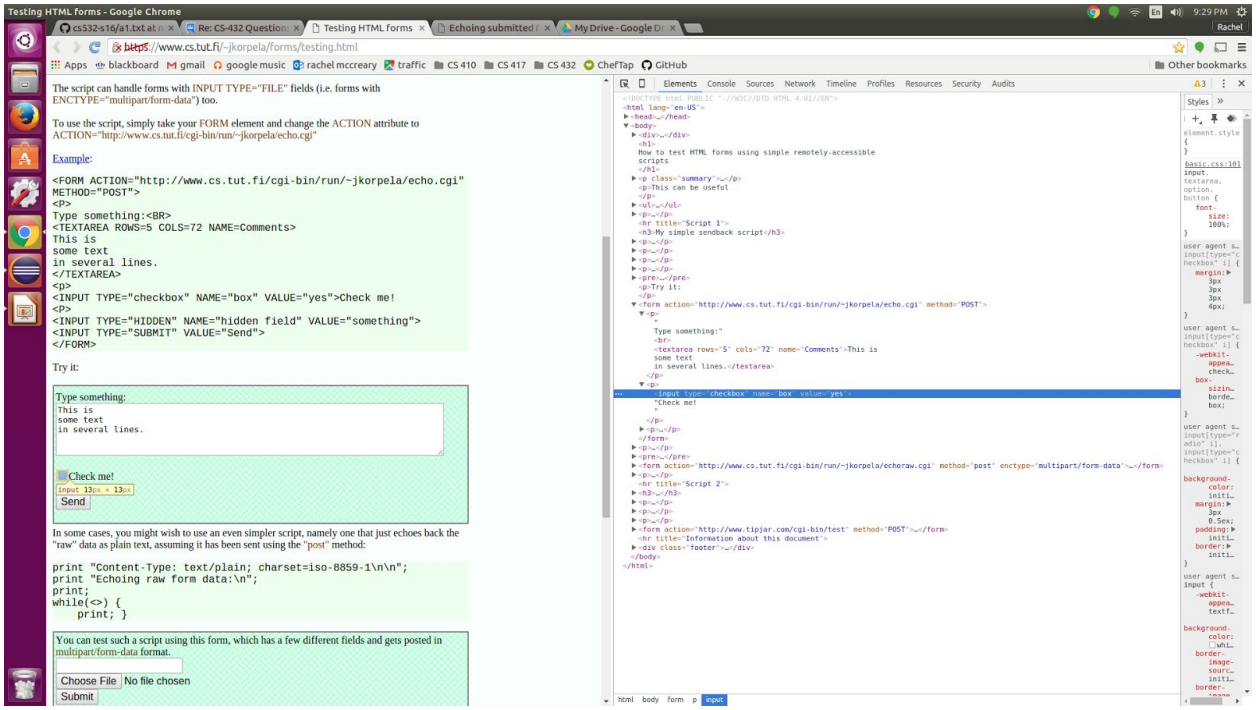
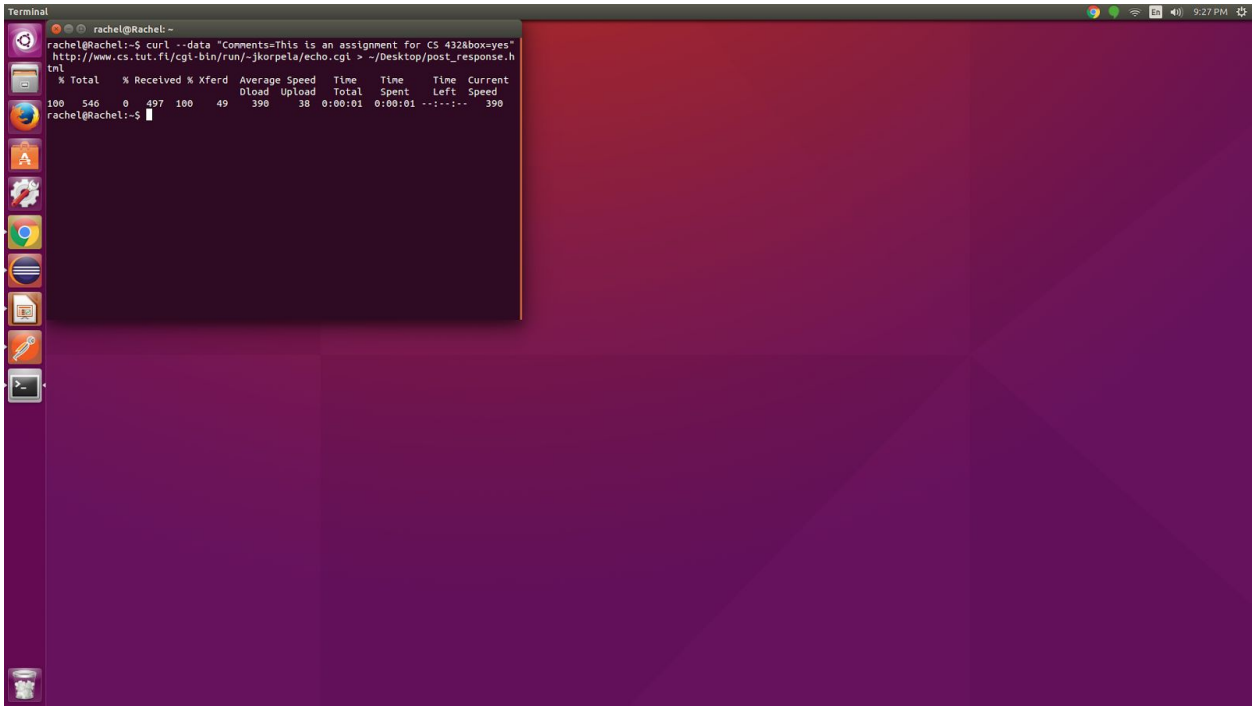
RACHEL MCCREARY
CS 432 ASSIGNMENT 1 REPORT

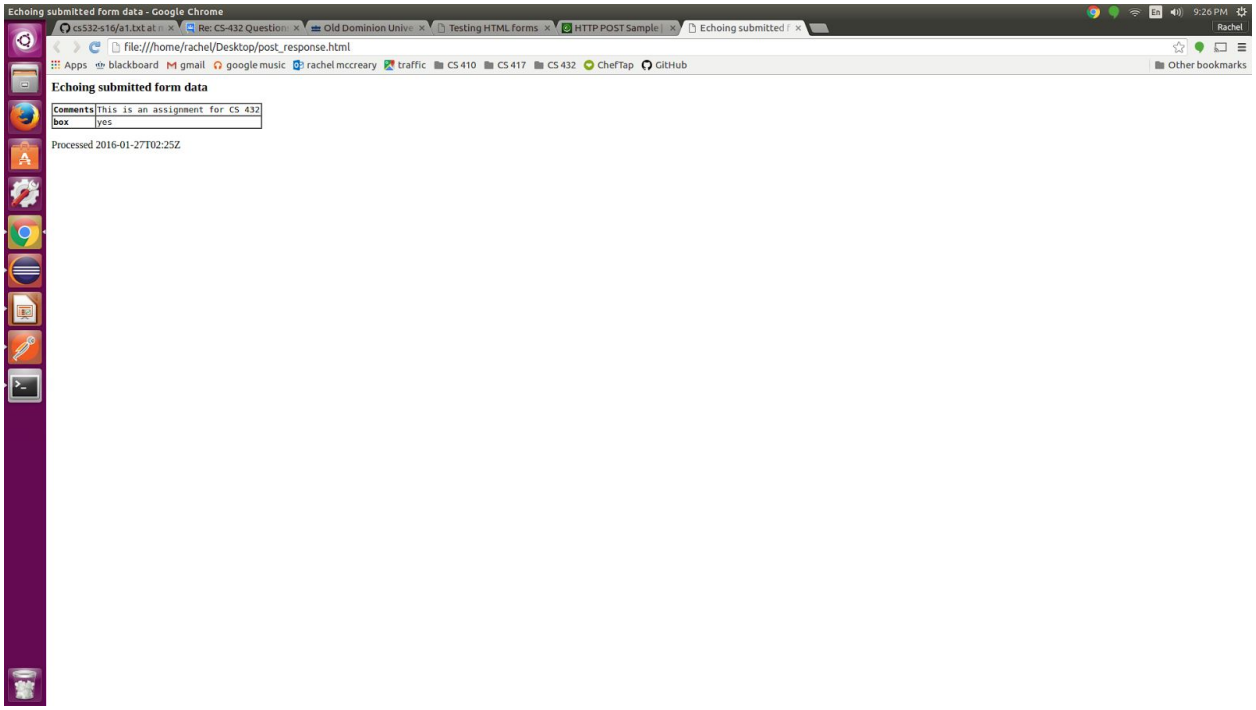
PART 1

I tried a POST here: https://www.cs.tut.fi/~jkorpela/forms/testing.html





and received this response:



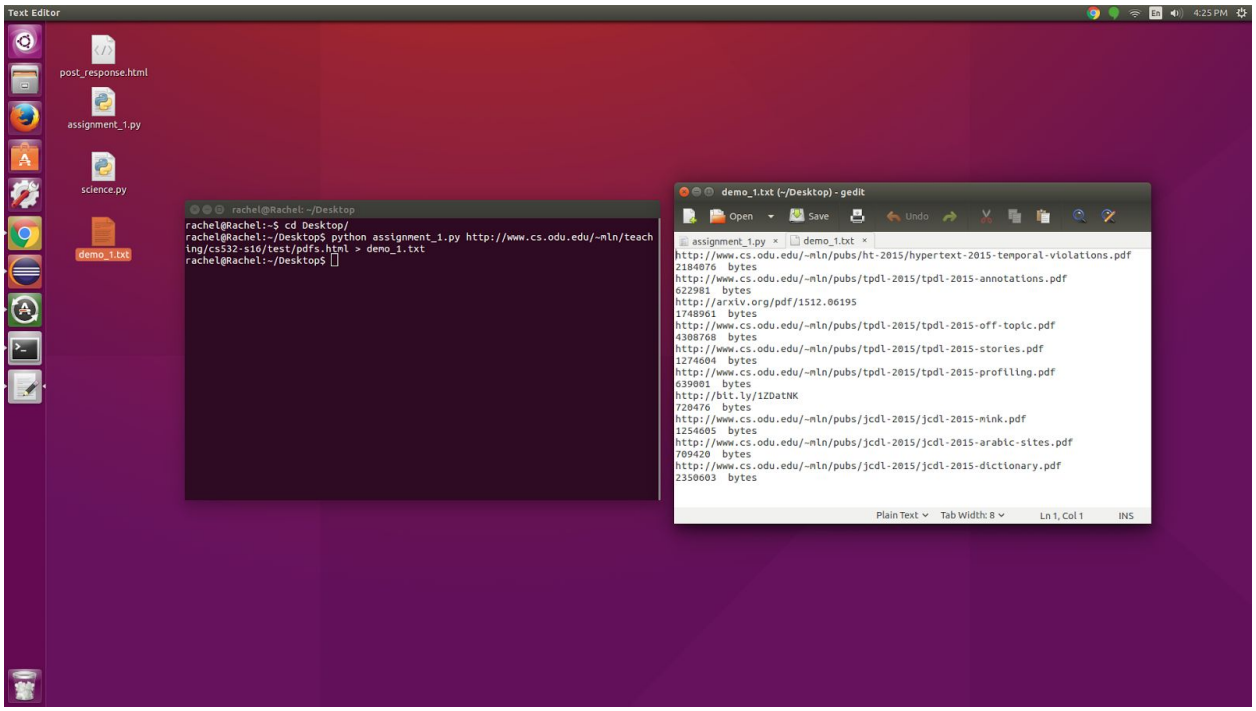
PART 2

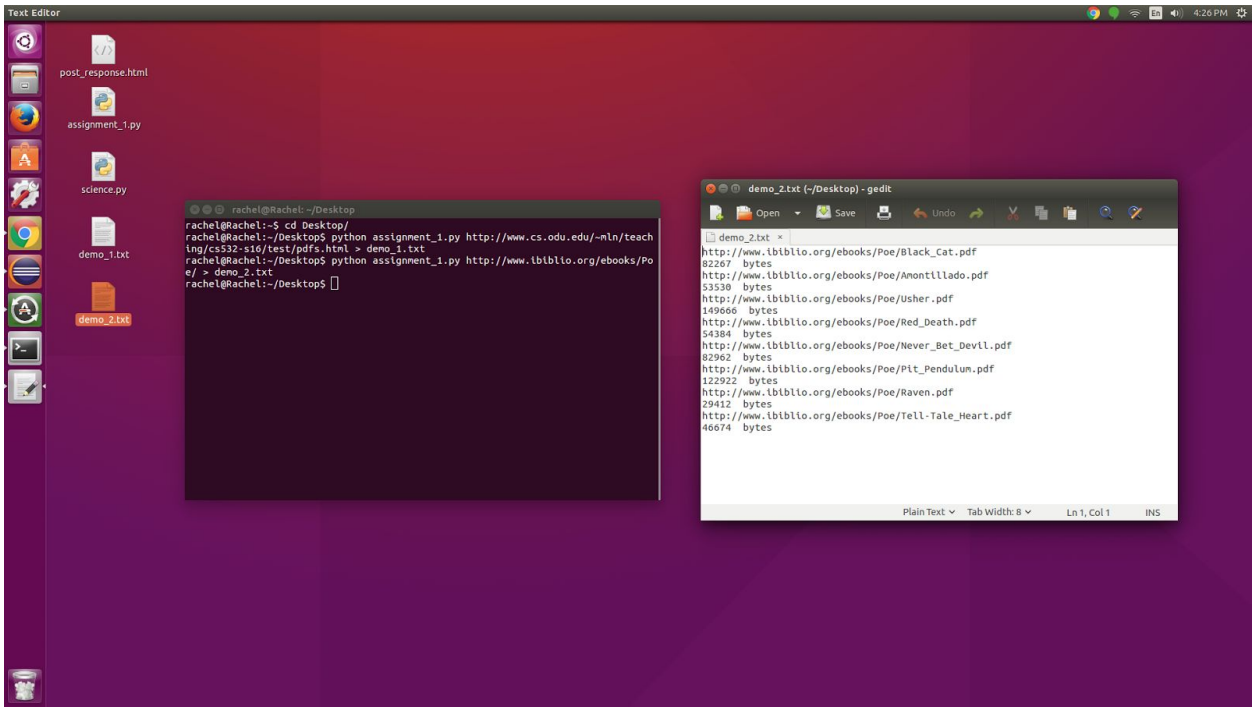
This program takes a web page as a command line argument, extracts all the links from that page, lists all links that result in PDF files, and prints out the size (in bytes) for each of the links. This was my first time programming in Python, so feedback is appreciated. (The actual file will also be attached)

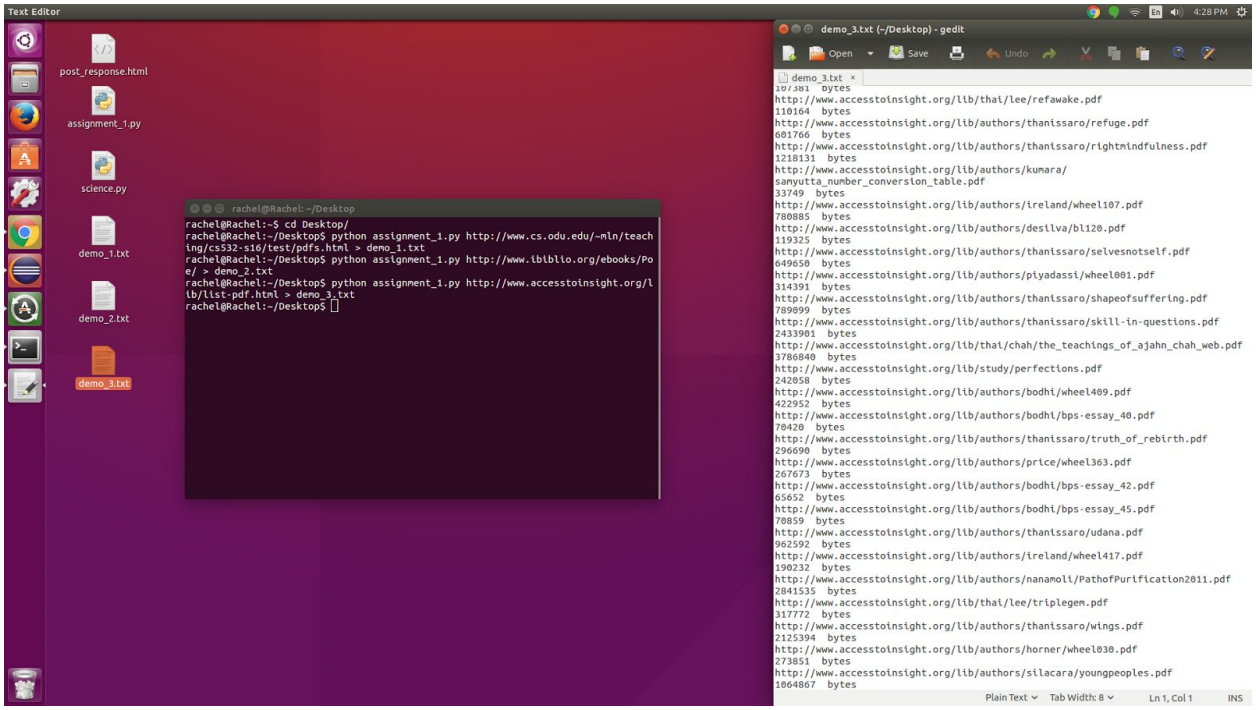
```
# RACHEL MCCREARY
# CS 432 ASSIGNMENT 1
from bs4 import BeautifulSoup
import urllib2
import urlparse
import sys
if __name__ == "__main__":
  for arg in sys.argv[1:]:
    uri = arg
  page = urllib2.urlopen(uri)
  soup = BeautifulSoup(page.read(), 'html.parser')
  for link in soup.find_all('a'):
    href = link.get('href')
    if href != None:
      if href.startswith("http") == False:
        href = urlparse.urljoin(uri, href)
        response = urllib2.urlopen(href)
        status_code = response.info().getheader('Status')
        content_type = response.info().getheader('Content-Type')
        if content_type == "application/pdf":
          print href
          size_of_pdf = response.info().getheader('Content-Length')
          print size_of_pdf, " bytes"
```

Libraries used: sys (for command line arguments)
BeautifulSoup (for extracting links)
urllib2 (for opening links and getting content type & length from headers)
urlparse (for dealing with redirects)

Trying it out on three different URIs:







PART 3

For this part, I mainly used these snippets in the the Broder et al. paper for reference:

“Figure 9: Connectivity of the web: one can pass from any node of IN through SCC to any node of OUT. Hanging off IN and OUT are TENDRILS containing nodes that are reachable from portions of IN, or that can reach portions of OUT, without passage through SCC. It is possible for a TENDRIL hanging off from IN to be hooked into a TENDRIL leading into OUT, forming a TUBE -- a passage from a portion of IN to a portion of OUT without touching SCC.”

“This connected web breaks naturally into four pieces. The first piece is a central core, all of whose pages can reach one another along directed links -- this "giant strongly connected component" (SCC) is at the heart of the web. The second and third pieces are called IN and OUT. IN consists of pages that can reach the SCC, but cannot be reached from it - possibly new sites that people have not yet discovered and linked to. OUT consists of pages that are accessible from the SCC, but do not link back to it, such as corporate websites that contain only internal links. Finally, the TENDRILS contain pages that cannot reach the SCC, and cannot be reached from the SCC.”

I ended up with these values:

IN: **M, O, P**
SCC: **A, B, C, G**
OUT: **D, H**
TENDRILS: **I, J, K, L, N**
TUBES: **N**
DISCONNECTED: **E, F**

My bow-tie graph: