## Project Proposal: Classification of Faculty directory pages and homepages

Author: Sameen Shaukat

Net ID: shaukat2

**Overview:**

ExpertSearch System is a search engine that is developed by students of University of Illinois to facilitate faculty search based on their research interests. The goal for this project is to enhance or aid the ExpertSearch System by writing a classification utility to identify good faculty directory and faculty webpage URLs. This utility/functionality can be added to the existing search system later on.

**Goals:**

Goals for this project are

- Identifying faculty directory URLs
- Identifying faculty webpage URLs

**Proposed Solution:**

Following methodology will be used as a starting point to achieve above stated goals.

1. Use data provided in Assignment MP2.1 and MP2.3 labeled as good examples and collect data for bad URLs and classify these as bad examples.
2. Based on this data, scrape these URLs to pre-process data.
3. Automatic or manual feature extraction based on data collected in step 2. This will be decided on the basis of results achieved.
4. Develop a utility in Python to identify good URLs from bad ones. This will involve thorough analysis in order to choose the best classification technique.
5. Integrate this utility in ExpertSearch System. (This is optional and time-dependent)

These steps will be used as a reference point and there is a possibility of addition of extra steps along the way of project completion.

Initial plan is to build a separate utility that can perform the above tasks and then to integrate it into the ExpertSearch System later on depending on time availability. It's expected that item no 2, 3 and 4 will be most time consuming and take up to 70% of the project time.

The results will be demonstrated by executing performance testing of the classification model for each task in project report. A utility which will return the results of classification model when URL is entered will also be provided to demonstrate the working of this system.