CIS 8538. Text Mining and Language Processing

INSTRUCTOR:
Dr. Alexander Yates

EDUCATIONAL OBJECTIVES

This course will cover a broad overview of problems and techniques in text mining and natural language processing. It will also provide in-depth coverage of the latest natural language processing research in selected topics.

The overview part of the course will cover problems in:

- **Text Mining**: Text classification, topic clustering, pattern learning, and the vector space model.
- **Natural Language Processing**: Part-of-speech tagging, parsing, partial parsing, meaning representations, semantic role labeling, semantic parsing, and pragmatics.
- **Information Retrieval**: Building indexes, data compression, representation of queries and documents, and similarity functions.
- **Information Extraction**: Building hierarchies of knowledge (ontologies), determining the relationships that exist between entities in text, and discovering implicit knowledge from text.

The in-depth part of the course will focus on the latest research in unsupervised information extraction. This part of the course will cover such techniques as pointwise mutual information, pattern-matching, bootstrapping, Hidden Markov Models, Conditional Random Fields, and language modeling techniques, among others.

PREREQUISITES

Math3031 (Introduction to Probability Theory) or Math3033 (Probability Theory for Computer Science), or an equivalent undergraduate-level understanding of probability, statistics, and linear algebra.

TEXT

The textbook for this course is not identified. The course will be structured extensively from the research literature. URLs for readings will be handed out during class.
GRADING

- In-class Participation: 10%
- Midterm: 20%
- Final Project: 30%
- Final Exam: 40%

EXAMS AND QUIZZES

All exams and quizzes are closed book. Their content is cumulative, i.e. they address the material from the entire semester up to the day of the exam. If a student misses the midterm for an emergency [as agreed with instructor], there will be no makeup exam: the quizzes and final project will become proportionally more important. If a student misses the midterm without previous agreement and without definitive proof as to the medical or legal reasons, he or she will get a zero for that exam. Quizzes that are missed will not be made up.

FINAL PROJECT

Several project ideas will be suggested during the course of the semester, but students are free to suggest their own, especially if they relate to their current research. Students will be expected to come up with innovative, novel solutions to problems in text mining and language processing.

Course projects will be undertaken individually or in small teams (2-3 students). Each student on a team will receive the same grade for the project; it is up to the team members to divide the work fairly.

OUTLINE

Introduction to Natural Language Processing (NLP)

Overview of the field; defining problems --- machine translation, question answering, information retrieval; history of approaches to NLP --- early logic-based approaches, heuristic techniques, statistical approaches, and approaches combining statistics and logic.

Statistical Techniques for NLP

Bayesian probability and statistics, directed and undirected graphical models, logistic regression, Hidden Markov Models, Conditional Random Fields, Expectation-Maximization, mixture models, statistical relational learning.

Text Mining
Text classification, topic clustering, pattern learning, and the vector space model.

**Information Retrieval**

Overview of a modern search engine, index construction, the vector space model, TF-IDF, methods for scoring and performing term weighting, pseudo-relevance feedback, collaborative filtering.

**Traditional Information Extraction**

Pattern matching, manual pattern declaration, pattern learning, and bootstrapping; named-entity recognition, binary relation extraction, and n-ary relation extraction; supervised semantic role labeling.

**Open Information Extraction**

Unsupervised named-entity extraction, set expansion, and class and attribute extraction; unsupervised, open-domain relation extraction; efficient processing of large corpora for unsupervised relation extraction; handling polysemy and synonymy.

**Advanced Topics in Information Extraction**

Sentiment extraction; cross-document coreference resolution; open semantic role labeling and generalization capacity of supervised NLP; open semantic parsing; Markov Logic models of text mining.