Essex Summer School 2024, Quantitative Text Analysis

- Instructor: Dr. Martijn Schoonvelde
  - martijn.schoonvelde@rug.nl
  - Office hours: by appointment (via Zoom)
  - Meetings: Daily 09:50am–1:40pm (BST)

Course introduction

With the massive availability of text data on the web, social scientists increasingly recognize quantitative text analysis (or "text as data") as a promising approach for analyzing various kinds of social and political phenomena. This course introduces participants to a variety of its methods and tools. We discuss their underlying theoretical assumptions including various ways of representing text, substantive applications of these methods, and their implementation in the R statistical programming language. The meetings – which combine lectures and coding sessions – will be hands-on, dealing with practical issues in each step of the research process.

Learning Outcomes

Participants will understand fundamental concepts in quantitative text analysis research design such as inter-coder agreement, reliability, validation, accuracy, and precision. Participants will learn to convert texts into informative feature matrices and to analyze those matrices using statistical methods. Participants will learn to apply these methods to a text corpus in support of a substantive research question. Furthermore, participants will be able to critically evaluate (social science) research that relies on quantitative text analysis methods.

Participation and communication

I expect that you come to our meetings prepared, having read required papers, and ready to discuss your questions, criticisms and thoughts. To facilitate communication and interaction we will make use of a dedicated Slack channel at `https://essqta24.slack.com` for which I will send you an invitation via email.

Literature

Throughout the course we will read papers from political science and other social sciences. For further background on quantitative text analysis and natural language processing I recommend the following books:

- Daniel Jurafsky and James H. Martin (2020). Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. 3rd edition. `https://web.stanford.edu/~jurafsky/slp3/`

- Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze (2008). An Introduction to Information Retrieval. New York: Cambridge University Press. `https://nlp.stanford.edu/IR-book/information-retrieval-book.html`
- Grimmer, J., Roberts, M.E. and Stewart, B.M. (2022). Text as data: A new framework for machine learning and the social sciences. Princeton University Press.
- van Atteveldt, W., Trilling, D. and Calderon, C.A., (2022). Computational Analysis of Communication. John Wiley & Sons. `https://cssbook.net/`

Software

In this module we will use R. Students will need to have R and RStudio installed on their computers / laptops. Students who have not used R at all are advised to work their way through the beginner courses listed on `https://education.rstudio.com/`. Another fantastic resource is R for Data Science by Hadley Wickham and Garett Grolemund. This book is available at `https://r4ds.had.co.nz/`.

Course outline

** This outline serves a general plan for the course; deviations (announced) may be necessary. To keep the workload manageable we'll stick to 2 readings a day but during our meetings (sometimes 3) we'll discuss other papers as well.*

Day 1 – 9 July

- What is quantitative text analysis? What will you learn in this course? Developing a corpus.
  - **Required reading**:
    - Benoit (2020). Text as Data: An Overview. Handbook of Research Methods in Political Science and International Relations. Ed. by L. Curini and R. Franzese. Thousand Oaks: Sage: 461–497.
    - Grimmer, J., Roberts, M.E., and Stewart, B.M. (2022). Text as Data: A New Framework for Machine Learning and the Social Sciences. Princeton: Princeton University Press: chapter 4.

Day 2 – 10 July

- Core assumptions in quantitative text analysis. Representations of text. Preprocessing and feature selection.
  - **Required reading**:
    - Baden, C., Pipal, C., Schoonvelde, M. & van der Velden, M.A.G., (2022). Three Gaps in Computational Text Analysis Methods for Social Sciences: A Research Agenda. *Communication Methods and Measures, 16(1)*: pp. 1–18.
    - Benoit, K., Watanabe, K., Wang, H, Nulty, P., Obeng, A., Müller, & Matsuo, A. (2018). Quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software, 3(30)*, 774.

Day 3 – 11 July

- Advanced text representations. Word embeddings.
  - **Required reading**:
    - Grimmer, J ., Roberts, M.E., and Stewart, B.M. (2022). Text as Data: A New Framework for Machine Learning and the Social Sciences. Princeton: Princeton University Press: chapter 8.
    - Rodriguez, P.L. and Spirling, A., (2022). Word embeddings: What works, what doesn't, and how to tell the difference for applied research. *The Journal of Politics, 84(1)*: pp.101–115.

Day 4 – 12 July

- What can we do with dictionaries and how can we validate them? Sensitivity and specificity.
  - **Required reading**:
    - Grimmer, J.., Roberts, M.E., and Stewart, B.M. (2022). Text as Data: A New Framework for Machine Learning and the Social Sciences. Princeton: Princeton University Press: chapter 16.
    - Rauh, C., 2018. Validating a sentiment dictionary for German political language–a workbench note. *Journal of Information Technology & Politics, 15(4)*: pp.319–343.

Day 5 – 15 July

- Human coding (or machine coding) and document classification using supervised machine learning. Evaluating a classifier.
  - **Required reading**:
    - Daniel Jurafsky and James H. Martin (2020). Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. 3rd edition: Chapter 4
    - Gilardi, F., Alizadeh, M., & Kubli, M. (2023). "ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks". *Proceedings of the National Academy of Sciences of the United States of America 120 (3):* e2305016120.

Day 6 – 16 July

- Supervised, semi-supervised and unsupervised approaches to place text on an underlying dimension.
  - **Required reading**:
    - Schwemmer, C. and Wieczorek, O., (2020). The methodological divide of sociology: Evidence from two decades of journal publications. *Sociology, 54(1)*: pp.3–21.
    - Watanabe, K., (2021). Latent semantic scaling: A semisupervised text analysis technique for new domains and languages. Communication Methods and Measures, 15(2), pp.81–102.

Day 7 – 17 July

- Understanding topic models. Discussing their pros and cons.
  - **Required reading**:

- Roberts, M et al. (2014). Structural topic models for open-ended survey responses. *American Journal of Political Science, 58(4)*, 1064–1082.
- Grimmer, J.., Roberts, M.E., and Stewart, B.M. (2022). Text as Data: A New Framework for Machine Learning and the Social Sciences. Princeton: Princeton University Press: chapter 13.

Day 8 – 18 July

- New developments in data. Machine translation. Automated speech recognition. Images as data.
  - **Required reading**:
    - Proksch, S.O., Wratil, C. and Wäckerle, J., (2019). Testing the validity of automatic speech recognition for political text analysis. *Political Analysis*, 1–21
    - De Vries, E., Schoonvelde, M. & Schumacher, G., (2018). No longer lost in translation: Evidence that Google Translate works for comparative bag-of-words text applications. *Political Analysis, 26(4)*, 417–430.
    - Schwemmer, C., Unger, S. and Heiberger, R., (2023). Automated image analysis for studying online behaviour. In: *Research Handbook on Digital Sociology.*

Day 9 – 19 July

- Deep learning. Transfer Learning. LLMs. Concluding remarks
  - **Required reading**:
    - Laurer, M., Van Atteveldt, W., Casas, A. & Welbers, K., (2024). Less annotating, more classifying: Addressing the data scarcity issue of supervised machine learning with deep transfer learning and BERT-NLI. *Political Analysis, 32(1)* pp. 84–100.
    - Chan, C.H., (2023). grafzahl: fine-tuning Transformers for text data from within R. *Computational Communication Research, 5(1)* p.76–84.
    - Bail, C.A., (2024). Can Generative AI improve social science? *Proceedings of the National Academy of Sciences, 121(21)* p.e2314021121.