CS395T: Large-Scale Data Mining
Fall 2001

Class times: TTh: 11:00am-12:30 pm, Taylor Hall 2.106
Instructor: Inderjit Dhillon
Homepage: http://www.cs.utexas.edu/users/inderjit
E-mail: inderjit@cs.utexas.edu
Office: ACES 2.332, Ph: 471-9725
Office Hrs: Tue 2-3pm. Other times by appointment **only**.
Class URL: http://www.cs.utexas.edu/users/inderjit/courses/dm2001/

TA: Yuqiang Guan
E-mail: yguan@cs.utexas.edu
Phone: 232-7420
Office Hours: Thurs 1:30-3pm at PAI 3.04N

Prerequisites: Consent of the instructor. You are expected to know the basics (undergraduate level) of linear algebra and statistics.

Outline: Recent times have seen an explosive growth in the raw data available electronically. The data occurs in various forms, e.g., as text, image, video or numeric data; both in public domains and in private corporations. Data mining is the automatic discovery of interesting patterns and relationships in very large data sets. This graduate course will focus on scalable algorithms for data mining, with an emphasis on web mining. Topics covered will include (i) information retrieval — query retrieval, clustering and classification, (ii) link analysis on the internet (like Google), (iii) applications of machine learning algorithms to text and web mining, and bio-informatics. A basic knowledge of linear algebra and familiarity with basic methods such as clustering, classification, graph partitioning will be helpful.

A substantial portion of the class will be homeworks and a research project, where students will choose a well-defined problem. Emphasis will be on implementing algorithms that could lay the foundations of a usable system, such as, a text categorizer.

Grading:
10 + 30 pts: Class Project (First submission + Final submission).
40 pts: Homeworks.
10 pts: Presentation.
10 pts: Attendance and participation in class discussions.

References:
Pattern Classification by R. Duda, P. Hart and D. Stork, John Wiley and Sons, November 2000.
Foundations of Statistical Natural Language Processing by C. Manning and H. Schutze, MIT Press, 1999.
Modern Information Retrieval by R. Baeza-Yates and B. Ribeiro-Neto,ACM Press, 1999.

Disabilities statement: "The University of Texas at Austin provides upon request appropriate academic accommodations for qualified students with disabilities. For more information, contact the Office of the Dean of Students at 471-6259, 471-4641 TTY."