14D010

Text Mining for Social Sciences

Spring Term - 3 ECTS

Elective Course

Prof. Hannes Mueller
Prof. Ruben Durante
Prof. Nandan Rao

Prerequisites to Enroll

Python basic level

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. Lectures will be complemented with hands-on exercises, working with text data in Python. Time will be balanced between A) theory and techniques of text mining and B) case studies from contemporary research in the social sciences.

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

Information Retrieval

- Word count methods
- TF-IDF weighting
- Cosine similarity

Text Preprocessing

- Scikit-learn interfaces
- Tokenizing, stemming and lemmatization, stop-word removal
- Regular expressions

Supervised Learning

- Naive Bayes vs. logistic regression

14D010

Text Mining for Social Sciences

Spring Term - 3 ECTS

Elective Course

Prof. Hannes Mueller
Prof. Ruben Durante
Prof. Nandan Rao

- Generative vs. discriminative models
- Sentiment analysis
- Domain transfer

Continuous Metric Spaces

- Latent Semantic Analysis
- Neural network structure and backpropagation
- Neural embeddings - Word2Vec / GloVe
- Going from word embeddings to document embeddings
- Transformers / BERT

Topic Modeling

- Mixed-membership modeling and 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
- Analysis of Political Influence Campaigns (Trolls)
- Detecting who copies whom in news reporting
- Analysis of politicians speeches and judges' sentencing decisions

Required Activities

20 hours of lecture, 5 hours of practical sessions. Problem sets with coding exercises.

14D010

Text Mining for Social Sciences

Spring Term - 3 ECTS

Elective Course

Prof. Hannes Mueller
Prof. Ruben Durante
Prof. Nandan Rao

Evaluation

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

Competences

☒ Construct a global vision of the situation of the problem based on knowledge of the synergies between advanced statistical methods, computing and business analysis to generate added value.

☐ Modeling and predicting high-dimensional data with advanced statistical methods in the field of data science in order to improve strategic decision making.

☒ Apply the knowledge of programming languages, computer programs and advanced services in the Cloud to solve the problems that are presented to the data scientist.

☒ Solve the real problems that arise in the fields of study through the accurate analysis of the data.

☐ Visualize and interact with high-dimensional data in order to contextualize the information and facilitate subsequent decision-making.

☒ Communicate with conviction in English the results and implications of the required analytical study using a language related to the receiver.

☐ Work in a heterogeneous team of researchers in the field of the economic analyst using specific group techniques.

☐ Make use of personal data knowing the limits of it, its legal consequences and the practical repercussions of it.

☒ Own and understand knowledge that provides a basis or opportunity to be original in the development and / or application of ideas, often in a research context.

☒ That students know how to apply the acquired knowledge and their ability to solve problems in new or unfamiliar environments within broader (or multidisciplinary) contexts related to their area of study.

☐ That the students be able to integrate knowledge and face the complexity of making judgments based on information that, being incomplete or limited, include reflections on the social and ethical responsibilities linked to the application of their knowledge and judgments.

☒ That the students know to communicate their conclusions and the knowledge and last reasons that sustain them to specialized and non-specialized publics in a clear and unambiguous way.

☒ That students have the learning skills that allow them to continue studying in a way that will be largely self-directed or autonomous.

Learning Outcomes

☒ Apply mathematical and computational analysis of social, business and economic networks knowing the theory and optimization algorithms.

☐ Work with databases and cloud computing.

☒ Model Big Data information using data mining techniques.

☒ Visually display Big Data information using data mining techniques.

☒ Work with Big Data information using data mining techniques.

☒ Express in computer language the resolution of complex problems with high-dimensional data.

☐ Apply mathematical and statistical analysis using economic theory in complex problems with high-dimensional data.

☒ Create visualizations of information according to each type of data.

☐ Sort the information in a visual and understanding mode from the selection and qualification of the data.

☒ Treat high-dimensional data environments knowing their limitations and how to present the results.

☐ Present information visually and in an orderly manner to improve decision making.

☐ Answer the question "And then what do we do?" Based on the information obtained and presented.

☐ Collaborate in a computing environment that requires structuring and planning.

☒ Apply mathematical theory and statistics on data sets from disparate disciplines.

14D010

Text Mining for Social Sciences

Spring Term - 3 ECTS

Elective Course

Prof. Hannes Mueller
Prof. Ruben Durante
Prof. Nandan Rao

☐ Know the restrictions and considerations of the use of personal data in relation to the Organic Law of Data Protection.

14D010

Text Mining for Social Sciences

Spring Term - 3 ECTS

Elective Course

Prof. Hannes Mueller
Prof. Ruben Durante
Prof. Nandan Rao

Materials

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.

Cage, Julia, Nicolas Herve, and Marie-Luce Viaud (2017) The Production of Information in an Online World: Is Copy Right? CEPR Discussion Papers DP12066.

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

Hansen, Stephen, Michael McMahon and Andrea Prat (2018) Transparency and Deliberation within the FOMC: a Computational Linguistics Approach. Quarterly Journal of Economics, 133 (2).

Ash Elliot, Daniel Chen, and Suresh Naidu (2019) Ideas Have Consequences: The Impact of Law and Economics on American Justice, Working Paper.

Kelly Bryan T., Dimitris Papanikolaou, Amit Seru, and Matt Taddy (2020), Measuring Technological Innovation over the Long Run, NBER Working Paper No. 25266.

Original journal articles from computer science and finance/economics.

Lecture notes.

Gentzkow Matthew and Jesse Shapiro (2010), What Drives Media Slant, Econometrica, vol. 78, n.1., pp. 35-71.

Hassan, Tarek A., Stephan Hollander, Laurence van Lent and Ahmed Tahoun (2017) Firm-Level Political Risk: Measurement and Effects. R&R QJE.

Mueller, Hannes, and Christopher Rauh (2018) Reading Between the Lines: Prediction of Political Violence Using Newspaper Text. American Political Science Review. https://doi.org/10.1017/S0003055417000570

Vegard H. Larsen and Leif Anders Thorsrud, (2015) The Value of News. Working Papers No 6/2015. Centre for Applied Macro- and Petroleum economics (CAMP).