**Report for CSE 5311 – Programming Project**

Name : Sarvesh Sadhoo   
Project : **Web Crawler**  Date : November 26,2013

**Project Specification:**In this project the Web Crawler perform the following functionalities:

1. Find the diameter of the underlying graph (i.e. the longest shortest path), and print out the URLs of the two pages at the two ends of the diameter as well as the diameter (path distance) itself.

2. For each page found, it prints out the URL of the page, the outbound links of the page and the inbound link is of the page.

**Development Environment:**Languages : Python 2.7Text Editor/IDE : Sublime Text and PyCharmOperating System : Windows 7External Libraries Used: Beautiful Soup and NetworkX

**Steps on how to compile and run program:**

1. Install or use Python 2.7 version from the python website: <http://www.python.org/download/>
2. Download and Install Beautiful Soup from [http://www.crummy.com/software/BeautifulSoup/#Download](http://www.crummy.com/software/BeautifulSoup/%23Download)
3. Download and Install NetworkX from <http://networkx.github.io/documentation/latest/install.html>
4. To run the code, open the command prompt and set the path where the python file is present.
5. Give the command 'python<space> filename.py . The program will start running in the cmd.

**Algorithm Used:**   
1. Breadth First Search for crawling the website cse.uta.edu  
2. Dijkstra’s Algorithm for finding the shortest path between the nodes (link).

**Project/Program Features:**1. Web crawler is able to crawl 116 unique links starting from cse.uta.edu domain.  
2. Since there are 116 links which can be considered nodes in a graph, there are 2149 edges between these nodes.  
3. The web crawler handles pages that are not a hyperlink and are out of cse.uta.edu domain.

**Specific Observations:**   
Since the output of the program is quite large it is ideal to run the code in a IDE like PyCharm. The cmd might not be able to display the program output correctly.