# CSCI 6370 IR and Web Search Project

## Part 3: Indexing and Searching the Entire Corpus

### Due date: Monday, June 22 @ 11:59 PM.

### Note 1: The course project “MySearchEngine.com” is composed of 5 part with one part per week.

### Note 2: The project is a group project. That is, you can organize a group of at most 3 members to work together.

### Note 3: Only Python is allowed for completing the course project.

## Problem Description:

This part of the project is composed the following tasks:

### Task 1: Spidering and indexing

Now, you need to work on the entire document corpus “rfh.zip”. You can download “rfh.zip” from the Blackboard. For this task, you shall complete a spider which takes input file “index.html” from the folder “rhf/”. From this input html file, your spider shall be able to collect every document in “rhf/”. You only need to index “.html” or “.htm” files. For each web page, apply your indexer from Part 2 to index it.

Additional job: If you haven’t considered anchor texts in Part 2, please extract these and include them in your indices.

### Task 2: Testing and Revising of Interface and Search functions

Test your interfaces and search functions from Part 2. Make sure they work for the entire corpus.

Additional job: Make the search results (the links) clickable.

## How to submit your work?

Submit your python codes to Blackboard before the deadline.

One group only needs to submit one copy. Please make sure to include group members’ names and ID’s in your submission.

## Grading

The TA would do test runs of your program to check whether it works.