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 or equivalent

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. Model data mining problems and evaluate, visualize and communicate statistical models.
6. Perform systematic analysis of real world data mining problems end to end.
7. 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. 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. Privacy in Data Mining