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

## Project 3: Complete Search Engine (100 points)

**Due**: 4/22 at midnight per email to mhahsler@lyle.smu.edu

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

**Tasks:**

Create a web-based interface (using Java servlets, PHP or any other technology you can install on Linux) for your search engine with the following features:

1. Set up the web-server infrastructure [10 points]
2. Google-style interface:
    - Unlimited number of terms are automatically combined with a logical *AND*. [20 points]
    - - in front of a term means that the term should not occur in the result. [10 points]
    - "" creates a phrase query. [20 points]

    **Note:** Make sure that your search terms are correctly stemmed!
3. The results are sorted using *tfxidf*. **Note:** Remember that you can get the needed information from your positional index! [20 points]
4. The result shows the title of the page and the URL (see Google or Bing). [10 points]
5. The result has an excerpt of the text from the web page with the found terms highlighted (see Google or Bing). It is OK if the excerpt has only stemmed words and not the original words in the document (see bonus tasks). [10 points]

**Bonus Tasks**:

1. Crawl the whole SMU web space. [+10 points]
2. The result contains a maximum of 10 links with the possibility to page through more with a next button (requires session management or memoization). Also, if there are no or only a few results your search engine can drop terms to return more results. [+10 points]
3. The excerpt contains the original (unstemmed) text with the search terms highlighted. [+10 points]

**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, block or UML diagrams, etc.) and a description for each task. **Make sure that you have screen shots of how your search engine with examples for the tasks from above.**
4. If you worked in a team than you need a document stating who worked on what.