

Course Syllabus Part I DSC 550 - Data Mining

3 Credit Hours

Course Description

Data can often contain patterns and anomalies that only emerge at large scale. In this course, students explore techniques to mine and analyze large datasets to discover useful knowledge. Text mining, unstructured data, social networks, and other types of unsupervised data mining methods for data science are included.

Course Prerequisites

Recommend DSC 540

Course Objectives

Students who successfully complete this course should be able to:

1. Explain data mining techniques available for analyzing big data.
2. Transform data in preparation for data mining algorithms.
3. Construct natural language models for text analysis.
4. Recommend statistical approaches for evaluating big data.
5. Formulate implementation and automation strategies for data mining projects.

Grading Scale

93 – 100% = A	87 – 89% = B+	77 – 79% = C+	67 – 69% = D+
90 – 92% = A-	83 – 86% = B	73 – 76% = C	63 – 66% = D
	80 – 82% = B-	70 – 72% = C-	60 – 62% = D-
			0 – 59% = F

Topic Outline

- I. Text Preprocessing, Transformation, and Vectorization
 - a. Identifying paragraphs, sentences, and words
 - b. Segmenting text into tokens and n-grams
 - c. Feature extraction, transformation, and selection
- II. Text Classification
 - a. Naive Bayes
 - b. Logistic Regression
 - c. Cross-validation
 - d. Model validation and evaluation

- III. Handling Categorical Data, Text, Dates & Times
- IV. Topic Modeling and Document Similarity
 - a. Latent Dirichlet Allocation
 - b. Latent Semantic Analysis
- V. Context Aware Text Analysis
 - a. Part of speech (POS) tagging
 - b. Keyphrase extraction
 - c. Named-entity resolution
 - d. Word vectors
- VI. Graph Analysis
 - a. Social network analysis
 - b. Graph algorithms
 - c. Graph visualization
- VII. Unsupervised Learning
 - a. Principal Components Analysis
 - b. Hierarchical Clustering
 - c. Frequent Pattern Mining
 - d. Collaborative Filtering
- VIII. Model Evaluation and Selection
- IX. Real-World Implementation
 - a. Organizing a data mining project
 - b. Data infrastructure
 - c. Scaling
 - d. Ethics
- X. Advanced Topics in Data Mining
 - a. Neural networks and deep learning
 - b. Bayesian networks
 - c. Image, video, and speech processing