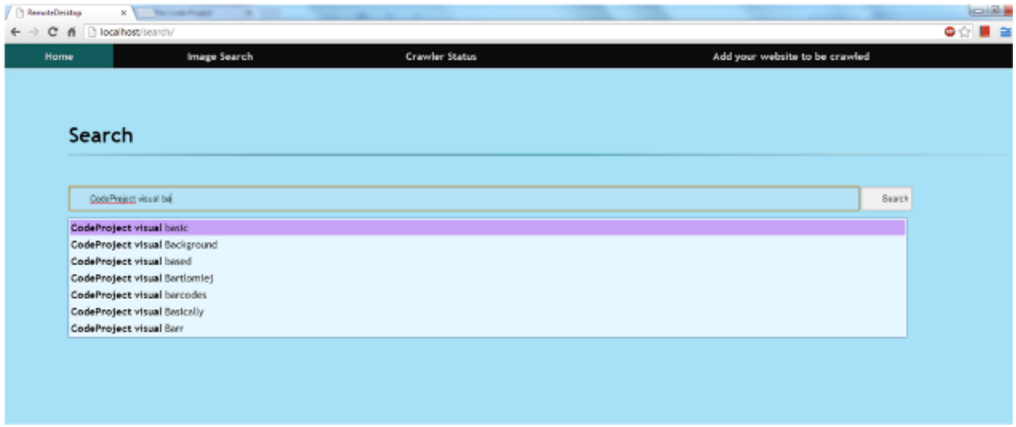
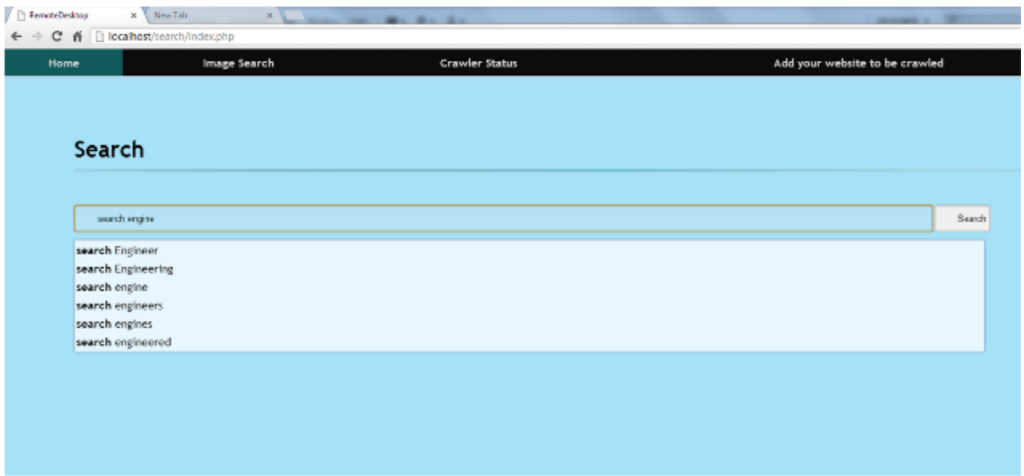
Making a Search Engine

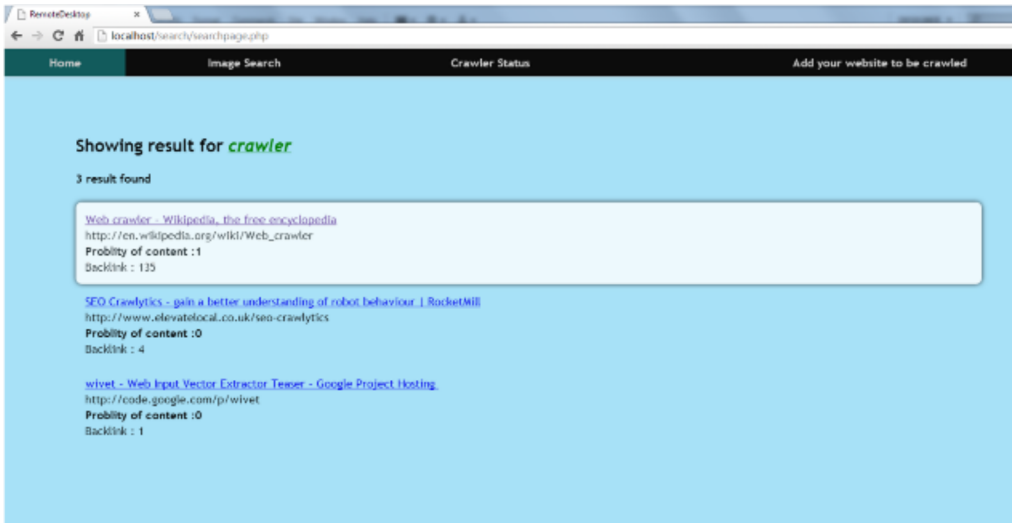
kburman6, 4 May 2013

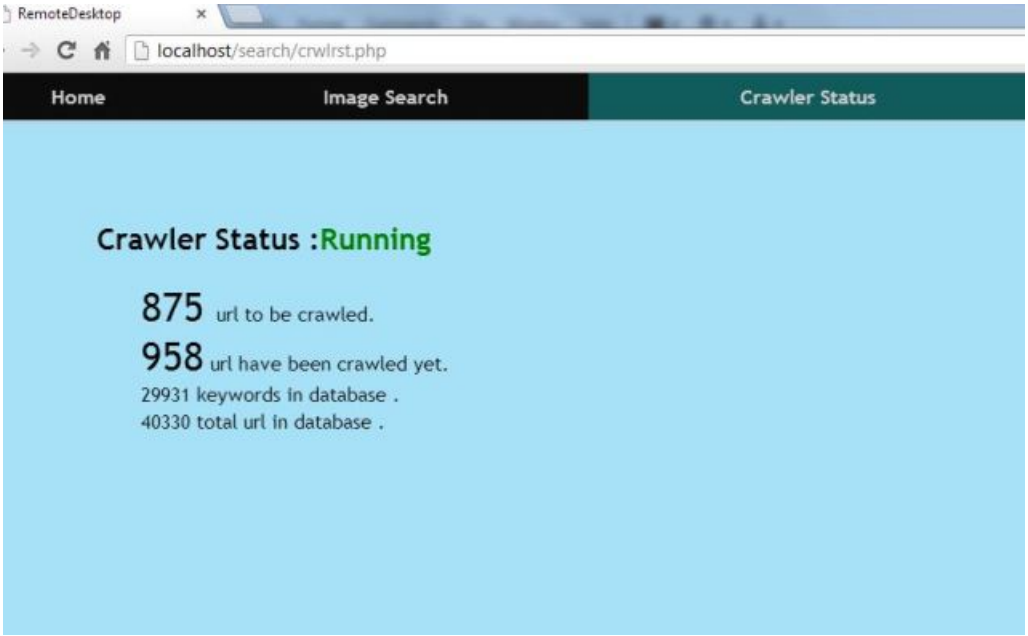
This article discusses the making of a search engine.

Download Executable files - 178.3 KB
Download source -21.5 KB
Download PHP file - 6.83 KB









Introduction

This project is still not complete. You must use it to crawl thousands of URLs because you may find that it crawls the same URL for the last 100 times. (This is because of some unidentified problem in conversion of relative to absolute URL.)

Like most search engines, this one also has a crawler whose basic aim is to retrieve the source code of a given URL and then break the content into words with which we can create an array of tag words which will represent the content of the site. It is not a fool proof method, but can work for sites with lots of words in it, like a blog or article or a discussion forum, etc.

For example:

TAG CLOUD of http://www.google.co.in

TAG CLOUD of http://www.wikipedia.org

TAG CLOUD of http://www.w3schools.com

It is clear how tag cloud can highlight the key words which could describe a given URL.

Now these keywords are stored in a database and then used to find the relevant URL for given keywords.

Background

It all started when my friend showed me his search engine with 4 URLs in an XML file. It seemed like an auto complete rather than a search engine, but later that night it was 2:00 am, and I couldn't sleep at all because of that auto complete feature with which I was too impressed. I wanted my own... there was a thunder storm of ideas in my mind. After 3 sleepless nights, on the 3rd day at 6:00 a.m., I was ready with my search engine working with 100 URLs in database... that was the time when I finally slept comfortably and full of satisfaction. It took me 3 days because every day I started from the beginning because I was not satisfied with the performance of the crawler or there was some problem.

Using the Code

I have basically divided every task into small parts so that the work could become easy. So you would find lots of classes in the project.

Some important classes are given below:

- `Panda`-> It retrieves the source code of a given URL using a `get_sourcecode(url)` from Module '`func`' and then passes it to `juicer`(class) for extracting the useful information and after the work is completed, reports to the `panda_manger` (boss of all the `panda`s) and again assigns a new job to `panda` if any.
- `panda_manger`-> It manages all the `panda`s and from this class we assign any web URL for crawling and it will automatically assign this work to any free `panda` and if no `panda` is free, then it creates a new `panda` and assigns the work to it. When a `panda` finishes, it works and reports back to `panda_manger` then `panda_manger` checks whether there is any URL left to be crawled, if any then it gives the command to crawl that URL to the `panda`.
- `juicer`->This could be said to be the main class for the whole crawler which extracts keywords from the source code of any website and then saves it into the database.
- `juicer`->`extract_juice`-> This class gets the source code and then converts it into `HtmlAgilityPack.HtmlDocument` using the method `LoadHtml(source)`. It is used because we need an HTML parser to scrap the text out of HTML (building a custom one will require a lot of time) and since it supports xpath for retrieving any element(s) from source, it simplifies our work a lot.
  Now we pass the `HtmlDocument.DocumentNode(.SelectNodes)` to the various methods of this class for extraction of some type of information.

```
Dim doc As New HtmlAgilityPack.HtmlDocument()
doc.LoadHtml(source)
process_texttag(doc.DocumentNode.SelectNodes("//meta"))
prcess_anchor(doc.DocumentNode.SelectNodes("//a"))
process_image(doc.DocumentNode.SelectNodes("//img"))
```

- `juicer`->`process_metatag`-> Currently, it just processes the meta tag containing keywords and then assigns every word a 45% to total word count in the document and 30% of total word count to the words which are mentioned in the keywords but can't be found in the document.
- `juicer`->`process_anchor`->This is the most important part if you want the search engine to automatically move to the next link without manually entering every link to the crawler and the hardest part was to convert the relative URL to absolute URL. But Microsoft saved me with the URI class which can easily be used to convert the URI.

Database Structure

At the starting point, we have a 4 tables stored in a database named as "`Crawler`" by default.

Table Name : "Keyword_index"

This table is used to provide the suggested result in the search box. It contains all the words which the crawler has encountered till now and no word is repeated.

```
mysql> select * from keyword_index where word like 'a%' ;
+--------------+-------+
| word         | freq  |
+--------------+-------+
| A            |    20 |
| About        |     1 |
| ac           |     5 |
| According    |     2 |
| accumsan     |     3 |
| Achievements |     1 |
| across       |     1 |
| address      |     4 |
| adipis       |     2 |
| adipiscing   |     4 |
| adiscing     |     1 |
| ado          |     2 |
| Airtel       |     3 |
| alais        |     1 |
| aliquam      |     2 |
| aliquet      |     1 |
| All          |     6 |
| amet         |     9 |
| an           |     3 |
| and          |    10 |
| annihilation |     2 |
| answer       |     2 |
| ante         |     1 |
| antiparticle |     3 |
| api          |     2 |
| arcu         |     3 |
| are          |     2 |
| as           |     2 |
| asp          |    13 |
| at           |     7 |
| atl          |    13 |
| Atom         |     1 |
| augue        |     2 |
| Available    |     2 |
| Axiata       |     3 |
| AXIS         |     3 |
+--------------+-------+
36 rows in set (0.00 sec)
```

For multiple keywords, we first break the keyword from " ", then find the urlhash for the word and then find what the other words are that urlhashes contain and show the words. {Not yet implemented because I was getting an unidentified error and if I used SQL `join`, then it took more than 1 minute to search the database for 1 keyword and time increased exponentially.}

Table Name :"Keyword_list"

This table contains all the words in a tag cloud of the given URL. And to identify this word belongs to which URL, we store the urlhash of that website with the word.

```
mysql> desc keyword_list;
+-------------+--------------+------+-----+---------+-------+
| Field       | Type         | Null | Key | Default | Extra |
+-------------+--------------+------+-----+---------+-------+
| word        | varchar(100) | NO   |     | NULL    |       |
| urlhash     | char(32)     | NO   |     | NULL    |       |
| freq        | int(3)       | NO   |     | NULL    |       |
| per_in_site | int(3)       | NO   |     | NULL    |       |
+-------------+--------------+------+-----+---------+-------+
4 rows in set (0.01 sec)
```

```
mysql> select * from keyword_list where word like 'a%' limit 10;
+-----------+----------------------------------+------+--------------+
| word      | urlhash                          | freq | per_in_site  |
+-----------+----------------------------------+------+--------------+
| A         | ead7b7cab3aa7349913987b7b176bb15 |    1 |            0 |
| All       | ead7b7cab3aa7349913987b7b176bb15 |    1 |            0 |
| at        | ead7b7cab3aa7349913987b7b176bb15 |    1 |            0 |
| AT        | ead7b7cab3aa7349913987b7b176bb15 |    1 |            0 |
| address   | 34a84aa0676b69fc8826328bdaa9967d |    1 |            0 |
| Address   | 34a84aa0676b69fc8826328bdaa9967d |    1 |            0 |
| All       | 34a84aa0676b69fc8826328bdaa9967d |    1 |            0 |
| abelardo  | 4d04edf6d9edc6d7f1b36b2b65028523 |    1 |            0 |
| About     | 4d04edf6d9edc6d7f1b36b2b65028523 |    1 |            0 |
| adara     | 4d04edf6d9edc6d7f1b36b2b65028523 |    1 |            0 |
+-----------+----------------------------------+------+--------------+
10 rows in set (0.00 sec)
```

Table Name :"url_webpage"

This table is used to store all the links which it finds in the pages that it crawled till now. In this table, we also use MD5 hash of URL to refer to that URL instead of the original URL because no one knows how long the URL which we may find can be. So we change it to MD5 because it does not mater how long the URL is, the MD5 will always be 32 chars long.

```
mysql> desc url_webpage;
+-------------+--------------+------+-----+---------+-------+
| Field       | Type         | Null | Key | Default | Extra |
+-------------+--------------+------+-----+---------+-------+
| urlhash     | char(32)     | NO   | PRI | NULL    |       |
| url         | varchar(250) | NO   | UNI | NULL    |       |
| state       | int(2)       | NO   |     | NULL    |       |
| crawl_date  | datetime     | YES  |     | NULL    |       |
| backlink    | int(3)       | NO   |     | NULL    |       |
| priority    | int(1)       | NO   |     | NULL    |       |
| title       | varchar(500) | NO   |     | NULL    |       |
+-------------+--------------+------+-----+---------+-------+
7 rows in set (0.01 sec)
```

urlhash	url	state	crawl_date	backlink	priority	title
02a5004b16154bb1b196aeab0b01f1d2	http://mitunacasserole.tumblr.com	2	(null)	8	0	Zagazo
c66a50425514bf186d7524b7cb020dc5	http://www.youtube.com/?hl=en&gl=IN&sa=N	2	(null)	0	0	YouTube
fc341ad7803b95f02cc12f3d032ec0cb	http://www.youtube.com/?gl=IN&tab=w1	2	(null)	0	0	YouTube
0026c5bef6ba23b999225a110431019	http://fucksocial.tumblr.com	2	(null)	0	0	your stupd
02132d1677ba445dbbe04a8c976f5d	http://somuchfortakingchances.tumblr.com/p	2	(null)	4	0	You have to love something before you
04c072d76623a043b6700f4eb01a29b	http://somuchfortakingchances.tumblr.com	2	(null)	7	0	You have to love something before you
055eead9bf5cc01d39693d0d5dc26092	http://somuchfortakingchances.tumblr.com/p	2	(null)	4	0	You have to love something before you
009e19597f358377468e2f8a8a40a44c	http://smile-youre-amazing.tumblr.com	2	(null)	0	0	you are loved
046d17e3396ad4325ed2b775d9fad48	http://foreverbillimfinityy.tumblr.com	2	(null)	1	0	YOU ARE BEAUTIFUL :)
00ebcc6f0a188c612c38fe8482d7607c	http://nepetaleijon.tumblr.com	2	(null)	1	0	Ya did clean
021c3cee2f60012690e0e83b8ea1ce8e	http://wowgreat.tumblr.com/post/4487167022	2	(null)	2	0	WOWGREAT - Joe Winter (via Printer
013163eb93f6fe697685c3295099cee7	http://steveholtvstheuniverse.tumblr.com	2	(null)	0	0	Worms - oh my god WORMS
01262fb3bce11e0b7cdfd3473850e967	http://selecionar.tumblr.com/post/453072911	2	(null)	3	0	Will you stay?
00ce4f2502fc5694bf32d21e89bc801d	http://wildthemes.tumblr.com	2	(null)	0	0	wildthemes
005c2b186c901f7c6eb922aa8e1b61db	http://live-in-these-moments.tumblr.com	2	(null)	1	0	Wherever you are, be all there.
006b0e64bb0588bf32139d7b1b9b7d0f	http://mishas-assbutts.tumblr.com/post/4531	2	(null)	2	0	Where you invest your love, you invest
019264a023d77c0506c493475780f03e	http://mishas-assbutts.tumblr.com	2	(null)	19	0	Where you invest your love, you invest
01d2f66daf3d04b65e9e8c000efc10d3	http://mishas-assbutts.tumblr.com/post/4531	2	(null)	3	0	Where you invest your love, you invest
00613e356e63813ee50e41d5f8f94b44	http://mylightinthedarknesskania.tumblr.com	2	(null)	2	0	Where there is love there is life
0446736e377e39c602d37672ca6a540b	http://th1s-f1re.tumblr.com/post/4504322235	2	(null)	0	0	Where Do I Put This Fire? (laturbo:
039709d65c62bf1aedba0da916bc70dc	http://th1s-f1re.tumblr.com	2	(null)	1	0	Where Do I Put This Fire?
0001c9e204a8fd6a0b292ff98d64d76b	http://feelsadthensmile.tumblr.com	2	(null)	3	0	When you feel sad... just smile :D
001d13f35c2df9d2975b416156c672ee	http://coolhipa.tumblr.com/post/45306599884	2	(null)	4	0	whatever
0371ebf56bf0a13ec7f72907e198b8f	http://coolhipa.tumblr.com	2	(null)	3	0	whatever
0188111a2c636d5d23ad6c187d2d177	http://whatalov3.tumblr.com	2	(null)	1	0	What
024884e0dce9f3c20a0018a17633155a	http://cheezburger.com/7132712448	2	(null)	16	0	What Was I Thinking? - Cheezburger
00b95a2eada08d4b385e8deff594397b	http://soy-un-dango.tumblr.com	2	(null)	8	0	What is love?
010e43b9a1876147d4bf2274b5b92be	http://soy-un-dango.tumblr.com/post/453083	2	(null)	4	0	What is love?

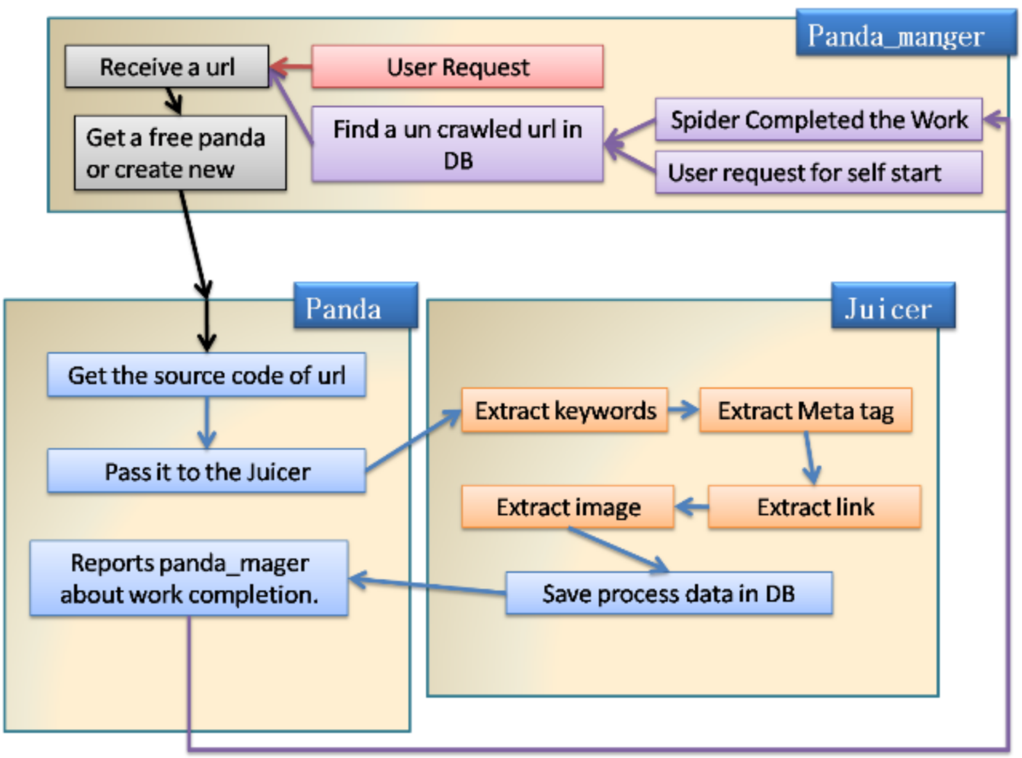
Table Name : "url_image"

This table is used to store all the links of the image which are found in the webpage it crawls. It will be used for image search, but I have not yet completed its processing and front end(PHP code).

So I will not discuss it right now. But I would like to tell that it will take a lot of processing power.

urlhash	url	state	crawl_date	backlink	priority	title
1cf5170c11c7ca47c1cc2baeea8f72	http://pixel.quantserve.com/pixel/p-19UtqE8ngoZbM.gif	0	(null)	600	0	
0ca61001f981d9970d28a4324789f5	http://assets.tumblr.com/images/default_avatar_16.gif	0	(null)	227	0	
0cb768f4d024451520e1dd67e2240	http://static.tumblr.com/lbe83dv/OUUltd858/spacer.gif	0	(null)	129	0	
3f64e606a849c4a60195691a4908d	http://assets.tumblr.com/images/x.gif	0	(null)	118	0	
07ed35b18a1810e614443879c0214	http://static.tumblr.com/uighh9x/OACIzwwao/reblog.png	0	(null)	52	0	
6a0a5ae1375ec0b32fec275df36383	http://s.ytimg.com/yts/img/pixel-vfl3z5WfW.gif	0	(null)	86	0	
4e1711aeb5d1bbbff8b4e39cd1c691	http://cur.cursors-4u.net/cursor.png	0	(null)	43	0	
776885f0c146cc4d9c36566a3869c	http://b.scorecardresearch.com/p?c1=2&c2=15742520&cv=2.0&cj=1	0	(null)	41	0	
602a8885e3f4f7bbcf5f632092ffc935	https://twimg0-a.akamaihd.net/profile_images/3007514026/8619c5cc363e6eebf26914d81f47	0	(null)	39	0	
0d303ed776779e3b5e00793e6ddd0	http://static.tumblr.com/tak5d5w/OBnlsu8qa/cllzyreblogbutton.png	0	(null)	38	0	
315cc288ce91471d2d945da198f33	https://pixel.quantserve.com/pixel/p-19UtqE0ngoZbM.gif	0	(null)	35	0	
33c86236ac07bf549d80f7d77da26a	http://assets.tumblr.com/images/iframe_reblog_alpha.png?4	0	(null)	34	0	
c59059eb1494d55f8d067fd4267537	http://assets.tumblr.com/images/default_avatar_40.gif	0	(null)	33	0	
84f78373479d3c02af515ba376518	http://l.yimg.com/g/images/spaceout.gif	0	(null)	29	0	
11bff35a4702b4152561c118f1c788	http://28.media.tumblr.com/tumblr_lpn6kseZUG1qm57imo1_500.png	0	(null)	28	0	
8097a0eb6f46bac05a4fcf3105009d	http://static.tumblr.com/i8axhtw/zGombktxi/rebloggg.png	0	(null)	28	0	
9eeb75f2ef3c81a6e404f76602085c	http://static.tumblr.com/i8axhtw/50mmbktzc/likeee.png	0	(null)	28	0	
ad547741d968368c25d36690bb574	http://static.tumblr.com/fkm6en6/vYolbws6r/heart_fill_8x7.png	0	(null)	28	0	
056a7482cedc3f7bc44d73e6693b1	http://uploads.neatorama.com/vosa/theme/pet/media/loading.gif	0	(null)	26	0	
058ea77048e877cd9f5cad8c56e7a	http://i42.tinypic.com/29uv9rp.jpg	0	(null)	25	0	
96b7b420c3d435dd9604a531580b6	http://school.discoveryeducation.com/images/lessonplans/spacer.gif	0	(null)	21	0	
066e1c2bbe85035c930e78b068b42	https://s.chzbgr.com/s/release_20130313.5/img/chz-logo-reg.png	0	(null)	19	0	
24b1c17f9c2b190d8cc8179092f73d	https://si0.twimg.com/profile_images/3007514026/8619c5cc363e6eebf26914d81f47f4e4_nor	0	(null)	19	0	
47466b5874181a828464af6a347c1	http://static.tumblr.com/j1tjexd/p25m8ecft/reblog.png	0	(null)	19	0	
4d0b30d378017433131c67a9e56bb	http://static.tumblr.com/j1tjexd/a8m8eogy/like.png	0	(null)	19	0	
8b7f738eca7ed49fc825f286a34a8d	https://twimg0-a.akamaihd.net/profile_images/2515236595/13t4ghoqffkfnvusfb2b_normal.jpe	0	(null)	19	0	
a3d7f048b7dc2c9c4a7f8dd5beb3cd	https://si0.twimg.com/profile_images/2515236595/13t4ghoqffkfnvusfb2b_normal.jpeg	0	(null)	19	0	
[illegible]	[illegible]	0	(null)	17	0	

How the Crawler Works



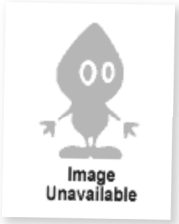
Points of Interest

It was very annoying to get rid of the relative URL. I tried several ways of resolving it, but every one ends up with some bug. Then I got a magic class, named as URI which solved all my problems, but while writing this article I should crawl codeproject.com and show the result for CodeProject search, but then a problem hit my crawler after identifying that I have deleted my database of URL, otherwise I would have posted a screenshot of that. It was something like */search.aspx* (some text)(same text repeated again)(and again, increasing with each crawl). It may be a problem with my code. I will try later to identify this problem and post the solution.

```
<hr /><h3>Currently Crawler is not following robots.txt so be careful while crawling.
</h3><h3>It is you responsibility if crawling a site which is not to be
crawled.   </h3><p>Sorry, but currently i am working on a new model of crawler. </p>
```

License

About the Author



kburman6
Student
India

I just love coding. But due to my studies it became very tough for me to manage both.

Comments and Discussions

57 messages have been posted for this article Visit **https://www.codeproject.com/Articles/563869/Making-a-Search-Engine** to post and view comments on this article, or click **here** to get a print view with messages.