Dynamic change detection in topics based on rolling LDAs

Jonas Rieger, Kai-Robin Lange, Jonathan Flossdorf, Carsten Jentsch

Text2Story'22 Workshop, Stavanger

04/10/2022

technische universität dortmund

DORTMUND CENTER FOR DATA-BASED MEDIA ANALYSIS

fakultät statistik

Motivation

change detection in topics

- narrative extraction is intensively researched,
- detection of change may indicate narrative shift,
- temporal localization,
- topical localization,

topic models

- used in many application fields,
- modeling idea intuitive:
  - a set of texts is clustered into topics,
  - each text is seen as a mixture of several topics,
  - each topic is in turn characterized by its word distribution.

Latent Dirichlet Allocation Blei et al. (2003), Griffiths und Steyvers (2004)

$$W_n^{(m)} \mid T_n^{(m)}, \phi_k \sim \text{Discrete}(\phi_k), \quad \phi_k \sim \text{Dirichlet}(\eta),$$
$$T_n^{(m)} \mid \boldsymbol{\theta}_m \sim \text{Discrete}(\boldsymbol{\theta}_m), \quad \boldsymbol{\theta}_m \sim \text{Dirichlet}(\alpha)$$

with $\left(W_n^{(m)}, T_n^{(m)}\right) =$ (word, topic) at position n in text m.

- probabilistic topic model,
- latent topics,
- soft cluster,
- results in:
  - topic distributions $\hat{\theta}_m, m = 1, ..., M$ for each text,
  - word distributions $\hat{\phi}_k, k = 1, ..., K$ for each topic,
- K: number of topics (hyperparameter), M: number of texts,
- α, η determine heterogeneity of texts and topics (hyperparameter).

RollingLDA Rieger et al. (2021)

idea

- sequential modeling (chunks of texts),
- ability to add (update the model with) new texts,
- creates time-consistent time series,
- rolling memory,
- preserves overall topic structure (hypertopic),
- prevents from just fitting to existing topics,
- allows for (small) changes in topics,
- hyperparameters for customization of memory length, chunk size, ...

Illustration of the modeling idea later in context of the dataset used.

RollingLDA Rieger et al. (2021)

idea

- sequential modeling (chunks of texts),
- ability to add (update the model with) new texts,
- creates time-consistent time series,
- rolling memory,
- preserves overall topic structure (hypertopic),
- prevents from just fitting to existing topics,
- allows for (small) changes in topics,
- hyperparameters for customization of memory length, chunk size, ...

Illustration of the modeling idea later in context of the dataset used.

Idea of change detection

procedure

- similarity of current to previous count vectors of word assignments,
- resample "expected" word count vectors,
- realized similarity vs. "expected" similarity under "stable" conditions,
- minor changes are expected,
- extraordinary changes should be detected.

Set of changes

$$C_k^t = \left\{ u \mid 0 < u \leq t \leq T : \cos\left(\boldsymbol{n}_{k|u}, \boldsymbol{n}_{k|(u - z_k^u):(u-1)} \right) < q_k^t \right\} \cup 0,$$

- k: topic number,
- $t \in \{0, \ldots, T\}$: time point,
- $\boldsymbol{n}_{k|u}$: count vector of topic k at time point u,
- q_k^t: threshold, least acceptable similarity under stable conditions.

- 0 included for technical reason,
- $z_k^t = \min\left\{ z_{\max}, t - \max C_k^{t-1} \right\}$: run length without a change,
- $z_{\max}$: hyperparameter, maximum length of the reference period.

Set of changes

$$C_k^t = \left\{ u \mid 0 < u \leq t \leq T : \cos\left(\boldsymbol{n}_{k|u}, \boldsymbol{n}_{k|(u - z_k^u):(u-1)} \right) < q_k^t \right\} \cup 0,$$

- k: topic number,
- $t \in \{0, \dots, T\}$: time point,
- $\boldsymbol{n}_{k|u}$: count vector of topic k at time point u,
- q_k^t: threshold, least acceptable similarity under stable conditions.

- 0 included for technical reason,
- $z_k^t = \min\left\{ z_{\max}, t - \max C_k^{t-1} \right\}$: run length without a change,
- $z_{\max}$: hyperparameter, maximum length of the reference period.

How to determine q_k^t (and $z_{\max}$)?

Dynamic thresholds

comparison to stable condition

$$\tilde{\phi}_k^{(t)} = (1-p)\,\hat{\phi}_k^{(t-z_k^t):(t-1)} + p\,\hat{\phi}_k^{(t)},$$

- $\hat{\phi}_k^{(t)}$: word distribution estimator for topic k at time point t,
- p: hyperparameter, sensitivity of change detection,

- resample (R times) word count vectors $\tilde{\boldsymbol{n}}_{k|t}$ using $\tilde{\phi}_k^{(t)}$,
- q_k^t is 0.01 quantile of realized similarities

$$\cos\left(\tilde{\boldsymbol{n}}_{k|t}, \boldsymbol{n}_{k|(t-z_k^t):(t-1)}\right),$$

- repeat procedure for all topics and all time points.

Data and parameters

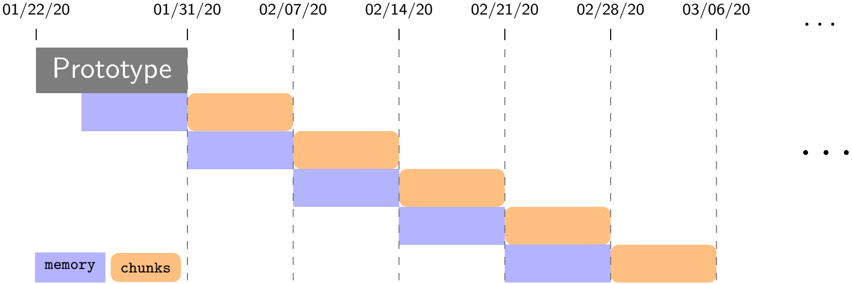
data

- TLS-Covid19 data set (Pasquali et al. 2021),
- Covid-19 related liveblog articles of CNN,
- 27 432 texts from January 22nd 2020 until December 12th 2021,
- common NLP preprocessing; including lemmatization,

parameter

- LDA: $K = 12\ (8, \ldots, 20)$, $\alpha = \eta = 1/K$,
- RollingLDA:
  - weekly updates (data `chunks`),
  - previous week as `memory`,
  - initialization with first ten days via LDAPrototype (Rieger et al. 2022),
- $z_{\max} = 4\ (1, \ldots, 20)$, $p = 0.85\ (0.5, \ldots 0.8, 0.81, \ldots, 0.90)$.

Modeling procedure



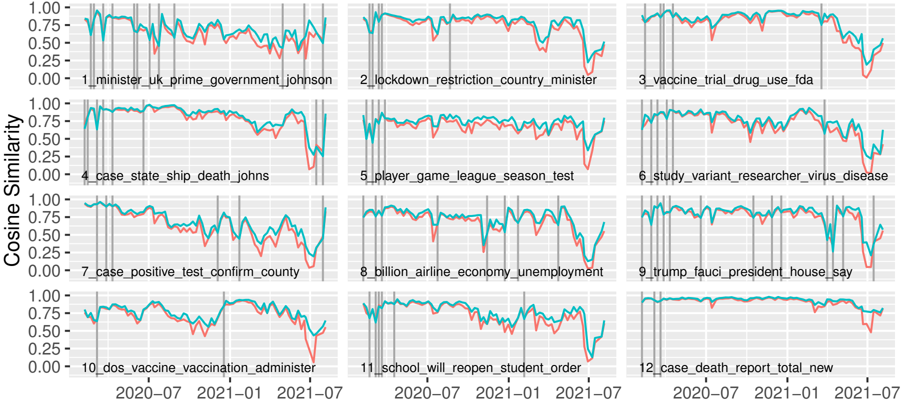
- prototype period serves as $t = 0$,
  - 10 days,
  - 605 texts,
- updating the model every 7 days (`chunks`)
  - with the previous 7 days serving as `memory`.

Evaluation

human assisted evaluation

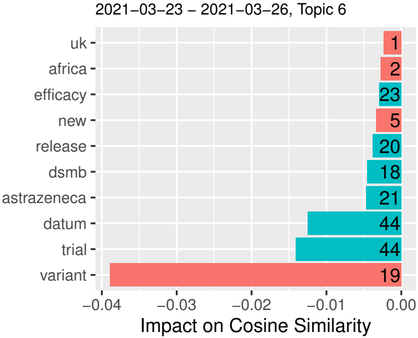
- set of changes per topic,
- precision calculation human assisted:
  - plausibility judgment using eye-balling,
  - timestamp, external knowledge,
  - time-dependent top words,
  - word impacts, (see Analysis)
- recall not calculated:
  - need of gold standard,
  - grateful for tips regarding data sets including reliable target variable in this particular setting.

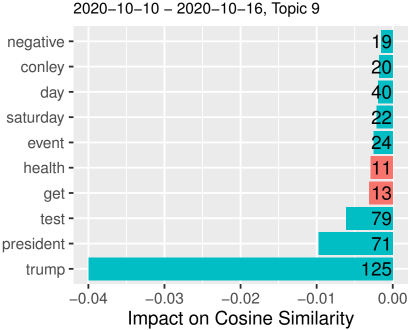
Monitoring time series



- red: threshold (0.01 quantile of "expected" similarity),
- blue: realized similarity,
- 71% precision (55/78), 78% since April 2020.

Insights with word impacts





- leave-one-out cosine impacts of words,
- red: reduced frequency, blue: increased frequency,
- 2021-03-25: trial about efficiency of AstraZeneca vaccine,
- 2020-10-12: Trump recovers from Covid-19.

Discussion

- GitHub repository `github.com/JonasRieger/topicalchanges` for additional analyses,
  - Guardian, Observador, Publico,
  - various parameters,
- p tuning parameter,

Outlook

- improve evaluation,
- comparison to other methods (maybe need of slight modifications),
- extraction of narratives and determination of their life span,
- additional exploration tools.

Bibliography

- Blei, Ng, Jordan (2003). Latent Dirichlet allocation. JMLR 3, 993–1022.
- Griffiths, Steyvers (2004). Finding scientific topics. PNAS 101.1, 5228–5235.
- Pasquali, Campos, Ribeiro, Santana, Jorge, Jatowt (2021). TLS-Covid19: A new annotated corpus for timeline summarization. ECIR, 497–512.
- Rieger, Jentsch, Rahnenführer (2021). RollingLDA: An Update Algorithm of Latent Dirichlet Allocation to Construct Consistent Time Series from Textual Data. Findings of EMNLP, 2337–2347.
- Rieger, Jentsch, Rahnenführer (2022). LDAPrototype: A Model Selection Algorithm to Improve Reliability of Latent Dirichlet Allocation. Preprint.