

CSCE 5200: Information Retrieval and Web Search

Spring 2012

TTh 12:30PM-01:50PM

Syllabus

Instructor:	Rada Mihalcea
Office:	Research Park, F228, tel: 940-369-7630
Email:	rada at cs.unt.edu
Class hours:	TTh 12:30-01:50pm
Office hours:	Th 03:00-05:00pm or by appointment. Anytime electronically.
Teaching assistant:	Veronica Pérez-Rosas
Class web page:	http://www.cs.unt.edu/~rada/CSCE5200
Course description:	This course will cover traditional material, as well as recent advances in Information Retrieval (IR), the study of indexing, processing, and querying textual data. Basic retrieval models, algorithms, and IR system implementations will be covered. The course will also address more advanced topics in "intelligent" IR, including Natural Language Processing techniques, and "smart" Web agents.
Textbook:	Introduction to Information Retrieval (available online) <i>Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze</i>
Recommended readings:	Readings in Information Retrieval <i>K.Sparck Jones and P. Willett</i> Modern Information Retrieval <i>Ricardo Baeza-Yates and Berthier Ribeiro-Neto</i>
Course work and grading (tentative)	Homeworks: 30% Exam I: 20% Exam II: 20% Project: 25% Class participation: 5% Each assignment will specify the material to be turned in. All programs must be written in Perl. Assignments are due by 11:59pm on the due date. Assignments may be turned in up to 3 days late, with a penalty of 10% for each day late. No credit will be given after 3 days. There will be no final exam for this class. The final is replaced by the project.
Academic honesty:	Each program and homework assignment must be worked on individually. A submission carries with it an implicit statement that the submission is your own work. You may discuss the requirements and syntactical issues, but not solutions or designs. Violations may result in failure of the course.

Topics (tentative)	Course overview Introduction to IR models and methods Perl tutorial Text analysis / Web spidering Text properties Vector-based model Boolean model Probabilistic model; other IR models IR evaluation and IR test collections Relevance feedback, query expansion Web search: link based and content based Query-based and content sensitive link analysis Search engine technologies Text classification and clustering Question answering on offline and online collections Cross-language IR Personalized IR Web 2.0: wikis, blogs, etc.
-----------------------	--