

Project Report

Mini Search Engine

Yogita Ramesh Bugade | 10476645

A trie (pronounced "try") is a tree-based data structure used to store strings to facilitate quick pattern matching. Tries are most used in information retrieval. The term "trie" is derived from the word "retrieval." We are given a collection S of strings, all defined using the same alphabet, in an information retrieval application, such as a search for a certain DNA sequence in a genomic database.

Theorem:

Let X be a text string of n characters from a d -dimensional alphabet. With the suffix trie of X , which consumes $O(n)$ space and can be created in $O(dn)$ time, we may conduct pattern matching queries on X in $O(dm)$ time, where m is the length of the pattern.

One of the applications that used trie data structure is as follows:

Search Engines:

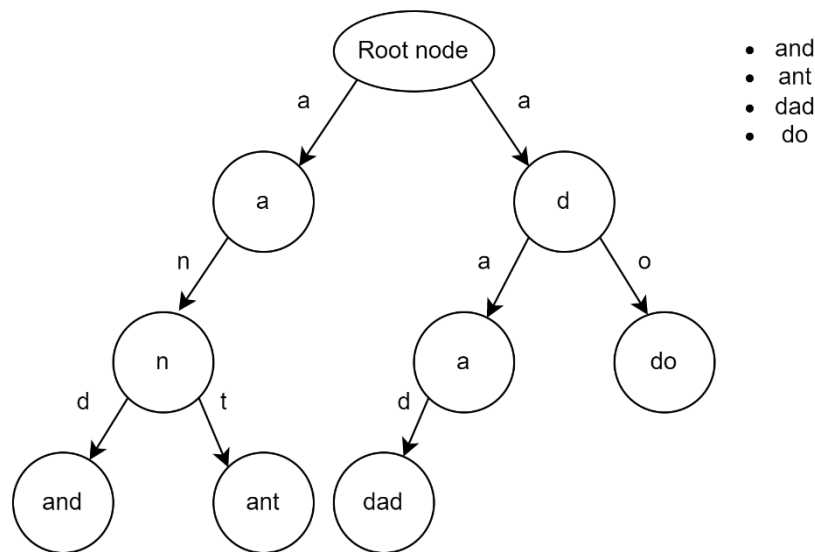
The World Wide Web has a massive library of text documents (web pages). Information about these pages is obtained by a computer called a web crawler, which then stores this information in a particular dictionary database. An online search engine enables users to access relevant information from this database, finding relevant web pages containing provided keywords. We give a simple form of a search engine in this section.

Search Engine using Trie Data Structure:

Approach:

- I constructed the simplified Search Engine specified in Section 23.6.4 for the pages of a tiny Web site. • I used all the words in the site's pages as index terms, eliminating stop words such as articles, prepositions, and pronouns, which I stored in a file and worked on.
- The trie data structure is implemented using the HashMap, but there are many alternative techniques for implementing tries, such as arrays and linked lists, but for our situation, the HashMap would perform more efficiently, therefore I used it.
- I have used Jsoup to work with HTML and to read the web pages in from the internet using Jsoup jar mentioned by our professor (Make sure to install the Jsoup jar before running the program from the below site) <https://jsoup.org/download>

Trie is a sort of k -ary search tree that is used to store and search for a certain key in a set. Search complexity can be reduced to an ideal level using Trie (key length). A well-balanced binary search tree will require time proportional to $M * \log N$, where M is the maximum string length and N is the number of keys in the tree. The key can be found in $O(M)$ time using Trie.



Ranking is based on the number of occurrences of the word on the page:

Continue looping using `String#indexOf (str, fromIndex)`, whereas the String may be discovered using the second argument, which specifies the index to start looking from.

Searching a word using trie data structure:

Generic explanation:

The insert procedure is analogous to searching for a key. It merely compares the characters and then proceeds down. The search may conclude due to the end of a string or the absence of a key in the trie. In the former scenario, if the last node's `isEndofWord` field is true, the key exists in the trie. Because the key is not included in the trie, the search finishes without evaluating all of its characters in the second case. Hashset is used to count the number of times a word appears on each webpage. The HashMap would display their word count, and if the term is not found in any of the documents, the count would be 0. To eliminate the stop words, as we read each word from the paragraph of the webpage, we feed that word to the stopword function, which has a list of all stop words (`stopwords.txt`). If a match is detected, the word is not added to the trie data structure or the hash map. In addition, to show the total of all the terms found in all the documents, I kept a second map that shows the words as well as their total occurrences found in all the documents.

Following Data structures and technologies are used to develop this Web Search Engine:

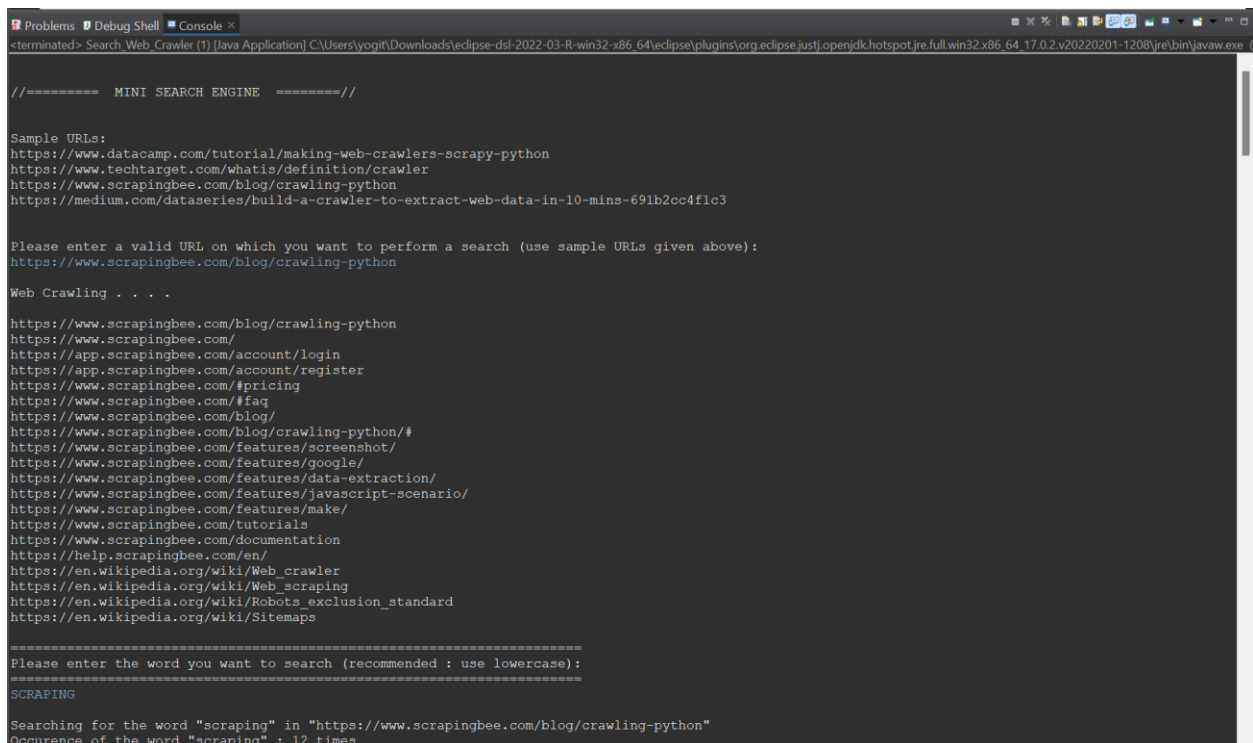
- 1) Web Crawler (Using Trie Data Structure, LinkedList)
- 2) Ranking occurrence (Using Hashset, List, Set, `String#indexOf`)
- 3) Word Search (Using Trie Data Structure, Hashset, LinkedList)
- 4) Conversion of HTML Data to Text (Using JSoup)
- 5) Execution time (Using `currentTimeMillis`)
- 6) StopWords (Using Arrays, List)

For Multiple Keywords:

For multiple keywords, we try to search the sites containing the word and order them first by the number of keywords in the search query, and then by the number of individual occurrences.

Approach, Algorithms used and code walkthrough:

- Sample URLs are provided in the Input.txt file which are used as input to crawl the web pages. Website class contains search_url method which sets the page limit size to less than 20. This class later calls the web_crawl method present in the Website_Search class. The web_crawl method uses Jsoup which later uses built-in Java classes Connection/URLConnection to connect to the URL. It later fetches the pages using href which is an attribute of the anchor tag, which is also used to identify sections within a document and prints the links as displayed below. It adds all the fetched URLs to a LinkedList data structure.
- After crawling through the websites, it prompts for the word to be searched. The searched word is converted to lowercase. If the searched word is equal to any of the stop words mentioned in StopWords java class file, it sends a message that it is a stop word. Please try some other word, which is done by using the isStopWord method in the StopWords class file. All the articles, pronouns, prepositions, HTML, alphabets are stored in the StopWords class file.



```
//===== MINI SEARCH ENGINE =====//

Sample URLs:
https://www.datacamp.com/tutorial/making-web-crawlers-scrappy-python
https://www.techtarget.com/whatis/definition/crawler
https://www.scrapingbee.com/blog/crawling-python
https://medium.com/dataseries/build-a-crawler-to-extract-web-data-in-10-mins-691b2cc4f1c3

Please enter a valid URL on which you want to perform a search (use sample URLs given above):
https://www.scrapingbee.com/blog/crawling-python

Web Crawling . . . .

https://www.scrapingbee.com/blog/crawling-python
https://www.scrapingbee.com/
https://app.scrapingbee.com/account/login
https://app.scrapingbee.com/account/register
https://www.scrapingbee.com/#pricing
https://www.scrapingbee.com/#faq
https://www.scrapingbee.com/blog/
https://www.scrapingbee.com/blog/crawling-python/#
https://www.scrapingbee.com/features/screenshot/
https://www.scrapingbee.com/features/google/
https://www.scrapingbee.com/features/data-extraction/
https://www.scrapingbee.com/features/javascript-scenario/
https://www.scrapingbee.com/features/make/
https://www.scrapingbee.com/tutorials
https://www.scrapingbee.com/documentation
https://help.scrapingbee.com/en/
https://en.wikipedia.org/wiki/Web_crawler
https://en.wikipedia.org/wiki/Web_scraping
https://en.wikipedia.org/wiki/Robots_exclusion_standard
https://en.wikipedia.org/wiki/Sitemaps

Please enter the word you want to search (recommended : use lowercase):
SCRAPING

Searching for the word "scraping" in "https://www.scrapingbee.com/blog/crawling-python"
Occurrence of the word "scraping" : 12 times
```

- The URLs that contain the searched word are fetched using the search_for_word method used in the Website_Search class file. The search_for_word method converts the word into an array and later calls the search_words method, which creates a HashSet output, later calling the method search_word which converts the string into characters and later uses trie data structure mentioned in the Trie class, which later adds all the URLs matching the word.

- Once the urls are fetched that contains the searched word. Using a url at a time the number of occurrences are then calculated, to find the total count, using String#indexOf (str, fromIndex) while the String can be found using the second argument that indicates the index to start searching from. Note: Occurrence of exact words searched a counted. Later it counts the total number of urls using urls. size().

```

Problems | Debug Shell | Console x
<terminated> Search Web Crawler (1) [Java Application] C:\Users\yogit\Downloads\ eclipse-dsl-2022-03-R-win32-x86_64\ eclipse\plugins\org.eclipse.justi.openjdk.hotspot.jre.full.win32.x86_64.17.0.2.v20220201-1208\jre\bin\javaw.exe

Searching for the word "scraping" in "https://www.scrapingbee.com/blog/crawling-python"
Occurrence of the word "scraping" : 12 times

Searching for the word "scraping" in "https://en.wikipedia.org/wiki/Sitemaps"
Searching for the word "scraping" in "https://www.scrapingbee.com/features/data-extraction/"
Occurrence of the word "scraping" : 15 times

Searching for the word "scraping" in "https://www.scrapingbee.com/#faq"
Occurrence of the word "scraping" : 19 times

Searching for the word "scraping" in "https://www.scrapingbee.com/"
Occurrence of the word "scraping" : 19 times

Searching for the word "scraping" in "https://www.scrapingbee.com/features/screenshot/"
Occurrence of the word "scraping" : 9 times

Searching for the word "scraping" in "https://app.scrapingbee.com/account/register"
Searching for the word "scraping" in "https://en.wikipedia.org/wiki/Web_scraping"
Occurrence of the word "scraping" : 51 times

Searching for the word "scraping" in "https://www.scrapingbee.com/#pricing"
Occurrence of the word "scraping" : 19 times

Searching for the word "scraping" in "https://www.scrapingbee.com/features/google/"
Occurrence of the word "scraping" : 8 times

Searching for the word "scraping" in "https://www.scrapingbee.com/features/make/"
Occurrence of the word "scraping" : 8 times

Searching for the word "scraping" in "https://www.scrapingbee.com/blog/crawling-python/#"
Occurrence of the word "scraping" : 12 times

Searching for the word "scraping" in "https://www.scrapingbee.com/tutorials"
Occurrence of the word "scraping" : 3 times

Searching for the word "scraping" in "https://www.scrapingbee.com/features/javascript-scenario/"
Occurrence of the word "scraping" : 9 times

Searching for the word "scraping" in "https://www.scrapingbee.com/blog/"
Occurrence of the word "scraping" : 28 times

Searching for the word "scraping" in "https://en.wikipedia.org/wiki/Robots_exclusion_standard"
Searching for the word "scraping" in "https://en.wikipedia.org/wiki/Web_crawler"
Occurrence of the word "scraping" : 4 times

```

- List of all the urls is displayed using the iterator else it displays message No urls contains the word.
- At the end, Execution time completion message and time required to complete the execution is displayed in minutes.
- Use exit word to end the execution

```

Problems | Debug Shell | Console x
<terminated> Search Web Crawler (1) [Java Application] C:\Users\yogit\Downloads\ eclipse-dsl-2022-03-R-win32-x86_64\ eclipse\plugins\org.eclipse.justi.openjdk.hotspot.jre.full.win32.x86_64.17.0.2.v20220201-1208\jre\bin\javaw.exe

Searching for the word "scraping" in "https://en.wikipedia.org/wiki/Robots_exclusion_standard"
Occurrence of the word "scraping" : 4 times

Searching for the word "scraping" in "https://www.scrapingbee.com/documentation"
Occurrence of the word "scraping" : 334 times

Number of URLs that contain the word "scraping" : 18

List of the URLs that contains the word "scraping":
https://www.scrapingbee.com/blog/crawling-python
https://en.wikipedia.org/wiki/Sitemaps
https://www.scrapingbee.com/features/data-extraction/
https://www.scrapingbee.com/#faq
https://www.scrapingbee.com/
https://www.scrapingbee.com/features/screenshot/
https://app.scrapingbee.com/account/register
https://en.wikipedia.org/wiki/Web_scraping
https://www.scrapingbee.com/#pricing
https://www.scrapingbee.com/features/google/
https://www.scrapingbee.com/features/make/
https://www.scrapingbee.com/blog/crawling-python/#
https://www.scrapingbee.com/tutorials
https://www.scrapingbee.com/features/javascript-scenario/
https://www.scrapingbee.com/blog/
https://en.wikipedia.org/wiki/Robots_exclusion_standard
https://en.wikipedia.org/wiki/Web_crawler
https://www.scrapingbee.com/documentation

Execution Completed

Total time required for execution : 0.22656666 minutes

Please enter the word you want to search (recommended : use lowercase):
THE

"the" is one of the stop word.Try searching so other word.

Please enter the word you want to search (recommended : use lowercase):

```

Sample Execution Screenshots

```
Problems | Debug Shell | Console x
<terminated> Search_Web_Crawler (1) [Java Application] C:\Users\yogit\Downloads\ eclipse-dsl-2022-03-R-win32-x86_64\ eclipse\plugins\org.eclipse.justi.openjdk.hotspot.jre.full.win32.x86_64.17.0.2.v20220201-1208\jre\bin\javaw.exe
Please enter the word you want to search (recommended : use lowercase):
=====
is

Searching for the word "is" in "https://www.scrapingbee.com/blog/crawling-python"
Occurence of the word "is" : 76 times

Searching for the word "is" in "https://www.scrapingbee.com/features/data-extraction/"
Occurence of the word "is" : 17 times

Searching for the word "is" in "https://en.wikipedia.org/wiki/Sitemaps"
Occurence of the word "is" : 126 times

Searching for the word "is" in "https://www.scrapingbee.com/#faq"
Occurence of the word "is" : 21 times

Searching for the word "is" in "https://www.scrapingbee.com/"
Occurence of the word "is" : 21 times

Searching for the word "is" in "https://www.scrapingbee.com/features/screenshot/"
Occurence of the word "is" : 10 times

Searching for the word "is" in "https://www.scrapingbee.com/#pricing"
Occurence of the word "is" : 21 times

Searching for the word "is" in "https://www.scrapingbee.com/features/google/"
Occurence of the word "is" : 49 times

Searching for the word "is" in "https://www.scrapingbee.com/features/make/"
Occurence of the word "is" : 13 times

Searching for the word "is" in "https://www.scrapingbee.com/blog/crawling-python/#"
Occurence of the word "is" : 76 times

Searching for the word "is" in "https://www.scrapingbee.com/features/javascript-scenario/"
Occurence of the word "is" : 10 times

Searching for the word "is" in "https://en.wikipedia.org/wiki/Robots_exclusion_standard"
Occurence of the word "is" : 118 times

Searching for the word "is" in "https://en.wikipedia.org/wiki/Web_crawler"
Occurence of the word "is" : 274 times

Number of URLs that contain the word "is" : 13
```

```
Problems | Debug Shell | Console x
<terminated> Search_Web_Crawler (1) [Java Application] C:\Users\yogit\Downloads\ eclipse-dsl-2022-03-R-win32-x86_64\ eclipse\plugins\org.eclipse.justi.openjdk.hotspot.jre.full.win32.x86_64.17.0.2.v20220201-1208\jre\bin\javaw.exe
List of the URLs that contains the word "is":
https://www.scrapingbee.com/blog/crawling-python
https://www.scrapingbee.com/features/data-extraction/
https://en.wikipedia.org/wiki/Sitemaps
https://www.scrapingbee.com/#faq
https://www.scrapingbee.com/
https://www.scrapingbee.com/features/screenshot/
https://www.scrapingbee.com/#pricing
https://www.scrapingbee.com/features/google/
https://www.scrapingbee.com/features/make/
https://www.scrapingbee.com/blog/crawling-python/#
https://www.scrapingbee.com/features/javascript-scenario/
https://en.wikipedia.org/wiki/Robots_exclusion_standard
https://en.wikipedia.org/wiki/Web_crawler

Execution Completed

Total time required for execution : 0.4096333 minutes

=====
Please enter the word you want to search (recommended : use lowercase):
=====
for

"for" is one of the stop word.Try searching so other word.

=====
Please enter the word you want to search (recommended : use lowercase):
=====
bee

Searching for the word "bee" in "https://en.wikipedia.org/wiki/Sitemaps"
Searching for the word "bee" in "https://en.wikipedia.org/wiki/Robots_exclusion_standard"
Occurence of the word "bee" : 2 times

Searching for the word "bee" in "https://en.wikipedia.org/wiki/Web_crawler"
Occurence of the word "bee" : 5 times

Number of URLs that contain the word "bee" : 3

List of the URLs that contains the word "bee":
https://en.wikipedia.org/wiki/Sitemaps
https://en.wikipedia.org/wiki/Robots_exclusion_standard
```

```

Problems | Debug Shell | Console x
<terminated> Search Web Crawler (1) [Java Application] C:\Users\yogit\Downloads\ eclipse-dsl-2022-03-R-win32-x86_64\eclipse\plugins\org.eclipse.justi.openjdk.hotspot.jre.full.win32.x86_64.17.0.2.v20220201-1208\jre\bin\javaw.exe
List of the URLs that contains the word "bee":
https://en.wikipedia.org/wiki/Sitemaps
https://en.wikipedia.org/wiki/Robots_exclusion_standard
https://en.wikipedia.org/wiki/Web_crawler

Execution Completed

Total time required for execution : 1.5146667 minutes

=====
Please enter the word you want to search (recommended : use lowercase):
=====
crawling

Searching for the word "crawling" in "https://www.scrapingbee.com/blog/crawling-python"
Occurrence of the word "crawling" : 18 times

Searching for the word "crawling" in "https://en.wikipedia.org/wiki/Sitemaps"
Occurrence of the word "crawling" : 3 times

Searching for the word "crawling" in "https://www.scrapingbee.com/features/data-extraction/"
Occurrence of the word "crawling" : 2 times

Searching for the word "crawling" in "https://www.scrapingbee.com/#faq"
Searching for the word "crawling" in "https://www.scrapingbee.com/"
Searching for the word "crawling" in "https://www.scrapingbee.com/features/screenshot/"
Searching for the word "crawling" in "https://app.scrapingbee.com/account/register"
Searching for the word "crawling" in "https://en.wikipedia.org/wiki/Web_scraping"
Occurrence of the word "crawling" : 4 times

Searching for the word "crawling" in "https://www.scrapingbee.com/#pricing"
Searching for the word "crawling" in "https://www.scrapingbee.com/features/google/"
Searching for the word "crawling" in "https://www.scrapingbee.com/features/make/"
Searching for the word "crawling" in "https://www.scrapingbee.com/blog/crawling-python/#"
Occurrence of the word "crawling" : 18 times

Searching for the word "crawling" in "https://www.scrapingbee.com/tutorials"
Searching for the word "crawling" in "https://www.scrapingbee.com/features/javascript-scenario/"
Searching for the word "crawling" in "https://www.scrapingbee.com/blog/"
Searching for the word "crawling" in "https://en.wikipedia.org/wiki/Robots_exclusion_standard"
Occurrence of the word "crawling" : 3 times

Searching for the word "crawling" in "https://en.wikipedia.org/wiki/Web_crawler"
Occurrence of the word "crawling" : 52 times

```

```

Problems | Debug Shell | Console x
<terminated> Search Web Crawler (1) [Java Application] C:\Users\yogit\Downloads\ eclipse-dsl-2022-03-R-win32-x86_64\eclipse\plugins\org.eclipse.justi.openjdk.hotspot.jre.full.win32.x86_64.17.0.2.v20220201-1208\jre\bin\javaw.exe
https://www.scrapingbee.com/features/make/
https://www.scrapingbee.com/blog/crawling-python/#
https://www.scrapingbee.com/tutorials
https://www.scrapingbee.com/features/javascript-scenario/
https://www.scrapingbee.com/blog/
https://en.wikipedia.org/wiki/Robots_exclusion_standard
https://en.wikipedia.org/wiki/Web_crawler
https://www.scrapingbee.com/documentation

Execution Completed

Total time required for execution : 1.7164333 minutes

=====
Please enter the word you want to search (recommended : use lowercase):
=====
crawling BEE

Searching for the word "crawling bee" in "https://www.scrapingbee.com/blog/crawling-python"
Searching for the word "crawling bee" in "https://en.wikipedia.org/wiki/Sitemaps"
Searching for the word "crawling bee" in "https://www.scrapingbee.com/features/data-extraction/"
Searching for the word "crawling bee" in "https://www.scrapingbee.com/#faq"
Searching for the word "crawling bee" in "https://www.scrapingbee.com/"
Searching for the word "crawling bee" in "https://www.scrapingbee.com/features/screenshot/"
Searching for the word "crawling bee" in "https://app.scrapingbee.com/account/register"
Searching for the word "crawling bee" in "https://en.wikipedia.org/wiki/Web_scraping"
Searching for the word "crawling bee" in "https://www.scrapingbee.com/#pricing"
Searching for the word "crawling bee" in "https://www.scrapingbee.com/features/google/"
Searching for the word "crawling bee" in "https://www.scrapingbee.com/features/make/"
Searching for the word "crawling bee" in "https://www.scrapingbee.com/blog/crawling-python/#"
Searching for the word "crawling bee" in "https://www.scrapingbee.com/tutorials"
Searching for the word "crawling bee" in "https://www.scrapingbee.com/features/javascript-scenario/"
Searching for the word "crawling bee" in "https://www.scrapingbee.com/blog/"
Searching for the word "crawling bee" in "https://en.wikipedia.org/wiki/Robots_exclusion_standard"
Searching for the word "crawling bee" in "https://en.wikipedia.org/wiki/Web_crawler"
Searching for the word "crawling bee" in "https://www.scrapingbee.com/documentation"

Number of URLs that contain the word "crawling bee" : 18

List of the URLs that contains the word "crawling bee":
https://www.scrapingbee.com/blog/crawling-python
https://en.wikipedia.org/wiki/Sitemaps
https://www.scrapingbee.com/features/data-extraction/
https://www.scrapingbee.com/#faq
https://www.scrapingbee.com/

```


Boundary conditions:

```
Problems | Debug Shell | Console x
Search Web_Crawler (1) [Java Application] C:\Users\yogit\Downloads\eclipse-dsl-2022-03-R-win32-x86_64\eclipse\p

https://www.scrapingbee.com/blog/crawling-python
https://www.scrapingbee.com/
https://app.scrapingbee.com/account/login
https://app.scrapingbee.com/account/register
https://www.scrapingbee.com/#pricing
https://www.scrapingbee.com/#faq
https://www.scrapingbee.com/blog/
https://www.scrapingbee.com/blog/crawling-python/#
https://www.scrapingbee.com/features/screenshot/
https://www.scrapingbee.com/features/google/
https://www.scrapingbee.com/features/data-extraction/
https://www.scrapingbee.com/features/javascript-scenario/
https://www.scrapingbee.com/features/make/
https://www.scrapingbee.com/tutorials
https://www.scrapingbee.com/documentation
https://help.scrapingbee.com/en/
https://en.wikipedia.org/wiki/Web_crawler
https://en.wikipedia.org/wiki/Web_scraping
https://en.wikipedia.org/wiki/Robots_exclusion_standard
https://en.wikipedia.org/wiki/Sitemaps

=====
Please enter the word you want to search (recommended : use lowercase):
=====
.

Number of URLs that contain the word "." : 0

List of the URLs that contains the word ".":
No URL contains the word "."

Execution Completed

Total time required for execution : 0.29313332 minutes

=====
Please enter the word you want to search (recommended : use lowercase):
=====
```

```
Execution Completed

Total time required for execution : 1.5197166 minutes

=====
Please enter the word you want to search (recommended : use lowercase):
=====
874

Number of URLs that contain the word "874" : 0

List of the URLs that contains the word "874":
No URL contains the word "874"

Execution Completed

Total time required for execution : 1.9011999 minutes

=====
Please enter the word you want to search (recommended : use lowercase):
=====
6878

Number of URLs that contain the word "6878" : 0

List of the URLs that contains the word "6878":
No URL contains the word "6878"

Execution Completed

Total time required for execution : 2.0219166 minutes

=====
Please enter the word you want to search (recommended : use lowercase):
=====
```

```
Problems | Debug Shell | Console x
<terminated> Search Web Crawler (1) [Java Application] C:\Users\yogit\Downloads\eclipse-dsl-2022-03-R-win32-x86_64\eclipse\plugins\org.eclipse.justi.openjdk.hotspot.jre.full.win32.x86_64.17.0.2.v20220201-1208\jre\bin\javaw.exe
Searching for the word "crawling bee" in "https://www.scrapingbee.com/features/javascript-scenario/"
Searching for the word "crawling bee" in "https://www.scrapingbee.com/blog/"
Searching for the word "crawling bee" in "https://en.wikipedia.org/wiki/Robots_exclusion_standard"
Searching for the word "crawling bee" in "https://en.wikipedia.org/wiki/Web_crawler"
Searching for the word "crawling bee" in "https://www.scrapingbee.com/documentation"

Number of URLs that contain the word "crawling bee" : 18

List of the URLs that contains the word "crawling bee":
https://www.scrapingbee.com/blog/crawling-python
https://en.wikipedia.org/wiki/Sitemaps
https://www.scrapingbee.com/features/data-extraction/
https://www.scrapingbee.com/#faq
https://www.scrapingbee.com/
https://www.scrapingbee.com/features/screenshot/
https://app.scrapingbee.com/account/register
https://en.wikipedia.org/wiki/Web_scraping
https://www.scrapingbee.com/#pricing
https://www.scrapingbee.com/features/google/
https://www.scrapingbee.com/features/make/
https://www.scrapingbee.com/blog/crawling-python/#
https://www.scrapingbee.com/tutorials
https://www.scrapingbee.com/features/javascript-scenario/
https://www.scrapingbee.com/blog/
https://en.wikipedia.org/wiki/Robots_exclusion_standard
https://en.wikipedia.org/wiki/Web_crawler
https://www.scrapingbee.com/documentation

Execution Completed

Total time required for execution : 1.87975 minutes

=====
Please enter the word you want to search (recommended : use lowercase):
SHE
=====
"she" is one of the stop word.Try searching so other word.

=====
Please enter the word you want to search (recommended : use lowercase):
EXIT
=====
```

Testing Boundary Conditions:


```
Problems Javadoc Declaration Console x
<terminated> Search Web Crawler [Java Application] C:\Users\yogit\Downloads\eclipse-dsl-2022-03-R-win32-x86_64\eclipse\plugins\org.eclipse.justj.openjdk

//===== MINI SEARCH ENGINE =====//

Sample URLs:
https://www.datacamp.com/tutorial/making-web-crawlers-scrapy-python
https://www.techtarget.com/whatis/definition/crawler
https://www.scrapingbee.com/blog/crawling-python
https://medium.com/dataseries/build-a-crawler-to-extract-web-data-in-10-mins-691b2cc4f1c3

Please enter a valid URL on which you want to perform a search (use sample URLs given above):
https://www.datacamp.com/tutorial/making-web-crawlers-scrapy-python

Web Crawling . . . .

https://www.datacamp.com/tutorial/making-web-crawlers-scrapy-python
https://www.datacamp.com/tutorial/making-web-crawlers-scrapy-python#main
https://www.datacamp.com/
https://www.datacamp.com/careers
https://www.datacamp.com/blog
https://www.datacamp.com/podcast
https://www.datacamp.com/tutorial
https://www.datacamp.com/cheat-sheet
https://www.datacamp.com/tutorial/category/for-business
https://www.datacamp.com/tutorial/category/git
https://www.datacamp.com/tutorial/category/julia
https://www.datacamp.com/tutorial/category/power-bi
https://www.datacamp.com/tutorial/category/python
https://www.datacamp.com/tutorial/category/r-programming
https://www.datacamp.com/tutorial/category/scala
https://www.datacamp.com/tutorial/category/spreadsheets
https://www.datacamp.com/tutorial/category/sql
https://www.datacamp.com/tutorial/category/tableau
https://www.datacamp.com/tutorial/category/ai
https://www.datacamp.com/tutorial/category/big-data

=====
Please enter the word you want to search (recommended : use lowercase):
=====
datacamp

Searching for the word "datacamp" in "https://www.datacamp.com/tutorial/category/r-programming"
```

```
Problems Javadoc Declaration Console ×
<terminated> Search_Web_Crawler [Java Application] C:\Users\yogit\Downloads\eclipse-dsl-2022-03-R-win32-x86_64\eclipse\plugins\org.eclipse.justj.openjdk.b
=====
datacamp

Searching for the word "datacamp" in "https://www.datacamp.com/tutorial/category/r-programming"
Searching for the word "datacamp" in "https://www.datacamp.com/tutorial/making-web-crawlers-scrapy-python"
Searching for the word "datacamp" in "https://www.datacamp.com/tutorial"
Searching for the word "datacamp" in "https://www.datacamp.com/tutorial/making-web-crawlers-scrapy-python#main"
Searching for the word "datacamp" in "https://www.datacamp.com/tutorial/category/big-data"
Searching for the word "datacamp" in "https://www.datacamp.com/tutorial/category/spreadsheets"
Searching for the word "datacamp" in "https://www.datacamp.com/tutorial/category/tableau"
Searching for the word "datacamp" in "https://www.datacamp.com/tutorial/category/sql"
Searching for the word "datacamp" in "https://www.datacamp.com/podcast"
Searching for the word "datacamp" in "https://www.datacamp.com/tutorial/category/ai"
Searching for the word "datacamp" in "https://www.datacamp.com/tutorial/category/scala"
Searching for the word "datacamp" in "https://www.datacamp.com/tutorial/category/git"
Searching for the word "datacamp" in "https://www.datacamp.com/careers"
Searching for the word "datacamp" in "https://www.datacamp.com/blog"
Searching for the word "datacamp" in "https://www.datacamp.com/tutorial/category/for-business"
Searching for the word "datacamp" in "https://www.datacamp.com/tutorial/category/python"
Searching for the word "datacamp" in "https://www.datacamp.com/cheat-sheet"
Searching for the word "datacamp" in "https://www.datacamp.com/"
Searching for the word "datacamp" in "https://www.datacamp.com/tutorial/category/power-bi"
Searching for the word "datacamp" in "https://www.datacamp.com/tutorial/category/julia"

Number of URLs that contain the word "datacamp" : 20

List of the URLs that contains the word "datacamp":
https://www.datacamp.com/tutorial/category/r-programming
https://www.datacamp.com/tutorial/making-web-crawlers-scrapy-python
https://www.datacamp.com/tutorial
https://www.datacamp.com/tutorial/making-web-crawlers-scrapy-python#main
https://www.datacamp.com/tutorial/category/big-data
https://www.datacamp.com/tutorial/category/spreadsheets
https://www.datacamp.com/tutorial/category/tableau
https://www.datacamp.com/tutorial/category/sql
https://www.datacamp.com/podcast
https://www.datacamp.com/tutorial/category/ai
https://www.datacamp.com/tutorial/category/scala
https://www.datacamp.com/tutorial/category/git
https://www.datacamp.com/careers
https://www.datacamp.com/blog
https://www.datacamp.com/tutorial/category/for-business
https://www.datacamp.com/tutorial/category/python
https://www.datacamp.com/cheat-sheet
https://www.datacamp.com/
```

```
Problems Javadoc Declaration Console x
<terminated> Search_Web_Crawler [Java Application] C:\Users\yogit\Downloads\eclipse-dsl-2022-03-R-win32-x86_64\eclipse\plugins\org.eclipse.justj.openjdk
Execution Completed

Total time required for execution : 1.0182834 minutes

=====
Please enter the word you want to search (recommended : use lowercase):
=====
python

Searching for the word "python" in "https://www.datacamp.com/tutorial/making-web-crawlers-scrapy-python"
Occurence of the word "python" : 5 times

Searching for the word "python" in "https://www.datacamp.com/tutorial/making-web-crawlers-scrapy-python"
Occurence of the word "python" : 5 times

Number of URLs that contain the word "python" : 2

List of the URLs that contains the word "python":
https://www.datacamp.com/tutorial/making-web-crawlers-scrapy-python
https://www.datacamp.com/tutorial/making-web-crawlers-scrapy-python#main

Execution Completed

Total time required for execution : 1.3607333 minutes

=====
Please enter the word you want to search (recommended : use lowercase):
=====
an

"an" is one of the stop word.Try searching so other word.

=====
Please enter the word you want to search (recommended : use lowercase):
=====
you

"you" is one of the stop word.Try searching so other word.

=====
Please enter the word you want to search (recommended : use lowercase):
=====
from
```

```
Problems Javadoc Declaration Console x
<terminated> Search_Web_Crawler [Java Application] C:\Users\yogit\Downloads\eclipse-dsl-2022-03-R-win32-x86_64\eclipse\plugins\org.eclipse.justj.openjdk

=====
Please enter the word you want to search (recommended : use lowercase):
=====
78

Number of URLs that contain the word "78" : 0

List of the URLs that contains the word "78":
No URL contains the word "78"

Execution Completed

Total time required for execution : 2.6808002 minutes

=====
Please enter the word you want to search (recommended : use lowercase):
=====
DataCamp

Searching for the word "datacamp" in "https://www.datacamp.com/tutorial/category/r-programming"
Searching for the word "datacamp" in "https://www.datacamp.com/tutorial/making-web-crawlers-scrapy-python"
Searching for the word "datacamp" in "https://www.datacamp.com/tutorial"
Searching for the word "datacamp" in "https://www.datacamp.com/tutorial/making-web-crawlers-scrapy-python"
Searching for the word "datacamp" in "https://www.datacamp.com/tutorial/category/big-data"
Searching for the word "datacamp" in "https://www.datacamp.com/tutorial/category/spreadsheets"
Searching for the word "datacamp" in "https://www.datacamp.com/tutorial/category/tableau"
Searching for the word "datacamp" in "https://www.datacamp.com/tutorial/category/sql"
Searching for the word "datacamp" in "https://www.datacamp.com/podcast"
Searching for the word "datacamp" in "https://www.datacamp.com/tutorial/category/ai"
Searching for the word "datacamp" in "https://www.datacamp.com/tutorial/category/scala"
Searching for the word "datacamp" in "https://www.datacamp.com/tutorial/category/git"
Searching for the word "datacamp" in "https://www.datacamp.com/careers"
Searching for the word "datacamp" in "https://www.datacamp.com/blog"
Searching for the word "datacamp" in "https://www.datacamp.com/tutorial/category/for-business"
Searching for the word "datacamp" in "https://www.datacamp.com/tutorial/category/python"
Searching for the word "datacamp" in "https://www.datacamp.com/cheat-sheet"
Searching for the word "datacamp" in "https://www.datacamp.com/"
Searching for the word "datacamp" in "https://www.datacamp.com/tutorial/category/power-bi"
Searching for the word "datacamp" in "https://www.datacamp.com/tutorial/category/julia"

Number of URLs that contain the word "datacamp" : 20
```

```
Problems Javadoc Declaration Console X
<terminated> Search_Web_Crawler [Java Application] C:\Users\yogit\Downloads\eclipse-dsl-2022-03-R-win32-x86_64\eclipse\plugins\org.eclipse.justj.openjdk
https://www.datacamp.com/tutorial/category/ai
https://www.datacamp.com/tutorial/category/scala
https://www.datacamp.com/tutorial/category/git
https://www.datacamp.com/careers
https://www.datacamp.com/blog
https://www.datacamp.com/tutorial/category/for-business
https://www.datacamp.com/tutorial/category/python
https://www.datacamp.com/cheat-sheet
https://www.datacamp.com/
https://www.datacamp.com/tutorial/category/power-bi
https://www.datacamp.com/tutorial/category/julia

Execution Completed

Total time required for execution : 2.8165832 minutes

=====
Please enter the word you want to search (recommended : use lowercase):
=====
PYTHON

Searching for the word "python" in "https://www.datacamp.com/tutorial/making-web-crawlers-scrapy-python"
Occurence of the word "python" : 5 times

Searching for the word "python" in "https://www.datacamp.com/tutorial/making-web-crawlers-scrapy-python"
Occurence of the word "python" : 5 times

Number of URLs that contain the word "python" : 2

List of the URLs that contains the word "python":
https://www.datacamp.com/tutorial/making-web-crawlers-scrapy-python
https://www.datacamp.com/tutorial/making-web-crawlers-scrapy-python#main

Execution Completed

Total time required for execution : 2.9636333 minutes

=====
Please enter the word you want to search (recommended : use lowercase):
=====
>

">" is one of the stop word.Try searching so other word.
```

```
Problems Javadoc Declaration Console x
<terminated> Search_Web_Crawler [Java Application] C:\Users\yogit\Downloads\eclipse-dsl-2022-03-R-win32-x86_64\eclipse\plugins\org.eclipse.justj.openjdk.

">" is one of the stop word.Try searching so other word.

=====
Please enter the word you want to search (recommended : use lowercase):
=====
datacamp python

Searching for the word "datacamp python" in "https://www.datacamp.com/tutorial/category/r-programming"
Searching for the word "datacamp python" in "https://www.datacamp.com/tutorial/making-web-crawlers-scrappy-python"
Searching for the word "datacamp python" in "https://www.datacamp.com/tutorial"
Searching for the word "datacamp python" in "https://www.datacamp.com/tutorial/making-web-crawlers-scrappy-python#main"
Searching for the word "datacamp python" in "https://www.datacamp.com/tutorial/category/big-data"
Searching for the word "datacamp python" in "https://www.datacamp.com/tutorial/category/spreadsheets"
Searching for the word "datacamp python" in "https://www.datacamp.com/tutorial/category/tableau"
Searching for the word "datacamp python" in "https://www.datacamp.com/tutorial/category/sql"
Searching for the word "datacamp python" in "https://www.datacamp.com/podcast"
Searching for the word "datacamp python" in "https://www.datacamp.com/tutorial/category/ai"
Searching for the word "datacamp python" in "https://www.datacamp.com/tutorial/category/scala"
Searching for the word "datacamp python" in "https://www.datacamp.com/tutorial/category/git"
Searching for the word "datacamp python" in "https://www.datacamp.com/careers"
Searching for the word "datacamp python" in "https://www.datacamp.com/blog"
Searching for the word "datacamp python" in "https://www.datacamp.com/tutorial/category/for-business"
Searching for the word "datacamp python" in "https://www.datacamp.com/tutorial/category/python"
Searching for the word "datacamp python" in "https://www.datacamp.com/cheat-sheet"
Searching for the word "datacamp python" in "https://www.datacamp.com/"
Searching for the word "datacamp python" in "https://www.datacamp.com/tutorial/category/power-bi"
Searching for the word "datacamp python" in "https://www.datacamp.com/tutorial/category/julia"

Number of URLs that contain the word "datacamp python" : 20

List of the URLs that contains the word "datacamp python":
https://www.datacamp.com/tutorial/category/r-programming
https://www.datacamp.com/tutorial/making-web-crawlers-scrappy-python
https://www.datacamp.com/tutorial
https://www.datacamp.com/tutorial/making-web-crawlers-scrappy-python#main
https://www.datacamp.com/tutorial/category/big-data
https://www.datacamp.com/tutorial/category/spreadsheets
https://www.datacamp.com/tutorial/category/tableau
https://www.datacamp.com/tutorial/category/sql
https://www.datacamp.com/podcast
https://www.datacamp.com/tutorial/category/ai
https://www.datacamp.com/tutorial/category/scala
https://www.datacamp.com/tutorial/category/git
```

```
Problems Javadoc Declaration Console ×
<terminated> Search_Web_Crawler [Java Application] C:\Users\yogit\Downloads\eclipse-dsl-2022-03-R-win32-x86_64\eclipse\plugins\org.eclipse.justj.openjdk
Execution Completed

Total time required for execution : 3.5482168 minutes

=====
Please enter the word you want to search (recommended : use lowercase):
=====
robot.txt

Searching for the word "robot.txt" in "https://www.datacamp.com/tutorial/category/r-programming"
Searching for the word "robot.txt" in "https://www.datacamp.com/tutorial/category/sql"
Searching for the word "robot.txt" in "https://www.datacamp.com/tutorial"
Searching for the word "robot.txt" in "https://www.datacamp.com/tutorial/category/ai"
Searching for the word "robot.txt" in "https://www.datacamp.com/tutorial/category/big-data"
Searching for the word "robot.txt" in "https://www.datacamp.com/careers"
Searching for the word "robot.txt" in "https://www.datacamp.com/blog"
Searching for the word "robot.txt" in "https://www.datacamp.com/tutorial/category/for-business"
Searching for the word "robot.txt" in "https://www.datacamp.com/cheat-sheet"
Searching for the word "robot.txt" in "https://www.datacamp.com/"

Number of URLs that contain the word "robot.txt" : 10

List of the URLs that contains the word "robot.txt":
https://www.datacamp.com/tutorial/category/r-programming
https://www.datacamp.com/tutorial/category/sql
https://www.datacamp.com/tutorial
https://www.datacamp.com/tutorial/category/ai
https://www.datacamp.com/tutorial/category/big-data
https://www.datacamp.com/careers
https://www.datacamp.com/blog
https://www.datacamp.com/tutorial/category/for-business
https://www.datacamp.com/cheat-sheet
https://www.datacamp.com/

Execution Completed

Total time required for execution : 3.6290002 minutes

=====
Please enter the word you want to search (recommended : use lowercase):
=====
exit
```