, which Stox thought might partly reflect the changed search-result presentation.

After the block was removed, the Bing-searches article quickly regained some featured snippets.

What should website owners learn from this?

The experiment does not prove that robots.txt always causes a particular traffic or ranking decline. Stox explicitly noted that he could not confidently attribute every observed change to the blocking itself. It does, however, illustrate a crucial risk: **blocking a page from crawling does not reliably remove it from search results, and it can affect how that result is presented.** If a page should remain indexed and competitive in organic search, make sure your robots.txt rules do not accidentally block it.

The experiment is documented by Ahrefs here: [The Story of Blocking 2 High-Ranking Pages with Robots.txt](#).

Attribution note: This is an experiment conducted and published by Ahrefs, not a case study performed by TSI Digital Solution.

Robots.txt Best Practices

A reliable robots.txt setup starts with a clear objective. Do not add rules simply because another website uses them.

1. Keep valuable pages crawlable

Review the URLs that support your business goals: service pages, product categories, important products, landing pages, and useful blog content.

Confirm that the relevant search engine can crawl these URLs. A rule that unintentionally blocks a high-value section can undermine your wider SEO work.

2. Be careful with wildcard rules

A rule such as Disallow: / can prevent a compliant crawler from accessing every URL covered by that robots.txt group.

This can be appropriate for a site that genuinely should not be crawled by a particular bot. It is dangerous when copied from a staging environment into production without review.

Before every website launch or migration, check for broad disallow rules and confirm that they are intentional.

3. Do not block resources without checking the consequences

Search engines may need access to CSS, JavaScript, images, and other resources to understand a page properly.

Blocking essential resources can make it harder for Google to render and interpret your content. Restrict resources only when there is a clear reason and the effect has been evaluated.

4. Include your XML sitemap when appropriate

A sitemap can help search engines discover URLs you want them to consider for crawling and indexing.

You can identify its location in robots.txt with an absolute URL:

Sitemap: <https://www.example.com/sitemap.xml>

The sitemap and robots.txt should work together. A URL included in your sitemap but blocked from crawling may send conflicting signals about your intended SEO setup.

5. Treat robots.txt as public information

Do not use disallow rules as a way to conceal private areas. The file itself is publicly accessible, and its contents can reveal paths that you would prefer not to advertise.

Use proper authentication, authorization, and server-side protection for confidential resources.

6. Review robots.txt after technical changes

Check the file after:

- Website migrations and redesigns
- CMS or SEO plugin changes
- Changes to ecommerce filters or URL parameters
- Staging-to-production deployments
- Changes to sitemap locations
- Technical SEO fixes involving crawling and indexing

A small configuration change can affect a large number of URLs, so it deserves a place in your technical release checklist.

How to Test Robots.txt and Troubleshoot Problems

If you suspect robots.txt is not working as expected, do not start by adding more directives. First establish what the file contains, what the crawler can access, and what outcome you actually want.

1. Open the live file

Visit: `https://www.yourdomain.com/robots.txt`

Replace yourdomain.com with your actual domain. Check that the file loads correctly and that it contains the rules you intended to publish.

Remember that different hosts or protocols can have separate robots.txt files. Check the exact website origin used by the URLs you are auditing.

2. Inspect important URLs

List your key service pages, product pages, category pages, and high-value articles.

Review the applicable User-agent, Disallow, and Allow rules. Watch for broad patterns that may match more URLs than expected.

For Google-specific checks, use the relevant robots.txt testing facilities available in Google Search Console and follow Google's current documentation.

3. Check indexing separately from crawling

Use URL Inspection in Google Search Console to investigate an affected URL.

Check whether Google can access the page, whether it has been indexed, and whether the reported indexing status indicates a robots.txt restriction. A URL being blocked from crawling and a URL being excluded from the index are different situations.

4. Review the wider technical setup

If the page is crawlable but not indexed, investigate other possible causes: noindex directives, canonical tags, redirects, server responses, internal links, content quality, and whether the page provides distinct value.

Robots.txt is only one part of the diagnosis.

5. Validate the fix

After changing the file, recheck the affected paths and monitor relevant Google Search Console reports. Allow time for search engines to recrawl and process the change.

If robots.txt appears not to be working, confirm that you are editing the correct file, that the rule matches the intended URL, and that the crawler is

one that respects the protocol. Do not assume that every bot follows Google's interpretation.

Can Robots.txt Improve Crawl Budget?

It can help in the right circumstances, but it is not a universal SEO shortcut.

Crawl budget is most relevant to very large websites, websites with rapidly changing content, and sites where search engines face many low-value or duplicate URLs. Most small websites do not need an elaborate crawl-budget strategy.

For a large ecommerce website, reducing crawling of endless filter combinations or other low-value URL patterns may help search engines spend fewer requests on those URLs. But the correct solution depends on how the site generates URLs and which pages should be discovered and indexed. Blocking URLs also does not guarantee that Google will never discover or index them. If the goal is to manage duplicates, consolidate signals, or remove pages from search results, review the appropriate canonicalization and indexing methods as well.

Robots.txt and AI Crawlers: What Website Owners Should Know

Robots.txt is no longer discussed only in the context of traditional search engines. Website owners may also use it to communicate preferences to crawlers associated with AI training, retrieval, or search products.

The 2025 Web Almanac's Generative AI chapter reported that the GPTBot user-agent appeared in around **4.5% of sampled robots.txt files**, up from **2.6% in 2024**. This is evidence of increased use of AI-crawler rules in the measured sample—not a measure of AI citations, website traffic, or the percentage of all websites that receive AI visibility.

Before adding AI-specific rules, identify which crawler you are addressing and what that crawler is used for. Different bots may have different purposes, and a restriction intended for one use should not be assumed to affect every AI-related service.

Also remember:

- Robots.txt rules are voluntary instructions for compliant crawlers, not a technical security barrier.
- Allowing a crawler does not guarantee that your content will be used, cited, or surfaced by an AI product.
- Blocking one crawler does not necessarily block every crawler operated by the same company.
- Decisions about access should reflect your content rights, privacy obligations, commercial priorities, and desired discoverability.

For current measurement context, see the [2025 Web Almanac's Generative AI chapter](#).

What the Web Almanac Data Says About Robots.txt

The HTTP Archive's Web Almanac offers a useful snapshot of how websites implement technical SEO practices at scale. Its 2024 SEO chapter reported that robots.txt requests returned HTTP 200 responses for **83.9% of sampled mobile sites and 83.5% of sampled desktop sites**. In 2022, the corresponding figures were 82.4% for mobile and 81.5% for desktop.

These figures describe the observed HTTP response status—not whether every file had ideal rules or whether every site was correctly configured.

The same chapter found that approximately 97.8% of robots.txt files in the sample were no larger than 100 KB. Google documents a 500 KiB parsing limit, so file size is rarely the first issue most website owners should investigate.

The practical lesson is not to chase a particular statistic. It is to verify that your own file loads, contains deliberate rules, and supports the website's crawling and indexing objectives.

Source: [HTTP Archive Web Almanac 2024 – SEO chapter](#).

How TSI Digital Solution Can Help With Technical SEO

Robots.txt is one file, but the problems around it often involve several connected parts of a website: crawlability, indexation, canonical tags, sitemap quality, redirects, site structure, and page performance.

A focused technical SEO review should establish which URLs matter most to the business, whether search engines can crawl them, and what is preventing the right pages from being discovered or indexed.

At TSI Digital Solution, our [crawl and technical services](#) include technical SEO, crawling, schema, and Core Web Vitals. If your site has recently been redesigned, migrated, or expanded, a structured review can help identify issues before they undermine your organic visibility.

Find the crawl and indexing issues holding your website back

Request a technical quote through our [online technical quote page](#), which includes a free website audit. When you submit your request, ask for your assessment to focus on:

- Robots.txt rules that may block important service, category, product, or blog URLs
- Pages that are indexed, excluded, or blocked unexpectedly
- Sitemap URLs that conflict with crawl or indexation settings

- Canonical tags, redirects, and internal-link paths that may confuse discovery
- Technical issues that should be prioritized by business impact

The aim is a clearer action plan: what is wrong, which pages are affected, and which fixes deserve attention first.

Final Thoughts: A Small File With a Big Technical Role

Robots.txt has survived from the early web into an era of complex websites, search engines, and AI crawlers. Its longevity is a reminder that technical SEO often depends on understanding the fundamentals—not simply adopting the newest tool or trend.

Use robots.txt to manage crawling where appropriate. Use noindex when your objective is to exclude a crawlable page from supported search indexes. Use access controls to protect private information, and review canonical tags, sitemaps, and internal links as part of the wider index-hygiene process. Most importantly, test your configuration after every significant website change.

A healthy robots.txt file does not guarantee rankings. It helps prevent avoidable crawling mistakes so your broader SEO strategy has a stronger technical foundation.

If you are unsure whether search engines can reach the pages that drive your business, start with a technical review before a small configuration issue becomes a larger visibility problem.

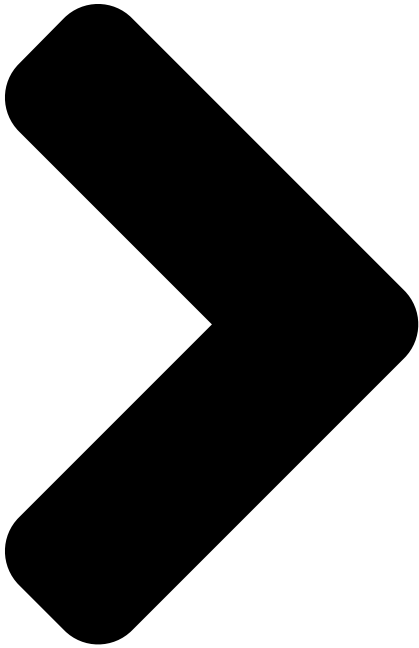
References and further reading

1. IETF – RFC 9309: Robots Exclusion Protocol
2. Google Search Central – Introduction to robots.txt
3. Google Search Central – Block search indexing with noindex
4. Google Search Central – How Google interprets robots.txt
5. Ahrefs – The Story of Blocking 2 High-Ranking Pages With Robots.txt
6. HTTP Archive – Web Almanac 2024: SEO
7. HTTP Archive – Web Almanac 2025: Generative AI

Frequently Asked Questions (FAQ)

What is robots.txt used for in SEO?

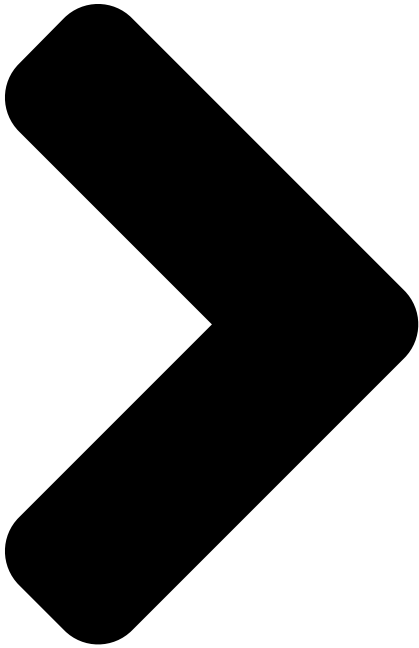




Robots.txt tells compliant crawlers which URL paths they may or may not access. It is commonly used to manage crawling, but it is not a reliable method for removing web pages from search results.

Does robots.txt prevent Google from indexing a page?

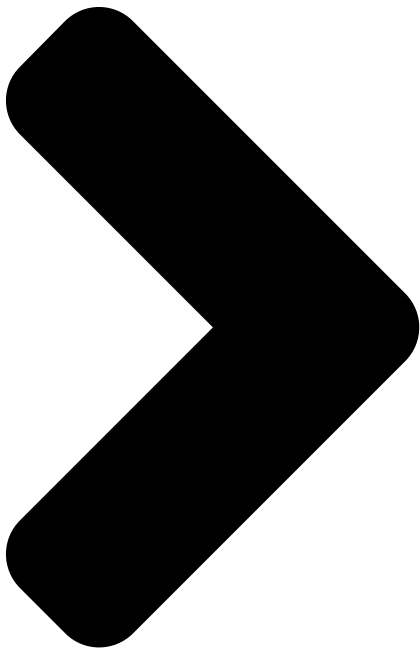




Not necessarily. Google may index a blocked URL if it discovers the URL through other sources. If you need to remove a page from Google Search, use an appropriate method such as a crawlable `noindex` directive, authentication for private content, or removal of the page.

How do I check my robots.txt file?

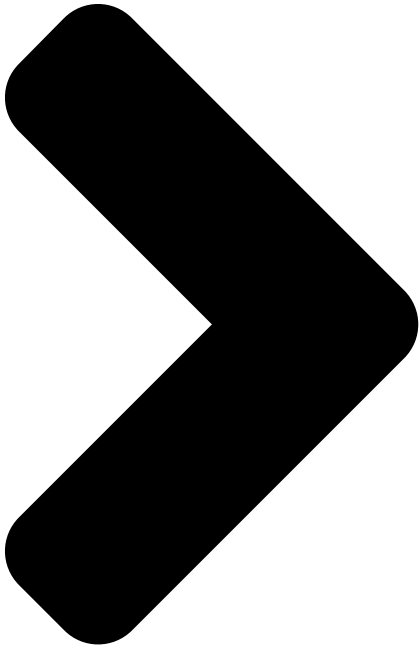




Open `https://yourdomain.com/robots.txt` and review the live rules. Then use Google Search Console and the relevant testing tools to investigate whether important URLs can be crawled and what their indexing status is.

Can robots.txt improve crawl budget?

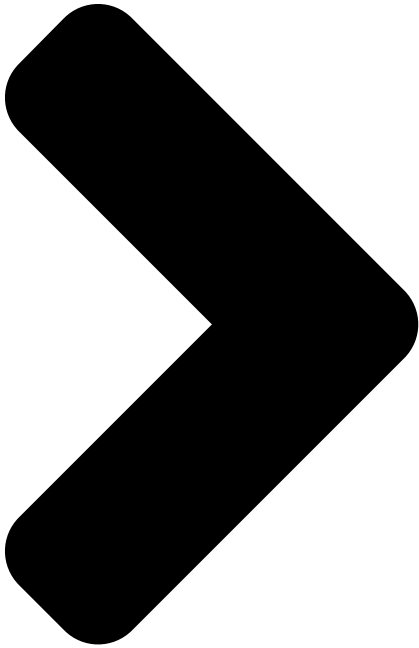




It can help large or complex websites reduce crawling of selected low-value URL patterns. However, most small websites do not need advanced crawl-budget management, and blocking URLs does not automatically solve indexing or duplicate-content problems.

What should I do if robots.txt is not working?

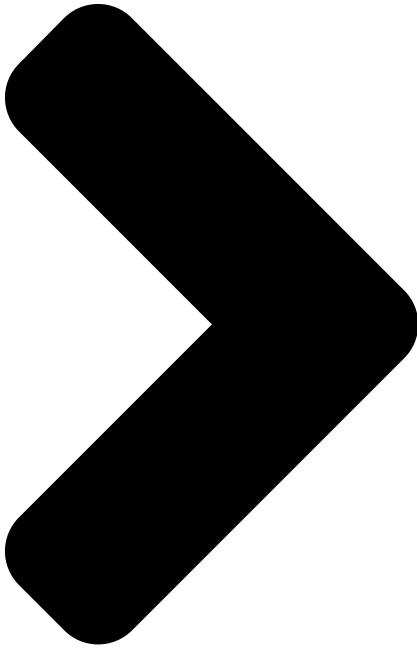




It depends on your goals and the specific crawler. Review the crawler's purpose and your content-use preferences before adding a rule. Remember that robots.txt is not a security control, and allowing a crawler does not guarantee AI search visibility.

Should I block AI crawlers in robots.txt?





Confirm that you are editing the correct file for the URL's origin, check the syntax and matching rules, and verify that the crawler respects the protocol. Then inspect crawling and indexing separately in Search Console. Some bots may ignore robots.txt instructions.

Robots.txt tells compliant crawlers which URL paths they may or may not access. It is commonly used to manage crawling, but it is not a reliable method for removing web pages from search results.

Not necessarily. Google may index a blocked URL if it discovers the URL through other sources. If you need to remove a page from Google Search, use an appropriate method such as a crawlable noindex directive, authentication for private content, or removal of the page.

Open <https://yourdomain.com/robots.txt> and review the live rules. Then use Google Search Console and the relevant testing tools to investigate whether important URLs can be crawled and what their indexing status is.

It can help large or complex websites reduce crawling of selected low-value URL patterns. However, most small websites do not need advanced crawl-budget management, and blocking URLs does not automatically solve indexing or duplicate-content problems.

It depends on your goals and the specific crawler. Review the crawler's purpose and your content-use preferences before adding a rule. Remember that robots.txt is not a security control, and allowing a crawler does not guarantee AI search visibility.

Confirm that you are editing the correct file for the URL's origin, check the syntax and matching rules, and verify that the crawler respects the protocol. Then inspect crawling and indexing separately in Search Console. Some bots may ignore robots.txt instructions.

[Reach Out to Us](#)

Find What's Blocking Your SEO Growth

Get in touch with [TSI Digital Solution](#) to get a free website audit and identify crawlability, indexing, and technical SEO issues holding your website back.

Leave a Reply

Logged in as [TSI Digital Solution](#). [Edit your profile](#). [Log out?](#) Required fields are marked *

Message*

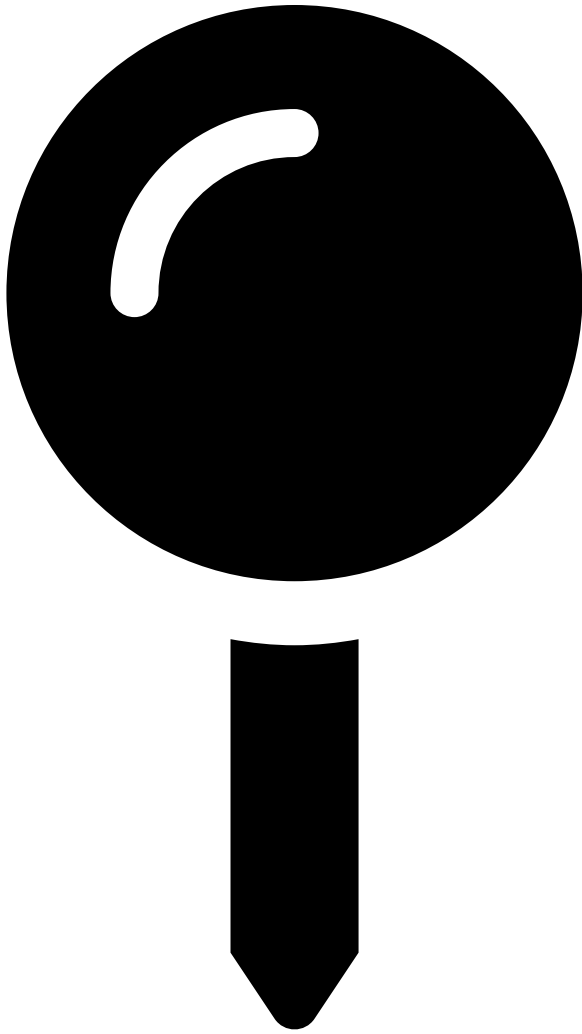
[Go Back >](#)

[Get in Touch](#)



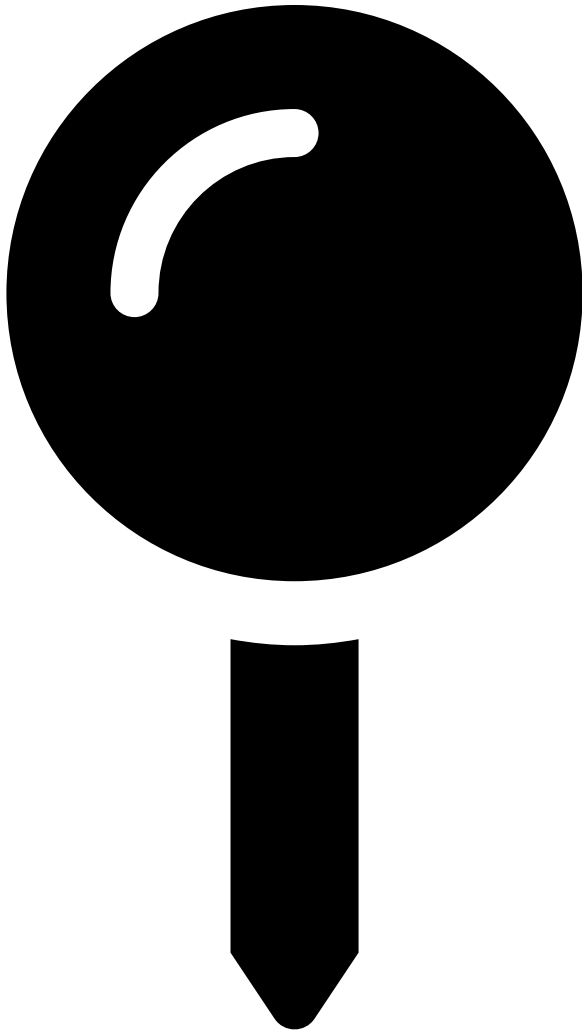
We Reflect Your Wishes

Contact



TSI Digital Solution
(Brand of PT Tripple SoRa Indonesia)

Jl. Sunset Road No.815 Seminyak, Kuta, Badung, Bali – 80361, Indonesia

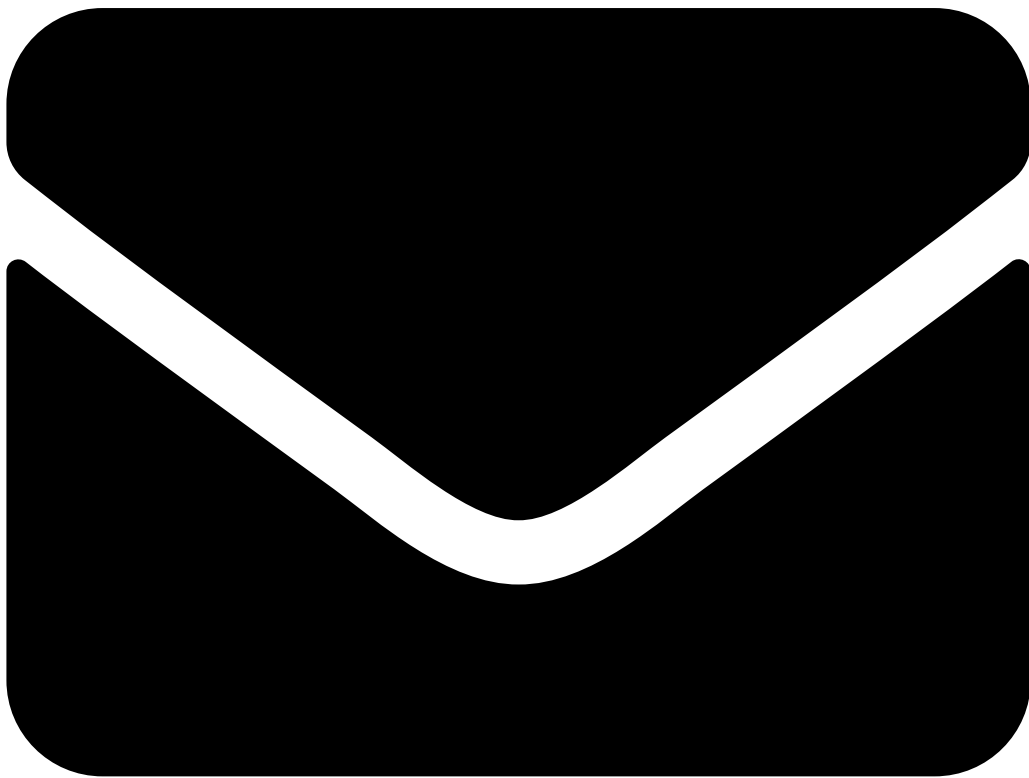


TSI Digital Solution
(Brand of PT Tripple SoRa Indonesia)

Jl. Sunset Road No.815 Seminyak, Kuta, Badung, Bali – 80361, Indonesia



+(62) 813-3936-1587



contact@tsidigitalsolution.com

Services

- Websites / e-Commerce
- Apps
- AI Agents
- Technical/SEO
- Branding
- Social Media
- Graphic Design

- Copywriting

- Photo-& Videography



tsidigitalsolution.com
tsidigitalsolution.be
tsidigitalsolution.nl

Copyright © 2022 –

TSI Digital Solution | All rights reserved.

Edit Template