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

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, Virtual Event, CA, USA. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3558100.3563844

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 $X \in \mathbb{R}^{n \times m}$ (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 $W \in \mathbb{R}^{m \times k}$ and $H \in \mathbb{R}^{k \times n}$, such that $X_{ij} = \sum_{k} W_{ik} H_{kj}$, and the low-rank $k \ll n, m$. $X$ can be factorized using multiplicative update, a minimization algorithm with a non-negativity constraint, that minimizes the non-convex objective $||X - WH||^2_F$, where $||..||_F$ denotes the Frobenius norm [11]. Here the columns of $W$ represent the topics and the rows of $H$ 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 a 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-SPLIT is publicly available¹ [3].

2 RELEVANT WORK

Topic modeling provides a convenient way to analyze large amounts of text data. The goal is to segment sets 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 $X$ with word-context matrix $M \in \mathbb{R}_{+}^{N \times M}$. 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 ($M$) and term-document ($X$) 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.

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

3 DATASET AND PRE-PROCESSING

We apply SeNMFk-SPLIT on the abstracts of AI/ML scientific literature uploaded to arXiv² [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 $X \in \mathbb{R}_{+}^{25,869 \times 168,177}$. The semantics of the text are represented with the word-context/co-occurrence matrix $M \in \mathbb{R}_{+}^{25,869 \times 25,869}$, where the values represent the number of times two words co-occur in a predetermined window length of $w = 100$ tokens. We normalize $M$ 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 associated to the TF-IDF. Given the TF-IDF matrix $X$ and normalized word-context SPMII matrix $M$, SeNMFk extracts the topics - the columns of the matrix $W$, and the coordinates of the documents - the columns of the matrix $H$, as well as a secondary mixing matrix $G$. SeNMFk is performed by solving the joint optimization problem:

$$\min_{W \in \mathbb{R}_{+}^{N \times k}, H \in \mathbb{R}_{+}^{k \times M}} \frac{1}{2}\|X - WH\|^2_F + \alpha\|M - WHG\|^2_F. \quad (1)$$

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

SeNMFk-SPLIT includes the following optimizations,

$$W_1, H_1 = \min_{W_1 \in \mathbb{R}_{+}^{N \times k_1}, H_1 \in \mathbb{R}_{+}^{k_1 \times M}} \|X - W_1H_1\|^2_F \quad (2)$$

$$W_2, H_2 = \min_{W_2 \in \mathbb{R}_{+}^{N \times k_2}, H_2 \in \mathbb{R}_{+}^{k_2 \times M}} \|M - W_2H_2\|^2_F \quad (3)$$

Finally,

$$W, H = \min_{W \in \mathbb{R}_{+}^{N \times k_1}, H \in \mathbb{R}_{+}^{k_1 \times M}, k_1 \leq (k_1 + k_2)} \|W_1 |W_2 - WH|\|^2_F \quad (4)$$

First, SeNMFk-SPLIT estimates the number of latent features $k_1$ and $k_2$ in Equations 2, 3, and extracts the factor matrices $[W_1, H_1]$, and $[W_2, H_2]$ via pyDNMFk. Next, it concatenates the normalized topic matrices $W_1$ and $W_2$ to find the common $k$ factors ($k \leq (k_1 + k_2)$), and to avoid taking into account factors that are co-linear, or/and

²arXiv Dataset: https://www.kaggle.com/datasets/Cornell-University/arxiv
linear combinations of other factors. To do that, the concatenated matrix \([W_1|W_2]\) is factorized by pyDNMFk, which results in a new topic matrix \(W\), with common latent topics, and the corresponding mixing matrix \(H\), Equation 4. The final decomposition of \(X\) is achieved by learning the coordinates of the documents in \(X\) in the space of the new common topics through a regression,

\[
H = \underset{H \in \mathbb{R}^{k \times M}}{\text{minimize}} ||X - WH||_F^2
\]  

(5)

To obtain the final clusters of documents, we assign each document to a given topic via the \(\text{argmax}\) operation along the columns of the matrix \(H\). \(\text{Argmax}\) is a common operation for probabilistic multi-class classification. Since the columns of \(H\) 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 topics Linguistic Analysis (43) and Evolutionary Algorithms (44) from Table 1, while a significant number of documents belong to
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, graph, node | Graph Theory |
| 4     | state, quantum, classical, machine, circuit | Quantum Computing |
| 5     | noisy, sen, unlabeled, supervised, label | Weakly Supervised Systems |
| 6     | vqa, reason, question, query, answer | Visual Query Answering |
| 7     | methods, classification, extract, selection, feat... | 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    | action, policies, reward, policy, reinforcement | Adversarial Attacks |
| 15    | loss, classifiers, classifier, class, classifier | Reinforcement Learning |
| 16    | space, architecture, query, nas, search | Classification Systems |
| 17    | track, scene, video, visual, object | Neural Architecture Search |
| 18    | rule, make, structure, decision, tree | Computer Vision |
| 19    | nlp, multilingual, natural, languages, language | Decision Trees and Random Forests |
| 20    | instance, consider, edit, point, art | Natural Language Processing |
| 21    | path, robot, act, motion, plan | Rule-Based Systems |
| 22    | mechanism, self, attention, sequence, transformer | Computer Vision |
| 23    | end, aes, speaker, recognition, speech | Natural Language Theory |
| 24    | communication, agent, environment, agents, multi | Speech Recognition |
| 25    | gaussian, linear, regression, kernel, kernels | Multi-agent Systems |
| 26    | relation, entities, semantic, information, entity | Kernel Methods |
| 27    | fairness, fair, group, fairness, bias | Information Theory |
| 28    | generative, representation, representations, la... | Fairness in AI |
| 29    | objective, solve, algorithms, problems, optimiza... | Representation Learning |
| 30    | item, recommendation, item, users, user | Optimization problems |
| 31    | semantics, reason, program, logic, rule | Recommendation Systems |
| 32    | matrices, rank, matrix, tensor, sparse | Reasoning Systems |
| 33    | evaluation, dialogue, generate, human, generation | Tensor Factors |
| 34    | sgd, descent, gradient, stochastic, convergence | Dialogue Systems |
| 35    | distribution, uncertainty, caption, inference, b... | Optimizing Systems |
| 36    | bandit, band, arm, bandit, regret |Carolan's Algorithm |
| 37    | predict, temporal, prediction, series, forecast | Multi-armed Bandits |
| 38    | schneider, max, supervision, expectations, adp... | Predictive Analytics |
| 39    | gecco, umda, lehre, binary, dang | Linear Algebra |
| 40    | safety, dynamics, drive, systems, control | Evolutionary Algorithms |

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 SeNMF 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.