Alexander Wilson

Assignment 1

1)

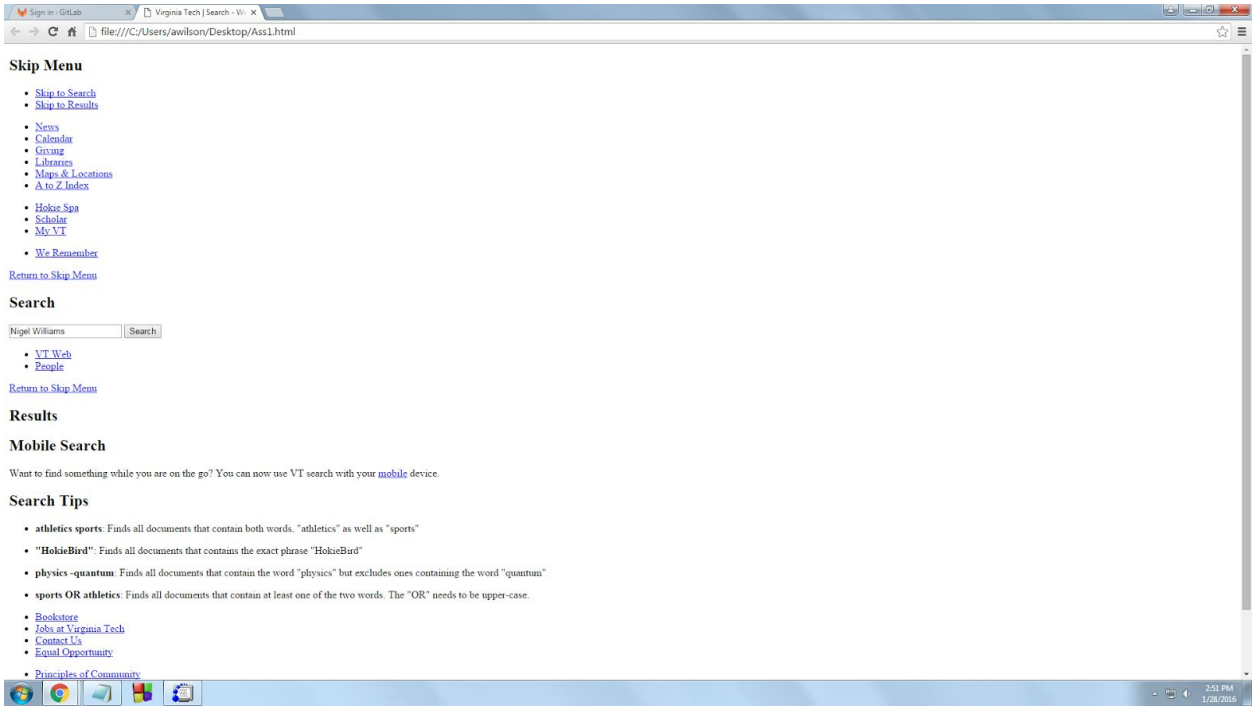
```
sirius:~> curl --data "q=Nigel Williams" http://search.vt.edu/search/pages.html
<!DOCTYPE html>


<html lang="en">
<head>
<meta http-equiv="Content-Type" content="text/html; charset=utf-8" />
<meta http-equiv="X-UA-Compatible" content="IE=edge" />
<meta name="viewport" content="width=device-width, initial-scale=1.0" />
<title>Virginia Tech | Search - Web</title>
<link rel="shortcut icon" type="image/x-icon" href="//www.assets.cms.vt.edu/images/favicon.ico"
/>
<link rel="stylesheet" href="/search/assets/css/base.css" type="text/css" media="screen" />
<link rel="stylesheet" href="/search/assets/css/enhanced.css" type="text/css" media="screen" />
<script type="text/javascript"
      src="//www.assets.cms.vt.edu/jquery/archives/jquery-1.10.latest.min.js"></script>
<script type="text/javascript" src="/search/assets/js/search_utils.js"></script>
<script type="text/javascript" src="/search/assets/js/search_pages.js"></script>
</head>
<body>
<div class="vt_skip">
      <h2><a id="vt-skipto-menu">Skip Menu</a></h2>
      <ul>
            <li><a href="#vt-skipto-search">Skip to Search</a></li>
            <li><a href="#vt-skipto-results">Skip to Results</a></li>
      </ul>
</div>
      <div id="container">
            <div id="header_container">
                  <!-- BEGIN HEADER -->
<div id="header">
      <div id="vt_logo">
            <a id="vt_home_btn" title="Virginia Tech" href="http://www.vt.edu"></a>
      </div>
      <div id="vt_utilities">
            <ul id="vt_toplinks">
                  <li><a href="http://www.vtnews.vt.edu">News</a></li>
                  <li><a href="http://www.calendar.vt.edu">Calendar</a></li>
                  <li><a href="http://www.givingto.vt.edu">Giving</a></li>
                  <li><a href="http://www.lib.vt.edu">Libraries</a></li>
```

```
                <li><a href="http://maps.vt.edu">Maps &amp;
                        Locations</a></li>
                <li><a href="http://www.vt.edu/az_index/index.html">A to Z Index</a></li>
            </ul>
            <ul id="vt_student_tools">
                <li><a

href="https://banweb.banner.vt.edu/ssb/prod/twbkwbis.P_WWWLogin">Hokie
                        Spa</a></li>
                <li><a href="https://scholar.vt.edu/portal">Scholar</a></li>
                <li><a href="https://my.vt.edu/">My VT</a></li>
            </ul>
            <ul id="vt_we_remember">
                <li><a href="http://www.weremember.vt.edu">We Remember</a></li>
            </ul>
        </div>
</div>
<!-- END HEADER -->
            </div>
            <div id="content_container">
                <div id="content">
<div class="vt_skip">
        <a href="#vt-skipto-menu">Return to Skip Menu</a>
        <h2><a id="vt-skipto-search">Search</a></h2>
</div>
                        <div id="vt_search_block">
                            <form action="#" onSubmit="return executeQuery()" method="get"
name="vt_search_form" id="vt_header_search_form">
                                <input type="text" maxlength="50" placeholder="Search pages
and people" name="q"
                                    value="Nigel Williams" id="vt_search_box"
autocomplete="off"/>
                                <button id="vt_go_button">
                                    <span class="vt_skip">Search</span>
                                </button>
                            </form>
                        </div>
                        <div id="navigation">
                            <ul>
                                <li class="current"><a href="#">VT Web</a></li>
                                <li>
```

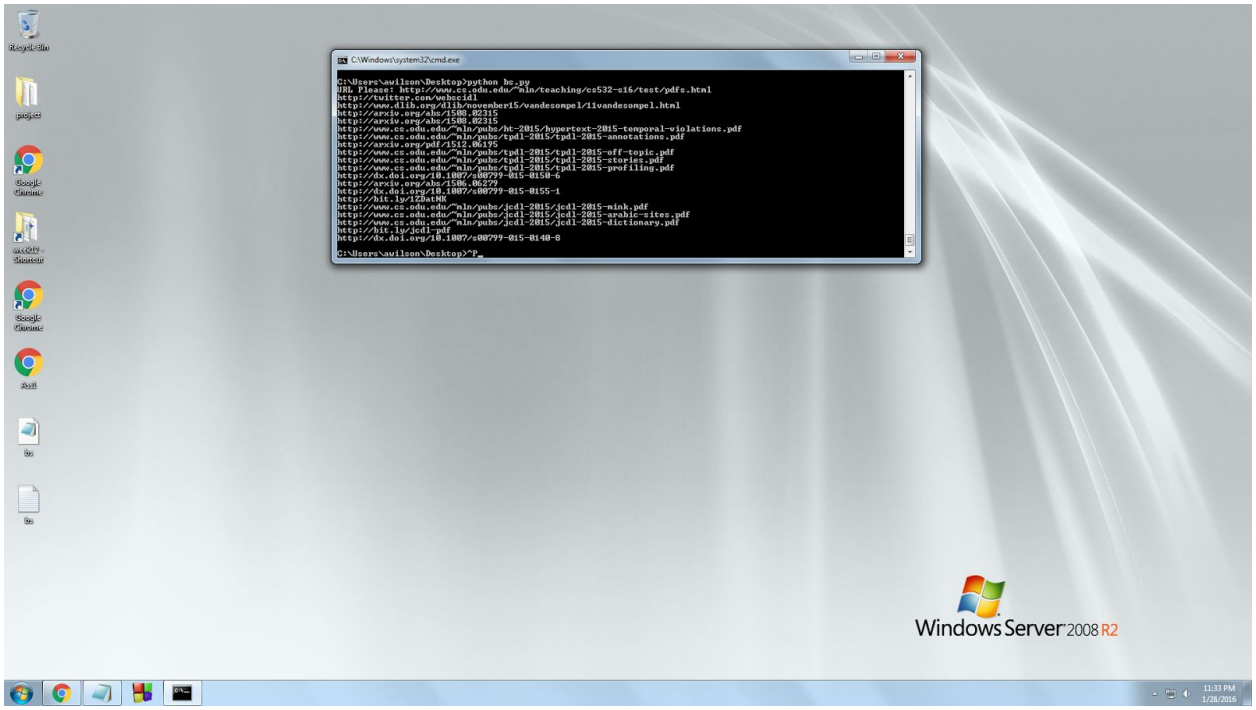
```
                                <a
href="people.html;jsessionid=E79F304A8A153027C8C3C7BA58A98F2E.mt-prod-3?q=Nigel+W
illiams" id="vt-people-nav">People</a>
                                </li>
                                </ul>
                        </div>
<div class="vt_skip">
        <a href="#vt-skipto-menu">Return to Skip Menu</a>
        <h2><a id="vt-skipto-results">Results</a></h2>
</div>
                        <div id="results">
                                <div id="vt_gcse_script">
                    <noscript>
                    <div class="noscript">
                                It looks like you have JavaScript turned off.  See search results
                                <a
href="http://www.google.com?cx=012042020361247179657:wmrvw9b99ug&cof=FORID:11&ie=
UTF-8&q=Nigel Williams">here</a>.
                        </div>
                    </noscript>
                    <div id="vt_gcse_results" class="gcse-searchresults-only" data-resultsetsize="7"
data-gname="vt_gcse_results">
                    </div>

                        </div>
                        </div>
                        <div id="rb_content">
                                <h2>Mobile Search</h2>
                                <p>
                                        Want to find something while you are on the go? You can now
use VT
                                        search with your <a href="/search/m">mobile</a> device.
                                </p>

                                <h2>Search Tips</h2>
                                <ul>
                                        <li>
                                                <p>
                                                        <strong>athletics sports</strong>: Finds all documents
that
                                                        contain both words, "athletics" as well as "sports"
                                                </p>
                                        </li>
```

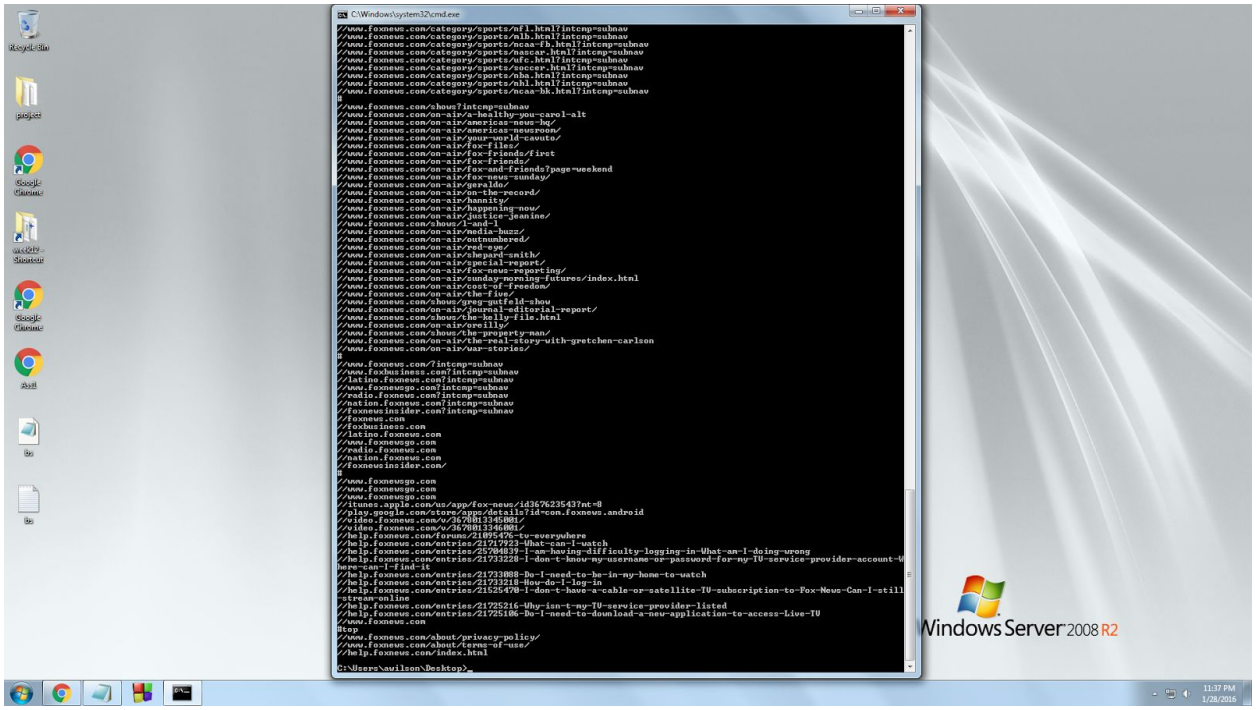
```
                                <li>
                                    <p>
                                        <strong>"HokieBird"</strong>: Finds all documents that
contains
                                        the exact phrase "HokieBird"
                                    </p>
                                </li>
                                <li>
                                    <p>
                                        <strong>physics -quantum</strong>: Finds all
documents that
                                        contain the word "physics" but excludes ones
containing the word
                                        "quantum"
                                    </p>
                                </li>
                                <li>
                                    <p>
                                        <strong>sports OR athletics</strong>: Finds all
documents that
                                        contain at least one of the two words. The "OR" needs
to be
                                        upper-case.
                                    </p>
                                </li>
                            </ul>
                        </div>
                    </div>
                </div>
                <div id="footer_container">


<!-- BEGIN FOOTER -->
<div id="footer">
    <ul>
        <li><a href="http://www.bookstore.vt.edu">Bookstore</a></li>
        <li><a href="http://www.jobs.vt.edu/">Jobs at Virginia Tech</a></li>
        <li><a href="http://www.vt.edu/contacts/">Contact Us</a></li>
        <li><a href="http://www.vt.edu/about/equal-opportunity.html">Equal
                Opportunity</a></li>
    </ul>
    <ul>
        <li><a
```

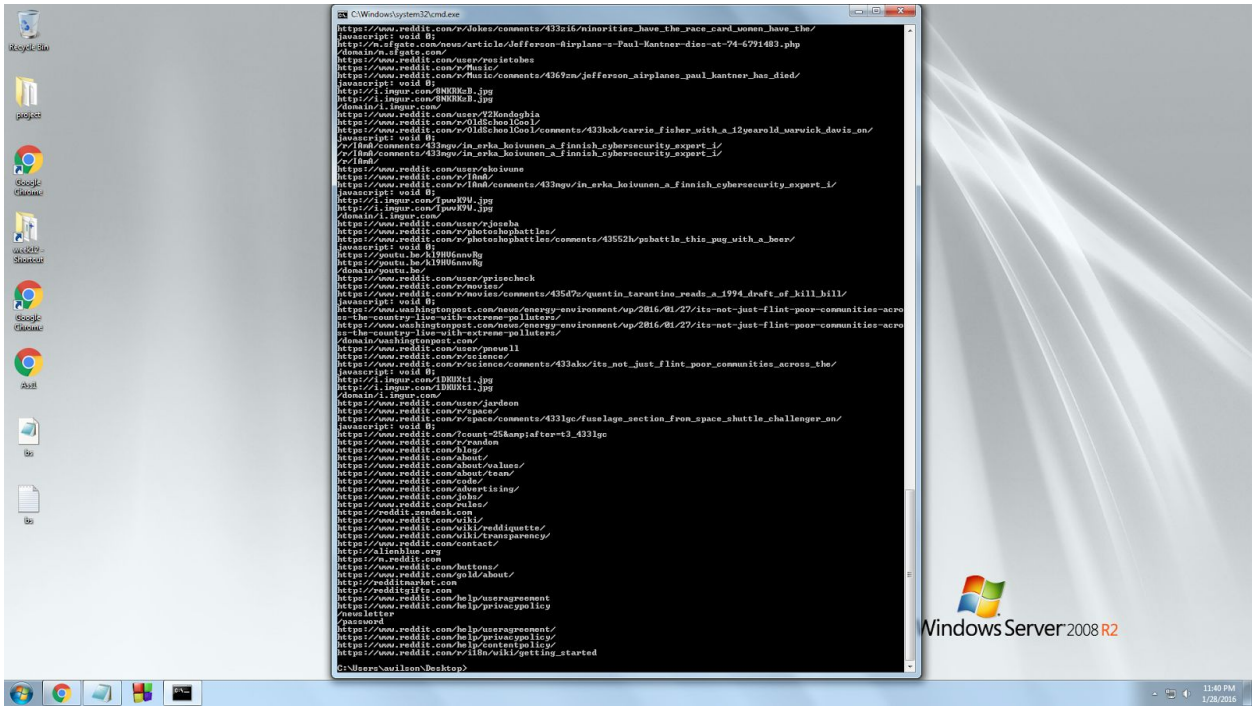
```
                href="http://www.vt.edu/diversity/principles-of-community.html">Principles
                    of Community</a></li>
            <li><a href="http://www.vt.edu/about/privacy.html">Privacy
                    Statement</a></li>
            <li><a href="http://www.vt.edu/about/acceptable-use.html">Acceptable
                    Use Policy</a></li>
            <li><a href="http://www.vt.edu/about/accessibility.html">Accessibility</a>
            </li>
        </ul>
        <p>&copy; 2016 Virginia Polytechnic Institute and State University. <span
id="version-number">VT Search: 2.4.3</span></p>
</div>
<!-- END FOOTER -->
            </div>
        </div>
<script type="text/javascript">
var gaJsHost = (("https:" == document.location.protocol) ? "https://ssl." : "http://www.");
document.write(unescape("%3Cscript src='" + gaJsHost + "google-analytics.com/ga.js'
type='text/javascript'%3E%3C/script%3E"));
</script>
<script type="text/javascript">
var pageTracker = _gat._getTracker("UA-5217491-2");
pageTracker._trackPageview();
</script>
</body>
</html>
```


Sign in · GitLab
Virginia Tech | Search - Vi
file:///C:/Users/awilson/Desktop/Ass1.html
Skip Menu
Skip to Search
Skip to Results
News
Calendar
Giving
Libraries
Maps & Locations
A to Z Index
Hokie Spa
Scholar
My VT
We Remember
Return to Skip Menu
Search
Nigel Williams
Search
VT Web
People
Return to Skip Menu
Results
Mobile Search
Want to find something while you are on the go? You can now use VT search with your mobile device.
Search Tips
athletics sports: Finds all documents that contain both words, "athletics" as well as "sports"
"HokieBird": Finds all documents that contains the exact phrase "HokieBird"
physics -quantum: Finds all documents that contain the word "physics" but excludes ones containing the word "quantum"
sports OR athletics: Finds all documents that contain at least one of the two words. The "OR" needs to be upper-case.
Bookstore
Jobs at Virginia Tech
Contact Us
Equal Opportunity
Principles of Community
2:51 PM
1/28/2016

2)


C:\Windows\system32\cmd.exe
C:\Users\awilson\Desktop>python bs.py
URL Please: http://www.cs.odu.edu/~mln/teaching/cs532-s16/test/pdfs.html
http://twitter.com/webscidl
http://www.dlib.org/dlib/november15/vandesompel/11vandesompel.html
http://arxiv.org/abs/1508.02315
http://arxiv.org/abs/1508.02315
http://www.cs.odu.edu/~mln/pubs/ht-2015/hypertext-2015-temporal-violations.pdf
http://www.cs.odu.edu/~mln/pubs/tpdl-2015/tpdl-2015-annotations.pdf
http://arxiv.org/pdf/1512.06195
http://www.cs.odu.edu/~mln/pubs/tpdl-2015/tpdl-2015-off-topic.pdf
http://www.cs.odu.edu/~mln/pubs/tpdl-2015/tpdl-2015-stories.pdf
http://www.cs.odu.edu/~mln/pubs/tpdl-2015/tpdl-2015-profiling.pdf
http://dx.doi.org/10.1007/s00799-015-0150-6
http://arxiv.org/abs/1506.06279
http://dx.doi.org/10.1007/s00799-015-0155-1
http://bit.ly/1Z3atMX
http://www.cs.odu.edu/~mln/pubs/jcdl-2015/jcdl-2015-mink.pdf
http://www.cs.odu.edu/~mln/pubs/jcdl-2015/jcdl-2015-arabic-sites.pdf
http://www.cs.odu.edu/~mln/pubs/jcdl-2015/jcdl-2015-dictionary.pdf
http://bit.ly/jcdl-pdf
http://dx.doi.org/10.1007/s00799-015-0140-8
C:\Users\awilson\Desktop>^P_
Windows Server 2008 R2


C:\Windows\system32\cmd.exe
//www.foxnews.com/category/sports/nfl.html?intcmp=subnav
//www.foxnews.com/category/sports/mlb.html?intcmp=subnav
//www.foxnews.com/category/sports/ncaa-fb.html?intcmp=subnav
//www.foxnews.com/category/sports/nascar.html?intcmp=subnav
//www.foxnews.com/category/sports/ufc.html?intcmp=subnav
//www.foxnews.com/category/sports/soccer.html?intcmp=subnav
//www.foxnews.com/category/sports/nba.html?intcmp=subnav
//www.foxnews.com/category/sports/nhl.html?intcmp=subnav
//www.foxnews.com/category/sports/ncaa-bk.html?intcmp=subnav
#
//www.foxnews.com/shows?intcmp=subnav
//www.foxnews.com/on-air/a-healthy-you-carol-alt
//www.foxnews.com/on-air/americas-news-hq/
//www.foxnews.com/on-air/americas-newsroom/
//www.foxnews.com/on-air/your-world-cavuto/
//www.foxnews.com/on-air/fox-files/
//www.foxnews.com/on-air/fox-friends/first
//www.foxnews.com/on-air/fox-friends/
//www.foxnews.com/on-air/fox-and-friends?page=weekend
//www.foxnews.com/on-air/fox-news-sunday/
//www.foxnews.com/on-air/gerald/
//www.foxnews.com/on-air/on-the-record/
//www.foxnews.com/on-air/hannity/
//www.foxnews.com/on-air/happening-now/
//www.foxnews.com/on-air/justice-jeanine/
//www.foxnews.com/shows/1-and-1
//www.foxnews.com/on-air/media-buzz/
//www.foxnews.com/on-air/outnumbered/
//www.foxnews.com/on-air/red-eye/
//www.foxnews.com/on-air/shepard-smith/
//www.foxnews.com/on-air/special-report/
//www.foxnews.com/on-air/fox-news-reporting/
//www.foxnews.com/on-air/sunday-morning-futures/index.html
//www.foxnews.com/on-air/cost-of-freedom/
//www.foxnews.com/on-air/the-five/
//www.foxnews.com/shows/greg-gutfeld-show
//www.foxnews.com/on-air/journal-editorial-report/
//www.foxnews.com/shows/the-kelly-file.html
//www.foxnews.com/on-air/oreilly/
//www.foxnews.com/shows/the-property-man/
//www.foxnews.com/on-air/the-real-story-with-gretchen-carlson
//www.foxnews.com/on-air/war-stories/
#
//www.foxnews.com/?intcmp=subnav
//www.foxbusiness.com?intcmp=subnav
//latino.foxnews.com?intcmp=subnav
//www.foxnewsgo.com?intcmp=subnav
//radio.foxnews.com?intcmp=subnav
//nation.foxnews.com?intcmp=subnav
//foxnewsinsider.com?intcmp=subnav
//foxnews.com
//foxbusiness.com
//latino.foxnews.com
//www.foxnewsgo.com
//radio.foxnews.com
//nation.foxnews.com
//foxnewsinsider.com/
#
//www.foxnewsgo.com
//www.foxnewsgo.com
//www.foxnewsgo.com
//itunes.apple.com/us/app/fox-news/id367623543?mt=8
//play.google.com/store/apps/details?id=com.foxnews.android
//video.foxnews.com/v/3670813345001/
//video.foxnews.com/v/3670813346001/
//help.foxnews.com/forums/21095476-tv-everywhere
//help.foxnews.com/entries/21717923-What-can-I-watch
//help.foxnews.com/entries/21704839-I-am-having-difficulty-logging-in-What-an-I-doing-wrong
//help.foxnews.com/entries/21733228-I-don-t-know-my-username-or-password-for-my-TV-service-provider-account-W
here-can-I-find-it
//help.foxnews.com/entries/21733088-Do-I-need-to-be-in-my-home-to-watch
//help.foxnews.com/entries/21733218-How-do-I-log-in
//help.foxnews.com/entries/21525470-I-don-t-have-a-cable-or-satellite-TV-subscription-to-Fox-News-Can-I-still
stream-online
//help.foxnews.com/entries/21725216-Why-isn-t-my-TV-service-provider-listed
//help.foxnews.com/entries/21725106-Do-I-need-to-download-a-new-application-to-access-Live-TV
//www.foxnews.com
$top
//www.foxnews.com/about/privacy-policy/
//www.foxnews.com/about/terms-of-use/
//help.foxnews.com/index.html
C:\Users\awilson\Desktop>_
Windows Server 2008 R2




```
from bs4 import BeautifulSoup          #used to declare the location of the library and to import it
import requests                        #used to import the requests library to be used for declaring the
urls
import urllib2                         #The main library used

url = raw_input("URL Please: ")        #To allow the user to imput their own URL
response = requests.get(url)  #Setting response to be the name of the urls

page = str(BeautifulSoup(response.content, 'html.parser'))  #used to read the html file and to
access BS4


def urlextractor(page):                #definiton for the website being parsed
    start_link = page.find('href')     #the inital link used to find links inside the page
    if start_link == -1:               #If there is no links then no URI's exist
        return None, 0
    start_quote = page.find("", start_link)       #starts the url extraction process
    end_quote = page.find("", start_quote + 1) #
    url = page[start_quote + 1: end_quote]
    return url, end_quote

while True:      #the loop to determine if a url will be produced
    url, n = urlextractor(page)
```

```
page = page[n:]
if url:                     #if it exisits print if not then go back to the begging until
    print url    #there are no more start links
else:
    break
```

3)

IN: O, P, M
OUT: A, B, C, G
SCC: D, H
Tendrils: I, J, K, L
Tubes: N
Disconnected: E, F