Editor Matching for Academic Journals Through Rich Semantic Network Development

Jack Prominski, Pragati Shah, and Rafael Alvarado
University of Virginia, jmp3mk, pvs3vf, rca2t@virginia.edu

Abstract - **The Public Library of Science (PLOS) is a major publisher of academic journals with a collection of over 200,000 scientific journal articles. As the library grows, a key operational challenge for PLOS is the matching of a submitted manuscript with one of the nearly 6,000 academic editors with whom they work. We frame this problem as one of reducing distance between editors and documents in a network, where editors and documents are modeled as nodes in a graph connected by shared features. We apply Natural Language Processing (NLP) and text mining methods, such as topic modeling and word embedding, to create new nodes and edges in the network to reduce its scale, thus reducing the distance between documents and editors. These new nodes and edges form a rich semantic network that may also be used to represent the thematic content of the corpus as a whole. Our solution iterates on PLOS's current internal tool for editor matching, PLOS Match, by recommending potential editors who have edited semantically similar manuscripts in the past. Improved quality of editor matches will increase the rate at which editors accept an invitation to participate in manuscript evaluation. This will streamline PLOS's internal review process and hence speed knowledge delivery.**

Index Terms - editor matching, natural language processing, PLOS, text mining, topic modeling, word embedding

INTRODUCTION

The Public Library of Science (PLOS) is a major publisher of academic journals and advocate of open access science. Its largest journal, PLOS ONE, publishes 20,000+ articles each year and receives many more submissions. A key operational challenge for PLOS is the matching of a submitted manuscript with one of the nearly 6,000 academic editors with whom they work. Academic editors must be subject matter experts on the topic of the submitted manuscript. Deep knowledge and experience in the article's subject leads to more productive feedback and more accurate evaluation, so this step is crucial in maintaining the efficiency and effectiveness of the manuscript review process. This matching is a bigger task than can be handled by a team of people and, with advances in Natural Language Processing (NLP), this task is particularly well suited for automation.

In 2014, PLOS launched a tool called PLOS Match that was developed in-house to address this issue. PLOS Match is fully integrated with Editorial Manager, the manuscript submission tracking system that PLOS uses for internal process management. PLOS Match uses a simple bag-of-words and tf-idf approach to determine similarity between a submitted manuscript and a corpus of articles associated with their editor partners. The tool then outputs a list of suggestions of academic editors best-suited to handle the manuscript.

PLOS leadership, with feedback from its editorial staff and academic editors, believe that the quality of these matches could be improved, and have been actively exploring solutions to this problem. The key performance indicator that PLOS has associated with this operational process and would like to improve upon is the academic editor acceptance rate. The acceptance rate is defined as the percentage of academic editors who accept the invitation when approached with a request to evaluate a manuscript.

We frame this problem as one of reducing distance between editors and documents in a network, where editors and documents are modeled as nodes in a graph connected by shared features. In this project we explore a survey of NLP and text mining methods to generate these features and develop a rich semantic network. We then apply this tool to the task of editor matching.

To begin, we developed an NLP pipeline for feature engineering. Using implementations from the Python module Gensim [1], we then explored topic modeling techniques such as Latent Dirichlet Allocation (LDA) and Latent Semantic Analysis (LSA), and a neural network approach through word embedding with doc2vec. Through these techniques we sought to develop a rich semantic network made up of quantitative measures of the relationships between documents in the corpus.

We then applied this knowledge base to the task of suggesting appropriate academic editors based on semantic similarity of the manuscript submission and the texts that editors have previously contributed to the PLOS corpus. At this stage we also conducted citation network analysis to aid in the task of final editor ranking.

After model building, we developed a front end web application that we delivered to PLOS allowing their team to interact with our models for further evaluation and eventual incorporation into PLOS Match.

RELATED WORK

Substantial prior research has been conducted in the area of content-based recommender systems for academic journal articles [2]. Many employ a tf-idf weighting scheme, and recommendations based on LDA topics [3] and vector space

similarity with LSA and doc2vec [4] have all been explored. Most approaches in this area, if evaluated at all, were evaluated offline or were not evaluated against a baseline, making them very difficult to compare [2]. Ground truth does not exist in such a subjective domain, and therefore model performance is very difficult to determine.

We were unable to find any substantial research on manuscript to editor matching, the primary application of this study. Some publishing houses have developed proprietary, in-house matching algorithms, though this research is outside of the public domain. Personal communication with editorial teams indicate that some journals, especially smaller ones, rely on manual matching by a staff member or repurposing tools designed for authors to find appropriate journals or collaborators, such as the Journal/Author Name Estimator (JANE) [5].

DATA

PLOS, an Open Access publisher, has made their corpus easily accessible and encourages researchers to perform analyses upon it [6]. For this project, we have accessed the PLOS corpus through the PubMed Central (PMC) Open Access (OA) subset [7]. This corpus, accessed in September 2017, supplied over 200,000 articles across nine PLOS journals in the Journal Article Tag Suite (JATS)-compliant XML document format. JATS defines a standard set of XML elements and attributes for describing the textual and graphical content of journal articles [8]. A standard XML format is crucial to allow our analysis to scale to the levels necessary to include PLOS's entire corpus. These documents contain full article text as well as article metadata, including contributor and publication information.

APPROACH

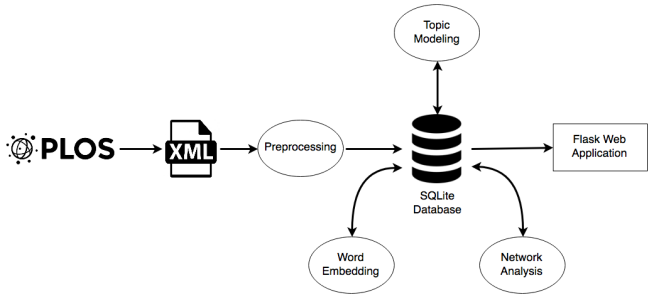

FIGURE I
PROJECT DATA PIPELINE

I. Preprocessing

We developed a data pipeline, shown in Figure I, to parse and extract relevant data from the raw XML documents and convert them into a tabular structure suitable for input into our models. We used the ElementTree XML API [9], which makes use of flexible container objects designed to store hierarchical data structures in memory. ElementTree represents the inherent tree structure of XML as a Python iterable object that can be traversed to extract element-level data We extracted the following contents from the PLOS articles and stored them in an SQLite database:

- Digital object identifier (DOI)
- PubMed Identifier
- Journal Name
- Authors
- Editors
- Subjects
- Title
- Date of Publishing
- Abstract
- Body

The abstract and body texts are then further processed for normalization. We used a standard NLP process, which included stripping punctuation, removing stopwords, tokenizing, and stemming the documents using the package NLTK [10]. From these features, we constructed term frequency and tf-idf matrices which served as inputs to our models.

II. Topic Modeling

Topic modeling is an unsupervised clustering technique for the discovery of latent semantic structures in a corpus. Each document is described in terms of one or more "topics," which are made up of the most common words in that topic. These document-topic assignments can then be used to measure document similarity. We trained topic models using varying input features and parameters. We varied the number of topics from 50 to 200 and varied our training features to include tf-idf matrices built from both article abstracts and full article bodies.

Latent Dirichlet Allocation (LDA) [11] is a probabilistic technique for topic modeling that uses a generative process to define topics as a group of words and documents as a mixture of topics.

These topics should be thematically coherent, and we can infer the subjects of these topics by examining the terms that make them up. As an example, the top five topics assigned to a target article "Measurement of Phospholipids May Improve Diagnostic Accuracy in Ovarian Cancer" along with the associated topic weights and terms are displayed in Table I. Note that the "terms" here are actually features that were generated with our preprocessing pipeline and include words that have been stemmed. These terms are groups of words that are likely to appear in documents that are assigned a high weight to that topic.

The first two topics indicate that the article likely contains a study about using statistical models for diagnosis, and the third topic indicates it is about a type of cancer. As each document is described in terms of these topic weights, similar articles in an LDA model are those which share similar topic weights.

TABLE I

LDA: TOP TOPIC ASSIGNMENTS FOR SAMPLE DOCUMENT: "MEASUREMENT OF PHOSPHOLIPIDS MAY IMPROVE DIAGNOSTIC ACCURACY IN OVARIAN CANCER"

Topic	Weight	Top 10 Terms
148	.183	measur, valid, correl, use, assess, discrimin, reliabl, biomark, test, specif
43	.140	model, predict, data, use, method, estim, base, approach, set, statist
97	.137	cancer, lung, prostat, donovani. kaposi colorect. gastric, pancreat. daughter, associ
161	.077	hospit, complic, surgery, center, case, surgic, congenit, oper, procedur, seek
25	.065	detect, use, test, method, sampl, sensit, assay, diagnost, result, can

TABLE II

LSA: MOST SIMILAR TO TARGET WITH COSINE SIMILARITY

cos(θ)	Article Title
Target	Measurement of Phospholipids May Improve Diagnostic Accuracy in Ovarian Cancer
.703	Development and Preliminary Evaluation of a Multivariate Index Assay for Ovarian Cancer
.671	Systematic Evaluation of Candidate Blood Markers for Detecting Ovarian Cancer
.601	Effects of Blood Collection Conditions on Ovarian Cancer Serum Markers
.600	Knowledge about Cervical Cancer and Associated Factors among 15-49 Year Old Women in Dessie Town, Northeast Ethiopia
.598	Comprehensive Serum Profiling for the Discovery of Epithelial Ovarian Cancer Biomarkers

Latent Semantic Analysis (LSA) [12], also known as latent semantic indexing, is another topic modeling technique that uses singular value decomposition (1) to reduce a sparse document-term co-occurance matrix X into a lower dimensional approximation that retains the relationships between documents. A diagonal matrix of singular values Σ reduces the dimensionality of the U (words) and V (documents) vectors and retains only those features that capture the most variation in the corpus.

$$X_k = U_k \Sigma_k V_k^T \tag{1}$$

Two document vectors (A and B) can then be compared using cosine similarity (2), with values ranging from 0 to 1 where higher values indicating greater similarity. In this domain, the cosine similarity between two documents will never be below zero, as the values of the term frequency matrix cannot be negative.

$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}} \tag{2}$$

To determine the most similar documents to a given target document, the cosine similarity between the target and every document in the corpus is calculated. Cosine similarity is preferred to Euclidean distance in this context as it is independent of document size. Table II shows an example of this similarity measure applied to the PLOS corpus. The top five most similar documents to the article about ovarian cancer diagnosis are displayed.

Using principal component analysis to reduce dimensionality of these vectors, we are able to examine these article vectors and their cosine similarity visually. Figure II shows that the vectors representing two articles discussing ovarian cancer are very similar. However, the vector for an article discussing the perceived attractiveness of facial features is almost orthogonal to the two other vectors. This makes intuitive sense, as the semantic similarity of the two ovarian cancer articles is likely very high, and the subject matter of the other article is not related.

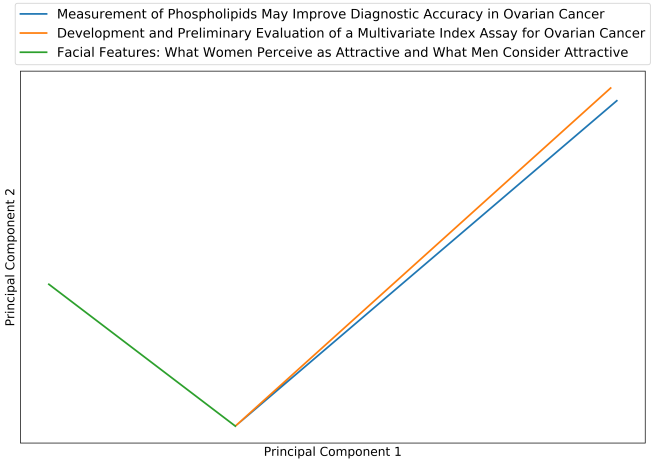

FIGURE II

VISUALIZATION OF LSA VECTORS FOR THREE SAMPLE DOCUMENTS

III. Word Embedding

We also represented documents using word embedding, another vector-space model. Rather than the singular value decomposition approach that LSA uses, word embedding employs a shallow neural network to represent words and documents as dense numerical vectors. The doc2vec model we used [13] is an extension of the popular word2vec word embedding model. Figure III illustrates how the model works. D represents the document context, and W represents the word context surrounding a target word. Word vectors are trained in the same method as in word2vec, and document vectors are then inferred with a Distributed

Memory Model of Paragraph Vectors (PV-DM) model, using stochastic gradient descent.

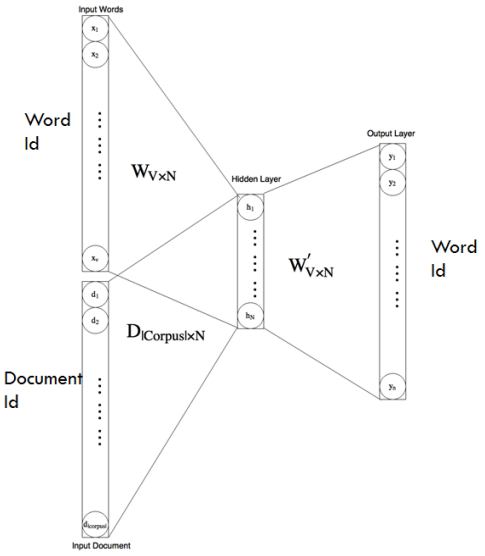

FIGURE III
WORD EMBEDDING DOC2VEC

After training this neural network, we can both compare document vectors to each other and infer the vector of an unseen document (in our case, manuscript submissions). As a point of comparison to the LSA model, Table III shows the most similar articles for the same ovarian cancer diagnosis article, as determined by cosine similarity between the document vectors generated with doc2vec.

TABLE III
DOC2VEC: MOST SIMILAR TO TARGET WITH COSINE SIMILARITY

cos(θ)	Article Title
Target	Measurement of Phospholipids May Improve Diagnostic Accuracy in Ovarian Cancer
.471	REG4 Is Highly Expressed in Mucinous Ovarian Cancer: A Potential Novel Serum Biomarker
.468	Prostate Cancer Associated Lipid Signatures in Serum Studied by ESI-Tandem Mass Spectrometryas Potential New Biomarkers
.464	Improving Clinical Risk Stratification at Diagnosis in Primary Prostate Cancer: A Prognostic Modelling Study
.435	Validation of LRG1 as a Potential Biomarker for Detection of Epithelial Ovarian Cancer by a Blinded Study

IV. Citation Analysis

Citations of research papers are a rich source of information as they embody the existing connections in academia. In our approach, we connected the citations of the PLOS corpus using a network graph. The key idea was to use the inherent characteristics of this citation network as auxiliary predictors to the semantic features. In past approaches, recommendation engines based on citation networks have used co-citation and co-reference factors to improve their recommendation [14]. However, our approach is slightly different as our ultimate intent is not to build a recommender system. The nodes of our citation network are all a combination of all the PLOS research papers and their respective citations. The edges are directed, and represent a citation from a PLOS research paper the referenced paper. We make use of features like the number of other PLOS research papers referencing a PLOS paper, and betweenness centrality of the PLOS research paper. These features can be used to order the output of the editor recommendation model by giving a greater preference to editors who have edited papers with a higher betweenness centrality. Using these features is beneficial as it leverages the connections which are characteristic of an academic corpus and incorporates the popularity of a research paper implicitly, without using the number of citations. As our data was limited to the PLOS corpus, a richer network would incorporate all of a research paper's citations, as well as the references of these referenced papers.

IV. Editor Recommendations

Building upon the modeling work to generate the rich semantic network of articles and their semantic relationships, we generated editor suggestions using two general approaches. First was to simply find the most semantically similar articles in PLOS's corpus to a submission manuscript using the topic modeling and word embedding approaches discussed previously, and outputting the editors who edited those articles. This is the approach that PLOS Match currently uses to derive editor suggestions. However, it is possible that the individual who edited the single most semantically similar article is not the best candidate to edit the submitted manuscript. An editor with a larger body of work in the subject may be more appropriate.

To address this shortcoming, we also trained a model where the input matrices did not represent individual documents, but rather represented an editor's entire body of work. We concatenated the abstracts of all the articles that an editor had previously edited, and used the resulting matrices as the model inputs.

RESULTS

We developed a front-end web application using the Flask web framework to allow our partners at PLOS to test our approaches and evaluate the results. Users are able to input a submission abstract and obtain recommendations for individuals whom our model determines as most appropriate to serve as editors for the manuscript. They can also examine similar articles given an article DOI or string query. Figure IV shows screenshots of the application's user interface, including sample results.

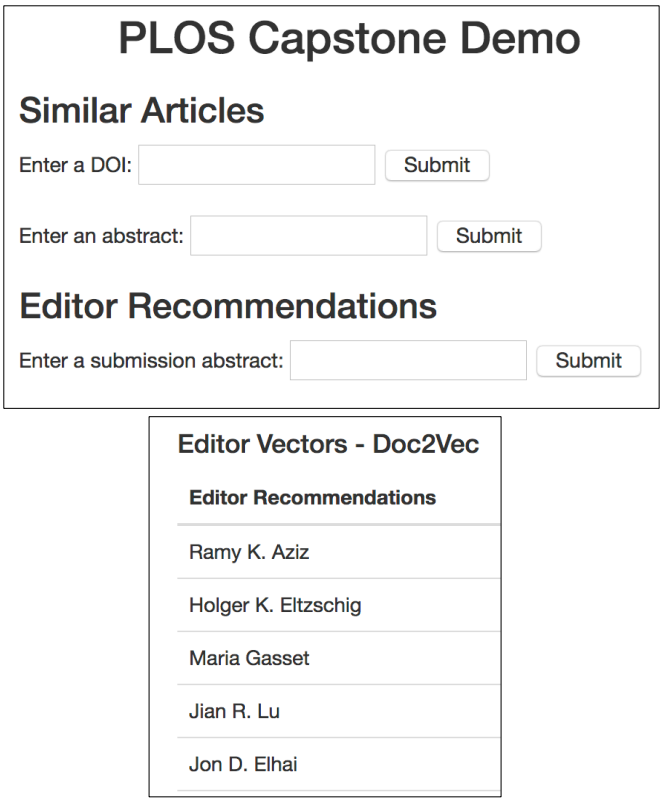

FIGURE IV
TOP: FLASK WEB APPLICATION HOMEPAGE
BOTTOM: SAMPLE RESULTS

Tables IV and V show editor recommendations for a sample abstract submission, sourced from the article about ovarian cancer diagnosis. It is disappointing that there is no overlap in the editor recommendations of the models, though given the large size of PLOS's corpus and the apparent thematic similarity of the similar articles, this is an acceptable outcome.

TABLE IV
SAMPLE RESULTS: DOC2VEC SINGLE ARTICLE EDITOR RECOMMENDATIONS

Editor Name)	Article Title
Gideon Schreiber	Barnase as a New Therapeutic Agent Triggering Apoptosis in Human Cancer Cells
Syed A. Aziz	Rapid Point-Of-Care Breath Test for Biomarkers of Breast Cancer and Abnormal Mammograms
Devanand Sarkar	Gastric Cancer Staging with Dual Energy Spectral CT Imaging
Arun Sreekumar	MALDI-ToF Mass Spectrometry for the Rapid Diagnosis of Cancerous Lung Nodules
Ramona Natacha	Validation of Reference Genes for Oral Cancer Detection Panels in a Prospective Blinded Cohort

TABLE V
SAMPLE RESULTS: LSA SINGLE ARTICLE EDITOR RECOMMENDATIONS

Editor Name)	Article Title
Massimo Ciccozzi	Ten-Year Mortality after a Breast Cancer Diagnosis in Women with Severe Mental Illness: A Danish Population-Based Cohort Study
Robert M Lafrenie	Breast Cancer Biology and Ethnic Disparities in Breast Cancer Mortality in New Zealand: A Cohort Study
Kalimuthusamy Natarajaseenivasan	Factors Associated with Uptake of Visual Inspection with Acetic Acid (VIA) for Cervical Cancer Screening in Western Kenya
Dimitrios Paraskevis	Knowledge about Cervical Cancer and Associated Factors among 15-49 Year Old Women in Dessie Town, Northeast Ethiopia Relationship between
Ali Montazeri	Cancer Worry and Stages of Adoption for Breast Cancer Screening among Korean Women

The unsupervised and subjective nature of this project make developing an objective measure of what constitutes a "good" editor recommendation very difficult. Additionally, we do not possess the subject matter expertise to qualitatively determine how appropriate an editor recommendation is outside of a cursory examination of their previous work. A possible solution is to design a controlled survey in which the PLOS's editorial team rates the quality of the editor recommendations for a sample of submissions as generated by our models and PLOS Match. Alternatively, a live A/B test could be conducted, whereby editors for some manuscript submissions are suggested by our proposed model. Then the editor acceptance rate (PLOS's KPI for this process) for our new model's suggestions is compared to a baseline control group where suggestions have been derived using PLOS Match.

FURTHER WORK

In addition to more fully evaluating the quality of our recommendations and eventual implementation of the models into PLOS Match, there are several areas in which further work can be done.

Through this project, we have developed several different models to match editors, but these models have been executed in isolation. There exists an opportunity to combine these approaches into an ensemble model to potentially further improve the recommendations. Incorporating our citation network analysis to contribute to final editor rankings could also be implemented to further enrich this model.

A further refinement of this model could allow for feedback from editors to be incorporated. The content that trains an editor vector could be refined or explicit rules could be stipulated to reflect editor preferences.

Additional model inputs could be also considered. In addition to including the manuscripts that an editor has edited, the articles that he or she has contributed to as an author could be explore.

PLOS has also recently developed a Python module called allofPLOS [15] that allows for live access to PLOS's entire corpus. As our project worked on a static copy of the corpus, implementing a pipeline that sourced newly published articles from allofPLOS, passed them to our preprocessing operations, and incorporated them into our models would be very beneficial and likely necessary to translate this project into a production environment.

The pipeline for this project is also generalized and can be implemented for any journal that is compliant with JATS

standards. All of the code for this project is openly available [16] and published under an MIT License. We welcome others to further refine and adapt our work.

ACKNOWLEDGMENT

We would like to thank the team at PLOS, including Veronique Kiermer, CJ Rayhill, and Elizabeth Seiver for their sponsorship and support of this project. We would also like to thank our project advisors from the University of Virginia Data Science Institute: Drs. Philip Bourne, L. P. Alonzi III, and Daniel Mietchen.

REFERENCES

[1] 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, Valletta, Malta, 2010, pp. 45–50.

[2] J. Beel, B. Gipp, S. Langer, & C. Breitinger. (2015). Research-paper recommender systems: a literature survey. International Journal on Digital Libraries, 17(4), 305–338. https://doi.org/10.1007/s00799-015-0156-0

[3] S. Bethard and D. Jurafsky, "Who should I cite: learning literature search models from citation behavior," in *Proceedings of the 19th ACM international conferenceon Information and knowledge management*, 2010, pp. 609–618.

[4] S. Gupta, and V. Varma. (2017). Scientific Article Recommendation by using Distributed Representations of Text and Graph. In Proceedings of the 26th International Conference on World Wide Web Companion - WWW '17 Companion. ACM Press. https://doi.org/10.1145/3041021.3053062

[5] M. J. Schuemie, J. A. Kors (2008). Jane: suggesting journals, finding experts. Bioinformatics, 24(5), 727–728. https://doi.org/10.1093/bioinformatics/btn006

[6] PLOS. *Text and Data Mining at PLOS.* Retrieved from https://www.plos.org/text-and-data-mining

[7] PubMed Central. *Open Access Subset.* Retrieved from https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/

[8] National Information Standards Organization (2015). *JATS: Journal Article Tag Suite, version 1.1* (ANSI/NISO Z39.96-2015). Retrieved from https://www.niso.org/publications/ansiniso-z3996-2015-jats-journal-article-tag-suite

[9] The ElementTree XML API (version 3.3). Retrieved from https://docs.python.org/3/library/xml.etree.elementtree.html

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

[11] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," J. Mach. Learn. Res., vol. 3, pp. 993–1022, Mar. 2003.

[12] S. Deerwester, S. T. Dumais, G. W. Furnas, T.K. Landauer and R. Harshman. (1990), Indexing by latent semantic analysis. J. Am. Soc. Inf. Sci., 41: 391-407.

[13] Q. Le and T. Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32* (ICML'14), Eric P. Xing and Tony Jebara (Eds.), Vol. 32. JMLR.org II-1188-II-1196.

[14] T. Huynh, K. Hoang, L. Do, H. Tran, H. Luong and S. Gauch, "Scientific publication recommendations based on collaborative citation networks," *2012 International Conference on Collaboration Technologies and Systems (CTS)*, Denver, CO, USA, 2012, pp. 316-321. doi: 10.1109/CTS.2012.6261069

[15] The Public Library of Open Science. allofplos (version 0.11.1). Retrieved from https://pypi.python.org/pypi/allofplos

[16] J. Prominski and P. Shah, PLOS Capstone. doi:10.5281/zenodo.1211801 https://github.com/jprominski/PLOS_Capstone

AUTHOR INFORMATION

Jack Prominski, M.S. Student, Data Science Institute, University of Virginia.

Pragati Shah, M.S. Student, Data Science Institute, University of Virginia.

Rafael Alvarado, General Faculty, Data Science Institute, University of Virginia.