**Question 2**

[Unless I can get my environment properly setup on my home computer, I am unable to properly execute code, though I am able to write it. So I present my code as it is, and will attempt to best explain how it works in theory. All code will still be written as I could get it to execute. i.e. There will be no pseudocode.]

The Python program used for this question is in the file “ParsePDFvs.py”

Most of the code is based off of code obtained from *elssar@altrawcode.com1* on github. The original purpose of the code was to download all the PDF files linked on a given webpage. The syntax needed to run this code from the command line is listed in the code itself, but is generally “ParsePDF\_v2.py [URI]” for this particular usage of the code.

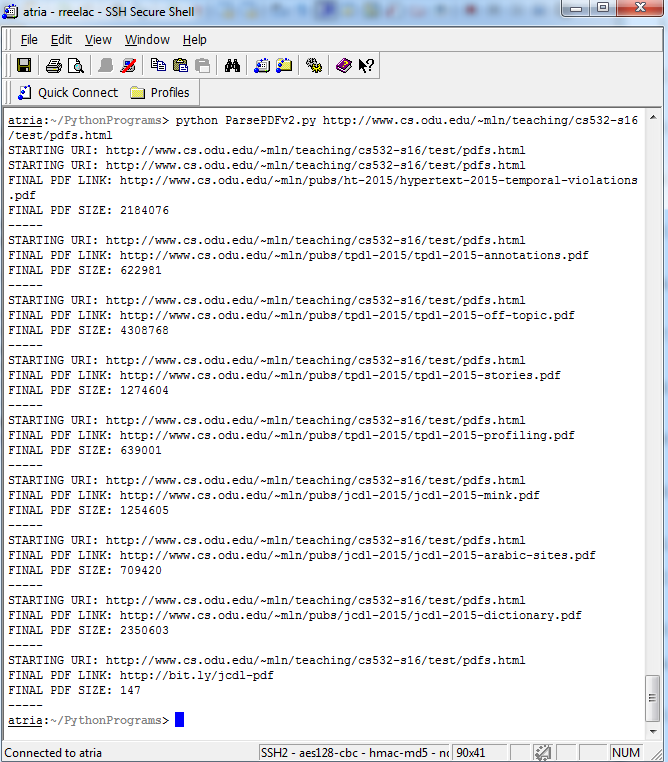
The first function “get\_page” on line 33 first checks to see if the URI is valid by checking for a status code of 200. This returns an error if no 200 code is not reached.

The second function “get\_all\_links” on line 39 searches the HTML code of the URI and gets any links it can find.

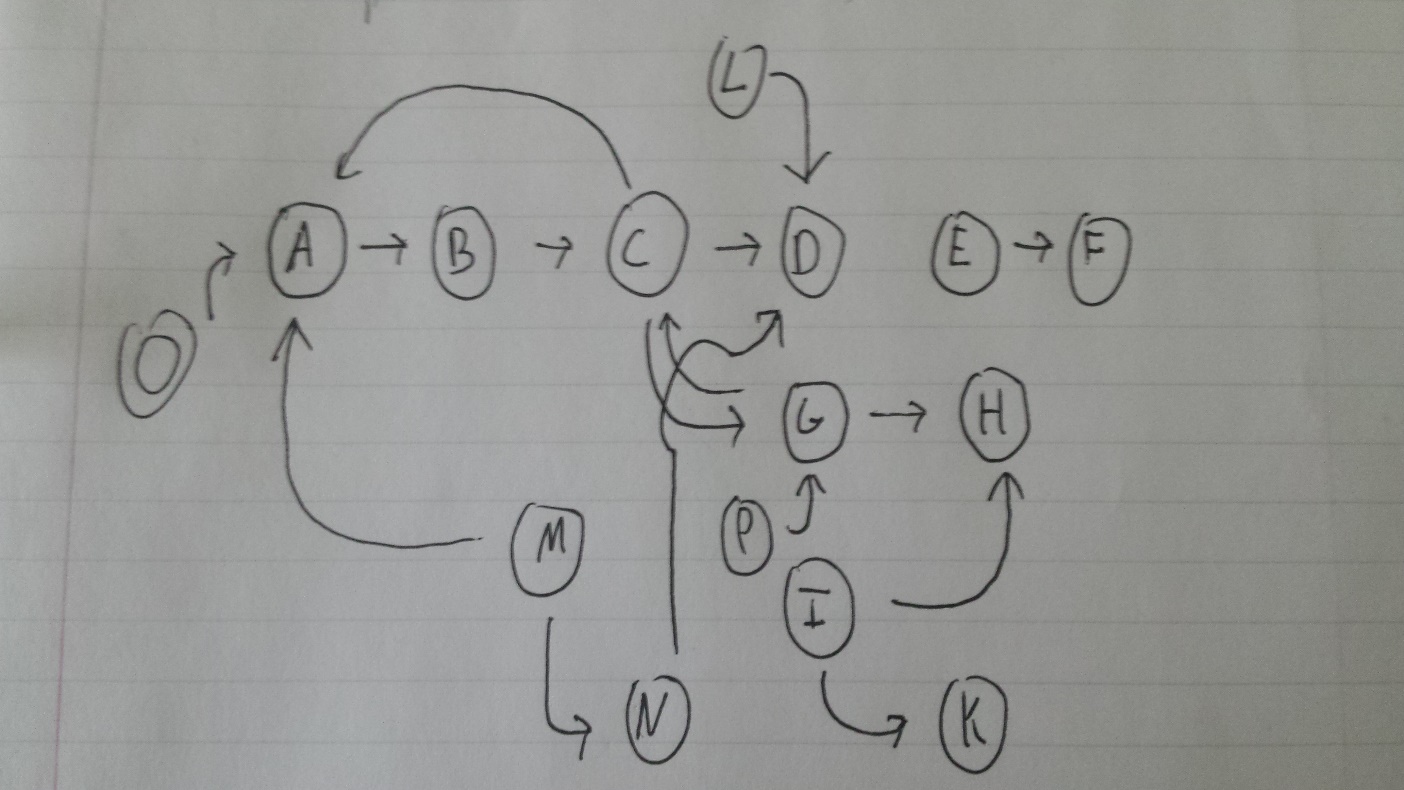
The primary function of the code “get\_pdf” on line 44 is where most of the work occurs. After grabbing and testing the URI and links with “get\_page” and “get\_all\_links”, the code then uses a FOR loop to search through the grabbed links any ending in “.pdf”.

Then, if a link returns a status code of 200 and matches the PDF file type, it prints the size of the PDF.

**Question 2**

****

**Question 3**

****

IN: *O, M*

IN components are starting points on the graph. They go into the SCC and nowhere else. They may have TENDRILS coming out of them that do not connect to the SCC.

SCC: *A, B, C, G*

SCC components are heavily connected with each other. In other words, it is possible to iterate through these components without leaving into the OUT. And they can only be accessed through the IN.

OUT: *D, H*

OUT Components exit the SCC and do not join back into it or IN.

TENDRILS: *I, K, L*

TENDRILS are components that are not within the SCC. They either come from somewhere else to join the IN, or go off from OUT.

TUBES: *N*

TUBES are connections that go directly from IN (i.e. M) to OUT (i.e. D) without ever joining the SCC.

DISCONNECTED: *E, F*

DISCONNECTED components are completely disconnected from the rest of the graph. They may connect to each other, but not to the IN, SCC, OUT, TENDRILS, or TUBES.

**References**

1. <https://gist.github.com/elssar/5160757>
2. https://gist.github.com/aio00/4463432