

---

Researchers, Instructors, & Staff Scholarship

---

5-2022

## AI-Informed Approaches to Keyword Generation, Text Summarization, and Document Clustering for Improved Resource Discovery

Charlie Harper  
*Case Western Reserve University*

Anne Kumer  
*Case Western Reserve University, anne.kumer@case.edu*

Shelby Stuart  
*Case Western Reserve University*


Evan Meszaros  
*Case Western Reserve University, evan.meszaros@case.edu*

Author(s) ORCID Identifier:

 <https://orcid.org/0000-0003-3781-5812>

Follow this and additional works at: <https://commons.case.edu/staffworks>

 <https://orcid.org/0000-0001-8791-5964>

 Part of the [Cataloging and Metadata Commons](#)

 <https://orcid.org/0000-0003-1946-6233>

---

 <https://orcid.org/0000-0002-9500-0294>

### Recommended Citation

Harper, Charlie; Kumer, Anne; Stuart, Shelby; and Meszaros, Evan, "AI-Informed Approaches to Keyword Generation, Text Summarization, and Document Clustering for Improved Resource Discovery" (2022).

*Researchers, Instructors, & Staff Scholarship*. 1.

<https://commons.case.edu/staffworks/1>

This Book Chapter is brought to you for free and open access by Scholarly Commons @ Case Western Reserve University. It has been accepted for inclusion in Researchers, Instructors, & Staff Scholarship by an authorized administrator of Scholarly Commons @ Case Western Reserve University. For more information, please contact [digitalcommons@case.edu](mailto:digitalcommons@case.edu).

CWRU authors have made this work freely available. [Please tell us](#) how this access has benefited or impacted you!



## Chapter 8

# AI-Informed Approaches to Metadata Tagging for Improved Resource Discovery

*Charlie Harper, Anne Kumer, Shelby Stuart, and Evan Meszaros*

## Introduction

Academic and cultural institutions are grappling with problems of how to organize, label, and search disparate bodies of texts. As aggregators, preservers, and disseminators of substantial repositories of digital texts, research libraries are naturally situated at the heart of these problems. This chapter explores how unsupervised machine learning may be used to capture and simplify the complexity and nuances of text. Traditional approaches to improving discoverability and accessibility of text through metadata and controlled vocabularies have time-tested strengths. As the volume of digital data explodes, the obstacles and limitations of traditional approaches become more pronounced, and machine learning “show(s) the potential to create efficiencies that smooth the path to access, enhancing description and expanding forms of discovery along the way.”<sup>1</sup> In light of the need for new approaches to metadata generation to facilitate discovery, the authors look at Doc2Vec and topic modelling with Latent Dirichlet Allocation (LDA) to explore their utility as assistive

tools for authors, librarians, and readers. The authors apply the two approaches to a corpus of electronic theses and dissertations (ETDs) completed at Ohio universities and colleges.\*

## Current Issues in Metadata and Discovery

Searchability is one of the greatest advantages that online documents have over their print counterparts, and surveys show that users view this as a vital feature when asked about using e-resources over print.<sup>2</sup> Metadata quality influences the searchability and the discoverability of e-resources. Research databases and discovery layers rely on proprietary algorithms to generate and order results in response to the user's query. Relevance ranking algorithms may compare the query to metadata fields such as subject headings, publication titles, abstracts, and (sometimes) full text in order to determine the results. Therefore, search engines will return resources with greater effectiveness and precision when they have complete metadata and a useful set of subject headings. High-quality metadata is also a key component in ensuring that the most relevant documents appear at the top of the result set, where the user is more likely to find them.<sup>3</sup>

Studies by Tina Gross and her colleagues have examined the efficacy of controlled vocabularies for resource discovery. They established that, whether or not a user sees them, the existence of controlled vocabulary metadata, which depends on carefully assigned subject headings, generally contributes to up to one-third more positive search results than if that metadata was not there.<sup>4</sup> The research landscape, however, has changed dramatically due to Google's powerful influence, and keyword searching has exploded in popularity. The millions of documents that are commonly returned by keyword searches may overwhelm the user, while subject searches are able to return smaller sets of documents that are often more tailored to a user's query. Concurrently, several LIS scholars find that unregulated author-generated keywords enhance searches if they are employed in addition to subject headings from widely used controlled vocabularies assigned by librarians.<sup>5</sup>

The most widely used library-controlled vocabulary, the Library of Congress Subject Headings (LCSH), is maintained on the principle of literary warrant.<sup>6</sup> This has historically meant that only topics published in books warrant inclusion in the vocabulary's authorized headings lists. Vocabularies like the LCSH are slow to add new, potentially dubious terminology, essentially "controlling" its terms by applying parameters for use. This principle neglects formats, such as articles and dissertations, where scholarship is typically first published.<sup>7</sup> A contrasting principle is user warrant, which is based on user preference, need, and search patterns. Leaving out the specialized knowledge of a document's author potentially lessens discoverability because the LCSH is slow to include new specialized subject terminology. ANSI/NISO standards present literary and user warrant as complementary and equally important for search and discovery.<sup>8</sup> Author-generated keywords may yield many irrelevant search results, which the restriction of a controlled

---

\* This study's data sets, python notebooks, and trained models are provided on OSF (<https://osf.io/r6yhp/>) and are licensed under Creative Commons Attribution-ShareAlike 4.0.

vocabulary mitigates. Conversely, a controlled vocabulary imposes conservatism in the face of shifting cultural standards, which is balanced by author-generated keywords.

## ETDs and Subject Metadata

For many universities and colleges, the transition from print to electronic theses and dissertations began in the mid-2000s. With this format change, librarians were able to harvest author-supplied keywords from the electronic submission forms that accompanied ETDs and include those in the dissertation's catalog record alongside cataloger-supplied descriptive subject headings to enhance search and discovery. When selecting keywords, authors tend to choose those that represent their experiences and expectations rather than terms that derive from "any kind of controlled indexing language or authority-controlled procedure."<sup>9</sup> Personal experiences and social motivators, such as opinion, expression, performance, and activism, can impact keyword choice and result in both overly specific and overly broad keywords.<sup>10</sup> As Yelton notes for MIT's ETD repository, "Most of [the author-assigned keywords] are so granular that they apply to only one thesis and therefore don't collocate anything."<sup>11</sup> An ETD cataloged with only highly specialized or overly broad keywords does little to enhance search and discovery.

At the same time, ETDs are particularly important when researching topics that are new and emerging. McCutcheon notes that while print theses and dissertations tend to receive little attention, "it's not uncommon for ETDs to be downloaded hundreds or thousands of times, from all over the world."<sup>12</sup> As gray literature, however, ETDs do not benefit from the kinds of support that are offered by commercial publishers. They lack, for instance, standard distribution channels and presence on major publishers' web platforms. In addition, ETDs are not necessarily indexed by major abstracting and indexing services, which can make them difficult to discover. ETDs are accordingly a prime dataset for projects that aim to improve metadata and increase discoverability. In order to address this problem, the authors elected to work with ETDs published at Ohio colleges and universities. These ETDs are hosted by OhioLINK,<sup>†</sup> a consortium of over one hundred academic institution members across the state of Ohio, and they have consumable metadata available through the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH).<sup>‡</sup> ETDs published on OhioLINK are globally accessible, free of charge, and frequently include born-digital PDFs.

The authors wrote a series of Python notebooks to generate a dataset of OhioLINK ETDs. First, the authors used Python's Sickle library to pull Dublin Core metadata for ETDs that were published and uploaded in 2019. From the Dublin Core XML results, the authors created one CSV of the title, abstract, publication date, source university/college, URI, and rights restrictions, as well as a second CSV of the keywords assigned to each ETD. The final dataset consisted of metadata for 3,316 ETDs from thirty-six Ohio universities and colleges and 13,141 non-unique keywords.

---

† See: <https://www.ohiolink.edu/>.

‡ See <https://www.openarchives.org/pmh/>.

Representation of the thirty-six Ohio universities and colleges was highly uneven within the dataset. For example, Ohio State University produced 843 ETDs, while smaller institutions produced only one. The different academic focuses of each institution likely means that the subject areas of the dataset are skewed. Keywords that occur over one hundred times give a sense of how the subjects trend (table 8.1). Since 85.88 percent (11,285) of keywords occur only once, however, this list should be read cautiously. Likewise, the length of the abstracts is highly varied, which may further bias the dataset toward particular subject areas.

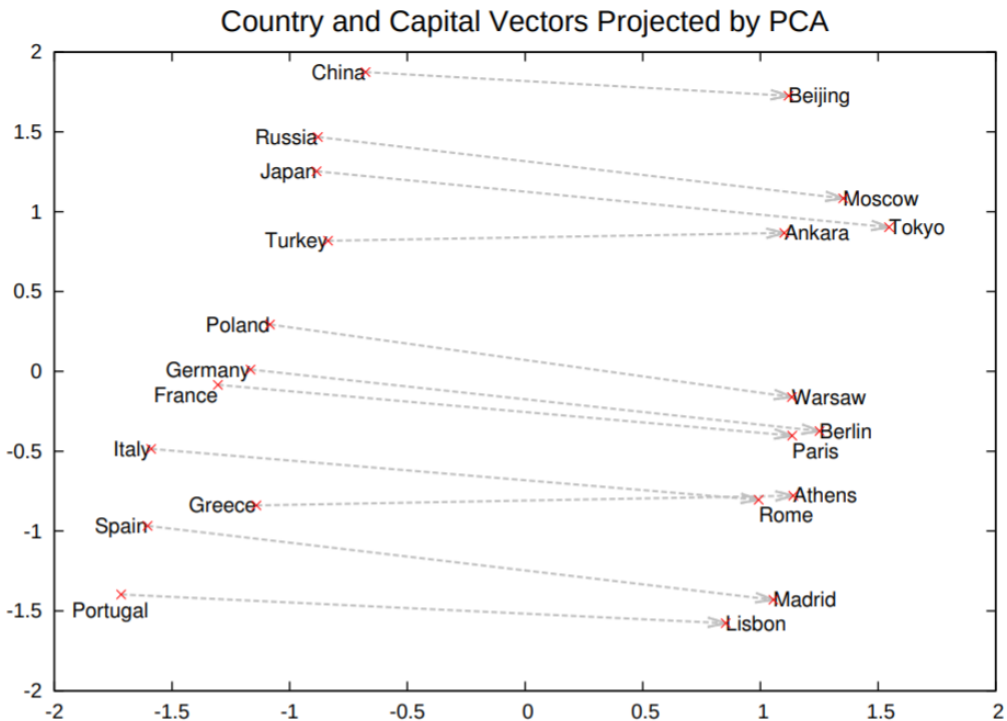
**Table 8.1**

Keywords that occur more than 100 times in the dataset of 3,316 ETDs. These keywords hint at how the content of the dataset may be skewed toward certain subjects.

Keyword	Occurrences
psychology	220
biology	175
education	169
mechanical engineering	154
chemistry	134
electrical engineering	133
computer science	128
communication	107
literature	106

## Tagging ETDs with Doc2Vec and DBPedia

Doc2Vec is an approach that learns to map units of text, such as sentences, paragraphs, or full documents, into a numerical vector space.<sup>13</sup> It is an extension of an earlier, and still frequently used, incarnation known as Word2Vec, which worked with single words.<sup>14</sup> Both Word2Vec and Doc2Vec are built on a neural network architecture that trains on a corpus of text and learns how to represent text as coordinates in a high-dimensional space.<sup>15</sup> The value of these learned coordinates is that the topology of the vector space in which the text is embedded holds information on the content or meaning of the text. For example, embedded texts that are located more closely should also show a closer semantic relationship. Mathematical connections between points can also reveal deeper linguistic relationships. With single words, one can discover antonyms, synonyms, declensions, or conjugations based on spatial relationships (figure 8.1). Doc2Vec extends Word2Vec's capabilities to texts of any length.



**Figure 8.1**

A classic example of how Word2Vec can capture meaning is the relationship between capital cities and countries learned from a corpus of text. The spatial relationship between the learned word embeddings for country and capital reflects the semantic relationships between the words in text.<sup>16</sup>

Doc2Vec has shown particular application in document retrieval systems, where a user can search for documents whose content or subject is related to an input document. In the library world, Yelton used Doc2Vec in her app, HAMLET, to calculate the similarity between graduate theses at MIT.<sup>17</sup> As Yelton notes, however, Doc2Vec cannot assign meaningful labels to related documents in the traditional sense of metadata keywords or subject headings, nor can it draw boundaries to create discrete categories of documents.<sup>18</sup> This is part of a larger issue with unsupervised machine learning, which reveals similarities in data but still requires humans to assign meaningful labels or keywords. In order to overcome this limitation and to automatically generate content-specific keywords, the authors trained Doc2Vec on a corpus of text generated from DBpedia, a large linked and already-labeled dataset.<sup>19</sup> The authors then tagged a sample of OhioLINK ETDs with the trained model to assess its effectiveness.

# DBPedia and Model Training

DBPedia\* is a knowledge base that classifies content using descriptive terms as well as the contextual relationships of its content. As a source for descriptive keywords, DBPedia has multiple strengths: it is crowdsourced and likely to remain more current than controlled vocabularies; its entries are internally linked to enhance semantic queries; it provides a URI, keyword, and abstract for each idea; its keywords are frequently multilingual; and a single abstract and URI can map to multiple keywords that capture the same idea.

The authors used Python’s SPARQLWrapper library† to gather three hierarchical levels of data from DBPedia’s SPARQL endpoint, which the authors termed page-level, subject-level, and concept-level. Page-level data is the finest grained and maps to a single entry with an abstract. Subject-level data is marked by the RDF verb “dct:subject-of” and aggregates related page-level data. Concept-level data is marked by the RDF verb “skos:broadener-of” and aggregates subject-level data. Neither subjects nor concepts possess abstracts. The three should respectively represent a continuum from more specific to more general ideas (figure 8.2).

The DBPedia dataset consisted of 4,935,271 pages.‡ Abstracts ranged from 1 to 168,193 words with an average of 525 words. Initial experiments with the entire body of abstracts showed poor results, which the authors attributed to the prevalence of shorter abstracts that did not convey enough meaning. Therefore, the authors removed all but the 75th through 99.9th percentile of abstracts based on word count. The authors felt the resulting range of 648 to 5,127 words was more reasonable. This subset of 1,230,980 abstracts constituted the training set for the Doc2Vec model.

The authors used Python’s Gensim library to build the Doc2Vec model.<sup>20</sup> Because model accuracy can be difficult to measure in unsupervised learning, the past work on Doc2Vec with Wikipedia, the computational time for training, and the authors’ interpretation of experimental results guided hyperparameter choices.<sup>21</sup> Ultimately, the authors chose to use a continuous bag of words with a vector space of 500 dimensions. DBPedia abstracts were preprocessed by removing non-alphanumeric characters, stopwording, and lemmatizing. Training took approximately 2.5 hours on an Amazon Web Services (AWS) r5.4xlarge instance. After training, a k-d tree was built from the embedded page vectors stored in the Doc2Vec model in order to speed the search for the closest (measured by Euclidean<sup>§</sup> distance) points in 500 dimensions.<sup>22</sup>

To test the efficacy of this approach, the authors tagged a selection of ETD abstracts with the page-level keywords that were closed in vector space. Tagging was accomplished by first embedding an ETD’s preprocessed abstract in 500-dimensional space with the trained Doc2Vec model and then searching the k-d tree for the n-nearest points, each of

---

\* See <https://wiki.dbpedia.org/>.

† See <https://rdflib.dev/sparqlwrapper/>.

‡ This study employed the DBPedia version 2016–10 release for page-level metadata and abstracts (<https://wiki.dbpedia.org/develop/datasets/dbpedia-version-2016-10>).

§ Euclidean distance extends the measure of distance as expressed in the Pythagorean Theorem to n-dimensions.

## About: Alan Turing

## Page Level

An Entity of Type : [scientist](#), from Named Graph : <http://dbpedia.org>, within Data Space : [dbpedia.org](#)

Alan Mathison Turing OBE FRS (*/tjʊərɪŋ/*; 23 June 1912 – 7 June 1954) was an English computer scientist, mathematician, logician, cryptanalyst and theoretical biologist. He was highly influential in the development of theoretical computer science, providing a formalisation of the concepts of algorithm and computation with the Turing machine, which can be considered a model of a general purpose computer. Turing is widely considered to be the father of theoretical computer science and artificial intelligence.

dct:subject

- [dbc:1912\\_births](#)
- [dbc:1954\\_deaths](#)
- [dbc:20th-century\\_mathematicians](#)
- [dbc:20th-century\\_philosophers](#)
- [dbc:Academics\\_of\\_the\\_University\\_of\\_Manchester](#)
- [dbc:Academics\\_of\\_the\\_University\\_of\\_Manchester\\_Institute\\_of\\_Science\\_and\\_Technology](#)
- [dbc:Alan\\_Turing](#)
- [dbc:Alumni\\_of\\_King's\\_College,\\_Cambridge](#)
- [dbc:Artificial\\_intelligence\\_researchers](#)
- [dbc:Atheist\\_philosophers](#)

## About: 20th-century mathematicians *Subject Level*

An Entity of Type : [Concept](#), from Named Graph : <http://dbpedia.org>, within Data Space : [dbpedia.org](#)

skos:broader

- [dbc:Mathematicians\\_by\\_century](#)
- [dbc:20th-century\\_scholars](#)
- [dbc:20th-century\\_people\\_by\\_occupation](#)

## Concept Level

### Figure 8.2

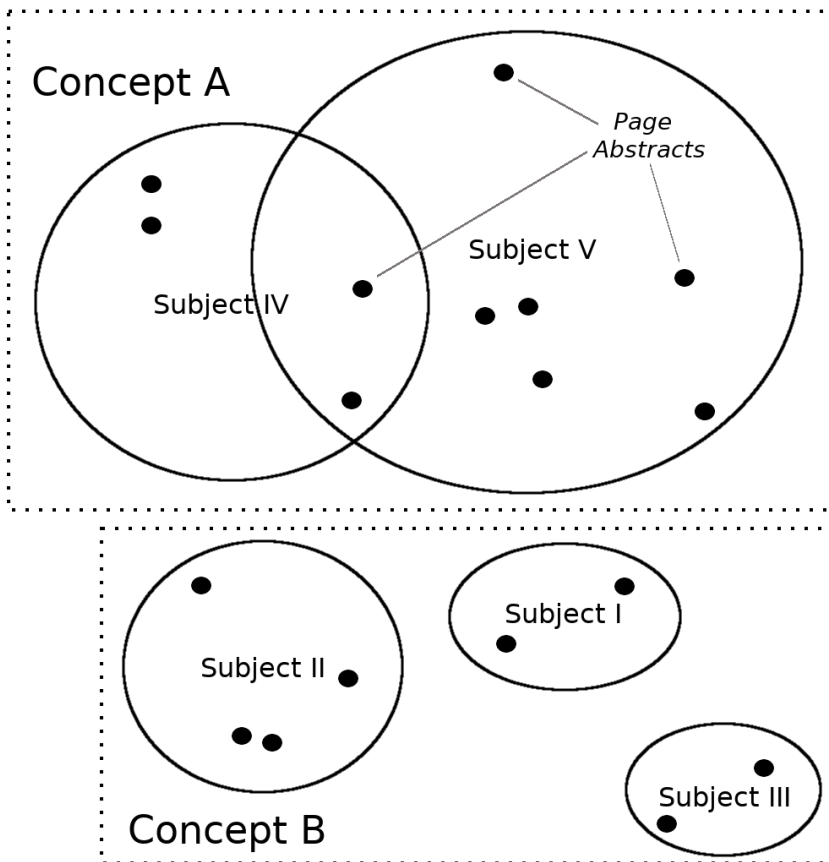
An example of a partial page with abstract ([http://dbpedia.org/page/Alan\\_Turing](http://dbpedia.org/page/Alan_Turing)) and a linked subject ([http://dbpedia.org/page/Category:20th-century\\_mathematicians](http://dbpedia.org/page/Category:20th-century_mathematicians)) with reference to its higher concepts.

which corresponds to one DBpedia page. The results were extremely poor and typically nonsensical. For example, one thesis on college students' perceptions of conservation efforts was tagged “Keg\_stand”! The authors concluded that the information contained at the page level was overly specific and that the vector space was likely too densely packed with points. To overcome this, the authors utilized the linked nature of DBpedia to move up to the subject (“dct:subject”) and concept (“skos:broader”) levels for tagging.

The subset of 1,230,980 abstracts linked to 728,752 subjects and 421,051 concepts. Subjects mapped to a range of 1 (e.g., “[Crocodile\\_Dundee\\_Films](#)”) to 177,622 (“[Living\\_People](#)”) page-level abstracts. Concepts mapped to a range of 1 (e.g., “[1130s\\_in\\_Europe](#)”)



to 5,063 (“Songs\_by\_songwriter”) subjects. Because of the interlinked nature of DBPedia, there is overlap between subject and concept keywords. To build a k-d tree for the subject level, the vectors of each subject’s pages were averaged together. For the concept level, the vectors of each concept’s subjects were averaged together (figure 8.3). The trained Doc2Vec model was unaltered.



**Figure 8.3**

Illustration of moving from page- to subject- to concept-level in the vector space using relationships stored in DBPedia. For example, Subject IV contains four pages with abstracts, represented by black dots. These four points, which in reality are 500-dimensional, are averaged together to create Subject IV, a new, 500-dimensional point. To create Concept A, Subject IV and Subject V are averaged together. As one moves from page to concept, the vector space becomes sparser and content should become more general. Note that pages can belong to multiple subjects, and subjects can belong to multiple concepts.

In order to test this approach for subject and concept tagging, the authors sampled 250 ETDs published in 2019. The sample was stratified by university/college in order to reflect the distribution of institutions in OhioLINK. The Doc2Vec model was used

to embed each ETD's preprocessed abstract into vector space and then the subject and concept k-d trees were searched to find the five nearest subjects and concepts as measured by Euclidean distance (table 8.2).

**Table 8.2**

An example of subject and concept DBpedia tags assigned to an ETD entitled, "Development of a Conformal Additive Manufacturing Process and its Application."<sup>23</sup>

	1	2	3	4	5
<b>Subject</b>	Nanotechnology	Materials_ science	Lithography_ (microfabrication)	Microtechnology	Semiconductor_ device_ fabrication
<b>Concept</b>	Microtechnology	Computer- aided_ engineering	Materials_ science	Forming_ processes	Instrumental_ analysis

## Results

The individual authors each rated 125 ETD's subject and concept tags to ensure that tags were always rated by two separate individuals. For simplicity, each rater marked the relevance of the tag -1 (not relevant), 0 (somewhat relevant), or 1 (relevant). The ratings were then averaged across raters. Averaged ratings for subjects were more relevant, on average, than for concepts. In both cases, moving from the first subject or concept (closest in space) to the fifth subject or concept (farther in space) showed a downward trend in ratings (table 8.3).

The mean subject rating was  $0.32548 \pm 0.057$ . The mean concept rating was  $0.23496 \pm 0.057$ . Subjects and concepts were, therefore, both ranked as being "somewhat relevant" on the whole to the ETDs. This result is far from perfect, but it is very promising. While page-level tagging produced no meaningful results, at the subject and concept level, this approach is capturing meaning and assigning viable keywords based only on an abstract.

**Table 8.3**

The mean and 95% confidence interval for subject and concept ratings based on a sample of 250 tagged ETDs.

	Mean	95% Lower	95% Upper
subject1	0.4630	0.379	0.547
subject2	0.3817	0.294	0.469
subject3	0.3471	0.261	0.433
subject4	0.2396	0.155	0.324
subject5	0.1942	0.102	0.286
concept1	0.3104	0.230	0.390
concept2	0.3389	0.257	0.420

**Table 8.3**

The mean and 95% confidence interval for subject and concept ratings based on a sample of 250 tagged ETDs.

	Mean	95% Lower	95% Upper
concept3	0.1925	0.102	0.282
concept4	0.1958	0.105	0.287
concept5	0.1208	0.035	0.207
subject_avg	0.32548	0.2684	0.3825
concept_avg	0.23496	0.1785	0.2915

## Finding Relevant Information with Topic Modeling

Topic modeling is a generative statistical approach that clusters related content. This term is commonly a stand-in for the more specific topic modeling algorithm, Latent Dirichlet Allocation, or LDA.<sup>24</sup> The approach is often employed in fields that engage with large corpora of textual data. In the academic library, researchers have already used topic modeling to cluster ETDs and government documents for enhanced discovery to generate alt-metrics by mining book reviews and to recommend tags for enhancing metadata records.<sup>25</sup>

Despite its value in certain applications, there are notable shortcomings with topic modeling. Foremost, “topic” is a misnomer. As a statistical method, LDA produces a statistical distribution of words that constitute a “topic” and a statistical distribution of “topics” across documents. Often, scholars will choose the top  $n$  words to represent a topic, but LDA does not produce a label for a topic, nor does it guarantee the top  $n$  words are meaningful to a human reader. Second, LDA requires a preset number of topics. There are methods to best determine this, but if a trained model is continuously applied to a naturally growing corpus, such as is the case with ETDs, the number of topics is unable to organically grow with the changing content.

For these reasons, the authors believe that topic modeling retains immense use for clustering fixed corpora of text but that it is less useful for a living corpus. While an approach like the combined Doc2Vec and DBpedia above is best situated to generate metadata to improve the discovery of resources within a large, living corpus of ETDs, topic modeling is better suited to enhance discovery of specific information within an ETD, which is, in effect, a fixed corpus.

## ETD Full Text and Model Training

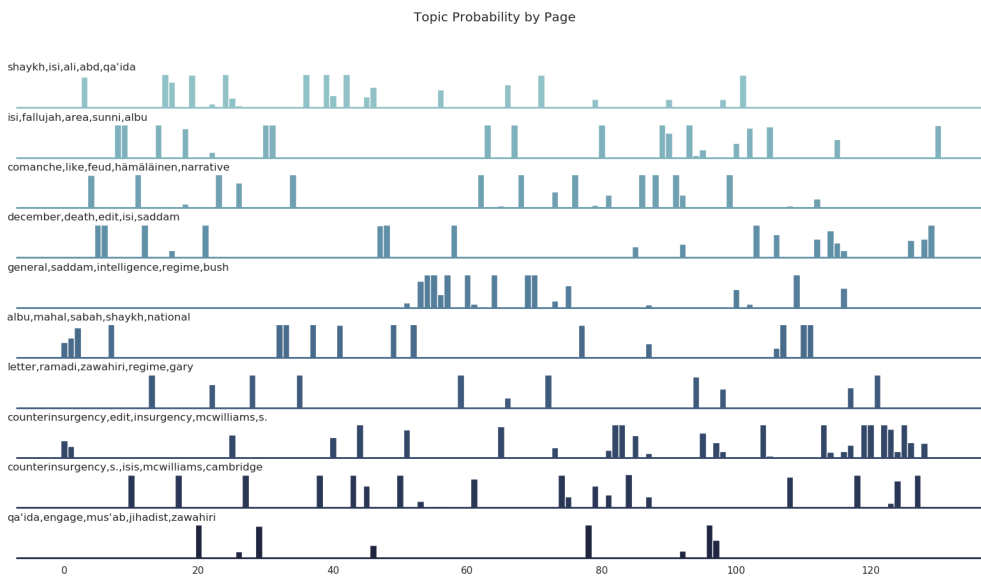
To exemplify the authors’ proposal that LDA is most useful for internal information discovery, topic models were trained on the full text of ten ETDs from the previous sample of 250. The full text was extracted from each PDF in Python. Because of difficulties in

working with non-standardized ETDs, the authors chose the page as the basic unit of analysis when training the topic models. No other preprocessing was done.

An LDA model was trained on each ETD's set of pages using the Gensim library. The number of topics was set at ten, which seemed reasonable to capture enough nuance in ETDs of variable length. The model used only words that appeared on at least five pages but fewer than 25 percent of pages. After training, a CSV of topic distributions for each page was generated and the top five words for each topic were stored. The LDA model was then discarded.

## Results

Assessing the results of topic models is difficult and requires specialized knowledge and deeper engagement with each ETD's content. Visualizing the results, however, does show the strong potential of this approach for assisting readers in finding information within an ETD. Figure 8.4 shows the visualization of topic distributions by page for an MA thesis entitled, "Enduring Failure: A Borderlands History of the Iraq War and its Aftermath."<sup>26</sup> Without hyperparameter tuning, the LDA model has produced generally good topics. The fifth topic, "general, saddam, intelligence, regime, bush," is an example of this. The topic is absent from the first portion of the text and clusters around pages in the fifties and sixties. If a reader were interested in the rhetoric, personalities, and intelligence that led up to the Iraq War, this would indicate that the reader should glance at these pages first.



**Figure 8.4**

The distribution of topics across the pages of an ETD on the Iraq War.<sup>27</sup> Ten topics are presented, from top to bottom, with the top five words for each topic. The mixture of each topic by page is shown from left to right.

# Conclusion

A Doc2Vec model trained on DBPedia's linked content and topic models trained on individual ETDs show promise as tools to enhance metadata and discovery. Both approaches outlined above warrant deeper study and the authors are pursuing ways to improve and better assess their efficacy. Regardless, these approaches seem well-poised to inform human metadata creation and discovery efforts but not to replace them. Although the Doc2Vec subject and concept tags were generally relevant to the ETDs' abstracts, there is substantial room for improvement and model tuning. In addition, finding ways to better tune topic models to individual ETDs would produce stronger results. In the course of this work, the authors made numerous observations that are guiding their ongoing work. Many observations additionally reflect deeper issues with the rising tide of machine learning in the library. Although only a handful of these can be enumerated here, the authors find it beneficial to conclude with the following:

1. It is difficult to judge model effectiveness. Rating machine-generated tags and topics require a baseline level of subject expertise and familiarity with terminology, which is especially important when documents in the sample set have been written by and for graduate-level researchers. Of the authors, those who had educational backgrounds in the social sciences and humanities struggled to assess the relevance of some tags assigned to, for instance, physics and engineering ETDs. It is, therefore, advisable to engage with subject-matter experts when assessing the effectiveness of machine learning approaches to tagging and discovery.
2. Linked data augment discovery. One oft-noted benefit of employing controlled subject headings is that they integrate the ETDs with materials that share the same subject but are published in different formats. This increases the visibility of the ETDs, which otherwise may only be retrievable by searching within a particular repository or library collection and exposes them to a much broader range of researchers.<sup>28</sup> Utilizing keywords drawn from DBPedia's linked data set may offer an additional way to interlink ETDs with other academic resources. Moreover, following links between keywords may facilitate the sort of serendipitous discovery that can occur when browsing print items on a library shelf.
3. All subjects are not created equal. Abstracts for humanities ETDs, such as those describing poetry collections, creative writing, theater productions, and others, were less likely to be assigned relevant tags. This could be related to the tendency of those abstracts to have smaller word counts than their STEM counterparts. Moreover, the authors observed a lack of accuracy and specificity in tagging ETDs that examine certain understudied communities and locations. Among the sampled ETDs, this issue seemed particularly common among those that focused on Latin America. For example, an ETD studying public performances in Colombia was tagged "Argentine Art," and one describing ecological research in the Peruvian Andes was tagged "Forestry in Brazil." As mentioned previously, the ETDs are likely biased toward certain subject areas as are the DBPedia abstracts.

These biases in datasets become reified in machine learning models and can contribute to results that show an even stronger bias.

4. Humans and machines need balance. Authors choose keywords from a place of ownership and perceived use of their scholarship, librarians apply subject headings in compliance with best practices and parameters for metadata quality control, and machine learning models select terms or topics according to patterns learned from human-supplied data. No one method is ideal, and a balance between the strengths and weaknesses of each is needed; the human capability to shift perspective and interpret words or phrases in different contexts is not directly replicated by machine learning methods, while a machine learning model's ability to rapidly process huge corpora cannot be directly replicated by a human. Mediating the differing roles and biases of author, librarian, and machine requires ongoing research and human devotion to consistency. Cataloging best practices remains essential for quality control when applying machine learning techniques to resource description.

## Endnotes

1. Thomas Padilla, *Responsible Operations: Data Science, Machine Learning, and AI in Libraries* (Dublin, OH: OCLC Research, 2019), 12, <https://doi.org/10.25333/xk7z-9g97>.
2. Gabrielle Wiersma and Esta Tovstiadi, "Inconsistencies between Academic E-Book Platforms: A Comparison of Metadata and Search Results," *portal: Libraries and the Academy* 17, no. 3 (2017): 628.
3. Robert Losee, "The Effect of Assigning a Metadata or Indexing Term on Document Ordering," *Journal of the American Society for Information Science and Technology* 64, no. 11 (2013).
4. Tina Gross, Arlene G. Taylor, and Daniel N. Joudrey, "Still a Lot to Lose: The Role of Controlled Vocabulary in Keyword Searching," *Cataloging & Classification Quarterly* 53, no. 1 (2015): 1–39, <https://doi.org/10.1080/01639374.2014.917447>.
5. Tom Steele and Nicole Sump-crethar, "Metadata for Electronic Theses and Dissertations: A Survey of Institutional Repositories," *Journal of Library Metadata* 16, no. 1 (2016): 53–68, <https://doi.org/10.1080/19386389.2016.1161462>; Gross, Taylor, and Joudrey, "Still a Lot to Lose"; Sevim McCutcheon, "Basic, Fuller, Fullest: Treatment Options for Electronic Theses and Dissertations," *Library Collections, Acquisitions, & Technical Services* 35, no. 2–3 (2011): 64–68, <https://doi.org/10.1016/j.lcats.2011.03.019>.
6. Cynthia R. Strader, "Author-Assigned Keywords Versus Library of Congress Subject Headings," *Library Resources and Technical Services* 53, no. 4 (2011): 243–51, <https://doi.org/10.5860/lrts.53n4.243>.
7. Strader, "Author-Assigned Keywords."
8. Ibid.
9. Michalis Gerolimos, "Tagging for Libraries: A Review of the Effectiveness of Tagging Systems for Library Catalogs," *Journal of Library Metadata* 13, no. 1 (2013): 41, <https://doi.org/10.1080/19386389.2013.778730>.
10. Alla Zollers, "Emerging Motivations for Tagging: Expression, Performance, and Activism," in *WWW 2007: Proceedings of the 16th International Conference on World Wide Web, Banff, (Alberta, Canada), May 8-12, 2007*, ed. Carrie Williamson and Mary E. Zurko (New York: ACM Press, 2007). [http://www.conference.org/www2007/workshops/paper\\_55.pdf](http://www.conference.org/www2007/workshops/paper_55.pdf).
11. Andromeda Yelton, "HAMLET: Neural-Net-Powered Prototypes for Library Discovery," in *Artificial Intelligence and Machine Learning in Libraries*, ed. Jason Griffey (Chicago: ALA TechSource, 2019), 12.
12. McCutcheon, "Basic, Fuller, Fullest," 64.

13. Quoc V. Le and Tomas Mikolov, “Distributed Representations of Sentences and Documents,” in *Proceedings of the 31st International Conference on Machine Learning, PMLR 32* (2014): 1188–96, <https://arxiv.org/abs/1405.4053>.
14. Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean, “Efficient Estimation of Word Representations in Vector Space,” Cornell University (2013), 1–12, <https://arxiv.org/abs/1301.3781>.
15. Hobson Lane, Hannes M. Hapke, and Cole Howard, *Natural Language Processing in Action: Understanding, Analyzing, and Generating Text with Python* (Shelter Island, NY: Manning Publications, 2019), 181–217.
16. Mikolov et al., “Distributed Representations of Words,” fig. 2.
17. Yelton, “HAMLET.”
18. *Ibid.*, 11–12.
19. M. Allahyari and K. Kochut, “Semantic Tagging Using Topic Models Exploiting Wikipedia Category Network,” in *2016 IEEE Tenth International Conference on Semantic Computing (ICSC)* (New York: IEEE Xplore, 2016), 63–70, <https://doi.org/10.1109/ICSC.2016.34>.
20. Radim Řehůřek and Petr Sojka, “Software Framework for Topic Modelling with Large Corpora,” in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks* (Valletta, Malta: ELRA, 2010), 45–50.
21. Jey H. Lau and Timothy Baldwin, “An Empirical Evaluation of Doc2vec with Practical Insights into Document Embedding Generation,” *Proceedings of the 1st Workshop on Representation Learning for NLP* (2016), 78–86.
22. G. Bonaccorso, *Hands-On Unsupervised Learning with Python: Implement Machine Learning and Deep Learning Models Using Scikit-Learn, TensorFlow, and More* (Birmingham: Packt Publishing, 2019), 66–71.
23. Faez Alkadi, “Development of a Conformal Additive Manufacturing Process and Its Application” (MA Thesis, University of Akron, 2019).
24. D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet Allocation,” *Journal of Machine Learning Research* 3 (2003): 993–1022.
25. Yelton, “HAMLET”; Jonathan Cain, “Using Topic Modeling to Enhance Access to Library Digital Collections,” *Journal of Web Librarianship* 10, no. 3 (2016); Zhou Qingqing and Zhang Chengzhi, “Measuring Book Impact via Content-Level Academic Review Mining,” *The Electronic Library* 38, no. 1 (2020): 138–54, <https://doi.org/10.1108/EL-08-2019-0184>; Suppawong Tuarob, Line C. Pouchard, and C. Lee Giles, “Automatic Tag Recommendation for Metadata Annotation Using Probabilistic Topic Modeling,” in *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL '13* (New York: Association for Computing Machinery, 2013), 239–48, <https://doi.org/10.1145/2467696.2467706>.
26. Terry T. Tait, “Enduring Failures: A Borderlands History of the Iraq War and Its Aftermath” (MA Thesis, Miami University, 2019).
27. Tait, “Enduring Failures.”
28. McCutcheon, “Basic, Fuller, Fullest.”

## Bibliography

- Alkadi, Faez. “Development of a Conformal Additive Manufacturing Process and Its Application.” MA Thesis, University of Akron, 2019.
- Allahyari, M., and K. Kochut. “Semantic Tagging Using Topic Models Exploiting Wikipedia Category Network.” In *2016 IEEE Tenth International Conference on Semantic Computing (ICSC)*, 63–70. New York: IEEE Xplore, 2016. <https://doi.org/10.1109/ICSC.2016.34>.
- Blei, D. M., A. Y. Ng, and M. I. Jordan. “Latent Dirichlet Allocation.” *Journal of Machine Learning Research* 3 (2003): 993–1022.
- Bonaccorso, G. *Hands-On Unsupervised Learning with Python: Implement Machine Learning and Deep Learning Models Using Scikit-Learn, TensorFlow, and More*. Birmingham: Packt Publishing, 2019.
- Cain, Jonathan. “Using Topic Modeling to Enhance Access to Library Digital Collections.” *Journal of Web Librarianship* 10, no. 3 (2016).

- Gerolimos, Michalis. "Tagging for Libraries: A Review of the Effectiveness of Tagging Systems for Library Catalogs." *Journal of Library Metadata* 13, no. 1 (2013): 36–58. <https://doi.org/10.1080/19386389.2013.778730>.
- Gross, Tina, Arlene G. Taylor, and Daniel N. Joudrey. "Still a Lot to Lose: The Role of Controlled Vocabulary in Keyword Searching." *Cataloging & Classification Quarterly* 53, no. 1 (2015): 1–39. <https://doi.org/10.1080/01639374.2014.917447>.
- Lane, Hobson, Hannes M. Hapke, and Cole Howard. *Natural Language Processing in Action: Understanding, Analyzing, and Generating Text with Python*. Shelter Island, NY: Manning Publications, 2019.
- Lau, Jey H., and Timothy Baldwin. "An Empirical Evaluation of Doc2vec with Practical Insights into Document Embedding Generation." *Proceedings of the 1st Workshop on Representation Learning for NLP* (2016), 78–86.
- Le, Quoc V., and Tomas Mikolov. "Distributed Representations of Sentences and Documents." In *Proceedings of the 31st International Conference on Machine Learning, PMLR 32, 2* (2014), 1188–96. <https://arxiv.org/abs/1405.4053>.
- Losee, Robert. "The Effect of Assigning a Metadata or Indexing Term on Document Ordering." *Journal of the American Society for Information Science and Technology* 64, no. 11 (2013).
- McCutcheon, Sevim. "Basic, Fuller, Fullest: Treatment Options for Electronic Theses and Dissertations." *Library Collections, Acquisitions, & Technical Services* 35, no. 2–3 (2011): 64–68. <https://doi.org/10.1016/j.lcats.2011.03.019>.
- Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. "Efficient Estimation of Word Representations in Vector Space," Cornell University (2013), 1–12. <https://arxiv.org/abs/1301.3781>.
- Mikolov, Tomas, Ilya Sutskever, Kai Chen, G. S. Corrado, and Jeffrey Dean. "Distributed Representations of Words and Phrases and Their Compositionality." *Advances in Neural Information Processing Systems* 26 (2013).
- Padilla, Thomas. "Responsible Operations: Data Science, Machine Learning, and AI in Libraries." Dublin, OH: OCLC Research, 2019. <https://doi.org/10.25333/xk7z-9g97>.
- Qingqing, Zhou, and Zhang Chengzhi. "Measuring Book Impact via Content-Level Academic Review Mining." *The Electronic Library* 38, no. 1 (2020): 138–54. <https://doi.org/10.1108/EL-08-2019-0184>.
- Řehůřek, Radim, and Petr Sojka. "Software Framework for Topic Modelling with Large Corpora." In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 45–50. Valletta, Malta: ELRA, 2010.
- Steele, Tom, and Nicole Sump-crethar. "Metadata for Electronic Theses and Dissertations: A Survey of Institutional Repositories." *Journal of Library Metadata* 16, no. 1 (2016): 53–68. <https://doi.org/10.1080/19386389.2016.1161462>.
- Strader, Cynthia R. "Author-Assigned Keywords versus Library of Congress Subject Headings." *Library Resources and Technical Services* 53, no. 4 (2011): 243–51. <https://doi.org/10.5860/Irts.53n4.243>.
- Tait, Terry T. "Enduring Failures: A Borderlands History of the Iraq War and Its Aftermath." MA Thesis, Miami University, 2019.
- Tuarob, Suppawong, Line C. Pouchard, and C. Lee Giles. "Automatic Tag Recommendation for Metadata Annotation Using Probabilistic Topic Modeling." In *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries*, 239–48. JCDL '13. New York: Association for Computing Machinery, 2013. <https://doi.org/10.1145/2467696.2467706>.
- Wiersma, Gabrielle, and Esta Tovstiadi. "Inconsistencies Between Academic E-Book Platforms: A Comparison of Metadata and Search Results." *portal: Libraries and the Academy* 17, no. 3 (2017): 617–48.
- Yelton, Andromeda. "HAMLET: Neural-Net-Powered Prototypes for Library Discovery." In *Artificial Intelligence and Machine Learning in Libraries*, edited by Jason Griffey, 10–15. Chicago: ALA TechSource, 2019.
- Zollers, Alla. "Emerging Motivations for Tagging: Expression, Performance, and Activism." In *WWW 2007: Proceedings of the 16th International Conference on World Wide Web, Banff, (Alberta, Canada), May 8-12, 2007*, edited by Carrie Williamson and Mary E. Zurko. New York: ACM Press, 2007. [http://wwwconference.org/www2007/workshops/paper\\_55.pdf](http://wwwconference.org/www2007/workshops/paper_55.pdf).



