CIS 833: Information Retrieval and Text Mining

Fall 2015

Course Information & Syllabus

Instructor:
Doina Caragea
Email: dcaragea@ksu.edu
Office: Nichols 227C
Office hours: Wed 1:30pm-2:30pm, or by appointment

Meeting time and place: TU 10am-11:15pm, Nichols 236
Midterm: Week of October 12

Targeted audience: Graduate students from Computer Science and related areas.

Prerequisites: Basic knowledge on probability and statistics, data structures and algorithms. Prior knowledge of Java. The Yahoo! Hadoop implementation of MapReduce will be used for programming assignments. However, prior experience with Hadoop is not required.

Course Description: Information Retrieval (IR) refers to the processing, indexing and querying of unstructured or loosely structured information. This course will focus on the theory and practice of search engines for retrieving textual information (including web documents). Basic and advanced topics in IR will be covered, with emphasis on newer technologies that go beyond simple keyword search. Programming assignments will provide hands-on experience with retrieval systems. More advanced research in IR will be stimulated through the means of a class project. Given the need for processing large amounts of text data in information retrieval, the course will also cover MapReduce basics and algorithm design.

Course Objectives:

- Learn about traditional and advanced topics in IR. Understand the technologies underlying search engines, how they work and when they fail.
- Focus on intelligent techniques that go beyond simple keyword search. Learn how machine learning and text mining can help the information retrieval process.
- Gain practical experience in IR by implementing simple ?proof-of-concept? retrieval systems.
- Gain practical experience with Yahoo!'s Hadoop implementation of the MapReduce technology - programming assignments will make use of Hadoop.
- Identify active research topics in IR and study one in detail, as part of the class project.

Targeted Topics: MapReduce paradigm, efficient text indexing, latent semantic indexing, Boolean and vector space retrieval models, probabilistic retrieval models (binary independence models and language models), evaluation, web spiders, link analysis (PageRank and HITS), relevance feedback and query expansion, text mining and categorization, topic detection and clustering, query-answering, etc.

Recommended textbooks:

- *Introduction to Information Retrieval*, by Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze (available online at: http://www-nlp.stanford.edu/IR-book/).
- *Data-Intensive Text Processing with MapReduce*, by Jimmy Lin and Chris Dyer (available online at: http://lintool.github.io/MapReduceAlgorithms/).
- Part of the course will draw on material from other books and recent research papers.

Course Work and Evaluation: There will be one midterm exam and one final exam for the course. Students will be evaluated based on exams, homework assignments (both theory and programming), class participation, and a term project. Students are highly encouraged to attend every lecture and to participate in class discussion. All assignments will be submitted through KSOL. Assignments are due by 11:59pm on the due date (generally, a week after they are assigned, unless otherwise noted). Late submission is highly discouraged. I will accept late submissions at my discretion, but there might be grading penalties. The specific grading scheme is shown below:

Section	Weight
Written homework assignments	15%
Programming assignments	15%
Midterm exam	25%
Final exam	25%
Project	15%
Class Participation	5%

Collaboration Policies: Students are encouraged to discuss the course material, concepts, and assignments, but they should write their answers independently. Your submission should reflect your own knowledge and you should be able to reproduce the material you turn in at any time. Sharing answers will not be tolerated. Plagiarism will not be tolerated either. Appropriate citations for any external sources used in your work are mandatory. Never use sentences or phrases taken directly from a paper you are reviewing.

Expectations for classroom conduct: All student activities in the University, including this course, are governed by the Student Judicial Conduct Code as outlined in the Student Governing Association By Laws, Article VI, Section 3, number 2. Students who engage in behavior that disrupts the learning environment may be asked to leave the class.

Other Policies: No make-up exams and no incompletes, unless there is a very serious reason.

Students with Disabilities: Any student with a disability who needs an accommodation or other assistance in this course should make an appointment to speak with one of the instructors as soon as possible.

Honor System: Kansas State University has an Honor System [http://www.k-state.edu/honor/] based on personal integrity, which is presumed to be a sufficient assurance that, in academic matters, one's work is performed honestly and without unauthorized assistance. Undergraduate and graduate students, when they register, acknowledge the jurisdiction of the K-State Honor System. The policies and procedures of the Honor System apply to all full and part-time students enrolled in undergraduate and graduate courses on-campus, off-campus, as well as on-line. A component vital to the Honor System is the inclusion of the Honor Pledge, which applies to all assignments, examinations, and other course work undertaken by students:
"*On my honor, as a student, I have neither given nor received unauthorized aid on this academic work.*"