# Essex Summer School 2022, Quantitative Text Analysis


– Instructor: dr Martijn Schoonvelde
  – martijn.schoonvelde@rug.nl
  – Office hours: by appointment via Zoom
  – Meetings: Daily 10:00am–1:30pm (BST)


## Course introduction

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

## Learning Outcomes

Participants will understand fundamental issues 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 uses automated 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 `essqta2022.slack.com` for which I will send you an invitation via email.

## Literature

We do not use a textbook in this course, but we will 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 who have not used R at all are advised to work their way through one of the free resources that are listed on https://www.rstudio.com/online-learning/#R. Another good resource is Quantitative Politics with R, developed by Erik Gahner Larsen and Zoltán Fazékas: http://qpolr.com/index.html. Students will need to have R and RStudio installed on their computers / laptops or they can work on a browser-based R session via R Studio Cloud: https://rstudio.cloud/. Another fantastic resource is R for Data Science bu 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 two readings a day but during our meetings we'll discuss other papers as well.*

## Day 1 – 11 July

- What is quantitative text analysis? What will you learn in this course?
  - **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., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis, 21(3)*, 267–297.

## Day 2 – 12 July

- Core assumptions in quantitative text analysis. Considering issues of measurement and validation.
  - **Required reading**:
    - Welbers, K., Van Atteveldt, W., & Benoit, K. (2017). Text analysis in R. *Communication Methods and Measures, 11(4)*, 245–265.
    - 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.

## Day 3 – 13 July

- Going from text to data. Preprocessing and feature selection. Deciding on the unit of observation and unit of analysis.
  - **Required reading**:
    * Denny, M.J. & Spirling, A., (2018). Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it. *Political Analysis, 26(2)*: pp.168–189.
    * 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 4 – 14 July

- Comparing documents in a corpus. Combining linguistic features and social science theories.
  - **Required reading**:
    * Peterson, A. & Spirling, A., (2018). Classification accuracy as a substantive quantity of interest: Measuring polarization in Westminster systems. *Political Analysis, 26(1)*: pp. 120–128.
    * Cross, J. & Hermansson, H., (2017). Legislative amendments and informal politics in the European Union: A text reuse approach. *European Union Politics, 18(4)*: 581–602.

## Day 5 – 15 July

- What are dictionaries and how can we validate them? Sensitivity and specificity.
  - **Required reading**:
    * Pennebaker J. & King, L. (1999) Linguistic styles: language use as an individual difference. *Journal of Personality and Social Psychology, 77(6)*, 1296–1312.
    * Young, L., & Soroka, S. (2012). Affective news: The automated coding of sentiment in political texts. *Political Communication, 29(2)*, 205–231.
  - **Coding Assignment 1 Due**

## Day 6 – 18 July

- Human coding and document classification using supervised machine learning.
  - **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
    * Müller, S., (2020). "Media coverage of campaign promises throughout the electoral cycle." *Political Communication*: 1–23.

## Day 7 – 19 July

- Supervised, semi-supervised and unsupervised approaches to place text on an underlying (political) dimension.
  - **Required reading**:
    * Slapin J. & Proksch S. (2008). A scaling model for estimating time-serial positions from texts. *American Journal of Political Science 52*, 705–722.
    * Hjorth, F., Klemmensen, R., Hobolt, S., Hansen, M. E., & Kurrild-Klitgaard, P. (2015). Computers, coders, and voters: Comparing automated methods for estimating party positions. Research & Politics, 2(2).

## Day 8 – 20 July

- Understanding topic models. What are their pros and cons?
  - **Required reading**:
    * Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM, 55(4)*, 77–84.
    * Roberts, M et al. (2014). Structural topic models for open-ended survey responses. *American Journal of Political Science, 58(4)*, 1064–1082.

## Day 9 – 21 July

- New developments in data. Images as data. Automated speech recognition. Machine translation.
  - **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.
    * Casas, A. & Williams, N.W., (2019). Images that matter: Online protests and the mobilizing role of pictures. *Political Research Quarterly, 72(2)*: 360–375.
  - **Coding Assignment 2 Due**

## Day 10 – 22 July

- Word embeddings. Concluding remarks.
  - Rudkowsky, E., Haselmayer, M., Wastian, M., Jenny, M., Emrich, Š. & Sedlmair, M., (2018). More than bags of words: Sentiment analysis with word embeddings. *Communication Methods and Measures, 12(2-3)*, 140–157.
  - Rodman, E., (2020). A Timely Intervention: Tracking the Changing Meanings of Political Concepts with Word Vectors. *Political Analysis, 28(1)*: pp. 87–111.