



OLD DOMINION UNIVERSITY

CS 432 WEB SCIENCE

# Assignment One

*Derek Goddeau*

Professor

Michael L. Nelson

January 26, 2017

---

# 1 POST to a form with curl

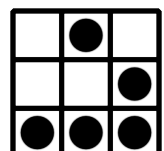
In order to submit POST data to a form using `curl` first it must be ensured that the form accepts POST data. This can be done by viewing the page source and verifying that the form tag has `method="post"` as in the nostarch.com search bar form tag shown somewhat abridged below.

```
<form action="/" method="post" id="search-theme-form">
<input name="search_theme_form" value="" class="form-text"/>
<input name="op" value="Search" class="form-submit"/>
<input type="hidden" name="form_build_id" value="form-6Skwd"/>
<input type="hidden" name="form_id" value="search_theme_form"/>
</form>
```

In order to craft the `curl` command the `-d` flag can be used along with the `"name=value"` pattern for each input to the form where `name` is copied from each input tag and `value` is changed in the fields where the default values are not desired.

```
curl -L -i -o results.html \
    -d "search_theme_form=$1" \
    -d "op=Search" \
    -d "form_build_id=form-6SkwdjCka872mUDOLyJspWzIHtkBGso7f5RMZ2fGr9U" \
    -d "form_id=search_theme_form" \
    https://www.nostarch.com/
```

The command `curl_post.sh car` will return a page with the search results for "car" on nostarch.com. Inspecting the output `results.html` the HTTP/1.1 200 OK after a single redirect and lack of a 405 Method not allowed error means the request was successful.



Search | No Starch Press

file:///home/datenstrom/workspace/cs532-s17/assignments/assignment\_one/ctSuchen

HTTP/1.1 302 Moved Temporarily Date: Sun, 22 Jan 2017 05:25:38 GMT Content-Type: text/html; charset=utf-8 Transfer-Encoding: chunked Connection: keep-alive Set-Cookie: \_\_cfduid=d12d05d49dd5a2620f839ba1f652b1b161485062738; expires=Mon, 22-Jan-18 05:25:38 GMT; path=/; domain=.nostarch.com; HttpOnly X-Powered-By: PHP/5.2.17 Expires: Sun, 19 Nov 1978 05:00:00 GMT Cache-Control: store, no-cache, must-revalidate, post-check=0, pre-check=0 Set-Cookie: SESS1ff143602f7518d305560cea1fca05f6=08c621a968633fd6cef8ace9d57d603b; expires=Tue, 14-Feb-2017 08:58:58 GMT; path=/; domain=.nostarch.com Last-Modified: Sun, 22 Jan 2017 05:25:38 GMT Location: https://www.nostarch.com/search/node/car Server: cloudflare-nginx CF-RAY: 32509624cee02432-IAD HTTP/1.1 200 OK Date: Sun, 22 Jan 2017 05:25:38 GMT Content-Type: text/html; charset=utf-8 Transfer-Encoding: chunked Connection: keep-alive Set-Cookie: \_\_cfduid=d12d05d49dd5a2620f839ba1f652b1b161485062738; expires=Mon, 22-Jan-18 05:25:38 GMT; path=/; domain=.nostarch.com; HttpOnly X-Powered-By: PHP/5.2.17 Expires: Sun, 19 Nov 1978 05:00:00 GMT Cache-Control: must-revalidate Set-Cookie: SESS1ff143602f7518d305560cea1fca05f6=d9cc6715f497a720833d02b1f190a234; expires=Tue, 14-Feb-2017 08:58:58 GMT; path=/; domain=.nostarch.com Last-Modified: Sun, 22 Jan 2017 05:16:27 GMT Server: cloudflare-nginx CF-RAY: 325096263f802432-IAD

Home

Search this site:

Search

Catalog

Media

Write for Us

About Us

Topics

Art & Design

General Interest

Hacking & Computer Security

Hardware / DIY

Kids

LEGO®

LEGO®

MINDSTORMS®

Linux & BSD

Manga

Programming

Python

Science & Math

System Administration

EARLY ACCESS

Free ebook edition with every print book purchased from nostarch.com!

Shopping cart

View your shopping cart.

User login

Log in

Create account

Search

Enter your keywords:

car

Search

Advanced search

Containing any of the words:

Containing the phrase:

Containing none of the words:

Catalog

Art, Photography, Design

Early Access

LEGO MINDSTORMS

Merchandise

Python

Business

For Kids

General Computing

Hardware and DIY

Only in the category(s):

Only of the type(s):

Blog entry

Newsletter issue

Page

Poll

Product

Product kit

Story

Advanced search

Want sweet Sign up for newslett

Subscribe

Search results

Car Hacker's Handbook

... and OpenGarages.org. Craig is a frequent speaker on car hacking and has run workshops at RSA, DEF CON, and other major security ... needs more hackers, and the world definitely needs more car hackers. We're all safer when the systems we depend upon are inspectable, ...

PAGE 2 OF 7

DEREK GODDEAU (DATENSTROM)  
OLD DOMINION UNIVERSITY

DATENSTROM.GITLAB.IO/INDEX

---

## 2 A Python program that finds PDFs

The Common House Spider can take any number of URIs as input optionally from a specified file with the `-f` flag, and use multiple threads using the `-t` flag. It outputs all PDF URIs on the page and the PDF size as reported by the server. Note that the `-u` or `--ugly` parameter must be passed to print first and last URI.

```
datenstrom@redacted$ python cli.py -t 2 www.nostarch.com/carhacking https://www.nostarch.com/blackhatpython -u
```

```
[*] Crawling pages:
```

```
www.nostarch.com/carhacking
```

```
https://www.nostarch.com/blackhatpython
```

```
[*] Spinning up with 2 threads
```

```
[*] Thread 1 discovered 3 PDF links for https://www.nostarch.com/blackhatpython
```

```
[*] Thread 1 removed 0 duplicate PDF files
```

```
First link: http://www.nostarch.com/download/BlackHatPython_ch07.pdf
```

```
Last link: https://www.nostarch.com/download/BlackHatPython_ch07.pdf
```

```
PDF size: 88339
```

```
First link: http://www.nostarch.com/download/BlackHatPython_dTOC.pdf
```

```
Last link: https://www.nostarch.com/download/BlackHatPython_dTOC.pdf
```

```
PDF size: 54377
```

```
First link: http://www.nostarch.com/download/BlackHatPython_Index.pdf
```

```
Last link: https://www.nostarch.com/download/BlackHatPython_Index.pdf
```

```
PDF size: 116530
```

```
[*] Thread 0 discovered 5 PDF links for www.nostarch.com/carhacking
```

```
[*] Thread 0 removed 1 duplicate PDF file
```

```
First link: http://www.nostarch.com/download/Car Hackers Handbook_sample_Chapter5.pdf
```

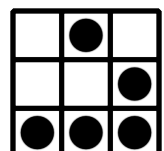
```
Last link: https://www.nostarch.com/download/Car%20Hackers%20Handbook_sample_Chapter5.pdf
```

```
PDF size: 1713557
```

```
First link: http://www.nostarch.com/download/Car Hackers Handbook_sample_Chapter5.pdf
```

```
Last link: https://www.nostarch.com/download/Car%20Hackers%20Handbook_sample_Chapter5.pdf
```

```
PDF size: 1713557
```



---

First link: [http://www.nostarch.com/download/Car Hackers Handbook\\_sample\\_dT0C.pdf](http://www.nostarch.com/download/Car Hackers Handbook_sample_dT0C.pdf)  
Last link: [https://www.nostarch.com/download/Car%20Hackers%20Handbook\\_sample\\_dT0C.pdf](https://www.nostarch.com/download/Car%20Hackers%20Handbook_sample_dT0C.pdf)  
PDF size: 594880

First link: [http://www.nostarch.com/download/Car Hackers Handbook\\_sample\\_index.pdf](http://www.nostarch.com/download/Car Hackers Handbook_sample_index.pdf)  
Last link: [https://www.nostarch.com/download/Car%20Hackers%20Handbook\\_sample\\_index.pdf](https://www.nostarch.com/download/Car%20Hackers%20Handbook_sample_index.pdf)  
PDF size: 660045

First link: [https://www.usenix.org/system/files/login/articles/login\\_summer16\\_19\\_books.pdf](https://www.usenix.org/system/files/login/articles/login_summer16_19_books.pdf)  
Last link: [https://www.usenix.org/system/files/login/articles/login\\_summer16\\_19\\_books.pdf](https://www.usenix.org/system/files/login/articles/login_summer16_19_books.pdf)  
PDF size: 81289

[\*] PDF links discovered in 20.1669859409 seconds

It is also both Python 2.6+ and Python 3 compatible:

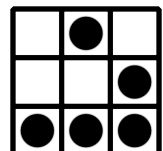
```
datenstrom@redacted$ python3 cli.py http://www.cs.odu.edu/~mln/teaching/cs532-s17/test/pdfs.html -u
[*] Crawling pages:
http://www.cs.odu.edu/~mln/teaching/cs532-s17/test/pdfs.html
[*] Spinning up with 1 thread
[*] Thread 0 discovered 11 PDF links for http://www.cs.odu.edu/~mln/teaching/cs532-s17/test/pdfs.html
[*] Thread 0 removed 0 duplicate PDF files

[*] Thread 0 discovered 11 PDF links for http://www.cs.odu.edu/~mln/teaching/cs532-s17/test/pdfs.html
[*] Thread 0 removed 0 duplicate PDF files
```

First link: <http://www.cs.odu.edu/~mln/pubs/ht-2015/hypertext-2015-temporal-violations.pdf>  
Last link: <http://www.cs.odu.edu/~mln/pubs/ht-2015/hypertext-2015-temporal-violations.pdf>  
PDF size: 2184076

First link: <http://www.cs.odu.edu/~mln/pubs/tpdl-2015/tpdl-2015-annotations.pdf>  
Last link: <http://www.cs.odu.edu/~mln/pubs/tpdl-2015/tpdl-2015-annotations.pdf>  
PDF size: 622981

First link: <http://arxiv.org/pdf/1512.06195>  
Last link: <https://arxiv.org/pdf/1512.06195.pdf>  
PDF size: 1748961



---

First link: <http://www.cs.odu.edu/~mln/pubs/tpdl-2015/tpdl-2015-off-topic.pdf>

Last link: <http://www.cs.odu.edu/~mln/pubs/tpdl-2015/tpdl-2015-off-topic.pdf>

PDF size: [4308768](#)

First link: <http://www.cs.odu.edu/~mln/pubs/tpdl-2015/tpdl-2015-stories.pdf>

Last link: <http://www.cs.odu.edu/~mln/pubs/tpdl-2015/tpdl-2015-stories.pdf>

PDF size: [1274604](#)

First link: <http://www.cs.odu.edu/~mln/pubs/tpdl-2015/tpdl-2015-profiling.pdf>

Last link: <http://www.cs.odu.edu/~mln/pubs/tpdl-2015/tpdl-2015-profiling.pdf>

PDF size: [639001](#)

First link: <http://www.cs.odu.edu/~mln/pubs/jcdl-2014/jcdl-2014-brunelle-damage.pdf>

Last link: <http://www.cs.odu.edu/~mln/pubs/jcdl-2014/jcdl-2014-brunelle-damage.pdf>

PDF size: [2205546](#)

First link: <http://bit.ly/1ZDatNK>

Last link: <http://www.cs.odu.edu/~mln/pubs/jcdl-2015/jcdl-2015-temporal-intention.pdf>

PDF size: [720476](#)

First link: <http://www.cs.odu.edu/~mln/pubs/jcdl-2015/jcdl-2015-mink.pdf>

Last link: <http://www.cs.odu.edu/~mln/pubs/jcdl-2015/jcdl-2015-mink.pdf>

PDF size: [1254605](#)

First link: <http://www.cs.odu.edu/~mln/pubs/jcdl-2015/jcdl-2015-arabic-sites.pdf>

Last link: <http://www.cs.odu.edu/~mln/pubs/jcdl-2015/jcdl-2015-arabic-sites.pdf>

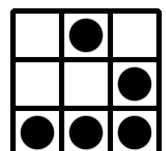
PDF size: [709420](#)

First link: <http://www.cs.odu.edu/~mln/pubs/jcdl-2015/jcdl-2015-dictionary.pdf>

Last link: <http://www.cs.odu.edu/~mln/pubs/jcdl-2015/jcdl-2015-dictionary.pdf>

PDF size: [2350603](#)

[\*] PDF links discovered in [14.306671047210693](#) seconds

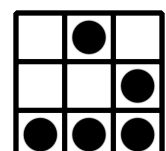
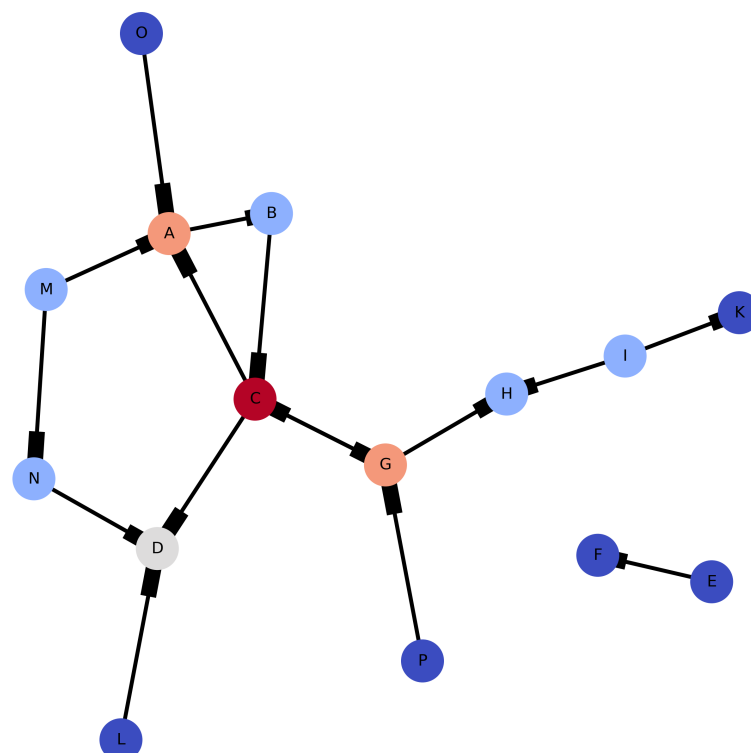


### 3 Graph Structure

The sample graph below is the dataset that will be used to demonstrate the **SCC**, **IN**, **OUT**, **DISCONNECTED**, **TUBES**, and **TENDRILS** components. The heatmapping in figure one is based on the degree for each node. Using this directed graph the single **SCC** component can be found, it contains all of the nodes which are reachable from eachother. In this sample graph these nodes are **A**, **B**, **C**, and **G** which are color coded red in figure 2.

Once the SCC has been discovered, the IN and OUT components can be found. These consist of the nodes that link only into or out of the SCC respectively. The IN component consists of nodes O, M, and P which are colored green in figure 2. The OUT components are H and D, yellow in figure 2.

Figure 1: Graph heatmap by node degree



---

The **DISCONNECTED** component contains all nodes unreachable from the other components, which are the grey nodes F and E. **TUBES** are nodes which connect **IN** and **OUT** nodes, there is only one node in this example N colored purple. Finally the **TENDRILS** are the blue nodes I, K, and L which shoot off of the **IN** and **OUT** components but do not directly interact with the **SCC**.

Figure 2: Graph components

