getsitemap

Release 0.1

capjamesg

Aug 30, 2023

CONTENTS

getsitemap is a simple Python library that retrieves all the URLs in the sitemaps associated with a website.

This library may be useful when building a web search crawler, an SEO validation tool, or a sitemap monitor.

You can download *getsitemap* using the following comamnd:

```
pip install getsitemap
```

See the documentation for *getsitemap* below.

CHAPTER

ONE

INDICES AND TABLES

- genindex
- modindex
- search

CHAPTER

TWO

GET ALL URLS IN A WEBSITE'S SITEMAPS

The *retrieve_sitemap_urls()* function returns all URLs found in a website's sitemaps.

This function:

1. Checks for *Sitemap* directives in a website's robots.txt file. All sitemap found are crawled recursively.
2. Checks for the presence of a sitemap.xml file. If one is found, it is crawled recursively.
3. Merges the results of all checks to return either a list of all URLs or a dictionary that maps each URL to the sitemap in which it was found.

To get a list of all sitemaps in a website, you can append *.keys()* to the result of this function, as long as you specify *as_flat_list=False* in the command arguments.

Please note this function may take time to run if there are a lot of sitemaps to crawl. This is because a network request has to be made for each URL.

CHAPTER THREE

GET ALL URLS IN A SINGLE SITEMAP

The *get_individual_sitemap()* function returns all URLs found in a single sitemap.

With the *recurse=True* argument, this function will also crawl all sitemaps found in the sitemap and do so recursively.

If *recurse=False*, this function will return only the list of URLs in the provided sitemap file. This will include sitemap files if you use this function on a sitemap index.

CHAPTER

FOUR

CHANGELOG

All notable changes to this project will be documented in this file.

The format is based on Keep a Changelog, and this project adheres to Semantic Versioning.

4.1 [0.1.1] - 2022-10-09

4.1.1 Added

- Refactored `get_individual_sitemap` to allow use as a public function.
- Documentation for the `get_individual_sitemap` function.

4.2 [0.1.0] - 2022-10-09

4.2.1 Added

- Initial release of `getsitemap`.
- `retrieve_sitemap_urls` to retrieve all the URLs from a sitemap.
- Documentation for the `retrieve_sitemap_urls` function.