CHAPTER 4

Topic Modeling as a Tool for Resource Discovery

SHAWN GOODWIN, ATLA, AND EVAN KUEHN, NORTH PARK UNIVERSITY

As theological librarians look toward future developments in religious studies disciplines, many of the humanistic interpretive questions asked by researchers will remain the same. The biblical scholar will continue to explain the textual, philological, or ideological/theological coherence of biblical texts even as new methods for doing so are developed. Within systematic theology, the classic formula of *faith seeking understanding* articulated by Anselm of Canterbury has remained applicable in twentieth-century theologies and will remain a touchpoint in the future.

What will change, and what are currently in the process of immense change, are the methodological and technological aspects of theological research that allow us to ask and answer increasingly complex questions about sacred texts and religious communication. Although the digital humanities are sometimes (and often rightly) maligned by theologians as merely faddish, there are many examples of how computational methods are opening new possibilities for textual analysis, especially in biblical studies but also increasingly in theological research (see Anderson 2018; Robinson 2019). The fundamental problems of theology are not changed by digital methods, but the tools we have at our disposal for engaging in theological research have changed.

We are interested in investigating how digital humanities tools can address new problems of complexity within theology. In particular, we are interested in how topic models can be useful for determining new directions in theological research. In this paper, we will demonstrate how topic models can be used as a tool for resource discovery in emerging fields of study.

What Is a Topic Model?

Topic modeling is a statistical approach to grouping discrete collections of words based on similarity. The two dominant approaches used today are LDA (Latent Dirchlet Analysis) and NMF (Non-negative Matrix Factorization). "Topics" are groups of words that occur in proximity within a text. Each word is given a statistical weight that aligns with one of the topics.

Topics have often been used in digital humanities research to identify patterns in discourse for analysis (see Saxton n.d.). This work supplements the more traditional close reading approaches of literary and historical work by using computational methods of "reading," and is especially helpful for working with large bodies of texts. For example, Jeri Wieringa (2019) has used topic modeling to describe and visualize the relationship between end-times expectations and gender in early Seventh Day Adventist literature. Wieringa's research examined 31 periodicals spanning 77 years of publication, analyzing this literature at a depth that would not be possible with traditional methods of reading.

This is how topic models are typically used in digital humanities. In our case, we wanted to employ such models earlier in the process. If topic models are able to identify research-relevant patterns in texts, then could they also be used to recognize the research-relevance of texts for traditional (i.e., non-digital) modes of text analysis? To employ topic models in this way, we trained a topic model on a smaller, more recent, specific corpus that we selected for relevance using typical (i.e., keyword search) methods of discovery, and then used that model to filter works from a much larger historical corpus of political theology texts to identify any promising matches for the specific topics we were interested in.

Identifying New Knowledge in Theology

Resource discovery is often a solitary task, performed by theologians (or, if they are well-funded, by their research assistants) in preparation for a particular project. The description and organization of theological literature that makes discovery possible in the first place, however, involve a highly interconnected set of processes that are, in turn, sensitive to the changing nature of the research literature itself. When new constellations of research knowledge are produced, typical ways of describing research knowledge need to be adjusted. There can be a delay in learning what adjustments are most appropriate. There can also be an inability, because of various constraints, to go back to existing material and reorganize it in a way that might be more suitable under new circumstances. Soumenin and Toivannen (2016) have recently examined the helpfulness of

unsupervised machine learning techniques for mapping new scientific knowledge in such situations, noting the inherent limitations of traditional descriptive metadata for identifying new constellations of scientific work:

> Preexisting categories of science provide a finite definition of new knowledge, fitting knowledge that is by definition infinite and new to the world into preexisting categories and coordinates[…] They are best at monitoring the behavior of known and defined bodies of knowledge, but lend themselves poorly–if at all–to correctly identifying the emergence of truly new epistemic bodies of knowledge. (Soumenin and Toivannen 2016, 2464)

Scientists working in physical, natural, and social scientific disciplines are well aware of the complex shifting ground upon which they work, and so computational approaches in these fields are already well established. Theology, along with other humanities disciplines, tends to lag behind in its embrace of digital humanities approaches. Where it does employ computational methods, it tends to use them for text mining, textual analysis, and visualization, rather than to map new knowledge.

Theological researchers often have much more traditionalist, even nostalgic, conceptions of their discipline and do not understand theology as a field in which emergent problems fundamentally change the nature of theological knowledge. For the good of the discipline, though, theological librarians need not grant that this problematic self-understanding of theological research is the case. They should investigate ways to most effectively engage with emerging constellations of theological problems so that new research is not overly restrained by descriptive schema that do not adequately map onto new theological questions.

New or dynamic fields of study present obvious challenges for mapping scientific knowledge, but they also present challenges for the researcher related to resource discovery of existing knowledge. Typical theological research is conducted with ready-made maps of knowledge available in the catalog and database metadata, but research in new fields may lack adequate descriptive metadata. Either the description of new texts is simply lacking, or it is inadequate because it does not capture new terminology or logical relationships that make these theological texts novel, much less connect these new relationships with older ones. In these situations, the theologian is left to map the new territory for themselves in an ad hoc fashion.

It is also difficult to identify which older texts might be applicable to the new theological situation, especially when new terminology is employed. For instance, searching a term like “Dreamer” (as in the DACA program) will not turn up any meaningfully related texts from the twentieth century, although there are surely

older texts that communicate relevant concepts and ideas using different words. We propose that topic modelling is a tool that can help librarians and researchers alike as they tackle complex domains of new theological knowledge. Topic modelling can connect these domains with existing theological texts on the basis of patterns in these two otherwise discontinuous discourses.

We have focused this study on the emerging theological subfield of migration studies. Political discourse in the United States about the treatment of immigrants from Latin America, as well as the influx of refugees at a global level, most notably as a result of the Syrian Civil War, have made questions of migration and refugee identity an important, prominent, and growing subfield of political theology. As a result, terminology and research questions related to migration studies in theology are not as well established as more traditional fields such as christology, ecclesiology, or church history. For researchers conducting literature reviews and seeking out source material for the production of new knowledge in this and similar fields, it will be important to have discovery tools that are able to recognize and describe the relevance of sources in new and more complex ways.

Topic Models for Resource Discovery: Theology and Migration

Creating the Model

First, we established a small corpus of known texts on theology and migration, from which we could derive topics that would guide our discovery of unknown texts in this subject area. We identified a small corpus of representative texts (monographs, edited volumes, and journal articles) published from 2010–19, using search keywords of [migration OR refugee OR immigrant] with [theology OR religion].[1]

Using PDFs of these texts, we created objects from the content with stop words removed (e.g., articles, commonly used words, etc.) and obvious misspellings and misdivided words fixed. We trained a model using the LDA (Latent Dirchlet Allocation) algorithm on these texts in order to generate topics that would be coherent but not have significant overlap. The visualization of these topics in figure 1 shows which words are most definitive of the topic and a visualized proximity of each topic to the others. This visualization also shows the dominance of particular words and the size of the topic in the entire corpus.

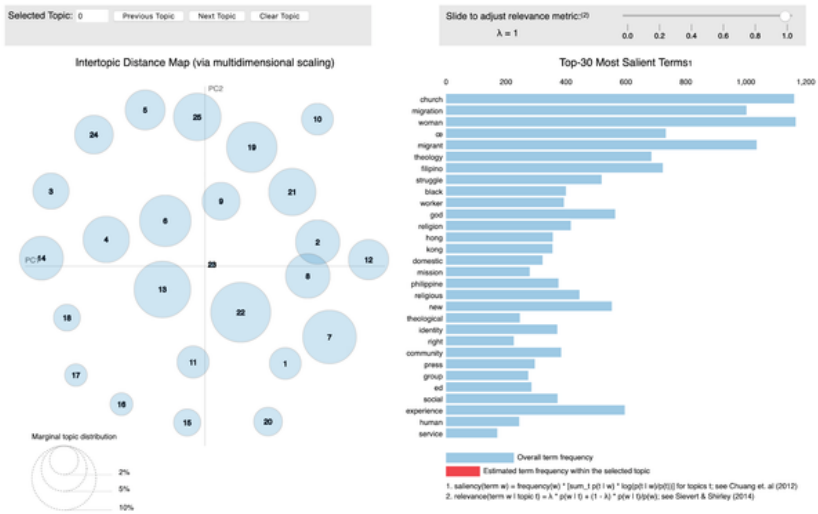

Figure 1

This figure shows a visual representation of the topic model we developed. An interactive version of this model can be viewed on *efkuehn.github.io/topicmodeldiscovery/docs/visualising_topic_model.html*.

One of the problems with topic modeling is that, because it is an unsupervised clustering method, sometimes the computer sees connections that are not obvious or, at the very least, are not *semantic* clusters. A topic model is a blunt tool, but we picked six of these topics that we thought might be helpful in discovering books over the past 100 years that might build on the topic we had chosen. We gave these topics headings in order to give them some sort of identifying description. The headings were derived from the word clusters as well as the pages that were best represented by these topics. We chose these topics because we thought they were coherent and might provide interesting analysis when looked at in the political theology corpus generated from HathiTrust.

These topics are:

– topic number: 0

  - heading: Black Experience
  - key terms: 'black, experience, life, mean, like, make, point, american, challenge, relation'

– topic number: 1

  - heading: Context of Migrant Experience

    - key terms: ‘identity, challenge, term, experience, context, question, migrant, people, state’

– topic number: 3

    - heading: Communal Experience
    - key terms: ‘migrant, country, home, community, family, experience, life, economic, new, reality’

– topic number: 5

    - heading: Social, Political, Economic Migrations
    - key terms: ‘social, political, economic, immigrant, society, cultural, perspective, issue, people, life’

– topic number: 6

    - heading: Immigration and American Christianity
    - key terms: ‘church, christian, american, immigrant, community, role, state, faith’

– topic number: 11

    - heading: Religion and Culture
    - key terms: ‘religion, religious, culture, cultural, Christian, identity, faith, experience, example, time’

These are the only six topics we looked for in the HathiTrust corpus that we had identified. When using topic models for resource discovery, the researcher will need to make their own decision on what level of breadth or specificity will best fit the scope of their project.

Applying the Topic Model

The HathiTrust collection was created by searching for all works that might be related to Political Theology.[2] The collection on HathiTrust was further filtered by matching the OCLC numbers with connection to pull out valid subject headings. This left a series of about 9,000 books. These books were further filtered to exclude any that did not have an English language tag. The final count of HathiTrust books we gathered was 6,260. One of the limitations of the topic model we are using is that it can only be used for a single language. Although some work has been done

on topic models in multilingual contexts, this is an area where the digital humanities will need to improve upon current options.[3]

Distribution of Topics

We grouped the records by decade and then proceeded to count all of the pages that were dominated by the specific topics. Sorting the data this way gives a nice overview of the corpus (figure 2). These counts correspond to the amount of materials we had from each year. It also shows that Topic 5 is the most common in this corpus. When we went back and looked at the keywords for this topics, it became clear why these words showed up so commonly. Many of these books aren't necessarily about migration exactly, but society, politics, and economics are covered thoroughly in many of the books in our corpus.

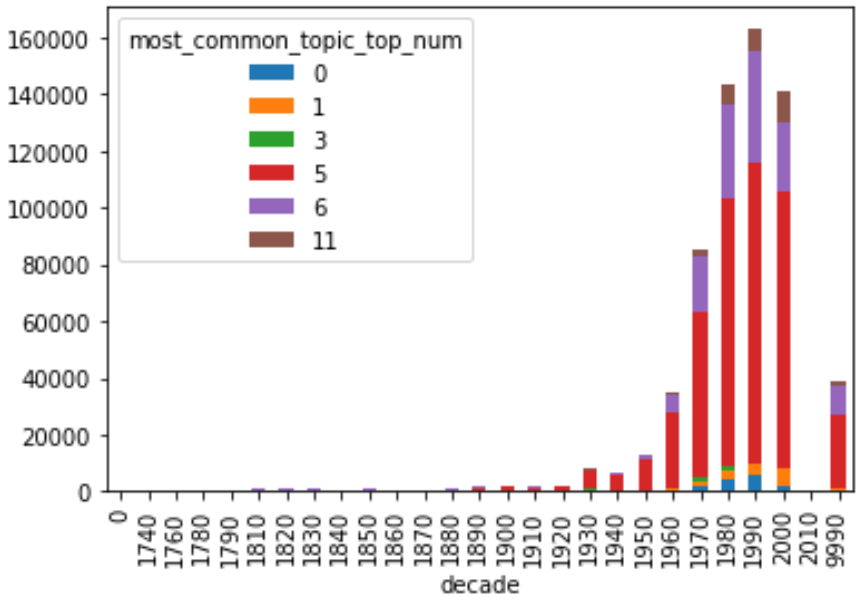

Figure 2

Topic Average Distribution

We also averaged the topic fit percentage across the corpus for each decade. One interesting aspect of this chart (figure 3) is the decades that don't include any instances of one or more of the topics. It would be worth investigating further if this is just a weakness in our corpus or if it reflects a trend in the period.

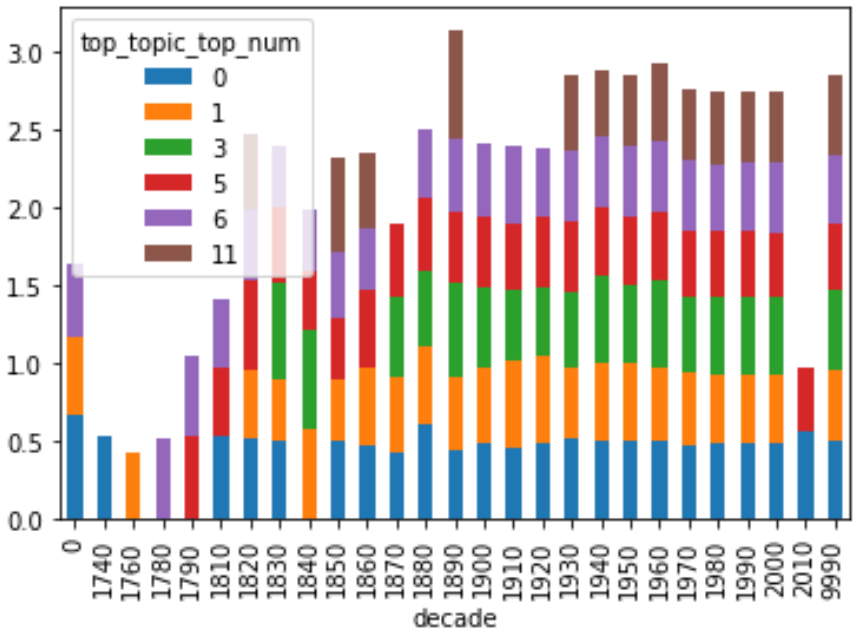

Figure 3

Using the Topics For Discovery

We filtered the topic matches based on which pages had the greatest percentage of a match, as well as having the most pages that had that topic as the dominant topic. These topics produced a lot of noise, but in that noise many interesting potential books also appeared in the results. The following books seem to relate well or suggest interesting, if non-obvious, connections with our area of study:

– Topic 0: Black Experience

- *Selected Black American, African, and Caribbean Authors: A Bio-bibliography* / compiled by James A. Page and Jae Min Roh. 1985.

> *Subjects:* Authors, Black Biography Dictionaries. | Caribbean literature Black authors Bio-bibliography. | African literature Bio-Bibliography. | African American authors Biography Dictionaries. | American literature African American authors Bio-bibliography.

- *Afro-American Life, History and Culture* / developed for USIS Programs by the Collections Development Branch, Library Programs Division, Office of Cultural Centers and Resources, Bureau of Educational and Cultural Affairs, United States Information Agency. 1985.

  > *Subjects:* African Americans Social conditions Bibliography. | African Americans Bibliography.

– Topic 1: Context of Migrant Experience

  - *Who's Who in American Jewry*, v. 3, 1938–1939. 1939.

    > *Subjects:* Jews Biography Periodicals. | Jews United States Biography Periodicals.

  - *Joy of the Worm* / Sargeson, Frank. 1969.[4]

– Topic 5: Social, Political, Economic Migration

  - *The Blackwell Companion to Globalization* / edited by George Ritzer. 2007.

    > *Subjects:* Internationalisation | Globalization.

  - *Social Aspects of Alienation: An Annotated Bibliography* / Mary H. Lystad. 1969.

    > *Subjects:* Social Problems abstracts. | Social Isolation abstracts. | Alienation (Social psychology) Bibliography.

  - *Bystanders to the Holocaust* / edited with an introduction by Michael R. Marrus, v. 3. 1989.

    > *Subjects:* Jews United States Politics and government. | Jews Palestine Politics and government. | Jewish refugees. | Holocaust, Jewish (1939–1945) Public opinion.

– Topic 6: Immigration and American Christianity

  - *The Indian Church* / Virag Pachpore. 2001.

    > *Subjects:* Christianity India.

- *Dalits in India: Religion as a Source of Bondage or Liberation with Special Reference to Christians* / James Massey. 1995.

  > *Subjects:* Dalits India Religion. | Discrimination India. | Caste India. | Christians India.

- *An Introduction to the Reformed Tradition: A Way of Being the Christian community* / John H. Leith. 1977.

  > *Subjects:* Reformed Church Doctrines

– Topic 11: Religion and Culture

- *The Enlightenment, An Interpretation: The Rise of Modern Paganism* / Gay, Peter (1923–). 1966.

  > *Subjects:* Philosophy History. | Enlightenment. | Europe Intellectual life.

- *Predicting Religion: Christian, Secular, and Alternative Futures* / edited by Grace Davie, Paul Heelas, Linda Woodhead. 2003.

  > *Subjects:* Twenty-first century Forecasts. | Christianity Forecasting. | Religion Forecasting.

- *Ernst Troeltsch and the Future of Theology* / edited by John Powell Clayton. 1976.

  > *Subjects:* Troeltsch, Ernst, 1865–1923 Congresses.

Applications

This project demonstrated the potential usefulness of topic modeling for exploring a larger corpus and for providing a supplement to traditional library metadata. However, it also illustrates some challenges to be aware of. The parameters of a topic model can vastly improve its ability to provide helpful results for research. Two parameters that we should have improved on are the filter extremes and the number of topics. Filter extremes are a type of limiter for a model, functioning in a similar manner to stop words. The researcher can set filter extremes to cut out any words that do not occur in very many documents, or on the other hand to cut out words that occur in most or all of the documents. Each of these cut-offs is simply a predetermined numerical value. This can help the model from relying too much on

the extremes of language use. However, because we built our model on such a small corpus and then applied it to a much larger, historical corpus, the extremes of our training set may have actually been part of what we were looking for. In light of this difficulty, another way of improving this study would be to more carefully curate both the training set and the larger corpus explored using the topic model. Many of the works in this larger corpus have little or nothing to do with theology, and we should remove some of them. In addition, our training corpus could have been more carefully selected around a specific topic and made larger. Both of these steps could vastly improve the results of our experiment.

A further step could be to filter the larger HathiTrust corpus on a constellation of topics. We could use two or three topics that create an interesting look at a topic. One way this could work is to train the topic model so that migration and theology are distinct and coherent topics, and look for a work that has a predominance of both of those topics. If we were to do this, we would need to record more than just the top-ranked topic for a document, but also the second and third as well. This approach would also be a promising way to build on our current project by focusing the scope of the topics used.

JSTOR's Text Analyzer also uses a topic model to match uploaded papers to additionally interesting ones.[5] However, JSTOR's method is nearly the reverse of ours: JSTOR starts with their large corpus, and then tries to fit the new paper into the model. Instead, we start with a smaller data set and try to filter things out. JSTOR's approach is a more traditional use for topic modeling. However, there are other algorithms for matching texts that might provide better approaches for discovery. Algorithms like Google's PageRank algorithm could also be leveraged for digital humanities projects like ours.[6]

Work has also been done on using topic models for query expansion, and our own project can be understood in terms of query expansion, insofar as our initial selection of a text corpus for training our topic model was an initial query, which we then expanded for use in further resource discovery (see Yi and Allen 2009). Some of the strategies mentioned above for curating the data set, as well as constraining the model, will help in guiding the task of query expansion.

One benefit of using topic modelling for resource discovery, in contrast to more typical uses of topic models in digital humanities, is what can be called the "low stakes" of this use. Bernard Schmidt (2012) has offered caution about the helpfulness of topic models for discovering conceptual patterns in text corpora, point out that, "excitement about the use of topic models for discovery needs to be tempered with skepticism about how often the unexpected juxtapositions LDA creates will be helpful, and how often merely surprising. A poorly supervised machine learning algorithm is like a bad research assistant. It might produce some unexpected constellations that show flickers of deeper truths; but it will also

produce tedious, inexplicable, or misleading results." At the stage of resource discovery (rather than "discovery" within textual analysis itself), this possibility of merely apparent pattern recognition is still present. That said, a researcher who is going through an initial selection of relevant sources is not performing textual analysis, but rather is merely identifying texts using analysis of topics as a preliminary way to judge relevance. This use of topic modelling will rarely, if ever, be paired with a topic model's typical use of text analysis later on in the research project, since a corpus of texts determined at such a level of specificity at the stage of resource discovery will be too small and selective to be used for genuine analysis using topic models. Recall from some of the examples above, topic models in the digital humanities usually examine long runs of periodicals or large amounts of longitudinal data, rather than bibliographies of works preselected as relevant to a particular research project.

The low stakes of resource discovery do not give the researcher license to use topic modeling without proper care, however. Resource discovery is "high stakes" in its own ways. It takes time to teach an algorithm how to function properly and time for it to process large amounts of literature. When texts are identified using topic models, a false positive may not lead to faulty analysis in published research; more likely it will be recognized as irrelevant and discarded. But this time wasted working with irrelevant texts that do not help the research process is a real cost of topic modeling for resource discovery. Whether such wrong turns and wasted time are any more present using this method than they are in resource discovery using subject authorities or single keyword searches is an open question. But this is a cost of which researchers should be cognizant.

Topic modeling for resource discovery is a tool that should be used when the significance of the research project warrants such measures. It is also a tool that should be further developed by information specialists and even formally incorporated into library or database discovery layers. Ideally, theologians who are open to digital humanities methodologies and theological librarians who are equipped to engage at a deeper level with the content of emerging fields of study will work together to improve upon these and other new tools for theological research.

Works Cited

Anderson, Clifford. 2018. "Digital Humanities and the Future of Theology: What Potential Does Digital Humanities have to Shape the Practice of Theology? Are There Theological Questions at Stake?" *Cursor: Zeitschrift für explorative*

Theologie, 25 July 2018. cursor.pubpub.org/pub/anderson-digitalhumanities–2018.

Kuehn, Evan and Shawn Goodwin. 2018. "Indexing the Theologico-Political." *Atla Summary of Proceedings* 72: 168–72. doi.org/10.31046/proceedings.2018.129

Robinson, Matthew Ryan. 2019. "Embedded, Not Plugged-In: Digital Humanities and Fair Participation in Systematic Theological Research." *Open Theology* 5, no. 1: 66–79. doi.org/10.1515/opth-2019-0005.

Saxton, Micah. n.d. *Best Practices for Topic Modelling*. Accessed July 8, 2020. msaxton.github.io/topic-model-best-practices/.

Schmidt, Benjamin. 2012. "Words Alone: Dismantling Topic Models in the Humanities." *Journal of Digital Humanities* 2, no. 1. journalofdigitalhumanities.org/2–1/words-alone-by-benjamin-m-schmidt/.

Suominen, Arho and Hannes Toivanen. 2016. "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, no. 10: 2464–76.

Vulić, Ivan, Wim de Smet, Jie Tang, and Marie-Francine Moens. 2015. "Probabilistic Topic Modeling in Multilingual Settings: An Overview of Its Methodology and Applications." *Information Processing & Management* 51, no. 1: 111–47. doi.org/10.1016/j.ipm.2014.08.003.

Wieringa, Jeri E. 2019. "A Gospel of Health and Salvation: Modeling the Religious Culture of Seventh-day Adventism, 1843–1920." PhD diss. George Mason University. dissertation.jeriwieringa.com/.

Yi, Xing and James Allen. 2009. "A Comparative Study of Utilizing Topic Models for Information Retrieval." *Proceedings of the European Conference on Information Retrieval* 5478: 29–41.

Notes

1. The full project site for Topic Modeling as a Tool for Resource Discovery is available at efkuehn.github.io/topicmodeldiscovery/. The bibliography for the textual corpus is available at docs.google.com/document/d/1sXHkN6WsW_SwG5xLSRPLS-DPqbwbQCHcB8ErrIbxu0M/edit?usp=sharing.
2. The HathiTrust collection can be viewed here: babel.hathitrust.org/cgi/mb?a = listis&c = 1154484. This collection was originally created for Evan Kuehn and Shawn Goodwin, "Indexing the Theologico-Political," *Atla Summary of Proceedings* 72 (2018): 168–72, doi.org/10.31046/proceedings.2018.129.

3. Some work has been done on multilingual models. For example, this model creates a bilingual LDA model: Ivan Vulić et al., "Probabilistic Topic Modeling in Multilingual Settings: An Overview of Its Methodology and Applications," *Information Processing & Management* 51, no. 1 (January 1, 2015): 111–47, doi.org/10.1016/j.ipm.2014.08.003. In general, more work needs to be done in multilingual and low-resource language natural language processing techniques.
4. This is a novel that does not have any fulltext available in HathiTrust, or for that matter any description readily available online. The source itself may not end up being helpful, but it is interesting insofar as it is the sort of text that a typical search for migration-related literature would not turn up.
5. See JSTOR Labs, *Test Analyzer: About*, www.jstor.org/analyze/about.
6. For example, see this implementation in Python: github.com/ashkonf/PageRank.