Evidence Based Library and Information Practice

Research Article

First-Year Students and the Framework: Using Topic Modeling to Analyze Student Understanding of the Framework for Information Literacy for Higher Education

Melissa Harden
First Year Experience Librarian
Hesburgh Libraries
University of Notre Dame
Notre Dame, Indiana, United States of America
Email: mharden@nd.edu

Abstract

Objective – The Framework for Information Literacy for Higher Education has generated a significant amount of discussion among academic librarians; however, few have discussed the potential impact on learning when students interact directly with the Framework itself. At the University of Notre Dame, over 1,900 first-year students completed an information literacy assignment in their required first-year experience course. Students read a condensed version of the Framework, then wrote a response discussing how a frame of their choosing was reflected in an assigned reading. The goal of this exploratory study was to determine if the students demonstrated an understanding of the themes and concepts in the Framework based on this assignment.

Methods – Topic modeling, a method for discovering topics contained in a corpus of text, was used to explore the themes that emerged in the students' responses to this assignment and assess the degree to which they connect to frames in the Framework. The model receives no information about the Framework prior to the analysis; it only uses the students' words to form topics.

Results – The responses formed several topics that are recognizable as related to the frames from the Framework, suggesting that students were able to engage effectively and meaningfully with the language of the Framework. Because the topic model does not know anything about the Framework, the fact that the responses formed topics that are recognizable as frames suggests that students internalized the concepts in the Framework well enough to express them in their own writing.

Conclusion – This research provides insight regarding the impact that the Framework may have on student understanding of information literacy concepts.

Introduction

The Framework for Information Literacy for Higher Education has generated a significant amount of discussion and interest since it was first introduced. Librarians have shared ideas for lesson plans and collaboration with other campus partners to help students understand the concepts described in the Framework. However, little is known about the direct role of the Framework on student learning. In other words, what happens when students are presented with language from the Framework and asked to respond to it? This article describes a study that explores the themes that emerge in student writing from an assignment based on the Framework.

Beginning with the 2015-2016 academic year, all first-year students at the University of Notre Dame, almost 2,000 in total, enroll in a two-semester first-year experience course. Throughout the course, students engage with a number of important topics, including developing successful study habits, financial literacy, mental and physical health and wellbeing, and information literacy. During the Fall 2016 semester, the students completed an information literacy assignment that was created by four librarians at the university. In this assignment, the students read an article about one researcher's description of the troublesome and often challenging aspects of the research process (Tompkins, 1986) and watched a TED Talk about online filter bubbles and their effect on our understanding of truth and knowledge

(Pariser, 2011). Additionally, the students read an adapted and condensed version of the Framework for Information Literacy for Higher Education and wrote a short response of approximately 200 words to a prompt in which they were asked to make connections between the article they read and the Framework (see Appendix A for the full text of the prompt). The librarians who created the information literacy assignment for the course provided a modified shorter version of the Framework for the students to read because they thought the original document contained too much jargon. While some of the language in the adapted version was simplified or reduced, the underlying concepts described in the Framework mostly remained the same.

The responses to the assignment provide a wealth of information available for analysis to explore students' understanding of the Framework. However, because nearly 2,000 students completed the assignment, manually coding the themes in each written response would be an unwieldy, time-consuming task with many opportunities for human error. Therefore, topic modeling was used to analyze the responses. Topic modeling is a useful method in this case because it can provide information about the patterns in the text with minimal structure or guidance, and potential bias, from the researcher.

The goal of this exploratory study was to determine if the students demonstrated an understanding of the themes and concepts in the

Framework based on this assignment. The responses formed several topics that are recognizable as related to the frames from the Framework. Further, the frames that most commonly appeared in the topics generated by the model relate well to the content of the article they read, indicating meaningful interpretation of the assignment. Because the topic model did not receive information about the Framework itself, but rather formed topics based solely on students' writing, this suggests that students internalized the concepts in the Framework well enough to express them in their own writing. The remainder of this paper reviews past literature on the Framework and topic modeling, describes the procedures used for data collection and cleaning, provides details on modeling the texts of student responses, and presents an analysis of the results.

Literature Review

The Framework

The introduction of the Framework marked a significant shift in the practice and discussion of library instruction. Although there has been support for the Framework since its inception, many expressed concerns about its potential value and structure. For example, Demspey, Dalal, Dokus, Charles, and Scharf (2015) considered whether the Framework would "receive the same kind of widespread recognition and endorsement as the *Standards*" by national organizations and agencies because the document does not outline standardized learning outcomes (p. 165). Wilkinson (2014) expressed concern about the use of threshold concepts in the Framework, and Beilin (2015) noted that threshold concepts "may end up functioning as the means to merely reinforce disciplinary boundaries and institutional hierarchies." However, many librarians used the introduction of the Framework to reflect on and reevaluate their teaching (Burgess, 2015) and to explore how theory impacts their practice (Holliday, 2017). Specifically, the introduction of the Framework provided an opportunity to

focus on fostering critical thinking, drawing deeper connections (Pagowsky, 2015), and creating opportunities for students to grapple with the complexity of information and its context (Seeber, 2015). Additionally, the new document provided grounds for talking with campus partners and establishing, or reestablishing, librarians' place on campus as educators and collaborators (Gibson & Jacobson, 2014; Lancaster, Callender, & Heinz, 2016; McClure, 2016). A beneficial feature of the Framework is its adaptability to meet individual campus needs. For example, Witek (2016) describes making connections between her institution's core curriculum and information literacy outcomes, and Oakleaf (2014) offers strategies for developing assessments appropriate for the local context. In a specific example, Swanson (2017) co-led a six-week course where disciplinary faculty engaged deeply with the Framework and discussed how the frames could influence their teaching.

Even though survey results indicate that there is variation in the degree to which librarians say the Framework influences their instruction (Julien, Gross, & Lathan, 2018), there are many examples of lessons and suggestions for incorporating concepts described in the Framework in instruction, whether it be in one-shot sessions or full-semester courses (for example, Bauder & Rod, 2016; Bravender, McClure, & Schaub, 2015; Jacobson & Gibson, 2015; Mays, 2016). Kuglitsch (2015) discusses how the threshold concept theory underpinning the Framework can be used to situate information literacy in the disciplines.

There are several examples of the Framework influencing engagement with first-year students. Librarians at DePaul University used the Framework as a lens to analyze student responses to a library orientation activity (Dempsey & Jagman, 2016), while librarians at Texas Tech University collaborated with a composition instructor to redesign library activities for a first-year composition course, drawing heavily upon the Framework

(Lancaster et al., 2016). As a part of the redesign, students were presented with the language of relevant frames from the Framework within specific activities. Even though students are not the intended audience for the document, they presented students with the frames in order to provide context for "the why behind the instruction and, more importantly, the why behind a productive research process" (Lancaster et al., 2016, p. 90). However, the authors noted that because the Framework is "highly theoretical" and the language is "too dense for most freshmen students," they suggested providing a paraphrased version of the Framework that is easier to understand to students in the future (p. 91).

In a study documented in two articles, Rachel E. Scott (2017a, 2017b) explored whether student engagement with Framework language improved their understanding of information literacy concepts. To assess engagement, students enrolled in a credit-bearing course focused on research methods took a pre-test at the beginning of the semester (Scott, 2017a). Each of the frames in the Framework was presented to the students, followed by two open-ended questions relating to each frame. Student responses to the questions were analyzed to determine how first-year students "respond to the language and concepts of the Framework" and whether or not the "concepts used in the Frames fit in with undergraduates' existing understanding of research practices" (Scott, 2017a, p. 2). In the second article describing this project, Scott (2017b) outlines the results of the post-test the students took at the end of the semester, after having engaged with the frames throughout the course. The same survey instrument was used in the post-test as was used in the pre-test. Overall, the students in the course showed improvement in their responses from the pre-test to the post-test. The author noted that, "[s]everal responses indicate that students have internalized the Frames and are not merely parroting the language provided in the questions" (Scott, 2017b, p. 25). As a result, Scott concludes that student interaction

with the language of the Framework had a positive effect on their understanding of information literacy concepts.

This paper builds on previous work by exploring the themes that emerge in actual student writing about the Framework. Topic modeling is used to identify and analyze the themes (i.e., topics) running through the text of students' responses. This method has the advantage of allowing a wide range of potential topics to emerge based on how students reacted to the Framework. It provides a useful complement to the researcher-directed forms of inquiry seen in past work, such as surveys.

Topic Modeling

Probabilistic topic modeling is useful for analyzing large amounts of textual data. Probabilistic topic models are "statistical methods that analyze the words of the original texts to discover the themes that run through them, how those themes are connected to each other, and how they change over time" (Blei, 2012, pp. 77-78; see also Blei, Ng, & Jordan, 2003). Topic modeling allows for analyzing themes that emerge in a group of texts, collectively referred to as a corpus. Topics are clusters of words that appear together frequently throughout the corpus, allowing researchers to identify the hidden structure of the texts. The topics that emerge in the analysis may not be apparent to a researcher making a manual observation of the corpus. This study uses Latent Dirichlet Allocation (LDA), a specific type of topic model, which "treats each document as a mixture of topics, and each topic as a mixture of words" (Silge & Robinson, 2017, p. 90). In the case of the present study, the topics that emerge from the corpus of student responses provide information about how students wrote about the Framework and the way they saw its concepts reflected in an assigned reading. A key benefit of this method is that the model detects the topics and brings forth the underlying structure of the text archive, allowing for an exploratory analysis of

the text while minimizing any effect a researcher's preconceived ideas might have on the analysis.

There are several examples of the use of topic modeling to assess written student responses. For example, Chen, Yu, Zhang, and Yu (2016) analyzed the reflective writing in journal responses of pre-service teachers. This analysis allowed them to not only see how much of the course content the students incorporated into their reflections but also other themes that mattered to the students. Southavilay, Yacef, Reimann, and Calvo (2013) used topic modeling to create topic evolution charts to better understand how the topics in collaborative student writing changed over time when students used Google Docs to write together. Kakkonen, Myller, and Sutinen (2006) and Rahimi and Litman (2016) describe employing LDA to assist with automatic scoring of essays. Because of the high enrollment of Massive Open Online Courses (MOOCs), topic modeling is useful for quickly analyzing course forums, a task that would be nearly impossible to do manually. Several articles describe the application of topic modeling in these environments to assist with assessment of content in course forums (Ramesh, Goldwasser, Huang, Daumé III, & Getoor, 2014; Reich, Tingley, Leder-Luis, Roberts, & Stewart, 2014; Vytasek, Wise, & Woloshen, 2017).

Topic modeling is widely used to analyze texts in a variety of other fields, including political science (Grimmer, 2010), robotics (Girdhar, Giguère, & Dudek, 2014), and biology (Zheng, McLean Jr., & Lu, 2006), but has been used narrowly in the study of libraries or library services. For example, topic modeling has been used to enhance findability and access to digital library collections (Cain, 2016; Hagedorn, Kargela, Noh, & Newman, 2011; Newman, Hagedorn, & Chemudugunta, 2007; Mimno & McCallum, 2007).

Aims

This paper describes an exploratory study designed to determine if students demonstrate an understanding of the themes and concepts outlined in the Framework based on the assignment described above. Additionally, analyzing the assignment presents an opportunity to identify and address weaknesses in the assignment in future iterations of the course, or in the design of similar assignments at other universities.

A key contribution of this paper is a demonstration of how topic modeling can be used by librarians to analyze student work as it relates to information literacy. Past research describes how librarians are using the Framework to facilitate student understanding of information literacy concepts. However, there is little scholarship discussing student engagement with the Framework itself or the direct impact it has on student learning. This study provides an opportunity to explore on a large scale how students engage with concepts from the Framework.

While this study is similar in many ways to the work done by Scott (2017a, 2017b), mentioned earlier, there are a few key differences. First, the present study assesses just over 1,900 student responses (the whole first-year cohort) instead of a single class. Second, the students in this study engaged with the concepts of the Framework for one assignment in their first-year experience course rather than over the course of a whole semester. In other words, the focus is on a larger group of students at one point in time rather than a small group over the course of a semester. These two distinct approaches complement one another to improve our understanding of whether engaging with the Framework improves students' understanding of information literacy concepts.

Methods

As a part of the assignment, the students read the article "'Indians': Textualism, Morality, and the Problem of History" by Jane Tompkins. In this article, the author describes her experience attempting to reconcile the descriptions of events related to interactions between Native Americans and early colonists in books and articles written by historians. After failing to identify a definitive truth in these conflicting secondary sources, the author turned to primary sources, hoping to find unbiased truth in these sources. After reading several primary source documents representing different perspectives of individuals, the author realized the challenge of identifying a singular, unbiased truth about these events. After reading this article and the abridged version of the Framework, the students were asked to write a response selecting one frame and discussing how its themes are reflected in the Tompkins article.

Students entered their responses to the information literacy assignment into the campus learning management system. After receiving Institutional Review Board (IRB) approval for access, the responses for each section of the course were retrieved. All identifying information was removed from the files, and a unique, anonymous identification number was assigned to each response. While going through the responses to de-identify them, duplicate responses and responses to different assignments that had been erroneously uploaded to that week's assignment were removed. Additionally, several students did not complete the information literacy assignment, resulting in blank rows in the file. These lines were also removed. After all blank, duplicate, or erroneous responses were removed, 1,914 responses remained.

The assignment contained two questions to which the students responded (see the assignment prompt in Appendix A). When students entered their responses into the learning management system, these answers were not separated into different fields, meaning each student's response contained the answers to both questions. However, the responses to the first question, which asked about the Framework, were the focus of this study. Student answers to the second question were separated from answers to the first question, isolating the text to be analyzed. The resulting files were ingested and merged in the R statistical environment (R Core Team, 2018).

Once loaded into R, further steps were taken to prepare the data for analysis, including removing English language stop words, punctuation, separators, and special characters, and rendering all words in lowercase. Additionally, a lemmatization operator was run on the corpus. This process groups words based on a common root word, regardless of different inflectional endings. For example, the word "constructed" becomes "construct" in the corpus after lemmatization. Similarly, "is," "are," and other forms of this verb become "be." Lemmatization removes "semantic duplicates," allowing for better interpretation of the topic model results (Risueño, 2017).

Next, the sparsity threshold for the analysis was set. This number determines the threshold for removing uncommon terms from the corpus (Benoit, 2018). Words with values that are larger than the threshold are removed from the corpus. This retains words that allow for meaningful interpretation of students' responses while removing words that are too specific to add substantive value. Please see Appendix C for more details on setting the sparsity threshold.

Another necessary choice the researcher must make as an initial step is to select the number of topics to return in the model. Because researchers can take different approaches in determining the optimal number of topics to request from the model, the R package called ldatuning was used as a guide in this decision (Murzintcev, 2016). After analyzing the output of the ldatuning package, it was determined that 10 topics is the optimal number for this corpus.

Thus, the model identifies how words across the corpus cluster into 10 topics. Please see Appendix C for detailed information on the process taken to determine the optimal number of topics.

Results

After these preliminary steps, the corpus was analyzed using LDA. The top five words, or terms, associated with each of the 10 topics are found in Table 1.

Table 1
Terms in Each Topic

Topic Label	Terms
Topic 1	tompkins, indian, account, reflect, academic
Topic 2	idea, people, work, think, important
Topic 3	scholarly, process, knowledge, research, purpose
Topic 4	research, understand, topic, learn, good
Topic 5	first, authority, bias, construct, author
Topic 6	source, find, look, tompkins, subject
Topic 7	information, value, give, important, take
Topic 8	different, perspective, view, point, fact
Topic 9	american, tompkins, native, history, write
Topic 10	research, question, new, process, academic

See Appendix B for graphs of the probability of each term appearing in each topic.

Discussion

Analyzing the Topics

Each of the topics contains a set of terms that are frequently clustered near each other throughout the student responses. The combinations of terms in each topic provide insight into which frames students wrote about and how students described various frames. This is the basis of the analysis of student understanding of the Framework in this study.

The ordering of the topics listed in the table is random (e.g., Topic 1 is not necessarily more important than Topic 10). The terms listed within each topic appear in the order of greatest to least probability of each term being generated from that topic. In the analysis that follows, the topics that appear to be most closely related to the Framework are discussed first. Analysis includes a discussion of the terms in the topic, along with an example student response showing some of the terms in context.

The terms in Topic 5 ("first," "authority," "bias," "construct," and "author") are terms that are related to the description of the frame "Authority is Constructed and Contextual." The presence of this frame as one of the topics means that many students wrote about it in their responses. The article the students read touches on several themes described in this frame, including the importance of considering the authority of various sources and seeking different perspectives in the research process. The presence of this frame as a topic suggests that the students are able to successfully recognize and identify its key concepts reflected in a document other than the Framework. For example, one student wrote in their response, "One aspect that stands out in Tompkins' article is her exploration of multiple sources and, especially, the bias found in each. This connects to the First Frame [*sic*] which describes that

authority is both constructed and contextual." In this example response, the student interprets the frame "Authority is Constructed and Contextual" to be about examining multiple sources and their potential bias. Because the term "bias" appears near the terms "authority" and "construct" in the corpus (and are thus found together in Topic 5), it is apparent that many students referred to bias when discussing this frame. Student understanding of the social construction of authority, and thus the need to assess a source's credibility, can be inferred from their use of the term "bias" in this context.

The terms in Topic 8 are also related to the frame "Authority is Constructed and Contextual." The description of this frame in the student version of the Framework contains the sentence "Keep an open mind to varied and even conflicting perspectives." Topic 8 contains the terms "different," "perspective," "view," and "point," indicating that many students wrote about seeking different perspectives or points of view during the research process. Given the content of Tompkins' article, it comes as no surprise that many students wrote about this practice. The following student response provides an example:

> The first frame in "The Framework for Information Literacy for Higher Education" [Authority is Constructed and Contextual] parallels with Tompkins' account of her research project when Tompkins initially mentions the conflicting arguments of multiple sources. She discovers that there is controversy because the points of views of the observer differ…For example, Tompkins notes that Miller describes America as "vacant". There were Indians on the land when they arrived, and Miller would be mistaken if he truly believed there was no one there; however, Tompkins explains the reasoning behind Miller's word choice. Miller simply did not deem the Indians as important enough to be noted. . .. Thus, when constructing research, multiple

perspectives must be accounted for, not just the observer's.

The following excerpt from a student response also reflects how Topic 5 and Topic 8 are related:

> I think the first frame that states that "Authority is Constructed and Contextual" relates well to Tompkins essay. She writes about how the accounts of relations between Indians and early immigrants in New England are influenced by the person writing them and the period of time and social construct the author comes from. Also, many of these perspectives conflict with each other and some are even irreconcilable.

In both of these example responses, the students are able to identify how an individual's account of an event is influenced by their place and time in society. Additionally, they both reference the importance of seeking out multiple perspectives rather than relying on only one interpretation of an event. These responses exemplify how students connected the concepts in the frame "Authority is Constructed and Contextual" with the importance of recognizing bias and seeking out multiple perspectives in order to create as complete of an understanding around a topic as possible.

Interestingly, other students discussed the importance of seeking multiple perspectives on a topic or issue in their responses when discussing other frames. For example, one student wrote about the importance of reviewing multiple sources while discussing how the frame "Research as Inquiry" is reflected in the Tompkins article.

> Jane Tompkins' account of her academic research process reflects the fourth frame, "Research is Inquiry." Her first exposure to works of historians and secondary sources caused her to question the contradicting accounts with points of view so opposing that she was unable to reconcile them in any

way. She then mentions her turning towards primary sources where she found the same problem of conflicting viewpoints. These experiences caused her to view research as an iterative and collaborative exploration rather than a means to find a satisfying and definitive end result…Her understanding that "research is inquiry" helped her to not only accept unanswered questions, but also to formulate new questions in response to information she acquires.

As in the previous example response, this student picked up on the importance of weighing different perspectives in the research process. While these student responses discuss a similar theme (the importance of seeking out multiple perspectives on a topic) using the same terms, they are doing so while referring to different frames. This reflects the interrelatedness of the frames. As the Framework states, it is a "cluster of interconnected core concepts" (Association of College & Research Libraries, 2015, "Introduction"). These responses highlight how students are discussing similar themes using some of the same keywords but referring to different frames.

In the modified version of the Framework provided to the students, the frame "Information Creation as a Process" was reworded as "Knowledge is Both a Process and a Product." Two of the terms in Topic 3 ("knowledge" and "process") relate to this frame. Some language in the descriptive text of this frame was also modified. Therefore a few of the other terms that appear in Topic 3, such as "research," "scholarly," and "learn" are words that appear in the description of the student version of the frame but are not part of the wording of this frame in the original Framework document. While these keywords do not necessarily reflect main points of the original wording of the frame "Information Creation as a Process," the students were able to successfully respond to the version of the frame that they read. For example, one student wrote,

One of the frames from the framework was the "knowledge as both a process and a product". This frame reflects the idea that in pursuit of research, the process of trying to find answers will create more questions. This is very evident in Tompkins' account of her own academic research process because Tompkins had dwelled into a world that she had thought she had known a lot about, but instead opened a whole new world as she gained so much new information in her research.

The theme of this example student response better matches the general theme of the frame "Research as Inquiry" in the original Framework. In analyzing student responses, it became apparent that there were many similarities between the descriptions of the frames "Knowledge is Both a Process and a Product" and "Research as Inquiry" in the adapted student version of the Framework. Based on this analysis, in future iterations of this assignment, it is recommended that the adapted version of the frame "Information Creation as a Process" be reviewed and updated to ensure fidelity between the original wording and the adapted version.

The terms in Topic 7 ("information," "value," "give," "important," "take") suggest many students wrote about the frame "Information has Value" as it relates to Tompkins' article. The following student response is an example of the application of this frame to the reading.

One frame discussed within the reading, "Frameworks [*sic*] for Information for Literacy for Higher Education," focuses on the idea that information has value. In other words, information can be used in many different avenues to raise voices and induce change. This skill can be done by not only respecting the original merit of other's information, but also by giving yourself credit for being a contributor to the outflow of information, not just a "bystander." Within Tompkins' account of her academic

research progress, she discusses how a sentence in one of Perry Miller's writings stopped her dead. This sentence discussed the vacancy of the American wilderness prior to European civilization…I think Tompkins' shock from reading this sentence is a good example of the framework on the value of information because it shows how even though we don't agree with someone, we should still acknowledge the information of others and learn from its value.

This example response refers to two different aspects of the value of information described in the Framework. First, the student acknowledges that information is valuable in that it can be used to bring about change and to "raise voices." Additionally, the student refers to the value of citation and giving credit to the original work used in research.

Topic 1, Topic 4, and Topic 9 all contain terms related to the content of the Tompkins article but not necessarily to the Framework. Because the students were asked to describe connections between one frame and the content of the article, it is no surprise that topics would surface containing terms related to the description of Tompkins' research experience outlined in the assigned article. Many of the terms in Topic 10 are terms that also appear in the original assignment prompt. Therefore, this topic likely arises from students restating the question in their responses, thus those terms would be clustered together frequently. Finally, Topic 2 and Topic 6 do not collectively contain any keywords that are thematically important to the corpus. While the goal is to minimize the number of topics with little meaning, it is not uncommon for a few to remain.
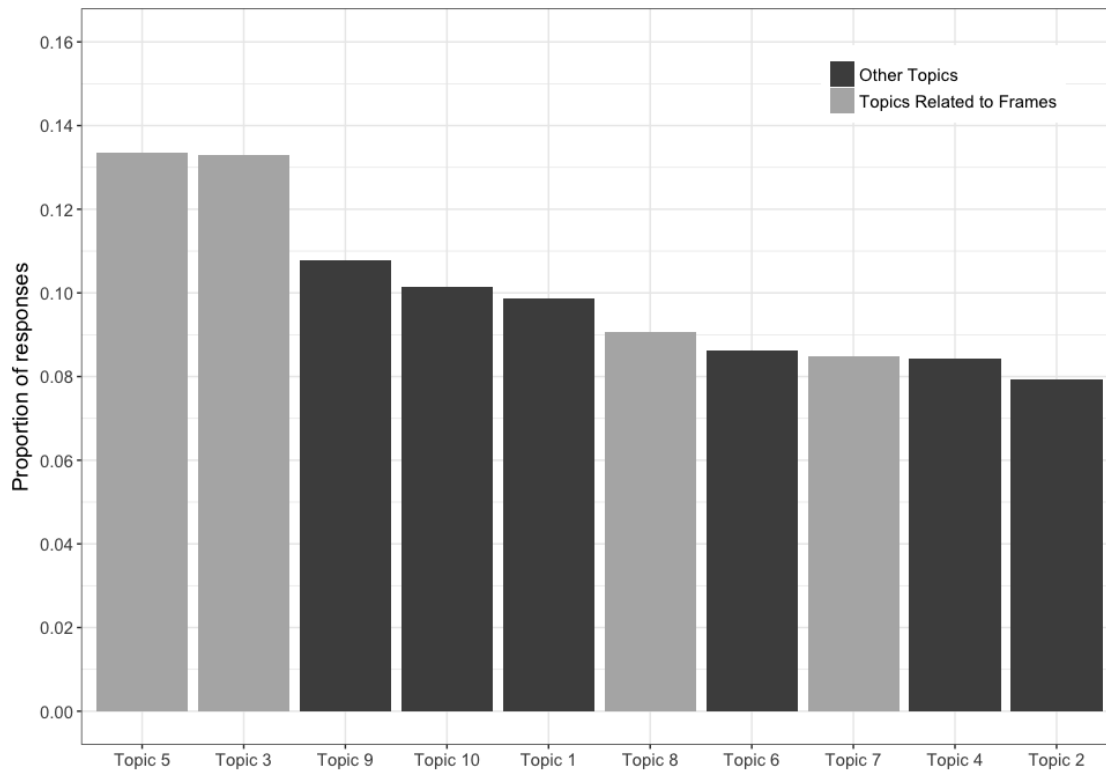

Figure 1
Proportion of responses discussing each topic

Topic Popularity

The analysis presented to this point has been focused on identifying and understanding the topics themselves. However, the LDA model also assigns each document, or each student response in the case of the present study, with topic probabilities, which are estimates of the "proportion of words from that document that are generated from that topic" (Silge & Robinson, 2017, p. 95). The higher a given topic probability, the more likely a document discusses that topic. These probabilities were used to assess the relative popularity of each topic among students' responses. For each response, the topic(s) associated with the highest topic probability were identified. Then, for each of the 10 topics, the proportion of responses in the corpus for which that topic produced the highest probability was computed. Figure 1 presents these proportions on the y-axis for each of the 10 topics on the x-axis.

This graph shows that, on average, Topic 5 and Topic 3 were the most popular among student responses. Topic 5 and Topic 3 both contain terms related to frames in the Framework. Further, these topics contain terms that are connected to the frames "Authority is Constructed and Contextual" and "Knowledge is Both a Process and a Product" ("Information Creation as a Process" in the original Framework), whose themes are present in the assigned reading by Tompkins. The intent of the assignment was for students to engage with the Framework and describe how it applies to the article by Tompkins. This analysis suggests that, on average, the students accomplished this objective.

Limitations

This exploratory analysis has a few limitations. The assignment analyzed in the present study was administered at one university. Therefore, the results may not be generalizable. Additionally, this analysis looked at one assignment involving direct engagement with

the Framework. It does not assess the transfer of knowledge of the skills and concepts described in the Framework to other situations. This analysis also captures only one point in time. As a result, it does not provide information about any potential growth in student understanding or skills related to information literacy. In the adapted version of the Framework that the students read, the frame "Information Creation as a Process" was reworded as "Knowledge is Both a Process and a Product." When this frame was adapted for the assignment, the way in which the language in the descriptive text was modified made it similar to the descriptive text of another frame, "Research as Inquiry." The similarity between the two frames in the student version may have had an effect on student understanding of the frames and the subsequent analysis. As noted previously, there were some concerns about the Framework at the time it was introduced, and students are not the intended audience of the document. Thus, this method is not the only way to assess student understanding of information literacy concepts.

Conclusion

In this exploratory study, topic modeling was used to analyze the text of first-year undergraduates' responses to an assignment on the Framework for Information Literacy for Higher Education. Topic modeling, as it was employed here, is an unsupervised, unstructured method which imposes few assumptions about what the researcher is going to find ahead of time. Three frames are easily recognizable in the topics. Based on the topics generated from the topic model and examining sample student responses, the analysis suggests that these first-year students made meaningful connections between the concepts described in the Framework and the assigned reading. Much like the results found by Scott (2017b), students were not simply interacting with the language of the Framework in a perfunctory manner. Rather, as seen in example responses, students made meaningful connections between the assigned reading and the frame about which they chose

to write. Thus, there appears to be benefit in asking students to apply the concepts in the Framework to an additional reading. Overall, this research provides valuable information for evaluating the assignment's effectiveness locally, as well as general insight regarding student understanding of the Framework.

References

AlSumait, L., Barbará, D., Gentle, J., & Domeniconi, C. (2009). Topic significance ranking of LDA generative models. In W. Buntine, M. Grobelnik, D. Mladenić, & J. Shawe-Taylor (Eds.). *Machine learning and knowledge discovery in databases*. ECML PKDD 2009. Lecture Notes in Computer Science, vol. 5781. Berlin: Springer. https://doi.org/10.1007/978-3-642-04180-8_22

Arun, R., Suresh, V., Veni Madhavan, C. E., & Narasimha Murthy, M. N. (2010). On finding the natural number of topics with Latent Dirichlet Allocation: Some observations. In M. J. Zaki, J. X. Yu, B. Ravindran, & V. Pudi (Eds.) *Advances in knowledge discovery and data mining*. PAKDD 2010. Lecture Notes in Computer Science, vol. 6118. Berlin: Springer. https://doi.org/10.1007/978-3-642-13657-3_43

Association of College & Research Libraries. (2015). *Framework for information literacy for higher education*. Retrieved from http://www.ala.org/acrl/standards/ilframework

Bauder, J., & Rod, C. (2016). Crossing thresholds: Critical information literacy pedagogy and the ACRL framework. *College & Undergraduate Libraries, 23*(3), 252-264. https://doi.org/10.1080/10691316.2015.1025323

Beilin, I. (2015, Feb. 25). Beyond the threshold: Conformity, resistance, and the ACRL Information Literacy Framework for Higher Education. *In the Library with the Lead Pipe*. Retrieved from http://www.inthelibrarywiththeleadpipe.org/2015/beyond-the-threshold-conformity-resistance-and-the-aclr-information-literacy-framework-for-higher-education/

Benoit, K. (2018). quanteda: Quantitative analysis of textual data. R package version 0.99.22. Retrieved from http://quanteda.io

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research, 3*(4/5), 993-1022.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM, 55*(4), 77-84. https://doi.org/10.1145/2133806.2133826

Bravender, P., McClure, H. A., & Schaub, G. (2015). *Teaching information literacy threshold concepts: Lesson plans for librarians*. Chicago, IL: Association of College and Research Libraries.

Burgess, C. (2015). Teaching students, not standards: Threshold crossings for students and instructors alike. *Partnership: The Canadian Journal of Library and Information Practice and Research, 10*(1). https://doi.org/10.21083/partnership.v10i1.3440

Cain, J. O. (2016). Using topic modeling to enhance access to library digital collections. *Journal of Web Librarianship, 10*(3), 210-225. https://doi.org/10.1080/19322909.2016.1193455

Cao, J., Xia, T., Li, J., Zhang, Y., & Tang, S. (2009). A density-based method for adaptive LDA model selection. *Neurocomputing, 72*(7-9), 1775–1781. https://doi.org/10.1016/j.neucom.2008.06.011

Chen, Y., Yu, B., Zhang, X., & Yu, Y. (2016). Topic modeling for evaluating students' reflective writing: A case study of pre-service teachers' journals. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge* (pp. 1-5). New York, NY: ACM. https://doi.org/10.1145/2883851.2883951

Dempsey, M. E., Dalal, H., Dokus, L. R., Charles, L. H., & Scharf, D. (2015). Continuing the conversation: Questions about the Framework. *Communications in Information Literacy, 9*(2), 164-175. https://doi.org/10.15760/comminfolit.2015.9.2.193

Dempsey, P. R., & Jagman, H. (2016). "I felt like such a freshman": First-year students crossing the library threshold. *portal: Libraries and the Academy, 16*(1), 89-107. https://doi.org/10.1353/pla.2016.0011

Deveaud, R., SanJuan, E., & Bellot, P. (2014). Accurate and effective latent concept modeling for ad hoc information retrieval. *Document Numérique, 17*(1), 61–84. https://doi.org/10.3166/dn.17.1.61-84

Gibson, C., & Jacobson, T. E. (2014). Informing and extending the draft ACRL information literacy framework for higher education: An overview and avenues for research. *College & Research Libraries, 75*(3), 250-254. https://doi.org/10.5860/0750250

Girdhar, Y., Giguère, P., & Dudek, G. (2014). Autonomous adaptive exploration using realtime online spatiotemporal topic modeling. *The International Journal of Robotics Research, 33*(4), 645-657. https://doi.org/10.1177/0278364913507325

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences, 101*(suppl. 1), 5228-5235. https://doi.org/10.1073/pnas.0307752101

Grimmer, J. (2010). A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in Senate press releases. *Political Analysis, 18*(1), 1-35. https://doi.org/10.1093/pan/mpp034

Hagedorn, K., Kargela, M., Noh, Y., & Newman, D. (2011). A new way to find: Testing the use of clustering topics in digital libraries. *D-Lib Magazine, 17*(9/10). https://doi.org/10.1045/september2011-hagedorn

Holliday, W. (2017). Frame works: Using metaphor in theory and practice in information literacy. *Communications in Information Literacy, 11*(1), 4-20. https://doi.org/10.15760/comminfolit.2017.11.1.44

Jacobson, T. E., & Gibson, C. (2015). First thoughts on implementing the Framework for Information Literacy. *Communications in Information Literacy, 9*(2), 102-110. https://doi.org/10.15760/comminfolit.2015.9.2.187

Julien, H., Gross, M., & Latham, D. (2018). Survey of information literacy instructional practices in U.S. academic libraries. *College & Research Libraries, 79*(2), 179. https://doi.org/10.5860/crl.79.2.179

Kakkonen, T., Myller, N., & Sutinen, E. (2006). Applying Latent Dirichlet Allocation to

automatic essay grading. In T. Salakoski, F. Ginter, S. Pyysalo, & T. Pahikkala (Eds.) *Advances in natural language processing*. Lecture Notes in Computer Science, vol. 4139. Berlin: Springer. https://doi.org/10.1007/11816508_13

Kuglitsch, R. Z. (2015). Teaching for transfer: Reconciling the Framework with disciplinary information literacy. *portal: Libraries and the Academy, 15*(3), 457-470. https://doi.org/10.1353/pla.2015.0040

Lancaster, A., Callender, D., & Heinz, L. (2016). Bridging the gap: New thresholds and opportunities for collaboration. In R. McClure (Ed.), *Rewired: Research-Writing partnerships within the frameworks* (pp. 85-101). Chicago: Association of College and Research Libraries.

Mays, D. A. (2016). Using ACRL's framework to support the evolving needs of today's college students. *College & Undergraduate Libraries, 23*(4), 353-362. https://doi.org/10.1080/10691316.2015.10 68720

McClure, R. (Ed.). (2016). *Rewired: Research-writing partnerships within the frameworks*. Chicago, IL: Association of College and Research Libraries.

Mimno, D., & McCallum, A. (2007, June). Organizing the OCA: learning faceted subjects from a library of digital books. In *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 376-385). New York, NY: ACM. https://doi.org/10.1145/1255175.1255249

Murzintcev, N. (2016). ldatuning: Tuning of the Latent Dirichlet Allocation models parameters. R package version 0.2.0. https://CRAN.R-project.org/package=ldatuning

Newman, D., Hagedorn, K., Chemudugunta, C., & Smyth, P. (2007, June). Subject metadata enrichment using statistical topic models. In *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 366-375). New York, NY: ACM. https://doi.org/10.1145/1255175.1255248

Oakleaf, M. (2014). A roadmap for assessing student learning using the new Framework for Information Literacy for Higher Education. *The Journal of Academic Librarianship, 40*(5), 510-514. https://doi.org/10.1016/j.acalib.2014.08.0 01

Pagowsky, N. (2015). A pedagogy of inquiry. *Communications in Information Literacy, 9*(2), 136-144. https://doi.org/10.15760/comminfolit.201 5.9.2.190

Rahimi, Z., & Litman, D. (2016). Automatically extracting topical components for a response-to-text writing assessment. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 277–282). Stroudsburg, PA: Association for Computational Linguistics. https://doi.org/10.18653/v1/w16-0532

Ramesh, A., Goldwasser, D., Huang, B., Daumé III, H., & Getoor, L. (2014). Understanding MOOC discussion forums using seeded LDA. In *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications* (pp 28-33). Stroudsburg, PA: Association for Computational Linguistics. https://doi.org/10.3115/v1/w14-1804

Reich, J., Tingley, D. H., Leder-Luis, J., Roberts, M., & Stewart, B. (2014). Computer-assisted reading and discovery for student generated text in massive open

online courses. *HarvardX Working Paper Series Number 6*. https://doi.org/10.2139/ssrn.2499725

R Core Team (2018). R: A Language and Environment for Statistical Computing. Vienna, Austria: The R Project for Statistical Computing. https://www.r-project.org/.

Risueño, T. (2017, June 15). Lemmatization to enhance topic modeling results. In *Bitext*. Retrieved from https://blog.bitext.com/lemmatization-to-enhance-topic-modeling-results

Scott, R. E. (2017). Part 1. If we frame it, they will respond: Undergraduate student responses to the Framework for Information Literacy for Higher Education. *The Reference Librarian, 58*(1), 1-18. https://doi.org/10.1080/02763877.2016.1196470

Scott, R. E. (2017). Part 2. If we frame it, they will respond: Student responses to the Framework for Information Literacy for Higher Education. *The Reference Librarian, 58*(1), 19-32. https://doi.org/10.1080/02763877.2016.1196471

Seeber, K. P. (2015). This is really happening: Criticality and discussions of context in ACRL's Framework for Information Literacy. *Communications in Information Literacy, 9*(2), 157-163. https://doi.org/10.15760/comminfolit.2015.9.2.192

Silge, J., & Robinson, D. (2017). *Text mining with R: A tidy approach*. Sebastopol, CA: O'Reilly.

Southavilay, V., Yacef, K., Reimann, P., & Calvo, R. A. (2013). Analysis of collaborative writing processes using revision maps and probabilistic topic models. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge* (pp. 38-47). New York, NY: ACM. https://doi.org/10.1145/2460296.2460307

Swanson, T. (2017). Sharing the ACRL Framework with faculty: Opening campus conversations. *College & Research Libraries News, 78*(1), 12-48. https://doi.org/10.5860/crln.78.1.9600

Ted Talk: Pariser, E. (2011, March). *Eli Pariser: Beware online "filter bubbles"* [Video file]. Retrieved from https://www.ted.com/talks/eli_pariser_beware_online_filter_bubbles

Tompkins, J. (1986). "Indians": Textualism, morality, and the problem of history. *Critical Inquiry, 13*(1), 101-119. https://doi.org/10.1086/448376

Vytasek, J. M., Wise, A. F., & Woloshen, S. (2017). Topic models to support instructors in MOOC forums. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference* (pp. 610-611). New York, NY: ACM. https://doi.org/10.1145/3027385.3029486

Wilkinson, L. (2014, June 19). The problem with threshold concepts [Blog post]. Retrieved from https://senseandreference.wordpress.com/2014/06/19/the-problem-with-threshold-concepts/

Witek, D. (2016). Becoming gardeners: Seeding local curricula with the ACRL Framework for Information Literacy. *College & Research Libraries News, 77*(10), 504-508. https://doi.org/10.5860/crln.77.10.9572

Zheng, B., McLean Jr., D. C., & Lu, X. (2006). Identifying biological concepts from a protein-related corpus with a probabilistic topic model. *BMC Bioinformatics, 7*(58). https://doi.org/10.1186/1471-2105-7-58
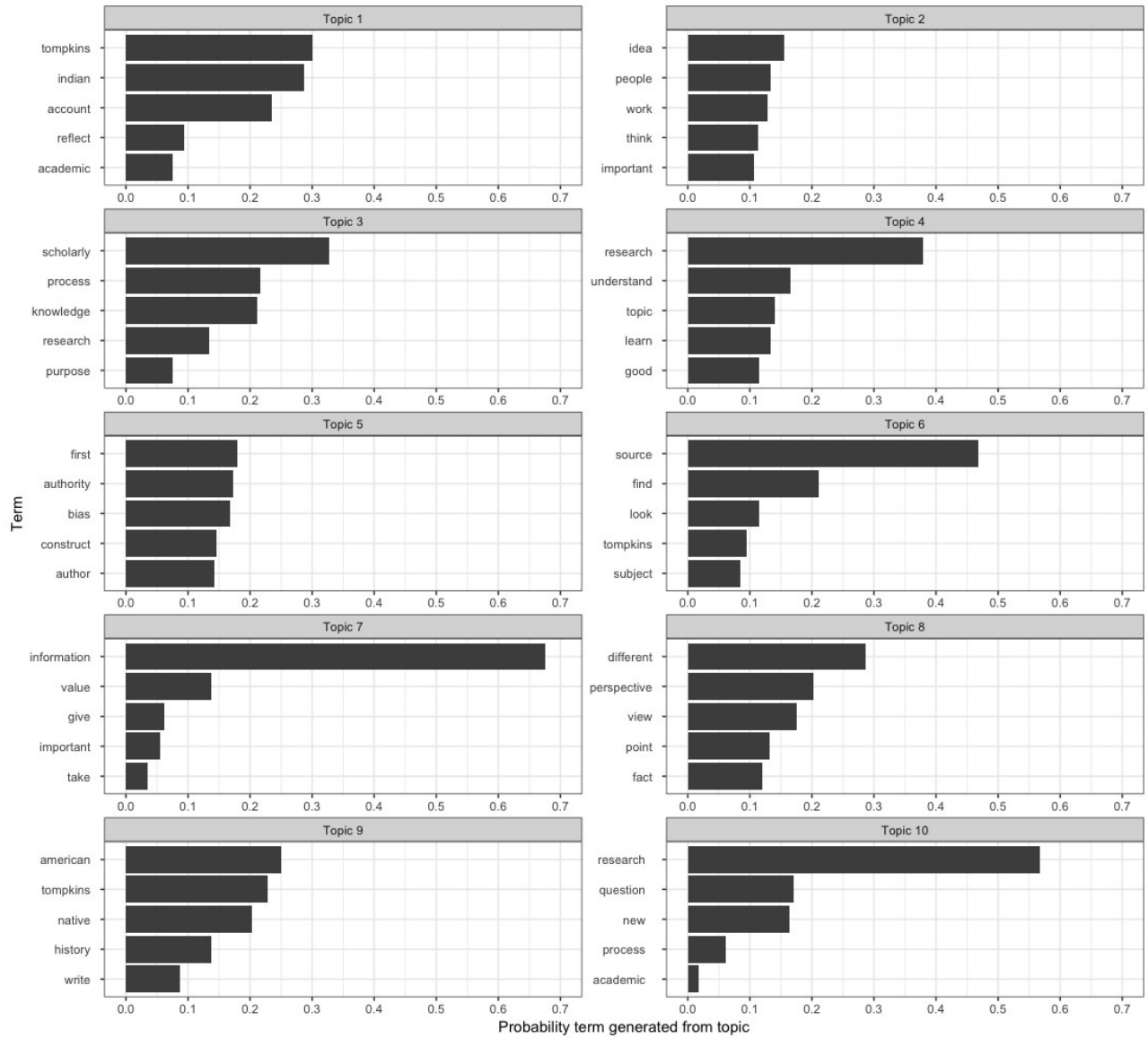
Appendix A
Information Literacy Assignment

1. Watch: Beware Online "Filter" Bubbles, TED Talk, by Eli Pariser

2. Read: "Indians": Textualism, Morality, and the Problem of History, Article, by Jane Tompkins

3. Read: The Framework for Information Literacy for Higher Education, Document, Literacy Framework (Association for College and Research Libraries)

In your reading for this week, we ask you to reflect carefully on the purpose, value, and process of scholarly research. First, choose *at least one* of the six frames from "The Framework for Information Literacy for Higher Education" and discuss how that frame is reflected in Tompkins' account of her own academic research process. You might, for example, describe how Tompkins' back-and-forth journey between primary, secondary, and tertiary sources reflects the truth of the second Frame ("Knowledge is Both a Process and a Product") in which knowledge must be understood as "a process of discovery rather than mere reaffirmation of prior held beliefs." Second, consider your own nascent student career here at Notre Dame. What controversial issue might you choose to explore? What might you expect to discover? How might you negotiate any contradictions you find?

Author note: When the student version of the Framework was adapted from the original, several modifications were made to the frame "Information Creation as a Process." First, the title of the frame was changed to "Knowledge is Both a Process and a Product." Additionally, the modified description of this frame contains many similarities to the description of the frame "Research as Inquiry." The titles of the other frames remained the same as in the original. Similarly, while the descriptions of the other frames were shortened, the language was not changed significantly.

Appendix B
Probability of Each Term Belonging to Topics

Appendix C
Methodological Details

Sparsity Threshold

The sparsity of a particular term is defined as $1-\frac{f}{c}$ where f is the number of times the word appears in the corpus and c is the number of documents in the corpus (Benoit, 2018). For instance, the sparsity of a term that appears 15 times in a corpus of 100 documents is $1-\frac{15}{100} = 0.85$.

Selecting the Number of Topics

The R package called ldatuning was used as a guide in deciding how many topics to have the model return (Murzintcev, 2016). This package applies four different methods for determining an appropriate number of topics for LDA in a given corpus and presents the results in one graph for comparison (see Figure C1). The four methods are described in Griffiths and Steyvers (2004), Cao, Xia, Li, Zhang, and Tang (2009), Arun, Suresh, Veni Madhavan, and Narasimha Murthy (2010), and Deveaud, SanJuan, and Bellot (2014). In the models developed by Cao et al. and Arun et al., the number of topics closest to zero on the y-axis are considered optimal for a given corpus, whereas in the models developed by Griffiths and Steyvers and Deveaud et al., the number of topics that are closest to 1.00 are considered optimal.
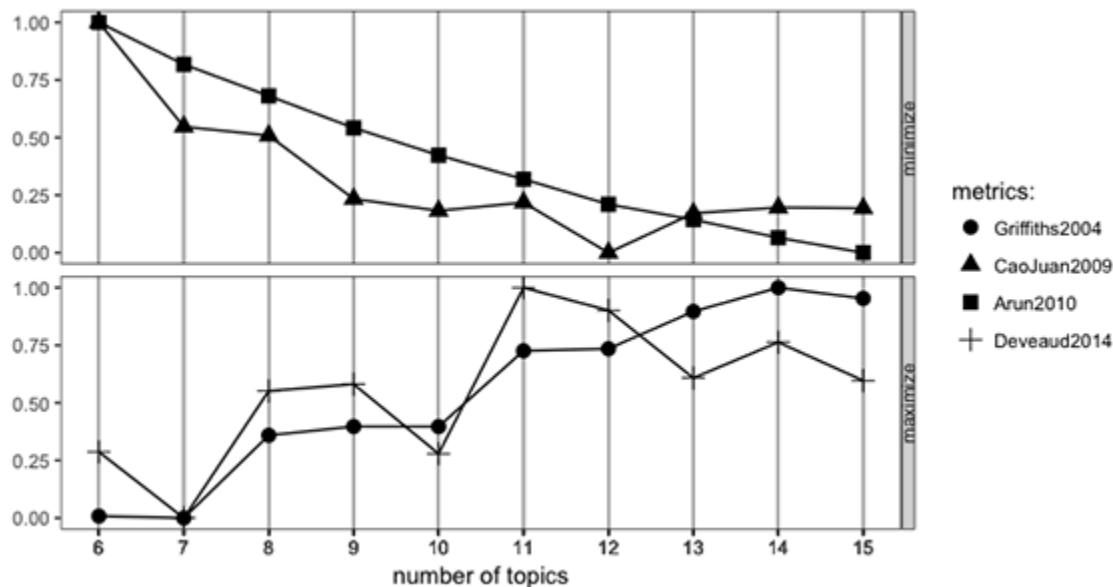

Figure C1
Topic selection measures.

The point on the graph in Figure C1 where the four measures show the most convergence indicates 12 is the optimal number of topics. However, this choice produced several substantive topics and a few topics that contained common words across documents and seemed to be meaningless. It is not unusual to have a few topics that are thematically unimportant; however, the goal is to find as many meaningful topics as possible (AlSumait, Barbará, Gentle, & Domeniconi, 2009). Running the model with 10 topics produced more meaningful topics and fewer that were meaningless, indicating that 10 is the optimal number of topics for this corpus.