COVID-19 Multidimensional Kaggle Literature Organization

Maksim E. Eren*
meren1@umbc.edu
USA

Nick Solovyev*
sonic1@umbc.edu
USA

Chris Hamer*
chamer1@umbc.edu
USA

Renee McDonald*
rp53139@umbc.edu
USA

Boian S. Alexandrov†
boian@lanl.gov
USA

Charles Nicholas*
nicholas@umbc.edu
USA

ABSTRACT

The unprecedented outbreak of Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2), or COVID-19, continues to be a significant worldwide problem. As a result, a surge of new COVID-19 related research has followed suit. The growing number of publications requires document organization methods to identify relevant information. In this paper, we expand upon our previous work with clustering the CORD-19 dataset by applying multidimensional analysis methods. Tensor factorization is a powerful unsupervised learning method capable of discovering hidden patterns in a document corpus. We show that a higher-order representation of the corpus allows for the simultaneous grouping of similar articles, relevant journals, authors with similar research interests, and topic keywords. These groupings are identified within and among the latent components extracted via tensor decomposition. We further demonstrate the application of this method with a publicly available interactive visualization of the dataset.

CCS CONCEPTS

• **Information systems** → **Document topic models**; • **Computing methodologies** → **Information extraction**; **Topic modeling**.

KEYWORDS

COVID-19, tensor factorization, CP decomposition, document organization

ACM Reference Format:
Maksim E. Eren, Nick Solovyev, Chris Hamer, Renee McDonald, Boian S. Alexandrov, and Charles Nicholas. 2021. COVID-19 Multidimensional Kaggle Literature Organization. In *ACM Symposium on Document Engineering 2021 (DocEng '21), August 24–27, 2021, Limerick, Ireland*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3469096.3474927

*Department of Computer Science and Electrical Engineering, UMBC
†Theoretical Division, Los Alamos National Laboratory

DocEng '21, August 24–27, 2021, Limerick, Ireland

ACM ISBN 978-1-4503-8596-1/21/08…$15.00
https://doi.org/10.1145/3469096.3474927

1 INTRODUCTION

The world continues the fight against the Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2), or COVID-19. Now, more than a year removed from the start of the pandemic, COVID-19 research has not dwindled. As of July 2021, a Google Scholar search for "covid 19" shows 117,000 new publications in the last seven months. The online statistical document analysis tool, *Dimensions Database*, estimates over 534,000 new COVID-19 publications since 2020 [5]. This overwhelming quantity of data makes the discovery of better document organization methods even more urgent. Document organization methods can assist with new research through information retrieval and sharing.

The focus of this paper is organizing the *COVID-19 Open Research Dataset* (CORD-19) [18][1]. CORD-19 is a collection of over 400,000 scholarly articles about COVID-19 and related diseases. In our previous work on this dataset, we showed that investigation of the CORD-19 corpus can be simplified through clustering and dimensionality reduction using *t-SNE*, *PCA*, and *k-means* [7]. The Kaggle notebook from our prior research has attracted great interest in the data science community[2]. In this paper, we continue to tackle the CORD-19 organization problem with a different approach by applying a multidimensional analysis method.

Tensor decomposition is an unsupervised learning method capable of extracting multifaceted latent patterns from a dataset. Scientific papers are a type of data that is naturally multidimensional and can be represented as a tensor. Analyzing documents in a higher dimensional space allows for simultaneously finding correlations across each dimension. This approach provides for a more natural representation of the data in comparison to traditional matrix factorization methods. Specifically, we build a tensor with dimensions corresponding to author, title, journal, and keywords in the paper to characterize the documents in the CORD-19 dataset.

In our work, we show that authors with similar research interests, relevant articles, and related journals can be grouped in and among the latent components through tensor factorization. At the same time, by representing the corpus vocabulary with keywords as a tensor dimension, we can identify the topic keywords for the papers. To the best of our knowledge, we are the first to use tensor analysis to organize the CORD-19 dataset. Finally, we present our results on a publicly available interactive visualization of the components we extracted via tensor decomposition[3].

[1]Dataset is available at https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge
[2]Kaggle notebook for the prior work is available at https://www.kaggle.com/maksimeren/covid-19-literature-clustering
[3]Interactive visualization is available at https://maksimekin.github.io/CORD19-Tensor/

2 RELEVANT WORK

Tensor and matrix decomposition and their application to text analysis is an area that has been widely studied. In this section we present a brief summary of related research.

An important and difficult problem for both matrix and tensor decomposition is determining the number of latent topics in a corpus. Vangara et al. factorize an ensemble of term frequency inverse document frequency (TF-IDF) matrices and apply k-means clustering to identify the rank of those matrices [16], using a distributed software package [3, 4]. In their work, the number of latent topics is chosen to be the rank that returns the best combination of a high *silhouette score* and a low *relative reconstruction error* for the factorization. In comparison, we utilize the *cosine similarity* score to identify the distinct topics over an ensemble of ranks rather than identifying a single optimal rank.

Larsen et al. introduce random sampling methods for faster computation of large tensors in [13]. They apply their methods to analyse *Reddit*[4] posts using a tensor with dimensions *Subreddit* x *User* x *Word*, where the entries in the tensor are $log(1 + \text{word count})$. They observe that when the highest values in latent factors for *Subreddit* dimension are plotted, similar subreddits[5] group together in the components. Similarly, we extract n elements with the highest values in the latent factors to determine groupings, and use the log of word counts as tensor entries.

Tensor decomposition has been applied to medical data in previous studies including identification of chronic diseases [17] and analysis of neurodevelopmental disorders [15]. Tensors have also been used to organize biomedical texts. Drakopoulos et al. built a tensor with the dimensions *Term* x *Keyword* x *Document* which is a generalization of the *term-document* matrix. They use TF-IDF values as tensor entries, and the clustering is done using k-means [6]. In our work, we perform analysis over a four-dimensional tensor and find that components extracted via factorization can separate the documents, authors, and journals into groups and extract topic keywords.

Several document analysis methods applied to the CORD-19 dataset have been presented previously [8]. Grotheer et al. demonstrated the use of hierarchical non-negative matrix factorization in [8]. In their analysis, the corpus is represented as a *Word* x *Document* matrix. Given a number of topics, they decomposed the matrix into the *dictionary* and *coding* matrices. From there, the documents can be sorted into topic matrices using the *coding matrix*. The process can be repeated with the "leaf" matrices codifying the corpus into a hierarchical tree [8]. This approach, while yielding promising results on the CORD-19 dataset, is limited since the information is represented in two dimensional space. With higher dimensional analysis, we can observe the latent information regarding each dimension including the topic keywords which get extracted in an unsupervised manner.

The remainder of the paper consists of a description of our tensor decomposition methods, the results of applying those methods to the CORD-19 dataset, and a concluding discussion regarding the effectiveness of our strategy.

[4] *Reddit* is an online content sharing platform.
[5] *Subreddits* are topic specific discussion boards in *Reddit*.

3 METHODS

Most of the pre-processing steps presented in Section 3.1 improve upon the data cleaning methods from our prior work on the same dataset [7]. Then, Section 3.2 describes the details of tensor construction and analysis of the latent factors.

3.1 Data Pre-processing

Before cleaning, the CORD-19 corpus consists of over 400,000 scholarly articles. We first drop the documents that lack a text body and/or are written in a language other than English. Next, the title and the journal name are shifted to lower-case, and numbers and special characters are removed. The duplicate papers (i.e. documents with the same title or documents with the same abstract) are dropped. After these steps, we are left with 128,359 unique articles published in 10,321 journals by 105,300 distinct first authors in the corpus. The stop-words are then removed from the articles and the text is tokenized using SpaCy biomedical [6] [14] and English parsers [9]. The word tokens are used to remove names, words with numerical values, and special characters. We then identify and remove DNA sequence patterns using the standard Python Regex function. Finally, we remove typos and nonsense words using the *nonsense* Python library [10, 11]. The SpaCy library was also used to perform lemmatization. After the text is processed, the documents in the dataset contain a total of 821,410 unique words.

3.2 Tensor Factorization

In our analysis, we use the Canonical Polyadic (CP) decomposition. For a d dimensional tensor, CP decomposition is written as:

$$\mathcal{X} \approx \mathcal{M} = [\![\lambda \; ; \; \mathbf{A}^{(1)}, \mathbf{A}^{(2)}, \cdots, \mathbf{A}^{(d)}]\!] \tag{1}$$

where $\mathcal{M}$ is the low-rank R approximation of $\mathcal{X}$, and computed as:

$$\mathcal{M} \equiv \sum_{r=1}^{R} \lambda_r \mathbf{a}_r^{(1)} \circ \mathbf{a}_r^{(2)} \circ \cdots \circ \mathbf{a}_r^{(d)} \tag{2}$$

where $\circ$ is the outer product of the latent factors $\mathbf{a}_r^{(d)}$ that are normalized to sum up to 1, and the weight is absorbed by each λ_r. Finally, $\mathbf{A}^{(d)}$ is the set of R latent factor vectors for dimension d:

$$\mathbf{A}^{(d)} = \{\mathbf{a}_1^{(d)}, \mathbf{a}_2^{(d)}, \ldots, \mathbf{a}_R^{(d)}\} \tag{3}$$

For further information on tensors and CP in particular, see [12]. We build an order four tensor $\mathcal{X}$ shaped *105300* x *128359* x *10321* x *821410*, where the dimensions of the tensor are *1st Author* x *Document* x *Journal* x *Words*. Here the *Document* dimension represents the title of the articles. An entry in this tensor is $\mathcal{X}_{a,p,j,w} = log(1 + x)$, where x is number of times the first author a used the word w in document p that was published in journal j. There are approximately $1.15x10^{20}$ entries in $\mathcal{X}$, but only 63,418,308 (or $5.53x10^{-11}$%) of them are non-zero. Hence, we exploit this extreme sparsity, and store $\mathcal{X}$ in *COO* format where only the non-zero entries are represented by a list of coordinates along with the list of non-zero values for each coordinate.

[6] SciSpacy is used for text tokenization.

Figure 1: Component 7

Figure 2: Component 16

Figure 3: Component 21

Figure 4: Component 56

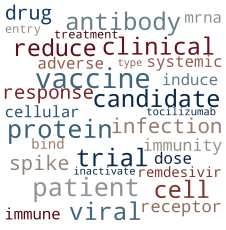

Figure 5: Component 66

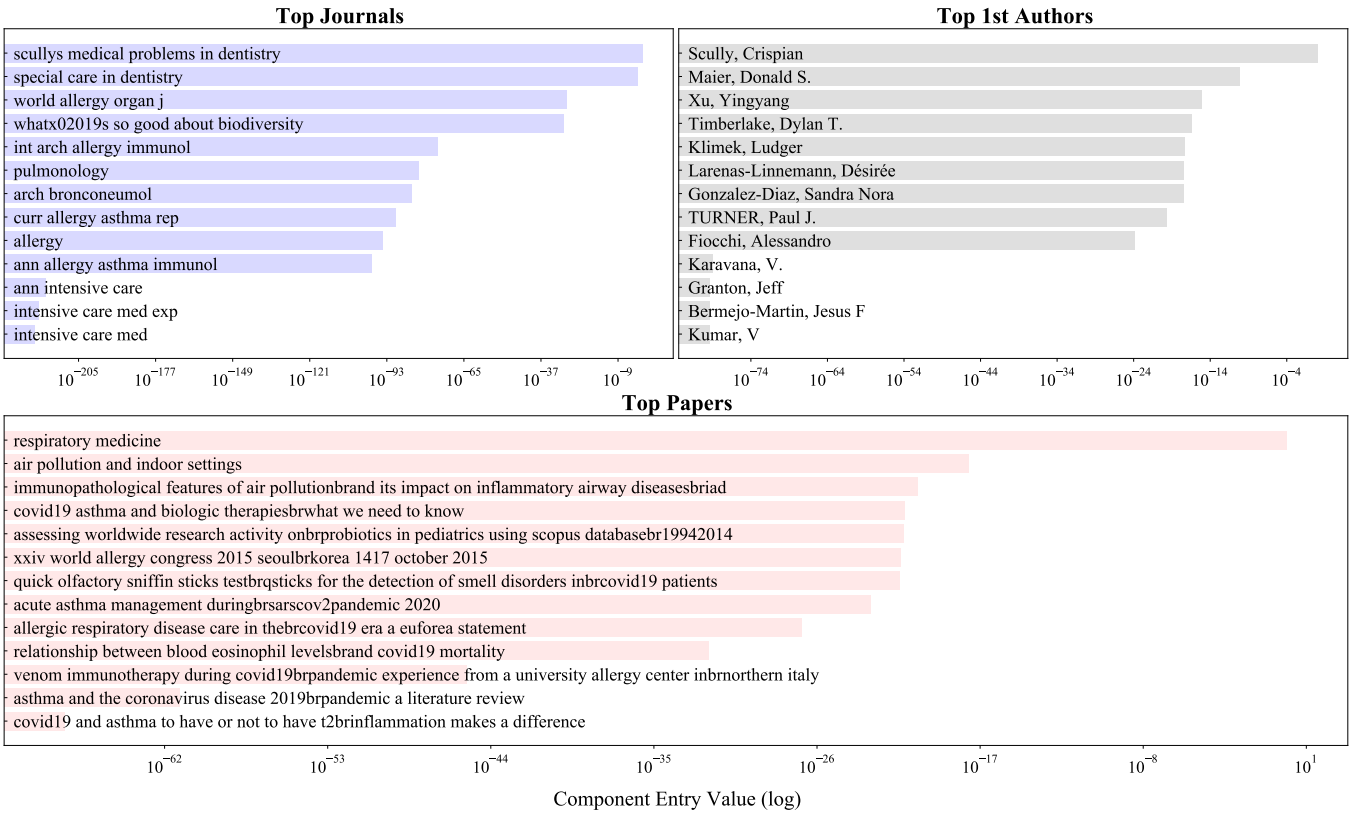

Figure 6: The top 13 values in the latent factors of component 56. The topic keywords for this component are shown in Figure 4 as a word-cloud. Tensor decomposition placed the papers and journals related to air, asthma, and pollution in this component. The *Journals* factor shows similar journals in close proximity. The authors in this component include the ones who have published the listed papers in same or relevant journals.

We factorize this tensor with the CP Alternating Least Squares (CP-ALS) algorithm [1, 2][7]. $\mathcal{X}$ is decomposed for ranks $R = i \in K : K = \{20, 40, 60, 80, 100, 120, 200\}$. Since the extracted latent factors for the *Words* dimension represent the topic keywords, we employ the *cosine similarity* scores when comparing each of the factor vectors $\mathbf{a}_r^{(\text{Words})}$, where $r \in \{1, \cdots, i\}$ and $i \in K$, to limit the components to unique topics. We use the *cosine similarity* metric because it measures how much two factors point in the same "direction" (i.e. how similar their entry value distributions are).

After filtering the components with a (somewhat arbitrary) *cosine similarity* score threshold of 0.35 or higher, we select 73 out of 620 components. During our analysis, we look at each of the 73 components, and plot the entries with the highest values for three of the latent factor vectors $\{\mathbf{a}_r^{(1st\ Author)}, \mathbf{a}_r^{(\text{Document})}, \mathbf{a}_r^{(\text{Journal})}\}$ to observe the groupings of relevant documents, journals, and authors. The latent factors for the fourth dimension, $\mathbf{a}_r^{(\text{Words})}$, are used to form the word-clouds representing the topic keywords.

[7] CP-ALS is available with MATLAB Tensor Toolbox: https://www.tensortoolbox.org/cp_als_doc.html

4 RESULTS

Modeling the corpus in a multidimensional space allowed us to analyze the results using the information from each dimension simultaneously. Another benefit of using tensor decomposition over the various "black-box" machine learning methods is the interpretability of the results. The values in the extracted latent factors can indicate meaningful relationships in the data. We therefore report our findings by manually inspecting the paper, journal, author, and word groupings in the components.

We first look at the word clouds of topic keywords obtained from the *Words* dimension in each component. In Figures 1, 2, 3, 4, and 5 we see that the relevant words are grouped. Specifically, Figure 1 includes terms pertaining to public health. Terms concerning education during the pandemic are grouped in Figure 2. Figure 3 contains information regarding aviation security. Terms related to vaccination are collected in Figure 5. Indeed, the extracted topic keywords semantically parallel the documents, journals, and author publications grouped in each of the respective components. We provide an example of this with Component 56 in Figure 6 which focuses on respiratory studies as shown in Figure 4.

Component 56 is one of our more interesting results. As the keywords indicate, the papers and journals listed in this component focus on respiratory issues caused and/or exacerbated by COVID-19. Using our interactive visualization, we identify patterns preserved by the tensor factorization. Journals, papers, and author publications in this component address topics such as "asthma", "air pollution", and "allergy". While the majority of the papers belong to a single journal, "World Allergy Organization Journal", we were also able to cluster other relevant journals such as "International Archives of Allergy and Immunology" and "Current Allergy and Asthma Reports". We also notice distinct groupings of journals concerning other research fields in the same component.

Several components grouped together single authors who specialize in a niche area of research. For instance, one component grouped together a single author "Ruwantissa Abeyratne" who has produced several articles on aviation law. Furthermore, CORD-19 includes textbooks divided into individual chapters and our methodology produced several components that identified and grouped together textbooks by the same author. More prolific authors (most notably "Theodore Tulchinsky") have multiple components dedicated to their textbooks. Despite this, we were able to extract relevant keywords for these components. For instance, the word cloud for Tulchinsky's components includes words pertaining to public health policy.

5 CONCLUSION

In this paper, we expanded upon our prior work in organizing the COVID-19 literature using tensor analysis. We showed that using a higher-order representation of the documents allow capturing of the topic keywords for the papers found in the CORD-19 corpus. We also observed groupings of similar articles, journals, and researchers within and among the components obtained by taking the CP decomposition of our four dimensional CORD-19 tensor. Future work can consider using a non-negative tensor factorization algorithm instead of CP-ALS to improve interpretability, and explore different combinations of tensor dimensions.

6 ACKNOWLEDGMENT

This manuscript has been approved for unlimited release and has been assigned LA-UR-21-25094. This research was partially funded by the Los Alamos National Laboratory (LANL) Laboratory Directed Research and Development (LDRD) grant 20190020DR and LANL Institutional Computing Program, supported by the U.S. Department of Energy National Nuclear Security Administration under Contract No. 89233218CNA000001.

REFERENCES

[1] Brett W. Bader, Tamara G. Kolda, et al. 2015. MATLAB Tensor Toolbox Version 2.6. Available online. http://www.sandia.gov/~tgkolda/TensorToolbox/

[2] Casey Battaglino, G. Ballard, and T. Kolda. 2018. A Practical Randomized CP Tensor Decomposition. *SIAM J. Matrix Anal. Appl.* 39 (2018), 876–901.

[3] Manish Bhattarai, Gopinath Chennupati, Erik Skau, Raviteja Vangara, Hristo Djidjev, and Boian S Alexandrov. 2020. Distributed Non-Negative Tensor Train Decomposition. , 10 pages.

[4] Manish Bhattarai, Ben Nebgen, Erik Skau, Maksim Eren, Gopinath Chennupati, Raviteja Vangara, Hristo Djidjev, John Patchett, Jim Ahrens, and Boian ALexandrov. 2021. pyDNMFk: Python Distributed Non Negative Matrix Factorization. https://github.com/lanl/pyDNMFk. https://doi.org/10.5281/zenodo.4722448

[5] Dimensions. [n.d.]. COVID-19 Report: Publications, Clinical Trials, Funding. https://reports.dimensions.ai/covid-19/. Accessed on 04.30.2021.

[6] Georgios Drakopoulos, Andreas Kanavos, Ioannis Karydis, Spyros Sioutas, and Aristidis G. Vrahatis. 2017. Tensor-Based Semantically-Aware Topic Clustering of Biomedical Documents. *Computation* 5, 3 (2017). https://doi.org/10.3390/computation5030034

[7] Maksim Ekin Eren, Nick Solovyev, Edward Raff, Charles Nicholas, and Ben Johnson. 2020. COVID-19 Kaggle Literature Organization. In *Proceedings of the ACM Symposium on Document Engineering 2020* (Virtual Event, CA, USA) *(DocEng '20)*. Association for Computing Machinery, New York, NY, USA, Article 15, 4 pages. https://doi.org/10.1145/3395027.3419591

[8] Rachel Grotheer, Longxiu Huang, Yihuan Huang, Alona Kryshchenko, Oleksandr Kryshchenko, Pengyu Li, Xia Li, Elizaveta Rebrova, Kyung Ha, and Deanna Needell. 2020. COVID-19 Literature Topic-Based Search via Hierarchical NMF. (2020). https://www.aclweb.org/anthology/2020.nlpcovid19-2.4.pdf

[9] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. *spaCy: Industrial-strength Natural Language Processing in Python.* https://doi.org/10.5281/zenodo.1212303

[10] Michael Hucka. 2018. Nostril: A nonsense string evaluator written in Python. *Journal of Open Source Software* 3, 25 (2018), 596. https://doi.org/10.21105/joss.00596

[11] Michael Hucka. 2019. *casics/nostril: Version 1.2.0 – Change license to LGPL.* CaltechDATA. https://doi.org/10.22002/D1.1313

[12] Tamara G. Kolda and Brett W. Bader. 2009. Tensor Decompositions and Applications. *SIAM Rev.* 51, 3 (September 2009), 455–500. https://doi.org/10.1137/07070111X

[13] Brett W Larsen and Tamara G Kolda. 2020. Practical leverage-based sampling for low-rank tensor decomposition. *arXiv preprint arXiv:2006.16438* (2020).

[14] Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. *CoRR* abs/1902.07669 (2019). arXiv:1902.07669 http://arxiv.org/abs/1902.07669

[15] Hamdi S.M.and Wu Y.and Boubrahimi S.F.and Angryk R.and Krishnamurthy L.C.and Morris R. 2018. Tensor Decomposition for Neurodevelopmental Disorder Prediction. *Brain Informatics. BI 2018. Lecture Notes in Computer Science* 11309 (2018).

[16] Raviteja Vangara, Erik Skau, Gopinath Chennupati, Hristo Djidjev, Thomas Tierney, James P. Smith, Manish Bhattarai, Valentin G. Stanev, and Boian S. Alexandrov. 2020. Semantic Nonnegative Matrix Factorization with Automatic Model Determination for Topic Modeling. In *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*. 328–335. https://doi.org/10.1109/ICMLA51294.2020.00060

[17] Haolin Wang, Qingpeng Zhang, Frank Youhua Chen, Eman Yee Man Leung, Eliza Lai Yi Wong, and Eng-Kiong Yeoh. 2019. Tensor Factorization-based Prediction with an Application to Estimating the Risk of Chronic Diseases. *bioRxiv* (2019). https://doi.org/10.1101/810556 arXiv:https://www.biorxiv.org/content/early/2019/10/18/810556.full.pdf

[18] Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Darrin Eide, Kathryn Funk, Rodney Michael Kinney, Ziyang Liu, William Merrill, P. Mooney, D. Murdick, Devvret Rishi, J. Sheehan, Zhihong Shen, Brandon Stilson, Alex D Wade, Kuansan Wang, Christopher Wilhelm, Boya Xie, Douglas A. Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. 2020. CORD-19: The COVID-19 Open Research Dataset. *ArXiv* (2020).