Text analysis in R
(Helge Liebert)

Course description
Much of human knowledge is stored in unstructured formats. Processing and analyzing unstructured text data is an elementary part of both research in modern social science and data science in industry.

This course teaches methods to process and analyze unstructured text data. The first part of the lecture reviews tools and concepts for processing text data and introduces the fundamentals. The second part focuses on different representation concepts underlying the transformation of unstructured text data into structured formats suited for statistical analysis. The last part introduces statistical models suited for the analysis of text data, focusing on both supervised models for prediction as well as unsupervised models which make it possible to discover structure in unlabeled text data.

Throughout the course, I try to emphasize real-world applications of the techniques in research and industry. The methods taught in class are applied to example data sets using the statistical software R. All class material will be provided on a dedicated website.

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

Course outline
1. Introduction
2. Regular expressions and pattern matching
3. Representing text as data
4. Supervised models for text data
5. Unsupervised models for text data
6. Information retrieval and distributional language models

Prerequisites
Knowledge of undergraduate statistics and econometrics. Basic knowledge of R. A basic understanding of predictive modeling concepts (e.g., a class on computational statistics) is helpful, but not required.

Dates
February 15-17, 2021, 8.30-16.30

Evaluation
Data analysis assignment and report.

Course website
All class material and links to online tutorials are available here.
https://github.com/hliebert/course-text-analysis-in-r

PC lab material
Lab session will be in R (pointers to equivalent Python libraries and some code can be made available for those that are interested). Tutorial material will be accessible online and run self-contained on Binder from any device.

References

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/.