

HW 1 - Web Science Intro

Prashant Tomar

1/29/2023

Q1

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.

Answer

Figure 1 represents the plotted direct graph from the question.

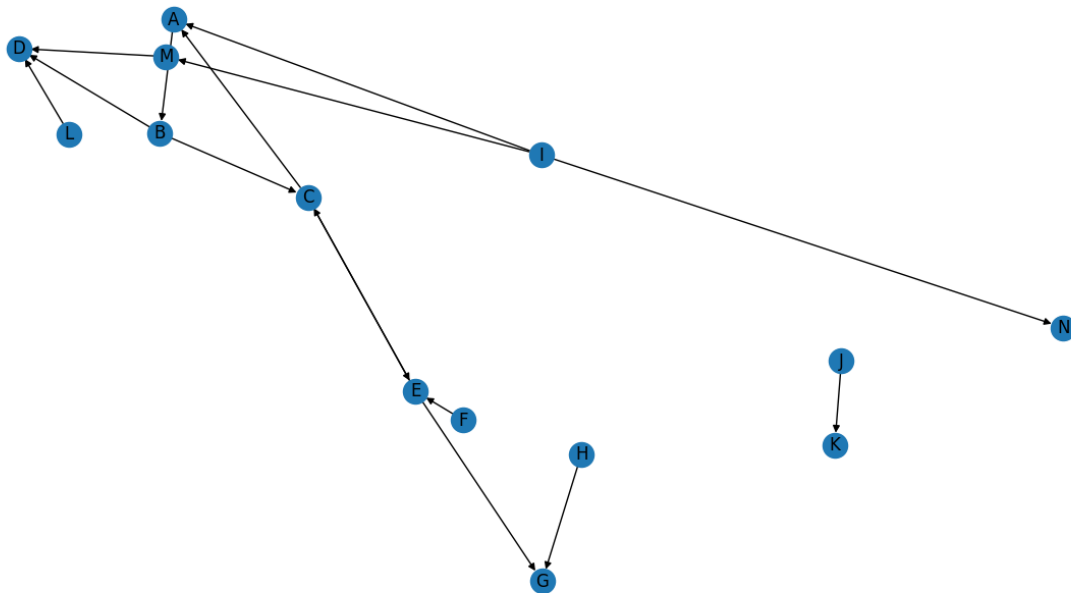


Figure 1: Directed Graph

Here is the python code to generate the directed graph.

```
1 import networkx as nx
2 import matplotlib.pyplot as plt
3
4 # Create an empty directed graph
```

```

5 G = nx.DiGraph()
6
7 # Add edges to the graph
8 G.add_edge("A", "B")
9 G.add_edge("B", "C")
10 G.add_edge("B", "D")
11 G.add_edge("C", "A")
12 G.add_edge("C", "E")
13 G.add_edge("E", "C")
14 G.add_edge("E", "G")
15 G.add_edge("F", "E")
16 G.add_edge("H", "G")
17 G.add_edge("I", "A")
18 G.add_edge("I", "M")
19 G.add_edge("I", "N")
20 G.add_edge("J", "K")
21 G.add_edge("L", "D")
22 G.add_edge("M", "D")
23
24 # Draw the graph
25 nx.draw(G, with_labels=True)
26 plt.show()

```

Listing 1: Python sample code loaded from file

Table 1 shows a nodes that are listed in the following categories.

Table 1: Simple Table

S.No.	Categories	Nodes
1	SCC:	A,B,C,E
2	IN:	D,G
3	OUT:	F,H,I,L
4	Tendrils:	N
5	Tubes:	M
6	Disconnected:	J,K

Q2

Demonstrate that you know how to use curl and are familiar with the available options

Answer

PART (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. This shows the directed web page in a browser.

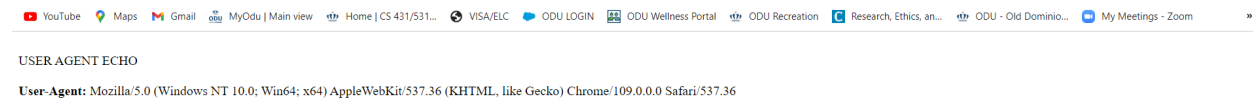


Figure 2: Webpage shows the "User-Agent" HTTP request header that your web browser sends to the web server

This is the URI loaded in the browser:

PART (b) In a single curl command, request the URI, shows the HTTP response headers, following any redirects, and change the User-Agent HTTP request field to "CS432/532". Shown on command line.

```
1 > curl -I -L -H "User-Agent: CS432/532" http://www.cs.odu.edu/~mweigle/courses/cs532/ua_echo.php
```

Listing 2: Python example copied into the LaTeX

In this CURL command -I: Show only the HTTP response headers, -L: Follow any redirects, -H: Set an HTTP request header

PART (c) In this single curl command, request the URI, follows any redirects, the User-Agent HTTP request field has been changed to "CS432/532", and then the HTML output been displayed below.

```
1 > curl -I -L -H "User-Agent: CS432/532" http://www.cs.odu.edu/~mweigle/courses/cs532/ua_echo.php --output index.html
```

Listing 3: Python example copied into the LaTeX

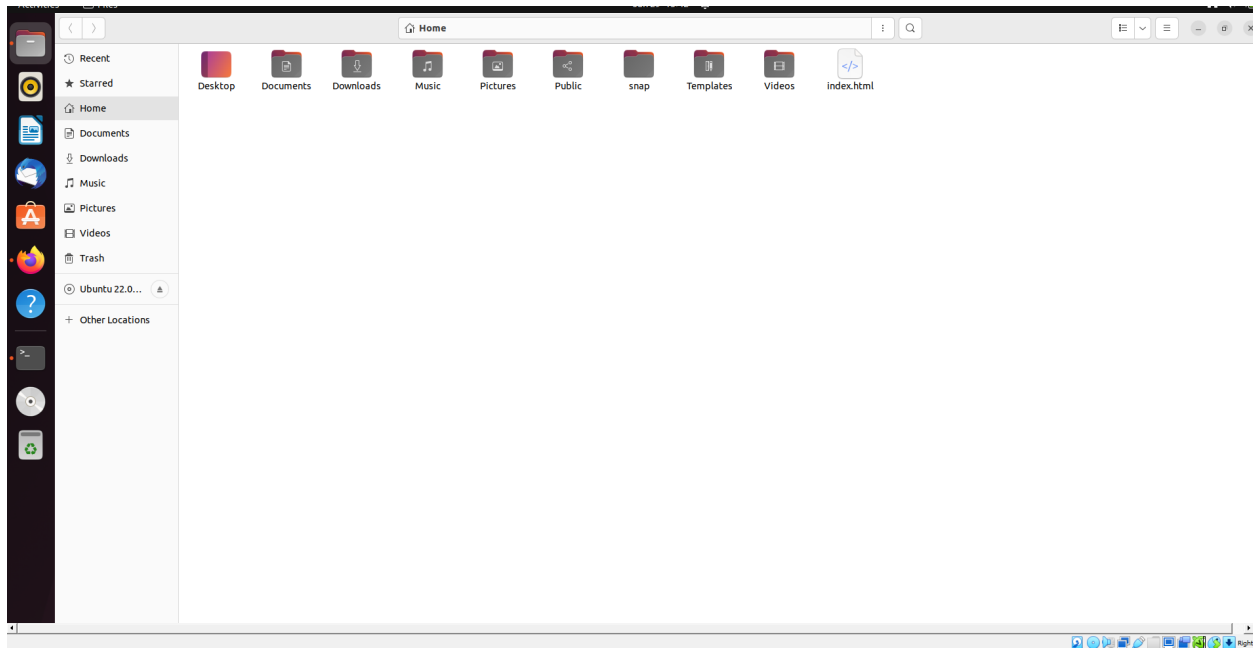


Figure 5: HTML file on local machine

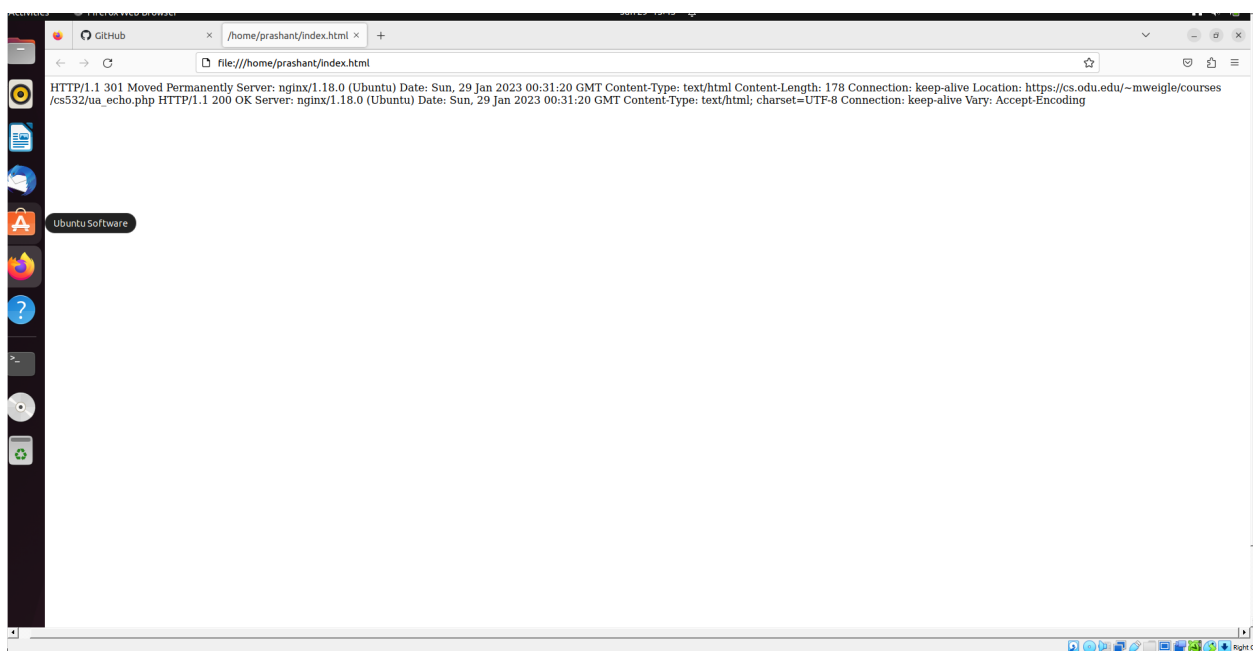


Figure 6: User-Agent HTTP request field to "CS432/532"

Q3

Write a Python program to find links to PDFs in a webpage

Answer

Here are some prints of the original URI (found in the source of the original HTML), the final URI (after any redirects), and the number of bytes in the PDF file.

Here is the python code to read the pdf from uri's.

```
1 import requests
2 from bs4 import BeautifulSoup
3
4 def get_pdf_links(url):
5     res = requests.get(url)
6     soup = BeautifulSoup(res.content, "html.parser")
7     links = [link.get("href") for link in soup.find_all("a")]
8     pdf_links = []
9     for link in links:
10         try:
11             if link.endswith(".pdf"):
12                 response = requests.head(link, allow_redirects=True)
13                 content_type = response.headers.get("Content-Type", "")
14                 .lower()
15                 if "pdf" in content_type:
16                     pdf_links.append((link, response.url, response.
17 headers.get("Content-Length", 0)))
18         except:
19             pass
20     return pdf_links
21
22 if __name__ == "__main__":
23     import sys
24     if len(sys.argv) < 2:
25         print("Usage: python findPDFs.py URL")
26         sys.exit(1)
27     url = sys.argv[1]
28     pdf_links = get_pdf_links(url)
29     for link, final_url, size in pdf_links:
30         print("URI:", link)
31         print("Final URI:", final_url)
32         print("Content Length:", size, "bytes")
33         print()
```

Listing 4: Python sample code loaded from file

Here are the screenshots of the webpages after executing the python commands),

```

PS C:\Users\prshn\OneDrive\Desktop\Web Science\WS HW 1> python main.py https://www.cs.odu.edu/~mmeigle/courses/cs532/pdfs.html
URI: http://www.cs.odu.edu/~mln/pubs/ht-2018/hypertext-2018-mala-bootstrapping.pdf
Final URI: https://www.cs.odu.edu/~mln/pubs/ht-2018/hypertext-2018-mala-bootstrapping.pdf
Content Length: 994153 bytes

URI: http://www.cs.odu.edu/~mln/pubs/ipres-2018/ipres-2018-atkins-news-similarity.pdf
Final URI: https://www.cs.odu.edu/~mln/pubs/ipres-2018/ipres-2018-atkins-news-similarity.pdf
Content Length: 1895885 bytes

URI: http://www.cs.odu.edu/~mln/pubs/ipres-2018/ipres-2018-jones-off-topic.pdf
Final URI: https://www.cs.odu.edu/~mln/pubs/ipres-2018/ipres-2018-jones-off-topic.pdf
Content Length: 3119285 bytes

URI: http://www.cs.odu.edu/~mln/pubs/ipres-2018/ipres-2018-jones-archiveit.pdf
Final URI: https://www.cs.odu.edu/~mln/pubs/ipres-2018/ipres-2018-jones-archiveit.pdf
Content Length: 2639215 bytes

URI: http://www.cs.odu.edu/~mln/pubs/jcdl-2018/jcdl-2018-mala-scraping-serps-seeds.pdf
Final URI: https://www.cs.odu.edu/~mln/pubs/jcdl-2018/jcdl-2018-mala-scraping-serps-seeds.pdf
Content Length: 2172494 bytes

URI: http://www.cs.odu.edu/~mln/pubs/jcdl-2018/jcdl-2018-kelly-private-public-web-archives.pdf
Final URI: https://www.cs.odu.edu/~mln/pubs/jcdl-2018/jcdl-2018-kelly-private-public-web-archives.pdf
Content Length: 2553579 bytes

URI: http://www.cs.odu.edu/~mln/pubs/jcdl-2018/jcdl-2018-aturban-archivenow.pdf
Final URI: https://www.cs.odu.edu/~mln/pubs/jcdl-2018/jcdl-2018-aturban-archivenow.pdf
Content Length: 3998654 bytes

URI: http://www.cs.odu.edu/~mln/pubs/jcdl-2018/jcdl-2018-alam-archive-banner.pdf
Final URI: https://www.cs.odu.edu/~mln/pubs/jcdl-2018/jcdl-2018-alam-archive-banner.pdf
Content Length: 596800 bytes

PS C:\Users\prshn\OneDrive\Desktop\Web Science\WS HW 1>

```

Figure 7: PDF:1

```

PS C:\Users\prshn\OneDrive\Desktop\532> python main.py https://mmeigle.github.io/publications/
URI: https://www.cs.odu.edu/~mmeigle/papers/jones-sigweb21.pdf
Final URI: https://www.cs.odu.edu/~mmeigle/papers/jones-sigweb21.pdf
Content Length: 2089905 bytes

URI: https://www.cs.odu.edu/~mln/pubs/ipres-2018/ipres-2018-atkins-news-similarity.pdf
Final URI: https://www.cs.odu.edu/~mln/pubs/ipres-2018/ipres-2018-atkins-news-similarity.pdf
Content Length: 1895885 bytes

URI: https://www.cs.odu.edu/~mln/pubs/ipres-2018/ipres-2018-jones-off-topic.pdf
Final URI: https://www.cs.odu.edu/~mln/pubs/ipres-2018/ipres-2018-jones-off-topic.pdf
Content Length: 3119205 bytes

URI: https://www.cs.odu.edu/~mln/pubs/ipres-2018/ipres-2018-jones-archiveit.pdf
Final URI: https://www.cs.odu.edu/~mln/pubs/ipres-2018/ipres-2018-jones-archiveit.pdf
Content Length: 2639215 bytes

URI: https://www.cs.odu.edu/~mln/pubs/ht-2018/hypertext-2018-mala-bootstrapping.pdf
Final URI: https://www.cs.odu.edu/~mln/pubs/ht-2018/hypertext-2018-mala-bootstrapping.pdf
Content Length: 994153 bytes

URI: https://www.cs.odu.edu/~mln/pubs/jcdl-2018/jcdl-2018-aturban-archivenow.pdf
Final URI: https://www.cs.odu.edu/~mln/pubs/jcdl-2018/jcdl-2018-aturban-archivenow.pdf
Content Length: 3998654 bytes

URI: https://www.cs.odu.edu/~mln/pubs/jcdl-2018/jcdl-2018-alam-archive-banner.pdf
Final URI: https://www.cs.odu.edu/~mln/pubs/jcdl-2018/jcdl-2018-alam-archive-banner.pdf
Content Length: 596800 bytes

URI: https://www.cs.odu.edu/~mln/pubs/jcdl-2018/jcdl-2018-kelly-private-public-web-archives.pdf
Final URI: https://www.cs.odu.edu/~mln/pubs/jcdl-2018/jcdl-2018-kelly-private-public-web-archives.pdf
Content Length: 2553579 bytes

URI: https://www.cs.odu.edu/~mmeigle/papers/mccoy-kdd18.pdf
Final URI: https://www.cs.odu.edu/~mmeigle/papers/mccoy-kdd18.pdf
Content Length: 1002468 bytes

URI: https://www.cs.odu.edu/~mln/pubs/jcdl-2018/jcdl-2018-mala-scraping-serps-seeds.pdf
Final URI: https://www.cs.odu.edu/~mln/pubs/jcdl-2018/jcdl-2018-mala-scraping-serps-seeds.pdf
Content Length: 2172494 bytes

URI: http://www.cs.odu.edu/~mmeigle/papers/alkwai-tois17-preprint.pdf
Final URI: https://www.cs.odu.edu/~mmeigle/papers/alkwai-tois17-preprint.pdf
Content Length: 1439568 bytes

URI: http://www.cs.odu.edu/~mmeigle/papers/brunelle-jcd117.pdf
Final URI: https://www.cs.odu.edu/~mmeigle/papers/brunelle-jcd117.pdf
Content Length: 1276346 bytes

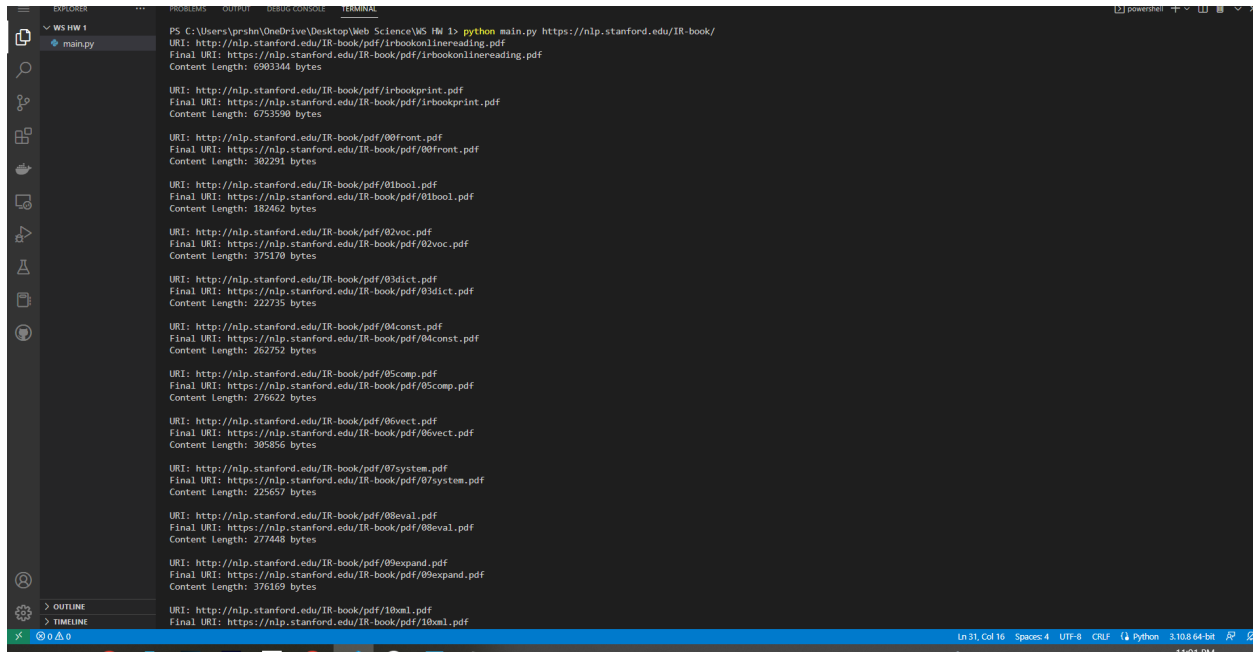
URI: http://www.cs.odu.edu/~mmeigle/papers/alam-jcd117.pdf
Final URI: https://www.cs.odu.edu/~mmeigle/papers/alam-jcd117.pdf
Content Length: 596800 bytes

```

Figure 8: PDF:2

References

In this homework I have referred to python docs for installation of libraries and google developer doc for running the code.



```
PS C:\Users\jprshh\OneDrive\Desktop\Web Science\WS HW 1> python main.py https://nlp.stanford.edu/IR-book/
URI: http://nlp.stanford.edu/IR-book/pdf/lrbookonlinereading.pdf
Final URI: https://nlp.stanford.edu/IR-book/pdf/lrbookonlinereading.pdf
Content Length: 6903344 bytes

URI: http://nlp.stanford.edu/IR-book/pdf/lrbookprint.pdf
Final URI: https://nlp.stanford.edu/IR-book/pdf/lrbookprint.pdf
Content Length: 6753590 bytes

URI: http://nlp.stanford.edu/IR-book/pdf/00front.pdf
Final URI: https://nlp.stanford.edu/IR-book/pdf/00front.pdf
Content Length: 302291 bytes

URI: http://nlp.stanford.edu/IR-book/pdf/01bool.pdf
Final URI: https://nlp.stanford.edu/IR-book/pdf/01bool.pdf
Content Length: 182462 bytes

URI: http://nlp.stanford.edu/IR-book/pdf/02voc.pdf
Final URI: https://nlp.stanford.edu/IR-book/pdf/02voc.pdf
Content Length: 375170 bytes

URI: http://nlp.stanford.edu/IR-book/pdf/03dict.pdf
Final URI: https://nlp.stanford.edu/IR-book/pdf/03dict.pdf
Content Length: 222725 bytes

URI: http://nlp.stanford.edu/IR-book/pdf/04const.pdf
Final URI: https://nlp.stanford.edu/IR-book/pdf/04const.pdf
Content Length: 262752 bytes

URI: http://nlp.stanford.edu/IR-book/pdf/05comp.pdf
Final URI: https://nlp.stanford.edu/IR-book/pdf/05comp.pdf
Content Length: 276622 bytes

URI: http://nlp.stanford.edu/IR-book/pdf/06vect.pdf
Final URI: https://nlp.stanford.edu/IR-book/pdf/06vect.pdf
Content Length: 305856 bytes

URI: http://nlp.stanford.edu/IR-book/pdf/07system.pdf
Final URI: https://nlp.stanford.edu/IR-book/pdf/07system.pdf
Content Length: 225657 bytes

URI: http://nlp.stanford.edu/IR-book/pdf/08eval.pdf
Final URI: https://nlp.stanford.edu/IR-book/pdf/08eval.pdf
Content Length: 277448 bytes

URI: http://nlp.stanford.edu/IR-book/pdf/09expand.pdf
Final URI: https://nlp.stanford.edu/IR-book/pdf/09expand.pdf
Content Length: 376169 bytes

URI: http://nlp.stanford.edu/IR-book/pdf/10xml.pdf
Final URI: https://nlp.stanford.edu/IR-book/pdf/10xml.pdf
```

Figure 9: PDF:3

- Insert Reference 1, <https://docs.python-requests.org/en/v0.8.4/api/>
- Insert Reference 2, <https://developers.google.com/edu/python/regular-expressions>