Computational social sciences

Malo Jan & Luis Sattelmayer

This course is an introduction to *Computational Social Sciences* (CSS). Over the last two decades, the amount of data available in text format, the methods and computational power available to analyze it, have drastically increased. This course aims to introduce participants to the use of computational methods to answer questions of the social sciences. The objective is to demystify their complexity and to show how social scientists, both qualitative and quantitative, can take advantage of data collected from the internet, and text analysis methods. The class covers a variety of different methods used in CSS, including web scraping, text mining, topic modelling, word embeddings, and supervised machine learning using Transformer models and LLMs.

Objectives

This class has three main goals. First, we want to help students understand how to address data limitations by automatically collecting textual data and choosing the right corpora for their research interests. Second, we aim to expose students to various practical methods for analyzing text quantitatively, enabling them to conduct their own research and potentially prepare for more extensive projects like a Master's thesis. Lastly, we hope that students will improve their programming skills and grasp quantitative reasoning, which can be applied to handle different types of data, not just text.

The course will take place over a span of five days and integrates lectures with hands-on lab-sessions to apply the methods. Each morning session, lasting for two hours, is dedicated to presenting comprehensive content on the methods employed. This includes an exploration of the underlying logic, advantages, disadvantages, and typical applications of the methods discussed. In the afternoon, the first session will focus on demonstrating the practical application of these methods. The final session of each day is designed to actively engage students in applying the knowledge they have acquired throughout the day.

Pre-requisites

The course requires basic knowledge of RStudio. In terms of coding skills, this course picks up where the RStudio lab sessions for the Quantitative Methods II of the School of Research lecture ended. The final day introduces students to a text analysis application in Python, but no knowledge of this language is required.

Day 1 : Computational social sciences and webscraping

10h-12h : Introduction to Computational social sciences

To begin with, the instructors will present the objectives of the course and what CSS are. An overview of the methods and their different objectives will be presented, as well as existing applications in political science.

13h30-14h45 : Introduction to webscraping

In this session, students will be introduced to web scraping for collecting data on the web. The course will cover the logic of the method, the ethics of web scraping and its application in RStudio with the `rvest` package.

15h15-16h45 : Application in RStudio

References

- Edelmann, A., Wolff, T., Montagne, D., & Bail, C. A. (2020). Computational social science and sociology. Annual Review of Sociology, 46, 61-81.
- 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.
- Grimmer, J., Roberts, M. E., & Stewart, B. M. (2021). Machine learning for social science: An agnostic approach. Annual Review of Political Science, 24, 395-419.

Day 2 : Manipulating text

10h-12h : Transforming text to numbers

This session is designed to familiarize participants with the process of converting textual information into numerical representations. The lecture covers essential concepts in Natural Language Processing, including word representations, tokenization, document-term matrices, and tf-idf. Students will gain insights into word counts, dictionary methods, and their associated limitations.

13h30-14h45 : Manipulating text in RStudio

This session will be devoted to lean how to use the methods presented in the morning. Students will learn how to manipulate text in RStudio with `stringr` and converting text into numbers with the `tidytext` package.

15h15-16h45 : Applications in RStudio

References

- Van Atteveldt, W., Trilling, D., & Calderón, C. A. (2022). Computational analysis of communication. John Wiley & Sons.

Day 3 : Supervised methods

10h-12h : Introduction to supervised methods

In this session, we will explore supervised machine learning methods designed to generate measurements using text in the field of social sciences. The lecture will delve into the objectives of supervised learning and elaborate on the evaluation and validation of supervised models, shedding light on important use cases.

13h30-14h45 : Supervised machine learning in practice

15h15-16h45 : Applications in RStudio

References

- Bonikowski, B., Luo, Y., & Stuhler, O. (2022). Politics as Usual? Measuring Populism, Nationalism, and Authoritarianism in US Presidential Campaigns (1952–2020) with Neural Language Models. Sociological Methods & Research, 51(4), 1721-1787.
- Hvitfeldt, E., & Silge, J. (2021). Supervised machine learning for text analysis in R. CRC Press.
- Kuhn, M., & Silge, J. (2022). Tidy Modeling with R. " O'Reilly Media, Inc.".
- Peterson, A., & Spirling, A. (2018). Classification accuracy as a substantive quantity of interest: Measuring polarization in westminster systems. Political Analysis, 26(1), 120-128.

Day 4 : Unsupervised learning

10h-12h : Introduction to unsupervised text analysis

In this session, participants will delve into the fundamentals of unsupervised text analysis. The lecture will explore the logic behind machine learning, distinguishing between unsupervised and supervised approaches. Additionally, the session will provide an introduction to two unsupervised methods: topic modeling and word embeddings.

13h30-14h45 : Topic modelling and word embeddings in practice

15h15-16h45 : Applications in RStudio

References

- 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), 168-189.
- DiMaggio, P., Nag, M., & Blei, D. (2013). Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of US government arts funding. Poetics, 41(6), 570-606.
- Finseraas, H., Høyland, B., & Søyland, M. G. (2021). Climate politics in hard times: How local economic shocks influence MPs attention to climate change. European Journal of Political Research, 60(3), 738-747.
- Rodriguez, Pedro L., and Arthur Spirling. "Word embeddings: What works, what doesn't, and how to tell the difference for applied research." The Journal of Politics 84.1 (2022): 101-115.
- Rheault, L., & Cochrane, C. (2020). Word embeddings for the analysis of ideological placement in parliamentary corpora. Political Analysis, 28(1), 112-133.

Day 5 : Large language models

10h-12h : Overview of Large Language Models (LLMs)

This session will provide students with an introduction to cutting-edge methods in text analysis utilizing Large Language Models such as BERT or GPT. The lecture will explore the distinctions between these advanced models and their predecessors, emphasizing how artificial intelligence can contribute to addressing social sciences' questions.

13h30-14h45 : LLMs in practice

This practical session aims to guide students through the usage of Large Language Models in a straightforward manner. Utilizing the Huggingface Python library and Google Colab, participants will gain hands-on experience in applying LLMs.

15h15-16h45 : Application in Python

References

- Do, Salomé, Étienne Ollion, and Rubing Shen. "The Augmented Social Scientist: Using Sequential Transfer Learning to Annotate Millions of Texts with Human-Level Accuracy." Sociological Methods & Research (2022): 00491241221134526.

General references

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), 1-18.

Boehmke, B., & Greenwell, B. M. (2019). Hands-on machine learning with R. CRC press.

Gandrud, C. (2013). Reproducible research with R and R studio. CRC Press.

Grimmer, J., Roberts, M. E., & Stewart, B. M. (2022). Text as data: A new framework for machine learning and the social sciences. Princeton University Press.

Hvitfeldt, E., & Silge, J. (2021). Supervised machine learning for text analysis in R. CRC Press.

Kuhn, M., & Silge, J. (2022). Tidy Modeling with R. " O'Reilly Media, Inc.".

Licht, H. (2022). Cross-Lingual Classification of Political Texts Using Multilingual Sentence Embeddings. Political Analysis, 1-14.

Salganik, M. J. (2019). Bit by bit: Social research in the digital age. Princeton University Press.

Silge, J., & Robinson, D. (2017). Text mining with R: A tidy approach. " O'Reilly Media, Inc.".