Interactive Distillation of Large Single-Topic Corpora of Scientific Papers

Nicholas Solovyev
Theoretical Division, LANL
Los Alamos, USA
nks@lanl.gov

Ryan Barron
Theoretical Division, LANL
Los Alamos, USA
barron@lanl.gov

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

Maksim E. Eren
Advanced Research in Cyber Systems, LANL
Los Alamos, USA
maksim@lanl.gov

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

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

***Abstract*—Highly specific datasets of scientific literature are important for both research and education. However, it is difficult to build such datasets at scale. A common approach is to build these datasets reductively by applying topic modeling on an established corpus and selecting specific topics. A more robust but time-consuming approach is to build the dataset constructively in which a subject matter expert (SME) handpicks documents. This method does not scale and is prone to error as the dataset grows. Here we showcase a new tool, based on machine learning, for constructively generating targeted datasets of scientific literature. Given a small initial "core" corpus of papers, we build a citation network of documents. At each step of the citation network, we generate text embeddings and visualize the embeddings through dimensionality reduction. Papers are kept in the dataset if they are "similar" to the core or are otherwise pruned through human-in-the-loop selection. Additional insight into the papers is gained through sub-topic modeling using SeNMFk. We demonstrate our new tool for literature review by applying it to two different fields in machine learning.**

***Index Terms*—transformers, nlp, non-negative matrix factorization, data visualization**

I. Introduction

Literature review is an integral task of scientific research. The job often involves manually identifying papers based on keyword searches and following relevant citations. The organization of highly specific scientific literature datasets and application of data analysis techniques may provide deeper insight and discover new research directions. However, curating such highly specific datasets of scientific literature requires the time-consuming help of a subject matter expert (SME). Here, we introduce a new machine learning-based (ML) assistant tool that builds highly specific scientific literature datasets. Bibliographic Utility Network Information Expansion (BUNIE) streamlines literature review with an intuitive system while enhancing the specificity of the papers using ML techniques and human-in-the-loop procedures.

In this work, we contribute a novel approach to the scientific dataset expansion problem by integrating transformer-based document embeddings with human-in-the-loop pruning to generate targeted scientific datasets. We then use non-negative matrix factorization (NMF) with automatic model determination (NMFk) for modeling the topics in these papers to further refinement [1]. Our approach is unique in its inclusion of a human-in-the-loop for enhancing and distilling the extracted topics, such that the corpus of papers is narrowed down via an interactive process. To the best of our knowledge, this iterative method is the first to offer users the ability to analyze the topic modeling results and apply their feedback to enhance the literature review procedure by steering the ML output. The feedback loop enables the users to grow and refine the results until a targeted dataset of a specific size is reached, providing a unique and interactive solution to large-scale literature review.

The process begins with a small number of core papers selected from a topic of interest by an SME. At this initial stage, the topic may not fully align with the user's objectives and is likely incomplete. The core papers are used as a reference to obtain an additional set of relevant documents that increase the size and enhance the specificity of the existing dataset. The additional documents are selected using a citation network formed from the existing papers in the dataset. The expansion results are pruned using multiple methods, including an interactive selection by the user, document embedding similarity metrics, and topic modeling.

In contrast to the traditional static approach of computing the topics, our approach is iterative and dynamic. It allows for repetition of the refinement cycle, growing the dataset with each iteration. This enables the creation of large but specific datasets, ideal for training large language models. Through this interactive, user-driven approach, we empower users to steer the topic extraction process directly, ensuring that the results are tailored to their specific requirements. This paper demonstrates our novel tool by showcasing two different use cases on two different core datasets.

Our contributions include:

- Introducing a novel paper selection and visualization tool for scientific dataset curation and literature review.
- Utilizing text embeddings together with dimensionality reduction techniques to model the documents.

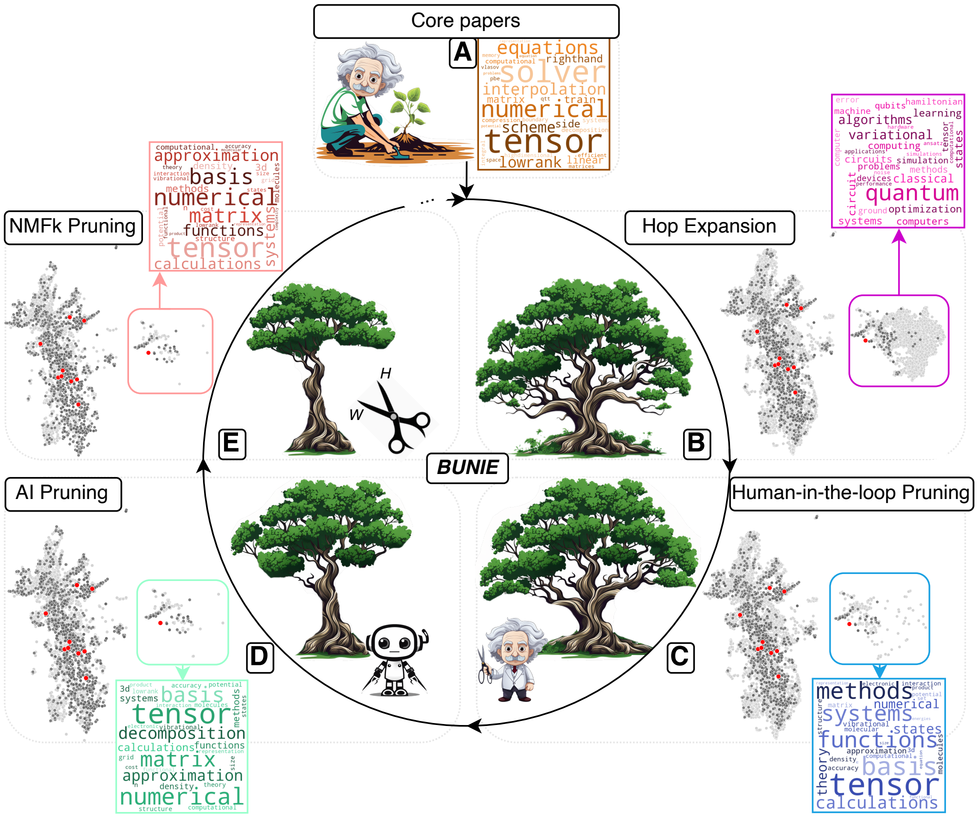

Fig. 1: BUNIE's distillation pipeline. **Panel A**: The SME selects specific papers, shown as a bag-of-words wordcloud. **Panel B**: The dataset expands through the citation network. The wordcloud of a selected subset shows the how the vocabulary differs from the core. **Panel C**: Human-in-the-loop pruning removes data. The wordcloud begins to resemble the core. **Panel D**: An embedding heuristic removes papers far from the core. The dataset becomes more compact. The wordcloud approaches parity with the original. **Panel E**: SeNMFk topic modeling produces H-clustering factorization to remove clusters lacking a core paper. A dense, yet substantial set of documents indicates a successful distillation. The wordcloud closely resembles the original. As indicated by the **Ellipse** iterations can be made for refinement or the data extracted in repose for downstream analysis.

- Integrating our machine learning approach to scientific literature with human-in-the-loop procedures for refining and guiding text modeling.
- Demonstrating the capabilities of our tool by applying it to the scientific literature in two fields of research.

II. Related Work

This section summarizes techniques and prior works applied in forging a highly-specific dataset of research papers.

1) Topic Modeling & Tensor Decomposition: A common approach to topic modeling is Non-Negative Matrix Factorization (NMF) [2], [3], and when applied to a document-word matrix identifies latent patterns of the corpus. Semantic NMF with automatic model determination (SeNMFk) is an NMF extension incorporating the text's semantic structure leveraged in [4]–[6]. Term frequency–inverse document frequency (TF–IDF) and the co-occurrence/word-context matrices are often used. Although common, more advanced methods exist.

2) Document Embeddings & Transformers: Vector representations of a text were previously used for dimensional mapping, cross-comparisons, and similarity analysis [7]. Common models for learning word embeddings have been Global Vectors for Word Representation (GloVe) [8] and Word2Vec [9]. Transformers have been used for large language models (LLM) as internal states, a popular example is the Bidirectional Encoder Representations from Transformers (BERT) [10]. Here, the SciNCL transformer generates document embeddings [11], where the document mapping includes citation data.

3) Data Visualization and Tools: Several publicly available citation network and topic modeling tools are available to explore and analyze research papers, such as Topic Modeling Tool [12], and Stanford Topic Modeling Toolbox [13]. While the tools gather topical data from inputs, they lack visualization. Alternatively, 'Connected Papers' (CP) has visuals and is a resource for discovering scientific literature [14]. While CP is useful for co-citation and bibliographic literature exploration, our tool advances bibliographic utility by creating a specialized document dataset leveraging a citation network coupled with human-in-the-loop and machine learning. Another research visualization tool, designed to handle the apex of Covid-19 research production, is explained in Ref. [15]. The tool cleaned the text (tokenized, removed stop-words, & punctuation & capitalization), constructed a TF-IDF matrix, then reduced dimensions for graphing through t-distributed stochastic neighbor embedding (t-SNE). Here, we use Uniform Manifold Approximation and Projection (UMAP) [16] to reduce 768-dimensional embeddings output by SciNCL [11] to 2D.

4) Human in the Loop: User feedback has recently been adopted into several schemes, including OpenAI's ChatGPT [17] and Google's BARD [18]. In Ref. [19], a knowledge graph is built from text-prompting a user with feedback in every response to provide an acceptable retail-item recommendation. Differences between BUNIE and Ref. [19] exist, as the study's structure provides one recommendation, whereas BUNIE offers a dataset. Interactive modes also differ–ours uses click-and-drag selection to delete rather than prompts. Furthermore, BUNIE removes papers at the HITL phase whereas [19] requests positive and negative feedback about recommendations.

A HITL work more similar to BUNIE in Ref. [20] aims to build labeled image datasets for Computer Vision (CV) applications. A user labels a few images, which are extrapolated to all in the image's cluster, then the images are model-evaluated for reassignment. Like BUNIE, the process is iterated to convergence but differs in direct user influence of datum retention.

III. Method

The utility of BUNIE comes from the combination of being able to quickly expand a dataset of publications by traversing the citation network in combination with being able to curate the dataset at scale by effectively using document text embeddings. As depicted in Figure 1, the workflow is cyclical, requiring iterative steps of document acquisition and refinement. The ultimate goal is to create an extensive collection of scientific literature centered around a specific topic, using a small, hand-picked set of relevant papers as the starting point.

*1) **Selecting the Core**:* First, the user provides BUNIE with a set of "core" papers, comprised of a unifying theme or topic — the anchor of the dataset. A subject matter expert (SME) should select and/or review this core to ensure quality and relevance. It is important to remember that BUNIE expands the dataset by traversing the citation network of known papers. A single, well-cited document may produce an extensive dataset after a few expansions following the citation network, while a collection of less frequently cited documents may yield a more limited network. The core papers are provided to BUNIE using unique paper identifiers such as DOI. Using the SemanticScholar API [21], BUNIE extracts basic information such as the title, abstract, citations, and references.

*2) **Expanding the Dataset**:* With the core established, the user can grow the dataset by making a "hop" within the citation network. This network is a directed graph of publications and their respective citations. If we denote a document, a, as belonging to a set of documents X and a document, b, belonging to the set of their citation X^c, we can say that $a \rightarrow b$ if and only if b cites a. In this context, a hop can be defined as $X := X \cup X^c$. A second hop would then also incorporate the citations from the documents in X^c, which was acquired from the first hop. The number of hops performed is left to the user's discretion, thereby controlling the scale of dataset expansion. The process can continue until the dataset reaches a desired size or until the entire citation network has been traversed. BUNIE also offers the capability to form the citation network with the edges reversed, using references as the basis instead of citations. This feature can be particularly useful when the core consists of relatively new or infrequently cited publications.

*3) **Pruning the Dataset**:* Given the interconnected nature of the citation network, not every paper found through the hop process will be relevant to the core. For example, a highly influential publication may be cited as an acknowledgment in subsequent studies focusing on entirely new issues. Thus, it is crucial to perform pruning at each hop along the citation network to prevent irrelevant topics from propagating within the growing dataset. In BUNIE, pruning is accomplished through a combination of the following three techniques.

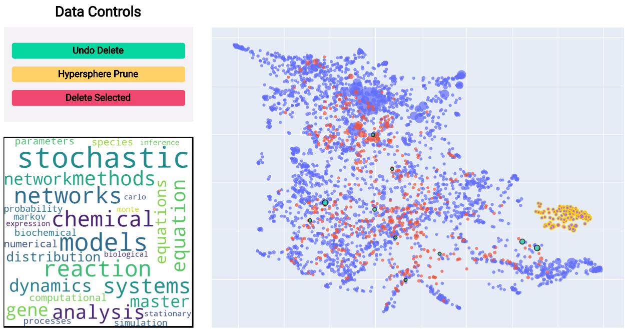

Fig. 2: BUNIE GUI screenshot during HITL pruning after two hops. Core papers are green, 1st hop red, 2nd hop blue. The SME can be seen selecting the papers in yellow and their corresponding bag-of-words wordcloud is generated to the left.

A. Human-in-the-Loop (HITL) Pruning:

Textual similarity comparison presents a substantial challenge for humans and computational algorithms. To simplify this task, we employ SciNCL [11] to transform the aggregated titles and abstracts of the dataset into 768-dimensional embeddings. These high-dimensional embeddings are then reduced to a two-dimensional projection using UMAP [16]. Semantically similar papers tend to cluster together in this two-dimensional space, providing an intuitive visual representation of the dataset's structure. This process simplifies manual content comparison.

To aid in the manual analysis and document pruning, we have designed a graphical user interface (GUI) to quickly select and examine many papers using the UMAP visualized projection of the embeddings, as shown in Figure 3. The SME can highlight papers by drawing a custom lasso or rectangle over the projected papers. The tool then generates a bag-of-words wordcloud to show the most frequent vocabulary in the chosen paper set. For a finer-grain analysis, the GUI provides a data table displaying all known data fields for the selected papers that an SME can analyze.

B. Automatic Pruning of Document Embeddings

While UMAP is useful for document visualization and enabling HITL pruning, a significant portion of the embedding structure is lost. To counteract this loss, we introduce a method for pruning the document embeddings in their original high-dimensional space. Each core paper in the dataset is considered specialized within its field. Therefore, the embeddings of the new papers added through the citation network are evaluated for their proximity to each of the core paper embeddings. The intuition is that the embeddings of relevant papers should reside ”close” to one or more of the embeddings of the core papers. In mathematical notation, given a set of core papers $P = \{p_1, .., p_n\}$ and their corresponding embeddings $E = \{e_1, .., e_n\}$, the radius ρ for each hypersphere is calculated as: *First*, compute the pairwise Euclidean distances for all embeddings in E, forming a set $D = d(e_i, e_j) : e_i, e_j \in E, i \neq j$ where $d(e_i, e_j)$ denotes the Euclidean distance between embeddings e_i and e_j. *Second*, the median Euclidean distance of all core embeddings, denoted as $\rho = \text{median}(D)$, is a threshold for new documents. Each core embedding becomes a hypersphere center with radius ρ and included papers must fall into atleast one of these spaces.

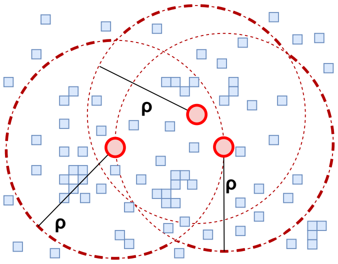

Fig. 3: 2-dimensional representation of hypersphere pruning. Core papers are red dots, citations/ references are blue squares, and hyperspheres are dotted red circles.

C. Pruning through Topic Modeling

To further ensure topic cohesion, we perform topic modeling on the pruned dataset. We utilize Semantic non-Negative Matrix Factorization with automatic model determination (SeNMFk) [6]. Given the documents, we form a term frequency-inverse document frequency (TF-IDF) matrix $\mathbf{X} \in \mathbb{R}_+^{\mathrm{m}\times\mathrm{n}}$ and an shifted positive pointwise mutual information (SPPMI) matrix $\mathbf{S} \in \mathbb{R}_+^{\mathrm{m}\times\mathrm{m}}$ which encodes the semantic structure of the data (where m is the number of tokens in the vocabulary and n is the number of documents). We then jointly factorize $\mathbf{X}$ and $\mathbf{S}$ to produce 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 $\mathbf{X}_{ij} \approx \sum_s^k \mathbf{W}_{is}\mathbf{H}_{sj}$. Here, $\mathbf{W}$ represents the distribution of words across different topics, and $\mathbf{H}$ describes how these topics are distributed across the documents. We use $\mathbf{H}$ to associate each document with the topic that it contributes to the most. The dataset can then be pruned by discarding documents from topics that do not feature any core documents.

It is important to conduct robust pre-processing of the documents to establish a meaningful vocabulary. Our pre-processing procedure removes common stop-words, symbols, newline characters, HTML tags, non-ASCII characters, e-mail addresses, and copyright statements. There are instances where specific tokens or phrases denote unique terms in the chosen domain. While these terms might appear in different forms (such as spelling, acronym, or hyphenation), all forms signify the same concept. Standard pre-processing may split a multi-token term into separate tokens, which can destroy potentially crucial meaning. However, given that an SME initially chooses the core papers, the SME can also pinpoint important terms and their assorted forms. Once these terms are identified, we consolidate all forms of each term into a singular entity. In the case of multi-token terms, we retain either the acronym or a hyphenated version to ensure that the term's meaning is preserved in the TF-IDF and SPPMI matrices. In our tensors literature example, we substitute *tensor-train* with $\{$*TT, tensor train*$\}$ and *partial-differential-equation* with *PDE* and all other various forms. Another strategy that we employ at this step involves reusing the same vocabulary for every hop. The vocabulary derived from the core papers is consistently applied at each pruning decomposition. Consequently, less relevant papers (those using a significantly different vocabulary than the core) are represented as sparse entries in the TF-IDF matrix, reducing their influence on the decomposition. This step also enhances computational efficiency as the vocabulary dimension remains constant and does not grow with the number of documents.

Through these methods, BUNIE effectively enhances the thematic coherence of the dataset while maintaining topical alignment with the original core. This results in a significantly larger, interconnected dataset that retains the integrity of the original subject matter, ready for more in-depth exploration or application. Furthermore, to quantify the efficacy of our approach, we employed a compactness score, which is a metric that evaluates how closely the documents in the dataset are related to each other in terms of the topics they cover. The compactness score of a dataset is calculated using cosine similarity between the document embeddings. In mathematical terms, given a set of document embeddings $E = \{e_1, e_2, ..., e_n\}$, the compactness score is given by:

$$C = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} \frac{e_i \cdot e_j}{|e_i|_2 |e_j|_2}, \tag{1}$$

where n is the total number of documents, e_i and e_j are the embeddings of the i^{th} and j^{th} document, $\cdot$ denotes the dot product, and $| \cdot |_2$ denotes the Euclidean norm. In measuring topic coherence using document embeddings, the cosine similarity between two embeddings, which ranges between -1 and 1, provides a measure of semantic alignment. A negative cosine similarity score, implying that the documents are semantically opposed, is an unlikely scenario within a specific topic. Therefore, we constrain the compactness score to fall between 0 and 1 to facilitate a meaningful quantification of topic coherence or alignment, accomplished by taking the absolute value of the cosine similarity. Higher values suggest a greater topic similarity between documents. The final compactness score, a value also ranging between 0 and 1, is computed as the average cosine similarity across all pairs of documents in the dataset. By this measure, a higher compactness score indicates a more coherent or well-aligned set of documents regarding their topical content.

IV. Results

This section presents two applications of BUNIE.

A. Expanding Targeted Dataset

We first applied BUNIE to 10 papers hand-picked by an SME on a specific topic. These publications were influential papers in solving integral equations using tensor-train decomposition. With the ”core” established, we sought to expand the dataset along the citation network. After the first hop, 632 papers were found. Using the visualization, papers not matching the topic were quickly pruned. While these papers tangentially addressed tensor decomposition, they failed to engage with the specific issues highlighted in the core papers. Automatic pruning was applied next through embeddings and SeNMFk. 411 papers remained after pruning in the first hop, including the original 10 core papers.

For such a minimal subset of papers, it was feasible to use the two-dimensional projection of the document embeddings in conjunction with bag-of-words word clouds to promptly identify the outlying papers. However, upon the second citation network expansion, the dataset rapidly grew to more than 8,000 papers. At this stage, the automatic pruning of the citation network became paramount. After pruning the second hops papers, a third hop was performed. After pruning, the final result came to 3,915 papers. This data flow demonstrates how BUNIE effectively combines human intuition with algorithmic utility to create a focused, relevant scientific dataset.

As demonstrated in Table I, BUNIE's iterative topic expansion and alignment increases the compactness score of the dataset. While the first expansion added many documents, many unrelated documents were included, causing the compactness score to drop from 0.894 to 0.823. The subsequent automatic pruning based on hypersphere proximity to the core document embeddings effectively increased the compactness score to 0.860, by eliminating less relevant documents, reducing the total document count to 4625. Following the hypersphere pruning, we aligned topics by applying SeNMFk and selecting relevant subtopics, further refining the dataset. This increased the compactness score and resulted in a more manageable dataset containing 3915 documents. The increase in compactness score at each stage of the BUNIE process demonstrates the method's effectiveness in maintaining topic cohesion while expanding the dataset from a small set of core papers.

TABLE I: Compactness Scores

Dataset	Compactness	Num. Documents
Tensors		
Core Papers	0.894	10
3-Hops, No Pruning	0.823	10338
3-Hops, After Hypersphere Pruning	0.860	4625
3-Hops, After SeNMFk Pruning	0.861	3915
Audio Processing		
Core Papers	0.913	68
4-Hops, No Pruning	0.798	15294
4-Hops, After Hypersphere Pruning	0.861	1987
4-Hops, After SeNMFk Pruning	0.861	1081

B. Exploratory Data Expansion

In recent years, the paper “Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context” [22] has drawn significant attention across multiple research domains. Given this impact, BUNIE allows exploration of the different domains influenced by the paper, made possible by topic modeling and visualizing text embedding projections.

We identified the associated topics of five prominent clusters: audio processing, computer vision, speech processing, natural language processing, and proteins. The audio processing cluster contained music terms from 68 papers. These papers were treated as a new core, then expanded through four hops along the citation and reference network, resulting in 15,294 papers. Following the expansion, the dataset was pruned through hyper-sphere calculation, retaining only papers within at least one of the 68 first-hop paper hyperspheres. At this point, the dataset contained 1,987 papers. SeNMFk then made topic clusters, preserving 1,081 papers from 8 core-containing clusters of 19 total. Top words from retained clusters were music, attention, generative, lyric, video, score, learn, and emotion. Table I shows dataset compactness increased with each pruning step. The core's 0.913 compactness decreased to 0.798 after four expansion hops. Hypersphere pruning increased the score to 0.861, indicating the removal of off-topic papers. Compactness remained stable after SeNMFk pruning, suggesting relevant papers were retained.

Notably, retained paper distributions per hypersphere pruned embedding mappings and SeNMFk decompositions will not always align with a human curator's intuitive UMAP-reduced selections. The discrepancy highlights the unique value of human judgment with algorithmic tools in dataset curations.

V. Conclusion

This work contributes a novel system to build scientific datasets. With minimal input, we are able to iteratively build a dataset of scientific literature anchored on the core subject provided by an SME. At each step, the dataset is enlarged through the citation network and subsequently pruned using three separate methods, including one with human-in-the-loop. The result is an expanded dataset of work relevant to the core.

Promising future work is to seed an initial topic specification. The system would then iterate autonomously, filtering out documents and recalculating topic estimates to achieve topic distillation based on reinforcement learning. Auto-distillation could dynamically adapt the topic extraction and refinement based on continuous feedback on the topic's state. The system's efficiency and accuracy could improve over time, leading to more precise and reliable topic distillation.

Additional considerations include various embedding methods and a 'synthetic' core paper to serve as a foundation for automated topic alignment. Graph neural networks to understand citation relationships can also be explored, furthering insights into the structure and interconnections of the literature. These enhancements would augment the effectiveness of BUNIE, further assisting researchers in the quest for knowledge.

VI. Acknowledgment

This manuscript has been assigned LA-UR-23-30223. This research was funded by DOE National Nuclear Security Administration (NNSA) - DNN R&D and Los Alamos National Laboratory (LANL) Laboratory Directed Research and Development (LDRD) grant 20190020DR supported by LANL's Institutional Computing Program, and by the U.S. Department of Energy National Nuclear Security Administration under Contract No. 89233218CNA000001.

References

[1] B. S. Alexandrov, V. Vesselinov, and K. Ø.. Rasmussen, "SmartTensors unsupervised AI platform for big-data analytics," Los Alamos National Lab.(LANL), Los Alamos, NM (United States), Tech. Rep., 2021, lA-UR-21-25064. [Online]. Available: https://www.lanl.gov/collaboration/smart-tensors/

[2] P. Paatero and U. Tapper, "Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values," *Environmetrics*, vol. 5, pp. 111–126, 1994.

[3] W. Xu, X. Liu, and Y. Gong, "Document clustering based on non-negative matrix factorization," 07 2003, pp. 267–273.

[4] R. Vangara, E. Skau, G. Chennupati, H. Djidjev, T. Tierney, J. P. Smith, M. Bhattarai, V. G. Stanev, and B. S. Alexandrov, "Semantic nonnegative matrix factorization with automatic model determination for topic modeling," in *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 2020, pp. 328–335.

[5] R. Vangara, M. Bhattarai, E. Skau, G. Chennupati, H. Djidjev, T. Tierney, J. P. Smith, V. G. Stanev, and B. S. Alexandrov, "Finding the number of latent topics with semantic non-negative matrix factorization," *IEEE Access*, vol. 9, pp. 117 217–117 231, 2021.

[6] M. E. Eren, N. Solovyev, M. Bhattarai, K. Ø. Rasmussen, C. Nicholas, and B. S. Alexandrov, "Senmfk-split: Large corpora topic modeling by semantic non-negative matrix factorization with automatic model selection," in *Proceedings of the 22nd ACM Symposium on Document Engineering*, ser. DocEng '22. New York, NY, USA: Association for Computing Machinery, 2022. [Online]. Available: https://doi.org/10.1145/3558100.3563844

[7] M. Kusner, Y. Sun, N. Kolkin, and K. Weinberger, "From word embeddings to document distances," in *Proceedings of the 32nd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, F. Bach and D. Blei, Eds., vol. 37. Lille, France: PMLR, 07–09 Jul 2015, pp. 957–966. [Online]. Available: https://proceedings.mlr.press/v37/kusnerb15.html

[8] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[9] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013.

[10] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018. [Online]. Available: http://arxiv.org/abs/1810.04805

[11] M. Ostendorff, N. Rethmeier, I. Augenstein, B. Gipp, and G. Rehm, "Neighborhood Contrastive Learning for Scientific Document Representations with Citation Embeddings," in *The 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP 2022)*. Abu Dhabi: Association for Computational Linguistics, December 2022, 7-11 December 2022. Accepted for publication.

[12] J. S. Enderle, "Topic modeling tool," https://senderle.github.io/topic-modeling-tool/documentation/2017/01/06/quickstart.html, 2023, accessed: June 1, 2023.

[13] E. R. Daniel Ramage, "Stanford topic modeling toolbox," https://downloads.cs.stanford.edu/nlp/software/tmt/tmt-0.4/, 2009, accessed: June 1, 2023.

[14] E. Alex, E. Smolyansky, I. Harpaz, and P. Sahar, "Connected papers," https://www.connectedpapers.com, 2023, accessed: June 1, 2023.

[15] M. E. Eren, N. Solovyev, E. Raff, C. Nicholas, and B. Johnson, "Covid-19 kaggle literature organization," in *Proceedings of the ACM Symposium on Document Engineering 2020*, ser. DocEng '20. New York, NY, USA: Association for Computing Machinery, 2020. [Online]. Available: https://doi.org/10.1145/3395027.3419591

[16] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2018.

[17] OpenAI, "GPT-3.5-based ChatGPT," 2021. [Online]. Available: https://openai.com

[18] Google, "Google Bard," 2023. [Online]. Available: https://bard.google.com/

[19] Z. Fu, Y. Xian, Y. Zhu, S. Xu, Z. Li, G. de Melo, and Y. Zhang, "Hoops: Human-in-the-loop graph reasoning for conversational recommendation," in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 2415ÔÇô2421. [Online]. Available: https://doi.org/10.1145/3404835.3463247

[20] F. Yu, A. Seff, Y. Zhang, S. Song, T. Funkhouser, and J. Xiao, "Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop," 2016.

[21] R. M. Kinney, C. Anastasiades, R. Authur, I. Beltagy, J. Bragg, A. Buraczynski, I. Cachola, S. Candra, Y. Chandrasekhar, A. Cohan, M. Crawford, D. Downey, J. Dunkelberger, O. Etzioni, R. Evans, S. Feldman, J. Gorney, D. W. Graham, F. Hu, R. Huff, D. King, S. Kohlmeier, B. Kuehl, M. Langan, D. Lin, H. Liu, K. Lo, J. Lochner, K. MacMillan, T. Murray, C. Newell, S. Rao, S. Rohatgi, P. L. Sayre, Z. Shen, A. Singh, L. Soldaini, S. Subramanian, A. Tanaka, A. D. Wade, L. M. Wagner, L. L. Wang, C. Wilhelm, C. Wu, J. Yang, A. Zamarron, M. van Zuylen, and D. S. Weld, "The semantic scholar open data platform," *ArXiv*, vol. abs/2301.10140, 2023.

[22] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. Le, and R. Salakhutdinov, "Transformer-xl: Attentive language models beyond a fixed-length context," 01 2019, pp. 2978–2988.