Detecting Short-lasting Topics Using Nonnegative Tensor Decomposition

Lara Kassab*, Alona Kryshchenko, Hanbaek Lyu, Denali Molitor, Deanna Needell, Elizaveta Rebrova

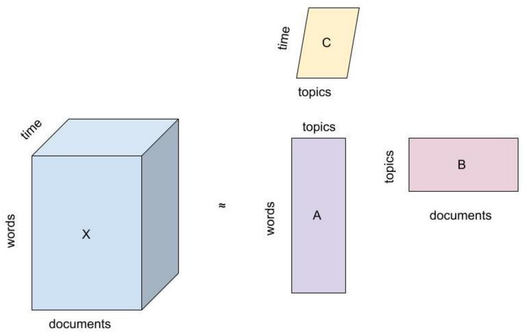
Pacific Northwest National Laboratories*

Abstract

Temporal data (such as news articles or Twitter feeds) often consists of a mixture of long-lasting trends and popular but short-lasting topics of interest. A truly successful topic modeling strategy should be able to detect both types of topics and clearly locate them in time. In this work, we compare the variability of topic lengths discovered by several well-known topic modeling methods including latent Dirichlet allocation (LDA), nonnegative matrix factorization (NMF), as well as its tensor counterparts based on the nonnegative CANDECOMP/PARAFAC tensor decomposition (NCPD and Online NCPD). We demonstrate that only tensor-based methods with the dedicated mode for tracking time evolution successfully detect short-lasting topics. Furthermore, these methods are considerably more accurate in discovering the points in time when topics appeared and disappeared compared to the matrix-based methods such as LDA and NMF. We propose quantitative ways to measure the topic length and demonstrate the ability of NCPD (as well as its online variant), to discover short and long-lasting temporal topics in semi-synthetic and real-world data including news headlines and COVID-19 related tweets.

Method

Nonnegative CP Tensor Decomposition (NCPD) is a tool for decomposing higher-dimensional data tensors into interpretable lower-dimensional representations. NCPD factorizes a tensor into a sum of nonnegative component rank-one tensors, defined as outer products of nonnegative vectors [1, 3]. More precisely, given a third-order tensor $\mathcal{X} \in \mathbb{R}_+^{n_1 \times n_2 \times n_3}$ and a fixed integer $r > 0$, the approximate NCPD of $\mathcal{X}$ seeks matrices $\mathbf{A} \in \mathbb{R}_+^{n_1 \times r}, \mathbf{B} \in \mathbb{R}_+^{n_2 \times r}, \mathbf{C} \in \mathbb{R}_+^{n_3 \times r}$, such that $\mathcal{X} \approx \sum_{k=1}^{r} a_k \otimes b_k \otimes c_k$, where the nonnegative vectors a_k, b_k, and c_k are the columns of $\mathbf{A}, \mathbf{B}$, and $\mathbf{C}$, respectively. The matrices $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ are referred to as the NCPD *factor matrices*.



One can note that given a large amount of time-stamped documents, such as news articles or tweets, topic evolution frequently happens not from one document to the next in time, but rather from a batch of nearly simultaneous documents to the next. Therefore, one can encode the entire corpus of documents as a *3-dimensional tensor* where the three modes correspond to *words*, *relatively simultaneous documents*, and *time*, respectively.

We believe the role of nonnegativity constraint on the temporal mode is crucial for the NCPD-type methods to be able to detect short-lasting topics. Indeed, NMF is well-known to be able to extract spatially localized features when applied to image data [4] by using nonnegativity constraint on the spatial mode. Being a 3D analogue of NMF, NCPD should be able to extract spatio-temporally localized features, which correspond to 'short-lasting' (temporally localized) 'topics' (spatially localized features) in our context of dynamic topic modeling.

For large tensors, the computational cost of applying NCPD to the entire tensor may be large compared to the LDA or NMF. To address this concern, we also apply a recently proposed online version of NCPD (ONCPD) [5].

Quantifying lengths of topics

For a fraction $\alpha \in [0, 1]$ and the topic τ, its α-*effective length* denoted by $\ell_\alpha(\tau)$, is defined as

$$\ell_\alpha(\tau) := \max_{i \in [n]} \begin{cases} \min\{l \mid \sum_{i:i+l} > \alpha\} & \text{if } \sum_{i:n} > \alpha, \\ 0 & \text{otherwise,} \end{cases} \qquad \text{where } \sum_{i_1:i_2} := \sum_{j=i_1}^{i_2} \tilde{\mathbf{T}}[\tau, j].$$

The matrix $\tilde{\mathbf{T}} \in \mathbb{R}_+^{r \times n}$ is the matrix representing the dynamics of the topics over time with the rows normalized to add up to 1. Informally, the metric captures how many consecutive days are required for each topic to include a certain proportion of its whole "mass".

Semi-synthetic Benchmark Dataset

Semi-synthetic 20 Newsgroups Dataset:

- The 20 Newsgroups dataset is a collection of documents divided into six groups partitioned into subjects, with a total of 20 subgroups.
- We consider only five categories: "Atheism", "Space", "Baseball", "For Sale", and "Windows X" with a total of 1040 documents.

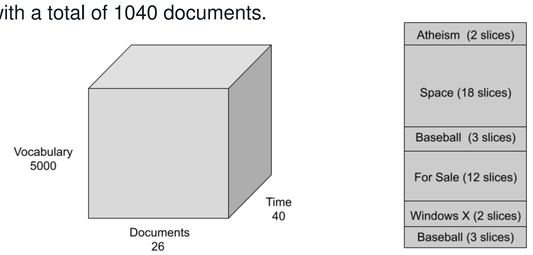


Fig. 2: Semi-synthetic 20 Newsgroups tensor construction.

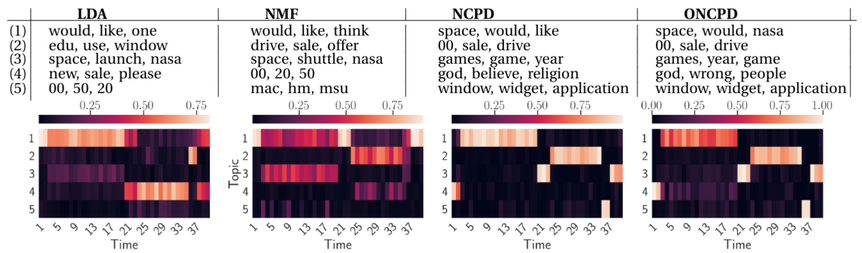


Fig. 3: The learned topics and prevalence of each extracted topic from the semi-synthetic 20 Newsgroups dataset for the four methods. NCPD and ONCPD identify topics associated with each subject and accurately indicates the temporal occurrence of each subject, while NMF and LDA learn topics that are prevalent during time slices associated with multiple subjects.

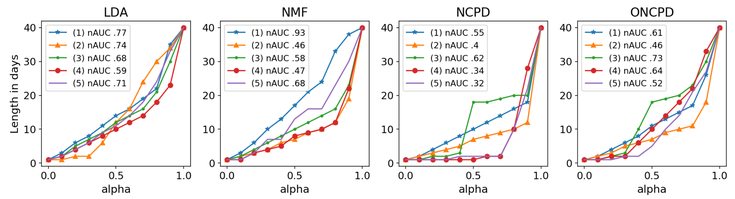


Fig. 4: Plot of the α-effective lengths of all 5 topics against $\alpha \in (0, 1)$ of the 20news dataset over LDA, NMF, NCPD, and ONCPD methods. The normalized area under the curve (nAUC) is given for each topic in the legend. The legends contain topic numbers referring to the topic numbers in Figure 3.

With an elbow method, NCPD discovers two short-lasting topics (topics 4 and 5) with the 0.7-effective length of one day, two topics (topics 2 and 1) of 0.9-effective lengths of 10 and 18 days, respectively, and one topic (topic 3) of 0.9-effective length of 20 days that also has 0.4-effective length of only 2 days (which is, precisely the lengths of these created topics).

Real-world Datasets

Twitter COVID-19 data:

- We consider Twitter text data related to the COVID-19 pandemic from Feb. 1 to May 1 of 2020 [2].
- Specifically, we consider the top 1000 retweeted English tweets from each day.

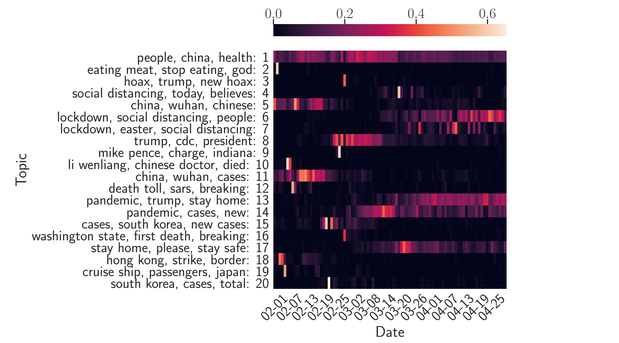


Fig. 5: The normalized factor matrix of NCPD with rank 20. Each column of the heatmap indicates the distribution over the extracted topics for each day. NCPD detects various short-lasting events (e.g. Topics 2, 3, 9, 10, 16, 19).

Million Headlines Dataset:

- Dataset containing news headlines published from the years 2003 to 2019 sourced from the Australian news source ABC.
- We consider 700 headlines randomly selected per month with a total of 142,100 headlines in the entire dataset.

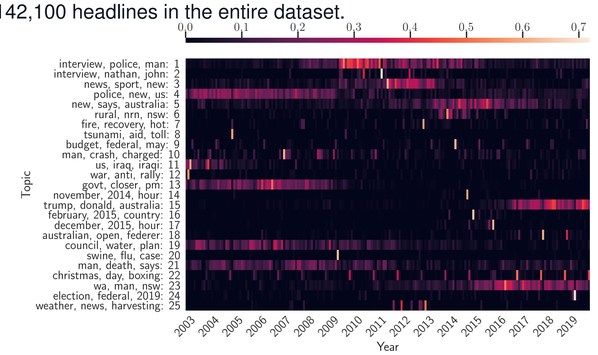


Fig. 6: The normalized factor matrix of NCPD on the News Headlines dataset with rank 25. NCPD discovers a range of short-lasting (Topics 8, 20, 24) and periodic events (e.g. Topic 22).

References

[1] J Douglas Carroll and Jih-Jie Chang. "Analysis of Individual Differences in Multidimensional Scaling via an N-way Generalization of "Eckart-Young" Decomposition". In: *Psychometrika* 35.3 (1970), pp. 283–319.

[2] Emily Chen, Kristina Lerman, and Emilio Ferrara. "Tracking Social Media Discourse About the COVID-19 Pandemic: Development of a Public Coronavirus Twitter Data Set". In: *JMIR Public Health and Surveillance* 6.2 (2020), e19273.

[3] Richard A Harshman et al. "Foundations of the PARAFAC Procedure: Models and Conditions for an "Explanatory" Multimodal Factor Analysis". In: (1970).

[4] Daniel D Lee and H Sebastian Seung. "Learning the Parts of Objects by Non-negative Matrix Factorization". In: *Nature* 401.6755 (1999), p. 788.

[5] Christopher Strohmeier, Hanbaek Lyu, and Deanna Needell. "Online nonnegative tensor factorization and CP-Dictionary Learning for Markovian data". In: *arXiv preprint arXiv:2009.07612* (2020).