



**Batch:B1 Roll No.: 16010421119 Experiment No.: 5**

**Aim:** Robot.txt and sitemap creation’

**Resources needed:** Internet, any seo sitemap generator tool, any seo robot.txt generator tool and MS-office

**Theory:**

**Sitemap**

In simple terms, a Sitemap is an XML file that is full of your individual webpage’s URLs. It’s like an archive of every webpage in your website. This file should be easily discoverable in your site in order for search engine crawlers to stumble upon it.

What is a Sitemap for?

A Sitemap is usually used for the purpose of letting the search engine crawlers follow the links to all your individual webpages so that it won’t miss out on anything.

Sometimes we leave out URLs or hide them from all visible pages because we don’t exactly want some of the users to go there. As a result, some of these URLs are uncrawlable to search engine spiders.



We can still leave those URLs hidden from some users without having to lose out on those pages not being crawled by search engine spiders through including them in an XML Sitemap.

**Robots.txt**

Robots.txt is a text file webmasters create to instruct web robots (typically search engine robots) how to crawl pages on their website. The robots.txt file is part of the the robots exclusion protocol (REP), a group of web standards that regulate how robots crawl the web, access and index content, and serve that content up to users. The REP also includes directives like [meta robots](https://moz.com/learn/seo/robots-meta-directives), as well as page-, subdirectory-, or site-wide instructions for how search engines should treat links (such as “follow” or “nofollow”).

In practice, robots.txt files indicate whether certain user agents (web-crawling software) can or cannot crawl parts of a website. These crawl instructions are specified by “disallowing” or “allowing” the behavior of certain (or all) user agents.

**Basic format:**

User-agent: [user-agent name]

Disallow: [URL string not to be crawled]

Here are a few examples of robots.txt in action for a [www.example.com](http://www.example.com/) site: Robots**.txt file URL:** [**www.example.com/robots.txt**](http://www.example.com/robots.txt)

* *Blocking all web crawlers from all content*

User-agent: \* Disallow: /

Using this syntax in a robots.txt file would tell all web crawlers not to crawl any pages on [www.example.com](http://www.example.com/), including the homepage.

* *Allowing all web crawlers access to all content*

User-agent: \* Disallow:

Using this syntax in a robots.txt file tells web crawlers to crawl all pages on [www.example.com](http://www.example.com/), including the homepage.

* Blocking a specific web crawler from a specific folder

User-agent: Googlebot

Disallow: /example-subfolder/

This syntax tells only Google’s crawler (user-agent name Googlebot) not to crawl any pages that contain the URL string [www.example.com/example-subfolder/.](http://www.example.com/example-subfolder/)

* *Blocking a specific web crawler from a specific web page*

User-agent: Bingbot

Disallow: /example-subfolder/blocked-page.html

This syntax tells only Bing’s crawler (user-agent name Bing) to avoid crawling the specific page at [www.example.com/example-subfolder/blocked-page.](http://www.example.com/example-subfolder/blocked-page)

**Results: (Description of selected tool in prescribed format)**

|  |  |
| --- | --- |
| Name Of the Tool (Include Company  Name,website etc. | [**https://www.xml-sitemaps.com/details-www.arya-nai**](http://www.xml-sitemaps.com/details-www.arya-nair.in-524e329e)**r**[**.in-524e329e**](http://www.xml-sitemaps.com/details-www.arya-nair.in-524e329e) **2.html** |
| License/ Open Source | **License** |
| Explanation of Tool |  |
| Procedure | 1. How tool accepts the input?   we just have to enter the website link to the tool   1. How tool processes the data?   the tool crawls through the website and finds all the paths   1. How tool displays the output/result?     This XML file does not appear to have any style information associated with it. The document tree is shown below. |

|  |  |
| --- | --- |
|  | <urlset [xmlns="http://www.sitemaps.org/schemas/sitemap/0.9"](http://www.sitemaps.org/schemas/sitemap/0.9) [xmlns:xsi="http://www](http://www.w3.org/2001/XMLSchema-instance).w3.or[g/2001/XMLSchema-instance"](http://www.w3.org/2001/XMLSchema-instance) [xsi:schemaLocation="http://www](http://www.sitemaps.org/schemas/sitemap/0.9).sitemaps.or[g/schemas/sitemap/0.9](http://www.sitemaps.org/schemas/sitemap/0.9) [http://www.sitemaps.org/schemas/sitemap/0.9/sitemap.xsd">](http://www.sitemaps.org/schemas/sitemap/0.9/sitemap.xsd)  <!-- created with Free Online Sitemap Generator [www.xml-sitemaps.com](http://www.xml-sitemaps.com/) -->  <url>  [<loc>https://www](http://www.arya-nair.in/).arya-nair[.in/</loc>](http://www.arya-nair.in/)  <lastmod>2024-09-03T07:00:10+00:00</lastmod>  </url>  </urlset>  (Attach result Screen shot of tool screen. Specify the sitemap code. Also specify the location of where sitemap is uploaded and screenshot displaying site map on webpage ) |

1. **Create robot.txt having reference of sitemap.xml**



**Update the existing robots.txt** with the sitemap.xml in the following way

Added sitemap to the website as well as update robots.txt

**Questions:**

1. Explain the content duplication issues from SEO point of view.

Certainly! Here’s a concise overview of content duplication issues from an SEO perspective:

# Content Duplication Issues

* 1. **Ranking Dilution**: Duplicate content splits ranking signals, making it harder for search engines to determine the most relevant page, potentially lowering overall rankings.
  2. **Indexing Problems**: Duplicate content can waste crawl budget and prevent important pages from being indexed effectively.
  3. **User Experience**: Users may find duplicate content confusing or low-quality, affecting engagement and trust.
  4. **Potential Penalties**: While not directly penalized, duplicate content can lead to lower

rankings and visibility issues.

* 1. **Link Equity Impact**: Backlinks may be spread across duplicate pages, reducing the strength of each page.

# Best Practices

* + - **Use Canonical Tags**: Indicate the preferred version of content.
    - **301 Redirects**: Redirect duplicate pages to the main version.
    - **Ensure Unique Content**: Keep each page's content distinct and valuable.
    - **Consistent URL Structure**: Avoid duplicates from URL parameters and variations.
    - **Monitor and Audit**: Regularly check for and resolve internal duplicate content.

Addressing these issues helps improve SEO performance and user experience.

**Outcomes:**



**CO2 Understanding search engine optimization essentials**

**Conclusion: (Conclusion to be based on the objectives and outcomes achieved)**

**Added sitemap.xml as well as update robots.txt file in the website**

**Grade: AA / AB / BB / BC / CC / CD /DD**

**Signature of faculty in-charge with date**

**References:**

1. Eric Greenberg, Alexander Kates, “Strategic Digital Marketing: Top Digital Experts Share the Formula for Tangible Returns on Your Marketing Investments” McGraw Hill Education 1st edition, 16 August 2013.
2. Jan Zimmerman “Web Marketing For Dummies” Willy Publishing 3rd Edition, 2011.
3. Jan Zimmerman, Deborah Ng, “ Social Media Marketing All-in-One For Dummies” Willy Publishing 4th Edition, 2017.
4. Eric Enge, Stephan Spencer, Jessie Stricchiola, Rand Fishkin, “The Art of SEO”, 2nd Edition Mastering Search Engine Optimization O'Reilly Media 2nd Edition, 2012
5. John I Jerkovic, “SEO Warriors” O'Reilly Media 1st edition, 2009

