**Report for CSE 5311 – Programming Project – Group 1 Question 2**

Name : Sarvesh Sadhoo and Arjun ShettyUTA ID : 1000980763 and 100099694  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 Beautiful Soup from [http://www.crummy.com/software/BeautifulSoup/#Download](http://www.crummy.com/software/BeautifulSoup/%23Download)
3. Download 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.

**Installing Beautiful Soup:**After downloading it and run python setup.py install from the directory that you have unzipped Beautiful Soup in (assuming that you have added Python to your system PATH - if you haven't and you don't want to you can run C:\Path\To\Python27\python "C:\Path\To\BeautifulSoup\setup.py" install)

**Installing NetworkX:**  
Go to the link mentioned above and select the link (<https://pypi.python.org/pypi/networkx>) under quick install and download [networkx-1.8.1.zip](https://pypi.python.org/packages/source/n/networkx/networkx-1.8.1.zip#md5=4a90328898d9db29bbb12ef3e6223bab) ([md5](https://pypi.python.org/pypi?:action=show_md5&digest=4a90328898d9db29bbb12ef3e6223bab)) and install it similarly as Beautiful Soup.

**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).

**Data Structure Used:  
LIST**: The list type is a container that holds a number of other objects, in a given order. The list type implements the sequence protocol, and also allows you to add and remove objects from the sequence.

**DICTIONARY:** A dictionary is mutable and is another container type that can store any number of Python objects, including other container types. Dictionaries consist of pairs (called items) of keys and their corresponding values.

Python dictionaries are also known as associative arrays or hash tables. The general syntax of a dictionary is as follows:

dict = {'Alice': '2341', 'Beth': '9102', 'Cecil': '3258'}

**Project/Program Features:**1. The 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.

**References:**

1. <http://www.python.org/>  
2. <https://www.udacity.com/course/cs101>  
3. http://www.tutorialspoint.com/python   
4. <http://www.codecademy.com/learn>  
5. <http://www.slideshare.net/poonamkenkre/web-crawler-14590800>

**IMPORTANT NOTE:**  
It might be possible that certain libraries will not function properly or they might not configure in the system. In that case we would request you to contact us.