Introduction to topic modeling

Tutorial

Jonas Rieger
https://jonasrieger.github.io/

SuSe 2023

Formalities

- English
- course number: 09-71-D.1-1d

- successful participation via
  - **active** participation
  - development of own research question (in teams of two)
- examination performance (optional)
  - elaborate the project for a written term paper

Organization

- in person
- 4 dates:
  - Fr. 12.05.2023 10:00–16:30 GW2 B1400
  - Sa. 13.05.2023 9:00–12:30 GW2 B1400
  - Fr. 16.06.2023 10:00–16:30 LINZ4 60070
  - Sa. 17.06.2023 9:00–12:30 GW2 B1400
- please register via Stud.IP

- lecture-like parts + hands-on parts
- please bring your own device (with R installed)

Contents

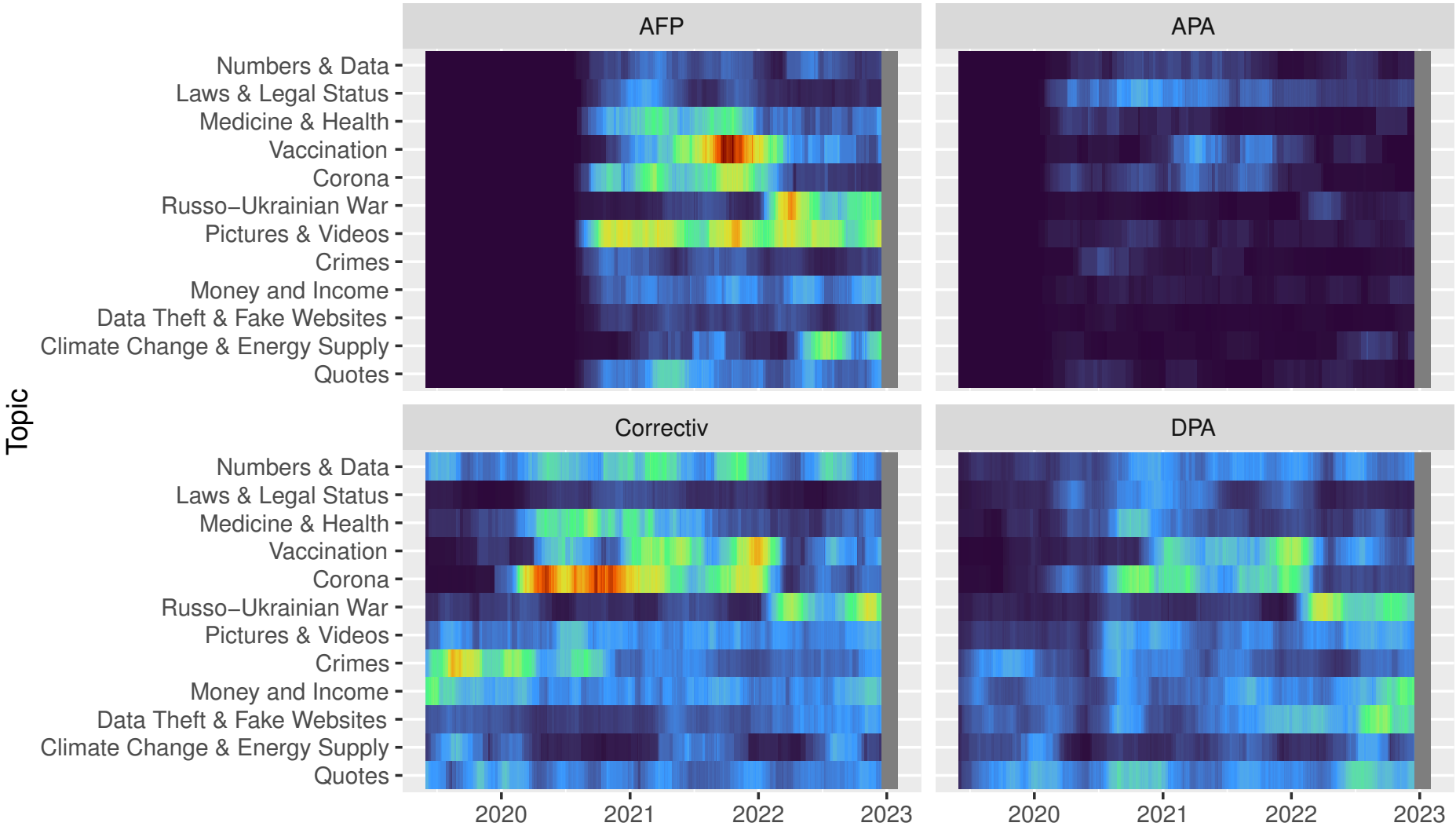
- text data handling and preprocessing
  (e.g., tokenization, stopwords, stemming, lemmatization)
- foundation of matrix factorization/dimension reduction techniques:
  - principal component analysis (PCA)
  - singular value decomposition (SVD)
  - latent semantic analysis (LSA)
- (probabilistic) topic models, mainly:
  - **latent Dirichlet allocation (LDA)**
  - structural topic model (STM)
- short digression to (classical) neural topic models
- transformer based topic models, i.e., BERTopic (which is also *neural*)
  - teaser on methodological idea (no worries!)
  - pros and cons; when to use
- discussion on differences of LDA, STM, BERTopic
- application of LDA, STM, BERTopic for real world questions

Preparation

- install (and be familiar with) R
- read "What We Can Do and Cannot Do with Topic Modeling: A Systematic Review" by Chen et al. (2023) DOI: 10.1080/19312458.2023.2167965.

- optional: same as for R for Python
  - not necessary, since this is only for playing around with BERTopic (which is also possible without installing Python using Google Colab)
- optional: install (and setup) R package rtweet
- optional: additional literature
  - Blei (2012). Probabilistic Topic Models. DOI: 10.1145/2133806.2133826.
  - please ask, if you want more suggestions :)

Example: Topics in German Fact-Checks



see also `https://gadmo.eu/`

Questions

rieger@statistik.tu-dortmund.de

You can also reach me at
jonrie@uni-bremen.de