

CST 386/486-10 Information Retrieval (Spring 2009)

Time and location

- Monday, 6:00 pm - 8:30 pm
- Chicago Campus, GB 307

Instructor

Dr. Evgeny Dantsin

- Email: edantsin@roosevelt.edu
- Webpage: cs.roosevelt.edu/~dantsin/
- Office hours: Monday, 4:30 pm - 6:00 pm, GB 506B

Course description

Theory and practice of information retrieval with emphasis on applications to web search. The course covers traditional information retrieval topics (retrieval models, indexing, classification, clustering, etc) and more recent techniques (ranking of web pages, recommender systems, etc).

Textbook

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze.

Introduction to Information Retrieval.

Cambridge University Press, 2008. ISBN-10: 0521865719; ISBN-13: 978-0-521-86571-5

Prerequisites

CST 333/433 Database Systems

Tentative schedule

Date	Topics	Homework	Textbook
01/26	Class introduction. The Boolean retrieval model. Indexing. Conjunctive queries and intersection of postings lists.		1.1-1.5
02/02	Document representation: major steps. Skip pointers. Indexes for phrase queries. Biword indexes. Positional indexes. Proximity queries and proximity intersection.	HW-1	2.1-2.5
02/09	Dictionary data structures: hashing and trees. Wildcard queries. Permuterm indexes and k-gram indexes. Spelling correction. Edit distance and its computation. Jaccard coefficients. Phonetic correction and soundex reduction.		3.1-3.5
02/16	Index compression. Statistical properties of natural languages. Dictionary compression. Postings file compression.	HW-2	5.1-5.5
02/23	Metadata and weighted zones. Term weighting. The vector space model. Computing the cosine score.		6.1-6.3, 6.4.1
03/02	Faster computation of vector space scoring. Evaluation in information retrieval. Relevance feedback.	HW-3	7.1, 8.1, 8.3, 9.1.1
03/09	Text classification. Bayes' theorem and the naive-Bayes approach. The Bernoulli model. Classification in the vector space model.		13.1-13.3, 14.1-14.2

	Rocchio classification.		
03/23	Midterm exam (open books).		
03/30	Classification based on kNN. Linear and nonlinear classification. Clustering. K-means.	HW-4	14.3-14.4, 16
04/06	The Web as a graph. Size and structure of the Web graph. Link analysis and authorities of web pages. Ranking algorithms.		19.1-19.2, 19.5, 21.1
04/13	PageRank computation. HITS: hubs and authorities.	HW-5	
04/20	IR techniques and recommender systems. Decentralized search.		
04/27	Students' presentations and class discussions.		
05/04	Students' presentations and class discussions. Preparation for the final exam.		
05/11	Final exam (open books).		

Assignments and grading

Grades will be determined by the total number of points earned on the following assignments:

- homework assignments and in-class quizzes (total maximum: 50 points);
- midterm exam (maximum: 15 points);
- final exam (maximum: 15 points).

In addition to the above assignments, graduate students must select an additional topic in information retrieval (in conjunction with the instructor), research this topic, and make a class presentation (20 points). No late homework will be accepted; no make-ups will be given. Instances of academic dishonesty will be handled as described in University policies. Grades will be assigned according to the following scale (the plus/minus grading system is not used in this course):

A	B	C	D	F
≥ 90%	≥ 75%	≥ 60%	≥ 45%	< 45%

The last day to withdraw with a "W" grade is 04/06/09.

Lecture notes

Lecture notes, slides, homework assignments, and other course materials will be posted on [Blackboard](#) after every lecture.