# Text Mining and Classification

## Rationale

With the explosion of textual data on the world-wide web, text mining has become an important area of research. Text sources such as blogs, literature, social media, web pages and news articles can be analyzed to learn patterns, opinions, trends, and ideas. Text mining is a sub-area of data mining that deals with unstructured text. Algorithms have been developed for learning from unstructured text, and these often have practical applications in areas such as health care, advertising and homeland security.

## Description

Text mining can be defined as the process of finding or learning patterns from textual data to aid in decision making. This course will include the study of different representations of textual data and the algorithms used to glean new information from the data. It encompasses ideas from many other areas in computer science including artificial intelligence, machine learning, databases, information retrieval, and natural language processing.This class will primarily focus on the statistical methods for text mining, including machine learning techniques that are used to facilitate decision making.

## Topic List

The topic lists may include but is not limited to:

- Text Data Representation
  - Bag of Words
  - Named Entities
  - Relationships
- Text Categorization
  - Rule-based classifiers

  - Decision trees
  - Nearest neighbor
  - Maximum margin classifiers
  - Probabilistic classifiers
  - Semi-supervised Learning using EM
- Text Clustering
  - Hierarchical clustering
  - K-means clustering
  - Dimensionality Reduction
  - Latent semantic indexing
- Topic Modeling
  - pLSI
  - LDA
- Information Retrieval and Text Mining
  - Key Word Search
  - Indices
  - Link Analysis
  - Text Mining from Social Media
- Sentiment Analysis

# Learning Goals

Students should be able to:

- Demonstrate an understanding of the algorithms that were taught in class.
- Use current text mining software with practical, real-world data sets in a way that aids decision making.

## Assessment

Homework sets and a final exam with questions that target the learning goals will be used to assess student knowledge. In addition, a semester project will be used to assess student ability to use text mining packages, and to choose appropriate representations, algorithms, and testing methodology.