Matthew Gomez
CS 432 – Web Science
Assignment 1
1/28/16

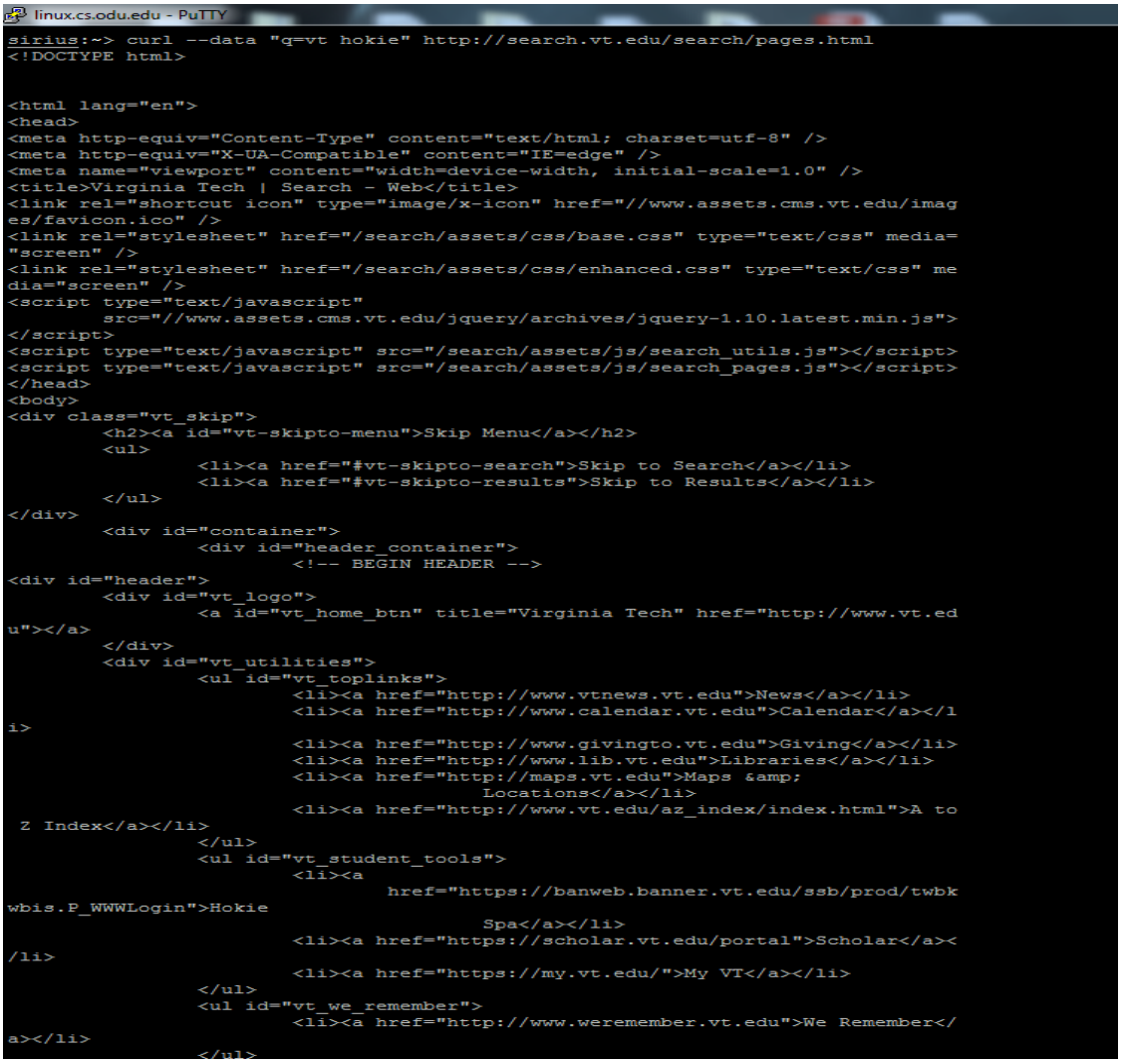
Program Narrative

For the first assignment of this course, we were tasked to write a program written in Python, which should be able to read any URI from and able to extra the links present on the page. Redirections are also required to be shown via command line if such a task is present. The program is also required to pull all PDF files laying around in the links alongside their content length; bytes size of the files.

The program consists of two functions, first function titled “LinkExtractor”, extracts the links directly from the page the specific URI directs to. The second function titled “LinkHandler” serves the purpose of getting a request and response while also handling the pdf extraction. As well as reporting their individual byte sizes.

PART 1

1. Curl command : curl --data "q=vt hokie" 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>

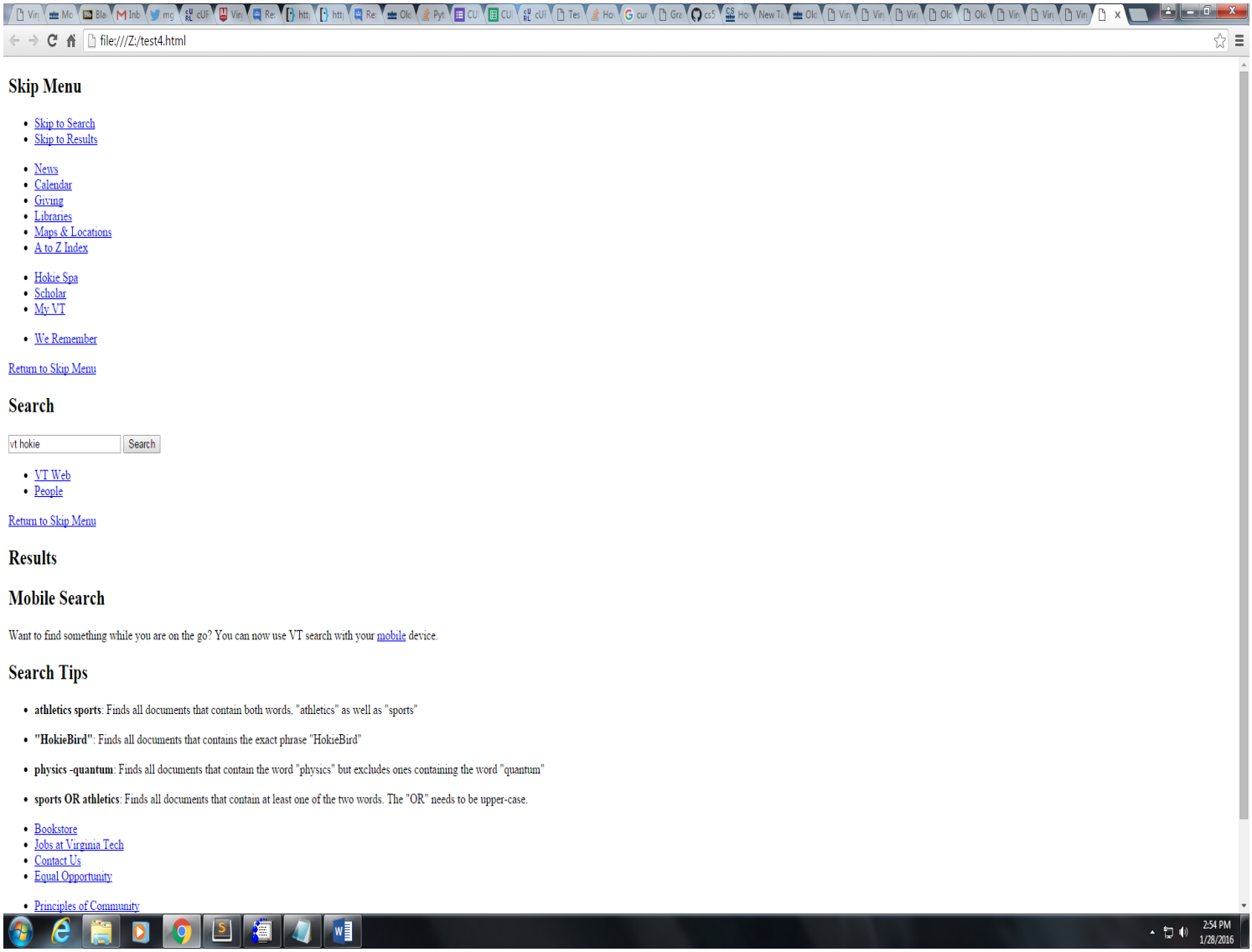
```
<script type="text/javascript" src="/search/assets/js/search_utils.js"></script>
<script type="text/javascript" src="/search/assets/js/search_pages.js"></script>
</head>
<body>
<div class="vt_skip">
	<h2><a id="vt-skipto-menu">Skip Menu</a></h2>
	<ul>
		<li><a href="#vt-skipto-search">Skip to Search</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="vt hokie" 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=0A02594B4BAC966B5EA2A1520BB386C6.mt-prod-4?q=vt+hokie"
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=vt hokie">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>
```

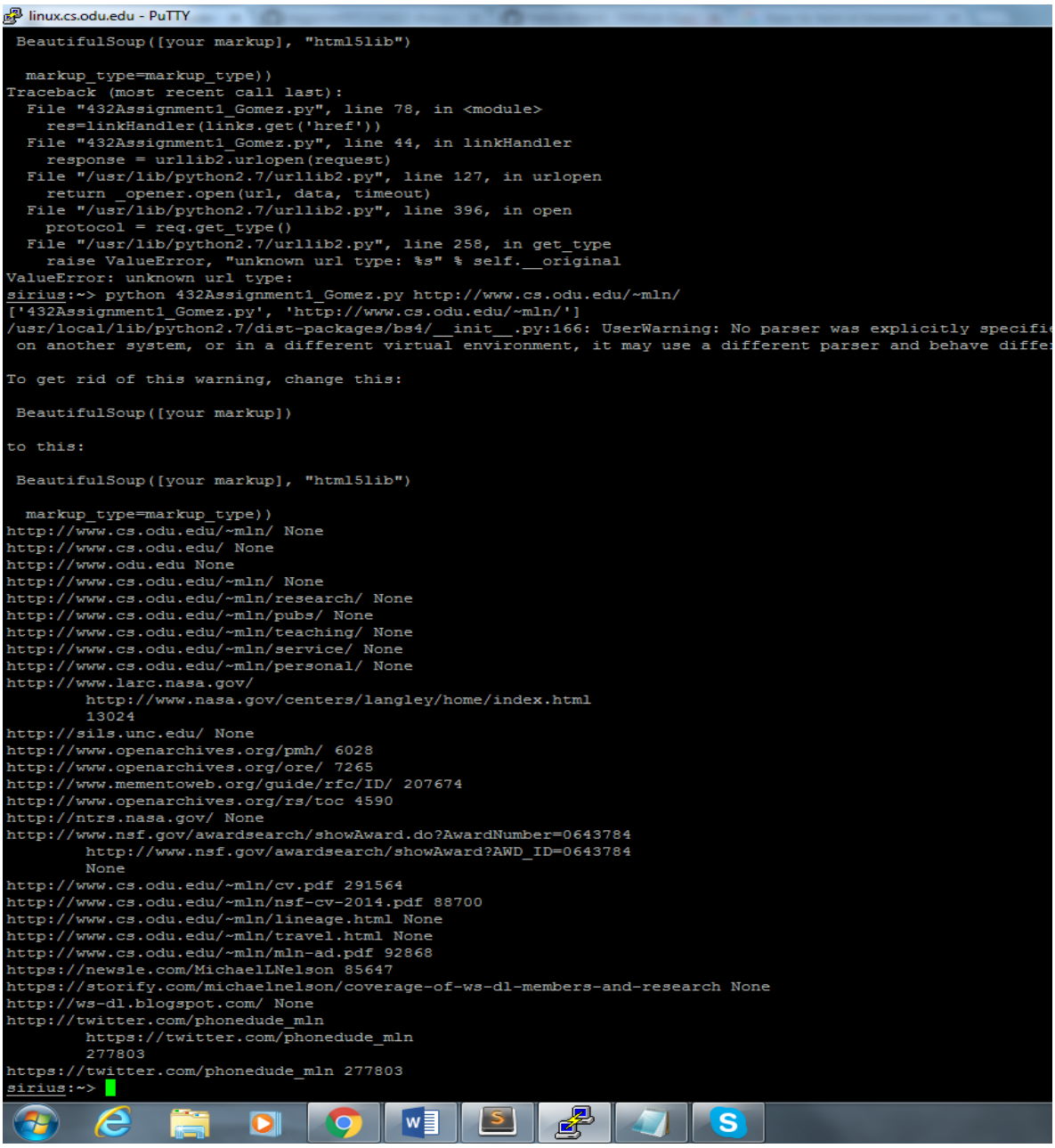
```
            </div>
        </div>
        <div id="footer_container">


<!-- 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 Output form Curl Command

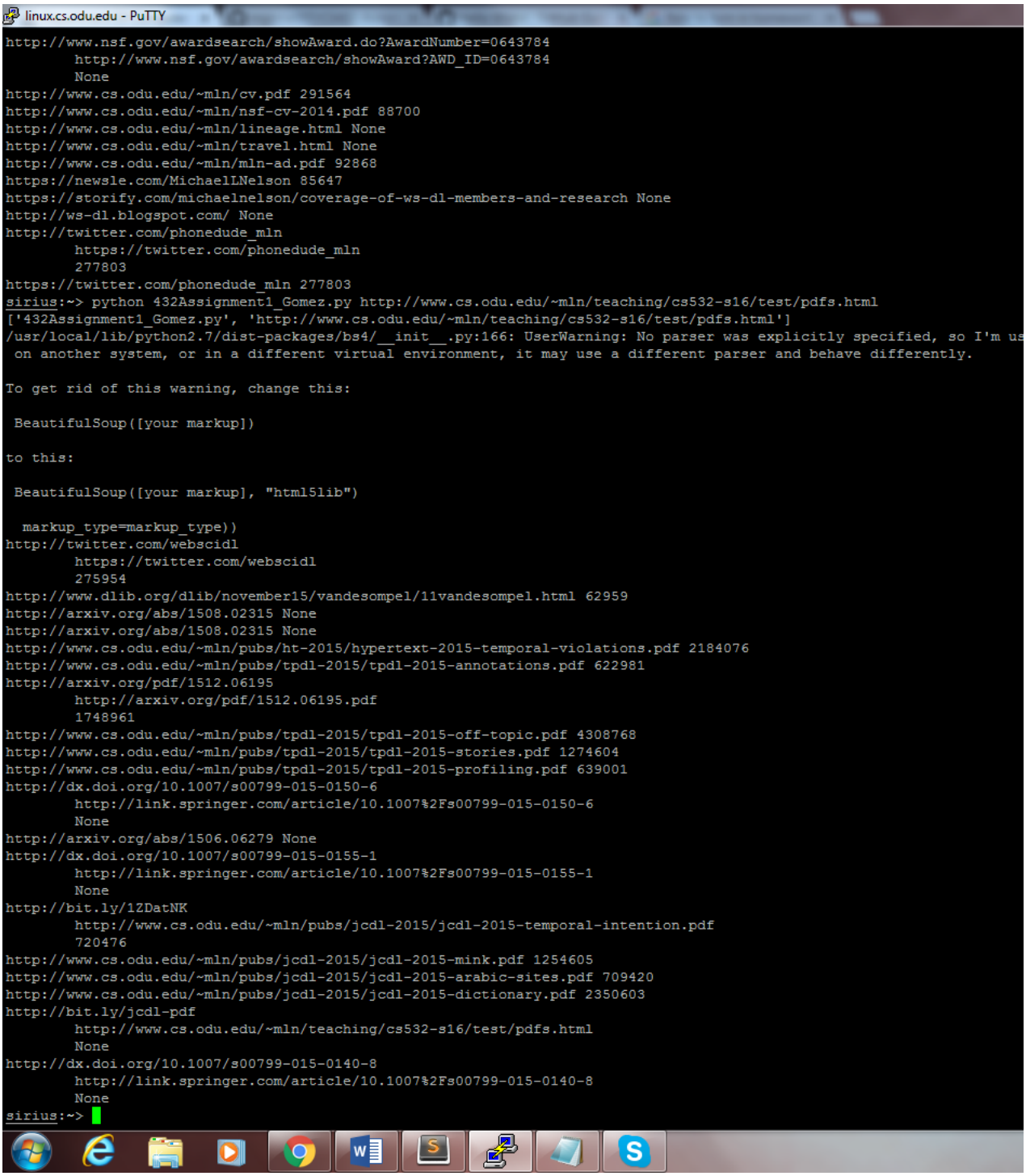


PART 2- URIs

http://www.cs.odu.edu/~mln/



http://www.cs.odu.edu/~mln/teaching/cs532-s16/test/pdfs.html



PART 3

Bow-Tie Graph Values:

IN: O, P, M

SCC: A, B, C, G

OUT: D, H

Tendrils: I, J, L

Tubes: N

Disconnected: E, F, K