WEB CRAWLER

Introduction:

- **Definition:** A Web crawler, sometimes called a spider or spider bot and often shortened to crawler, is an Internet bot that systematically browses the World Wide Web, typically for the purpose of Web indexing.

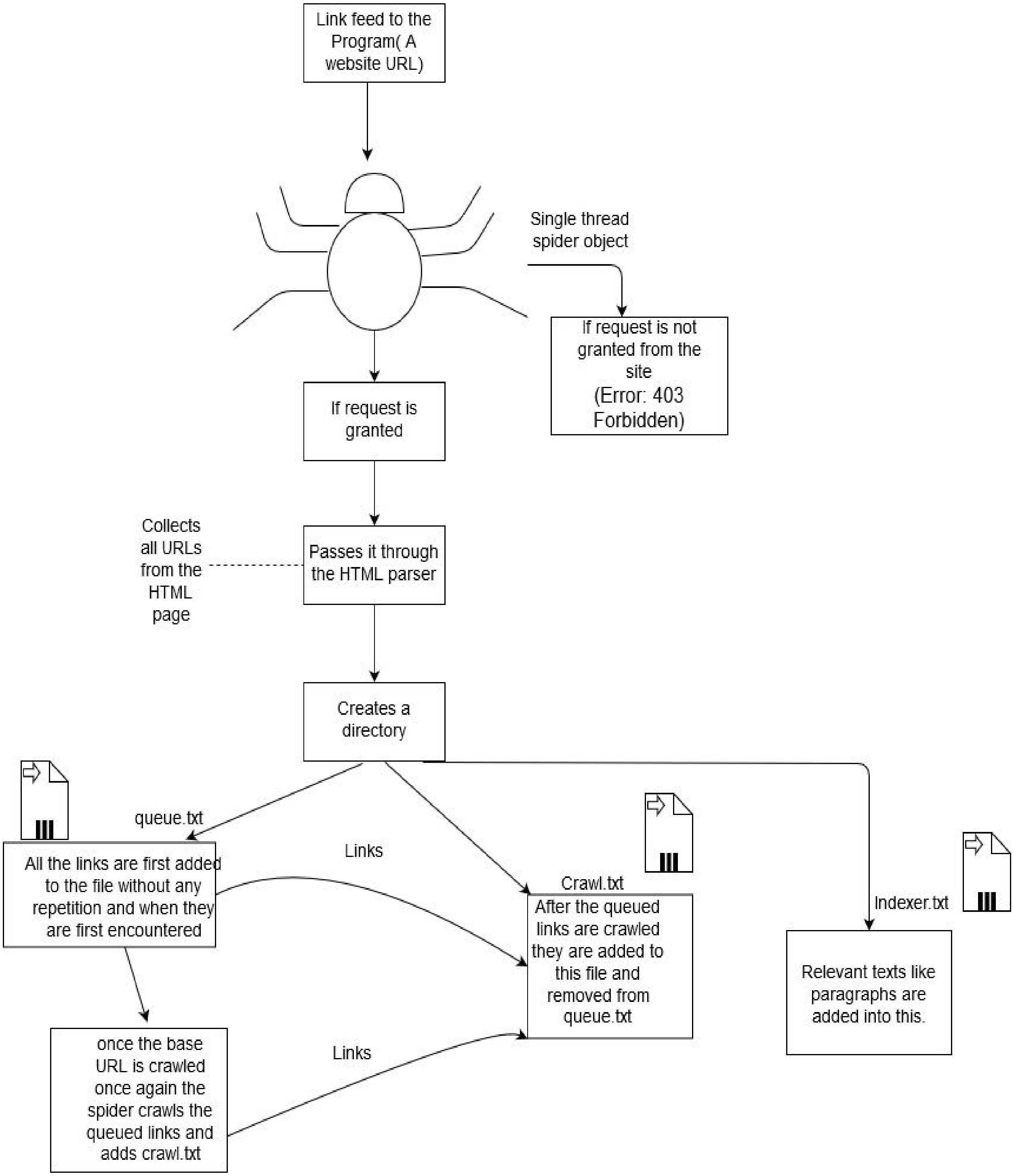
- **Uses:** Web crawlers are mainly used to create a copy of all the visited pages for later processing by a search engine, that will index the downloaded pages to provide fast searches. Crawlers can also be used for automating maintenance tasks on a Web site, such as checking links or validating HTML code.

- **Features:** The key features of Web Crawler implemented in the project are:

    1. The contents of the web page are the corpus from which we retrieve the data later.

    2. The user requests the web page to provide relevant links and text paragraphs and collects them as response.

    3. In this crawler we are crawling through its HTML page and collecting all the relevant links and paragraphs and storing them in different files.


Link feed to the Program(A website URL)
Single thread spider object
If request is not granted from the site (Error: 403 Forbidden)
If request is granted
Collects all URLs from the HTML page
Passes it through the HTML parser
Creates a directory
queue.txt
All the links are first added to the file without any repetition and when they are first encountered
Links
Crawl.txt
After the queued links are crawled they are added to this file and removed from queue.txt
Indexer.txt
Relevant texts like paragraphs are added into this.
once the base URL is crawled once again the spider crawls the queued links and adds crawl.txt
Links

Approach and Workflow-

Crawlers can be of many types,we have implemented the one which would a website URL as the user input and give us a file collecting all the links in it and relevant text inside each page of that particular webpage.
We have implemented a two level crawler.In the first level it gets the url and crawls it and after getting all the links in a file in the second level we are again crawling every individual link.
We have used the python programming language to implement the crawler.
Files in our Code

1.main.py
2.functions.py
3.htmlpars.py
4.crawler.py
5.textinfo.py

main.py is for the user reference where one will give the URL of a particular website as the input to the program.

crawler.py has the spider class that is responsible for collecting the user fed link, parse it using htmlparser.py and use the functions.py to get all the possible URLs.

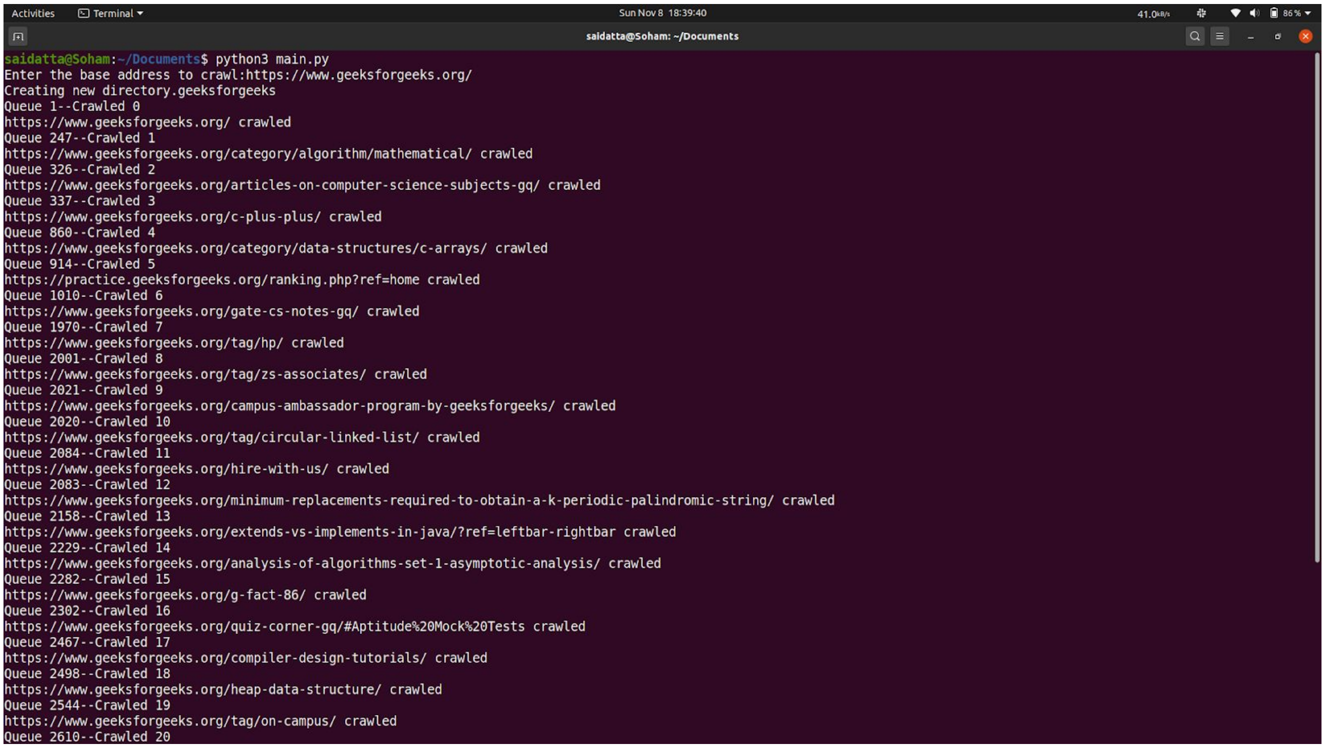
functions.py contains some user defined functions to create a directory,create queue and crawl text files,appending and deleting from files,files to set and set file conversions, getting project names and domain and subdomain names.

htmlpars.py will create an object of find_link( ) class which is a child class of HTMLParser,By getting a link it will parse its HTML page and get all the **<a>** tags with attribute type **“href”=** and return it in the form of set( ).These are the hidden links in a website in the from of buttons, texts, images which are navigable redirects.

textinfo.py is used to extract all the **<p>** tag values so as to create a corpus of text data that would be later indexed.

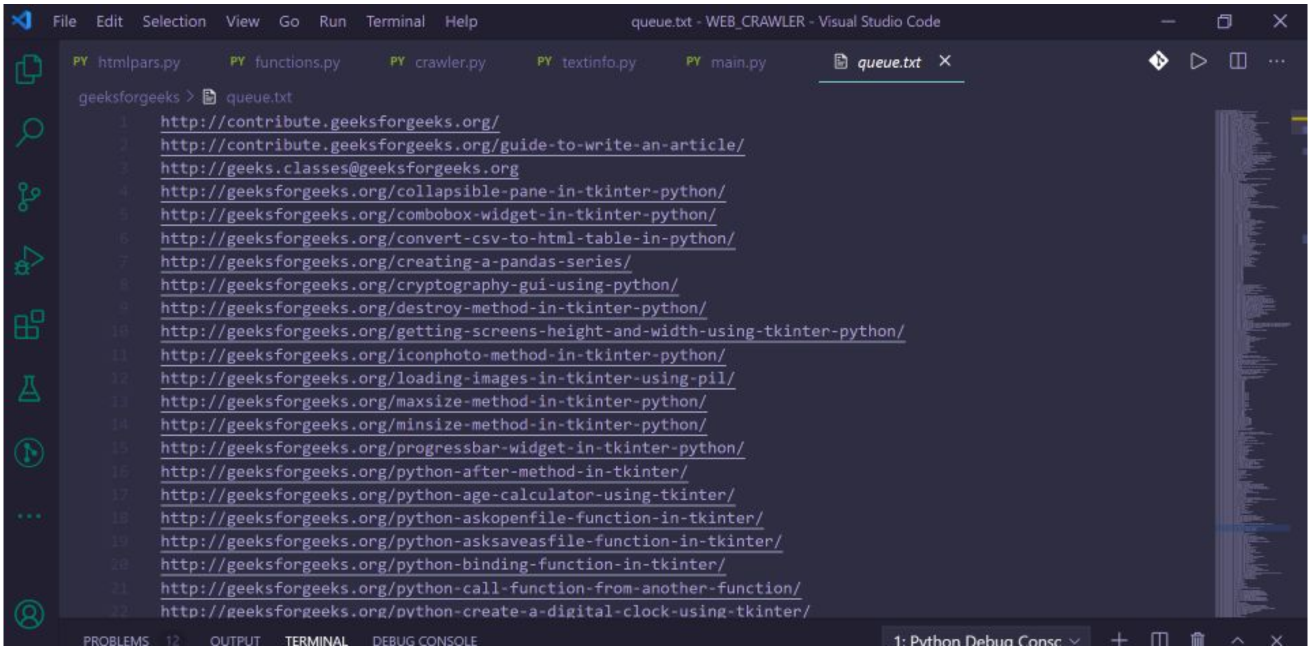
Finally we will get a Project directory consisting of three files queue.txt,crawl.txt and indexer.txt.

User Input:



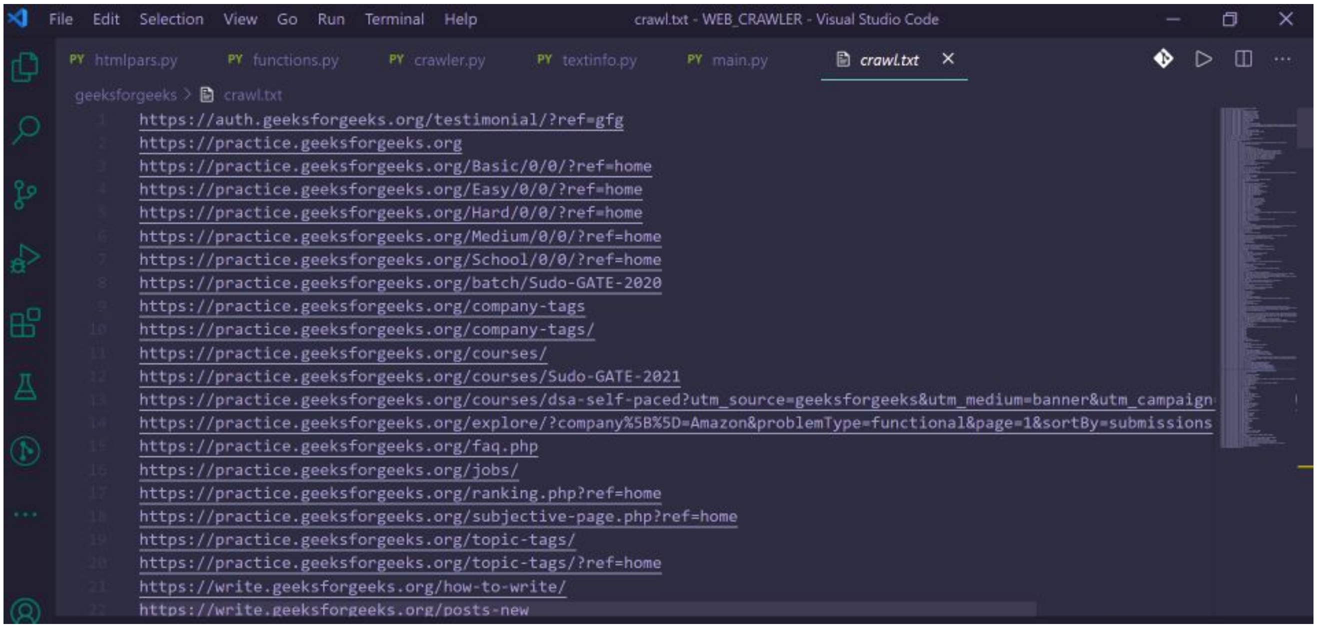
```
saidatta@Soham:~/Documents$ python3 main.py
Enter the base address to crawl:https://www.geeksforgeeks.org/
Creating new directory.geeksforgeeks
Queue 1--Crawled 0
https://www.geeksforgeeks.org/ crawled
Queue 247--Crawled 1
https://www.geeksforgeeks.org/category/algorithm/mathematical/ crawled
Queue 326--Crawled 2
https://www.geeksforgeeks.org/articles-on-computer-science-subjects-gq/ crawled
Queue 337--Crawled 3
https://www.geeksforgeeks.org/c-plus-plus/ crawled
Queue 860--Crawled 4
https://www.geeksforgeeks.org/category/data-structures/c-arrays/ crawled
Queue 914--Crawled 5
https://practice.geeksforgeeks.org/ranking.php?ref=home crawled
Queue 1010--Crawled 6
https://www.geeksforgeeks.org/gate-cs-notes-gq/ crawled
Queue 1970--Crawled 7
https://www.geeksforgeeks.org/tag/hp/ crawled
Queue 2001--Crawled 8
https://www.geeksforgeeks.org/tag/zs-associates/ crawled
Queue 2021--Crawled 9
https://www.geeksforgeeks.org/campus-ambassador-program-by-geeksforgeeks/ crawled
Queue 2020--Crawled 10
https://www.geeksforgeeks.org/tag/circular-linked-list/ crawled
Queue 2084--Crawled 11
https://www.geeksforgeeks.org/hire-with-us/ crawled
Queue 2083--Crawled 12
https://www.geeksforgeeks.org/minimum-replacements-required-to-obtain-a-k-periodic-palindromic-string/ crawled
Queue 2158--Crawled 13
https://www.geeksforgeeks.org/extends-vs-implements-in-java/?ref=leftbar-rightbar crawled
Queue 2229--Crawled 14
https://www.geeksforgeeks.org/analysis-of-algorithms-set-1-asymptotic-analysis/ crawled
Queue 2282--Crawled 15
https://www.geeksforgeeks.org/g-fact-86/ crawled
Queue 2302--Crawled 16
https://www.geeksforgeeks.org/quiz-corner-gq/#Aptitude%20Mock%20Tests crawled
Queue 2467--Crawled 17
https://www.geeksforgeeks.org/compiler-design-tutorials/ crawled
Queue 2498--Crawled 18
https://www.geeksforgeeks.org/heap-data-structure/ crawled
Queue 2544--Crawled 19
https://www.geeksforgeeks.org/tag/on-campus/ crawled
Queue 2610--Crawled 20
```

1. **queue.txt** contains all the URLs that we get after crawling the base URL which can be further crawled.



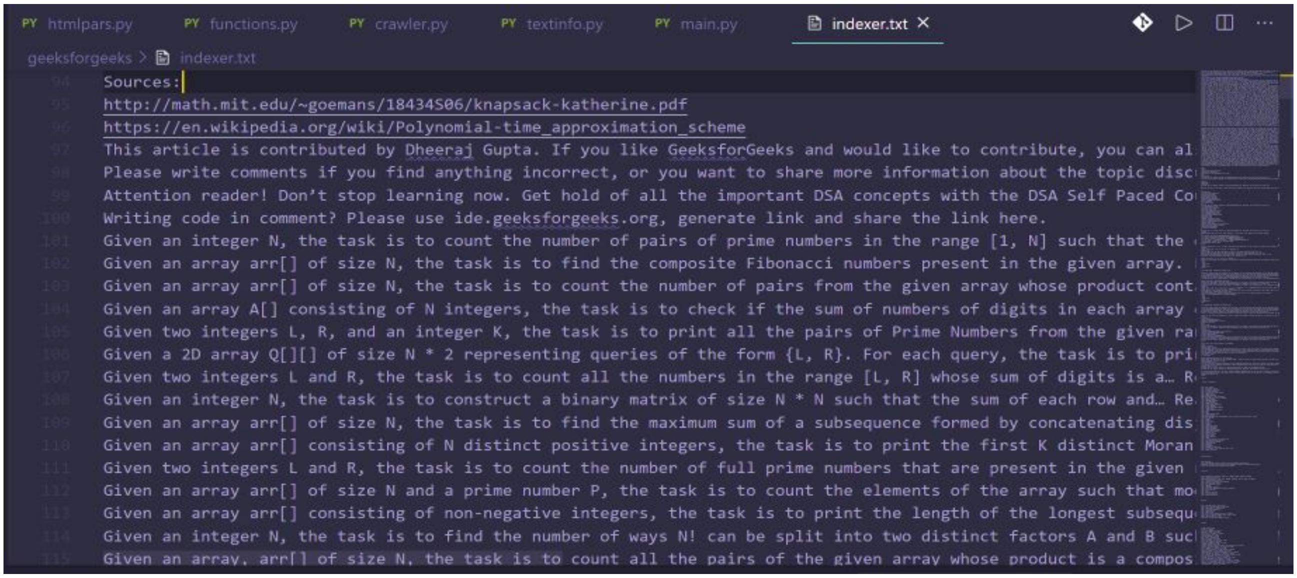
```
http://contribute.geeksforgeeks.org/
http://contribute.geeksforgeeks.org/guide-to-write-an-article/
http://geeks.classes@geeksforgeeks.org
http://geeksforgeeks.org/collapsible-pane-in-tkinter-python/
http://geeksforgeeks.org/combobox-widget-in-tkinter-python/
http://geeksforgeeks.org/convert-csv-to-html-table-in-python/
http://geeksforgeeks.org/creating-a-pandas-series/
http://geeksforgeeks.org/cryptography-gui-using-python/
http://geeksforgeeks.org/destroy-method-in-tkinter-python/
http://geeksforgeeks.org/getting-screens-height-and-width-using-tkinter-python/
http://geeksforgeeks.org/iconphoto-method-in-tkinter-python/
http://geeksforgeeks.org/loading-images-in-tkinter-using-pil/
http://geeksforgeeks.org/maxsize-method-in-tkinter-python/
http://geeksforgeeks.org/minsize-method-in-tkinter-python/
http://geeksforgeeks.org/progressbar-widget-in-tkinter-python/
http://geeksforgeeks.org/python-after-method-in-tkinter/
http://geeksforgeeks.org/python-age-calculator-using-tkinter/
http://geeksforgeeks.org/python-askopenfile-function-in-tkinter/
http://geeksforgeeks.org/python-asksaveasfile-function-in-tkinter/
http://geeksforgeeks.org/python-binding-function-in-tkinter/
http://geeksforgeeks.org/python-call-function-from-another-function/
http://geeksforgeeks.org/python-create-a-digital-clock-using-tkinter/
```

2. **crawl.text** is the file having all the URLs which are crawled.



```
https://auth.geeksforgeeks.org/testimonial/?ref=gfg
https://practice.geeksforgeeks.org
https://practice.geeksforgeeks.org/Basic/0/0/?ref=home
https://practice.geeksforgeeks.org/Easy/0/0/?ref=home
https://practice.geeksforgeeks.org/Hard/0/0/?ref=home
https://practice.geeksforgeeks.org/Medium/0/0/?ref=home
https://practice.geeksforgeeks.org/School/0/0/?ref=home
https://practice.geeksforgeeks.org/batch/Sudo-GATE-2020
https://practice.geeksforgeeks.org/company-tags
https://practice.geeksforgeeks.org/company-tags/
https://practice.geeksforgeeks.org/courses/
https://practice.geeksforgeeks.org/courses/Sudo-GATE-2021
https://practice.geeksforgeeks.org/courses/dsa-self-paced?utm_source=geeksforgeeks&utm_medium=banner&utm_campaign
https://practice.geeksforgeeks.org/explore/?company%5B%5D=Amazon&problemType=functional&page=1&sortBy=submissions
https://practice.geeksforgeeks.org/faq.php
https://practice.geeksforgeeks.org/jobs/
https://practice.geeksforgeeks.org/ranking.php?ref=home
https://practice.geeksforgeeks.org/subjective-page.php?ref=home
https://practice.geeksforgeeks.org/topic-tags/
https://practice.geeksforgeeks.org/topic-tags/?ref=home
https://write.geeksforgeeks.org/how-to-write/
https://write.geeksforgeeks.org/posts-new
```

3. **indexer.txt** contains the all relevant text inside a particular link



```
Sources:
http://math.mit.edu/~goemans/18434S06/knapsack-katherine.pdf
https://en.wikipedia.org/wiki/Polynomial-time_approximation_scheme
This article is contributed by Dheeraj Gupta. If you like GeeksforGeeks and would like to contribute, you can al
Please write comments if you find anything incorrect, or you want to share more information about the topic disc
Attention reader! Don't stop learning now. Get hold of all the important DSA concepts with the DSA Self Paced Co
Writing code in comment? Please use ide.geeksforgeeks.org, generate link and share the link here.
Given an integer N, the task is to count the number of pairs of prime numbers in the range [1, N] such that the
Given an array arr[] of size N, the task is to find the composite Fibonacci numbers present in the given array.
Given an array arr[] of size N, the task is to count the number of pairs from the given array whose product cont
Given an array A[] consisting of N integers, the task is to check if the sum of numbers of digits in each array
Given two integers L, R, and an integer K, the task is to print all the pairs of Prime Numbers from the given ra
Given a 2D array Q[][] of size N * 2 representing queries of the form {L, R}. For each query, the task is to pri
Given two integers L and R, the task is to count all the numbers in the range [L, R] whose sum of digits is a… R
Given an integer N, the task is to construct a binary matrix of size N * N such that the sum of each row and… Re
Given an array arr[] of size N, the task is to find the maximum sum of a subsequence formed by concatenating dis
Given an array arr[] consisting of N distinct positive integers, the task is to print the first K distinct Moran
Given two integers L and R, the task is to count the number of full prime numbers that are present in the given
Given an array arr[] of size N and a prime number P, the task is to count the elements of the array such that mo
Given an array arr[] consisting of non-negative integers, the task is to print the length of the longest subsequ
Given an integer N, the task is to find the number of ways N! can be split into two distinct factors A and B suc
Given an array, arr[] of size N, the task is to count all the pairs of the given array whose product is a compos
```

Shortcoming of our web crawler:

The web crawler is not able to crawl certain websites due to **403 Forbidden error,** we have implemented an exception handling for that.
The program being single threaded takes some time to crawl the website.
It will not work for files such as pdfs, text files,images to extract links in that.

By:
Rohit Kumar Behera
Saanidhya Vats
Saidatta Sahu
Samprita Sahoo