

CS 5319 Syllabus

Topics in Language Processing: Search Engine Technologies

Spring 2009

Cross-listed as Special Topics in Computer Science 4390

Monday & Wednesday, 4:30—5:50, Computer Science room 321

Instructor: Nigel Ward
Office: Comp 206
Phone: 747-6827
E-mail: nigel@cs.utep.edu
Office Hours: Wednesdays 2:30-3:30, Fridays 1:30-2:30, and whenever my door is open, or by appointment

Course Web Site: <http://www.cs.utep.edu/nigel/search/>

Course Objectives: This class will cover the theory and practice of building search engines, including standard methods, advanced techniques, and emerging technologies. Students will acquire a solid understanding of the basic concepts, models and algorithms for information retrieval, and gain experience with common tools and techniques, including many applicable more widely. In the final project students will combine these in novel ways to prototype new functionality.

In addition, this course will give students hands-on experience in several practical skills of real-world value, including modeling, project management, high-level design, scripting, and system integration.

Format: This course will be primarily lecture-based, but with several class sessions devoted to in-class design exercises, student project presentations, and student-led discussions.

Textbook: *Introduction to Information Retrieval*. Christopher D. Manning, Prabhakar Raghavan and Hinrich Schuetze, Cambridge University Press. 2008. (henceforth IIR) Chapters tagged below with an s are to be skimmed. Advanced sections should generally be skipped.

We will cover about half of the material in the book. There will be additional readings; these will be handed out in class.

Grading: Approximate weighting:
assignments 45%
final 20%
tests 30%
quizzes and participation 5%

Assignments and tests will be challenging. Grading will be on a points-earned basis (points above zero), rather than a points-off basis (points below expectation), thus the mapping from points to letter grades will be non-standard.

Cooperation among students and among teams is encouraged, but not to the extent that it interferes with each individual's understanding or with learning-by-doing. Help given and received from other students and sources should be noted in the assignment write-up.

As always, students will follow the UTEP Standards of Conduct, available at <http://studentaffairs.utep.edu/Default.aspx?tabid=4386>. And, as always, if you have or suspect a disability and need accommodations you should contact the Disabled Student Services Office (DSSO) at 747-5148 or at dss@utep.edu or visit Room 106 Union East Building.

Students are expected to be punctual. Assignments due at the start of class will be collected after a one minute grace period; late assignments will receive at most two-thirds credit. Assignments are to be submitted as hardcopy unless otherwise specified.

No make-up exams or assignments will be given except under the conditions set forth in the Catalog. Students are free to attend class or not, bearing in mind that absence may annoy other students, interfere with learning, and result in a lower grade.

Students taking this as an undergraduate class will have a reduced set of assignments.

Important Dates:

Test 1	February 18
Spring Break	March 16-20
Test 2	March 30
Final Exam	May 11, 4-6:45

Likely Assignments

Likely Due Dates

A. Familiarization with an Open Source Search Engine (est. 8 hours)	Jan 28 / Feb 4
B. Evaluate Search Engine Results (2)	February 2
C. Search Engine Optimization (3)	March 7 / April 4
D. Observe a Search Engine User (2)	early March
E. Reverse-Engineer a Specialized Search Engine (4)	early April
F. Build a Meta-Search Engine (6)	mid February
G. Present a Paper (required for graduate students only) (5)	February - April
I. Project Concept (2)	February 11
J. Project Design (8)	March 9
K. Final Project (20)	March 23 / April 22
Mini-Assignment 1. Structured Search, using the Web of Science (1)	Jan 26
Mini-Assignment 2. The Google File System (1)	early March
Mini-Assignment 3. Query Expansion (1)	mid April
Mini-Assignment 4. A Question for the Final Examination (1)	May 6
Readings/Discussion Question Assignments (5)	February - April

Likely Topics

Chapters

- 1. Introduction** (1 lecture)
 - a. Course Overview
 - b. Search Engine Architecture Overview

- 2. Information Retrieval Basics** (4)
 - a. Boolean retrieval IIR 1
 - b. The Vector Model IIR 6, 14
 - c. Term Weighting “
 - d. Probabilistic Models IIR 11
 - e. Combination of Evidence, Meta-Search IIR 7

- 3. Web Search Basics** (4)
 - a. Characteristics of the Web IIR 19
 - b. User Tasks and User Characteristics “
 - c. Performance Metrics IIR 8
 - d. Link Analysis IIR 21

- 4. Implementation Issues** (5)
 - a. Crawling IIR 20
 - b. Document Purification IIR 2, 3
 - c. Data Structures (Professor Freudenthal) IIR 4, 5s
 - d. Caching and Parallelism “

- 5. Search in Context** (3) IIR 19
 - a. User Behavior, User Interfaces
 - b. Webmaster Behavior and Ethics
 - c. Commercial Considerations
 - d. Special-Purpose Search

- 6. Advanced Features** (3)
 - a. Personalized Search
 - b. Document Clustering (Professor Fuentes) IIR 16, 17s
 - c. Query Expansion IIR 9
 - d. Using Post-Search Behavior

- 7. Advanced Topics** (4)
 - a. Synonyms and Dimensionality Reduction IIR 18
 - b. Information Extraction and Semantic Search IIR 10
 - c. Recommendation Systems
 - d. Text Classification IIR 13s
 - e. Using Language Models IIR 12s
 - f. Speech Recognition and Audio Search
 - g. Machine Translation and Cross-Language Search

- 8. Projects** (4)
 - a. Aims and Scope
 - b. Design and Implementation
 - c. Evaluation

- 9. Review** (1)