Mirko Wegemann (he/him)	Dr. Eva Krejcova (she/her)
PhD Candidate	Max Weber Fellow
European University Institute	European University Institute
mirko.wegemann@eui.eu	eva.krejcova@eui.eu

EUI DEPARTMENT OF POLITICAL AND SOCIAL SCIENCES

Third term 2023/2024

Computational Text Analysis

Seminar Dates

The seminar takes place on four dates between 30/05/2024 and 04/06/2024. Sessions 1-3 will be taught by Mirko Wegemann, session 4 by Dr. Eva Krejcova. The teaching assistant for this course is Sara Dybesland (she/her).

Schedule

Input session	Lab session
30/05/2024, 9:00-11:00 (SR 2)	30/05/2024, 13:00-15:00 (SR 2)
31/05/2024, 10:00-12:00 (SR 2)	31/05/2024, 13:00-15:00 (SR 2)
03/06/2024, 10:00-12:00 (SR 2)	03/06/2024, 13:00-15:00 (SR 2)
04/06/2024, 10:00-13:00 (SR 2)	No lab session (but longer input!)

Course Overview

This course provides an introduction into computational text analysis. It introduces a wide array of methods used in the world of ‘text-as-data’ such as web-scraping, unsupervised topic models, and supervised classification tasks. It provides students with an outlook on recent developments of the discipline, most notably the evolution of word embeddings, machine learning and large language models. Moreover, it introduces approaches treating images-as-data, encompassing techniques for image processing, classification and face recognition.

Credits can only be granted if participants actively participate in the input and in the lab sessions.

Learning Outcomes

At the end of the course, you have been introduced to key concepts of computational text analysis. More specifically, you will know how to gather textual data from websites and will be able to implement web-scraping techniques, adequate for different types of websites. With regard to textual analyses, you will have learnt about the pipeline to prepare textual data for standard, bags-of-words techniques of textual analysis. You will have made first experiences with unsupervised topic modelling. Furthermore, you will have used more advanced word embedding techniques in `R` and know how to create a deep-learning model. In a showcase session, you will also learn how to build your own transformer model in Python.
The last session focuses on images-as-data, combining lectures and practical sessions. You will acquire an understanding of the promises and pitfalls of using machine learning for automated visual content analysis. Lectures will cover the most recent developments in automated image and video analysis and their application in social sciences. In the practical sessions, you will learn how to use Python to implement different computer vision techniques, including image processing, image classification, and face recognition.

Requirements

A successful seminar participation grants students 10 credits. For a successful participation, students shall...

- actively attend all sections (including the lab sessions!) of the workshop
- apply the techniques they learnt in the workshop on their own research projects (during the lab sessions)

Inclusiveness

To promote inclusiveness, this seminar aims to use gender-inclusive language. Participants are invited to share their pronouns with the class.
Students are encouraged to submit feedback via Google Forms or by sending me a mail.

Readings

This is a method-intensive class. Readings are complementary, and may help you to better understand the underlying assumptions of a method.

Session 1: Web-scraping [30 May 2024]

Readings
Wickham, H., Çetinkaya-Rundel, M., & Grolemund, G. (2023). *R for Data Science*. "O'Reilly Media, Inc." – *chapter on web-scraping (free version online)*

Munzert, S. (2015). *Automated data collection with R: A practical guide to web scraping and text mining* (1st ed.). Wiley – *chapters on HTML, regular expressions and scraping the web are helpful for this seminar (although we will work with rvest as a package solution)*

Tutorials
Marble, W. (2016). Web-Scraping with R

For a tutorial on HTML: https://www.w3schools.com/html/default.asp

Session 2: Bags-of-words [31 May 2024]

Readings
Benoit, K. (2020). Text as data: An overview. In L. Curini & R. Franzese (Eds.), *The SAGE Handbook of Research Methods in Political Science and International Relations*. SAGE Publications Ltd. https://doi.org/10.4135/9781526486387 – *general overview into text analysis*

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. https://doi.org/10.1093/pan/mps028 – *another good and highly cited overview*

Roberts, M. E., Stewart, B. M., & Tingley, D. (2019). Stm: An R Package for Structural Topic Models. *Journal of statistical software*, *91*, 1–40. https://doi.org/10.18637/jss.v091.i02 – *intro into structured topic models*

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *The journal* *of machine learning research*, *3*(null), 993–1022 – *foundations of topic modelling (LDA)*

Blei, D. M., & Lafferty, J. D. (2005). Correlated topic models. *Proceedings of the 18th International Conference on Neural Information Processing Systems*, 147–154 – *further reading: correlated topic models*

Proksch, S.-O., Lowe, W., Wäckerle, J., & Soroka, S. (2019). Multilingual Sentiment Analysis: A New Approach to Measuring Conflict in Legislative Speeches. *Legislative studies quarterly*, *44*(1), 97–131. https://doi.org/10.1111/lsq.12218 – *further reading: application of sentiment analysis*

Watanabe, K., & Zhou, Y. (2022). Theory-Driven Analysis of Large Corpora: Semisupervised Topic Classification of the UN Speeches. *Social science computer review*, *40*(2), 346–366. https://doi.org/10.1177/0894439320907027 – *further reading: seeded topic models*

Eshima, S., Imai, K., & Sasaki, T. (2024). Keyword-Assisted Topic Models. *American journal of political science*, *68*(2), 730–750. https://doi.org/10.1111/ajps.12779 – *further reading: keyword-assisted topic models*

Watanabe, K., & Baturo, A. (2024). Seeded Sequential LDA: A Semi-Supervised Algorithm for Topic-Specific Analysis of Sentences. *Social science computer review*, *42*(1), 224–248. https://doi.org/10.1177/08944393231178605 – *further reading: Latent Semantic Scaling*

Tutorials

Introduction into quanteda

Introduction into stm

Session 3: Embeddings and machine learning [03 June 2024]

Readings

Rodriguez, P., & Spirling, A. (2021). Word Embeddings: What works, what doesn't, and how to tell the difference for applied research. *The journal of politics*. https://doi.org/10.1086/715162 – *word embeddings for the social sciences*

Rodriguez, P. L., Spirling, A., & Stewart, B. M. (2023). Embedding Regression: Models for Context-Specific Description and Inference. *American political science review*, *117*(4), 1255–1274. https://doi.org/10.1017/S0003055422001228 – *embedding regression*

Wankmüller, S. (2022). Introduction to Neural Transfer Learning With Transformers for Social Science Text Analysis. *Sociological methods & research*, 1–77. https://doi.org/10.1177/00491241221134527 – *explanation of transformer models*

Hoes, E., Altay, S., & Bermeo, J. (2023). Using ChatGPT to Fight Misinformation: ChatGPT Nails 72% of 12,000 Verified Claims. https://doi.org/10.31234/osf.io/qnjkf – *GPT to classify misinformation*

Yang, K.-C., & Menczer, F. (2023). Large Language Models Can Rate News Outlet Credibility. https://doi.org/10.48550/arXiv.2304.00228 – *GPT to classify the credibility of media outlets*

Törnberg, P. (2023). ChatGPT-4 Outperforms Experts and Crowd Workers in Annotating Political Twitter Messages with Zero-Shot Learning. https://doi.org/10.48550/arXiv.2304.06588 – *GPT against crowdcoding*

Reiss, M. (2023). *Testing the Reliability of ChatGPT for Text Annotation and Classification: A Cautionary Remark* (Preprint). Open Science Framework. https://doi.org/10.31219/osf.io/rvy5p – *on potential pitfalls of GPT*

Tutorials

Deep learning course by fastAI

NLP course by Hugging Face

Session 4: Images as data [04 June 2024]

Textbooks on Images as Data

Howse, J., & Minichino, J. (2020). *Learning OpenCV 4 Computer Vision with Python 3: Get to grips with tools, techniques, and algorithms for computer vision and machine learning.* Packt Publishing Ltd

Martínez, J. (2021). *TensorFlow 2.0 Computer Vision Cookbook: Implement machine learning solutions to overcome various computer vision challenges*. Packt Publishing

Recommended Literature

Cinar, A. C., & Kıbrıs, Ö. (2023). Persistence of voice pitch bias against policy differences. *Political science research and methods*, 1–15. https://doi.org/10.1017/psrm.2023.51

Boussalis, C., Coan, T. G., Holman, M. R., & Müller, S. (2021). Gender, Candidate Emotional Expression, and Voter Reactions During Televised Debates. *American political science review*, *115*(4), 1242–1257. https://doi.org/10.1017/S0003055421000666

Cantú, F. (2019). The Fingerprints of Fraud: Evidence from Mexico's 1988 Presidential Election. *American political science review*, *113*(3), 710–726. https://doi.org/10.1017/S0003055419000285

Dietrich, B. J. (2021). Using Motion Detection to Measure Social Polarization in the U.S. House of Representatives. *Political analysis*, *29*(2), 250–259. https://doi.org/10.1017/pan.2020.25

Peng, Y. (2018). Same Candidates, Different Faces: Uncovering Media Bias in Visual Portrayals of Presidential Candidates with Computer Vision. *Journal of communication*, *68*(5), 920–941. https://doi.org/10.1093/joc/jqy041

Steinert-Threlkeld, Z. C., Chan, A. M., & Joo, J. (2022). How State and Protester Violence Affect Protest Dynamics. *The journal of politics*, *84*(2), 798–813. https://doi.org/10.1086/715600

Torres, M., & Cantú, F. (2022). Learning to See: Convolutional Neural Networks for the Analysis of Social Science Data. *Political analysis*, *30*(1), 113–131. https://doi.org/10.1017/pan.2021.9

Zhang, H., & Peng, Y. (2022). Image Clustering: An Unsupervised Approach to Categorize Visual Data in Social Science Research. *Sociological methods & research*, 00491241221082603. https://doi.org/10.1177/00491241221082603

Software requirements

The data, scripts and slides will be made available on Github (link tba).

Please download the current version of `R` and `RStudio`. In R, use the following command to make sure to have the following packages installed `install.packages(c("tidyverse", "rvest", "RSelenium", "wdman", "httr2", "openxlsx", "quanteda", "quanteda.textplots", "quanteda.textstats", "quanteda.textmodels", "stm", "LSX", "text2vec", "uwot", "umap", "conText", "tensorflow", "keras", "torch", "reticulate", "devtools", "gptstudio", "TheOpenAIR"))`. Please also make sure to have Java installed. For session 3, it makes sense to have a version of Python installed to your device (download here). You can install Python packages (tensorflow) in R via *py_install("packagename")*. If you want download your own Manifesto Project data, install manifestoR.

For session 4, course materials will be shared with the participants in a Jupyter Notebook, with the code written in Python. However, there is no requirement for students to have Python installed, as everything will be run on Google Colab. Participants should have a Google account for using Google Colab and should have a basic familiarity with Markdown. While experience with programming in Python would be beneficial, it is not required for the course. All code will be provided by the instructor.

Participants who are not familiar with Markdown or Python and wish to do so can check the (very short) introductions on **Markdown** and **Colab**.