Cmpe 493 Introduction to Information Retrieval

There has been a striking growth in text data such as web pages, news articles, e-mail messages, social media data, 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 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.

Instructor: Arzucan Özgür, Office hours: Monday 14:00-15:00, Tuesday 13:00-15:00

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

Web site: Course content will be available at Moodle (the details will be provided).

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

Christopher D. Manning and Hinrich Schütze, *Foundations of Statistical Natural Language Processing*, MIT Press, 1999. http://nlp.stanford.edu/fsnlp/

Tentative List of 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)

Course Requirements:
The lectures will take place on Tuesdays between 16:00-17:00 and Wednesdays between 15:00-17:00. You are encouraged to attend and actively participate in the lectures.

The homework assignments will involve some programming where you will implement and test some of the techniques that we cover in class. You can use any programming language of your choice such as Java, Python, Perl, and etc.

The midterm and final exams will consist of problems covering the material in the lectures.

Each student (individually or in pairs) will be responsible for designing and completing a research project that demonstrates the ability to use concepts from the class. Each team will choose a project topic by selecting a recent scientific paper from an Information Retrieval conference or journal. The project will involve replicating the work done in the paper and proposing extensions/improvements to the existing work. The teams will give short project proposal and project final presentations in front of the class and submit a project report in the end of the semester describing the methods used to replicate the paper, the results obtained, the challenges encountered, and the proposed ways to improve/extend the current paper.

Grading:
- Midterm Exam: 15%
- Final Exam: 15%
- 3-4 Homework Assignments: 30%
- Project: 35%
- Class Participation: 5%