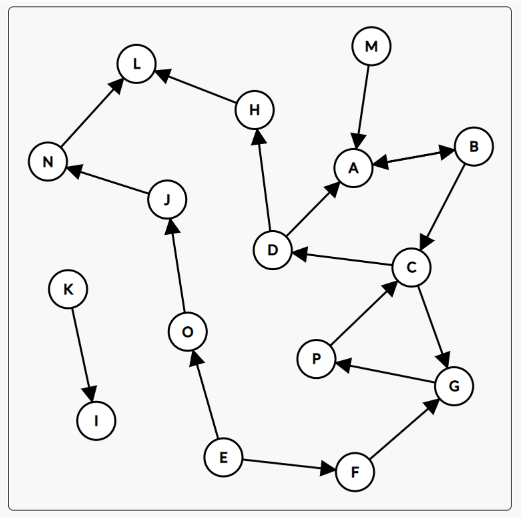
HW1 CS432

Ondra Torkilson

September 20, 2020

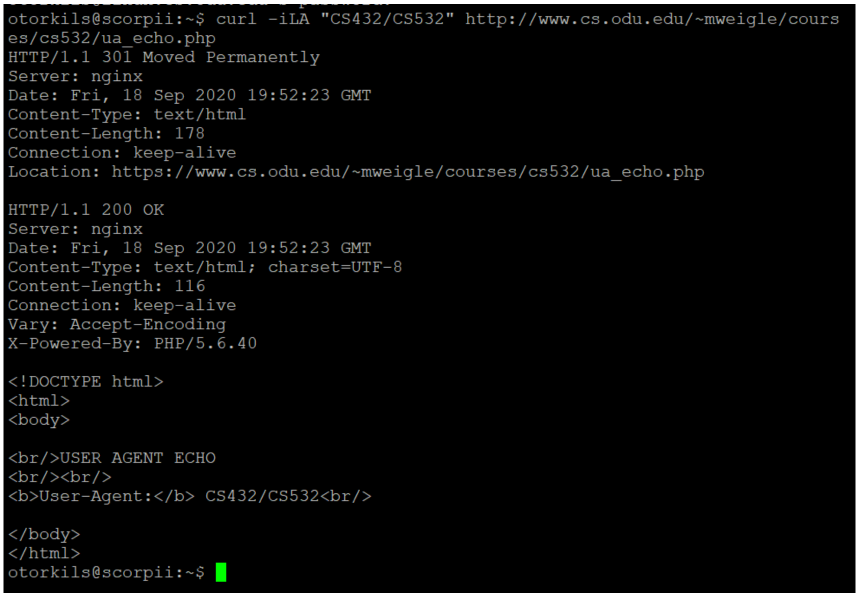
Question 1



Here is the directed graph that I created from the links provided. Within the SCC, each node can be reached by following a directed path from any other node. Meaning, no matter what node you start within the SCC, you can reach any other node within the SCC by following a series of nodes. A disconnected component is not connected to the main graph. The IN component can reach the SCC, but cannot be reached from the SCC, and the OUT component can be reached from the SCC, but cannot link back to it. A TENDRIL can travel away from IN, or into OUT, and a tendril that does both of those things is considered a TUBE from IN to OUT bypassing the SCC.

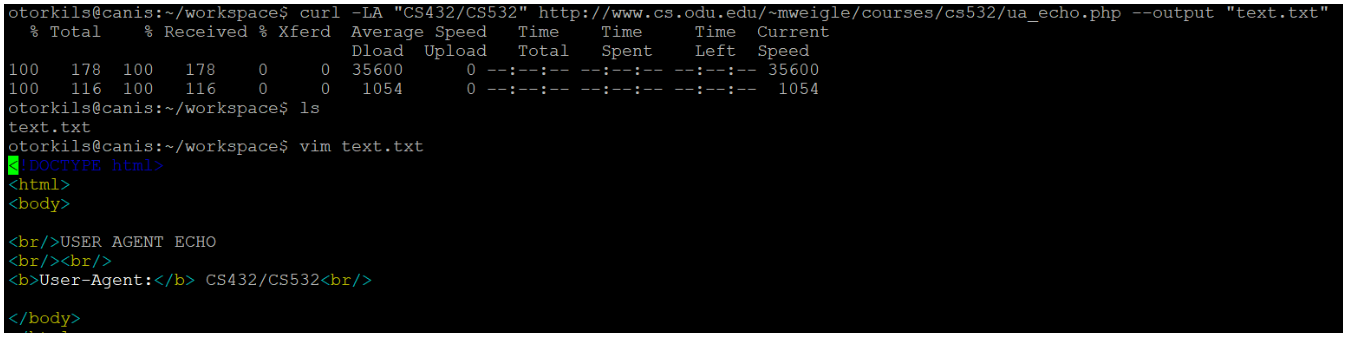
IN: E, F, M
SCC: A, B, C, D, G, and P.
OUT: H, L
TUBE: O, J, N forms a tube from IN to OUT without touching the SCC
DISCONNECTED: I,K

Question 2



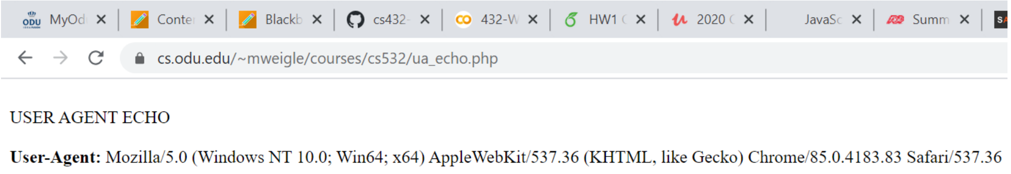
```
otorkils@scorpii:~$ curl -iLA "CS432/CS532" http://www.cs.odu.edu/~mweigle/cours
es/cs532/ua_echo.php
HTTP/1.1 301 Moved Permanently
Server: nginx
Date: Fri, 18 Sep 2020 19:52:23 GMT
Content-Type: text/html
Content-Length: 178
Connection: keep-alive
Location: https://www.cs.odu.edu/~mweigle/courses/cs532/ua_echo.php

HTTP/1.1 200 OK
Server: nginx
Date: Fri, 18 Sep 2020 19:52:23 GMT
Content-Type: text/html; charset=UTF-8
Content-Length: 116
Connection: keep-alive
Vary: Accept-Encoding
X-Powered-By: PHP/5.6.40

<!DOCTYPE html>
<html>
<body>

<br/>USER AGENT ECHO
<br/><br/>
<b>User-Agent:</b> CS432/CS532<br/>

</body>
</html>
otorkils@scorpii:~$
```

In Part A, I used the curl options i,L, and A. 'i' to return the headers and body of the HTML file, 'L' to follow any redirects, and 'A' to change the default User/Agent request to the specified "CS432/CS532" which you can see echoed in the body of the HTML.



```
otorkils@canis:~/workspace$ curl -LA "CS432/CS532" http://www.cs.odu.edu/~mweigle/courses/cs532/ua_echo.php --output "text.txt"
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100   178  100   178    0     0  35600      0 --:--:-- --:--:-- --:--:-- 35600
100   116  100   116    0     0   1054      0 --:--:-- --:--:-- --:--:--  1054
otorkils@canis:~/workspace$ ls
text.txt
otorkils@canis:~/workspace$ vim text.txt
<!DOCTYPE html>
<html>
<body>

<br/>USER AGENT ECHO
<br/><br/>
<b>User-Agent:</b> CS432/CS532<br/>

</body>
```

In Part B, I used the same command from above, minus the 'i' to exclude returning the headers. Then, after the URI, I specified using the option 'output "text.txt" ' for the response to be sent to the text.txt file. Then, using vim, I displayed the text.txt file to prove that my command worked.



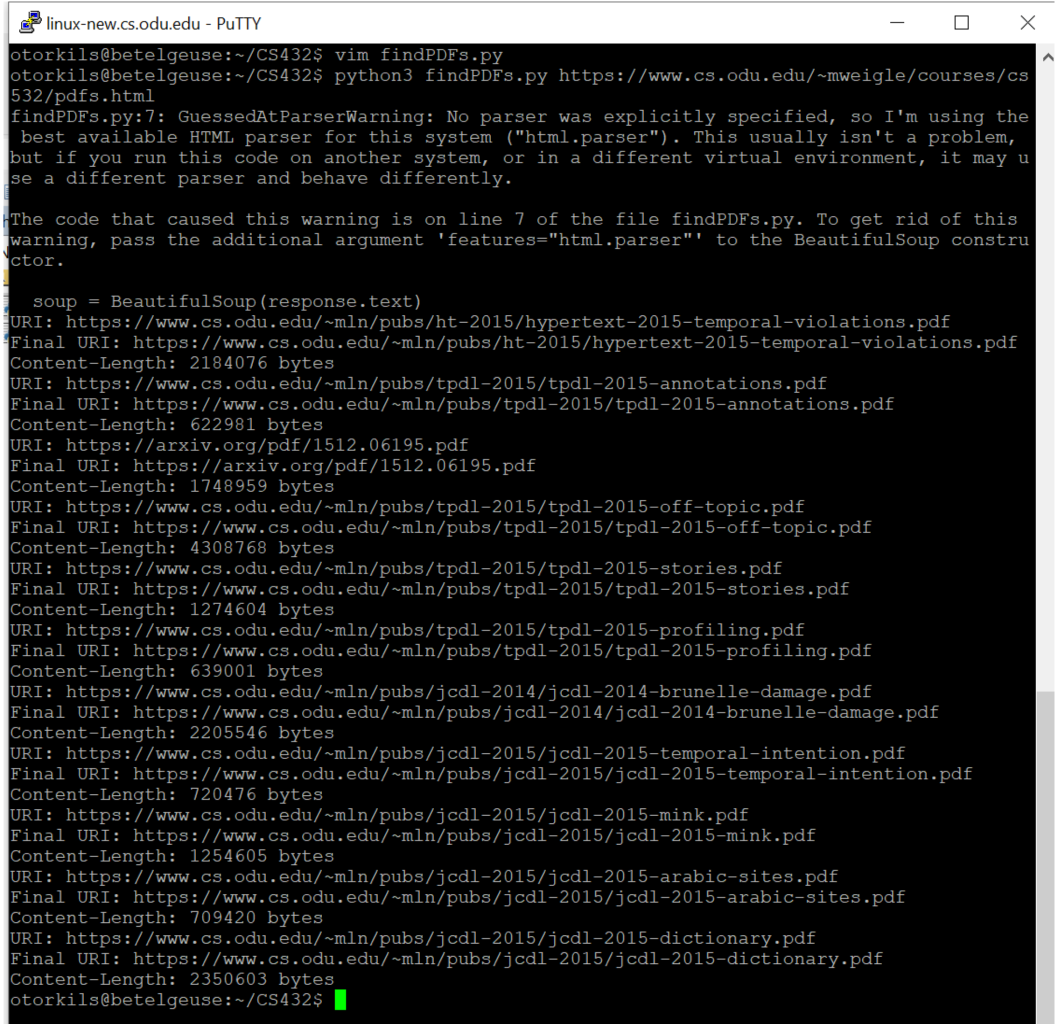
USER AGENT ECHO

User-Agent: Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/85.0.4183.83 Safari/537.36

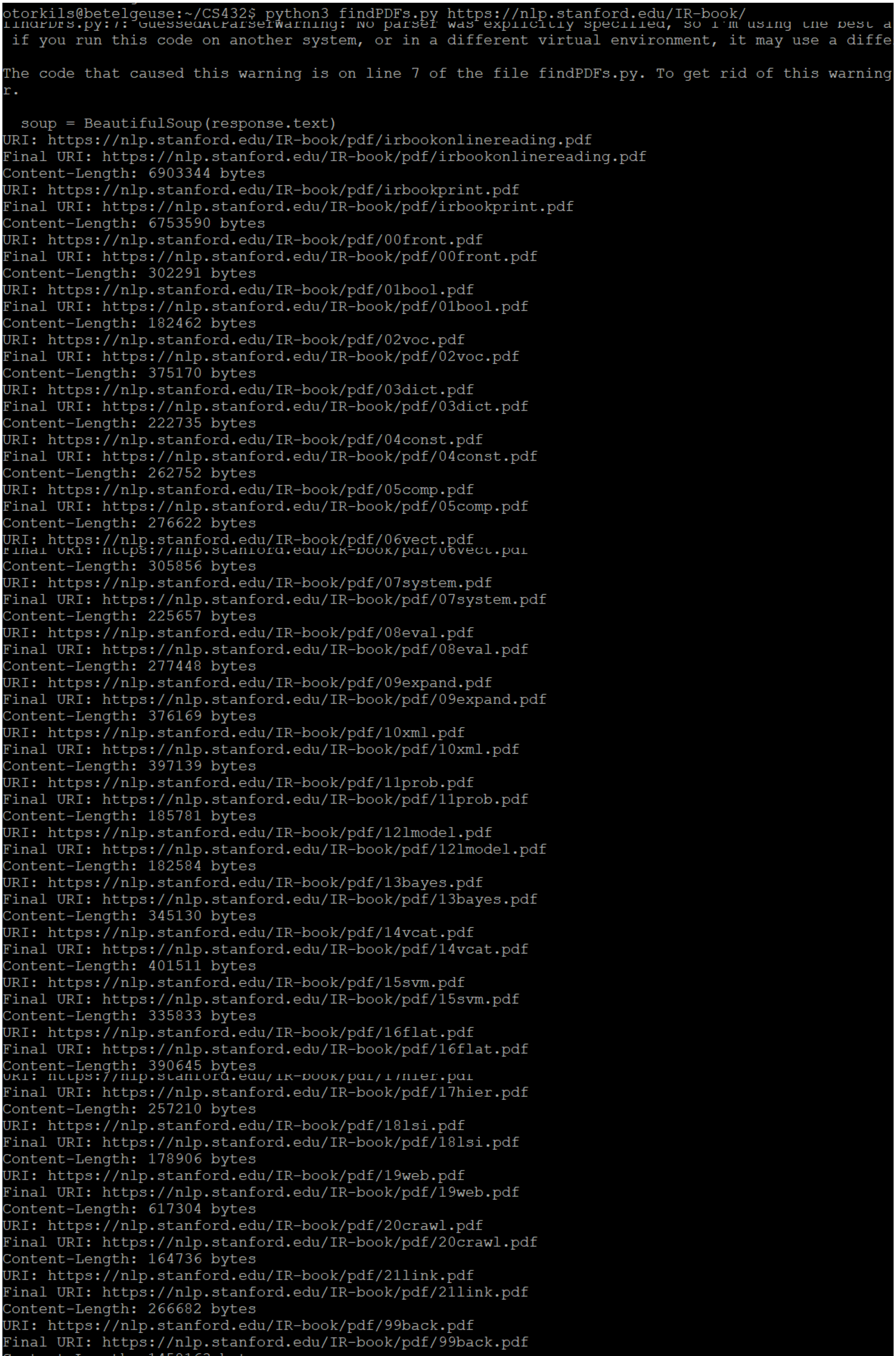
For Part C, I loaded the URI into my chrome browser.

Question 3

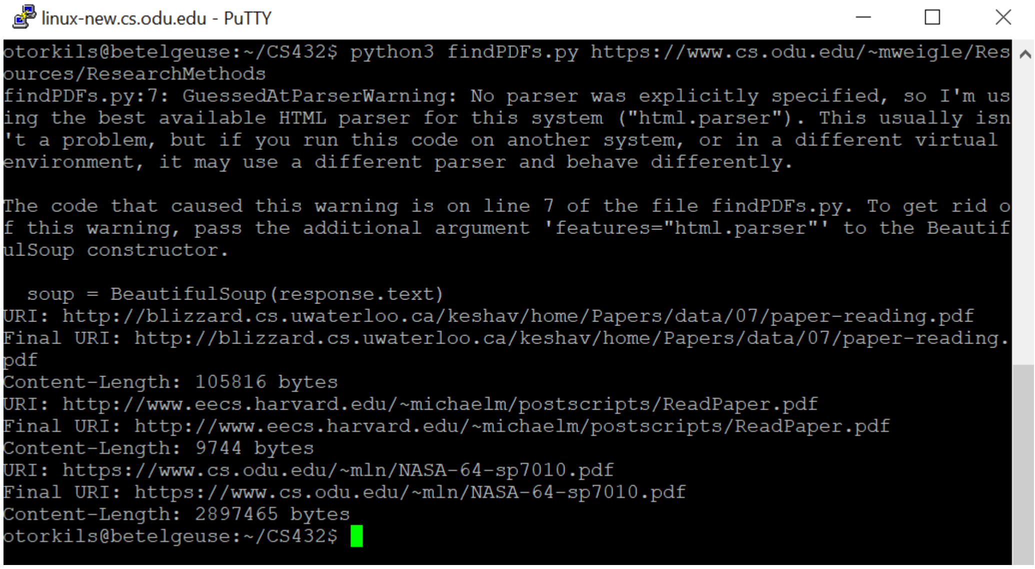
Here are the screenshots of my program findPDFs.py running three different webpages, each that contain a series of PDF links within them.



```
otorkils@betelgeuse:~/CS432$ vim findPDFs.py
otorkils@betelgeuse:~/CS432$ python3 findPDFs.py https://www.cs.odu.edu/~mweigle/courses/cs
532/pdfs.html
findPDFs.py:7: GuessedAtParserWarning: No parser was explicitly specified, so I'm using the
 best available HTML parser for this system ("html.parser"). This usually isn't a problem,
but if you run this code on another system, or in a different virtual environment, it may u
se a different parser and behave differently.

The code that caused this warning is on line 7 of the file findPDFs.py. To get rid of this
warning, pass the additional argument 'features="html.parser"' to the BeautifulSoup constru
ctor.

  soup = BeautifulSoup(response.text)
URI: https://www.cs.odu.edu/~mln/pubs/ht-2015/hypertext-2015-temporal-violations.pdf
Final URI: https://www.cs.odu.edu/~mln/pubs/ht-2015/hypertext-2015-temporal-violations.pdf
Content-Length: 2184076 bytes
URI: https://www.cs.odu.edu/~mln/pubs/tpdl-2015/tpdl-2015-annotations.pdf
Final URI: https://www.cs.odu.edu/~mln/pubs/tpdl-2015/tpdl-2015-annotations.pdf
Content-Length: 622981 bytes
URI: https://arxiv.org/pdf/1512.06195.pdf
Final URI: https://arxiv.org/pdf/1512.06195.pdf
Content-Length: 1748959 bytes
URI: https://www.cs.odu.edu/~mln/pubs/tpdl-2015/tpdl-2015-off-topic.pdf
Final URI: https://www.cs.odu.edu/~mln/pubs/tpdl-2015/tpdl-2015-off-topic.pdf
Content-Length: 4308768 bytes
URI: https://www.cs.odu.edu/~mln/pubs/tpdl-2015/tpdl-2015-stories.pdf
Final URI: https://www.cs.odu.edu/~mln/pubs/tpdl-2015/tpdl-2015-stories.pdf
Content-Length: 1274604 bytes
URI: https://www.cs.odu.edu/~mln/pubs/tpdl-2015/tpdl-2015-profiling.pdf
Final URI: https://www.cs.odu.edu/~mln/pubs/tpdl-2015/tpdl-2015-profiling.pdf
Content-Length: 639001 bytes
URI: https://www.cs.odu.edu/~mln/pubs/jcdl-2014/jcdl-2014-brunelle-damage.pdf
Final URI: https://www.cs.odu.edu/~mln/pubs/jcdl-2014/jcdl-2014-brunelle-damage.pdf
Content-Length: 2205546 bytes
URI: https://www.cs.odu.edu/~mln/pubs/jcdl-2015/jcdl-2015-temporal-intention.pdf
Final URI: https://www.cs.odu.edu/~mln/pubs/jcdl-2015/jcdl-2015-temporal-intention.pdf
Content-Length: 720476 bytes
URI: https://www.cs.odu.edu/~mln/pubs/jcdl-2015/jcdl-2015-mink.pdf
Final URI: https://www.cs.odu.edu/~mln/pubs/jcdl-2015/jcdl-2015-mink.pdf
Content-Length: 1254605 bytes
URI: https://www.cs.odu.edu/~mln/pubs/jcdl-2015/jcdl-2015-arabic-sites.pdf
Final URI: https://www.cs.odu.edu/~mln/pubs/jcdl-2015/jcdl-2015-arabic-sites.pdf
Content-Length: 709420 bytes
URI: https://www.cs.odu.edu/~mln/pubs/jcdl-2015/jcdl-2015-dictionary.pdf
Final URI: https://www.cs.odu.edu/~mln/pubs/jcdl-2015/jcdl-2015-dictionary.pdf
Content-Length: 2350603 bytes
otorkils@betelgeuse:~/CS432$ 
```

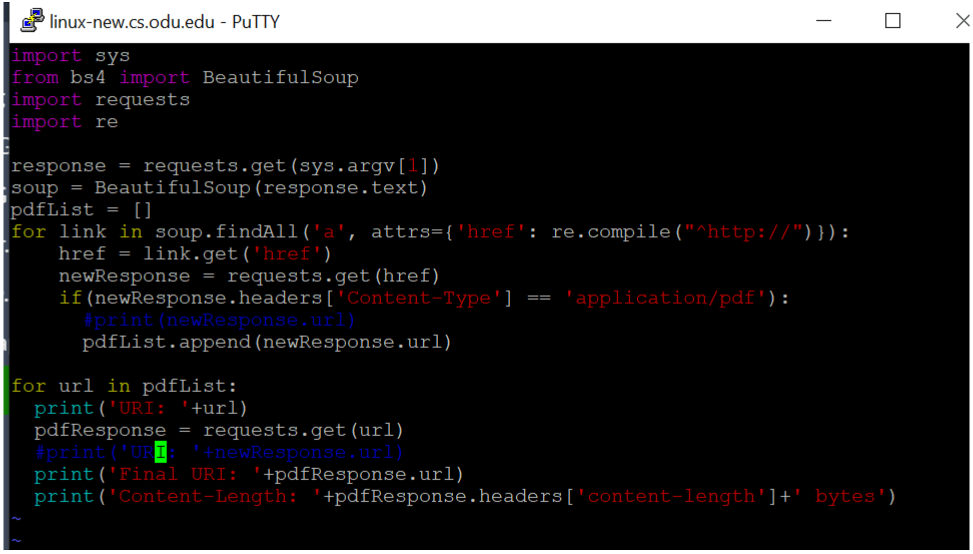


```
otorkils@betelgeuse:~/CS432$ python3 findPDFs.py https://nlp.stanford.edu/IR-book/
findPDFs.py:7: GuessedAtParserWarning: No parser was explicitly specified, so I'm using the best a
 if you run this code on another system, or in a different virtual environment, it may use a diffe

The code that caused this warning is on line 7 of the file findPDFs.py. To get rid of this warning
r.

  soup = BeautifulSoup(response.text)
URI: https://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf
Final URI: https://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf
Content-Length: 6903344 bytes
URI: https://nlp.stanford.edu/IR-book/pdf/irbookprint.pdf
Final URI: https://nlp.stanford.edu/IR-book/pdf/irbookprint.pdf
Content-Length: 6753590 bytes
URI: https://nlp.stanford.edu/IR-book/pdf/00front.pdf
Final URI: https://nlp.stanford.edu/IR-book/pdf/00front.pdf
Content-Length: 302291 bytes
URI: https://nlp.stanford.edu/IR-book/pdf/01bool.pdf
Final URI: https://nlp.stanford.edu/IR-book/pdf/01bool.pdf
Content-Length: 182462 bytes
URI: https://nlp.stanford.edu/IR-book/pdf/02voc.pdf
Final URI: https://nlp.stanford.edu/IR-book/pdf/02voc.pdf
Content-Length: 375170 bytes
URI: https://nlp.stanford.edu/IR-book/pdf/03dict.pdf
Final URI: https://nlp.stanford.edu/IR-book/pdf/03dict.pdf
Content-Length: 222735 bytes
URI: https://nlp.stanford.edu/IR-book/pdf/04const.pdf
Final URI: https://nlp.stanford.edu/IR-book/pdf/04const.pdf
Content-Length: 262752 bytes
URI: https://nlp.stanford.edu/IR-book/pdf/05comp.pdf
Final URI: https://nlp.stanford.edu/IR-book/pdf/05comp.pdf
Content-Length: 276622 bytes
URI: https://nlp.stanford.edu/IR-book/pdf/06vect.pdf
Final URI: https://nlp.stanford.edu/IR-book/pdf/06vect.pdf
Content-Length: 305856 bytes
URI: https://nlp.stanford.edu/IR-book/pdf/07system.pdf
Final URI: https://nlp.stanford.edu/IR-book/pdf/07system.pdf
Content-Length: 225657 bytes
URI: https://nlp.stanford.edu/IR-book/pdf/08eval.pdf
Final URI: https://nlp.stanford.edu/IR-book/pdf/08eval.pdf
Content-Length: 277448 bytes
URI: https://nlp.stanford.edu/IR-book/pdf/09expand.pdf
Final URI: https://nlp.stanford.edu/IR-book/pdf/09expand.pdf
Content-Length: 376169 bytes
URI: https://nlp.stanford.edu/IR-book/pdf/10xml.pdf
Final URI: https://nlp.stanford.edu/IR-book/pdf/10xml.pdf
Content-Length: 397139 bytes
URI: https://nlp.stanford.edu/IR-book/pdf/11prob.pdf
Final URI: https://nlp.stanford.edu/IR-book/pdf/11prob.pdf
Content-Length: 185781 bytes
URI: https://nlp.stanford.edu/IR-book/pdf/12lmodel.pdf
Final URI: https://nlp.stanford.edu/IR-book/pdf/12lmodel.pdf
Content-Length: 182584 bytes
URI: https://nlp.stanford.edu/IR-book/pdf/13bayes.pdf
Final URI: https://nlp.stanford.edu/IR-book/pdf/13bayes.pdf
Content-Length: 345130 bytes
URI: https://nlp.stanford.edu/IR-book/pdf/14vcat.pdf
Final URI: https://nlp.stanford.edu/IR-book/pdf/14vcat.pdf
Content-Length: 401511 bytes
URI: https://nlp.stanford.edu/IR-book/pdf/15svm.pdf
Final URI: https://nlp.stanford.edu/IR-book/pdf/15svm.pdf
Content-Length: 335833 bytes
URI: https://nlp.stanford.edu/IR-book/pdf/16flat.pdf
Final URI: https://nlp.stanford.edu/IR-book/pdf/16flat.pdf
Content-Length: 390645 bytes
URI: https://nlp.stanford.edu/IR-book/pdf/17hier.pdf
Final URI: https://nlp.stanford.edu/IR-book/pdf/17hier.pdf
Content-Length: 257210 bytes
URI: https://nlp.stanford.edu/IR-book/pdf/18lsi.pdf
Final URI: https://nlp.stanford.edu/IR-book/pdf/18lsi.pdf
Content-Length: 178906 bytes
URI: https://nlp.stanford.edu/IR-book/pdf/19web.pdf
Final URI: https://nlp.stanford.edu/IR-book/pdf/19web.pdf
Content-Length: 617304 bytes
URI: https://nlp.stanford.edu/IR-book/pdf/20crawl.pdf
Final URI: https://nlp.stanford.edu/IR-book/pdf/20crawl.pdf
Content-Length: 164736 bytes
URI: https://nlp.stanford.edu/IR-book/pdf/21link.pdf
Final URI: https://nlp.stanford.edu/IR-book/pdf/21link.pdf
Content-Length: 266682 bytes
URI: https://nlp.stanford.edu/IR-book/pdf/99back.pdf
Final URI: https://nlp.stanford.edu/IR-book/pdf/99back.pdf
Content-Length: 1450162 bytes
```



```
otorkils@betelgeuse:~/CS432$ python3 findPDFs.py https://www.cs.odu.edu/~mweigle/Res
ources/ResearchMethods
findPDFs.py:7: GuessedAtParserWarning: No parser was explicitly specified, so I'm us
ing the best available HTML parser for this system ("html.parser"). This usually isn
't a problem, but if you run this code on another system, or in a different virtual
environment, it may use a different parser and behave differently.

The code that caused this warning is on line 7 of the file findPDFs.py. To get rid o
f this warning, pass the additional argument 'features="html.parser"' to the Beautif
ulSoup constructor.

  soup = BeautifulSoup(response.text)
URI: http://blizzard.cs.uwaterloo.ca/keshav/home/Papers/data/07/paper-reading.pdf
Final URI: http://blizzard.cs.uwaterloo.ca/keshav/home/Papers/data/07/paper-reading.
pdf
Content-Length: 105816 bytes
URI: http://www.eecs.harvard.edu/~michaelm/postscripts/ReadPaper.pdf
Final URI: http://www.eecs.harvard.edu/~michaelm/postscripts/ReadPaper.pdf
Content-Length: 9744 bytes
URI: https://www.cs.odu.edu/~mln/NASA-64-sp7010.pdf
Final URI: https://www.cs.odu.edu/~mln/NASA-64-sp7010.pdf
Content-Length: 2897465 bytes
otorkils@betelgeuse:~/CS432$
```

My code is pictured below. The technique used to extract the links from the original webpage was inspired by the code segment in Google Collab authored by M.Weigle under the heading ”BeautifulSoup Library.”



```
import sys
from bs4 import BeautifulSoup
import requests
import re

response = requests.get(sys.argv[1])
soup = BeautifulSoup(response.text)
pdfList = []
for link in soup.findAll('a', attrs={'href': re.compile("^http://")}):
    href = link.get('href')
    newResponse = requests.get(href)
    if(newResponse.headers['Content-Type'] == 'application/pdf'):
      #print(newResponse.url)
      pdfList.append(newResponse.url)

for url in pdfList:
  print('URI: '+url)
  pdfResponse = requests.get(url)
  #print('URI: '+newResponse.url)
  print('Final URI: '+pdfResponse.url)
  print('Content-Length: '+pdfResponse.headers['content-length']+' bytes')
~
~
```