3 ECTS

14D010

Text Mining for Social Sciences

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 and finance. 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

Python Basics

- Basic data types and syntax
- Tour of the standard library
- Tour of relevant scientific computing and text processing packages

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

3 ECTS

14D010

Text Mining for Social Sciences

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

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

14D010

Text Mining for Social Sciences

3 ECTS

Manning, Raghavan, and Schütze (2009), *An Introduction to Information Retrieval.* Cambridge University Press.

McKinney (2012), *Python for Data Analysis*. O'Reilly.

Murphy (2012), *Machine Learning: a Probabilistic Perspective.* MIT Press.

Original journal articles from computer science and finance/economics.

Lecture notes.