57366-01 Methods for Unstructured Data, 3 KP
Graduate School of Business and Economics, Fall Term 2022

Dates and places

Tuesday	11.10.2022	08.30-12.00	PC-Lab S18 HG.37
Wednesday	19.10.2022	08.15-12.00	PC-Lab S18 HG.37
Thursday	20.10.2022	12.15-18.00	PC-Lab S18 HG.37
Wednesday	26.10.2022	08.15-12.00	PC-Lab S18 HG.37
Thursday	27.10.2022	12.15-18.00	PC-Lab S18 HG.37
Thursday	03.11.2022	12.15-18.00	PC-Lab S18 HG.37

Course website

All class material will be uploaded here:
https://github.com/hliebert/course-unstructured-data

Course description

Much of human knowledge is stored in unstructured formats. This course teaches methods to process and analyze unstructured data, focusing on text data. In the first part, we review tools required for processing text data. One lecture is dedicated to web scraping fundamentals. We then focus on different representation concepts for text data and study supervised models suited for text data, as well as unsupervised models which make it possible to discover structure in unlabeled data. In the last part we study vector space representations and distributional language models. Throughout the course, I try to emphasize real-world applications of the techniques in research and industry.

Course objective

A thorough understanding of the workflow, tools and models related to the analysis of text data.

Course outline

1. PC fundamentals
2. Regular expressions and pattern recognition
3. Web scraping
4. Representing text as data
5. Supervised models for text data
6. Unsupervised models for text data
7. Information retrieval and distributional language models

Prerequisites

Knowledge of graduate level statistics and econometrics. A basic understanding of predictive modeling concepts (e.g. a class on computational statistics) and basic knowledge of R are helpful, but not required.

PC lab material

Lab sessions will be in R. All tutorial material will be accessible online and run self-contained on Binder from any device. Students are encouraged to set up and work on their personal computers to familiarize themselves with the tools introduced in class.

References and material

The course does not adhere strictly to a single reference. References are pointed out in the course material. A preliminary list of general references is given below.

Gentzkow, M., Kelly, B., and Taddy, M. (2019). Text as Data. *Journal of Economic Literature* 57 (3), 535–574. https://doi.org/10/gf7rd5.

Jurafsky, D. and Martin, J. H. (2019). *Speech and Language Processing* (3rd ed. draft). https://web.stanford.edu/~jurafsky/slp3/.

Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer, New York. https://web.stanford.edu/~hastie/ElemStatLearn/.

Shotts, William E. (2019). *The Linux command line: a complete introduction*. Second edition. San Francisco: No Starch Press.