# CSE 7/5337: Information Retrieval and Web Search Spring 2012

## Project 1: Web crawler (100 points)

**Due**: 3/2 at noon per email to mhahsler@lyle.smu.edu

**Teams**: Individual student or teams of 2.

**Deliverables**:

1. Complete code in a compressed archive (zip, tgz, etc)
2. A readme file with complete installation, compilation and execution instructions
3. A document with documentation of your code (used software, UML diagrams, etc.) and the results for the questions below.
4. If you worked in a team than you need a document stating who worked on what.

Develop a Web crawler. Test your crawler only on the data in:

```
http://lyle.smu.edu/~mhahsler/crawler_test/michael.hahsler.net/
```

**Make sure that your crawler is not allowed to get out of this directory!!!**

1. Identify the key properties of a web crawler. Describe in detail how each of these properties is implemented in your code. [20 points]
2. Use your crawler to create a list of all pages in the test data and report all out-going links (leaving the directory crawler_test on the server) of the test data. [10 points]
3. How many PDF files are included in the test data? [10 points]
4. Use your crawler to identify all broken links in the test data. [10 points]
5. Make your crawler save a compressed version of each page. In this process give each page a unique document ID. [10 points]
6. Create a word frequency list from the compressed pages. Make sure that you do not count HTML markup. Report the 10 most common words. Order all words by frequency and plot the word frequency on the y axis and the rank on the x axis. [30 points]
7. Rerun the last step with stemming and a suitable stop word list. Compare the results. [10 points]