**CS 432/532 Web Science: Assignment #2**

Alexander C. Nwala

Bethany DeMerchant

16 February 2019

**Contents**

1. [Problem 1](#_Problem_1:)
   1. [Solution 1](#_Solution_1:)
2. [Problem 2](#_Problem_2:)
   1. [Solution 2](#_Solution_2:)
3. [Problem 3](#_Problem_3:)
   1. [Solution 3](#_Solution_3:)

# Background:

The Friendship Paradox is a social networking phenomenon which says that an average individual’s friends have more friends than the individual. This report demonstrates simple tests of this phenomenon performed using one account each on Facebook, Twitter, and LinkedIn:

Determine if the friendship paradox holds for [a given account]. Compute the mean, standard deviation, and median of the number of friends that… friends have. Create a graph of the number of friends (y-axis) and the friends (x-axis) themselves, sorted by number of friends (y-axis).

Horizontal

# Background (ii):

So I noticed a couple really, really big numbers and got curious about what would happen to the data if I cut them off. Since I had the mean and standard deviation in front of my face I ended up taking z-scores. Then I threw out all the data that was more than two standard deviations from the mean and did the whole stats panel over again. Then I figured since I had the data I may as well throw it in the report.

# Facebook:

Facebook requires a user to approve collection of information regarding their friend list, preventing the collection of live data. Instead, analysis was performed on data provided in a file acnwala-friendscount.csv.

# Analysis—Facebook:

Data data data! Data science. Yay.

# Analysis—Facebook (ii):

# Problem 2:

Write a Python program that (1) takes as a command line argument a web page, (2) extracts all the links from the page, (3) lists all the links that result in PDF files and prints out the bytes for each of the links. (4) Show that the program works on 3 different URIs, one of which needs to be:

http://www.cs.odu.edu/~mln/teaching/cs532-s17/test/pdfs.html

# Solution 2:

The solution was executed in Python using external modules Requests (http://docs.python-requests.org/en/master/) and BeautifulSoup (https://www.crummy.com/software/BeautifulSoup/) as well as built-in modules argparse and urllib.parse.

The argparse module was used to generate command-line interface and ensure proper functioning if a user fails to provide a URI argument (see Figure 2).

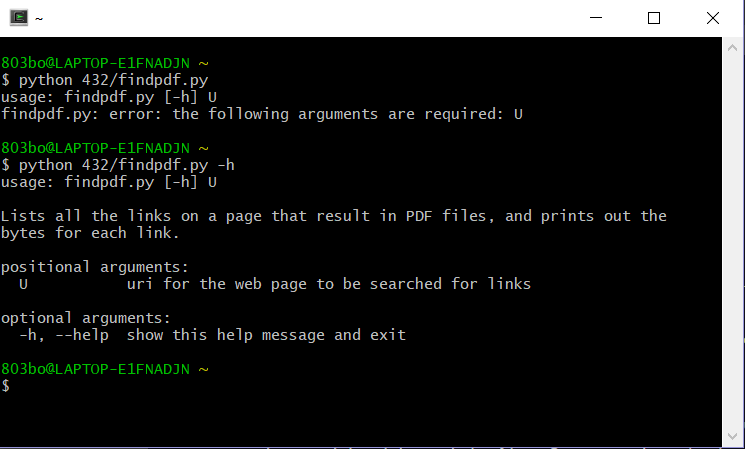


Figure 2: Command-Line Interface Using argparse

The URI provided as an argument is dereferenced using the Requests module, which performs a GET request and creates a response object containing the page contents. Assuming the page contents exist, they are then parsed by BeautifulSoup, which scans for the <a> tag which represents a hyperlink in HTML. For each <a> tag found, the URI address of the link is added to a list of all links in the page. Each address in the complete list is then checked with a HEAD request performed using the Requests module. If the header’s Content-Type header is “application/pdf”, the page’s URI and the value of the Content-Length header are added to a list of links resulting in PDF files. Finally, the finished list of links to PDF files is printed to the console.

The demonstration page (http://www.cs.odu.edu/~mln/teaching/cs532-s17/test/pdfs.html) contains twenty links, eleven of which lead to PDF files. Figure 3 shows that the program reported all eleven PDF files and no links to non-PDF pages, and also that it successfully followed redirects.

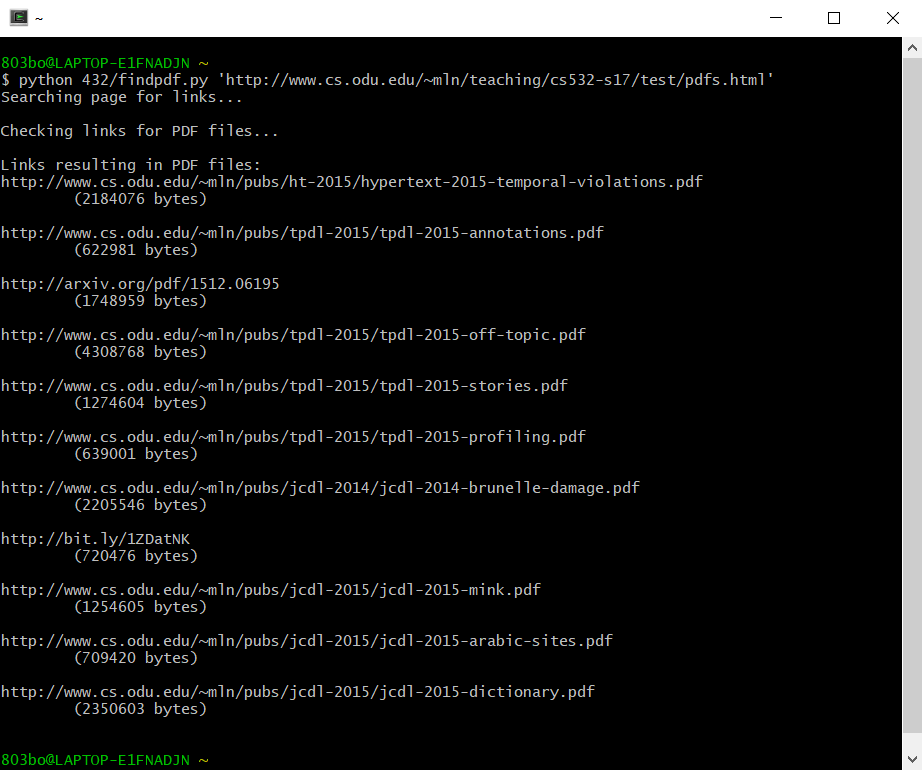


Figure 3: Results from Demonstration Page

The program was also tested on two Wikipedia articles and a test page created specifically for the purpose. The results can be seen in Figure 4.

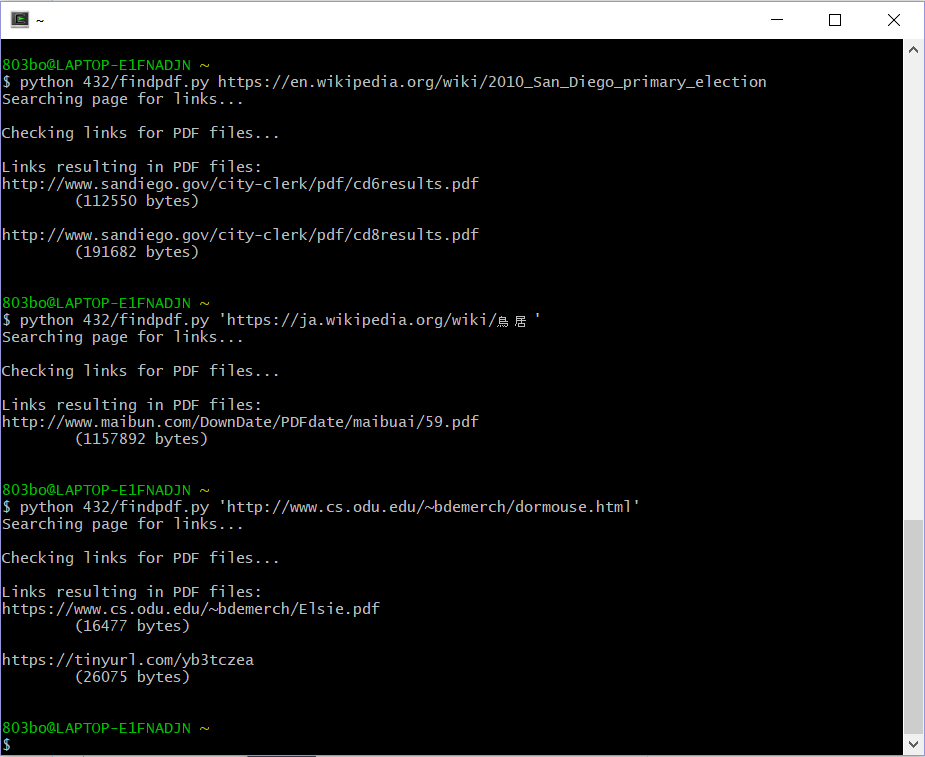


Figure 4: Results of Secondary Testing

# Problem 3:

Consider the "bow-tie" graph in the Broder et al. paper:

http://snap.stanford.edu/class/cs224w-readings/broder00bowtie.pdf

Now consider the following graph[[1]](#footnote-1):

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| A | B | C | C | C | E | G | G | I | I | L | M | M | N | O | P |
| ↓ | ↓ | ↓ | ↓ | ↓ | ↓ | ↓ | ↓ | ↓ | ↓ | ↓ | ↓ | ↓ | ↓ | ↓ | ↓ |
| B | C | D | A | G | F | C | H | H | K | D | A | N | D | A | G |

For the above graph, give the values for: IN, SCC, OUT, Tendrils, Tubes, Disconnected.

# Solution 3:

The first step to this solution is identifying the connections between the graph nodes (see Figure 5).

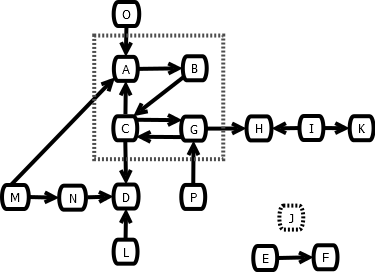


Figure 5: Directed Graph of Connections

The nodes in the dotted square (A, B, C, and G) form the SCC. Each of these nodes can be reached with a path of the other nodes from any other node in the square. Nodes O, M, and P are all IN. They each have no in-links and have out-links to other pages. Nodes H and D are OUT; they have only out-links. Nodes E, F, and J[[2]](#footnote-2) are disconnected. While E and F are connected to each other, none of them are connected to the rest of the network. Node L is a tendril, it has a link only to OUT (D). Node N is a tube; it is linked from IN and links to OUT but cannot reach the SCC. Node I is a tendril; it has only out-links to the other pages. Node K is then OUT, as defined in the secondary guide provided[[3]](#footnote-3); it has no out-links and its only in-link is from a tendril.

1. The graph has been reproduced horizontally in an effort to conserve space. [↑](#footnote-ref-1)
2. Node J, while not specifically referenced, is assumed to exist due to the existence of nodes I and K-P. [↑](#footnote-ref-2)
3. https://www.harding.edu/fmccown/classes/archive/comp475-s13/web-structure-homework.pdf [↑](#footnote-ref-3)