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 (arzucan.ozgur@boun.edu.tr)

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 and other textual resources

Web site: Course content will be available at Moodle.

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. (Draft of 3rd edition available at: https://web.stanford.edu/~jurafsky/slp3/)

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
- Probabilistic Models for information retrieval
- Text classification and clustering
- Latent semantic indexing

- Word Embeddings for IR
- Web search and crawling
- Link analysis (e.g. hubs and authorities, Google PageRank)

Course Requirements:
The lectures will take place on Mondays between 13:00-15:00 and Tuesdays between 11:00-12:00 through Zoom. You are encouraged to attend and actively participate in the lectures. The lecture videos will be recorded and posted on Moodle after the lecture.

The programming assignments will involve intermediate-level programming where you will implement and test some of the techniques that we cover in class using a programming language of your choice.

As term project, we will hold a shared task, where each team will design and implement a system related to IR, which will be evaluated at the end of the semester. The teams will give short project progress and project final presentations in front of the class describing the methods used, the results obtained, and the challenges encountered. Each team will consist of two people.

Grading:
- 4 or 5 Programming Assignments: 55%
- 2 Problem Sets: 10%
- Term Project: 35%