Computational Text Analysis Workshop

Instructor: Dr Theresa Gessler (University of Zurich)
Organising professor: Ellen M. Immergut
Administrative Assistant: Adele Battistini
Credits: 10

Course summary

Over the past years, the availability of new data through the digitalization of legal, political, journalistic corpora as well as the growth of online sources has allowed researchers to answer new research questions across political science.
The aim of this course is to introduce students to the quantitative analysis of textual data. We will cover both applications in recent empirical research and the implementation of text analysis techniques through hands-on experiences using the R statistical programming language.
The course will cover the collection of text data with web scraping techniques, text preprocessing, dictionaries and descriptive analysis of texts, as well as supervised and unsupervised learning methods to classify the content of text corpora.

Dates and format

- 18/11 9-11h
- 01/12 14-16h
- 08/12 14-16h
- 15/12 14-16h
- 22/12 14-16h (*to be discussed, may be postponed to January*)

All classes will be held via Zoom

Requirements

After each session, you will receive an exercise sheet with tasks based on the previous session. These will be accompanied by five lab sessions (dates and times to be determined in the first class). For credit, all classes and labs should be attended, and exercise sheets completed.

To follow the course, you will require an R installation (possibly with RStudio). A list of packages to be installed will be provided closer to the date but you may want to prepare by installing tidyverse, quanteda, rvest, caret and stm.

Related readings

These readings are by no means required for the course - they are meant to give you a starting point for getting into text analysis and finding interesting applications in your field.

• Grimmer, J., and B. M. Stewart. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." Political Analysis 21, no. 3 (July 1, 2013): 267–97. https://doi.org/10.1093/pan/mps028.

• Schoonvelde, Martijn, Gijs Schumacher, and Bert N. Bakker. "Friends With Text as Data Benefits: Assessing and Extending the Use of Automated Text Analysis in Political Science and Political Psychology." Journal of Social and Political Psychology 7, no. 1 (February 8, 2019): 124-143–143. https://doi.org/10.5964/jspp.v7i1.964.

• Gentzkow, Matthew, Bryan Kelly, and Matt Taddy. "Text as Data." Journal of Economic Literature 57, no. 3 (September 1, 2019): 535–74. https://doi.org/10.1257/jel.20181020.

• Atteveldt, Wouter van, and Tai-Quan Peng. "When Communication Meets Computation: Opportunities, Challenges, and Pitfalls in Computational Communication Science." Communication Methods and Measures 12, no. 2–3 (April 3, 2018): 81–92. https://doi.org/10.1080/19312458.2018.1458084.

• Denny, Matthew James, and Arthur Spirling. "Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It." SSRN Scholarly Paper. Rochester, NY: Social Science Research Network, January 25, 2017. https://papers.ssrn.com/abstract=2849145.

• James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani, eds. An Introduction to Statistical Learning: With Applications in R. Springer Texts in Statistics 103. New York: Springer, 2013.

• Monroe, B. L., and P. A. Schrodt. "Introduction to the Special Issue: The Statistical Analysis of Political Text." Political Analysis 16, no. 4 (October 4, 2008): 351–55. https://doi.org/10.1093/pan/mpn017.

• Fréchet, Nadjim, Justin Savoie, and Yannick Dufresne. "Analysis of Text-Analysis Syllabi: Building a Text-Analysis Syllabus Using Scaling." PS: Political Science & Politics, undefined/ed, 1–6. https://doi.org/10.1017/S1049096519001732.

• Munzert, Simon, Christian Rubba, Peter Meißner, and Dominic Nyhuis. Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining. Chichester, West Sussex, United Kingdom: John Wiley & Sons Inc, 2015.

• Benoit, Kenneth, Kohei Watanabe, Haiyan Wang, Paul Nulty, Adam Obeng, Stefan Müller, and Akitaka Matsuo. "Quanteda: An R Package for the Quantitative Analysis of Textual Data." Journal of Open Source Software 3, no. 30 (October 6, 2018): 774. https://doi.org/10.21105/joss.00774. Roberts, Margaret E., Brandon M. Stewart, and Dustin Tingley. "Stm: R Package for Structural Topic Models." Journal of Statistical Software, 2013.