

Course Code
COMP 4321

Course Title
Search Engines for Web and Enterprise Data

Course Description

Text retrieval models, vector space model, document ranking, performance evaluation; indexing, pattern matching, relevance feedback, clustering; web search engines, authority-based ranking; enterprise data management, content creation, meta data, taxonomy, ontology; semantic web, digital libraries and knowledge management applications. Prerequisite(s): COMP 2011 OR COMP 2012 OR COMP 2012H

List of Topics

1. Introduction and course overview	6. Document preprocessing
2. Business models	7. Query expansion and relevance feedback
3. Information retrieval models and Inverted Files	8. Machine learning for document ranking
4. Web-based information retrieval	9. Enterprise search
5. Retrieval effectiveness, benchmarking	10. Applications: text summarization

Textbook

C.D. Manning, R. Raghavan, and H. Schütze *Introduction to Information Retrieval*. Cambridge University Press, 2007.

- The pre-publication manuscript of the book and the lecture slides used in a Stanford course are [available online](#)

Reference books

C.J. van Rijsbergen *Information Retrieval*. 2nd Edition, Butterworth & Co (Publishers) Ltd, 1979. [Online Version](#).

[Web site for the textbook \[BR\] Baeza-Yates and Ribeiro-Neto](#). It does not have the book online, but it contains many useful resources and an errata.

[Web site for the reference book \[FB\] Bill Frakes and Ricardo Baeza-Yates](#). It does not have the book online, but it contains the source codes used in the book. The source code will be useful for your project.

R. Baeza-Yates and Berthier Ribeiro-Neto *Modern Information Retrieval*. Addison Wesley, Essex, England, 1999.

W.B. Frakes and R. Baeza-Yates (Eds.) *Information Retrieval: Data Structures and Algorithms*. Prentice-Hall, Englewood Cliffs, NJ, 1992.

G. Salton, *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley, Reading, MA, 1989.

G. Salton, and M.J. McGill, *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., New York, NY, 1983.

Grading Scheme

Item	Weight
Assignments	25%
Group Project	25%
Midterm Exam	15%
Final Exam	35%
Total	100%

Notes:

Course grades will normally fall within the following percentage bands:

A 15%
B 40%
C 40%
D/F 5%

There is no particular distribution within the subgrades of a grade but can be assumed to be equally divided.

Course Intended Learning Outcomes

On successful completion of this course, students are expected to be able to:

- (1) Design and implement a complete and functional search engine.
- (2) Test and evaluate the effectiveness of a search engine.
- (3) Identify the limitations of search engine technologies and develop solutions to meet application requirements.

Assessment Rubric

N/A