CMPE 59K Information Retrieval

There has been a striking growth in online text information such as web pages, news articles, e-mail messages, and scientific publications in the recent years. Developing tools for accessing, managing, and utilizing this huge amount of textual information is getting increasingly important. This course will cover the technology underlying the search engines, focusing on a wide range of topics including methods for processing, indexing, querying, and organizing textual data, as well as methods for web search, crawling, and link analysis.

Course Objectives:

- Understand how search engines work
- Learn to process, index, retrieve, and analyze textual data
- Learn to evaluate information retrieval systems
- Learn about web search, crawling and link analysis
- Build working systems that help users find useful information on the Web
- Learn about the state of the art in information retrieval research

Instructor: Arzucan Özgür

Textbook:
Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, *Introduction to Information Retrieval*, Cambridge University Press. 2008.
http://nlp.stanford.edu/IR-book/information-retrieval-book.html

Reference book (Optional):
Daniel Jurafsky and James H. Martin, SPEECH and LANGUAGE PROCESSING: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, Second Edition, 2008.

Web site: Course content is available at Moodle (http://moodle.cmpe.boun.edu.tr/). You should subscribe with your "boun" email address and the key provided in the first lecture.

Topics:

- Boolean model; text pre-processing; inverted indexes
- Approximate string matching and tolerant retrieval
- Index construction and compression
- Vector space model; text-similarity metrics; term weighting; ranked retrieval
- Evaluating information retrieval systems
- Relevance feedback; query expansion
- Language models for information retrieval
- Text classification and clustering
- Latent semantic indexing
- Web search and crawling
- Link analysis (e.g. hubs and authorities, Google PageRank)

Grading:

- 1 Paper presentation: 15%
- Paper discussions: 5%
- 2-3 Assignments: 20%
- Term Project: 35%
- 1 In-class Exam: 20%
- Class Participation: 5%