# Assignment 4 – Web Crawler

Maximum Points = 50

The purpose of this lab is to focus on the reading of data from a web document using the URL class. This assignment also requires a considerable use of the String class.

Most web documents are written in the HTML markup language (http://en.wikipedia.org/wiki/Html ). "A **Web crawler** is a computer program that browses the World Wide Web in a methodical, automated manner or in an orderly fashion. " (http://en.wikipedia.org/wiki/Webcrawler ) One of the features of a webcrawler is the ability to parse a web page and find all the links to other pages. You have been asked to write the program to implement this feature.

**BASIC ASSIGNMENT**

a) Design and implement a program that asks the user to enter URLs (e.g. http://csc.colstate.edu/summers/NOTES/1302/lab4.htm ), one at a time until the user tells the program to stop.

b) Design a class LinkFinder that finds all hyperlinks of the form

    *<a href="link">link text</a>*

   on the webpage at the specified URL

c) Once you have built the list of links on the page, print the list of links and associated link texts.

d) Make sure to include necessary constructors, accessors & mutators (gets/sets), and toString methods for all classes.

e) You can read the contents of a web page with the following sequence of commands:

```
String address = "http://csc.colstate.edu/summers/security.htm";
URL url = new URL (address);
Scanner in = new Scanner(url.openStream());
```

1. NOTE: Some of these methods may throw exceptions—check out the API documentation.
2. Throw an exception if you find a malformed link (e.g. missing a protocol).

EXTRA CREDIT: If your program follows the links that it finds and finds the links in those web pages as well.

## Sample Input / Output

```
Enter a URL (type quit to stop):
http://csc.colstate.edu/summers/security.htm
Links for http://csc.colstate.edu/summers/security.htm
address: http://csc.colstate.edu/notes/security.htm          link text: My Notes on Computer Crime,
Security,and Computer Viruses
address: http://csc.colstate.edu/ComputerCrime.html                 link text: Computer Crime / Incident
Handling
address: http://matrix0.members.beeb.net/iso-17799/                 link text: ISO 17799 - What is
iso17799 (the ISO Security Standard)?
address: http://www.unixtools.com/securecheck.html                  link text: Unix Computer Security
Checklist
:
Enter a URL (type quit to stop):
http://cs.colstate.edu/
Links for http://cs.colstate.edu/
```

```
address: http://text.usg.edu:8080/tt/http://cs.colstate.edu/           link text:  Text Only Version
address: http://www.columbusstate.edu/              link text: <img src="/images/school_name_sm.gif"
alt="Columbus State University" align="left" />
address: http://cs.colstate.edu/https://colstate8.view.usg.edu/               link text: CougarVIEW
:
Enter a URL (type quit to stop):
Quit
Goodbye
```

(Due before class on Wednesday, March 23, 2011) Submit a .doc file containing the UML class diagram showing inheritance for all the classes used in your program. [10 pts]

(Due before class on Wednesday, March 30, 2011) Submit your .java files containing your program to the dropbox in WebCT. [50 pts]

Grades are determined using the following scale:

- Runs correctly..…………………:___/10
- Correct output……..……………:___/10
- Design of output..………………:___/8
- Design of logic…………………:___/10
- Standards……………………….:___/7
- Documentation.………………...:___/5

Grading Rubric  (Word document)