

INTRODUCTION TO WEBSCIENCES: Assignment-1

Babitha Bokka

11 September 2014

Contents

1	Question 1	2
1.1	Approach Towards the Solution	2
1.1.1	Test 1	2
1.1.2	Test 2	2
1.2	Source Code	2
1.2.1	webTest.php	2
1.2.2	sucessPage.php	3
1.3	Solution	3
1.3.1	POST method	3
1.3.2	GET method	3
1.4	Output	4
1.4.1	response.htm	4
1.4.2	gresponse.htm	4
1.5	Screen Shots	12
2	Question 2	13
2.1	Description	13
2.2	Approach Towards the Solution	13
2.3	Python Code	13
2.3.1	A1.py	13
2.4	Output	16
3	Question 3	17
3.1	Problem Description:	17
3.2	Graph:	17
3.2.1	Diagram	17
3.3	Values of :	17
3.4	Definition	18
3.5	How are the nodes forming SCC, IN....?	18

1 Question 1

Demonstrate how to POST data to a webpage by using "curl" ,the server should take the arguments posted and generate a response. Take the response:

1. Save it into a file
2. Open the file in a browser and take a screen shot.

1.1 Approach Towards the Solution

"curl" is a command line tool to communicate with web servers.

1.1.1 Test 1

To demonstrate how "curl" can communicate with the webserver I have created a webpage which has to fields first and last name when you post the arguments by using the below command the server would take your commands and generate a response I saved the response in a file.

1.1.2 Test 2

The other way i tested the curl command is by posting the arguments(MIDAS ID:— ;Password:—) and got the response from the server.And if you observe the gresponse.htm there is an error message which explains server sends 200 OK but doesn't like to communicate with the curl command.

1.2 Source Code

1.2.1 webTest.php

```
<?php
?>
<html>
    <head>
        <meta http-equiv="content-type" content="text/html;
            charset=utf-8" />
        <title>Courses for Fall 2014</title>
        <link href="default.css" rel="stylesheet" type="text/css"
            />
    </head>
    <body>
        <div id="header">
            <h1>Babitha Bokka</h1>
        </div>
        <div id="content">
            <h2>Intro to web sciences</h2>
            <form action="sucessPage.php" method="get">
                First: <input type="text" name="fname"><
                    br>
```

```

Last: <input type="text" name="lname"><br
>
<input type="submit" name="submit" value
="ok">
</form>
</div>
</body>
</html>

```

1.2.2 sucessPage.php

```

<?php
?>
<html>
  <head>
    <meta http-equiv="content-type" content="text/html;
      charset=utf-8" />
    <title>Courses for Fall 2014</title>
    <link href="default.css" rel="stylesheet" type="text/css"
      />
  </head>
  <body>
    <div id="header">
      <h1>Babitha Bokka</h1>
    </div>
    <div id="content">
      <h1>Intro to web sciences</h1>
      <h2> Hello your are successfully Logged In .</h2>
    </div>
  </body>
</html>

```

1.3 Solution

1.3.1 POST method

Test 1:

```

curl -i -A "Mozilla/4.0" --data-urlencode "fname=babitha&lname=bokka&submit=ok"
http://www.cs.odu.edu/~bbokka/curlTest/sucessPage.php/ > response.htm

```

Test 2:

```

curl -i -A "Mozilla/4.0" --data-urlencode "j_username=bbokk001&j_password=***&submit=Login"

```

1.3.2 GET method

```

curl -A "Mozilla/4.0" "http://www.cs.odu.edu/~bbokka/curlTest/sucessPage.php?fname=babitha &

```

1.4 Output

1.4.1 response.htm

HTTP/1.1 200 OK

Server: nginx

Date: Sun, 07 Sep 2014 14:45:07 GMT

Content-Type: text/html

Transfer-Encoding: chunked

Connection: keep-alive

Vary: Accept-Encoding

```
<html>
  <head>
    <meta http-equiv="content-type" content="text/html;
      charset=utf-8" />
    <title>Courses for Fall 2014</title>
    <link href="default.css" rel="stylesheet" type="text/css"
      />
  </head>
  <body>
    <div id="header">
      <h1>Babitha Bokka</h1>
    </div>
    <div id="content">
      <h1>Intro to web sciences</h1>
      <h2> Hello your are successfully Logged In .</h2>
    </div>
  </body>
</html>
```

1.4.2 gresponse.htm

HTTP/1.1 200 OK

Set-Cookie: ODU_SHIBPROD_COOKIE=R3455529390; path=/; expires=Fri, 12-Sep-2014 14:10:36 GMT

Date: Thu, 11 Sep 2014 13:50:38 GMT

Set-Cookie: JSESSIONID=bAuFCZIMv8CDfv5ZNOwz.179; Path=/idp; Secure
Expires: 0

Cache-Control: no-cache, no-store, must-revalidate, max-age=0

Pragma: no-cache

X-FRAME-OPTIONS: DENY

Content-Type: text/html; charset=UTF-8

Transfer-Encoding: chunked

```
<!DOCTYPE HTML>
<!--[if lt IE 7]> <html lang="en-us" class="no-js ie6"> <![endif]-->
<!--[if IE 7]> <html lang="en-us" class="no-js ie7"> <![endif]-->
<!--[if IE 8]> <html lang="en-us" class="no-js ie8"> <![endif]-->
```

```

<!--[if IE 9]> <html lang="en-us" class="no-js ie9"> <![endif]-->
<!--[if gt IE 9]><!--> <html lang="en-us" class="no-js"> <!--<![endif]-->
<head>
    <meta charset="utf-8">
    <meta name="viewport" content="width=device-width, initial-scale
        =1.0">
    <meta name="keywords" content="">
    <meta name="description" content="">
    <script type="text/javascript" src="//ajax.googleapis.com/ajax/libs/
        jquery/1.8.1/jquery.min.js"></script>
<script type="text/javascript">
if (typeof jQuery == 'undefined') {
    document.write(unescape("%3Cscript src='//www.odu.edu/etc/designs/odu
        /clientlibs/libs/jquery-1.8.1.min.js' type='text/javascript' %3E%3C
        /script%3E"));
}
</script>

<link rel="stylesheet" href="//www.odu.edu/etc/designs/odu/clientlibs.css
    " type="text/css">
<script src="//www.odu.edu/etc/designs/odu/mobile.js"></script>

    <!--[if lt IE 9]>
        <script src="//html5shiv.googlecode.com/svn/trunk/html5.js"></
            script>
    <![endif]-->

    <link rel="icon" type="image/vnd.microsoft.icon" href="http://www.odu
        .edu/www.odu.edu/etc/designs/odu/favicon.ico">
    <link rel="shortcut icon" type="image/vnd.microsoft.icon" href="http
        ://www.odu.edu/www.odu.edu/etc/designs/odu/favicon.ico">

    <link rel="stylesheet" href="/css/shib.css" type="text/css">

<title>Monarch-Key Web Login - Old Dominion University</title>
</head>
<body class="page contentpage "
>
    <p class="accessible">[ <a href="#content">skip to content</a> ]</p>
    <input type="checkbox" id="uni-phone-menu-jump-toggle" />
    <div id="page-container">
        <header id="uni-header">
    <hgroup id="uni-logo">

```

```

<h1>
  <a href="http://www.odu.edu/">Old Dominion University</a></h1>
  >
  <h2>Idea Fusion</h2>
</hgroup>
<nav id="uni-find" title="Find People, Dates, and Information">
  <ul>
    <li>
      <a href="http://www.odu.edu/a-to-z">A to Z</a></li>
    <li>
      <a href="http://www.odu.edu/directory">Directories</a></li>
    <li>
      <a href="http://www.odu.edu/library">Libraries</a></li>
    <li>
      <a href="http://www.odu.edu/calendar">Calendars</a></li>
  </ul>
  <form action="http://www.odu.edu/search" method="get" id="uni-
    search">
    <label for="uni-search-bar" class="accessible">
      Search ODU
    </label>
    <input id="uni-search-bar" type="text" size="15" maxlength
      ="35" name="q" value="Search ODU" onfocus="if (value=='
        Search ODU') { value='';}" onblur="if (value=='') { value='
          Search ODU';}" >
    <input type="hidden" name="cx" value="006647748818914228700:
      idbduvl7ohw">
    <input type="hidden" name="cof" value="FORID:11" >
    <input type="hidden" name="ie" value="UTF-8">
    <button type="submit">
      Search
    </button>
  </form>
</nav>
<nav id="uni-nav" title="University Navigation">
  <ul>
    <li>
      <a href="http://www.odu.edu/about">About ODU</a>
    </li>
    <li>
      <a href="http://www.odu.edu/academics">Academics</a>
    </li>
    <li>
      <a href="http://www.odu.edu/life">University Life</a>
    </li>
    <li>

```

```

        <a href="http://www.odu.edu/admission">Admission & Aid
        </a>
    </li>
    <li>
        <a href="http://www.odu.edu/research">Research & Impact
        </a>
    </li>
    <li>
        <a href="http://www.odusports.com/">Athletics </a>
    </li>
</ul>
<label for="uni-phone-menu-jump-toggle" id="uni-phone-menu-jump">
    Find , Search , and Navigate</label>
<a id="uni-phone-menu-jump-anchor" href="#uni-phone-menu">Find ,
    Search , and Navigate</a>
</nav>
</header>

```

```

<section class="theme-1 second">

```

```

    <header style="background-image: url(//www.odu.edu/cq/external/shib/
        _jcr_content/headerimage.img.990.jpg)"><h1>Monarch-Key Web Login</
        h1>

```

```

</header>
<div id="content">

```

```

    <div class="container-16 clearfix">
        <div class="grid-10 page-content col-border">

```

```

<div class="loginContent">
    <div class="loginBox">
        <section>

            <hgroup>
                <h3>ERROR</h3>

            <p>
                An error occurred while processing your request. Please
                contact OCCS Help at <a href="https://fp.odu.edu">https
                ://fp.odu.edu</a> or
                <a href="mailto:occs-help@odu.edu">occs-help@odu.edu</a>.
            </p>
        <p>

```


Use of your browser's back button may cause specific errors that can be resolved by going back to your desired resource and trying to login again.

</hgroup>

</section>

</div>

</div>

</div>

<div class="grid-6 menu-1">

<section class="layout-2">

<header class="overlay"><h1>Integrated Services</h1></header>

Blackboard

CareerLink - Student

Center for Learning and Teaching

Distance Learning - Class Access and Archives

Google Mail

LeoOnline

myODU Portal

PAPERS

Virtual Library of Virginia - Public Broadcasting System Videos

</section>

<section class="layout-2">

<header class="overlay"><h1>Support</h1></header>

<p>Need Help? Please specify the service you are having problems with, and that you are trying to authenticate to Preproduction Monarch-Key Web Login.</p>

<footer>

```

    <p><a href="http://odu.edu/ts/helpdesk">ITS Help Desk</a></p>
  </footer>
</section>
<div class="logoFooter">
</div>

    </div>
  </div>
</div>
</section>

<footer id="uni-footer">
  <div class="container-16">
    <address class="grid-4">
      <span id="uni-footer-name">
        Old Dominion University
      </span>
      <br />Norfolk , VA 23529
      <br /><i class="footer-icon-mail"></i> <a href="http://www.
        odu.edu/about/contact">Contact & Mailing Info</a>
      <br /><i class="footer-icon-globe"></i> <a href="http://www.
        odu.edu/about/visitors/directions">Directions to Campus</a>
    </address>
    <nav id="uni-quicklinks" class="grid-4">
      <h1>Quick Links</h1>
      <ul>
        <li>
          <a href="http://www.odu.edu/admission">Apply Now</a></li>
        <li>
          <a href="http://www.odu.edu/about/support-odu">Support
            ODU</a></li>
        <li>
          <a href="http://www.odu.edu/about/visitors">Visitor 's
            Guide</a></li>
      </ul>
      <ul>
        <li>
          <a href="http://www.odu.edu/employment">Employment</a></
            li>
        <li>
          <a href="http://www.odu.edu/media">News & Media</a></
            li>
        <li>

```

```

        <a href="http://www.odu.edu/life/health-safety/alerts">
            Campus Alerts</a></li>
    </ul>
</nav>
<div id="uni-connect" class="grid-4">

    <p>
        <a href="http://www.odu.edu/link/c/connect">Connect with
            Old Dominion University</a>
    </p>
    <ul class="clearfix">
        <li class="facebook">
            <a href="http://www.facebook.com/Old.Dominion.
                University">Facebook</a></li>
        <li class="twitter">
            <a href="https://twitter.com/ODUnow">Twitter</a></li>
        <li class="pinterest">
            <a href="http://pinterest.com/OldDominion/">Pinterest
                </a></li>
        <li class="googleplus">
            <a href="https://plus.google.com/+
                olldominionuniversity">Google+</a></li>
        <li class="youtube">
            <a href="http://www.youtube.com/odu">YouTube</a></li>
        <li class="itunes">
            <a href="http://itunes.odu.edu/">iTunes</a></li>
    </ul>
</div>
<div id="uni-copyright" class="grid-3 prefix-1">
    <p>
        Copyright &copy; 2014
    </p>
    <p>
        Old Dominion University
    </p>
    <p>
        Updated: 5/22/2013
    </p>
    <p>
        <a href="http://www.odu.edu/about/contact">Contact Us</a>
        &bull;
        <a href="http://www.odu.edu/privacy">Privacy</a>
    </p>
</div>
</div>
</footer>
</div>

```

```

<div id="uni-phone-menu">
<nav id="uni-phone-nav" title="University Navigation">
  <ul>
    <li>
      <a href="http://www.odu.edu/about">About ODU</a>
    </li>
    <li>
      <a href="http://www.odu.edu/academics">Academics</a>
    </li>
    <li>
      <a href="http://www.odu.edu/life">University Life</a>
    </li>
    <li>
      <a href="http://www.odu.edu/admission">Admission & Aid</a>
    </li>
    <li>
      <a href="http://www.odu.edu/research">Research & Impact</a>
    </li>
    <li>
      <a href="http://www.odusports.com">Athletics</a>
    </li>
  </ul>
</nav>
<nav id="uni-phone-find" title="Find People, Dates, and Information">
  <ul>
    <li>
      <a href="http://www.odu.edu/a-to-z">A to Z</a></li>
    <li>
      <a href="http://www.odu.edu/directory">Directories</a></li>
    <li>
      <a href="http://www.odu.edu/library">Libraries</a></li>
    <li>
      <a href="http://www.odu.edu/calendar">Calendars</a></li>
  </ul>
  <form action="http://www.odu.edu/search" method="get" id="uni-
  phone-search" class="clearfix">
    <label for="uni-phone-search-bar" class="accessible">
      Search ODU
    </label>
    <input id="uni-phone-search-bar" type="text" size="15"
      maxlength="35" name="q" value="Search ODU" onfocus="if(
      value=='Search ODU'){ value='';}" onblur="if(value==''){
      value='Search ODU';}">
    <button type="submit">
      Search

```

```

        </button>
    </form>
</nav>
</div>
</body>
</html>

```

1.5 Screen Shots

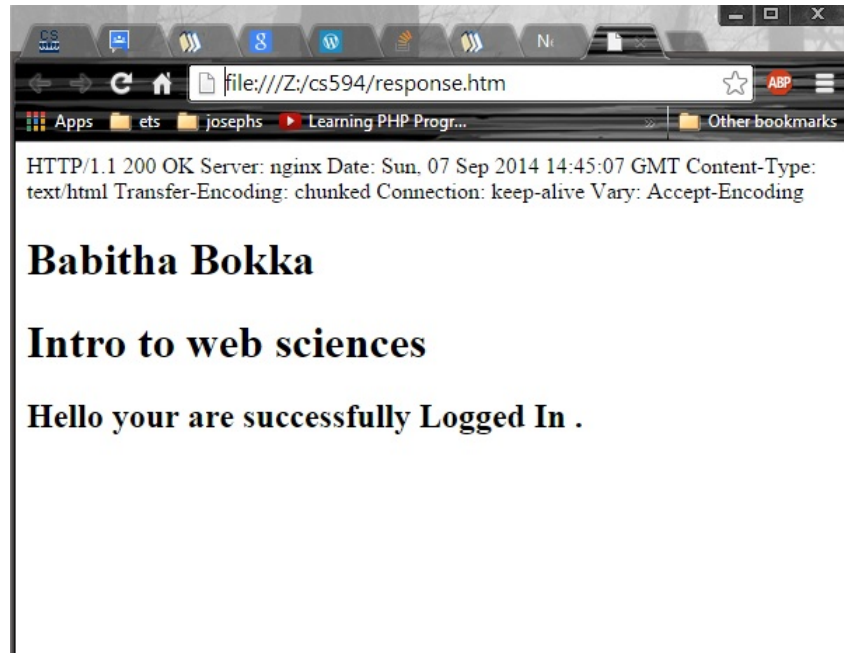


Figure 1: Test 1 Screen shot

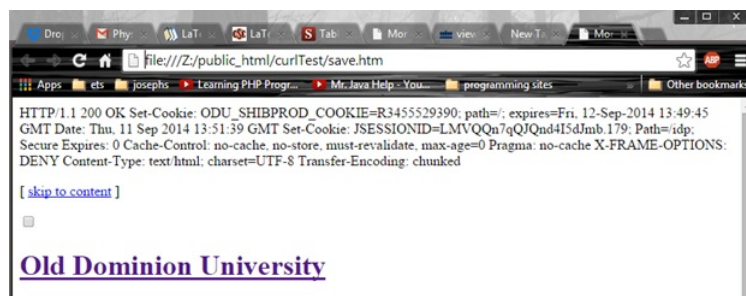


Figure 2: Test 2 Screen shot

2 Question 2

2.1 Description

Write a python program which takes the three arguments like

- 1.Team name
- 2.Time to Sleep
- 3.URI

and get the scores of the respective team and the opposite team waits for the time and gets the scores untill you hit ctrl+c.

2.2 Approach Towards the Solution

The aim of the program is to scrape sports webpages and get the scores of the given team. In order to scrape the webpages we have a beautiful library called BeautifulSoup which has many functions in it (contents, findChildren, findAll etc...). The below python program uses Requests and BeautifulSoup Library to get the contents of the webpage and extract the required data from the html. And to keep on updating the scores we halt the program by using time.sleep() function for the specified time. The below program gets scores from the respective URI's

<http://sports.yahoo.com/college-football/scoreboard/?week=2&conf=all>

<http://sports.yahoo.com/college-football/scoreboard/?week=1&conf=72>

2.3 Python Code

2.3.1 A1.py

```
#!/usr/bin/env python

import sys
import time
import requests
from bs4 import BeautifulSoup

#Main Function
def main():

    numOfArgs=len(sys.argv)

    if numOfArgs<4 or numOfArgs>4:

        print 'Usage: A1.py <university> <sec> < URI>'
        print 'e.g.: A1.py "old dominion" 60 http://sports.yahoo.com'
        sys.exit(1)

    print 'Number of arguments:', len(sys.argv), 'arguments.'
    univ = str(sys.argv[1])
    sec = int(sys.argv[2])
    uri = str(sys.argv[3])
```

```

print 'Team Name: ',univ
print 'Time to Sleep: ',sec
print 'URI: ',uri

response = requests.get(uri)
soup = BeautifulSoup(response.content)#gives you the html content of
    that page
tables = soup.findChildren('table')#finds all the children of type
    table
print "-" * 72
#print tables[1].prettify()
score_table = tables[1]#storing the results of the second table in
    score_table variable as our intersting stuff is in table
# when you extract data from web and use beautiful soup it is stored
    in the form of array nothing but list in python

while True:

    for row in score_table('tr', {'class' : 'game link' }):

        if univ.lower() in str(row).lower() :
            td_team_home = row('td', {'class' : 'home' })
            span_home     = td_team_home[0]('em')[0].contents[0]#the
                td_team_home is treated as a list so you have to get
                the contents of it

            td_team_away = row('td', {'class' : 'away' })
            span_away     = td_team_away[0]('em')[0].contents[0]

            td_score      = row('td', {'class' : 'score' })
            span_home_score = td_score[0]('span')[1].contents[0]
            span_away_score = td_score[0]('span')[0].contents[0]

            print "*" * 8
            print span_home
            print span_home_score
            print

            print span_away
            print span_away_score
            print

            print 'Press ctrl+c to exit getting the scores'

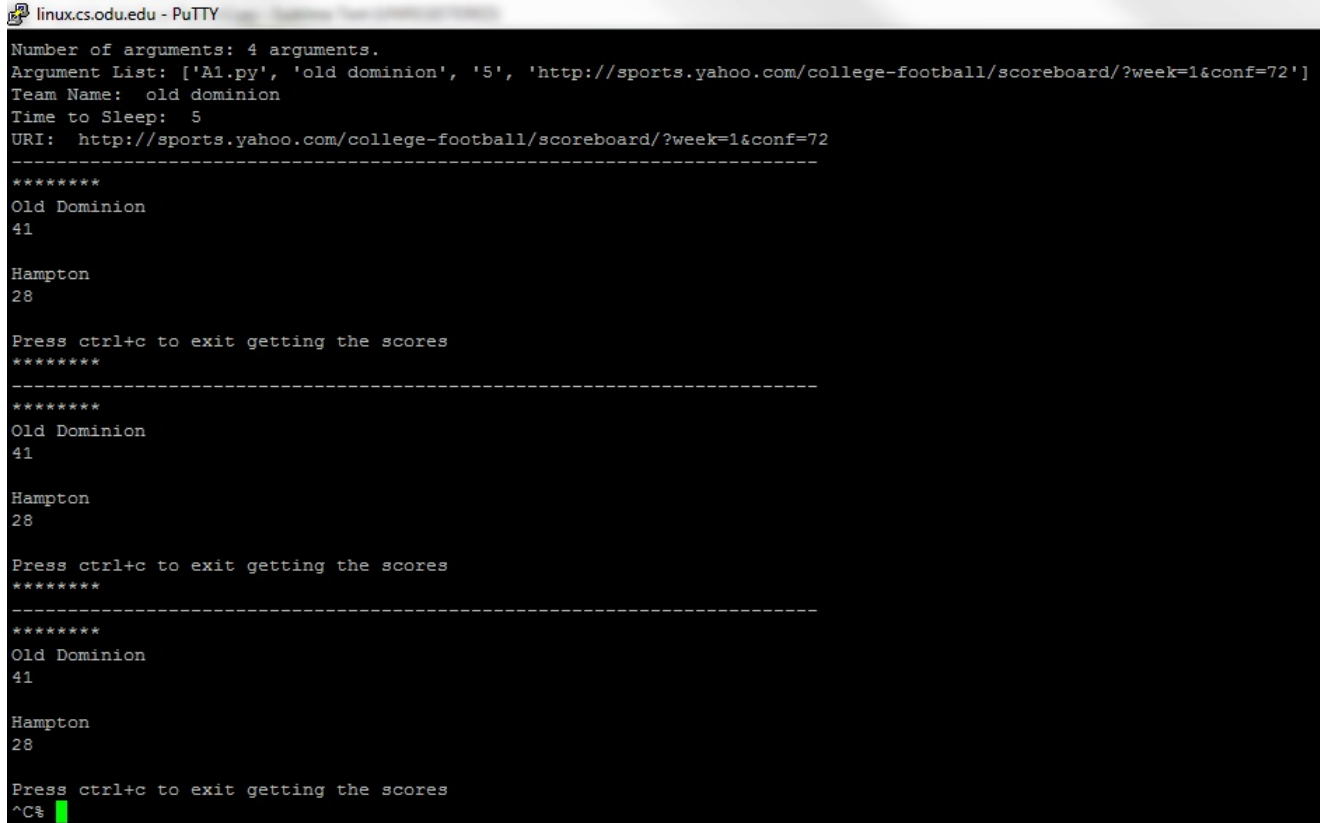
            time.sleep(sec) # delays for 60 seconds
            print "*" * 8

```

```
        print "-" * 72

if __name__ == "__main__":
    try:
        main()
    except KeyboardInterrupt:
        sys.exit(1)
```


2.4 Output



The screenshot shows a PuTTY terminal window with the title 'linux.cs.odu.edu - PuTTY'. The output of a Python program is displayed in white text on a black background. The program prints the number of arguments (4), the argument list, the team name ('old dominion'), and the time to sleep (5). It then prints the URI and a separator line. The program then prints the scores for 'Old Dominion' (41) and 'Hampton' (28). This sequence is repeated three times, with a prompt to press Ctrl+C to exit after each set of scores. The final prompt is followed by the characters '^C' and a green cursor.

```
linux.cs.odu.edu - PuTTY
Number of arguments: 4 arguments.
Argument List: ['A1.py', 'old dominion', '5', 'http://sports.yahoo.com/college-football/scoreboard/?week=1&conf=72']
Team Name: old dominion
Time to Sleep: 5
URI: http://sports.yahoo.com/college-football/scoreboard/?week=1&conf=72
-----
*****
Old Dominion
41

Hampton
28

Press ctrl+c to exit getting the scores
*****
-----
*****
Old Dominion
41

Hampton
28

Press ctrl+c to exit getting the scores
*****
-----
*****
Old Dominion
41

Hampton
28

Press ctrl+c to exit getting the scores
^C█
```

Figure 3: Python Program Output

3 Question 3

3.1 Problem Description:

Given a graph determine the values of
IN
SCC
OUT
Tendrils
Tubes
Disconnected

3.2 Graph:

3.2.1 Diagram

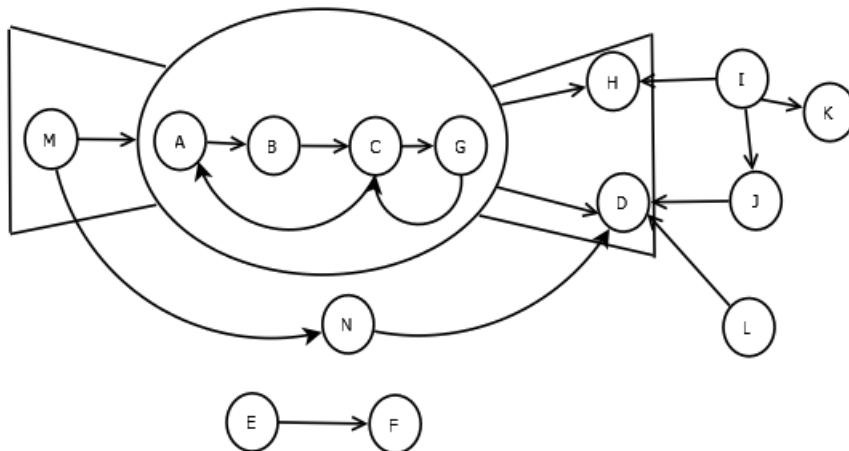


Figure 4: Given graph in bow tie structure of web

3.3 Values of :

IN:1
SCC:4
OUT:2
Tendrils:4
Tubes:1
Disconnected:2

3.4 Definition

From the broders paper and definition below :

Definition:

SCC: We say that a strongly connected component (SCC) in a directed graph is a subset of the nodes such that: (i) every node in the subset has a path to every other; and (ii) the subset is not part of some larger set with the property that every node can reach every other

IN: Nodes that can reach the giant SCC but cannot be reached from it i.e., nodes that are upstream of it.

OUT: Nodes that can be reached from the giant SCC but cannot reach it i.e., nodes are downstream of it.

Tendrils: The tendrils of the bow-tie consist of (a) the nodes reachable from IN that cannot reach the giant SCC, and (b) the nodes that can reach OUT but cannot be reached from the giant SCC. For example, the page My song lyrics in Figure 13.6 is an example of a tendril page, since its reachable from IN but has no path to the giant SCC. Its possible for a tendril node to satisfy both (a) and (b), in which case its part of a TUBE that travels from IN to OUT without touching the giant SCC. (For example, if the page My song lyrics happened to link to Blog post about Company Z in Figure 13.6, it would be part of a tube.)

Disconnected: Finally, there are nodes that would not have a path to the giant SCC even if we completely ignored the directions of the edges. These belong to none of the preceding categories.

3.5 How are the nodes forming SCC, IN....?

From the above definition SCC , has the strongly connected nodes if we apply this definition to the graph given that A , B, C, G are the strongly connected components where each and every node can be reached by every other node , the node M is the IN since that is only node which can reach the SCC and node H,D falls under OUT beacsue they are the only nodes that can be reached by SCC . The node N connects both IN and OUT so it forms the TUBE .Nodes E and F are not connected to SCC in any way so they are DISCONNECTED. Nodes I ,J,K,L are the OUT tendrils.

References

- [1] BeautifulSoup documentation.
- [2] Bibliography management with bibtex.
- [3] Extracting data in table using beautifulsoup.
- [4] Latex/floating, figures and captions.
- [5] The structure of the web.

□