# INLS 509

## Information Retrieval

Spring 2015
[Tuesday, 2:00-4:45]
Stone Center, RM 210

Instructor: Diane Kelly, Ph.D.
E-Mail: dianek@email.unc.edu
Phone: 919.962.8065
Office: Manning Hall, RM 10
Office Hours: Wednesday 2-3; Friday 1-2

## Description

The field of information retrieval (IR) is concerned with the analysis, organization, storage, and retrieval of unstructured and semi-structured data. In this course, we will focus mostly on text. While IR systems are often associated with Web search engines, IR applications also include digital library search, patent search, search for local businesses, and expert search, to name a few. Likewise, IR techniques (the underlying technology behind IR systems) are used to solve a wide range of problems, such as organizing documents into an ontology, recommending news stories to users, detecting spam, and predicting reading difficulty. This course will provide an overview of the theory, implementation, and evaluation of IR systems and IR techniques. In particular, we will explore how search engines work, how they "interpret" human language, what different users expect from them, how they are evaluated, why they sometimes fail, and how they might be improved in the future.

## Goals

- Improve students' understanding of how search technologies work and are evaluated.
- Help students think more critically about the power of search technologies.
- Increase students' familiarity with IR research and increase their confidence in reading this research.
- Improve students' quantitative and logical thinking skills.

## Pre-requisites and General Requirements

- There are no pre-requisites for this course if you are a SILS graduate student.  If you are an SILS undergraduate student, you should have completed INLS 201 (101).
- You are expected to read the assigned materials by the dates listed on the schedule.  Be prepared to ask questions and discuss the readings in class.
- You should have access to the readings during class.
- Please let me know in advance if you will not be able to attend class.

## Materials

- **Required Textbook**:  Croft, W. B., Metzler, D., & Strohman, T. (2009). *Search engines: Information Retrieval in Practice*.  Cambridge University Press.
- Some readings will be placed in the Sakai site or you will need to obtain them from UNC Library's online collections or from the open web.

## Assignments

**February 17, 2015**
Homework 1 Due

**March 17, 2015**
Literature Review Proposal Due

**April 7, 2015**
Homework 2 Due

**April 21, 2015**
Literature Review Due

**May 4, 2015 (Monday, 12-3 PM)**
Final Exam

## Evaluation

Your major assignments for this course are: Participation (10%), Homework (2@15% each), Literature Review (30%) and Final Exam (30%).

## Assignments

### Participation (10%)

You are an important part of this course. There are few things worse than teaching to a half-empty classroom. One thing worse than teaching to a half-empty classroom is teaching to a room full of people engaged in conversations with others or with their eyes glued to their computers! Because your presence is key to a successful class, your participation will be 'graded.' Your grade is based on several things: attendance, engagement and general participation.

- *Attendance.* You are expected to attend all classes and to arrive to class before it starts. You will lose participation points for excessive and unexcused absences and for arriving late to class. You should be seated and ready to start at 2:00 PM.
- *Behavior.* You are expected to behave *professionally*. This includes being courteous to your classmates and course instructor by not conversing with others during class lectures; turning off cell phones, pagers, and other devices that might disrupt class; using laptops and other devices to support current course activities only. Pay attention. Daydream infrequently. Do not sleep in class.
- *General Participation.* Class participation consists of doing the following: being prepared for class, making observations about the readings, asking questions, taking notes and actively listening.
- *Use of Laptops.* Use of laptops is strongly discouraged. Studies have shown that students who take notes on paper actually do better on exams. Not only do laptops provide many distractions, I have often observed that students work hard to type a verbatim transcript of what I am saying which means that they are not really paying attention to what I am saying, are not thinking about what I am saying, and are not being selective about what they record. It is also the case that we will work many problems on the board (think math and formulas): it can be difficult to quickly and accurately capture this with a laptop.

### Homework (2@15% each)

You will have two homework assignments that ask you questions about the readings and give you an opportunity to practice some of the problems we work in class. Homework should be submitted at the start of the class period when it is due. Late submissions will be subject to penalty. The penalty will be assessed based on the reason(s) that it was not turned in on time.

#### A Note about Collaboration

You are encouraged to learn from each other. However, all the work you hand in must be your own. This means that you cannot look at another student's answer and copy or re-word it as your own.

- If someone helps you with a homework assignment, please **write his/her name on the top of your homework**. This will not hurt you (provided your answer is your own), but it will help them.
- If you are the student giving help, don't give away the answer. Rather, help the student arrive at the answer him- or herself.
- If you are the student asking for help, don't ask for the answer. Rather, ask about the material. Your own answer must come from your own intuition. You must fully understand what you write and be able to explain your answer to the instructor.

### Literature Review (30%)

Note: Assignment description borrowed, in part, from Prof. Jaime Arguello's INLS 509 syllabus.

This assignment consists of a proposal (10%) and review (20%). The objective of this assignment is for you to gain an in-depth understanding of a particular area of Information Retrieval (IR) that is of interest to you.

Your **proposal** is due on **March 17** by the end of the day (i.e., midnight). In this proposal, you should identify the general topic of your literature review (one paragraph) and provide a *preliminary* list of 4-5 research papers you will review (your eventual target will be 8-10 papers). For each paper, you should provide a one-paragraph summary, including 2-3 sentences about why you selected the paper. If citation information is available for the paper, note this as well. You should try to pick the most important and well-known pieces of research to include and this can be a good indicator of impact.

This proposal is an important opportunity for you to get feedback from me about your topic, scope and resource selection. **Your proposal should be emailed to me.** You are encouraged to come by my office to discuss your preliminary ideas and findings BEFORE this assignment is due.

Your **review** is due on April 21. Your review should be about 15-20 double spaced pages, excluding reference and title pages, and discuss between 8-10 research papers. You should use APA style to format your paper. **Your review should be emailed to me.** Your paper should include the following information:

1. A description of the topic and the specific problem/issue you explored. What is the area/problem? Why is it important? Why is it difficult?
2. A survey of how others have attempted to solve the problem. Organize and present the reviewed literature in a way that shows you understand the different approaches to the problem. How are different solutions similar? How are they different? Do different solutions make different assumptions? What are those assumptions? How have solutions evolved with time? How have different approaches built on each other?
3. A survey of evaluation. How are solutions to the problem typically evaluated? Are there agreed upon metrics that are suitable to the task? What assumptions do these metrics make? Are there any specific conferences or workshops that specialize on the task?
4. *Relationship to things we've read or discussed in class. This is critical! Part of the purpose of this assignment is for you to show me what you've learned in class.*
5. What do you think? How do you see this area of IR progressing? What are the key problems that remain to be solved? How do you think this research will change people's lives? Your own view of the problem is very important. Make sure to allocate about 20% of your paper and presentation to this.

Around mid-February, I will distribute a list of suggested topics for this assignment, a starter-paper for each topic and a list of resources. We will also have the opportunity to discuss this assignment in class if you have questions.

## Final Exam (30%)

The Final Exam is cumulative and integrative, and will be closed-book and closed-note. The format of the exam questions will vary. You might also have a few multiple-choice questions. You will also have problems similar to those from the homework. The homework is designed to give you practice working different types of problems you will encounter on the exam and illustrate the different topics/issues that I think are most critical from the readings. During the final exam, you will be allowed to bring one-page of notes: these should be hand-written notes on one side of a blank piece of standard letter-sized paper.

## Grading

Grading is based on UNC Registrar Policy for undergraduate and graduate-level courses (http://registrar.unc.edu/academic-services/grades/explanation-of-grading-system/).

Graduate (H: 95-100%)

- Superior work: complete command of subject, unusual depth, great creativity or originality.

Undergraduate (A: 95-100%; A-: 90-94%))

- Mastery of course content at the highest level of attainment that can reasonably be expected of students at a given stage of development.

Graduate (P: 80-94%)

- Satisfactory performance that meets course requirements (expected to be the median grade of all students in the course).

Undergraduate (B: 80-89%)

- Strong performance demonstrating a high level of attainment for a student at a given stage of development.

Graduate (L: 70-79%)

- Unacceptable graduate performance: substandard in significant ways.

Undergraduate (C: 70-79%)

- A totally acceptable performance demonstrating an adequate level of attainment for a student at a given stage of development.

Undergraduate (D: 60-69%)

A marginal performance in the required exercises demonstrating a minimal passing level of attainment.

Graduate (F: 69% or less)

- Performance that is seriously deficient and unworthy of credit

Undergraduate (F: 59% or less)

- For whatever reason, an unacceptable performance. The F grade indicates that the student’s performance in the required exercises has revealed almost no understanding of the course content.

**January 13:  Introductions and Course Overview; Introduction to Information Retrieval; Architecture of a Search Engine**

- Croft, W. B., Metzler, D., & Strohman, T. (2009). *Search engines: Information Retrieval in Practice*.  Cambridge University Press. (Chapters 1 & 2)

**January 20: Getting Text: Crawls and Feeds**

- Croft, et al., Chapter 3

**January 27: Processing Text**

- Croft, et al., Chapter 4 (Note: SKIP section 4.5)

**February 3: Indexing and Query Processing**

- Manning, C. D., Raghavan, P., & Schutze, H. (2008). Introduction to information retrieval.  Cambridge University Press. (Chapter 1).  Available online at: **http://www-nlp.stanford.edu/IR-book/**
- Chapter 5: pgs. 125-140; 165-170; Chapter 6: 187-214
- **Web indexes: parallel and distributed; hierarchical.**

**February 10: Retrieval Models: Vector Space, Probabilistic and Language Modeling**

- Croft, et al., Chapter 7: 233-261

**February 17:  Retrieval Models: Vector Space, Probabilistic and Language Modeling; Catch-up**

- **DUE: HWK 1**

**February 24:  Web Search and Link Analysis**

- Easley, D. & Kleinberg, J. (2010).  *Networks, crowds and markets: Reasoning about a highly connected world* (Chapter 13)(Chapter 14: pgs. 397-417)(Chapter 18: pgs. 543-555).  Cambridge University Press.  Available online at: http://www.cs.cornell.edu/home/kleinber/networks-book/
- Croft, et al., Section 4.5

**March 3:  Evaluation & Experimentation: History, Frameworks and Measures**

- Robertson, S. (2008). On the history of evaluation in IR. *Journal of Information Science, 34*(4), 439-456. **[Library E-Journal]**
- Croft, et al., Chapter 8: pgs. 297-322.

**March 10: Spring Break**

**March 17:  Evaluation & Experimentation: Measures**

- Review Croft, et al. Chapter 8
- **DUE: Literature Review Proposals**

**March 24: Evaluation & Experimentation: Log Analysis**

- Dumais, S., Jeffries, R., Russell, D. M., Tang, D. & Teevan, J. (2014). Understanding user behavior through log data and analysis. J.S. Olson and W. Kellogg (Eds.), *Human Computer Interaction Ways of Knowing*. New York: Springer.
- Weber, I. & Jaimes, A. (2011). Who uses Web search for what and how. *Proceedings of the ACM Web Search and Data Mining Conference (WSDM '11)*, 15-24. **[ACM Digital Library]**
- Matthijs, N. & Radlinski, F. (2011). Personalizing Web search using long term browsing history. *Proceedings of the ACM Web Search and Data Mining Conference (WSDM '11)*, 25-34. **[ACM Digital Library]**

**March 31: No Class: Diane at European Conference on Information Retrieval (ECIR)**

**April 7: Interactive Information Retrieval and User Studies**

- Cool, C. & Belkin, N. J. (2011). Interactive information retrieval: history and background. In I. Ruthven & D. Kelly (Eds.) *Interactive Information Seeking, Behaviour and Retrieval*. Facet Publishing.
- Bennett, J. L. (1971). Interactive bibliographic search as a challenge to interface design. In D. E. Walker (Ed.) *Interactive bibliographic search: The user/computer interface*. Montvale, NJ: AFIPS Press.
- **DUE: HWK 2**

**April 14: Interactive Information Retrieval and User Studies**

- Kelly, D. (2009). Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends in Information Retrieval, 3*(1-2), Chapters 4-6; 9 & 10.
- Edwards, A., Kelly, D., & Azzopardi, L. (2015). The impact of query interface design on stress, workload and performance. *Proceedings of the European Conference on Information Retrieval (ECIR '15)*, Vienna, Austria.
- Qvarfordt, P., Golovchinsky, G., Dunnigan, T., & Agapie, E. (2013). Looking ahead: Query preview in exploratory search. *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '13)*, 243-252. **[ACM Digital Library]**

**April 21: Catch-up, Wrap-up, Review and The Future**

- Allan, J., Croft, B., Moffat, A., & Sanderson, M. (2012). Frontiers, challenges and opportunities for information retrieval: Report from SWIRL 2012. *SIGIR Forum, 46*(1), 2-32. Online: http://sigir.org/forum/2012J-TOC.html
- **DUE: Literature Review**

**May 4: Final Exam: 12:00-3:00 PM**