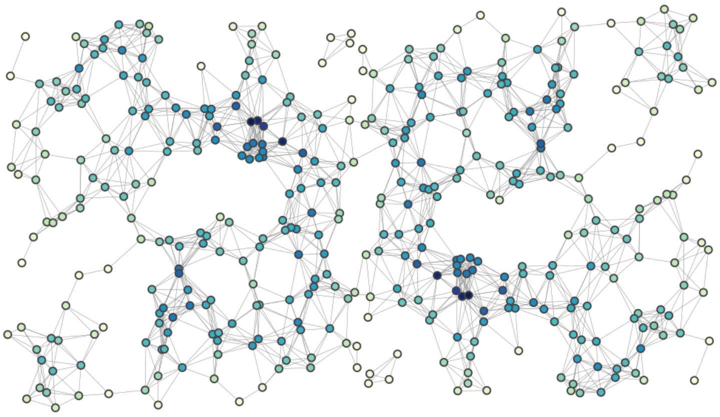


Old Dominion University

CS 432 Web Science

Assignment One

Derek Goddeau

Professor

Michael L. Nelson

January 26, 2017

1 POST to a from with `curl`

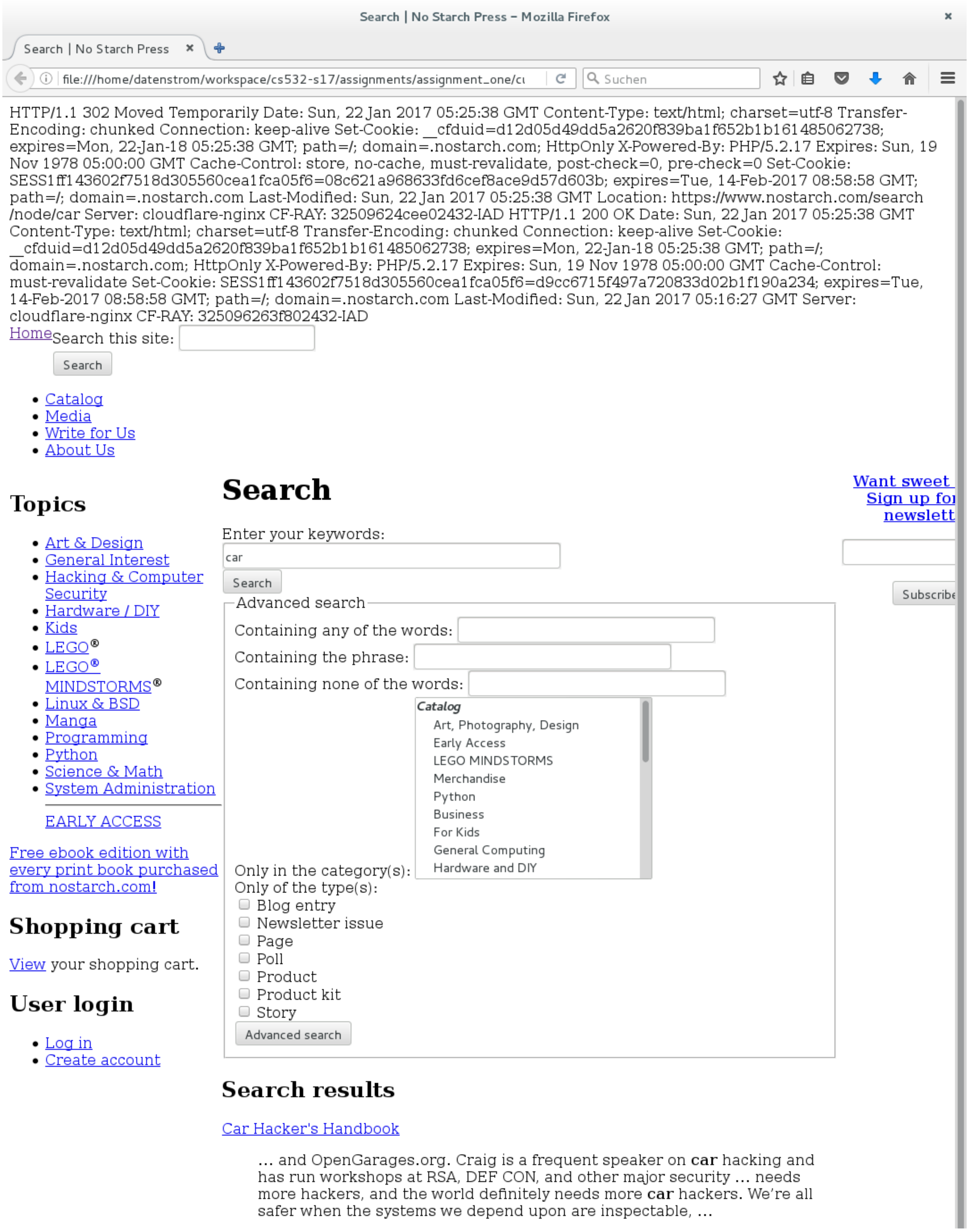
In order to submit `POST` data to a form using `curl` first it must be ensured that the form accepts `POST` data. This can be done by viewing the page source and verifying that the form tag has `method="post"` as in the nostarch.com search bar form tag shown somewhat abridged below.

```
<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-6SkwdjCka872mUD0LyJspWzIHtkBGso7f5RMZ2fGr9U" \
           -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 – Mozilla Firefox
Search | No Starch Press
file:///home/datenstrom/workspace/cs532-s17/assignments/assignment_one/cu
Suchen

```
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
```

HomeSearch this site:

- 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

Advanced search

Containing any of the words:

Containing the phrase:

Containing none of the words:

Only in the category(s):

Only of the type(s):

□ Blog entry
□ Newsletter issue
□ Page
□ Poll
□ Product
□ Product kit
□ Story

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, …

Want sweet
Sign up fo
newslett

2 A Python program that finds PDFs

The `Common House Spider` can take any number of URIs as input optionaly 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.

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

=========================================================  =============
PDF link                                                     size: bytes
=========================================================  =============
http://www.nostarch.com/download/BlackHatPython_ch07.pdf           88339
http://www.nostarch.com/download/BlackHatPython_Index.pdf         116530
http://www.nostarch.com/download/BlackHatPython_dTOC.pdf           54377
=========================================================  =============

[*] Thread 0 discovered 5 PDF links for www.nostarch.com/carhacking
[*] Thread 0 removed 1 duplicate PDF file

==============================================================================  =============
PDF link                                                                          size: bytes
==============================================================================  =============
http://www.nostarch.com/download/Car Hackers Handbook_sample_dTOC.pdf                  594880
https://www.usenix.org/system/files/login/articles/login_summer16_19_books.pdf          81289
http://www.nostarch.com/download/Car Hackers Handbook_sample_index.pdf                 660045
http://www.nostarch.com/download/Car Hackers Handbook_sample_Chapter5.pdf             1713557
==============================================================================  =============

[*] PDF links discovered in 20.1669859409 seconds
```

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

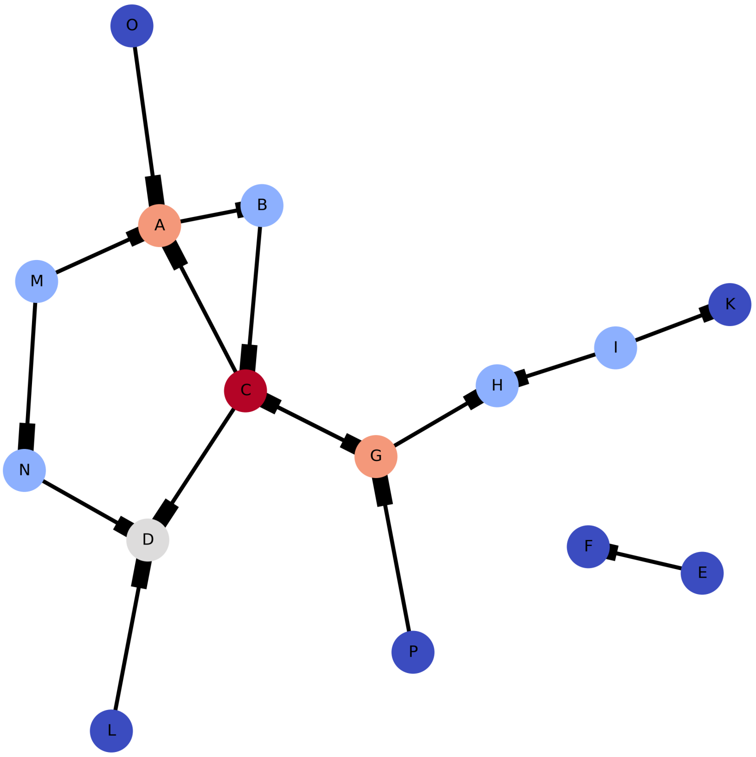
```
datenstrom@redacted$ python3 cli.py http://www.cs.odu.edu/~mln/teaching/cs532-s17/test/pdfs.html
[*] Crawling pages:
http://www.cs.odu.edu/~mln/teaching/cs532-s17/test/pdfs.html
[*] Spinning up with 1 thread
[*] Thread 0 discovered 11 PDF links for http://www.cs.odu.edu/~mln/teaching/cs532-s17/test/pdfs.html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


==============================================================================  =============
PDF link                                                                          size: bytes
==============================================================================  =============
http://arxiv.org/pdf/1512.06195                                                       1748961
http://bit.ly/1ZDatNK                                                                  720476
http://www.cs.odu.edu/~mln/pubs/jcdl-2015/jcdl-2015-mink.pdf                          1254605
http://www.cs.odu.edu/~mln/pubs/tpdl-2015/tpdl-2015-annotations.pdf                    622981
http://www.cs.odu.edu/~mln/pubs/tpdl-2015/tpdl-2015-off-topic.pdf                     4308768
http://www.cs.odu.edu/~mln/pubs/jcdl-2015/jcdl-2015-arabic-sites.pdf                   709420
http://www.cs.odu.edu/~mln/pubs/tpdl-2015/tpdl-2015-stories.pdf                       1274604
http://www.cs.odu.edu/~mln/pubs/jcdl-2014/jcdl-2014-brunelle-damage.pdf               2205546
http://www.cs.odu.edu/~mln/pubs/tpdl-2015/tpdl-2015-profiling.pdf                      639001
http://www.cs.odu.edu/~mln/pubs/jcdl-2015/jcdl-2015-dictionary.pdf                    2350603
http://www.cs.odu.edu/~mln/pubs/ht-2015/hypertext-2015-temporal-violations.pdf        2184076
==============================================================================  =============

[*] 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 heatmaping 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.

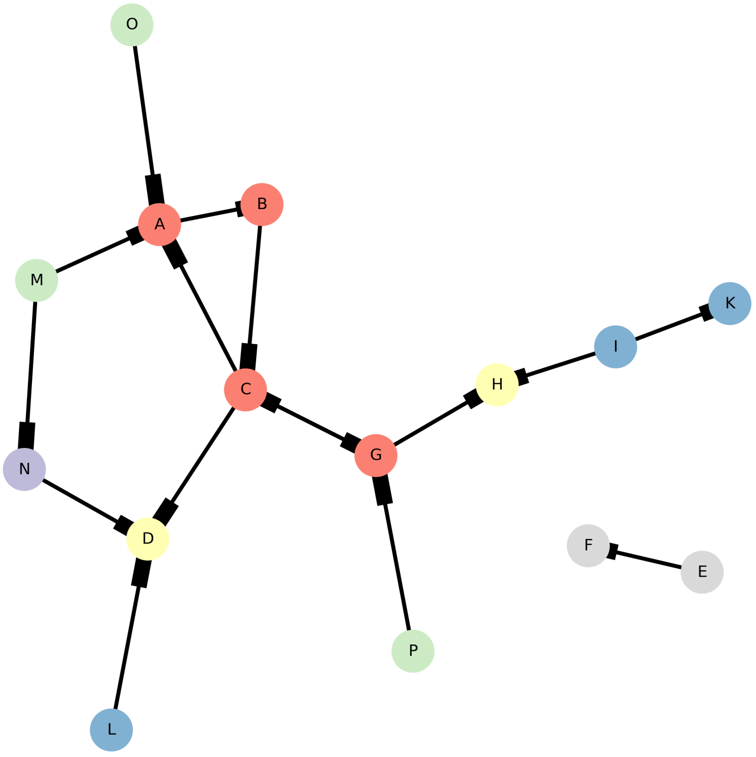
Once the `SCC` has been discovered, the `IN` and `OUT` components can be found. These consist of the nodes that link only into or out of the `SCC` respectively. The `IN` component consists of nodes `O`, `M`, and `P` which are colored green in figure 2. The `OUT` components are `H` and `D`, yellow in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


O
A
B
M
K
I
H
C
G
N
D
F
E
P
L