SeNMFk-SPLIT: Large Corpora Topic Modeling by Semantic Non-negative Matrix Factorization with Automatic Model Selection

Maksim E. Eren
maksim@lanl.gov
Theoretical Division, Los Alamos National Laboratory
USA

Nick Solovyev
nks@lanl.gov
Theoretical Division, Los Alamos National Laboratory
USA

Manish Bhattarai
ceodspspectrum@lanl.gov
Theoretical Division, Los Alamos National Laboratory
USA

Kim Ø. Rasmussen
kor@lanl.gov
Theoretical Division, Los Alamos National Laboratory
USA

Charles Nicholas
nicholas@umbc.edu
CSEE, University of Maryland, Baltimore County
USA

Boian S. Alexandrov
boian@lanl.gov
Theoretical Division, Los Alamos National Laboratory
USA

ABSTRACT

As the amount of text data continues to grow, topic modeling is serving an important role in understanding the content hidden by the overwhelming quantity of documents. One popular topic modeling approach is non-negative matrix factorization (NMF), an unsupervised machine learning (ML) method. Recently, Semantic NMF with automatic model selection (SeNMFk) has been proposed as a modification to NMF. In addition to heuristically estimating the number of topics, SeNMFk also incorporates the semantic structure of the text. This is performed by jointly factorizing the term frequency-inverse document frequency (TF-IDF) matrix with the co-occurrence/word-context matrix, the values of which represent the number of times two words co-occur in a predetermined window of the text. In this paper, we introduce a novel distributed method, SeNMFk-SPLIT, for semantic topic extraction suitable for large corpora. Contrary to SeNMFk, our method enables the joint factorization of large documents by decomposing the word-context and term-document matrices separately. We demonstrate the capability of SeNMFk-SPLIT by applying it to the entire artificial intelligence (AI) and ML scientific literature uploaded on arXiv.

CCS CONCEPTS

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

KEYWORDS

non-negative matrix factorization, topic modeling, document organization, model selection, semantic

DocEng '22, September 20–23, 2022, San Jose, California

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

ACM Reference Format:
Maksim E. Eren, Nick Solovyev, Manish Bhattarai, Kim Ø. Rasmussen, Charles Nicholas, and Boian S. Alexandrov. 2022. SeNMFk-SPLIT: Large Corpora Topic Modeling by Semantic Non-negative Matrix Factorization with Automatic Model Selection. In *ACM Symposium on Document Engineering 2022 (DocEng '22), September 20–23, 2022, San Jose, California.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/xxxxxx.xxxxxxx

1 INTRODUCTION

According to a recent report, 2.9 million research articles are published annually and that number is growing at the rate of 5% per year [24]. This substantial expansion of text data, including scientific literature, requires automated document engineering techniques that can handle large-scale data to produce actionable results. One important method for document organization is topic modeling, which maps a collection of documents into themes that summarize the hidden (or latent) features of those documents. By grouping text samples into topics via an unsupervised method, large datasets become easier to understand and process.

In our work, we base our methodology on non-negative matrix factorization (NMF), a popular unsupervised machine learning (ML) method for topic modeling. NMF performs a low-rank approximations of a given term frequency-inverse document frequency (TF-IDF) matrix $\mathbf{X} \in \mathbb{R}_{+}^{\mathrm{m}\times\mathrm{n}}$ (where m is the number of tokens in the vocabulary and n is the number of documents) by a product of two non-negative factor matrices $\mathbf{W} \in \mathbb{R}_{+}^{\mathrm{m}\times\mathrm{k}}$ and $\mathbf{H} \in \mathbb{R}_{+}^{\mathrm{k}\times\mathrm{n}}$, such that $\mathbf{X}_{ij} \approx \sum_{s}^{k} \mathbf{W}_{is}\mathbf{H}_{sj}$, and the low-rank $k \ll n, m$. $\mathbf{X}$ can be factorized using *multiplicative update*, a minimization algorithm with a non-negativity constraint, that minimizes the non-convex objective $||\mathbf{X} - \mathbf{W}\mathbf{H}||_F^2$, where $||...||_F$ denotes he Frobenius norm [11]. Here the columns of $\mathbf{W}$ represent the topics and the rows of $\mathbf{H}$ are the coordinates of the documents in the latent topic space. While NMF and its variants have been successfully applied to various corpora [8, 19, 23], classical NMF relies on manual selection of the number of latent topics k, which is crucial for identifying meaningful topics. Too small value of k can result on poor topic separation (*under-fitting*), and too large a value of k will result in noisy topics (*over-fitting*). In addition, traditional NMF does not incorporate semantics of the text, which is essential for extracting coherent topics [1, 16, 17].

Semantic NMF with automatic model determination (SeNMFk) is a topic modeling approach that incorporates semantics and estimates the number of latent features k to extract coherent and meaningful topics [22]. SeNMFk incorporates semantic information by joint factorization of the TF-IDF and co-occurrence/word-context matrices. In this paper, we introduce a new topic modeling method, named SeNMFk-SPLIT, that is designed for large-scale data. SeNMFk-SPLIT uses SPLIT where the TF-IDF and word-context matrices are factorized separately, which allows processing of smaller matrices with lower ranks.

We demonstrate the capability of SeNMFk-SPLIT by modeling topics on more than 168,000 abstracts about artificial intelligence (AI) and machine learning (ML) uploaded to arXiv. The distributed library used for SeNMFk-SPLITT is publicly available[1] [3].

2 RELEVANT WORK

Topic modeling provides a convenient way to analyze large amounts of text data. The goal is to segment set of documents on the basis of structural similarities between them. Commonly topic modeling techniques include Latent Semantic Analysis (LSA) [7], Probabilistic Latent Semantic Analysis (PLSA) [10], Latent Dirichlet Allocation (LDA) [5], and Non-negative Matrix Factorization (NMF) [25].

Bastani et al. used LDA to organize and analyze complaints filed with the Consumer Financial Protection Bureau [2]. Hazen applied PLSA to a dataset of transcribed telephone calls in an attempt to categorize and summarize conversations [9]. Most similar to this paper's dataset of choice, Rosca et al. used LDA on a collection of documents on AI-assisted legal research [15]. Rosca et al. manually determine 35 topics with the help of subject-matter-experts and make a conjecture on the evolution of AI-assisted legal research over the last 50 years. While PLSA and LDA are both probabilistic methods, NMF employs low-rank approximation where the matrix-based representation of the text is factorized to retrieve the manually selected k number of latent topics.

SeNMFk selects the number of latent topics k based on the stability of the extracted topics and includes semantics for the sake of extracting high-quality topics [22]. SeNMFk identifies the number of latent topics (automatic model selection) by analyzing sets of NMF minimizations via custom clustering and Silhouette statistics, used to estimate the robustness and accuracy of multiple NMF solutions for different values of the latent variable k [14]. In addition to the automatic model selection, SeNMFk incorporates semantic information from the text by jointly factorizing the term-document matrix $\mathbf{X}$ with word-context matrix $\mathbf{M} \in \mathbb{R}_+^{\mathrm{m}\times\mathrm{m}}$. Joint factorization yields more coherent, high quality topics [21]. However, the task of the joint factorization results in an increased memory and computational complexity, especially for a large number of topics and noisy data. In this work we introduce our method SPLIT for SeNMFk solutions. SPLIT enables joint factorization of large corpora by decomposing the word-context ($\mathbf{M}$) and term-document ($\mathbf{X}$) matrices separately first, then performing joint factorization of the resulting latent factors to find the common topics, and finally combines the latent information. This allows our approach to work on smaller matrices, breaking down the space complexity to separate operations.

[1] pyDNMFk: https://github.com/lanl/pyDNMFk

3 DATASET AND PRE-PROCESSING

We apply SeNMFk-SPLIT on the abstracts of AI/ML scientific literature uploaded to arXiv[2] [6]. We focus on papers that are self-reported categories, such as *cs.AI* (Artificial Intelligence), *cs.LG* (Machine Learning), *cs.CL* (Computation and Language), *cs.NE* (Neural and Evolutionary Computing), and *cs.MA* (Multiagent Systems). Our pre-processing procedure includes removal of common stop-words, symbols and next-line characters, non-ASCII characters, e-mail addresses, digits, and tags. We also apply lemmatization using the Python package *NLTK* [4], and exclude non-English abstracts using the Python implementation of *language-detection* software [18]. After removing the abstracts with less than 20 tokens, we obtain 168,177 documents in our corpus. Finally, when selecting the vocabulary we exclude tokens that appeared in less than five documents or in more than 50% of the corpus, further removing noise. After pre-processing, our final vocabulary includes 25,869 unique terms. We represent our corpus with a TF-IDF matrix $\mathbf{X} \in \mathbb{R}_+^{25,869\times168,177}$. The semantics of the text are represented with the word-context/co-occurrence matrix $\mathbf{M} \in \mathbb{R}_+^{25,869\times25,869}$ where the values represent the number of times two words co-occur in a predetermined window length of $w = 100$ tokens. We normalize $\mathbf{M}$ with Shifted Positive Point-wise Mutual Information (SPPMI) [12], with shift $s = 4$.

4 SENMFK-SPLIT

SeNMFk is a NMF method that automatically determines the number of latent topics and extracts coherent topics by exploiting the semantic representation encoded in the word-context matrix adjoined to the TF-IDF. Given the TF-IDF matrix $\mathbf{X}$ and normalized word-context SPPMI matrix $\mathbf{M}$, SeNMFk extracts the topics - the columns of the matrix $\mathbf{W}$, and the coordinates of the documents - the columns of the matrix $\mathbf{H}$, as well as a secondary mixing matrix $\mathbf{G}$. SeNMFk is performed by solving the joint optimization problem:

$$\underset{\mathbf{W}\in\mathbb{R}_+^{F\times k},\mathbf{H}\in\mathbb{R}_+^{k\times N},\mathbf{G}\in\mathbb{R}_+^{k\times F}}{\text{minimize}} \frac{1}{2}||\mathbf{X}-\mathbf{W}\mathbf{H}||_F^2 + \alpha||\mathbf{M}-\mathbf{W}\mathbf{G}||_F^2 \quad (1)$$

where $||.||_F^2$ is the Frobenius norm, and α is a regularization parameter controlling the weight of the semantic SPPMI in the decomposition. SeNMFk solves the above expression by concatenating the TF-IDF matrix $\mathbf{X}$ with the SPPMI matrix, $\mathbf{M}$, and applying pyDNMFk on the concatenation, [19].

SeNMFk-SPLIT includes the following optimizations,

$$\mathbf{W}_1, \mathbf{H}_1 = \underset{\mathbf{W}_1\in\mathbb{R}_+^{N\times k_1},\mathbf{H}_1\in\mathbb{R}_+^{k_1\times M}}{\text{minimize}} ||\mathbf{X}-\mathbf{W}_1\mathbf{H}_1||_F^2 \quad (2)$$

$$\mathbf{W}_2, \mathbf{H}_2 = \underset{\mathbf{W}_2\in\mathbb{R}_+^{N\times k_2},\mathbf{H}_2\in\mathbb{R}_+^{k_2\times N}}{\text{minimize}} ||\mathbf{M}-\mathbf{W}_2\mathbf{H}_2||_F^2 \quad (3)$$

Finally,

$$\mathbf{W}, \mathbf{H}^* = \underset{W\in\mathbb{R}_+^{N\times k},\mathbf{H}^*\in\mathbb{R}_+^{k\times(k_1+k_2)}}{\text{minimize}} ||[\mathbf{W}_1|\mathbf{W}_2] - \mathbf{W}\mathbf{H}^*||_F^2 \quad (4)$$

First, SeNMFk-SPLIT estimates the number of latent features k_1 and k_2 in Equations 2, 3, and extracts the factor matrices $[\mathbf{W}_1, \mathbf{H}_1]$, and

[2] arXiv Dataset: https://www.kaggle.com/datasets/Cornell-University/arxiv

$[\mathbf{W}_2, \mathbf{H}_2]$ via pyDNMFk. Next, it concatenates the normalized topic matrices $\mathbf{W}_1$ and $\mathbf{W}_2$ to find the common k factors ($k \le (k_1 + k_2)$), and to avoid taking into account factors that are co-linear, or/and linear combinations of other factors. To do that, the concatenated matrix $[\mathbf{W}_1|\mathbf{W}_2]$ is factorized by pyDNMFk, which results in a new topic matrix $\mathbf{W}$, with common latent topics, and the corresponding mixing matrix $\mathbf{H}^*$, Equation 4. The final decomposition of $\mathbf{X}$ is achieved by learning the coordinates of the documents in $\mathbf{X}$ in the space of the new common topics through a regression,

$$\mathbf{H} = \underset{\mathbf{H} \in \mathbb{R}_+^{k \times M}}{\text{minimize}} ||\mathbf{X} - \mathbf{W}\mathbf{H}||_F^2 \tag{5}$$

To obtain the final clusters of documents, we assign each document to a given topic via the *argmax* operation along the columns of the matrix $\mathbf{H}$. *Argmax* is a common operation for probabilistic multi-class classification. Since the columns of $\mathbf{H}$ are the coordinates of the documents in the space of the latent topics, here this assignment is equivalent to a soft clustering of the documents [20].

5 RESULTS

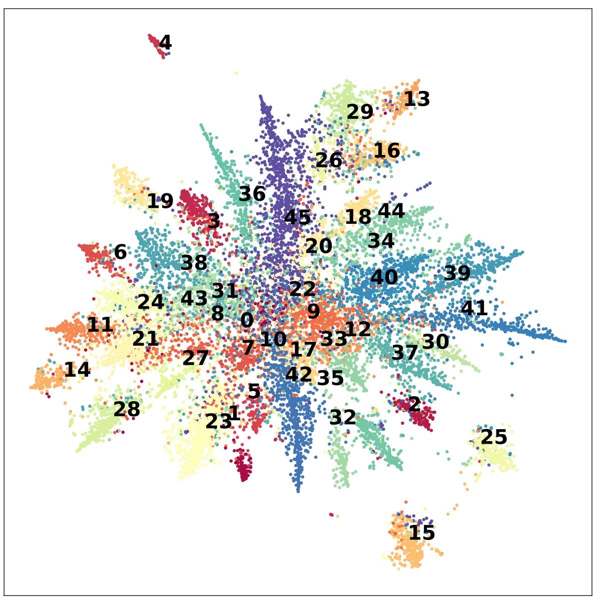

Figure 1: UMAP of H-clustering, where topic numbers are placed at cluster centroids, and color of each paper represents the topic assignments.

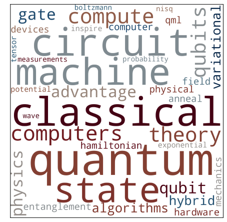

Figure 2: Quantum Computing (4)

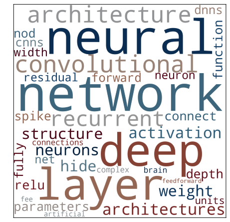

Figure 3: Deep Learning (9)

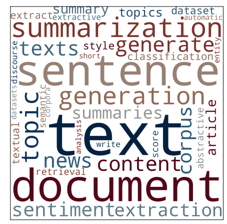

Figure 4: Document Engineering (24)

Figure 5: Tensor Factorization (37)

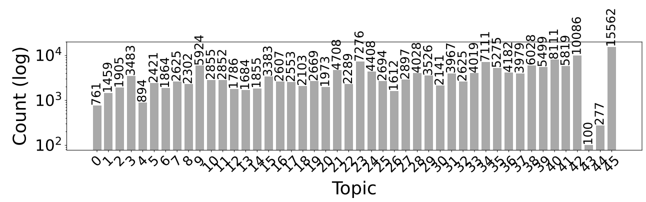

Figure 6: Distribution of the abstracts for the detected topics.

SeNMFk-SPLIT extracted 46 topics from our AI/ML corpus, where the 20 most prominent words from each topic are examined to determine the themes. The word clouds corresponding to these words are shown for four of the topics in Figures 2 3, 4, and 5. We have also listed the top five words and themes of each topic in Table 1. Word clouds and the top five words demonstrate that each topic features relevant words, highlighting the quality of the topics. In addition, we provide a Uniform Manifold Approximation and Projection (UMAP) [13] to visualise the clusters of abstract in Figure 1, where the topic number and the color of each document corresponds to different cluster obtained by **H**-clustering.

In Figure 1, documents with the same topic are grouped together. Furthermore, the clusters corresponding to related topics seem closer compared to unrelated ones. For example, topics Game Theory **(13)**, Reinforcement Learning **(16)**, Multi-agent Systems **(29)**, Robot Path Planning **(26)**, and Autonomous Vehicles **(45)** are related, and the clusters corresponding to these topics are overlapping. Similarly, the clusters corresponding to the topics Neural Machine Translation **(14)**, Word Embedding **(11)**, Attention Mechanism **(27)**, Natural Language Processing **(21)**, and Document Engineering **(24)**, are located in close proximity. This relationship can also be explored for the rest of the topics in Figure 1. In contrast, the clusters corresponding to the topics Quantum Computing **(4)**, Data Privacy **(25)**, and Adversarial Attacks **(15)** are further away from any other topic clusters as the research corresponding to these topics are emerging fields, and are relatively different directions when compared to the research associated with other topics. Manual investigation of a small set of documents also helps confirm the cogency of the document topic assignment. This demonstrates the ability of the SeNMFk-SPLIT to extract high-quality/interpretable topics.

Figure 6 displays the histogram of document topic assignments. Note the small number of documents published in the research

Table 1: Top 5 words in each topic, and the themes we have assigned to these topics based the top 20 words.

Topic	Top 5 Words	Theme
0	neural, datasets, performance, generative, model	Neural Network Theory
1	transfer, source, domains, domain, adaptation	Knowledge Generalization
2	unsupervised, spectral, algorithms, mean, cluster	Unsupervised Learning
3	structure, graph, nod, gnns, node	Graph Theory
4	state, quantum, classical, machine, circuit	Quantum Computing
5	noisy, semi, unlabeled, supervise, label	Weakly Supervised Systems
6	vqa, reason, question, query, answer	Visual Query Answering
7	methods, classification, extract, selection, fe...	Feature Extraction
8	performance, multi, transfer, task, meta	Meta-learning
9	convolutional, layer, network, deep, neural	Deep Learning
10	fine, tune, performance, pre, train	Model Tuning
11	sentence, semantic, embed, embeddings, word	Word Embedding
12	distributions, distribution, sample, generate, ...	Generative Systems
13	equilibrium, play, players, player, game	Game Theory
14	sentence, english, translation, nmt, machine	Neural Machine Translation
15	perturbations, examples, robustness, attack, ad...	Adversarial Attacks
16	action, policies, reward, policy, reinforcement	Reinforcement Learning
17	loss, classifiers, classifier, class, classific...	Classification Systems
18	space, architecture, query, nas, search	Neural Architecture Search
19	track, scene, video, visual, object	Computer Vision
20	rule, make, structure, decision, tree	Decision Trees and Random Forest
21	nlp, multilingual, natural, languages, language	Natural Language Processing
22	instance, consider, rule, point, set	Rule based Systems
23	caption, visual, medical, segmentation, image	Computer Vision
24	generation, sentence, document, summarization, ...	Document Engineering
25	clients, differential, private, federate, privacy	Data Privacy
26	path, robot, action, motion, plan	Robot Path Planning
27	mechanism, self, attention, sequence, transformer	Attention Mechanisms
28	end, asr, speaker, recognition, speech	Speech Recognition
29	communication, agent, environment, agents, multi	Multi-agent Systems
30	gaussian, linear, regression, kernel, kernels	Kernel Methods
31	relation, entities, semantic, information, entity	Information Theory
32	gender, fair, group, fairness, bias	Fairness in AI
33	generative, representation, representations, la...	Representation Learning
34	objective, solve, algorithms, problems, optimize...	Optimization problems
35	items, recommendation, item, users, user	Recommendation Systems
36	semantics, reason, program, logic, rule	Reasoning Systems
37	matrices, rank, matrix, tensor, sparse	Tensor Factorizations
38	evaluation, dialogue, generate, human, generation	Dialogue Systems
39	sgd, descent, gradient, stochastic, convergence	Optimizing Systems
40	distribution, uncertainty, causal, inference, b...	Casual AI
41	bandit, bind, arm, bound, regret	Multi-armed Bandits
42	predict, temporal, prediction, series, forecast	Predictive Analytics
43	schneider, snacs, supersenses, adpositions, adp...	Linguistic Analysis
44	gecco, umda, lehre, binval, dang	Evolutionary Algorithms
45	safety, dynamics, drive, systems, control	Autonomous Vehicles

topics Linguistic Analysis **(43)** and Evolutionary Algorithms **(44)** from Table 1, while a significant number of documents belong to the topics Computer Vision **(23)** and Autonomous Vehicles **(45)**. These statistics concur with the evolving research in the field of AI.

6 CONCLUSION

In this paper we have introduced SeNMFk-SPLIT, a new topic modeling method that performs automatic estimation of the number of topics and incorporates semantics in modeling. Our method expands upon SeNMFk with a design that can be utilized on large datasets. We demonstrated the capability of SeNMFk-SPLIT on a large corpus containing 168k+ abstracts from AI/ML scientific literature, and showed that our method extracts high-quality topics.

ACKNOWLEDGMENTS

This manuscript has been approved for unlimited release and has been assigned LA-UR-22-26571. This research was funded by the Los Alamos National Laboratory (LANL) Laboratory Directed Research and Development (LDRD) grant 20190020DR and the LANL Institutional Computing Program, supported by the U.S. Department of Energy National Nuclear Security Administration under Contract No. 89233218CNA000001.

REFERENCES

[1] Melissa Ailem, Aghiles Salah, and Mohamed Nadif. 2017. Non-negative matrix factorization meets word embedding. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Revelopment in Information Retrieval*. 1081–1084.

[2] Kaveh Bastani, Hamed Namavari, and Jeffrey Shaffer. 2019. Latent Dirichlet allocation (LDA) for topic modeling of the CFPB consumer complaints. *Expert Systems with Applications* 127 (2019), 256–271. https://doi.org/10.1016/j.eswa.2019.03.001

[3] 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

[4] Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.".

[5] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3, Jan (2003), 993–1022.

[6] Colin B. Clement, Matthew Bierbaum, Kevin P. O'Keeffe, and Alexander A. Alemi. 2019. On the Use of ArXiv as a Dataset. arXiv:1905.00075 [cs.IR]

[7] Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. 1990. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science (JASIS)* 41, 6 (1990), 391–407.

[8] Rundong Du, Da Kuang, Barry Drake, and Haesun Park. 2017. DC-NMF: nonnegative matrix factorization based on divide-and-conquer for fast clustering and topic modeling. *Journal of Global Optimization* 68, 4 (2017), 777–798.

[9] Timothy J Hazen. 2011. Latent topic modeling for audio corpus summarization. In *Twelfth Annual Conference of the International Speech Communication Association*.

[10] Thomas Hofmann. 2013. Probabilistic Latent Semantic Analysis. https://doi.org/10.48550/ARXIV.1301.6705

[11] Daniel D Lee and H Sebastian Seung. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature* 401, 6755 (1999), 788–791.

[12] Omer Levy and Yoav Goldberg. 2014. Neural word embedding as implicit matrix factorization. *Advances in neural information processing systems* 27 (2014).

[13] Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).

[14] Benjamin T Nebgen, Raviteja Vangara, Miguel A Hombrados-Herrera, Svetlana Kuksova, and Boian S Alexandrov. 2021. A neural network for determination of latent dimensionality in non-negative matrix factorization. *Machine Learning: Science and Technology* 2, 2 (2021), 025012.

[15] Constanta Rosca, Bogdan Covrig, Catalina Goanta, Gijs van Dijck, and Gerasimos Spanakis. 2020. Return of the AI: An Analysis of Legal Research on Artificial Intelligence using Topic Modeling. In *NLLP@KDD*.

[16] Aghiles Salah, Melissa Ailem, and Mohamed Nadif. 2018. Word co-occurrence regularized non-negative matrix tri-factorization for text data co-clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[17] Tian Shi, Kyeongpil Kang, Jaegul Choo, and Chandan K Reddy. 2018. Short-text topic modeling via non-negative matrix factorization enriched with local word-context correlations. In *Proceedings of the 2018 World Wide Web Conference*. 1105–1114.

[18] Nakatani Shuyo. 2010. Language Detection Library for Java. http://code.google.com/p/language-detection/

[19] Valentin Stanev, Erik Skau, Ichiro Takeuchi, and Boian S Alexandrov. 2021. Topic Analysis of Superconductivity Literature by Semantic Non-negative Matrix Factorization. In *International Conference on Large-Scale Scientific Computing*. Springer, 359–366.

[20] Raviteja Vangara, Manish Bhattarai, Erik Skau, Gopinath Chennupati, Hristo Djidjev, Tom Tierney, James P Smith, Valentin G Stanev, and Boian S Alexandrov. 2021. Finding the number of latent topics with semantic non-negative matrix factorization. *IEEE Access* 9 (2021), 117217–117231.

[21] Raviteja Vangara, Kim Ø Rasmussen, Dimiter N Petsev, Golan Bel, and Boian S Alexandrov. 2020. Identification of anomalous diffusion sources by unsupervised learning. *arXiv preprint arXiv:2010.02168* (2020).

[22] 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

[23] JianYu Wang and Xiao-Lei Zhang. 2021. Deep NMF topic modeling. *arXiv preprint arXiv:2102.12998* (2021).

[24] Karen White. 2021. Publications Output: US Trends and International Comparisons. *National Science Foundation* (2021).

[25] Wei Xu, Xin Liu, and Yihong Gong. 2003. Document clustering based on nonnegative matrix factorization. In *Proceedings of the 26th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 267–273.