Text Mining : Models and Algorithms

Winter Term - 6 ECTS

Mandatory Course

Prerequisites to Enrol

Python programming, Good understanding of Statistics and Econometric Concepts

Overview and Objectives

Text is complex data which is often available in abundance inside firms and public organizations. This course teaches how to use text to generate data. It starts with the collection of text from text documents like pdfs or webpages and ends with the actual use of data generated from text in different case studies. Particular focus will be on the different statistical models which allow the students to extract features from the text reaching from simple dictionary models to complex models like the latent dirichlet allocation (LDA) model. The lectures will be complemented by practical sessions in which students will build their own programs for analyzing real-world datasets.

Prerequisite reading / requirements

To be announced.

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

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

Class attendance, Completion of Homeworks, Exam

Evaluation

Examples: Exam (80%), homework (20%)

Competences

Be able to treat text as data.

Be aware of the many ways in which features can be extracted from the text and know which one is the best way for your decision making problem.

Be able to implement a data pipeline which goes from scraping the raw text, pre-treating it, extracting text features and using them as data.

Learning Outcomes

Programming python codes that scrape text from webpages.

Program python code that reads masses of text documents (csv, pdf, doc).

Use python to clean the text and make sure it is read in correctly.

Use python to generate a document term matrix from the cleaned text.

Use python to exclude stop words or rare words.

Lemmatize text in different languages.

Implement dictionary methods in python.

Implement weightings like tf/idf in python.

Implement naive bayes, support vector machines and k-nearest neighbors with text data in python.

Implement LSA with text data in python.

Implement simple mixture models using the EM algorithm in Python.

Implement mixed membership models like the LDA in Python.

Know how to use the data generated from text to analyse the text.

Text Mining : Models and Algorithms

Winter Term - 6 ECTS

Mandatory Course

Materials

Gentzkow, Matthew, Bryan T. Kelly, and Matt Taddy (2019) Text as Data. Journal of Economic Literature. Forthcoming.

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.