Proceedings of the INMM 61st Annual Meeting
July 12-16, 2020, Baltimore, MD, USA

TOPIC MODELING WITH NATURAL LANGUAGE PROCESSING FOR IDENTIFICATION OF NUCLEAR PROLIFERATION-RELEVANT SCIENTIFIC AND TECHNICAL PUBLICATIONS

Jonathan Bisila
Sandia National Laboratories

Daniel M. Dunlavy
Sandia National Laboratories

Zoe N. Gastelum
Sandia National Laboratories

Craig D. Ulmer
Sandia National Laboratories

ABSTRACT

Scientific and technical publications can provide relevant information regarding the technical capabilities of a state, the location of nuclear materials and related research activities within that state, and international partnerships and collaborations. Nuclear proliferation analysts monitor scientific and technical publications using complex word searches defined by fuel cycle experts as part of their collection and analysis of all potentially relevant information. These search strings have been refined over time by fuel cycle experts and other analysts but represent a top-down approach that is inherently defined by the requirement of term presence. In contrast, we are developing a bottom-up approach in which we develop topic models from a small number of expert refereed source documents to search similar topic space, with the hope that we can use this method to identify publications that are relevant to the proliferation detection problems space without necessarily conforming to the expert-derived rule base. We are comparing our results of various topic modeling and clustering techniques to a traditional analyst search strings to determine how well our methods work to find seed documents. We also present how our methods provide added benefit over traditional search by organizing the retrieved documents into topic-oriented clusters. Finally, we present distributions of author institutions to facilitate a broader perspective of the content of interest for analysts.

INTRODUCTION

Nuclear proliferation analysts monitor scientific and technical publications using complex queries defined by fuel cycle experts as part of their collection and analysis of all potentially relevant information. These queries are composed of search strings that have been refined over time by these experts and other analysts, but represent a top-down approach that is inherently limited by the requirement of specific terms being present in the publication data. This current approach requires constant manual adaptation of query terms to reflect changes in terminology in evolving scientific advances. Furthermore, using query-based information retrieval alone to identify relevant publications still requires a tremendous amount of time as well as much manual reasoning and triage of the information on the part of the analysts to identify emerging research activities of interest.

In contrast, we have developed a bottom-up, data-driven approach in which we use topic modeling and document clustering, with the hope that we can utilize this method to identify publications that are relevant to the proliferation detection problem space without necessarily conforming to the expert-derived rule base. With query-based retrieval, publications relevant to a query are identified, but further analysis of how those documents relate to one another is left to the analysts. With our approach, the retrieved documents are then organized by topic, such that analysts can quickly identify subsets of documents relevant to answering specific proliferation research activity questions. Furthermore, to facilitate identification of organizations and potential research partnerships focusing on particular proliferation topics, we present the distribution of author institutions across each of these topics. This approach provides analysts with a more organized, summarized view of publication data that can reduce the amount of time typically associated with scientific and technical research activity analyses.

The remainder of this paper is organized as follows. In the next section, we provide an overview of our approach to publication analysis along with a detailed description of our analysis framework. In the following section, we provide a case study illustrating how our approach can identify laboratories publishing in the area of nuclear fuel reprocessing, using publicly-available publication data from the U.S. Department of Energy. Finally, we provide concluding remarks and open questions related to our work and the potential role of our framework in nuclear proliferation analysis.

METHODS

In this section, we provide an overview of our publication analysis approach and detailed descriptions of the document analysis components we use in the framework implementing this approach. Figure 1 presents an illustration of the major components in our analysis framework. Starting with a collection of publication data and a user-defined query, we first apply several natural language processing (NLP) techniques for extracting and abstracting relevant information from individual documents. This information is transformed into a vector space model, a standard data abstraction that facilitates efficient and large-scale processing and analysis. With the data in this form, we perform query-based retrieval to identify a collection of documents of interest to the analyst. This is the typical stopping point for analysts, where manual inspection of individual documents would follow. The relationships identified by query-based retrieval leverage only simple term or phrase co-occurrence information. In our approach, we use topic modeling to create a representation of this post-query collection of documents that relates the documents to topics and the topics to the individual terms or phrases in the documents. By modeling documents in this way, we can then cluster them by topic. For each of the document clusters, we extract the set of authors' institutions and present the distribution of the count of documents per institution alongside some of the terms and phrases extracted from the topic model to provide a summary of the cluster. This summarized view of each cluster provides the means for analysts to quickly ascertain if the documents in each cluster can be useful in answering specific questions they have related to their query.

Note that the framework we present here is modularized to support the use of alternative methods and algorithms at the various steps in the pipeline illustrated in Figure 1.

In the following sections, we provide detailed descriptions of each of the steps in this framework. Our framework is implemented in the Python programming language, and references to specific packages used are provided in the context in which the steps are defined below.

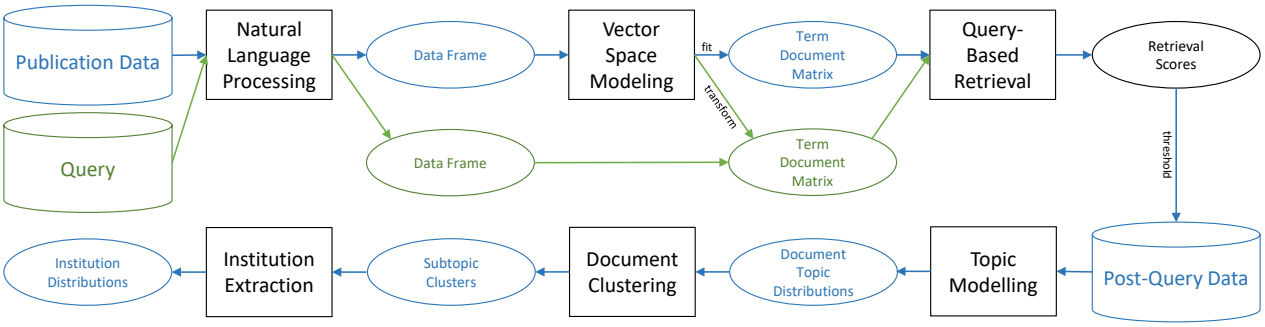


Figure 1: Analysis framework for the proposed method of identifying institutions publishing on proliferation-relevant topics. Cylinders, boxes, and ovals represent document data, data analysis steps, and analysis result data, respectively. Blue and green shapes and arrows represent publication and query data, respectively.

Natural Language Processing

We use standard NLP methods from the `nltk` Python package [3] to represent the publication data. Specifically, for each document we employ *lemmatization* [7] followed by *stemming* [14] of terms to reduce the impact of word inflection. Lemmatization maps words to roots that are part of the language, whereas stemming maps words to roots that may or may not belong to the language. These modifications of words allows us to capture and align variants of tense, number, gender, etc. By using both methods our analysis is flexible enough to map terms to both known roots and those that may be part of a change in language to describe emerging scientific advancement. For example, the words *proliferate*, *proliferation*, *proliferates*, etc., all share a common root (*proliferat*), and lemmatization and stemming are two methods that are used to map all these word variants to this common root.

We also remove common words—also known as *stopwords*—from the document representations, as such words often do not aid in distinguishing topical relationships between documents. Specifically, we use the list of words provided by Stone *et al.* [19], who demonstrated improved performance on several NLP tasks by removing the specific set of stopwords they recommend.

Vector Space Modeling

Vector space models (VSMs) are a common representation for text documents in many NLP methods, initially developed to support efficient, query-based information retrieval [17]. A collection of documents is represented by a VSM using a *term-document matrix*, where columns represent terms and rows represent documents, and the entry at row i and column j represents the "importance" of term j in document i. The measure of importance can be defined in many ways, and has been the subject of decades of research in the areas of information retrieval and, more generally, automated text analysis. In our framework, we currently set the entries of the term-document matrix to either the counts of terms or a weighted version of those counts based on the term frequency-inverse document frequency (TF-IDF) weighting scheme [18]. We use the VSM generators in the scikit-learn Python package [13] in our framework.

Query-Based Information Retrieval

Using a VSM to represent documents, query-based information retrieval can be performed efficiently by computing a single matrix-vector product [17]. The query is first represented as a single vector in the same vector space as that used to define the term-document matrix described in the previous section. The result of using this matrix-vector product approach is a score for each document that measures the relationship between the query and that document. When using counts in the term-document matrix, the score is the number of query terms that appear in the document; when using TF-IDF weighted counts, the score can be interpreted as the probability that the document is related to the query based on shared term usage [16]. A threshold on the score or a restriction on the maximum number of documents retrieved can be utilized to identify a selection of documents that will used for further analysis.

Note that query-based information retrieval using a VSM treats all terms in a query as being joined together with an OR conjunction. For example, the query, "nuclear proliferation research" would be equivalent to performing the query, "nuclear OR proliferation OR research" in an information retrieval system that supports Boolean operators [11].

Topic Modeling

Once we have identified the collection of documents relevant to the query as described above, we build a topic model from this collection using a method called Latent Dirichlet Allocation (LDA) [4]. LDA is a Bayesian model that relates documents to terms using latent variables called *topics*. In an LDA model, documents are represented as distributions across topics and topics are represented as distributions across terms. LDA provides document representation based on topics (i.e., mixtures of terms) rather than individual terms, supporting document similarity identification based on complex term usage. The number of topics is provided as input to LDA, thus there is a free parameter that must be determined prior to creating a topic model. Although many published results suggest choosing the number of topics to be in the range of $50-150$, some experimentation may be required in determining this number in practice for particular document collections. LDA has been shown to outperform other topic modeling approaches in document clustering tasks [24] and in detecting emerging scientific fields with high-accuracy [20]. Thus, we use LDA as the basis for modeling the topics in the post-query collection of documents.

In order to create topic models for large document collections, we use a computationally efficient, parallel implementation of LDA [9] from the `gensim` Python package [15] in our framework.

Document Clustering

Recently, the Louvain community detection algorithm for partitioning graphs [5] has been proposed as a method for clustering text documents [2]. We apply this approach in our framework as well, starting with the LDA topic model representations of documents as described above. We create an undirected document similarity graph whose vertices represent documents and weighted edges represent similarities between each pair of documents. The weight on the edge between vertices i and j is computed as $1-JSD(i,j)$, where $JSD(i,j)$ is the Jensen-Shannon divergence [12] between the document-topic probability distributions for documents i and j as computed from the LDA model. The Louvain algorithm is then applied to this graph to identify clusters of documents that are related

as follows: 1) documents within each cluster are strongly related by the LDA topics modeling those documents, and 2) documents in different clusters either do not share LDA topics in common or are only weakly related by the LDA topics modeling those documents.

Summaries of the clusters are provided in our framework utilizing the document-topic and topic-term probability distributions given by the LDA topic model. By using the document-topic distributions and the Jensen-Shannon divergence between all pairs of documents within the cluster, we provide a rank-ordered list of the documents that best summarize that cluster, as those documents are the ones that are most similar to all others. Next, using the topic-term distributions, we provide a rank-ordered list of the topics and terms that best summarize that cluster across all documents in that cluster.

Note that the use of LDA and Jensen-Shannon divergence in defining the underlying graph for clustering documents provides a probabilistic interpretation of the relationships between pairs of documents and thus can facilitate rigorous uncertainty quantification analysis. Such analysis is beyond the scope of this paper, but our use of this approach for document clustering is specifically motivated by the promise of quantifying and potentially reducing sources of variability in the topic clustering results, consequently leading to the most stable set of results for analysts.

We use the `community` Python package [1] for the Louvain implementation, which leverages the `networkx` Python package [8] in our framework.

Institution Extraction

Provided that the publication data contains the author institution data, we extract the institutions associated with the documents in each cluster and plot a histogram of the distribution of document counts per institution. The aim in doing so is to provide the analysts with a quick summary of institutions associated with each cluster. Along with the topic and term summaries described in the previous section, these histograms provide summaries of clusters based on both publication content and metadata (e.g., author institution information), aiding analysts in quickly finding the most relevant information related to their queries and the underlying questions they are trying to answer using this publication data.

Unfortunately, not all publication databases contain rich metadata, such as the mappings of authors to institutions, and thus this step may not always be performed. Furthermore, even when such information is available, specific metadata values may not be presented consistently across all documents. Although current research addressing these problems associated with entity resolution can help to address such inconsistencies [6], discussion of those methods and application in our framework is beyond the scope of this paper.

CASE STUDY: U.S. DEPARTMENT OF ENERGY PUBLICATIONS

To illustrate the utility of our framework, we present the following case study identifying institutional associations to the topic of *nuclear fuel reprocessing* using publication data from the U.S. Department of Energy.

Publication Data

We use a small subset of publication data from the U.S. Department of Energy Office of Science and Technical Information (OSTI) [21]. The OSTI database consists of scientific and technical reports, journal articles, patents, and presentations regarding research performed at the U.S. Department of Energy National Laboratories and associated institutions. We chose a random subset of 2,000 publications from the OSTI database published in 2017. Furthermore, we manually identified seven additional documents in the OSTI database published in 2017 that subject matter experts have confirmed are highly relevant to the topic of nuclear fuel reprocessing. These latter documents will be used to demonstrate how documents relevant to a query associated with nuclear fuel reprocessing are clustered together using our analysis framework.

We extracted the author, author institution, title, and abstract information from the OSTI publication data of use in this case study. The title and abstract for each document were combined and used as input to our analysis framework.

Although there are many documents in the OSTI database that do not have complete metadata information, in the case study presented here, we only consider documents with full metadata available. The average number of terms per document (title and abstract combined) over these 2,007 publications is 1,117.

Query Data

To create a query, we leveraged several publicly-available sources of information regarding nuclear fuel reprocessing. Specifically, we created a query using terms extracted from documents associated with this topic produced by the International Atomic Energy Agency [10], by the World Nuclear Association [23], and available on Wikipedia [22]. These documents represent a mix of subject matter expertise in this area. The resulting query terms are as follows:

```
alloy, anion, arms, boxes, bromine, cadmium, calcination, cell, ceric, chloride,
chopping, concrete, cuts, decladding, density, electrolysis, exchange, extraction,
fluoride, fluoridized, fuel, fuels, glass, glove, handling, hardened, hexone,
high, inconel, irradiated, isobutyl, ketone, lead, lithium, manipulator, master,
metal, methyl, MIBK, molten, nitrate, nuclear, oxidation, plutonium, precipitation,
purification, radiation, redox, reduction, remote, reprocess, rods, salts,
separation, shield, slave, solvent, spent, trifluoride, uranium, volatility, window
```

Results

We applied our analysis framework described in the previous section to the publication and query data identified above. For the vector space model, we used term count to populate the term-document matrix, and for the LDA topic model, we created a model with 50 topics. Of the 2,007 documents in the publication data, 1,281 documents matched the query to some extent (i.e., contained terms in the query data). Using query-based information retrieval alone, nuclear proliferation analysts at this point would need to read through these 1,281 documents (or at least the titles) to determine whether each document was relevant to the specific information related to nuclear fuel reprocessing

Table 1: Number of documents per cluster in the case study.

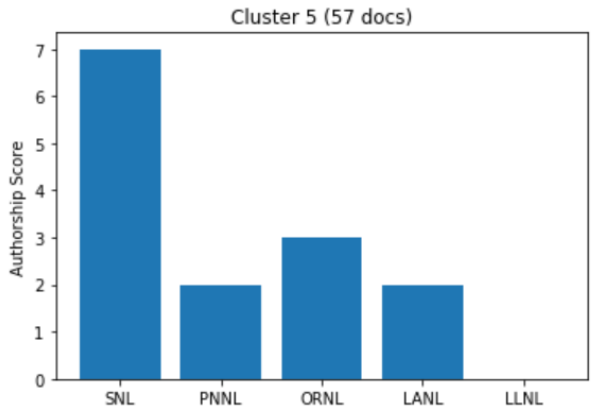
Cluster Number	Number of Documents
1	105
2	777
3	68
4	22
5	57
6	37
7	104
8	111

for their analysis. Using our framework, these 1,281 documents were clustered into eight clusters, distinguished by subtopics related to the main query data. Note that analysts currently use document retrieval systems that support more sophisticated queries than the Boolean OR representations we present here. However, we point out that manual inspection of the retrieved documents is still required to determine the relationships between documents and the topics of those documents across the entire retrieved set. The goal in using our framework is to automate that process for the analysts.

Table 1 presents the number of documents in each of the clusters. By clustering documents and presenting summaries of each cluster, we have reduced the number of documents that analysts may need to read to determine the publications they deem relevant to their analysis. One of the characteristics of the Louvain clustering algorithm is that it tends to identify a single large cluster that is a catch-all for documents that do not cluster well (i.e., that are related to many documents weakly in terms of Jensen-Shannon divergence and are not strongly related to any other particular subset of the data) and several smaller clusters whose documents are all strongly related. These latter clusters turn out to reflect clear subtopics represented in the post-query publication data. In Table 1, we see that Cluster 2 is this catch-all cluster, containing the majority of documents. Note that the smaller, subtopic clusters contain around 100 or fewer documents. By organizing the documents into clusters by topic, we have potentially reduced the number of documents to be further analyzed by an order of magnitude (assuming that there is a single subtopic that is of most interest to the analysts).

Figure 2 presents the summary for Cluster 5. We see the distribution of document count by author institution (i.e., a subset of U.S. Department of Energy National Laboratories presented here as examples), examples of the most relevant topic-term distributions associated with the documents in the cluster, and a sample of document titles from the cluster. The Authorship Score in the institutional distribution plot is the count of documents in the clusters in which at least one author was from that institution. The topic-term distributions in this figure present the top three topics with the top four terms for that respective topic. Both the number of topics and the number of terms shown can be adjusted by the user of our framework. As LDA topics are mixtures of terms (shown lemmatized and stemmed here), we present the probabilities associated with the terms in each topic to help analysts compare the relative importance of each individual term across the topics.

A summary such as the one presented in Figure 2 is provided for each cluster of documents identified using our analysis framework. One of the benefits of using this approach is that analysts need to read and analyze much less data (i.e. short summaries of topics, terms and titles) for a small number of clusters compared to the full set of documents returned using query-based information retrieval (e.g.,




```
Cluster 5 topics and terms

0.058*"compon" + 0.047*"contain" + 0.041*"polym" + 0.035*"solvent"
0.042*"layer" + 0.039*"shell" + 0.031*"flow" + 0.028*"electrod"
0.026*"modul" + 0.023*"optic" + 0.022*"system" + 0.018*"devic"

Cluster 5 titles

Production of alpha-hydroxy carboxylic acids and esters from higher sugars using tandem catalyst systems
Electrochemical Nucleation and Growth of Uranium and Plutonium from Molten Salts
Variants of polypeptides having cellulolytic enhancing activity and polynucleotides encoding same
```

Figure 2: Summary of documents in Cluster 5, which contains five of the seven documents most relevant to the query.

the full set of 1,281 documents in the case study presented here). As an example of this, we have determined that our seven expert-reviewed documents which serve as our experimental "tracers" appear in just two clusters: 5 of the tracer documents in Cluster 5 (57 documents), and 2 in Cluster 1 (105 documents). Together, these two clusters represent approximately 12% of the total 1,281 documents that were clustered into topically-relevant groups. While this is just one case study, it is possible that similar tracers could be scaled to larger data to help analysts prioritize clusters for review.

CONCLUSIONS

In this paper we proposed an analysis framework to aid analysts in finding associations between scientific topics and the institutions engaged in research in those areas. Through a single case study, we applied this framework to publication data from the U.S. Department of Energy to demonstrate how it can help analysts find relevant publications on topics related to nuclear proliferation research. Through our modular design of the framework, we have facilitated ease of much further analysis around uncertainty quantification, comparison of different clustering methods, and analysis of different vector space models. Finally, we believe that this framework can reduce the amount of time it takes analysts to triage and find relevant documents by reducing the number of documents through which they have to manually sort.

ACKNOWLEDGMENTS

This work was funded by the U.S. Department of Energy National Nuclear Security Administration's Office of Defense Nuclear Nonproliferation Research & Development (NA-22). Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

REFERENCES

[1] T. Aynaud. python-louvain 0.14: Louvain algorithm for community detection. https://github.com/taynaud/python-louvain, 2020 (accessed May 28, 2020).

[2] A. Beniwal, G. Roy, and S. Durga Bhavani. Text document clustering using community discovery approach. In D. V. Hung and M. D´Souza, editors, *Distributed Computing and Internet Technology*, pages 336–346. Springer International Publishing, 2020.

[3] S. Bird, E. Loper, and E. Klein. *Natural Language Processing with Python*. O'Reilly Media Inc., 2009.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, pages 993–1022, 2003.

[5] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.

[6] V. Christophides, V. Efthymiou, T. Palpanas, G. Papadakis, and K. Stefanidis. End-to-end entity resolution for big data: A survey. arXiv:1905.06397, 2019.

[7] C. Fellbaum, editor. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA, 1998.

[8] A. A. Hagberg, D. A. Schult, and P. J. Swart. Exploring network structure, dynamics, and function using networkx. In *Proceedings of the 7th Python in Science Conference*, pages 11–15, Pasadena, CA USA, 2008.

[9] M. Hoffman, F. R. Bach, and D. M. Blei. Online learning for latent dirichlet allocation. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 856–864. 2010.

[10] International Atomic Energy Agency. Spent fuel reprocessing options. Technical report, IAEA-TECDOC-1587, 2008.

[11] A. H. Lashkari, F. Mahdavi, and V. Ghomi. A boolean model in information retrieval for search engines. In *Proceedings of the 2009 International Conference on Information Management and Engineering*, pages 385–389, 2009.

[12] C. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, 1999.

[13] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[14] M. F. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

[15] R. Řehůřek and P. Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, May 2010.

[16] S. E. Robertson. Understanding inverse document frequency: on theoretical arguments for idf. *Journal of Documentation*, 60:503–520, 2004.

[17] G. Salton. *Automatic text processing: The transformation, analysis and retrieval of information by computer.* Addison-Wesley, Reading, MA, 1989.

[18] G. Salton and M. J. McGill. *Introduction to modern information retrieval*. McGraw-Hill, Inc., 1986.

[19] B. Stone, S. Dennis, and P. J. Kwantes. Comparing methods for single paragraph similarity analysis. *Topics in Cognitive Science*, 3(1):92–122, 2011.

[20] A. Suominen and H. Toivanen. Map of science with topic modeling: Comparison of unsupervised learning and human-assigned subject classification. *Journal of the Association for Information Science and Technology*, 67(10):2464–2476, 2016.

[21] U.S. Department of Energy. Office of Scientific and Technical Information (OSTI). http://osti.gov, 2020 (accessed May 28, 2020).

[22] Wikipedia. Nuclear reprocessing. http://en.wikipedia.org/wiki/Nuclear_reprocessing, 2020 (accessed May 28, 2020).

[23] World Nuclear Association. Processing of used nuclear fuel. http://world-nuclear.org, June 2018 (accessed May 28, 2020).

[24] C.-K. Yau, A. Porter, N. Newman, and A. Suominen. Clustering scientific documents with topic modeling. *Scientometrics*, 100:767–786, 2014.