DORTMUND CENTER FOR DATA-BASED MEDIA ANALYSIS

technische universität dortmund

fakultät statistik

Tools for Temporal Text Analysis

ttta: a python code collection for diachronic NLP tasks

Installation instructions: `github.com/K-RLange/ttta`

Kai-Robin Lange
Lars Grönberg
Niklas Benner
Imene Kolli
Aymane Hachcham
Jonas Rieger
Carsten Jentsch

1. Motivation

Politics reflect the discourse and the society at a given point in time. Just as society changes constantly, politics and the discourse surrounding them, do so as well.

Political data is inherently temporal

- Parliament debates
- Party manifestos
- News paper articles
- Election campaign speeches

↪ To handle temporal text data, we need specific tools. But:
Niche, scattered across GitHub, with different interfaces, in different programming languages.

Our goal

- Make it easier to access tools to work with temporal text data
- A code collection, to which researchers can contribute with their models
- Have a common interface to use for all models

In our demonstrations here, we will use Bundestag speeches from the SpeakGer data set.

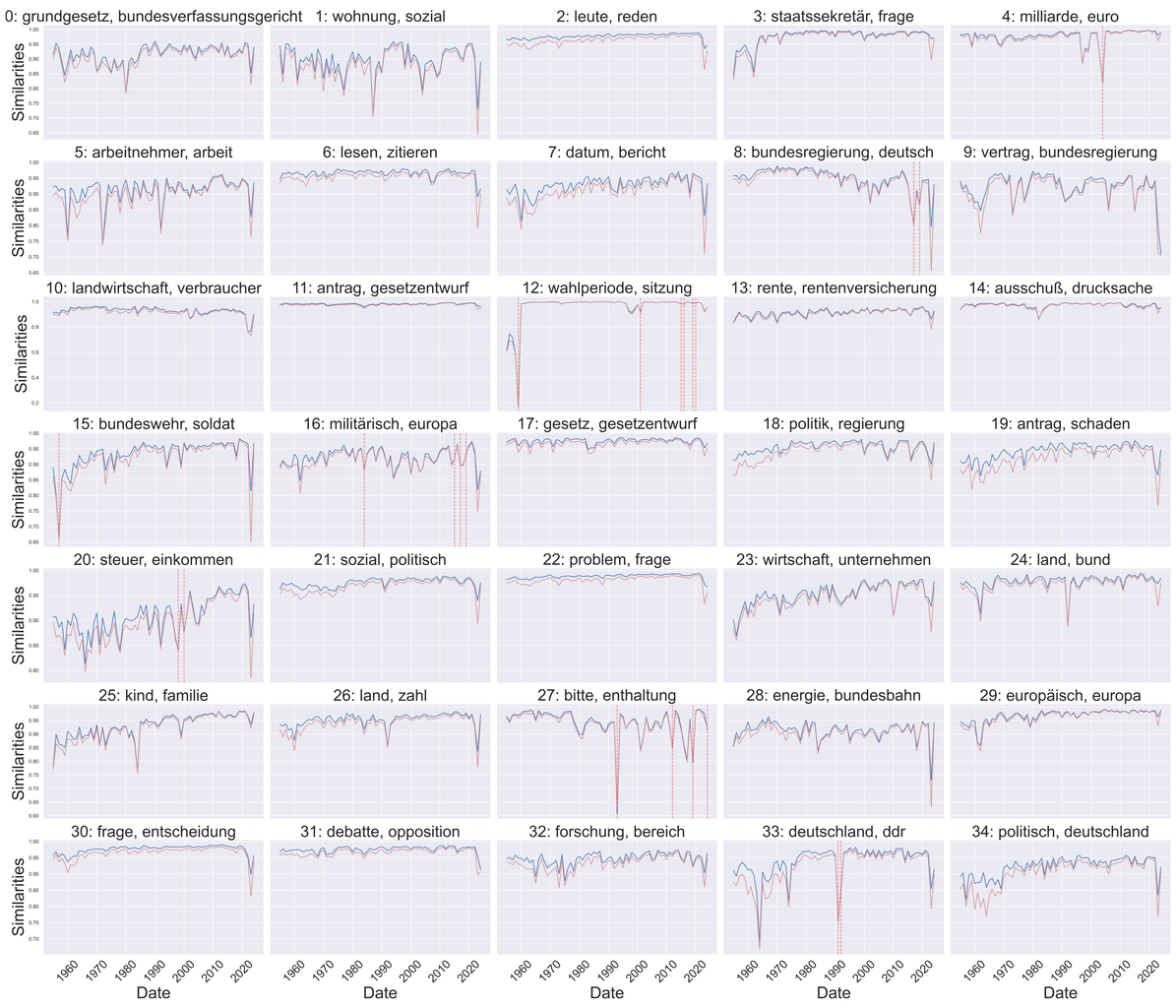
2. Contributing

This package is **not** complete. We aim to include many more models, with which you can help:
If you want to contribute an implementation, write a mail to kalange@statistik.tu-dortmund.de.
You will be fully credited for the model(s) you contribute.

3. Dynamic Topic Models and Topical Change detection

LDAPrototype, RollingLDA and Topical Changes

- LDAPrototype as a foundational model: a consistent, “average” LDA from multiple runs
- RollingLDA as a dynamic topic model: trains LDAs with a sliding window over time chunks
- Topical Changes: bootstrap test based change detection for each topic in RollingLDA



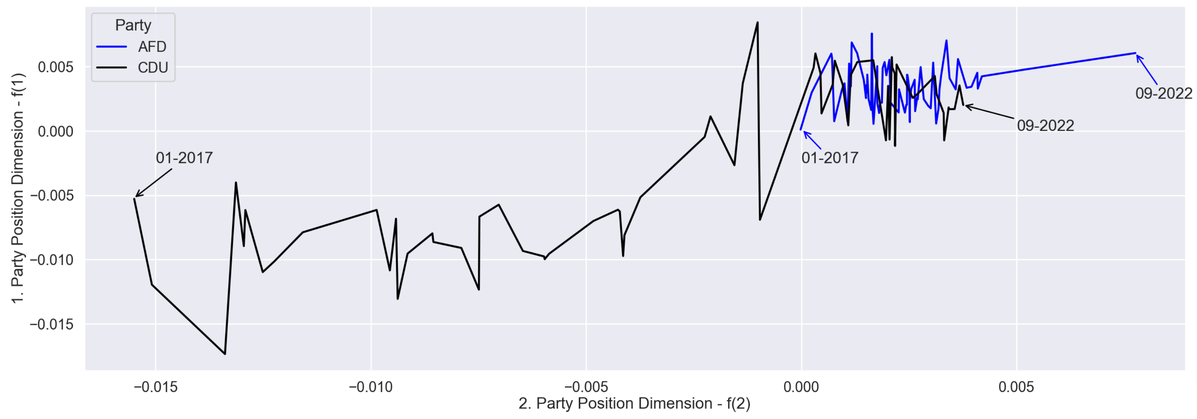
Interpretation

- Administrative changes in topics 12 and 27
- Introduction of compulsory military service in topic 15
- Peace demonstrations, Afghanistan war and other middle eastern wars in topic 16
- European financial crisis in topic 4
- German reunion in topic 33
- Tax reforms in topic 20

4. Document Scaling

Poisson Reduced Rank Models

- Poisson Reduced Rank Models can be used to estimate latent characteristics of documents
- Word weights can be modeled as time dependent or independent



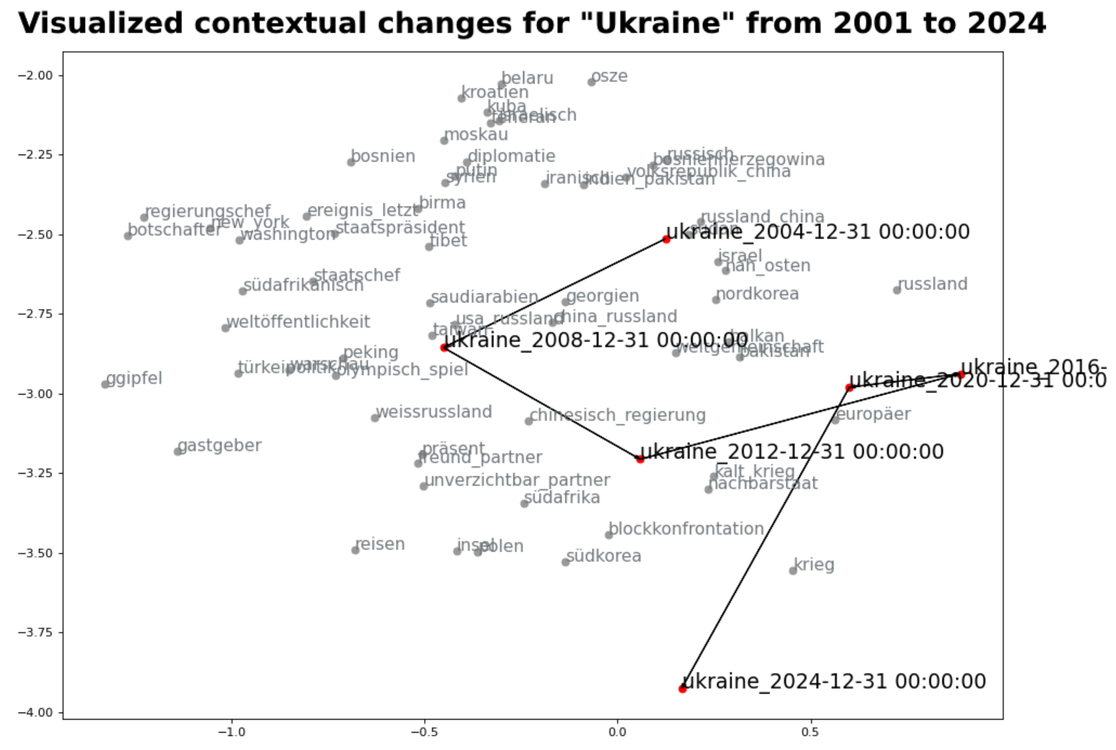
Interpretation

- Trajectory of latent party positions of parliament debates for AfD and CDU
- AfD’s position has remained relatively consistent
- CDU has changed its latent party position and approached the AfD
- Possible interpretation for the axes: conservative and right wing spectrum

5. Semantic Change Detection

Static or contextualized embeddings

- Contextualized: Differentiate word senses by using prototypical sense embeddings and detect changes in sense distribution
  ↪ Use pre-defined senses for a word and analyze the usage over time (e.g. dog whistles)
- Alignment: Training Word2Vec on each time chunk, then aligning all vector spaces
  ↪ Find big context changes, see if a word has changed or visualize word trajectories



Interpretation

- 2001-2004 (pre orange revolution in 2004):
  Part of a general “eastern block” along with the near east and the baltics
- 2005-2008 (post orange revolution in 2004):
  Close to Taiwan – similarity in conflict with Russia and China
- 2009-2012 (includes election of pro-russian president Wiktor Janukowitsch in 2010):
  Close to “cold war” and “block frontation”, as Ukraine’s government aligned with Russia
- 2013-2020 (includes annexiation of the Krim in 2014):
  Between Russia and Europe, but not close to “war”. No big change between 2016 and 2020.
- 2021-2024 (includes the Russian-Ukrainian war of 2022):
  Biggest observed change, “war” as its nearest neighbour.

6. References

[1] W. L. Hamilton, J. Leskovec, and D. Jurafsky. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2016.

[2] C. Jentsch, E. R. Lee, and E. Mammen. Time-dependent Poisson reduced rank models for political text data analysis. *Computational Statistics & Data Analysis*, 142, 2020.

[3] C. Jentsch, E. R. Lee, and E. Mammen. Poisson reduced-rank models with an application to political text data. *Biometrika*, 108(2), 2021.

[4] K.-R. Lange, J. Rieger, N. Benner, and C. Jentsch. Zeitenwenden: Detecting changes in the German political discourse. 2022.

[5] J. Rieger, C. Jentsch, and J. Rahnenführer. RollingLDA: An update algorithm of Latent Dirichlet Allocation to construct consistent time series from textual data. In *Findings Proceedings of the 2021 EMNLP-Conference*, 2021.

[6] J. Rieger, K.-R. Lange, J. Flossdorf, and C. Jentsch. Dynamic change detection in topics based on rolling LDAs. In *Proceedings of the Text2Story’22 Workshop*, 2022.

[7] J. Rieger, J. Rahnenführer, and C. Jentsch. Improving Latent Dirichlet Allocation: On Reliability of the Novel Method LDAPrototype. In *Natural Language Processing and Information Systems*, 2020.