

Find Content Duplication:


Description:

Create a simple website that allows users to input a URL and check for duplicate content on a given website. Users should have the following options:

- input the URL of the websites they want to check.
- start a scanning process.
- obtain a report that includes a list of scanned URLs, sections with potential duplicate content, and a summary of the percentage of duplication found.

Note: You have the freedom to choose how the pages of the application will look and add new features if you want.

Similar apps/tools:

-  **Introducing Copyscape**
<https://www.copyscape.com/>
- <https://youtu.be/VZohVRoCunA>
<https://www.siteliner.com/>

Coding resources:

1. <https://www.python.org/doc/>
2. <https://www.w3schools.com/python/>
3. <https://flask.palletsprojects.com/en/3.0.x/>
4. <https://www.geeksforgeeks.org/python-programming-language/>
5. <https://www.geeksforgeeks.org/search-engine-optimization-seo-basics/>

HTML & CSS:

1. <https://www.geeksforgeeks.org/html/>
2. <https://www.geeksforgeeks.org/css/>
3. <https://www.geeksforgeeks.org/javascript/>

Database Resources:

1. <https://www.postgresql.org/docs/>
2. <https://dev.mysql.com/doc/workbench/en/>

GIT & Github:

1. <https://git-scm.com/doc>
2. <https://docs.github.com/en/get-started/quickstart/hello-world>

Libraries and more resource:

1. Requests - <https://docs.python-requests.org/en/latest/>
2. Beautiful Soup - <https://www.crummy.com/software/BeautifulSoup/bs4/doc/>
3. URLLib - <https://docs.python.org/3/library/urllib.html>
4. Regex - <https://docs.python.org/3/library/re.html>
5. HTTP Libraries - <https://docs.python.org/3/library/http.client.html>
6. Multithreading or Multiprocessing -
<https://docs.python.org/3/library/threading.html>
<https://docs.python.org/3/library/multiprocessing.html>