Overview and Objectives
This course introduces various methods for analyzing text data. In addition to introducing mathematical and computational techniques for representing text quantitatively, a key objective is to show how such techniques can be fruitfully applied to questions in economics, finance and political science. The lectures will be complemented by practical sessions in which students will build their own programs for analyzing real-world datasets.

Prerequisites
The programming language for the course will be Python, and the course will assume a basic initial familiarity with it. The course will also assume knowledge of machine learning ideas acquired in previous courses in the program.

Course Outline

Text Mining Basics
- Regular expressions
- Tokenizing, stemming and lemmatization, stop-word removal
- Unigrams and N-grams

Word-counting approaches
- Term-document matrix
- Dictionary methods
- Tf-idf weighting

Neural Embeddings
- Neural network structure and backpropagation
- Word2Vec / GloVe
- Going from word embeddings to document embeddings
- Starspace
Text Mining for Social Sciences

**14D010**

**Vector Space Model**
- Documents as vectors
- Cosine similarity

**Supervised Learning**
- Naive Bayes
- Support Vector Machines
- K nearest neighbors

**Unsupervised Learning: Latent Semantic Analysis**
- Polysemy and synonomy
- Singular value decomposition
- LSA and similarity

**Unsupervised Learning: Topic Modeling**
- Mixture models and the EM algorithm
- Mixed-membership modeling and Latent Dirichlet Allocation

**Variational Inference**
- Mean field estimation
- Application to Latent Dirichlet Allocation

**Applications in Social Sciences**
- A literature overview of text as data
- Applications in economics and finance
- Applications in political science
- Presentation of existing datasets that are based on text

**Case Studies in Detail**
- Conflict forecasting with news text
- Tourist warnings with news text
- Find the troll (final project discussions)
14D010
Text Mining for Social Sciences

Required Activities
20 hours of lecture, 5 hours of practical sessions. Problem sets with theoretical questions and coding exercises.

Evaluation
Exercises and final project. There is no final exam for this course.

Materials
Manning, Raghavan, and Schütze (2009), An Introduction to Information Retrieval. Cambridge University Press.
Original journal articles from computer science and finance/economics.
Lecture notes.