

Srividya Majeti

Assignment 1

CS 532: Introduction to Web Science

Dr. Michael Nelson

Spring 2016

January 28, 2016

Contents

1	Question 1	1
2	Question 2	3
3	Question 3	7
	References	11

Question 1

Demonstrate that you know how to use “curl” well enough to correctly POST data to a form. Show that the HTML response that is returned is “correct”. That is, the server should take the arguments you POSTed and build a response accordingly. Save the HTML response to a file and then view that file in a browser and take a screen shot.

To POST data to a form using cURL, I did the following:

- I created a form using PHP with name as an input element and a submit button.
- When we open this PHP form in the browser and type any name, it displays a response saying “Welcome followed by the text”.
- By using the following cURL command we can POST data to the form which is received by the server and generates a response with the same message that we see on the browser. By using -o followed by parameter, I am outputting the HTML response to a file.

```
curl -d name=Srividya 'www.cs.odu.edu/~smajeti/postForm.php' -o output.html
```

```

1 <?php
2     if( $_POST["name"] ) {
3         if (preg_match("/^[A-Za-z'-]/",$_POST['name'] )) {
4             die ("invalid name and name should be alpha");
5         }
6         echo "Welcome ". $_POST['name'];
7         echo "\n";
8
9         exit();
10    }
11 ?>
12 <form method ="POST" action="<?php echo $_SERVER['PHP_SELF
13     '];?>">
14     <input type="text" name="name">
15     <input type="submit" name="submit" value="OK">
    </form>

```

Listing 1.1. “PHP script”

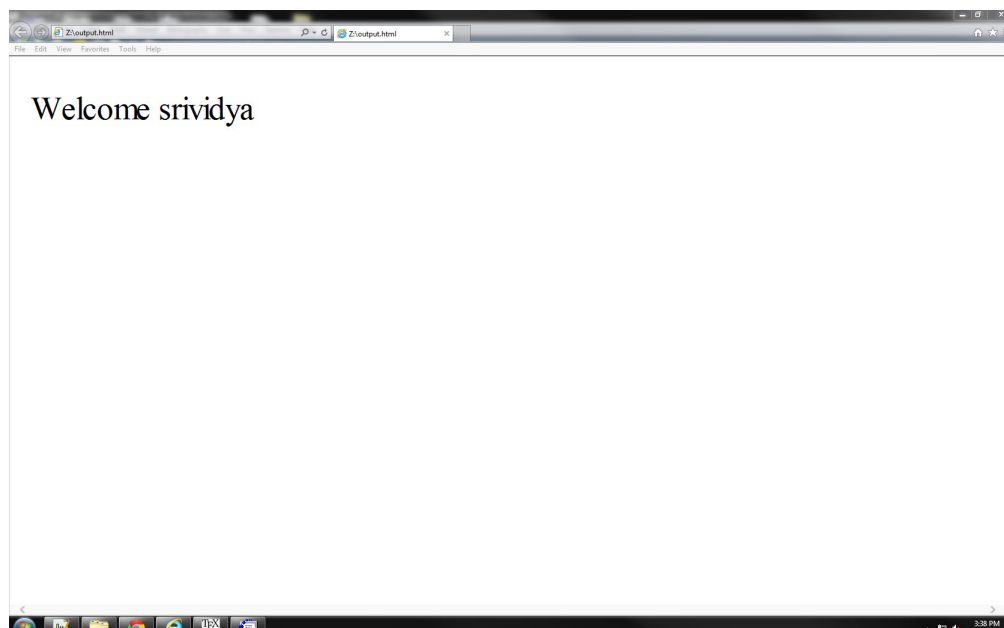


Fig. 1.1. HTML response saved into a file and viewed in browser.

Question 2

Write a Python program that:

- takes as a command line argument a web page
- extracts all the links from the page
- lists all the links that result in PDF files, and prints out the bytes for each of the links. (note: be sure to follow all the redirects until the link terminates with a “200 OK”.)
- show that the program works on 3 different URIs, one of which needs to be: `http://www.cs.odu.edu/~mln/teaching/cs532-s16/test/pdfs.html`

For solving the above problem I have written a program using Python. Following are the steps I’ve taken to solve the problem:

- The python script takes the web page as command line argument, finds all the links with “a” tag and passes its href to the function `getsize()`.
- This function fetches the content-length, content-type and status code for each link and extracts all the links that result in PDF files by checking if the content-type is “application/pdf” and if the status code is “200”.
- The output of the program prints the links that are PDF files and its size.

```

1 import urllib2
2 import sys #to pass link as command line argument
3 from bs4 import BeautifulSoup, SoupStrainer
4
5 uri= sys.argv[1]
6 request= urllib2.Request(uri)
7 response= urllib2.urlopen(request)
8 response.getcode()
9 soup = BeautifulSoup(response, "html5lib")
10
11 def getsize(link):
12     file= urllib2.urlopen(link)
13     size=file.headers.get("content-length")
14     type=file.headers.get("content-type")
15     status=file.getcode()
16     if status == 200 and type == "application/pdf":
17         print "found a url with pdf in the link:",
18             link
19         print "size:", size, "bytes"
20         print "status:", status
21     file.close()
22 for link in soup.findAll('a'):
23     getsize(link['href'])

```

Listing 2.1. “Python code for extracting links that are PDF files”

- To run the code, go to the folder where the python code is located. Type [python extractPDF.py <link>](#) to see the output.
- I tried to run the code for URIs which use relative path in the href, but my program wasn't returning any URI with PDFs as the href gets appended to the base URI. I did not handle this case. This can be handled by appending this relative href to the base URI.
- I ran the program on three different URIs.


```

linux.cs.odu.edu - Vidya - SSH Secure Shell
File Edit View Window Help
Quick Connect Profiles

$ python 2.py https://en.wikipedia.org/wiki/Car
pdf link: http://www.wales.ac.uk/geiriadur/pdf/GPC0018-02.pdf
size: 3682723 bytes
status: 200
pdf link: http://www.theicct.org/sites/default/files/publications/ICTI_fiscalpolicies_feb2011.pdf
size: 5201399 bytes
status: 200
pdf link: http://cst.uwinnipeg.ca/documents/Transport_Greenhouse.pdf
size: 746675 bytes
status: 200
pdf link: http://research.cibcwm.com/economic_public/download/sfeb09.pdf
size: 764503 bytes
status: 200

Connected to linux.cs.odu.edu  SSH2 - aes128-cbc - hmac-md5 - n: 97x13  NUM

```

Fig. 2.1. Output for URI <https://en.wikipedia.org/wiki/Car>

```

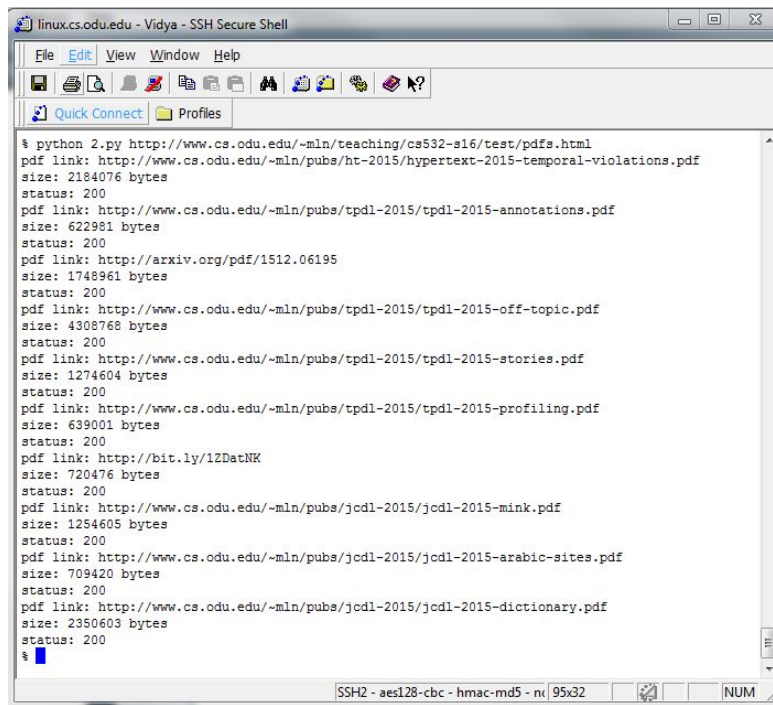
linux.cs.odu.edu - Vidya - SSH Secure Shell
File Edit View Window Help
Quick Connect Profiles

$ python 2.py http://www.cs.odu.edu/~nadeem/publications/index.html
pdf link: http://www.cs.odu.edu/~nadeem/papers/routing_mwcn.pdf
size: 156515 bytes
status: 200
pdf link: http://www.cs.odu.edu/~nadeem/papers/Rover_pwc.pdf
size: 490692 bytes
status: 200
pdf link: http://www.cs.odu.edu/~nadeem/papers/OPP_TVI.pdf
size: 1199726 bytes
status: 200
pdf link: http://www.cs.odu.edu/~nadeem/papers/reputation_TC.pdf
size: 2928938 bytes
status: 200
pdf link: http://www.cs.odu.edu/~nadeem/papers/VIIP_JSAC.pdf
size: 738801 bytes
status: 200
pdf link: http://www.cs.odu.edu/~nadeem/papers/LED_TMC.pdf
size: 2937058 bytes
status: 200
pdf link: http://www.cs.odu.edu/~nadeem/papers/context_TMC.pdf
size: 2686640 bytes
status: 200
pdf link: http://www.cs.odu.edu/~nadeem/papers/power_mobicom_page.pdf
size: 190482 bytes
status: 200
pdf link: http://www.cs.odu.edu/~nadeem/papers/Mobility_WCMC.pdf
size: 522425 bytes
status: 200
pdf link: http://www.cs.odu.edu/~nadeem/papers/TV_MC2R_web.pdf
size: 393748 bytes
status: 200
pdf link: http://www.cs.odu.edu/~nadeem/papers/IBN_MC2R.pdf
size: 89091 bytes
status: 200
pdf link: http://www.cs.odu.edu/~nadeem/papers/Rover_IEEE.pdf
size: 1170626 bytes
status: 200
pdf link: http://www.cs.odu.edu/~nadeem/papers/reputation_afri1.pdf

Connected to linux.cs.odu.edu  SSH2 - aes128-cbc - hmac-md5 - n: 93x38  NUM

```

Fig. 2.2. Output for URI <http://www.cs.odu.edu/~nadeem/publications/index.html>



```
linux.cs.odu.edu - Vidya - SSH Secure Shell
File Edit View Window Help
Quick Connect Profiles

% python 2.py http://www.cs.odu.edu/~mln/teaching/cs532-s16/test/pdfs.html
pdf link: http://www.cs.odu.edu/~mln/pubs/ht-2015/hypertext-2015-temporal-violations.pdf
size: 2184076 bytes
status: 200
pdf link: http://www.cs.odu.edu/~mln/pubs/tpdl-2015/tpdl-2015-annotations.pdf
size: 622981 bytes
status: 200
pdf link: http://arxiv.org/pdf/1512.06195
size: 1748961 bytes
status: 200
pdf link: http://www.cs.odu.edu/~mln/pubs/tpdl-2015/tpdl-2015-off-topic.pdf
size: 4308768 bytes
status: 200
pdf link: http://www.cs.odu.edu/~mln/pubs/tpdl-2015/tpdl-2015-stories.pdf
size: 1274604 bytes
status: 200
pdf link: http://www.cs.odu.edu/~mln/pubs/tpdl-2015/tpdl-2015-profiling.pdf
size: 639001 bytes
status: 200
pdf link: http://bit.ly/1ZDatNK
size: 720476 bytes
status: 200
pdf link: http://www.cs.odu.edu/~mln/pubs/jcdl-2015/jcdl-2015-mink.pdf
size: 1254605 bytes
status: 200
pdf link: http://www.cs.odu.edu/~mln/pubs/jcdl-2015/jcdl-2015-arabic-sites.pdf
size: 709420 bytes
status: 200
pdf link: http://www.cs.odu.edu/~mln/pubs/jcdl-2015/jcdl-2015-dictionary.pdf
size: 2350603 bytes
status: 200
%
```

Fig. 2.3. Output for URI <http://www.cs.odu.edu/~mln/teaching/cs532-s16/test/pdfs.html>

Question 3

Consider the “bow-tie” graph in the Broder et al. paper (fig 9):
<http://www9.org/w9cdrom/160/160.html>

Now consider the following graph:

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

For the above graph, give the values for:

IN:

SCC:

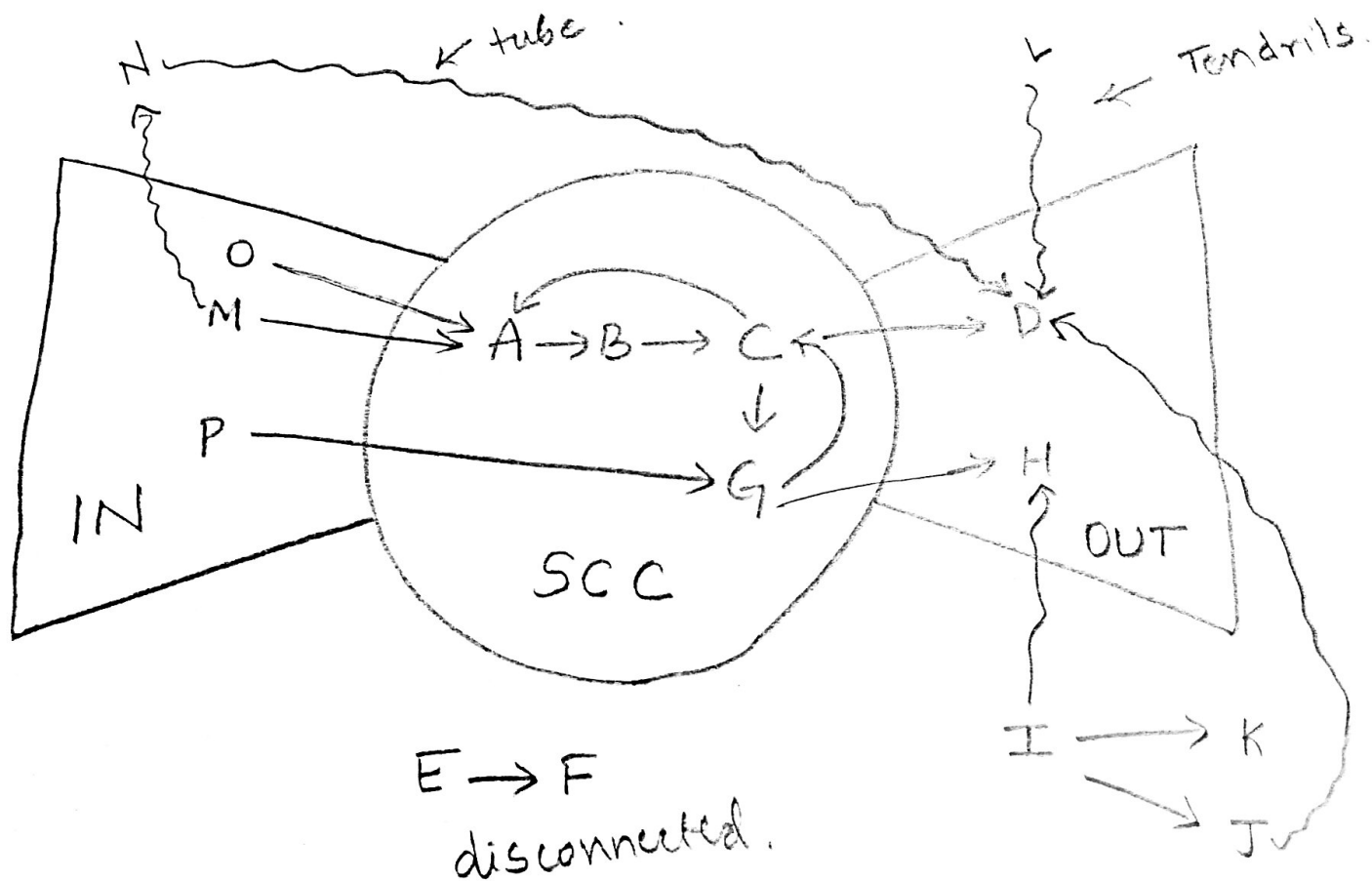
OUT:

Tendrils:

Tubes:

Disconnected:

- **SCC(heart of the web):** In our graph A, B, C, G are the strongly connected components. If we select any 2 nodes among A, B, C, G there exists a path between them.
- **IN:** O, M, P belongs to IN. These nodes can access SCC but they cannot be accessed from SCC.
- **OUT:** D, H belongs to OUT. These pages cannot access SCC but can be accessed from SCC.
- **TENDRILS:** I, J, L, N are tendrils. These pages cannot reach SCC and cannot be reached from SCC. N is a node that is reachable from portion of IN and I, J, L are nodes that can reach portions of OUT, without passing through SCC.
- **TUBES:** M, N, D form a tube.
- **DISCONNECTED:** E, F, K are disconnected nodes.



References

1. Inserting code listings in latex https://en.wikibooks.org/wiki/LaTeX/Source_Code_Listings
2. Extract links using BeautifulSoup <http://stackoverflow.com/questions/3075550/how-can-i-get-href-links-from-html-code>
3. Getting content-length <http://stackoverflow.com/questions/12317493/urllib2-urlopen-getting-the-size-of-the-content>
4. Inserting PDF in latex <http://tex.stackexchange.com/questions/149443/how-to-include-pdf-file-in-latex-doc-from-folder-location>