

HW #1 – Web Science Intro

Brandyn Winn

CS 432, Fall 2022

9/18/2022

Q1

Q1

Consider the “bow-tie” structure of the web in the Broder et al. paper “[Graph Structure in the Web](#)” that was described in Module 1.

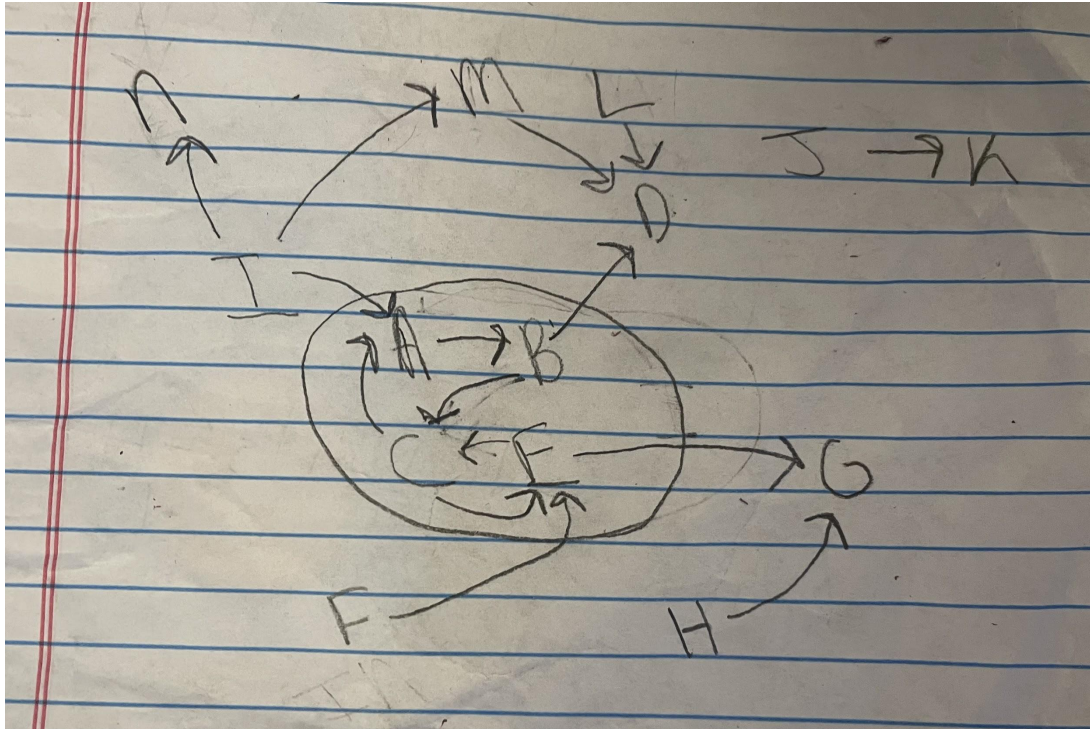
Now consider the following links:

```
A --> B
B --> C
B --> D
C --> A
C --> E
E --> C
E --> G
F --> E
H --> G
I --> A
I --> M
I --> N
J --> K
L --> D
M --> D
```

Draw the resulting [directed graph](#) (either sketch on paper or use another tool) showing how the nodes are connected to each other and include an image in your report. This does not need to fit into the bow-tie type diagram, but should look more similar to the graph on slide 24 from [Module-01 Web-Science-Architecture](#).

For the graph, list the nodes (in alphabetical order) that are each of the following categories:

- SCC:
- IN:
- OUT:
- Tendrils:
 - indicate if the tendril is reachable from IN or can reach OUT
- Tubes:
 - explain how the nodes serve as tubes
- Disconnected:



SCC:

- A
- B
- C
- E

IN:

- F
- I

OUT:

- D
- G

TENDRILS:

- H
- L

TUBES:

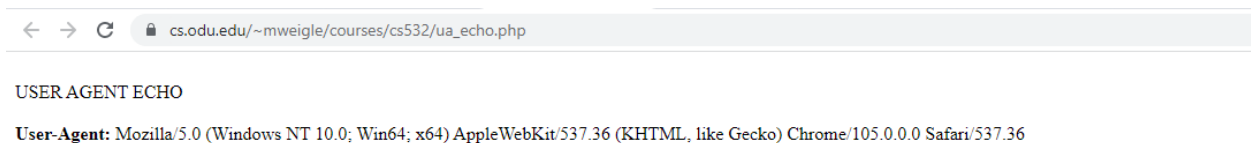
- M

DISCONNECTED:

- J
- K
- N

Q2: Demonstrate that you know how to use `curl` and are familiar with the available options.

a) First, load the URI directly in your browser and take a screenshot. The resulting webpage should show the "User-Agent" HTTP request header that your web browser sends to the web server.



b) In a single `curl` command, request the URI, show the HTTP response headers, follow any redirects, and change the User-Agent HTTP request field to "CS432/532". Show command you used and the result of your execution on the command line. (Either take a screenshot of your terminal or copy/paste into a code segment.)

References: <https://phoenixnap.com/kb/curl-command>

<https://phoenixnap.com/kb/curl-user-agent>

<code>-i, --include</code>	Specify that the output should include the HTTP response headers .
	Example:
	<code>curl -i https://example.com</code>

	Allow curl to follow any redirections .
<code>-L, --location</code>	Example:
	<code>curl -L https://example.com</code>

Change User Agent with curl

To change the curl user agent to a different browser, add the `-A` option with the wanted user agent string:

```
curl -A "user-agent-name-here" [URL]
```

```

Brandyns-MacBook-Air-2:CS432 Assignment#1 - Intro to Web Science brandynwin$ curl -i -L -A "CS432/532" http://www.cs.odu.edu/~mweigle/courses/cs532/ua_echo
.php
HTTP/1.1 301 Moved Permanently
Server: nginx/1.18.0 (Ubuntu)
Date: Sun, 18 Sep 2022 18:17:16 GMT
Content-Type: text/html
Content-Length: 178
Connection: keep-alive
Location: https://www.cs.odu.edu/~mweigle/courses/cs532/ua_echo.php

HTTP/1.1 200 OK
Server: nginx/1.18.0 (Ubuntu)
Date: Sun, 18 Sep 2022 18:17:16 GMT
Content-Type: text/html; charset=UTF-8
Transfer-Encoding: chunked
Connection: keep-alive
Vary: Accept-Encoding
Vary: Accept-Encoding

<!DOCTYPE html>
<html>
<body>

<br/>USER AGENT ECHO
<br/><br/>
<b>User-Agent:</b> CS432/532<br/>

</body>
</html>
Brandyns-MacBook-Air-2:CS432 Assignment#1 - Intro to Web Science brandynwin$

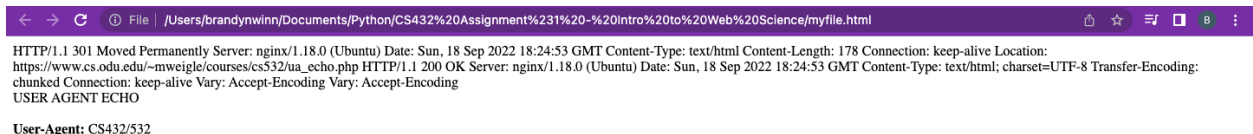
```

c) In a single `curl` command, request the URI, follow any redirects, change the User-Agent HTTP request field to "CS432/532", and save the HTML output to a file. Show the command you used and the result of your execution on the command line. View the HTML output file that was produced by `curl` in a web browser and take a screenshot.

```

Brandyns-MacBook-Air-2:CS432 Assignment#1 - Intro to Web Science brandynwin$ curl -i -L -A "CS432/532" http://www.cs.odu.edu/~mweigle/courses/cs532/ua_echo.php -o myfile.html
% Total % Received % Xferd Average Speed Time Time Time Current
      Dload Upload   Total   Spent    Left   Speed
 100  178  100  178    0    0  3368    0 --:--:-- --:--:-- --:--:-- 5933
 100  114    0  114    0    0  1181    0 --:--:-- --:--:-- --:--:-- 1181
Brandyns-MacBook-Air-2:CS432 Assignment#1 - Intro to Web Science brandynwin$

```



HTTP/1.1 301 Moved Permanently Server: nginx/1.18.0 (Ubuntu) Date: Sun, 18 Sep 2022 18:24:53 GMT Content-Type: text/html Content-Length: 178 Connection: keep-alive Location: https://www.cs.odu.edu/~mweigle/courses/cs532/ua_echo.php HTTP/1.1 200 OK Server: nginx/1.18.0 (Ubuntu) Date: Sun, 18 Sep 2022 18:24:53 GMT Content-Type: text/html; charset=UTF-8 Transfer-Encoding: chunked Connection: keep-alive Vary: Accept-Encoding Vary: Accept-Encoding
 USER AGENT ECHO
 User-Agent: CS432/532

Q3

Assignment1.py

```
1 from urllib.request import urlopen
2 from bs4 import BeautifulSoup
3 import requests
4 import re
5
6 url = input('Please enter URL:')
7
8 response = requests.get(url)
9
10 soup = BeautifulSoup(response.text, "html.parser")
11
12
13 urls = []
14
15 for links in soup.find_all('a', href=re.compile(r'(.pdf)')):
16     print("URI:" + links.get('href'))
17     print("Final URI: " + links.get('href'))
18     print("Content Length: {}".format(response.headers['Content-Length']
19 ) + " bytes")
20
21
22
23
```

```
Brandyns-MacBook-Air-2:CS432 Assignment#1 - Intro to Web Science brandynwinn$ python3 Assignment1.py
Please enter URL:https://www.cs.odu.edu/~mweigle/courses/cs532/pdfs.html
URI:http://www.cs.odu.edu/~mln/pubs/ht-2018/hypertext-2018-nwala-bootstrapping.pdf
Final URI: http://www.cs.odu.edu/~mln/pubs/ht-2018/hypertext-2018-nwala-bootstrapping.pdf
Content Length: 1122 bytes
URI:http://www.cs.odu.edu/~mln/pubs/ipres-2018/ipres-2018-atkins-news-similarity.pdf
Final URI: http://www.cs.odu.edu/~mln/pubs/ipres-2018/ipres-2018-atkins-news-similarity.pdf
Content Length: 1122 bytes
URI:http://www.cs.odu.edu/~mln/pubs/ipres-2018/ipres-2018-jones-off-topic.pdf
Final URI: http://www.cs.odu.edu/~mln/pubs/ipres-2018/ipres-2018-jones-off-topic.pdf
Content Length: 1122 bytes
URI:http://www.cs.odu.edu/~mln/pubs/ipres-2018/ipres-2018-jones-archiveit.pdf
Final URI: http://www.cs.odu.edu/~mln/pubs/ipres-2018/ipres-2018-jones-archiveit.pdf
Content Length: 1122 bytes
URI:http://www.cs.odu.edu/~mln/pubs/jcdl-2018/jcdl-2018-nwala-scraping-serps-seeds.pdf
Final URI: http://www.cs.odu.edu/~mln/pubs/jcdl-2018/jcdl-2018-nwala-scraping-serps-seeds.pdf
Content Length: 1122 bytes
URI:http://www.cs.odu.edu/~mln/pubs/jcdl-2018/jcdl-2018-kelly-private-public-web-archives.pdf
Final URI: http://www.cs.odu.edu/~mln/pubs/jcdl-2018/jcdl-2018-kelly-private-public-web-archives.pdf
Content Length: 1122 bytes
URI:http://www.cs.odu.edu/~mln/pubs/jcdl-2018/jcdl-2018-aturban-archivenow.pdf
Final URI: http://www.cs.odu.edu/~mln/pubs/jcdl-2018/jcdl-2018-aturban-archivenow.pdf
Content Length: 1122 bytes
URI:http://www.cs.odu.edu/~mln/pubs/jcdl-2018/jcdl-2018-alam-archive-banner.pdf
Final URI: http://www.cs.odu.edu/~mln/pubs/jcdl-2018/jcdl-2018-alam-archive-banner.pdf
Content Length: 1122 bytes
Brandyns-MacBook-Air-2:CS432 Assignment#1 - Intro to Web Science brandynwinn$
```