**CSE 5311 – Programming Project – Group 1**

**1. Web Crawler**

Write a web crawler/spider (click here for the Wikipedia definition) which crawls the cse.uta.edu (CSE) domain. E.g.: While starting from cse.uta.edu, if cse.uta.edu/faculty is encountered, it is followed. If www.uta.edu is found on cse.uta.edu, it is ignored since it is outside the CSE domain. A crawler is a tool which starts at a page and follows some algorithm to traverse to pages that can be reached by following hyperlinks. In the real world, crawlers are used primarily for traversing the structure of the domain and to create indices on the content of the underlying pages. For the purpose of this project the focus should be to only find the underlying structure of the document graph where each document represents a single vertex. Directed edges exist between vertices whenever there are hyperlinks connecting one document to the other. Thus, after the crawling procedure is complete, a directed graph representing the underlying structure should be the output of the crawl. The first task consists of writing a Breadth First Search crawler to crawl cse.uta.edu as mentioned above. This task requires proper and adequate data structures to represent the underlying graph. Parsing the right information from every page needs to be done carefully (Look at free crawler code bases available online). After the crawl is complete, the next task is to do the following information:

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, print out the URL of the page, the outbound links of the page and the inbound link is of the page.