Home
Tech
Entertainment
Business
Reviews
More
Science
Security
Geek
How To
DIY
Advertise

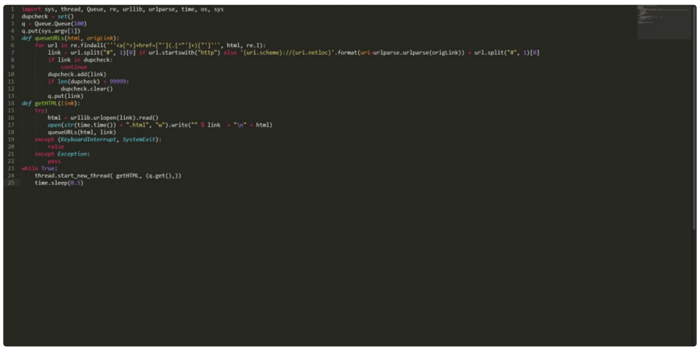
fossBytes *fresh bytes of technology and more*

DIY / HOW TO / TECH 12

FOLLOW:

How to Build a Basic Web Crawler in Python

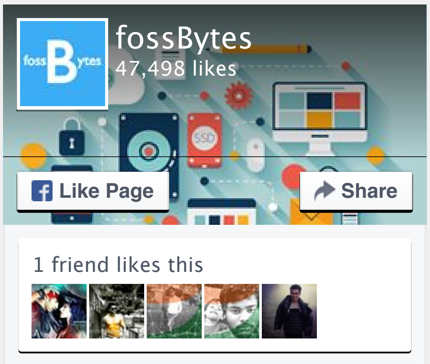
BY ANANDA VERMA · AUGUST 13, 2015



```
import sys, thread, Queue, re, urllib, urlparse, time, os, sys
dupcheck = set()
q = Queue.Queue(100)
q.put(sys.argv[1])
def queueURLs(html, origLink):
    for url in re.findall('''<a[^>]+href=["'](.[^"']+)["']''', html, re.I):
        link = url.split("#", 1)[0] if url.startswith("http") else '{uri.scheme}://{uri.netloc}'.format(uri=urlparse.urlparse(origLink)) + url.split("#", 1)[0]
        if link in dupcheck:
            continue
        dupcheck.add(link)
        if len(dupcheck) > 99999:
            dupcheck.clear()
        q.put(link)
def getHTML(link):
    try:
        html = urllib.urlopen(link).read()
        open(str(time.time()) + ".html", "w").write("" % link  + "\n" + html)
        queueURLs(html, link)
    except (KeyboardInterrupt, SystemExit):
        raise
    except Exception:
        pass
while True:
    thread.start_new_thread( getHTML, (q.get(),))
    time.sleep(0.5)
```

***Short Bytes:** Web crawler is a program that browses the Internet (World Wide Web) in a predetermined, configurable and automated manner and performs given action on crawled content. Search engines like Google and Yahoo use spidering as a means of providing up-to-date data.*

ebhose.io, a company which provides direct access to live data from



FOSSBYTES NEWSLETTER

Email *

☑ fossBytes Daily Top headlines, delivered daily

☑ fossBytes Monthly Most popular

hundreds of thousands of forums, news and blogs, on Aug 12, 2015, posted the articles describing a tiny, multi-threaded web crawler written in python. This python web crawler is capable of crawling the entire web for you. Ran Geva, the author of this tiny python web crawler says that:

> *I wrote as "Dirty", "Iffy", "Bad", "Not very good". I say, it gets the job done and downloads thousands of pages from multiple pages in a matter of hours. No setup is required, no external imports, just run the following python code with a seed site and sit back (or go do something else because it could take a few hours, or days depending on how much data you need).*

The python based multi-threaded crawler is pretty simple and very fast. It is capable of detecting and eliminating duplicate links and saving both source and link which can later be used in finding inbound and outbound links for calculating page rank. It is completely free and the code is listed below:

```
import sys, thread, Queue, re, urllib, ur
dupcheck = set()
q = Queue.Queue(100)
q.put(sys.argv[1])
def queueURLs(html, origLink):
    for url in re.findall('''<a[^>]+href=
```

stories, delivered end of every month

☑ fossBytes Weekly Most popular stories, delivered Monday morning

Subscribe

More From fossBytes

WHAT'S HOT

Happy Birthday Google : Some Fun Facts About the Company You Didn't Know

27 SEP, 2015

Yes, This GIF Takes 1,000 Years to Play Till the End

26 SEP, 2015

Stealth Dark Matter Could Be The Key to Universe's Missing Mass

28 SEP, 2015

```
        link = url.split("#", 1)[0] if ur
        if link in dupcheck:
            continue
        dupcheck.add(link)
        if len(dupcheck) > 99999:
            dupcheck.clear()
        q.put(link)
def getHTML(link):
    try:
        html = urllib.urlopen(link).read(
        open(str(time.time()) + ".html",
        queueURLs(html, link)
    except (KeyboardInterrupt, SystemExit
        raise
    except Exception:
        pass
while True:
    thread.start_new_thread( getHTML, (q.
    time.sleep(0.5)
```

Save the above code with some name lets say “myPythonCrawler.py”. To start crawling any website just type:

```
$ python myPythonCrawler.py http://fossby
```

Sit back and enjoy this web crawler in python. It will download the entire site for you.

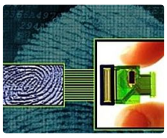
Do you like this dead simple python

us know in comments.

Get Pure Python Hacker Bundle here.

Download our Google chrome, Mozilla firefox and Opera extension to get instant updates -

3.1K 32

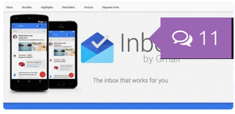
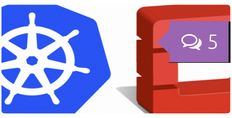
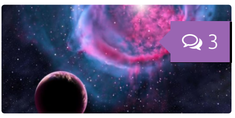


Tags: Crawler Crawler in Python Python web crawler

Ananda Verma

Writes for machines mostly. Sometimes for humans too.

YOU MAY ALSO LIKE...

All You Need to Know About Google Inbox – Reinvention of Email, Potential Gmail Killer

25 OCT, 2014

11

Google Joins OpenStack Foundation to Promote Open Source Technologies

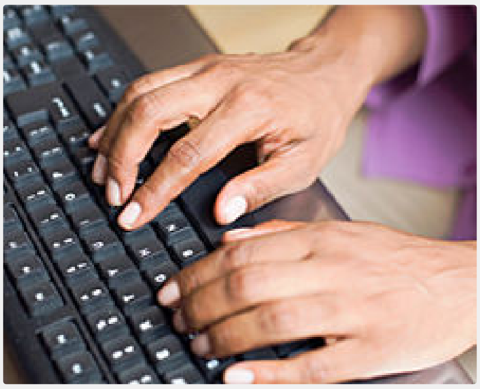
21 JUL, 2015

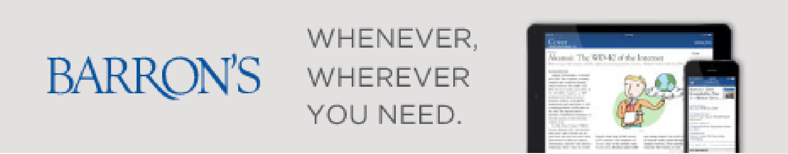
5

Eight New Planets Discovered In Goldilocks Zone: NASA

12 JAN, 2015

3





6 Comments

Sort by Top ▾

Add a comment...

Jayant Prabhakar · Bharati Vidyapeeth's College Of Engineering,Delhi

interesting material....

Like · Reply · Aug 19, 2015 12:22am

Martin Thugy · California High School

There is a simpler way with less code and 100% working!!!!

Like · Reply · Aug 14, 2015 1:27am

El Cora · Teacher at Escuela Miguel Hidalgo Y Costilla

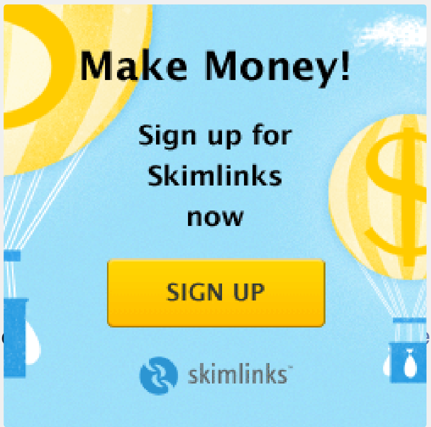
Very interesting material. Thanks

Like · Reply · Aug 13, 2015 10:13pm

fossBytes

You are welcome 🙂

Like · Reply · Aug 18, 2015 6:06pm



Anas Saeed · Government College Lahore

does not work as it gets stuck in the queueURLs function (here n s.com)

Like · Reply · Aug 13, 2015 2:28pm

Ran Geva · בית ספר הטכני בחיפה

The problem is that FossBytes uses CloudFlare to block crawlers. The first and only file the crawler downloads shows the following error:
Access denied
What happened?

The owner of this website (fossbytes.com) has banned your access based on your browser's signature

Like · Reply · Aug 13, 2015 3:59pm

Inspired Systems

Traceback (most recent call last):
File "pythonCrawler.py", line 1, in <module>
import sys, thread, Queue, re, urllib, urlparse, time, os, sys
ImportError: No module named 'thread'

Like · Reply · Aug 19, 2015 10:07pm

Jess Jukić · Swinburne University of Technology

_thread

Like · Reply · Sep 13, 2015 5:40am

Jean Paul Ruiz

File "web-crawler.py", line 7
link = url.split("#", 1)[0] if url.startswith("http") else '{uri.scheme}://{uri.netloc}'.format(uri=urlparse.urlparse(origLink)) + url.split("#", 1)[0]
^
SyntaxError: invalid syntax

Like · Reply · Aug 13, 2015 11:55pm

Ran Geva · בית ספר הטכני בחיפה

What python version are you using? This condition formatting works on python 2.6+

Like · Reply · Aug 14, 2015 2:48am

Jean Paul Ruiz

Ran Geva That must be the reason, Python 2.4.3. Do you know what would be the old supported format for this piece of code?

Like · Reply · Aug 14, 2015 2:55am

Ran Geva · בית ספר הטכני בחיפה

Jean Paul Ruiz It works since version 2.5:
https://mail.python.org/.../2005-September/056846.html

Like · Reply · Aug 16, 2015 1:37am

Show 1 more reply in this thread▼

Facebook Comments Plugin

Comments Community Login

Recommend Sort by Best

Start the discussion…

Be the first to comment.

DISQUS

ABOUT

- About Us
- Advertise With Us
- Contact Us
- fossBytes Team
- Privacy Policy
- Review Guidelines
- Sponsored Post Guidelines

TIMELINE

- Live Coverage: All Events
- Live from Google I/O 2015
- Live from Microsoft's Build 2015 keynote

SUBSCRIBE TO FOSSBYTES DAILY

Latest headlines delivered to you daily

Email *

Subscribe

MORE FROM FOSSBYTES

VN beta

Home
Tech
Entertainment
Business
Reviews
More
Science
Security
Geek

How To
DIY
Advertise