# IST657: BASICS OF INFORMATION RETRIEVAL SYSTEMS
## (AKA – Search Engines)

Fall 2008

Nancy McCracken
Research Associate Professor
School of Information Studies
Syracuse University

**Class Meetings**: Tuesdays, 11:00am to 1:50pm, 117 Hinds Hall

**Professor Contact Information**: njm@ecs.syr.edu, Office 209 Hinds Hall, x3955

**Course description**:

Information Retrieval (IR) is the area of Information Science that focuses on the electronic provision of information from textual, image, and sound databases to users in response to their information needs. IR possesses a long tradition of basic and applied research with emphasis on both the technical and human side of the provision of timely, accurate information. IR research is pursued in both academic and commercial organizations, such as Google, Yahoo and Microsoft, with increasing interaction between the two groups. IR is an area of resurgent research interest resulting from the broadly recognized fact that all of the information now residing on the web or on intranets is of little use if it cannot be effectively retrieved in response to users' needs.

This course will focus on text document retrieval. In the first half of the course, students will learn the full technical details of how IR is accomplished from both the theoretical and applied perspectives. This includes

- the processing of documents to build retrieval indexes,
- processing user queries and retrieving documents,
- improving retrieval results by query processing and ranking techniques,
- and methods for evaluating retrieval results.

In the second half of the course, there will be discussion of additional techniques for web searches and emerging technologies that incorporate retrieval:

- Using link structure in Google's pagerank algorithm
- Question answering
- Cross-language retrieval
- Retrieval on new text genre, such as email and blogs

In keeping with its status as a 600-level course, the course will provide opportunities for active student participation and initiative. Therefore, you should be prepared to summarize readings and discuss knowledgeably the topics assigned for each class, as well as discuss what you learn from your assignments and team projects.

Coursework will consist of weekly readings in preparation for discussion in class, small exercises, a mid-term exam, and a team project culminating with an end-of-semester presentations or papers. Intermediate reports will also be given in class.

While search engine tools and document collections will be made available for projects, each student team will be responsible for designing a problem scenario and set of tasks and experiments. Students will be assigned to teams based on their indicated interest in technologies, collections, and role that they wish to play in the team.

The search engine tools and document collections have not been finalized and are subject to availability. However, the search engine tools will include (and these require minimal or no programming):

- The LEMUR search engine, which can be run as a software package with configuration files
- The Lucene search engine, which can be deployed through a Java API

Training will be given to use these software platforms.

Document collections will include

- The TREC newswire document collection, with standard evaluation sets
- The Enron email collection
- Collections of web documents and blogs

In the case of the latter collections, project teams will need to define and carry out their own evaluations of retrieval results.

## Text and Materials:

The recommended textbook is the new textbook Introduction to Information Retrieval by Manning, Raghaven and Schutze, July 2008. This book should be available at the university bookstore or on-line sources. It is also available in PDF form on the web at http:// www-csli.stanford.edu/~hinrich/information-retrieval-book.html. Only partial chapters will be assigned from this book.

Sections of Steven Arnold's book, The Google Legacy, and additional papers by experts in the field will be provided for weekly readings. These additional readings will be posted on the iLMS a week in advance, or handed out if in hard copy format.

## Tentative Schedule of Class Topics

- Introduction to Information Retrieval
- Steps in the Information Retrieval Process, building the document index and processing queries for retrieval
- Information Retrieval Models: Boolean, Vector Space, Probabilistic and Language Models
- IR Performance Evaluation: recall and precision
- Introduction to Retrieval Software Packages: Lemur and Lucene
- Improving retrieval with Query Expansion and Relevance Feedback
- Web-based Retrieval Systems: hyperlinks in Google's PageRank algorithm, popularity, authority, large distributed indexes
- Cross-Language Information Retrieval

- Passage Retrieval and Question-Answering
- Future Directions  and Student Predictions
- Group Project Presentations or Poster Session