SIRCSS Summer School in Computational Text Analysis

June 10 – June 20, 2024, Linköpings Universitet, Campus Norrköping

Director of the Research School

Jacob Habinek, Institute for Analytical Sociology

Instructors

Måns Magnusson, Uppsala Universitet
Miriam Hurtado Bodell, Institute for Analytical Sociology

Guest Speakers

Robert Borges, Uppsala Universitet
Marco Kuhlmann, Linköpings Universitet
Väinö Yrjänäinen, Uppsala Universitet

Research Assistant

Hendrik Erz, Institute for Analytical Sociology

Course description

The course introduces standard methods to analyze textual data with applications in the social sciences. The focus is to introduce core methodologies and how these methods can be used to analyze textual data in the social sciences. The first week covers standard unsupervised methods, focusing on topic modeling and word embedding models. The supervised part focuses on transformer neural networks and large language models.

Prerequisites

1. An introductory machine learning course
2. Introductory knowledge of neural networks (supervised part)
3. Programming using Python and/or R

Course content

- Introduction: Text as Data
- Unsupervised methods
    - Topic models
    - Word embeddings
- Supervised methods
    - Transformer neural networks and encoder-based models (e.g. BERT)
    - Large language models/decoder-based models (e.g. GPT)

Course structure

The course will be conducted over two weeks. Then, each week on the core content (supervised and unsupervised methods) will contain two days of lectures combined with a literature seminar to present and discuss state-of-the-art empirical research using the methods discussed in the lecture. There will also be two lab sessions each week, where participants can get help from an RA to help solve labs that cover the content covered the previous day. On lab days, there is also time for self-study—reading the material for the literature seminars and project proposal seminars.

The last week will contain the presentation of the student's intended research proposal. The research proposals are handed in at the end of Summer.

Examination & Earning ECTS

Participation in the Summer School in Computational Text Analysis is worth **up to 6 ECTS**. Students who participate in the lectures, take part in the literature seminars (including a brief presentation on one of the assigned readings), and submit three completed laboratory reports are eligible to receive 3 ECTS (Pass / Fail).

Students who wish to receive a further 3 ECTS (Pass / Fail) must also complete an extended abstract or paper draft (2–4 pages) ideally closely connected to the student's research topic. Students will also present their projects during the second week of the summer school and receive feedback on the best path forward and how the work can otherwise be improved. The paper should be handed in to the instructors at the end of summer and must include the following elements:

- A clearly formulated research question;
- A convincing conceptual framework anchored in the relevant literature;
- An account of the data sources or data collection strategy;
- A description of the measurement strategy and methods of analysis, including the strengths and weaknesses of the research design;
- And, if applicable, the results of the analysis and a discussion of the outcome.

Reading material

Introduction

Chapter 2-4 in:
Grimmer, Justin, Margaret E. Roberts, and Brandon M. Stewart. Text as data: A new framework for machine learning and the social sciences. Princeton University Press, 2022.

Watanabe, K. Text as Data. Encyclopedia of Technology & Politics by Edward Elgar Publishing, 2021.
https://blog.koheiw.net/wp-content/uploads/2021/09/Text-as-Data.pdf

Chapter 1, 7, and 8 in:
Kracht, M. (2007). Introduction to linguistics. Los Angeles. LA: Hilgard Avenue, 2.

Eckert, P. (2012). Three waves of variation study: The emergence of meaning in the study of sociolinguistic variation. Annual Review of Anthropology, 41, 87–100.

Topic Models

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.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. Proceedings of the National Academy of Sciences, 101, 5228-5235.

Blei, D. M. (2012). Probabilistic topic models. Communications of the ACM, 55(4), 77-84.

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., ... & Rand, D. G. (2014). Structural topic models for open-ended survey responses. American journal of political science, 58(4), 1064-1082.

Word Embeddings

Kozlowski, A. C., Taddy, M., & Evans, J. A. (2019). The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review*, *84*(5), 905-949.

Stoltz, D. S., & Taylor, M. A. (2021). Cultural cartography with word embeddings. *Poetics*, *88*, 101567.

Rudolph, M., Ruiz, F., Mandt, S., & Blei, D. (2016). Expontential family embeddings. Advances in Neural Information Processing Systems, 29.

Rudolph, M., & Blei, D. (2018). Dynamic embeddings for language evolution. Proceedings of the 2018 World Wide Web Conference, 1003–1011.

Transformers/Encoders

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized Bert pretraining approach. arXiv preprint arXiv:1907.11692.

Hofmann, V., Pierrehumbert, J. B., & Schütze, H. (2020). Dynamic contextualized word embeddings. arXiv preprint arXiv:2010.12684.

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.

Do, S., Ollion, É., & Shen, R. (2022). The augmented social scientist: Using sequential transfer learning to annotate millions of texts with human-level accuracy. *Sociological Methods & Research*, 00491241221134526.

LLMs/Decoders

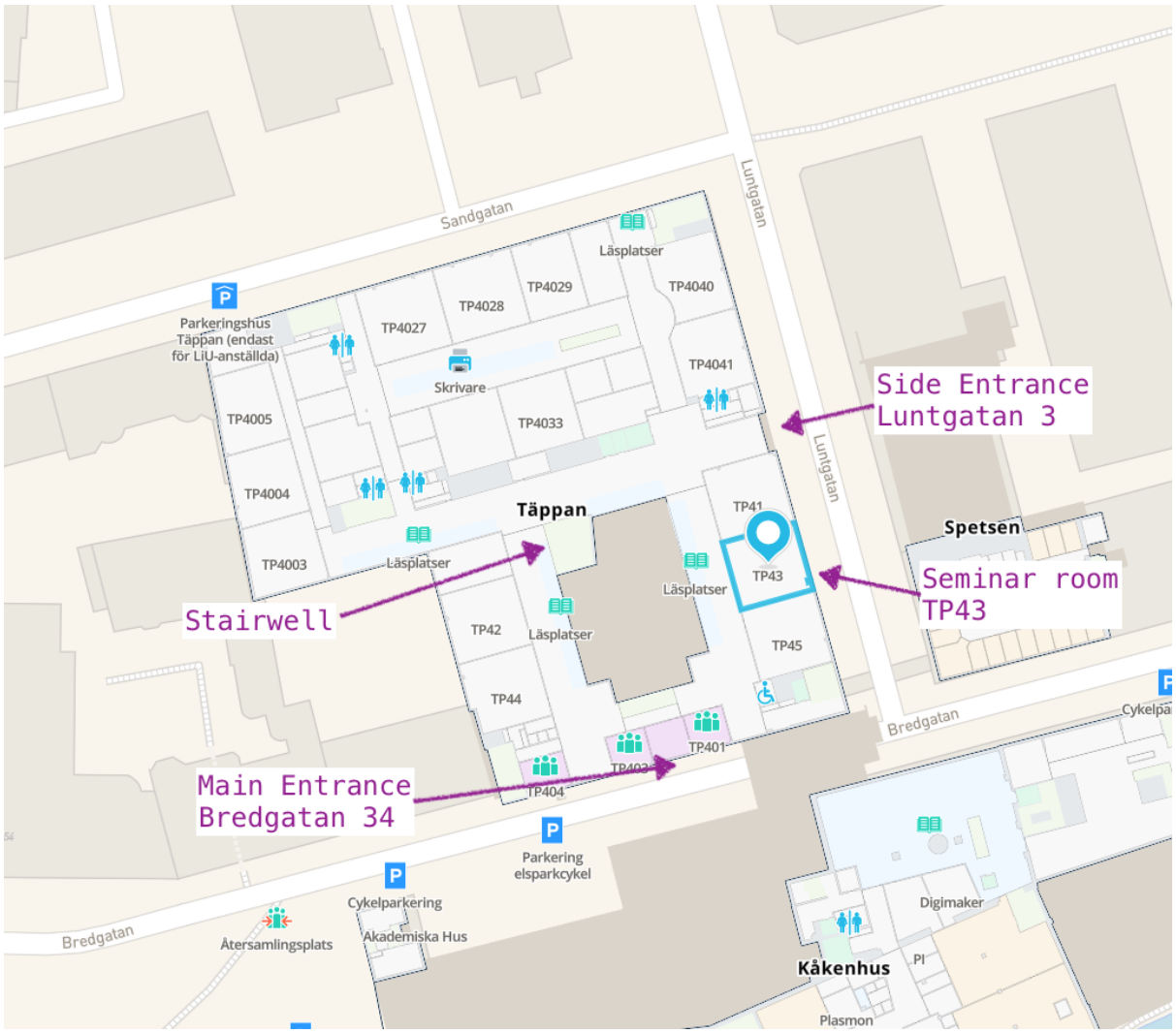
Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., & Yang, D. (2024). Can large language models transform computational social science?. *Computational Linguistics*, 1-55.

Atari, M., Xue, M. J., Park, P. S., Blasi, D., & Henrich, J. (2023). Which humans?.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., ... & Wen, J. R. (2023). A survey of large language models. arXiv preprint arXiv:2303.18223.

Summer School Schedule

All seminars will take place in room **TP43** (Bredgatan 34). Every day, there will be Fika (just coffee/tea) at 9.45 am. On all **non-lab days**, there will be additional Fika at 2.00 pm (coffee/tea and pastry).



Sunday **Hotel Check-In & Welcome on June 9**

15–18 Check-In: Hotel President (Vattengränden 11, 602 22 Norrköping)

Welcome from 18.00 onwards at Ölstugan Tullen
(Sankt Persgatan 80, 602 33 Norrköping)

Week One (June 10–14)

Monday **Introduction: Text as Data** (in the Social Sciences)

9.30–10 Introduction (Jacob Habinek)
10–12 Lecture (Robert Borges, Måns Magnusson, and Miriam Hurtado Bodell)
13–15 Lecture

Tuesday **Topic Models**
(Måns Magnusson and Miriam Hurtado Bodell)

9–12 Lecture
13–15 Literature Seminar on Topic Models

Wednesday **Lab/Study Day**

9–12 Lab session on Topic Models (Hendrik Erz; voluntary)

Thursday **Word Embeddings**
(Väinö Yrjänäinen and Miriam Hurtado Bodell)

9 –12 Lecture
13–15 Literature seminar on Word Embeddings

Friday **Lab/Study Day** (preparation for the literature seminar)

9–12 Lab session on Word Embeddings (Hendrik Erz; voluntary)

Week Two (June 17–20)

Monday **Transformers/Encoders**
(Måns Magnusson and Miriam Hurtado Bodell)

9–12 Lecture
13–15 Literature seminar on Transformer encoder models

Tuesday **Large Language Models/Decoders**
(Marco Kuhlmann and Miriam Hurtado Bodell)

9–12 Lecture
13–15 Literature seminar on Transformer decoder models/Large Language Models

Wednesday **Lab/Study Day**

9–15 Lab session on LLMs (Hendrik Erz; voluntary)

Thursday **Pre-research proposal presentations & Farewell**
(Måns Magnusson and Miriam Hurtado Bodell)

9–17 Pre-research proposal presentations

Farewell from 18.00 onwards at Ölstugan Tullen
(Sankt Persgatan 80, 602 33 Norrköping)

Friday **Check-Out on June 21**

Check-Out: Hotel President (until 12.00)