Tak Yeon Lee, Alison Smith, Kevin Seppi, Niklas Elmqvist, **Jordan Boyd-Graber**, and Leah Findlater. **The Human Touch: How Non-expert Users Perceive, Interpret, and Fix Topic Models**. *International Journal of Human-Computer Studies*, 2017, 27 pages.

```
    @article{Lee:Smith:Seppi:Elmqvist:Boyd-Graber:Findlater-2017,
Author = {Tak Yeon Lee and Alison Smith and Kevin Seppi and Niklas Elmqvist and Jordan Boyd-Graber and Lea
Journal = {International Journal of Human-Computer Studies},
Year = {2017},
Title = {The Human Touch: How Non-expert Users Perceive, Interpret, and Fix Topic Models},
Url = {http://cs.umd.edu/~jbg//docs/2017_ijhcs_human_touch.pdf},
}
```

Links:

- Journal [http://www.sciencedirect.com/science/article/pii/S1071581917300472]

The Human Touch: How Non-expert Users Perceive, Interpret, and Fix Topic Models

Tak Yeon Lee
UNIVERSITY OF MARYLAND, COLLEGE PARK, MD, USA
TYLEE@UMD.EDU

Alison Smith
UNIVERSITY OF MARYLAND, COLLEGE PARK, MD, USA
AMSMIT@CS.UMD.EDU

Kevin Seppi
BRIGHAM YOUNG UNIVERSITY, PROVO, UT, USA
KSEPPI@BYU.EDU

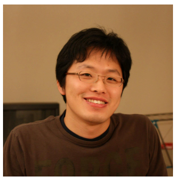
Niklas Elmqvist
UNIVERSITY OF MARYLAND, COLLEGE PARK, MD, USA
ELM@UMD.EDU

Jordan Boyd-Graber
UNIVERSITY OF COLORADO, BOULDER, BOULDER, CO, USA
JBG@BOYDGRABER.ORG

Leah Findlater
UNIVERSITY OF MARYLAND, COLLEGE PARK, MD, USA
LEAHKF@UMD.EDU

Corresponding Author
Tak Yeon Lee
UNIVERSITY OF MARYLAND, COLLEGE PARK, MD, USA
TYLEE@UMD.EDU

Vitae



Tak Yeon Lee

Tak Yeon Lee is a PhD student at the University of Maryland, College Park. He received his MSc in Design for Interaction from the Delft University of Technology, Netherlands. He holds a BSc in Industrial Design (major) and Computer Science (minor) from the Korea Advanced Institute of Science and Technology. His research focuses on mixed-initiative, symbiotic interaction between human users and intelligent systems.

Alison Smith

Alison Smith is a PhD student at the University of Maryland, College Park. She holds a BSc in Math and Computer Science from the College of William and Mary, and she received her MSc in Computer Science from University of Maryland, College Park. Ms. Smith is also the Lead Engineer of the Machine Learning Visualization Lab for Decisive Analytics Corporation where she designs user interfaces and visualizations for interacting with complex algorithms and their results. Her focus is on enhancing users' understanding and analysis of complex data without requiring expertise in data science or machine learning.

Kevin Seppi

Kevin Seppi is a professor at Brigham Young University where he works on probabilistic models in the context of human computer interaction. His work includes human robotic interaction, prototyping of intelligent devices, and natural language processing. He received his PhD in Operations Research from the University of Texas and holds Master's and Bachelor's degree in Computer Science from Santa Clara University and Brigham Young University respectively.

Niklas Elmqvist

Niklas Elmqvist received the PhD degree in 2006 from Chalmers University of Technology in Gothenburg, Sweden. He is an associate professor in the College of Information Studies at the University of Maryland, College Park, MD, USA. He was previously an assistant professor in the School of Electrical & Computer Engineering at the Purdue University in West Lafayette, IN. His research areas include data visualization, visual analytics, and human-computer interaction. He is a senior member of the IEEE, the IEEE Computer Society, and the ACM.

Jordan Boyd-Graber

Jordan Boyd-Graber is a seventh-year assistant professor in the Colorado Computer Science Department. Previously, He was a postdoc and then professor in Maryland's Computational Linguistics and Information Processing lab. He was a graduate student at Princeton with David Blei. His research focuses on making machine learning more useful, more interpretable, and able to learn and interact from humans. This helps users sift through decades of documents; discover when individuals lie, reframe, or change the topic in a conversation; or to compete against humans in games that are based in natural language.

Leah Findlater

Leah Findlater is an Assistant Professor in the College of Information Studies and the Institute for Advanced Computer Studies at the University of Maryland, College Park. Her research interests span mobile and wearable computing, accessibility, and interactive machine learning.

Abstract

Topic modeling is a common tool for understanding large bodies of text, but is typically provided as a “take it or leave it” proposition. Incorporating human knowledge in unsupervised learning is a promising approach to create high-quality topic models. Existing interactive systems and modeling algorithms support a wide range of refinement operations to express feedback. However, these systems’ interactions are primarily driven by algorithmic convenience, ignoring users who may lack expertise in topic modeling. To better understand how non-expert users understand, assess, and refine topics, we conducted two user studies—an in-person interview study and an online crowdsourced study. These studies demonstrate a disconnect between what non-expert users want and the complex, low-level operations that current interactive systems support. In particular, our findings include: (1) analysis of how non-expert users perceive topic models; (2) characterization of primary refinement operations expected by non-expert users and ordered by relative preference; (3) further evidence of the benefits of supporting users in directly refining a topic model; (4) design implications for future human-in-the-loop topic modeling interfaces.

Keywords: Topic modeling; User Study; Mixed-Initiative Interaction;

1. Introduction

Managing the collective knowledge captured in books, newspapers, webpages and other text sources has become less a problem of access and storage than one of search, comprehension, and discovery (Blei, 2012). Today’s users, even casual ones, need effective mechanisms to read, understand, and summarize these large text collections, from product reviews to news articles and government reports (Hotho et al., 2005).

Topic modeling is a popular technique to deal with this information overload by finding word co-occurrence patterns known as *topics* (Blei et al., 2003). Conceptually, a topic is a theme spanning multiple documents. While many other techniques exist to support users in working with large document collections—such as faceted browsing (English et al., 2002; Hearst, 2006), document clustering (Aggarwal and Zhai, 2012), and visualizations (Fortuna et al., 2005; Šilić and Bašić, 2010), topic modeling is an unsupervised method that offers a summary overview of the document collection and directed exploration via the discovered topics. A user can review the topics to get a sense of the corpus without reading all of the documents and can focus on a specific topic by exploring only the documents closely associated with it. For instance, a topic with the words “*beach, snow, vacation, miles, ski, downhill, air*” would likely be associated with documents about vacation getaways. Topic models are also used as the basis of a variety of analyses, such as differentiating language use (McFarland et al., 2013), analyzing topical changes over time (Malik et al., 2013), performing sentiment analysis (Titov and McDonald, 2008), and even recommending television shows (Hulu, 2011).

Despite this utility, topic modeling is not an out-of-the-box tool. To generate high-quality models, the user must specify parameters such as the number of topics, stop words, and hyper-parameters and may also need to do pre-processing such as lemmatizing and chunking. Not only are these methods indirect, the configuration will likely be inscrutable to users who are not topic-modeling experts. Moreover, the discovered topics do not always correlate with human judgments of topic quality and can appear “bad” from an end user’s perspective (Chang et al., 2009). Problems include topics with too many generic words (e.g*., “people, like, mr”*) (Boyd-Graber et al., 2014), topics with disparate or poorly connected words (Mimno et al., 2011), misaligned topics (Chuang et al., 2013a), irrelevant (Ramage et al., 2009) or missing associations between topics and documents (Daumé, 2009), and multiple nearly identical topics (Boyd-Graber et al., 2014). The presence of poor-quality topics has been cited as the primary obstacle to the acceptance of statistical topic models outside of the machine learning community (Mimno et al., 2011).

Addressing these issues of configuration complexity and topic model quality, researchers have recently proposed a variety of human-in-the-loop methods for users to directly manipulate and incrementally refine a model; this interactive approach is preferred by users over indirect configuration (Chuang et al., 2013b). Rather than reconfiguring and rebuilding a model when users are unsatisfied with its quality, they can improve it through operations such as adding or removing words in topics, splitting generic topics, or merging similar topics. Hu et al. (2014), for example, allow users to add, emphasize, and ignore words within topics, while Choo et al. (2013) allow users to adjust the weights of words within topics, merge and split topics, and create new topics. Additionally, iVisClustering (Lee et al., 2012) allows users to reassign documents to other topics. Human knowledge has also been incorporated into topic models in the form of semantic concepts (Chemudugunta et al., 2008), predefined topics (Zhai et al., 2009), or word pairs that should or should not belong together in a topic (Andrzejewski et al., 2009).

While these approaches are promising, a key limitation to the above work—both what makes for a "bad" topic model and human-in-the-loop methods—is that researchers have focused largely on algorithmic convenience rather than user needs, and—particularly—there is a dearth of studies with users who lack topic modeling experience. For example, the problems with topic model quality identified above come largely from self-reports of machine learning experts, with the exception of two domain experts in one paper (Mimno et al., 2011). Even the human-in-the-loop approaches offer few user studies. In a study with 20 users who had varied experience with topic models, Hu et al. (2014) showed that having access to their small set of refinement operations impacted how users searched for information in a corpus compared to using a traditional, static topic model. Another exception comes from Hoque and Carenini (2015), who found that users preferred a topic model interface that allows for splitting and merging of topics over a more traditional interface that does not allow for model refinement. While these studies begin to provide insight for a more user-centered approach to topic modeling, important questions remain unanswered: How do non-expert users interpret topic models and what problems do they perceive? Disregarding technical constraints, how would they *want* to fix or refine those problems? Based on these answers, what are the implications for making human-in-the-loop topic models more accessible to non-expert users?

We conducted two user studies with non-experts: a semi-structured interview study with 10 participants followed by an online crowdsourced experiment with 90 participants. The studies were exploratory and formative, meant to guide future algorithmic work on what refinement operations to support for human-in-the-loop topic modeling and how to prioritize those operations if need be. As such, we intentionally asked participants to interact with front-end user interfaces that *were not constrained* by current technical limitations of topic model implementations. For both studies, participants interacted with topic models built from a corpus of New York Times articles. In one case participants performed a theme summarization task that can act as a general first step for more complex tasks, such as for qualitative content analysis (Chuang et al., 2014), and in the second case participants performed the more constrained but common low-level task of refining individual topics. The interview study focused on exploring how non-experts *interpret*, *assess*, and want to *refine* topic models without being biased by technical constraints or suggested operations; users did not have preconceptions of how refinements might impact a topic model. For the online crowdsourced experiment, we then designed and implemented a user interface to support the six most popular refinement operations identified in the interviews: *Add words, Remove words, Merge words, Change word order*, *Remove documents*, and *Split topic*. We evaluated the frequency of use and subjective utility of the six operations by asking participants to refine individual topics. While the refinement operations were implemented only in the user interface (the backend model was not updated), this study allows us to compare the operations in terms of frequency of use and measure their perceived impact on topic quality; we also evaluate the quality of the refined topics with automatically computed topic coherence based on NPMI (Lau et al., 2014).

Combined, our studies demonstrate a disconnect between what non-expert users want (e.g., simply adding a word to a topic) and the low-level operations that human-in-the-loop topic modeling approaches currently support, (e.g., specifying a must-link constraint (Andrzejewski et al., 2009; Hu et al., 2014)). In addition to identifying a set of important refinement operations that should be included to support non-expert users, our findings highlight patterns in how non-experts interpret (and misinterpret) topics and apply refinement operations to individual topics. The online experiment also shows that, albeit without back-end updates, users believed their refinements significantly improved topic model quality; objective measures of topic coherence aligned with this perception. These findings should guide future algorithmic work on human-in-the-loop topic modeling.

This paper makes the following contributions: (1) showing how non-expert users perceive topic models, confirming several problems asserted by machine learning researchers; (2) enumerating non-expert users' desired topic-level refinement operations and the relative preference among these refinements, particularly with relevance to the tasks of summarizing themes in a corpus and refining individual topics; (3) providing further evidence of the benefits of allowing users to directly refine a topic model (confirming (Chuang et al., 2013a; Hoque and Carenini, 2015)), particularly in terms of the perception of topic quality; (4) providing additional implications for the design of future human-in-the-loop topic modeling interfaces, such as specific mixed-initiative support requested by users.

2. Related Work

We provide background on topic modeling as well as cover more work on user interaction with topic models to complement that already discussed in the Introduction (Section 1). Finally, we briefly touch on mixed-initiative interaction.

2.1 General Topic Modeling Background

Topic modeling helps users explore text data by discovering topics in a corpus, where each topic is represented as a set of words. Latent Dirichlet Allocation (LDA) (Blei et al., 2003)—the prototypical statistical topic model—is an unsupervised algorithm for topic modeling. LDA models each document in the corpus as a distribution of topics and each topic as a distribution of words using a "bag of words" approach that ignores the order of words in documents. Topic models help users answer factual questions (Hu et al., 2014), find relevant documents (Aletras et al., 2015), and create summaries of news articles (Haghighi and Vanderwende, 2009). Kim et al. (2015), for example, compared LDA to other statistical techniques and showed through a user study that LDA was the best for clustering documents. These studies, however, do not provide a rich understanding of how users more generally perceive a topic model and assess its quality—one of the goals of our interview study.

Instead, the developers of topic models often focus on evaluations from statistics and machine learning, such as perplexity (Stevens et al., 2012) and held-out likelihood (Wallach et al., 2009). However, Chang et al. (2009) contend that these measures encourage complexity, which is often at odds with user needs. For example, they showed that better predictive perplexity can result in overly specific, hard to understand topics. More recent approaches that use word co-occurrence information, such as pointwise mutual information (PMI) and log conditional probability (LCP) to compute topic coherence, more broadly agree with a user assessment of the semantic coherence of a topic (Mimno et al., 2011; Newman et al., 2010; Wallach et al., 2009). These observed topic coherence measurements deem topics to be more coherent if they contain words that are more commonly found together than apart in a reference corpus. In particular, Newman et al. (2010) explored the space of different observed coherence measures and found that among a set of individual measures, a method based on PMI for computing word co-occurrence using a Wikipedia reference corpus resulted in the largest correlation to human topic coherence ratings. Lau et al. (2014) extended this work by applying normalized pointwise information (NPMI (Bouma, 2009)) to reduce the bias of PMI for words of lower frequency. As such, we use an observed topic coherence metric based on NPMI as an objective assessment of the quality of refined topics in our study. While manual validation by experts can be time consuming, a few studies have incorporated more constrained user feedback. Chuang et al. (2013a), for example, asked non-expert users to rate whether pairs of topics "match", "partially match", or "don't match". However, while human evaluations remain the gold standard, automatic evaluations remain necessary to compare hundreds or thousands of topic models efficiently and quickly.

2.2 Supporting User Interaction with Topic Models

How to support users in interacting with topic models is an open area of research. Visualizations of static topic models typically allow the user to browse a single model of one corpus or to compare multiple models. The former often present lists of words ordered and/or sized by their probability to the topic (Chaney and Blei, 2012). The goals of these interfaces include visualizing correlated metadata (Alexander et al., 2014; Choo et al., 2013; Gardner et al., 2010; Gretarsson et al., 2012; Lee et al., 2012) providing access to the underlying documents (Chaney and Blei, 2012), or diagnosing poor topics (Chuang et al., 2013a). Comparative topic visualizations, in contrast, show how topics change over time (Malik et al., 2013) or with different models (Alexander and Gleicher, 2015; Chuang et al., 2015; Crossno et al., 2011). While these approaches can show how a topic has changed (e.g., after tweaking parameters and re-building the model), they do not solicit user feedback to make those changes. Recent work has also compared different representations for individual topics, such as word lists and word clouds (Alexander and Gleicher, 2015; Smith et al., 2016).

In contrast to the numerous visualizations of static topic models, less research has focused on allowing the user to refine the topic model. Hoque and Carenini (2015) summarize topic refinement operations that have been proposed in the literature, including: splitting a topic, merging multiple topics, removing a topic, assigning a label to a topic, changing the word-to-topic weightings, and specifying that a set of words must or must not appear together in the same topic. Among the systems surveyed, UTOPIAN (Choo et al., 2013) allows users to adjust the weights of words within a topic, merge and split topics, and create new topics based on documents or keywords. iVisClustering (Lee et al., 2012) shows topic-document associations in a scatter plot and lets users manually create or remove topics, merge or split topics, and reassign documents to another topic. Neither UTOPIAN nor iVisClustering provide user evaluations, so it is unclear how easy the interfaces are to use or how users would apply the refinement operations. Finally, Bakharia et al. (2016) allow users to specify configuration rules intended to seed, merge, and split topics as input when building new topic models on the same corpus, although the approach does not allow for iterative refinement of a single model over time.

Limited research provides user evaluations of human-in-the-loop topic modeling. Hu et al.'s (2014) *Interactive Topic Modeling* allows users to add, emphasize, or ignore words within topics. Their user study shows that incorporating user feedback into the model increases the accuracy of classifiers that use the improved topics as features. However, they also found that users sometimes created inscrutable correlations, such as connecting unrelated words, and that even sensible feedback did not always lead to successful topic changes. Second, ConVisIT (Hoque and Carenini, 2015) allows users to split and merge topics. In a controlled study, users preferred having these operations available to them compared to exploring a static topic model; results also suggest that splitting topics is preferable and more frequently used than merging topics. Compared to these studies, our user studies are more open-ended; the interview study does not constrain users to a preset group of refinement operations, while the crowdsourced study compares a relatively large set of operations. Moreover, the ConVisIT study can be considered complementary to ours, since it focused on refinements at the level of the entire topic model, while much of our focus is on topic-specific refinements (e.g., adding or removing words).

2.3 Mixed-initiative Interaction

Human-in-the-loop topic modeling is an example of *mixed-initiative* interaction, where the system and user work collaboratively to meet the user's goals (Horvitz, 1999). While mixed-initiative interaction does not by definition need to be paired with machine learning, this is increasingly the case. With interactive machine learning, the system can improve the model by soliciting user input and drawing the user's attention to areas of particular need (Amershi et al., 2011; Fails and Olsen, 2003). It has been applied to problems ranging from segmenting images (Fails and Olsen, 2003) to personalizing electronic musical instruments (Fiebrink et al., 2009) to grouping contacts in online social networks (Amershi et al., 2012).

Closest to our work are approaches for clustering or otherwise making sense of text-based data (Basu et al., 2010; Drucker et al., 2011; Hu et al., 2012). Drucker et al., for example, compared a mixed-initiative document clustering system that adapts over time as the user clusters documents to fully manual and automated approaches. Other examples include Cohn et al.'s (2003) semi-supervised approach that allows users to iteratively provide feedback to a document clustering algorithm and *meta clustering* (Caruana et al., 2006), which allows users to select the clustering that best fits their needs from among several alternate clusterings. Stumpf et al. (2007) also conducted a user study to understand what input users want to provide to a classification algorithm, with results including proposals for new features, reweighting of existing features, and larger-scale algorithm changes; similar to their work, we use an open-ended study method to solicit user feedback without being constrained by what is currently technically feasible. Interactive machine learning for visual analytics is also relevant to human-in-the-loop topic modeling. Within Mühlbacher et al.'s (2014) framework of strategies for increasing user involvement in algorithms for visual analytics, our work falls into the category of increasing "result control", that is, allowing the user to impact the result of the algorithm.

Any mixed-initiative or "intelligent" system introduces design challenges such as how to effectively weigh the costs and benefits of an automated action in light of uncertainty (Horvitz, 1999), to balance the user's and system's needs (e.g., not treating the user solely as input to the model) (Amershi et al., 2011), to maintain a sense of user control (Höök, 2000), and to provide transparency into the system's decision-making (Höök, 2000). Amershi et al. (2014) provide a set of guidelines for interactive machine learning that are particularly relevant to topic modeling, such as the need for immediate feedback and support for assessment of model quality. While our studies offer specific implications for the design of human-in-the-loop topic modeling interfaces, it will also be important to consider these general best practices in future designs.

3. Interview Study: Non-Expert User Understanding and Subjective Response to Topic Models

To investigate how non-expert users interpret, assess, and want to refine topic models, our initial study asked ten users without prior knowledge of topic modeling to inspect a topic model. We showed participants a model with 20 topics inferred from a corpus of news articles and asked them to identify a set of themes represented by the articles. Within this context, participants interacted with the topic model and described the meaning of each topic, assessed its quality, and proposed potential improvements. Although topic model quality and how users want to refine topic models are likely context dependent, understanding a model and summarizing themes in a corpus are useful steps toward other high-level tasks such as content analysis (Chuang et al., 2014). After reviewing the entire model, participants evaluated the utility of previously proposed refinement operations identified in Section 2.2. These interviews helped to identify the types of refinement operations that non-expert users find intuitive as well as broader implications for how to better support these users in interacting with topic models. This study also determined a set of refinement operations to examine in the crowdsourced study discussed later.

3.1 Method

We used a semi-structured interview approach and a topic model generated from a corpus of news articles.

Table 1. The interview study used a topic model of 20 topics from New York Times articles. Each topic was presented as a list of the top 20 topic words. The topics are ordered from highest to lowest automatically computed observed topic coherence (NPMI) score, with higher scores corresponding to topics that should be more interpretable.

ID	Topic Words	NPMI
I1	game season team coach games players play year league football points giants bowl teams player field won back time	0.19
I2	food restaurant wine restaurants chicken bar chef good meat menu sauce fish dinner place cream room chocolate minutes dining	0.19
I3	music mr theater dance street band jazz songs ballet musical song york album play performance ms hall west rock	0.12
I4	dr health drug research medical patients women study people disease drugs care cancer hospital researchers science percent found brain	0.10
I5	mr police court case law judge justice lawyer yesterday ms federal lawyers department trial death investigation state man officials	0.10
I6	company web site internet online mr technology microsoft video service advertising companies computer software digital apple media year marketing	0.10
I7	iraq american mr government iraqi military war officials iran troops united security baghdad bush states forces al hussein president	0.09
I8	school children students ms york university college schools church high parents education year child student family brooklyn class women	0.09
I9	mr company percent year business market million billion companies chief sales stock executive investors price financial money years executives	0.08
I10	mr bush president house democrats senator republican senate democratic congress political white party Iraq campaign Clinton administration war committee	0.08
I11	art museum street mr jan show work artists arts artist works american gallery center century exhibition modern ms made	0.07
I12	winter water www snow miles beach island park day travel year air airport west south north club ski trees	0.06
I13	state mr city health york tax officials year million money public federal system government spritzer care pay insurance plan	0.04
I14	film book mr movie books life world story films ms man war characters review movies director author wrote love	0.04
I15	united energy oil china states world mr government europe nuclear power country chinese gas countries european international trade global	0.04
I16	city building street house mr number square million town york market year avenue room park half weeks buildings area	0.03
I17	paid notice january deaths family york wife died beloved president father years st husband mother ny john survived late	0.02
I18	show mr television car million series ford corrections article news year fox cars radio executive ms viewers toyota nbc	0.01
I19	people time make years long don work made good end part world jan change put editor recent making important	0.01
I20	mr ms back time home day people year years life family don man night told father didn mother left	0.00

3.1.1 Participants

Ten participants (4 male, 6 female) were recruited via a university campus email list. They were on average 33.6 years old (*SD* = 13.6, range 22–59) and all were fluent English speakers. Their occupations varied: full-time college students with various majors (business, psychology, real estate development, environmental science, linguistics, and virology), salesperson, accountant, program management specialist, and office manager. Nine of the 10 participants stated that they read news articles more than once a week. Participants received $15 US for their participation. They are referred to as P1 to P10 throughout the paper.

3.1.2 Dataset, Topic Model, and User Interface

To ensure that the topic model is accessible to most users, we used a corpus of 7,156 *New York Times* articles published in January, 2007 (Sandhaus, 2008). In a pilot test participants took around five minutes to review each topic and could review about 20 topics without fatigue. Thus, we generated 20 topics using Mallet (McCallum, 2002), a popular open-source topic modeling toolkit, with standard stop words and hyper-parameter settings.[1] The set of topics is shown in Table 1; we refer to these topics by their IDs (I1–I20). For each topic we also computed the observed topic coherence, as discussed in Section 2.1, which is a measure of topic coherence based on word co-occurrence information. We used an NPMI-based metric (Lau et al., 2014) with a reference corpus of 2.3 Million Wikipedia articles and a sliding window of 10 words.[2]

We designed and built a web application for browsing topics (Figure 1). The application contained two scrollable panels, where the left panel displayed a list of the 20 topics in random order,[3] and the right panel displayed the documents associated with the currently selected topic ordered by the probability of the topic, z, given the document, d, $p(z|d)$. Each topic was represented by a list of its top 20 words, where each word, w, was ordered and sized by its probability for the topic, z, or $p(w|z)$; this information was available from Mallet's output.

While there are more complex visualizations for topic models, most have not been thoroughly vetted by users (Section 1). Instead, we displayed topics as a simple list. This technique is commonly used (e.g., (Chaney and Blei, 2012; Eisenstein et al., 2012a; Gardner et al., 2010; Hu et al., 2014)), is applicable to a wide range of tasks, and is fast and easy to understand for non-expert users (Smith et al., 2016). To interact with the model, users could (1) scroll through the topic list, (2) click a topic to display its top 20 documents,[4] (3) scroll through the document list, (4) click a document title to toggle its full text, and (5) hover the mouse cursor over a topic word to highlight the word's context within the displayed documents and to highlight other topics that include the word. Ten lab colleagues provided feedback on an early version of the interface.

3.2 Procedure and Task

Each interview session took up to 90 minutes. Participants were first introduced to the interface and the task scenario: identifying the main themes in a corpus. Topic models are used for a wide range of real-world tasks that may result in different refinement operation needs. The theme identification task was chosen because it is relatively generic and is an important step in larger tasks such as qualitative content analysis (i.e., theme identification and coding) (Chuang et al., 2014). In the context of theme identification, participants also had to repeatedly assess and suggest refinements for individual topics or subsets of topics—understanding and interacting with individual topics is a common component of many real-world topic modeling tasks (Alexander and Gleicher, 2015; Boyd-Graber et al., 2014). To increase approachability for our non-expert participants, the scenario used the terms "themes" and "news articles" instead of "topics" and "documents". The scenario description was general, so that findings would apply as widely as possible, and was worded so as not to draw attention to specific types of model flaws:

"Imagine that you were asked to come up with the main themes from the New York Times articles of January 2007. To help you get started, a computer program automatically discovered themes from the articles. [...] The set of themes should represent an overview of the articles, and each theme should be clear and understandable. Also, each theme should be associated with representative articles. However, you may notice that there is still room for improvement".

[1] $n = 50,\ \alpha = 0.1,\ \beta = 0.01$

[2] NPMI-based topic coherence computed using the implementation at https://github.com/jhlau/topic_interpretability.

[3] Topics were randomly ordered per user to ensure that equal attention would be paid to each topic throughout the study.

[4] To focus initial attention on topic words, documents were only shown after clicking the "unlock" button.

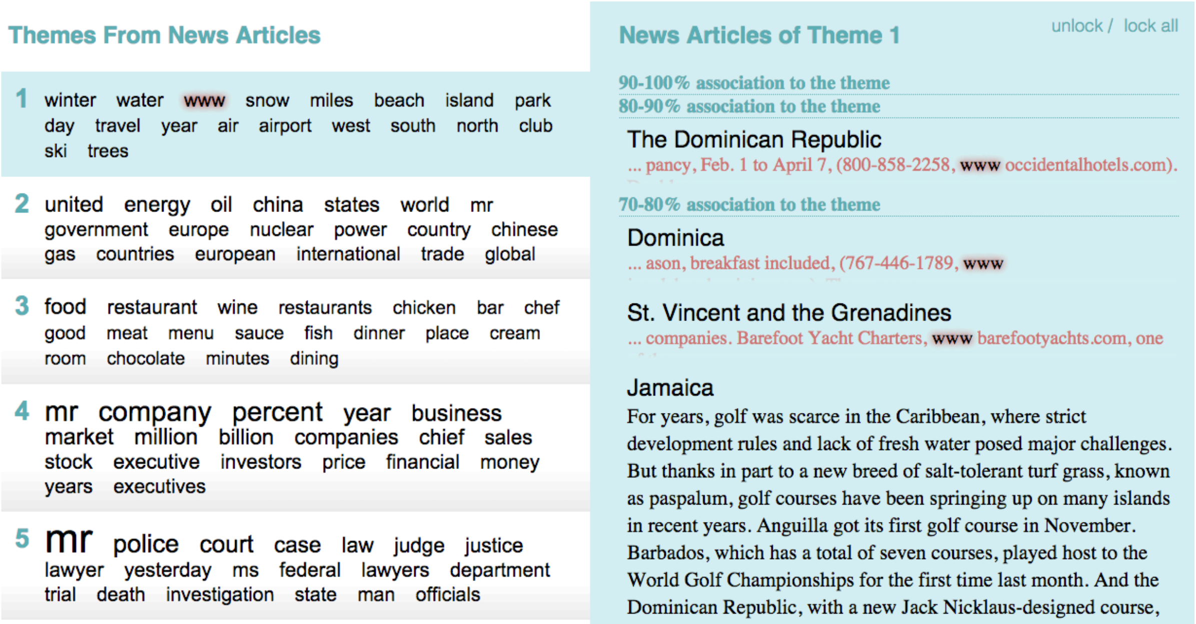
The 20 topics (Table 1) were ordered randomly within the interface per participant (see Figure 1). For each topic, participants:

1. Described the topic based only on the topic words and rated its *clarity* from 1 to 10. Associated documents were not shown.
2. Explored the associated documents, described them, and rated their consistency as a set from 1 to 10 (ignoring topic words).
3. Rated the strength of the relationship between the topic and the documents from 1 to 10.
4. Provided open-ended suggestions for how to improve the quality of the topic and associated documents.

While this study primarily focused on identifying refinements proposed by participants themselves, at the end of the session we also asked participants to use 7-point scales to rate the perceived effectiveness of nine previously identified refinement operations (Figure 2): merge similar topics into one, split words in a topic into groups, split documents in a topic into groups, remove a word from a topic, remove a word from every topic, remove irrelevant documents from a topic, and attach documents to a topic (Andrzejewski et al., 2009; Choo et al., 2013; Chuang et al., 2013b; Hoque and Carenini, 2015; Hu et al., 2014; Lee et al., 2012). The sessions were audio and screen recorded.

3.3 Data and Analysis

Seven participants analyzed all 20 topics, while three participants only analyzed 9, 11 or 13 topics before the time elapsed. This resulted in 173 total topic assessments, and on average 8.7 participants (*SD* = 1.4) assessed each topic. The audio was transcribed and coded. As exploratory work, we pursued an iterative analysis approach using a mixture of inductive and deductive codes (Braun and Clarke, 2006; Hruschka et al., 2004). For our primary focus of analyzing proposed topic model refinements, we created a codebook with 11 refinement operations based on topic


Themes From News Articles
News Articles of Theme 1
unlock / lock all
90-100% association to the theme
80-90% association to the theme
70-80% association to the theme

1 winter water www snow miles beach island park day travel year air airport west south north club ski trees

2 united energy oil china states world mr government europe nuclear power country chinese gas countries european international trade global

3 food restaurant wine restaurants chicken bar chef good meat menu sauce fish dinner place cream room chocolate minutes dining

4 mr company percent year business market million billion companies chief sales stock executive investors price financial money years executives

5 mr police court case law judge justice lawyer yesterday ms federal lawyers department trial death investigation state man officials

The Dominican Republic
... pancy, Feb. 1 to April 7, (800-858-2258, www occidentalhotels.com).

Dominica
... ason, breakfast included, (767-446-1789, www

St. Vincent and the Grenadines
... companies. Barefoot Yacht Charters, www barefootyachts.com, one

Jamaica
For years, golf was scarce in the Caribbean, where strict development rules and lack of fresh water posed major challenges. But thanks in part to a new breed of salt-tolerant turf grass, known as paspalum, golf courses have been springing up on many islands in recent years. Anguilla got its first golf course in November. Barbados, which has a total of seven courses, played host to the World Golf Championships for the first time last month. And the Dominican Republic, with a new Jack Nicklaus-designed course,

Figure 1. User interface for the interview study, showing the 20 topics extracted from a corpus of news articles. Topics were called "themes" during the study. There are two scrollable panels: the topics (left) and the documents associated with the currently selected topic (right). The documents are initially hidden, but shown as participants clicked the "unlock" button. Hovering over a topic word on the left highlights that word (e.g., "www") in context for the displayed articles (red font) and highlights it in other topics if applicable. Clicking a document title toggles its full text.

quality issues identified in past work and discussions among the research team. The codebook included: *add*, *remove*, and *replace words, change word order, capitalize proper nouns and honorifics, connect multi-word phrases*, *add* and *remove documents*, *split the topic*, and *remove the topic completely*. The unit of analysis was a *topic assessment*, that is, the set of open-ended comments that a participant made while improving an individual topic. Each topic assessment could be coded with 0 to multiple refinement operations. To calculate inter-rater reliability, two researchers independently coded 20 randomly selected topic assessments. Krippendorff's alpha was $\alpha = 0.67$. Disagreements were resolved and codes clarified through discussion, and a second round of coding on 20 new random topic assessments achieved higher inter-rater reliability ($\alpha = 0.91$). The first researcher then coded the remaining transcripts. For other qualitative findings where we were not interested in specific counts, one researcher employed the same process but without inter-rater reliability.

3.4 Findings

This section describes how our interview participants interpreted and assessed the topic model, identified problems, and suggested refinement operations. First, to provide context for the findings, we analyze topic quality and coherence.

3.4.1 Subjective Topic Quality and Topic Coherence

Participants also provided topic quality ratings on three metrics. All three metrics were reasonably high across all topics (10-point scales): clarity of the topic words (M = 7.8 out of 10, SD = 2.2), consistency of the document set (M = 8.0, SD = 2.0), and topic-document correlation (M = 7.5, SD = 2.3). However, Topics I19 and I20 were rated consistently poorly (average ratings of only 2.9 to 6.5 across the three measures), and, as will be seen below, two participants (P3 and P10) felt they should be removed entirely.

We also analyzed observed topic coherence, a common computed metric of topic quality (Newman et al., 2010). While many metrics exist for topic coherence, we chose to use a metric based on NPMI as described by Lau et al. (2014). NPMI-based topic coherence scores can range from -1.0 to 1.0, and higher values correspond to higher coherence. Table 1 shows the full list of topics used in this study, with coherence scores of on average 0.07 (SD = 0.05, *range* = 0–0.19). Consistent with prior work motivating the use of NPMI-based topic coherence, there was a significant correlation between coherence scores computed from the topic words and subjective ratings of the clarity of those topic words ($r = 0.28$, $p < .01$).

3.4.2 Challenges to Topic Interpretation

Participants described each topic after first seeing only on the topic words, then with access to the associated documents. Our analysis focuses on three challenges that arose.

Topics may be misinterpreted due to the "bag of words" presentation (Smith et al., 2017). While topics are often presented as ordered lists of words (e.g., (Chaney and Blei, 2012; Eisenstein et al., 2012b; Gardner et al., 2010; Hu et al., 2014)), our study confirms common issues with this presentation. Participants sometimes overlooked important words, read too much into words, assumed adjacent words went together, or incorrectly resolved polysemy. For example, in Topic I14, P2 and P4 overlooked book and movie-related words; P2 called the topic "*summaries of new movies*", while P4 called it "*talking about books [...] best-sellers*". In other cases, participants built phrases from adjacent words. Sometimes this was appropriate, such as for "*court case*" and "*death investigation*" in Topic I5. However, since topic words lack context when presented without the documents, users sometimes assumed incorrect meanings. For Topic I11, for example, three participants (P1, P4, P6) incorrectly thought the topic was about "*Mr. Jan*", not realizing "*jan*" was an abbreviation for January.

Topics may be hard to understand based on the topic words alone. Highlighting the importance of topic modeling interfaces that allow users to browse documents (Chaney and Blei, 2012; Gardner et al., 2010), there were many 'aha' moments where the participants interpreted a topic one way based on the words, then changed their interpretation after seeing the document set. For example, P8 was initially unsure of Topic I16's meaning, but after viewing the documents, exclaimed, "*Ah! Residential sales. I didn't expect it*". Participant P7 said in general, "*it was hard to understand some themes before seeing the articles, but once I saw the articles they all made sense*". Furthermore, P2 felt that including more words from the document titles in the topic words may address this issue.

Topics lacking coherence are hard to interpret. Even given that topics may be hard to understand based on the topic words alone, many participants were frustrated by seemingly random topics. For example, P3, when unable to

Table 2. Eleven refinements identified from the interview study, sorted by popularity. The crowdsourced experiment evaluates the top six of these, with *Connect phrases a*lso collapsed into *Merge words*.

Refinement Operation	Frequency of Suggestion	Participants (/10)	# Topics (/20)
Add words	57	10	17
Remove words	25	9	14
Change word order	16	7	13
Remove documents	15	8	8
Split topic	9	4	7
Connect phrases	5	4	5
Merge words	4	3	3
Replace words	4	4	2
Capitalize proper nouns & honorifics	2	1	2
Remove topic	2	2	2
Add documents	1	1	1

interpret Topic I19, said, "*I cannot tell what this is about. The words are so random*". And P10, when reviewing Topic I20, said, "*the terms are just all over the place*", and complained that in general topics with a wide range of words were difficult to interpret.

In addition to implications discussed in Section 5, these challenges highlight the importance of approaches to improve the "bag-of-words" presentation, such as handling multi-word phrases as single tokens (Wallach, 2006), providing context for words (Gardner et al., 2010), or displaying patterns within and across topics (Chuang et al., 2012).

3.4.3 Proposed Topic Refinement Operations

While not all participants had suggestions for improving all topics, 110 topic assessments (63.6% of the 173) included at least one refinement. On average, each participant made 14.0 suggestions (*SD* = 5.9), and each of the 20 topics received 7.0 suggestions (*SD* = 2.8, *range* = 2–13). Table 2 shows the 11 topic refinement operations that we identified through the qualitative coding. We describe each operation in order of popularity and provide examples of its use.

Add words was by far the most popular operation and was suggested 57 times overall—at least once for 17 of the 20 topics—and by all participants. Generic topics were narrowed by adding specific words from documents. For example, Topic I16 initially seems to be about generic locations, but the majority of users narrowed it to residential sales by adding "rental" (P2), "sales" (P2, P3, P4, P8, P9), "real estate" (P7, P9), or "listings" (P7). Another use case for *Add words* was to emphasize or strengthen a particular aspect of an already-specific topic. For example, Topic I5 was about criminal trials, but participants added words such as "crimes" (P2, P7, P9), "cases" (P4), "violence" (P6), and "criminals" (P9). In another example, P6 intensified the tone of Topic I1's sports theme by adding a word that was not taken from the documents: "*I'd like to add more words like 'compete', 'winning' and 'losing'. Some of the words are already talking about it, but I would intensify them like 'destroy'*".

Remove words was suggested by nine participants a total of 25 times (*M* = 2.5, *SD* = 2.3). Here, three general cases occurred. First, some topics contained meaningless words such as honorific titles (e.g., "mr") and common verbs or nouns (e.g., "made", "years")—words that an experienced user would typically add to a "stop words" list when building the model. For example, four participants (P3, P6, P8, P10) looking at Topic I20 immediately suggested removing "mr" and "ms", and P10 specifically said, "*I am annoyed by common words that don't help me understand the theme*". Second, participants removed words that did not match the documents. For example, P2 wanted to remove "brooklyn" from Topic I8 after not finding it in any of the top documents. Third, participants removed words when they thought the words were too specific given the set of documents. For example, while reviewing Topic I1, two participants (P3, P9) saw "football" and "giants" (a football team), and were surprised to see documents about baseball and basketball as well. They wanted to make the topic more generally about sports by removing the words specific to football.

Change word order was suggested a total of 16 times by seven participants (*M* = 2.3, *SD* = 2.6). Topic words were initially ordered by probability, but the order and font size did not always match participants' understanding of the meaning of the topic. In such cases, participants chose to promote or demote particular words. For example, P3 saw that Topic I4 was about medical research rather than general health, and said, *"I think 'research' should be emphasized more"*.

Remove documents was suggested 14 times (*M* = 1.5, *SD* = 1.2) by eight participants. For example, in Topic I13, P3 wanted to remove documents about insurance, saying, *"Then it becomes more about how health care is being used, bills that have been approved, and how public money is being used"*. While the interface only showed a subset of documents associated with each topic, participant comments revealed that unseen documents were sometimes considered for removal. For example, P10 wanted to remove documents about Senator Dodd from Topic I7, but wanted that removal to apply to all documents associated with the topic (*"like a Boolean search"*) rather than just the small set of top documents that he could see.

Split topic was suggested by four participants nine times (*M* = 0.9, *SD* = 1.3). We witnessed two scenarios for splitting a topic. First, three participants wanted to split topics because they contained two unrelated concepts (e.g., Topic I18 was about cars and TV shows). Another scenario was to create sub-topics under the current one. For example, two participants suggested splitting Topic I4, which was about general medical research, into more specific topics. From P10: *"I would refine it to have multiple sub [topics]. Like human body system, circulatory system, and nerve system"*. This scenario points to the potential utility of hierarchical topic models created by Griffiths and Tenenbaum (2004). Relevant to how an algorithmic implementation of split topic should function, at least some participants expected that splitting topic words would have a cascading effect, automatically updating remaining documents and topic words. For example, P10 wanted to split Topic I10, which is about medical research, into more specific topics based on only a few examples that did not necessarily match topic words and documents, such as *"human body system, circulatory system, and nerve system."*

Connect phrases was suggested by four participants to create multi-word phrases from topic words a total of five times (*M* = 0.5, *SD* = 0.7). For example, P10 found "white" and "house" in Topic I10 and wanted to replace them with "white house". P9, seeing the word "york", created "New York".

Merge word was suggested by three participants to merge multiple topic words a total of four times (*M* = 0.4, *SD* = 0.7) for three different topics. All cases involved merging words with the same root: "restaurant, restaurants" (P8), "art, arts, artist" (P7, P8), and "book, books" (P2).

Replace words was suggested by once each by four participants to replace one set of words with another. This refinement is essentially a compound operation, combining remove and add words, but was specifically framed as "replacement". For example, for topic I18, P10 said, "*Certainly take out car, cars, toyota, and ford. And put in some stations like CBS, ABC, and television shows*".

Capitalize proper nouns and honorifics was suggested twice by one participant. Note that all topic words were transformed to lower case for the study.

Remove topics was suggested by two participants, P7 and P3, who wanted to remove Topics I19 and I20 entirely; these topics had the lowest topic coherence.

Add documents was suggested by one participant who wanted to add documents about President Bush to Topic I10, although the participant could not point to an exact document to add because they did not have access to the full corpus.

3.4.4 Perceived Effectiveness of Pre-defined Refinements

After the open-ended topic assessment task, we asked participants to rate the perceived effectiveness (utility) of nine refinement operations the researchers enumerated before the first user study. Users rated all nine operations positively (all *M* > 4 on a 7-point scale, where 7 = very effective). The overall ratings (Figure 2) also reflect the popularity of the open-ended suggestions made by participants earlier in the session. One noticeable difference, however, is that although no participant suggested splitting off documents into a new theme, the response to that refinement was largely positive (*M* = 6.0, *SD* = 0.3).

3.5 Summary

In working with a topic model that allows for user input (i.e., human-in-the-loop), the ability to interpret, assess, and refine individual topics are each equally important steps. If a user misinterprets a topic, then any subsequent

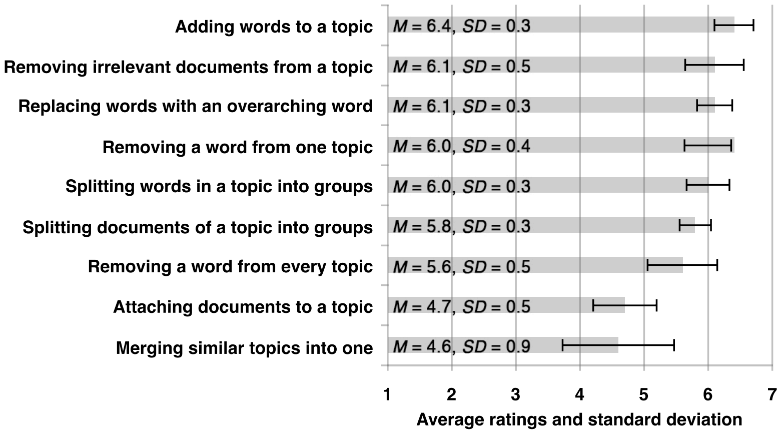


Figure 2. Average subjective effectiveness ratings on a set of previously proposed refinement operations (1: not effective at all, 7: very effective). The ratings generally reflect the frequency of open-ended suggestions made earlier in the interview study. N = 10; bars show standard error.

refinements will take the topic further from its original meaning. If a user incorrectly assesses a topic as poor quality, unnecessary time may be spent attempting to refine it, and finally refinements must achieve what the user intends to change in the model in an efficient manner. This study identified common issues faced by non-expert users during each of these steps, such as often needing both the topic words and documents to understand a topic—many existing tools provide only topic words to aid navigation to subsets of the corpus.

We also characterized how non-expert users want to refine a topic model when presented with no constraints or preconceived ideas about possible refinement operations. Among the more popular suggested operations, *Remove document*, in particular, has been suggested previously (Ramage et al., 2009) but is not to our knowledge supported in existing systems. Many of the other operations have been supported in previous systems, but participants expressed them in simple terms that often differ from current implementations, such as changing word order rather than word weightings; we further discuss implications for this finding in Section 5.

The interview study, however, only captured projected rather than actual use of refinement operations. Thus, we also conducted a larger crowdsourced experiment that allowed users to interactively apply the most popular refinements identified in the interviews. Of the 11 user-identified refinement operations suggested here, the next study includes the most popular: *Add words*, *Remove words*, *Change word order*, *Remove documents*, *Split topic*, and a combined *Connect phrases/Merge words*.

4. Crowdsourced Experiment: Comparing Topic Refinement Operations

To understand how non-expert users refine a topic model with the operations identified in the interview study, we evaluated the six most popular operations using a crowdsourced evaluation on Amazon Mechanical Turk (AMT). We excluded the five less popular operations for the following reasons: *Connect phrases* can be more simply collapsed into *Merge words*, *Replace words* can be achieved by *Remove words* + *Add words*, and the remaining three operations were used extremely rarely and/or, in the case of *Remove topics* and *Add documents*, were not applicable to the crowdsourced study task. As with the interview study, our goal was formative: to assess how users apply refinements *without the constraints* of current topic modeling implementations. We are unaware of an implementation that supports all six operations evaluated here. Moreover, issues such as latency of model updates and the instability of the model across iterations could have a strong negative impact on users. Thus, we only simulated the refinements in a prototypical user interface but did not update the underlying model; the tradeoffs of this decision are discussed in Section 5.

4.1 Method

Topic refinement is an open-ended task (as reflected in the range in user behavior in the interview study), so we opted for an online study to provide a large sample of diverse users.

4.1.1 Participants

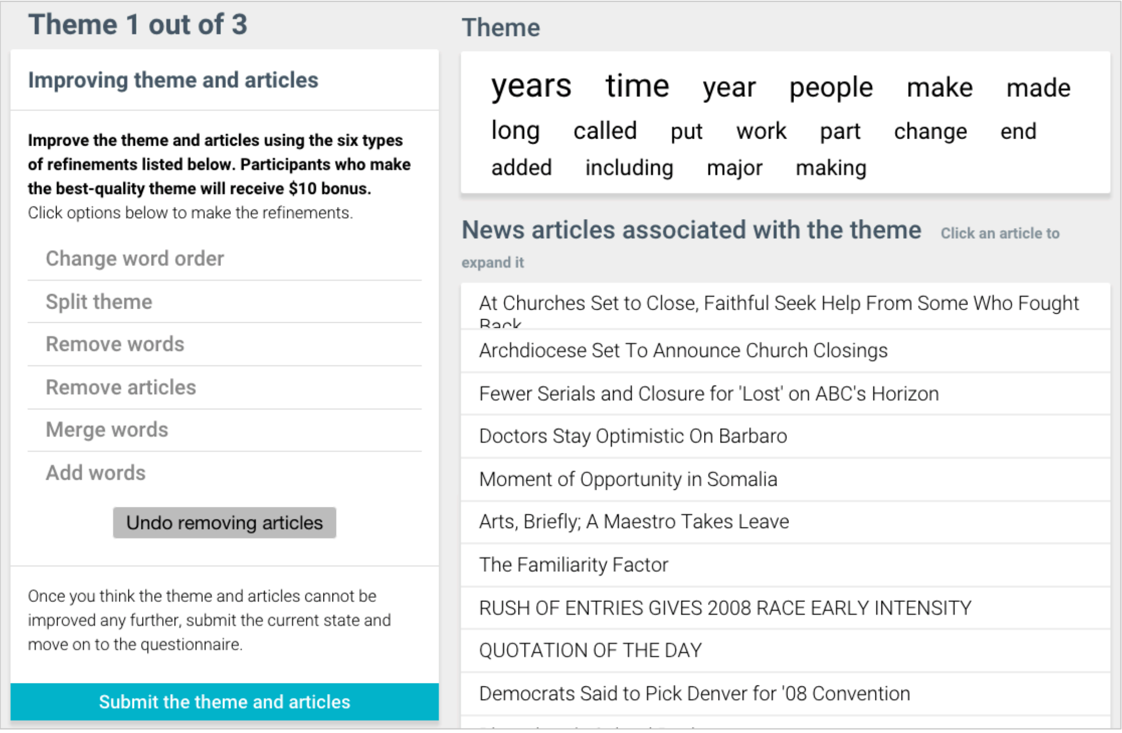
We recruited 90 participants from AMT. All participants are fluent English speakers and residents of the United States or Canada. We paid participants $3 for the session and a $10 bonus for the participant with the best improvements, as measured by an automatically computed topic coherence score.

4.1.2 Topic Model and Refinement User Interface

The NY Times news corpus was the same as in the prior study with the following changes to how we built the topic model: we increased the number of topics from 20 to 30 (to increase specificity) and removed words that interview participants deemed problematic (“dr”, “mr”, “ms”, “don”, “didn”). We did not want users to focus on these obvious flaws, potentially ignoring more interesting refinements.

We built a web application that displays the six refinements in random order in the left panel and a topic with its top 20 words and 40 associated documents on the right (Figure 3). Topics were again shown using a word list visualization to maintain consistency with Study 1 and because recent work has shown that, despite the drawbacks seen in Study 1, word lists are fast to interpret and result in a reasonable understanding a topic’s meaning compared to more complex topic visualizations (Smith et al., 2016). As with the interview study, topics and documents were called *themes* and *articles*. Clicking a document showed its full text. Clicking a refinement operation revealed a brief explanation of its use and any additional interface elements as shown in Figure 4:

- ***Add words*** shows a text input box. Users type a word in it and press Enter.
- ***Remove words*** shows a small ‘x’ icon next to each word. Users can click words to remove them.
- ***Merge words*** is a two-step process: (1) clicking words to be merged, (2) typing a new word and pressing Enter.



Theme 1 out of 3

Improving theme and articles

Improve the theme and articles using the six types of refinements listed below. Participants who make the best-quality theme will receive $10 bonus.
Click options below to make the refinements.

Once you think the theme and articles cannot be improved any further, submit the current state and move on to the questionnaire.

Theme

years time year people make made long called put work part change end added including major making

News articles associated with the theme Click an article to expand it

At Churches Set to Close, Faithful Seek Help From Some Who Fought Back
Archdiocese Set To Announce Church Closings
Fewer Serials and Closure for 'Lost' on ABC's Horizon
Doctors Stay Optimistic On Barbaro
Moment of Opportunity in Somalia
Arts, Briefly; A Maestro Takes Leave
The Familiarity Factor
RUSH OF ENTRIES GIVES 2008 RACE EARLY INTENSITY
QUOTATION OF THE DAY
Democrats Said to Pick Denver for '08 Convention

Figure 3. The experiment interface allows users to switch between the six supported refinement operations by clicking each one in the left panel. There is also a button to undo previous operations. The top 20 words of the theme and 40 associated articles are shown in the right panel. To complete the refining process for the current topic, users click the “Submit the theme and articles” button at the bottom.

- ***Change word order*** allows users to drag and drop words to re-order them in the topic list.
- ***Remove documents*** shows an ‘x’ next to each document, which users click once to remove the document or again to cancel that choice.
- ***Split topic*** allows users to drag words from the topic to a box below that represents a new topic.

Each refinement was immediately applied to the topic and documents, so that users could iteratively assess their refinements and improve upon them. Because of our focus on what users *want* rather than what is technically feasible with current human-in-the-loop topic modeling implementations, these changes were only reflected in the user interface and the underlying model did not update.

4.1.3 Procedure

Sessions took about 20 minutes and began with a task overview followed by a five-minute tutorial on topic models in general, the user interface, and the six refinement operations in random order. A mini-quiz followed each segment (e.g., identify the most frequent topic word) to reinforce content.

Participants then evaluated and refined three randomly selected topics. As mentioned in Section 3.2, understanding and interacting with an individual topic is an important subcomponent of many higher-level tasks (Alexander and Gleicher, 2015; Boyd-Graber et al., 2014). For each of the three topics, participants first provided a short (one-sentence) description and rated the topic on clarity of the topic words, consistency of the document set, and fit between the topic and documents (each question on a 5-point Likert scale). Participants were then asked to *“improve the theme and articles”* (Figure 3) until satisfied with the topic quality. Afterward, participants rated the now-refined topic using the same three Likert scales and described any extra information that would help improve the topic. Finally, after refining all three topics, the participant rated the usefulness of each of the six refinement operations on 5-point Likert scales. They also described any other ideas for refining topic models.

4.1.4 Study Design and Analysis

Each of the 90 participants viewed three randomly selected topics, with the constraint that each topic must appear nine times total across all participants. Nine participants thus refined each of the 30 original topics, resulting in 270 refined topics. We logged the time spent refining each topic.

4.2 Findings

This section reports the frequency and detailed usage for the refinement operations. Participants spent on average 280.0 seconds refining each of the 270 topics (*SD* = 217.4, *Median* = 208) and applied 34.8 refinements (*SD* = 28.7, *Median* = 25). The refined topics had fewer words (*M* = 13.7, *SD* = 4.7) and fewer documents (*M* = 35.4, *SD* = 6.6) on average than the original topics, which reflects the popularity of removing words and documents; the original topics had 20 words and 40 documents.

4.2.1 Quick-to-Apply Refinements Are Preferred

The three most popular refinements—*Remove words*, *Remove documents*, and *Change word order*—manipulate existing words and documents with a single mouse click or drag-and-drop. In contrast, *Add words*, which was the most popular operation in the interview study, was used less than half as often as *Remove words*. This is perhaps because adding a new word requires more effort than manipulating existing content. *Split topic*, an arguably complex operation, was the least frequently used.

Across the 270 refined topics, refinements were not applied uniformly (Friedman test: $\chi^2_{(5,\, N = 270)} = 338.40$, $p < .001$). Most post hoc pairwise comparisons using Wilcoxon signed ranks tests were significant at $p < .05$ after a Bonferroni adjustment; the only two non-significant differences were among the top three operations: *Remove documents* versus both *Remove words* and *Change word order*. Table 3 is thus an almost full ordering of operations.

Table 3. Comparison of refinement operations based on frequency of use per topic (N = 270) and subjective usefulness as rated by participants in the crowdsourced study (N = 90).

Refinement operations	Frequency of Use per Topic			Subjective Usefulness		
	M	*Median*	*SD*	*M*	*Median*	*SD*
Remove words	5.60	4	4.98	4.44	5	0.77
Remove documents	5.46	2	7.89	3.87	4	1.26
Change word order	3.50	2	3.84	3.74	4	1.25
Add words	2.06	1	3.38	3.71	4	1.23
Merge words	1.25	1	1.67	3.43	5	1.34
Split topic	0.94	0	2.20	2.23	2	1.26

Participants' perception of the usefulness of the operations is reflected in the frequency of their use. Overall, participants felt that some refinement operations were more useful than others; the refinement operations had a significant effect on Likert-scale ratings of usefulness (Friedman test: $\chi^2_{(5,\, N = 90)} = 145.50$, $p < .001$). *Remove words* was judged the most useful and *Split topic* the least useful of the six refinements. Post hoc pairwise comparisons with Wilcoxon signed ranks tests and a Bonferroni adjustment showed that *Remove words* was judged significantly more useful than *Merge words*, *Change word order*, and *Remove documents*, and that *Split topic* was less useful than all others operations (all $p < .05$).

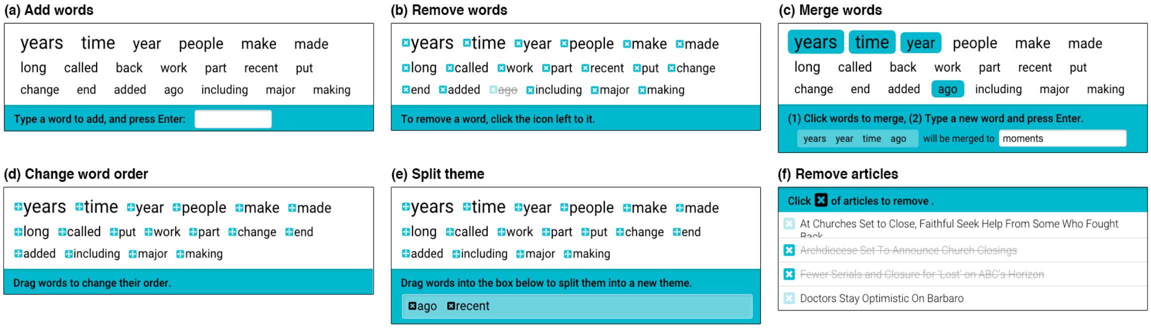
4.2.2 Detailed Refinement Operation Usage

This section analyzes detailed usage patterns of each of the refinements ordered by overall frequency of application.

Remove words was used with 91.5% of topics, to remove on average 6.1 words from each topic (*SD* = 4.9, *Median* = 5). Participants frequently removed generic words (e.g., "year", "million", "people", time), words that should be in bigrams (e.g., "york" of New York), and abbreviations (e.g., "st"). Words with lower probability for the topic were also more likely to be removed, with a significant correlation between word rank and removal frequency (Spearman's r_s =.77, $p < .001$).

Remove documents was used with 67.8% of topics. Participants removed on average 8.1 documents from each topic (*SD* = 8.4, *Median* = 5). As was the case with word removal, documents with lower probability for the topic were more likely to be removed, with a significant correlation between document rank and removal frequency (Spearman's r_s = .67, $p < .001$). This suggests that participants agreed with the topic model's ordering of documents.

Change word order was used in 76.7% of topics, with an average of 3.5 changes for each topic (*SD* = 3.8, *Median* = 2). Words were moved much more frequently to a higher probability position (81% of cases) than to a lower probability position. However, since participants focused on words near the top of the list (i.e., closer to rank 1), there was a significant negative correlation between the original word rank and the frequency with which a word's order was changed (Spearman's r_s = -.63, $p < .003$).



(a) Add words

years time year people make made
long called back work part recent put
change end added ago including major making

Type a word to add, and press Enter:

(b) Remove words

years time year people make made
long called work part recent put change
end added ~~ago~~ including major making

To remove a word, click the icon left to it.

(c) Merge words

years time year people make made
long called back work part recent put
change end added ago including major making

(1) Click words to merge, (2) Type a new word and press Enter.

years year time ago will be merged to moments

(d) Change word order

years time year people make made
long called put work part change end
added including major making

Drag words to change their order.

(e) Split theme

years time year people make made
long called work part put change end
added including major making

Drag words into the box below to split them into a new theme.

(f) Remove articles

Click ☒ of articles to remove .

☒ At Churches Set to Close, Faithful Seek Help From Some Who Fought Back

☒ ~~Archdiocese Set To Announce Church Closings~~

☒ ~~Fewer Serials and Closure for 'Lost' on ABC's Horizon~~

☒ Doctors Stay Optimistic On Barbaro

Figure 4. User interface implementation of the six refinement operations evaluated in the crowdsourced study, designed to support direct manipulation of topic words and documents. Participants had to first select an operation to view these tool details. Note that topics and documents were called "themes" and "articles" during the task.

Table 4. Part-of-speech tags for words used across the refinement operations. Nouns were commonly added, while verbs were often moved to a split topic or removed. Multi-word phrases were often added or created from merges. "Other" included abbreviations (e.g., "st") and incomplete words (e.g., "doesn").

	Original topic	Refined topic	Removed words	Merge words (from)	Merge words (to)	Change order	Split topic	Added words
Noun	74.8%	79.4%	67.5%	72.1%	77.4%	77.7%	73.5%	65.6%
Proper n.	7.6%	6.3%	10.3%	12.7%	4.5%	5.1%	8.3%	9.7%
Verb	7.4%	4.6%	10.6%	4.5%	1.5%	3.4%	9.1%	4.1%
Adjective	9.2%	7.5%	9.0%	8.6%	2.4%	9.6%	7.5%	10.8%
Adverb	0.3%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Phrase	0.0%	2.0%	0.5%	1.2%	14.2%	4.2%	1.6%	9.2%
Other	0.7%	0.2%	2.0%	0.8%	0.0%	0.0%	0.0%	0.0%
# words	1170	3709	1513	997	420	946	253	556

Add words was used in 58.1% of topics to add an average of 3.5 new words to each topic (*SD* = 3.8, *Median* = 2). Most added words could be found in the text of the 40 associated documents (88.7%), almost half occurred in the titles of those documents (41.4%). A small portion of the added words (13.7%) could also be found in the next 20 topic words (not shown to the user), while only 9.7% were not in the documents, titles or next 20 topic words. Multi-word phrases (e.g., "real estate transactions") were 9.2% of additions (Section 4.2.3).

Merge words was applied in 60.4% of topics to create an average of 2.1 words (*SD* = 1.7, *Median* = 1) from 3.4 original words (*SD* = 3.0, *Median* = 3) in each topic. The name of the merged word was often derived from the original words (*N* = 152, 45.1%): e.g., merging "government" and "officials" to "government". Manual coding identified four patterns: merging specific words to generic terms (e.g., "ford" and "toyota" to "manufacturers"), *N* = 259 (76.9%); merging inflected forms (e.g., "school" and "schools" and "art" and "artist", which lemmatization could also help with), *N* = 44 (13.1%); connecting phrases (e.g., merging "white" and "house" to "white house"), *N* = 25 (7.4%); and revising a single word (e.g., "rx" to "prescription"), *N* = 4 (1.2%). The finding that participants used *Merge words* to connect phrases supports our decision to combine the two separate refinements after the interview study.

Split topic was used in 25.6% of topics to move on average 3.7 (*SD* = 3.0) words from the original to a new topic.

4.2.3 *Part-of-Speech Tagging of Refined Words*

Although not included in the primary interview study findings, three participants in that study mentioned that verbs were less helpful than nouns as topic words. To quantitatively assess whether parts of speech (POS) affected refinement operations, we compared POS distributions across refinements (Table 4). Verbs were more frequently *removed* (10.6%) or *split* (9.1%) rather than *added* (4.1%) or *merged to* (1.5%), providing some evidence to support those comments in the interview study. The refined topics had fewer verbs (4.6%) than the original topics (7.4%). Although the bag-of-words representation does not support the visualization of multi-word phrases, our participants often added phrases (9.2%) or merged words into phrases (14.2%). These patterns may be useful for a mixed-initiative system to help automatically identify words (e.g., verbs) that a user may want to refine.

4.2.4 *Original and Refined Topic Quality*

While the refinement operations were only superficially applied and did not impact the underlying topic model, we are still interested in whether participants perceived their refinements to have a positive impact. Before and after refining each topic, participants assessed the topics on three 5-point Likert scales for clarity of the topic words, consistency of the document set, and correlation between the topic words and documents. Participants felt that their own refinements improved the topics on all three characteristics: from 3.6 (*SD* = 1.1) to 4.4 (*SD* = 0.7) for clarity, 3.6 (*SD* = 1.0) to 4.3 (*SD* = 0.6) for consistency, and 3.8 (*SD* = 0.9) to 4.5 (*SD* = 0.6) for the correlation between topic words and documents. These increases were statistically significant with Wilcoxon signed-rank tests (clarity: *Z* = 11.68, consistency: *Z* = 11.86, and correlation: *Z* = 11.62, all *p* < .001).

To more objectively assess quality, we compared topic coherence scores of the original and refined topics using an observed topic coherence metric based on NPMI as we did in the interview study (see Section 2.1 for a justification of the use of NPMI-based topic coherence). Again, because the refinements did not update the underlying model, these results should be considered preliminary. The coherence scores computed for the original topics ranged from 0.0 to 0.23, where higher values indicate more coherent topics. Unlike the original topics, each consisting of 20 words, refined topics may have different numbers of words. Observed topic coherence is computed as the sum of the pairwise PMI between topic words, so the number of words has a direct effect on the coherence score. We therefore adjusted the refined topics' coherence scores by dividing by the number of words in each topic. The resulting adjusted coherence scores for the refined topics ranged from -0.07 to 0.70. The 270 refined topics were more coherent than the original topics, with average coherence scores of 0.11 (*SD* = 0.10) versus 0.09 (*SD* = 0.06). A Wilcoxon signed-rank test showed this was a significant improvement ($Z = 3.54$, $p < .001$), suggesting that participants' refinements improved topic coherence.

4.2.5 Best and Worst Refined Topics

To supplement the analysis above, of the 270 total topics, we explored the 20 worst and 20 best refined topics identified from NPMI-based topic coherence scores. Even after normalizing based on the number of topic words, many of the topics identified as most coherent were succinct, such as "*season, game, football, team, coach, players, quarterback*". On average the worst (least coherent) topics had 17.9 words (*SD* = 6.2), while the best (most coherent) topics had 9.7 words (*SD* = 2.5); this difference was statistically significant with a Mann-Whitney U test ($U = 42$, $z = 4.26$, $p < .001$). However, short topics were not always coherent, such as "*People, recent, change, church, F.A.A*", which appeared in the worst set. Despite the difference in number of words, the number of refinements applied to the best and worst topics and the number of documents remaining were both similar. On average, participants applied 19.1 (*SD* = 14.4) refinements for the worst topics, compared to 21.7 (*SD* = 10.3) for the

Table 5. Three examples of topic refinements, with self-ratings and topic coherence. Topics C1 and C23 are examples where user refinements increased topic coherence (NPMI), whereas Topic C5 was the topic with the highest original coherence and user refinements tended to reduce its NPMI-based topic coherence score.

ID	Refinement	Topic Words (*Number of Documents)*	Subjective Ratings			Topic Coherence
			Clarity	Coherence	Correlation	NPMI
C1	Original	www winter water snow hotel travel island trees miles ski beach airport air ice mountain weather travelers resort airlines park *(40)*	3.1	3.1	3.4	0.06
	Refined 1	travel airlines winter hotel water island ski airport air weather resort park scenery[trees, mountain, ice, snow, beach] *(40)*	5.0	5.0	5.0	0.09
	Refined 2	winter water snow trees miles ski beach airport air ice mountain weather travelers resort airlines park Travel[travel, hotel, island] summer *(36)*	4.0	5.0	5.0	0.10
C23	Original	house building number square market room estate million street real weeks bathrooms apartment year bedroom half foot listed home property *(40)*	3.7	3.7	3.7	0.05
	Refined 1	real estate[real, estate] new york residential sales home sales apartment house building bathrooms bedroom home property square feet real estate transactions *(39)*	5.0	4.0	5.0	0.14
	Refined 2	property real estate[real, estate] house building market listed room street bathrooms apartment bedroom home rent *(36)*	4.0	4.0	4.0	0.17
C5	Original	food restaurant wine restaurants chicken chef menu bar sauce meat dinner good fish dining street cream minutes cheese wines chocolate *(40)*	3.9	3.9	4.0	0.23
	Refined 1	restaurants[restaurant, restaurants] food menu dinner Drinking[wine, bar[wine, wines],wine, wines] good chicken chef sauce meat cream cheese fish chocolate dining *(40)*	5.0	4.0	5.0	0.16
	Refined 2	sauce meat dinner good fish street cream wines cheese chocolate *(37)*	3.0	4.0	2.0	0.18

best topics. For number of documents, the worst topics had 36.1 on average (*SD* = 5.3), while the best topics had 35.9 (*SD* = 6.8). Mann-Whitney U tests revealed no significant differences between the best and worst topics on these two measures.

Although in general participants' refinements improved coherence scores, refinements consistently worsened topics with initially high coherence values. Table 5 shows examples of both improving and worsening the original topics from the point of view of computed topic coherence; for C5, the original topic with the highest coherence, only two out of nine participants were able to improve on that score. This finding could reflect that NPMI-based observed topic coherence is not a perfect measure of the *user's* interpretation of topic quality, particularly for topics with higher scores. However, it also suggests that a mixed-initiative approach that helps focus user attention away from topics that already have high computed coherence could help improve the overall model most quickly.

4.3 Open-ended Suggestions

At the end of the study, users were asked to provide open-ended suggestions for topic refinements. Of the 22 suggestions received, 10 participants provided ideas for new refinement operations, such as reordering documents by drag-and-drop (4 participants), splitting a topic by grouping documents (3 participants), creating sub-topics under a topic (2 participants), and an option to completely redefine poor topics based on their associated documents (1 participant).

The remaining suggestions were about the usability of the refinement process and many align with known challenges in mixed initiative interfaces. Two participants were overwhelmed by the amount of information, wanting fewer articles or topic words, such as: "*try to only have 5 [words] max or something, that way [...] we aren't bombarded with a ton of choices*". Two other participants wanted immediate feedback on the effects of their refinements, such as reordering the documents based on changes to the topic words: "*That way you could visually gauge, as you work, the effect you're having*". Participants also made several suggestions that motivate the need for the system to more actively support the refinement process. For example, four participants wondered if the system could automatically detect words with the same stem (e.g. "book" and "books") or multi-word phrases (e.g. "book" and "review"). Finally, another two participants wanted more contextual information such as the frequency of word occurrence in associated documents and the change of topic quality after each refinement operation. Several of these comments highlight the importance of further study where users are provided with the output from an updated model.

4.4 Summary

The crowdsourced experiment highlights users' preference for operations that are simple to apply (e.g., removing vs. adding a word) and shows that non-expert users felt they could effectively apply the given set of refinement operations to improve topics. While the refinements were only applied within the user interface rather than updating the underlying model, preliminary topic coherence analysis is also promising, suggesting that there was an objective improvement in topic coherence. More importantly though, this study provides concrete ideas for how to improve human-in-the-loop topic modeling for non-experts, such as what refinement operations to prioritize implementing if tradeoffs exist, patterns that could be leveraged for mixed-initiative support (e.g., perceived low utility of verbs as topic words), and the need for immediate feedback on how refinements cascade through the model. We discuss these and other implications further in the next section.

5. DISCUSSION

Our two formative studies are complementary and together provide evidence both for how non-expert users perceive topic models as well as how to make human-in-the-loop topic modeling approaches more accessible. Given a historical focus on the algorithm side of human-in-the-loop topic modeling, user-centered work, such as our two studies, is critical for achieving the broad impact that topic modeling has the potential to make. The interview study provides qualitative insight into how non-experts perceive and *want* to fix a topic model when given no constraints or preconceived notions about how to do so, while the crowdsourced study provides further quantitative evidence for how to prioritize these operations as well as ideas for how to more actively support users in their refinement process. These findings should be relevant to tasks where the user interacts directly with the topic model itself rather than applications where the topic model is used more indirectly (e.g., to generate recommendations for documents similar to a given example). Here, we discuss implications for the design of human-in-the-loop topic modeling approaches and future work on more active, mixed-initiative support for the refinement process, as well as limitations of the research.

5.1 Implications for Refinement Operations Intuitive to Non-Experts

We recommend a set of refinement operations to provide for non-expert users, which should be particularly useful for tasks similar to those used in our studies: theme identification and individual topic refinement. To direct future work, we also reflect on other potential refinements and questions of implementation.

Intuitive refinement set for non-expert users. We identify a variety of user-suggested refinements, which we condense to a set of six operations that should be supported for refining individual topics: *Add words*, *Remove words*, *Merge words*, *Change word order*, *Remove documents*, and *Split topic*. These topic-level refinements should also be combined with model-level refinements, including the *Remove topic* operation identified by participants in our interview study and the *Merge topics* operation found to be useful in a previous user study by Hoque and Carenini's (2015). While our interview study findings suggest that non-expert users strongly expect these specific operations to be available when interacting with a topic model, additional operations will also likely prove useful in different contexts. For example, one model-level concern not addressed in our study is *coverage*—if a user is highly familiar with a corpus, they may notice or more easily discover that some topics of interest are not represented in the model—a situation that may require the ability to *Create a new topic*. As well, for the six topic-level refinements, the crowdsourced study provides an almost entirely statistically significant ordering of the frequency of use and perceived utility of the operations, and this ordering can prioritize some refinements over others. However, different tasks, specificity of topics, or the ratio of topics to documents may impact usage patterns. For example, a task that requires answering a factual question will likely result in the user focusing on a small number of topics within the overall model, and not require as many model-level refinements as a theme identification task. As another example, users will likely employ *Split topic* more often when there is a small number of general topics compared to many specific topics. Grounded in the findings provided from our studies, future work should explore these possibilities.

Backend versus user-facing refinements. Refinement operations may not always need to be implemented on the model side and can instead be realized through changes to what the user views. For example, one use of *Merge words* is grouping similar hyponyms (e.g. "ford" and "toyota") under a hypernym (e.g. "manufacturers"). This change may be interpreted as a means of organizing the topic words to be easier for the user to read rather than a deeper specification that the words should always be linked together in the underlying model (e.g., using a *Must-link* constraint). To limit complexity for non-expert users, we intentionally did not distinguish in our studies between what was shown in the user interface and what might be stored in an underlying model—these were presented as a unified whole. Further work is needed to understand if users behave differently when this distinction is described and they are presented with the possibility of refining only the displayed topics in the user interface or also refining the backend model. In either case, with users being aware or unaware of the user interface vs. model distinction, further work is also needed to understand which refinement operations should be only user-facing and which should update the underlying model. Ultimately, the answer may depend on how refinements are implemented on the model side and the extent to which applying a refinement results in cascading effects that are unpredictable and unwelcome from the user's viewpoint.

Disconnect between user perceptions and current implementations. While most of the operations listed above are currently supported in human-in-the-loop topic modeling, how our participants expressed their needs demonstrates a disconnect between user perceptions and the technical details of many current implementations. For example, Hu et al. (2014) implement *Add words* and *Remove words* using underlying *Must-link* and *Cannot-link* constraints, but we did not find any evidence that non-expert users think about refinement operations in this way. As another example, sometimes existing low-level operations could theoretically support our participants' needs, but these operations do not directly match what is desired. For example, *Add words*, *Remove words*, and *Change word order* can be accomplished by a general operation that allows the user to tweak the weightings of words within a topic—as supported, for instance, with UTOPIAN (Choo et al., 2013) and iVisClustering (Lee et al., 2012). That weighting operation, however, is more complex and may not be as intuitive to a user who has little knowledge of how the underlying model works. Finally, *Connect phrases* and *Capitalize proper nouns and honorifics* can be supported through text pre-processing, but this solution does not allow the user to incrementally update a model as they are working with it. These disconnects point to the need for further algorithmic work in human-in-the-loop topic modeling that truly takes a human-centered approach.

Consideration of technical feasibility. To characterize how non-expert users assess and expect to be able to refine topic models, we purposefully employed a study method that was unconstrained by the limits of current human-in-the-loop topic modeling implementations. In practice, however, a balance will need to be achieved between user expectations and what is technically feasible from an algorithmic viewpoint. For example, with an LDA-based approach (e.g., Interactive Topic Modeling (Hu et al., 2014)), seemingly small changes on the user side could have unpredictable and nonsensical cascading side effects—an issue that was not considered by our non-expert users. In some cases these side effects may be inconsequential, while in other cases they could outweigh the benefits of offering the refinement operation in the first place. One possibility to allow users to manage cascading side effects is if the system can provide an estimate of the potential impacts of a refinement before it is applied. Latency is an additional concern, as waiting for the model to update may negatively affect the user workflow, and some refinements may take longer to incorporate than others. These questions emphasize the importance of revisiting our findings in the context of a fully interactive system.

5.2 The Potential for More Active, Mixed-Initiative Support

Even in our constrained study tasks, refining an individual topic took a non-trivial amount of time—on average several minutes. The topic models used in our studies only contained 20 or 30 topics, which would be considered small models for real-world applications that may have models with hundreds of topics, and may require many additional rounds of iterative refinements. Several ideas for future work on more actively supporting this refinement process emerge from our studies.

Direct users to topics with high refinement potential. Naïvely asking users to refine all topics is at best intractable and at worst counterproductive, nullifying the advantages of an unsupervised algorithm. As seen in the crowdsourced study, refining highly coherent topics can make those topics worse as measured by NPMI-based observed topic coherence; this may, however, be a limitation in the measurement's ability to capture how users perceive topic quality. Likewise, topics with low coherence are likely not useful to focus on and instead are candidates for complete removal (as shown by Topic I19 and I20 in the interview study). Instead, topics with average or just below average coherence may be the best place to focus, where the user can still interpret their intended meaning and suggest appropriate refinements.

Select an ideal-sized subset of words or documents to show per topic. In the interview study, showing only the top *N* (in our case 20) topic words or documents sometimes led to misconceptions about the topic that could have been overcome had more words or documents been shown. In addition to strategies that provide more context for topic words (Choo et al., 2013; Gardner et al., 2010; Lee et al., 2012), it may be useful to automatically determine how many words are needed to characterize a topic, using a probability cut-off, for example. Such a cut-off could also be applied to the number of documents shown. However, in both cases, the inclusion of some lower probability items may be useful for diversity.

Support common patterns of refinement. The differences between refinement use in the two studies also reflect a bias toward easier, less cognitively taxing refinements (e.g., removing words) in the crowdsourced study. How participants applied refinements can inform future interfaces that provide more active suggestions, thus minimizing effort. For example, potentially useful supports arising from the crowdsourced study include suggesting adding new words taken from document titles, suggesting removal of common verbs, and suggesting merging of words with the same stem (e.g. "book" and "books") or multi-word phrases (e.g. "book" and "review"). Automatic labeling techniques for selecting the most salient words to describe a topic (e.g., (Lau et al., 2011)) may also be adopted to suggest potential words to add. It may also be useful to assess the user's higher-level intent in refining a topic, such as making the topic more general, which could then lead to specific refinement suggestions.

Provide immediate feedback on the impact of user refinements. Providing feedback to users on the perhaps cascading changes that their refinements have on the underlying model is obviously necessary for any usable system. However, a key challenge that will likely arise is latency—for example, updates in the system used by Hu et al. (2014) took 5 to 50 seconds, a non-trivial amount of time for the user. If users expect more immediate feedback, as is suggested by the open-ended comments in the crowdsourced study, alternatives such as providing quick previews of potential changes before applying them may be needed. However, sufficiently reducing update latency for a truly interactive experience or, alternatively, providing accurate previews of those updates are both open areas of research.

Generalize refinement operations for unseen words and topics. In both studies, participants had access to only the top-*N* words and documents for each topic. However, we observed cases where participants expected their refinements would generalize to unseen words and documents. For example, when splitting topics or removing documents, one behavior was to first pick a few documents to specify a new subtopic; P18 found a few example words and documents about a subtopic, and said, "*I guess the theme can be separated into computer and TVs*". In practice, generalizing from the user's specific operation will be necessary for operations such as split topic and remove document so that users do not have to manually touch all documents in a corpus or all words associated with a topic—neither of which would be realistic and are in opposition to the goal of using a topic model in the first place.

5.3 Limitations

We made several simplifying assumptions that limit the scope of our findings. First, to allow for comparison across users, we used only a single topic model in each study. While the general pattern of findings will likely apply to other topic models (e.g., based on larger corpora or with more topics), it will be important to confirm the extent to which this is true. Further, the model used in Study 1 had 20 topics, while that used in Study 2 had 30 topics for the same corpus. We did not observe any clear impacts based on this difference, but it is possible that the number of topics in a model affects refinement operation usage, a question that should be examined in future work. While we generated topic models for our studies using LDA, the refinements suggested by non-expert users should be relevant to other types of topic modeling algorithms, due in part to our decision not to constrain users to existing implementations of refinement operations. However, how easily and how well each type of refinement can be implemented will of course depend on the particular modeling approach.

Second, while the task in Study 1 was intended to be reflective of a real-world summarization task, such as is necessary for content analysis (Chuang et al., 2014), the task used in Study 2 was more constrained and should be seen as an interim step within a larger task (Alexander and Gleicher, 2015); the two studies complement each other. In Study 2, users only viewed individual topics rather than the whole model, thus findings from that study are only applicable to topic-level refinements and need to be combined with model-level refinements in future work (e.g., splitting and merging topics (Hoque and Carenini, 2015)). As mentioned earlier in the Discussion, it is also likely that at least the frequency ordering of the refinement operations will differ for different real-world tasks, a question that should be explored in future work.

Third, we used the simple word list topic visualization because it is common as well as being easy and fast to interpret (Smith et al., 2016). However, more complex visualizations exist, such as a matrix view (Chuang et al., 2012) or network graph (Smith et al., 2016), and the different information those visualizations expose could impact the user's refinement process differently from the simple word list. Further work is needed to assess the extent to which this is the case.

Fourth, both studies involved non-expert users without any background in topic modeling and, in the case of Study 2, crowdsourced participants. While the decision to focus on non-expert users was intentional, as widespread use of interactive topic modeling requires adoption by non-expert users, studying more experienced users may give different insights. Expert users, for example, would have a better understanding of the technical feasibility of different operations and the potential cascading side effects of seemingly straightforward operations (e.g., *Add words*). It would also be interesting to compare the quality of models refined by novices versus experts—for example, analyzing what problems experts tackle first. For Study 2 we decided to recruit crowdsourced participants so that we could collect a larger amount of data than would have been feasible in person. The potential limitation is that these crowdworkers were paid for task completion, so may have been biased toward quick-to-apply refinements; however, efficiency is also generally good for usability, so while some of the detailed findings may change with in-person participants who have different motivations, we expect the overall patterns should hold.

Finally, a truly interactive topic modeling system updates the model with the constraints specified by the user's refinements. We did not feed the refinements into an iterative retraining process because we wanted to explore how people interpret and want to refine topic models without constraining them to current human-in-the-loop topic modeling implementations. However, the next step is to study similar questions with a fully interactive model. One issue that will need to be explored is the extent to which we can implement each of these refinement operations in the way that our participants envisioned them—some operations may be easier to support than others. Additional user interface issues will also need to be addressed with a fully interactive system, such as maintaining stability across iterations of a model and helping non-expert users to compare model versions. Lastly, propagating changes through a machine learning model can cause cascading changes to many topics; the idealized updates in our study avoided this complication. With a fully interactive system, however, these changes may cause the user to adapt their own behavior over time. Interacting with machine learning can generally be unpredictable and the interaction itself may cause both the learning system and the user to evolve their behaviors.

6. CONCLUSION

We reported on results from two user studies with non-expert users to explore how they interpret, assess, and wish to refine topic models without any constraints. The first, a face-to-face interview study, explored how users understand and wish to refine a topic model in an exploratory setting, and the second, a larger crowdsourced performance study, demonstrated how users prioritize these refinements and provided design ideas for actively supporting users in this process. In particular, we identified the primary refinements that should be supported by any interactive topic modeling system to support non-expert users and discussed how understanding user needs can guide the implementations of these refinements. We also identified areas for future work where more active, mixed-initiative support may be useful during the refinement process, such as pointing users to topics with high refinement potential and providing immediate feedback. Finally, future work should ensure that these findings hold for diverse models (in terms of quality and content) and when interacting with a full topic model instead of individual topics (as done in our crowdsourced study).

ACKNOWLEDGMENTS

This work was funded by NSF grants IIS-1409287 and IIS-1409739.

REFERENCES

Aggarwal, C.C., Zhai, C., 2012. A Survey of Text Clustering Algorithms, in: Aggarwal, C.C., Zhai, C. (Eds.), Mining Text Data. Springer US, pp. 77–128.

Aletras, N., Baldwin, T., Lau, J.H., Stevenson, M., 2015. Evaluating Topic Representations for Exploring Document Collections. J. Assoc. Inf. Sci. Technol. doi:10.1002/asi.23574

Alexander, E., Gleicher, M., 2015. Task-Driven Comparison of Topic Models. IEEE Trans. Vis. Comput. Graph. PP, 320–329. doi:10.1109/TVCG.2015.2467618

Alexander, E., Kohlmann, J., Valenza, R., Witmore, M., Gleicher, M., 2014. Serendip: Topic model-driven visual exploration of text corpora, in: 2014 IEEE Conference on Visual Analytics Science and Technology (VAST). Presented at the 2014 IEEE Conference on Visual Analytics Science and Technology (VAST), pp. 173–182. doi:10.1109/VAST.2014.7042493

Amershi, S., Cakmak, M., Knox, W.B., Kulesza, T., 2014. Power to the People: The Role of Humans in Interactive Machine Learning. AI Mag.

Amershi, S., Fogarty, J., Kapoor, A., Tan, D., 2011. Effective End-User Interaction with Machine Learning, in: Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence. pp. 1529–1532. doi:10.1145/2046396.2046416

Amershi, S., Fogarty, J., Weld, D., 2012. Regroup: Interactive Machine Learning for On-demand Group Creation in Social Networks, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '12. ACM, New York, NY, USA, pp. 21–30. doi:10.1145/2207676.2207680

Andrzejewski, D., Zhu, X., Craven, M., 2009. Incorporating Domain Knowledge into Topic Modeling via Dirichlet Forest Priors, in: Proceedings of the 26th Annual International Conference on Machine Learning, ICML '09. ACM, New York, NY, USA, pp. 25–32. doi:10.1145/1553374.1553378

Bakharia, A., Bruza, P., Watters, J., Narayan, B., Sitbon, L., 2016. Interactive Topic Modeling for aiding Qualitative Content Analysis, in: Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval. ACM, pp. 213–222.

Basu, S., Fisher, D., Drucker, S.M., Lu, H., 2010. Assisting Users with Clustering Tasks by Combining Metric Learning and Classification, in: Proceedings of the Twenty-Fourth Conference on Artificial Intelligence (AAAI 2010). American Association for Artificial Intelligence.

Blei, D.M., 2012. Probabilistic Topic Models. Commun ACM 55, 77–84. doi:10.1145/2133806.2133826

Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022. doi:10.1162/jmlr.2003.3.4-5.993

Bouma, G., 2009. Normalized (pointwise) mutual information in collocation extraction, in: From Form to Meaning: Processing Texts Automatically, Proceedings of the Biennial GSCL Conference 2009. Tübingen, pp. 31–40.

Boyd-Graber, J.L., Mimno, D., Newman, D., 2014. Care and Feeding of Topic Models, in: Handbook of Mixed Membership Models and Their Applications, Chapman & Hall/CRC Handbooks of Modern Statistical Methods. Chapman and Hall/CRC, pp. 225–254.

Braun, V., Clarke, V., 2006. Using thematic analysis in psychology. Qual. Res. Psychol. 3, 77–101. doi:10.1191/1478088706qp063oa

Caruana, R., Elhawary, M., Nguyen, N., Smith, C., 2006. Meta clustering, in: Sixth International Conference on Data Mining (ICDM'06). IEEE, pp. 107–118.

Chaney, A.J.-B., Blei, D.M., 2012. Visualizing Topic Models, in: Proceedings of the International Conference on Weblogs and Social Media. pp. 419--422.

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J.L., Blei, D.M., 2009. Reading Tea Leaves: How Humans Interpret Topic Models. Adv. Neural Inf. Process. Syst. 288–296.

Chemudugunta, C., Smyth, P., Steyvers, M., 2008. Combining Concept Hierarchies and Statistical Topic Models, in: Proceedings of the 17th ACM Conference on Information and Knowledge Management, CIKM '08. ACM, New York, NY, USA, pp. 1469–1470. doi:10.1145/1458082.1458337

Choo, J., Lee, C., Reddy, C.K., Park, H., 2013. UTOPIAN: User-driven topic modeling based on interactive nonnegative matrix factorization. IEEE Trans. Vis. Comput. Graph. 19, 1992–2001. doi:10.1109/TVCG.2013.212

Chuang, J., Gupta, S., Manning, C.D., Heer, J., 2013a. Topic Model Diagnostics: Assessing Domain Relevance via Topical Alignment, in: Proceedings of the 30th International Conference on Machine Learning (ICML-13). pp. 612–620.

Chuang, J., Hu, Y., Jin, A., Wilkerson, J.D., McFarland, D.A., Manning, C.D., Heer, J., 2013b. Document Exploration with Topic Modeling: Designing Interactive Visualizations to Support Effective Analysis Workflows. Conf. Neural Inf. Process. Syst. NIPS Workshop Top. Models Comput. Appl. Eval. 1--4.

Chuang, J., Manning, C.D., Heer, J., 2012. Termite: Visualization Techniques for Assessing Textual Topic Models, in: Proceedings of the International Working Conference on Advanced Visual Interfaces, AVI '12. ACM, New York, NY, USA, pp. 74–77. doi:10.1145/2254556.2254572

Chuang, J., Roberts, M.E., Stewart, B.M., Weiss, R., Tingley, D., Grimmer, J., Heer, J., 2015. TopicCheck: Interactive Alignment for Assessing Topic Model Stability. Work. Pap.

Chuang, J., Wilkerson, J.D., Weiss, R., Tingley, D., Stewart, B.M., Roberts, M.E., Poursabzi-Sangdeh, F., Grimmer, J., Findlater, L., Boyd-Graber, J., Heer, J., 2014. Computer-Assisted Content Analysis: Topic Models for Exploring Multiple Subjective Interpretations, in: NIPS Workshop on Human-Propelled Machine Learning. Montreal, Canada.

Cohn, D., Caruana, R., McCallum, A., 2003. Semi-supervised clustering with user feedback. Constrained Clust. Adv. Algorithms Theory Appl. 4, 17–32.

Crossno, P.J., Wilson, A.T., Shead, T.M., Dunlavy, D.M., 2011. TopicView: Visually Comparing Topic Models of Text Collections, in: 2011 23rd IEEE International Conference on Tools with Artificial Intelligence (ICTAI). Presented at the 2011 23rd IEEE International Conference on Tools with Artificial Intelligence (ICTAI), pp. 936–943. doi:10.1109/ICTAI.2011.162

Daumé, H., III, 2009. Markov Random Topic Fields, in: Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, ACLShort '09. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 293–296.

Drucker, S.M., Fisher, D., Basu, S., 2011. Helping Users Sort Faster with Adaptive Machine Learning Recommendations, in: Proceedings of the 13th IFIP TC 13 International Conference on Human-Computer Interaction - Volume Part III, INTERACT'11. Springer-Verlag, Berlin, Heidelberg, pp. 187–203.

Eisenstein, J., Chau, D.H., Kittur, A., Xing, E., 2012a. TopicViz: interactive topic exploration in document collections, in: Extended Abstracts of the ACM Conference on Human Factors in Computing Systems. ACM, pp. 2177–2182.

Eisenstein, J., Chau, D.H., Kittur, A., Xing, E., 2012b. TopicViz: Interactive topic exploration in document collections, in: Proceedings of the 2012 ACM Annual Conference Extended Abstracts on Human Factors in Computing Systems Extended Abstracts. pp. 2177–2182. doi:10.1145/2212776.2223772

English, J., Hearst, M., Sinha, R., Swearingen, K., Yee, K.-P., 2002. Flexible Search and Navigation Using Faceted Metadata.

Fails, J.A., Olsen, D.R., 2003. Interactive machine learning. Proc. 8th Int. Conf. Intell. User Interfaces IUI 03 39–45. doi:10.1145/604045.604056

Fiebrink, R., Trueman, D., Cook, P.R., 2009. A metainstrument for interactive, on-the-fly machine learning, in: Proceedings of New Interfaces for Musical Expression (NIME). Presented at the NIME, pp. 280–285.

Fortuna, B., Grobelnik, M., Mladenic, D., 2005. Visualization of text document corpus. Informatica 29, 497–502.

Gardner, M.J., Lutes, J., Lund, J., Hansen, J., Walker, D., Ringger, E., Seppi, K., 2010. The Topic Browser: An Interactive Tool for Browsing Topic Models. Proc. Workshop Chall. Data Vis. Held Conjunction 24th Annu. Conf. Neural Inf. Process. Syst. NIPS 2010 1–9.

Gretarsson, B., O'Donovan, J., Bostandjiev, S., Höllerer, T., Asuncion, A., Newman, D., Smyth, P., 2012. TopicNets: Visual Analysis of Large Text Corpora with Topic Modeling. ACM Trans Intell Syst Technol 3, 23:1–23:26. doi:10.1145/2089094.2089099

Griffiths, T.L., Tenenbaum, J.B., 2004. Hierarchical topic models and the nested Chinese restaurant process. Adv. Neural Inf. Process. Syst. 16, 17.

Haghighi, A., Vanderwende, L., 2009. Exploring Content Models for Multi-Document Summarization, in: Proceedings of HLT-NAACL 2009. Presented at the NAACL '09, Association for Computational Linguistics, pp. 362--370.

Hearst, M.A., 2006. Clustering Versus Faceted Categories for Information Exploration. Commun ACM 49, 59–61. doi:10.1145/1121949.1121983

Höök, K., 2000. Steps to take before Intelligent User Interfaces become real. Interact. Comput. 12, 409--426. doi:doi:10.1016/S0953-5438(99)00006-5

Hoque, E., Carenini, G., 2015. ConVisIT: Interactive Topic Modeling for Exploring Asynchronous Online Conversations, in: Proceedings of the 20th International Conference on Intelligent User Interfaces, IUI '15. ACM, New York, NY, USA, pp. 169–180. doi:10.1145/2678025.2701370

Horvitz, E., 1999. Principles of mixed-initiative user interfaces, in: CHI '99. ACM, New York, NY, USA, pp. 159–166. doi:10.1145/302979.303030

Hotho, A., Nürnberger, A., Paaß, G., 2005. A brief survey of text mining. LDV Forum - GLDV J. Comput. Linguist. Lang. Technol. 20, 19--62.

Hruschka, D.J., Schwartz, D., St.John, D.C., Picone-Decaro, E., Jenkins, R.A., Carey, J.W., 2004. Reliability in Coding Open-Ended Data: Lessons Learned from HIV Behavioral Research. Field Methods 16, 307–331. doi:10.1177/1525822X04266540

Hu, Y., Boyd-Graber, J., Satinoff, B., Smith, A., 2014. Interactive Topic Modeling. Mach Learn 95, 423–469. doi:10.1007/s10994-013-5413-0

Hu, Y., Milios, E.E., Blustein, J., Liu, S., 2012. Personalized Document Clustering with Dual Supervision, in: Proceedings of the 2012 ACM Symposium on Document Engineering, DocEng '12. ACM, New York, NY, USA, pp. 161–170. doi:10.1145/2361354.2361393

Hulu, 2011. Hulu's Recommendation System « Hulu Tech Blog [WWW Document]. URL http://tech.hulu.com/blog/2011/09/19/recommendation-system/ (accessed 9.14.15).

Kim, B., Patel, K., Rostamizadeh, A., Shah, J., 2015. Scalable and interpretable data representation for high-dimensional, complex data, in: Twenty-Ninth AAAI Conference on Artificial Intelligence. pp. 1763--1769.

Lau, J.H., Grieser, K., Newman, D., Baldwin, T., 2011. Automatic Labelling of Topic Models, in: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1, HLT '11. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 1536–1545.

Lau, J.H., Newman, D., Baldwin, T., 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. Proc. Assoc. Comput. Linguist. 530–539.

Lee, H., Kihm, J., Choo, J., Stasko, J., Park, H., 2012. iVisClustering: An Interactive Visual Document Clustering via Topic Modeling. Comput. Graph. Forum 31, 1155–1164. doi:10.1111/j.1467-8659.2012.03108.x

Malik, S., Smith, A., Hawes, T., Papadatos, P., Li, J., Dunne, C., Shneiderman, B., 2013. TopicFlow: Visualizing Topic Alignment of Twitter Data over Time, in: Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM '13. ACM, New York, NY, USA, pp. 720–726. doi:10.1145/2492517.2492639

McCallum, A.K., 2002. MALLET: A Machine Learning for Language Toolkit.

McFarland, D.A., Ramage, D., Chuang, J., Heer, J., Manning, C.D., Jurafsky, D., 2013. Differentiating language usage through topic models. Poetics, Topic Models and the Cultural Sciences 41, 607–625. doi:10.1016/j.poetic.2013.06.004

Mimno, D., Wallach, H.M., Talley, E., Leenders, M., McCallum, A., 2011. Optimizing Semantic Coherence in Topic Models. Proc. 2011 Conf. Empir. Methods Nat. Lang. Process. 262–272.

Mühlbacher, T., Piringer, H., Gratzl, S., Sedlmair, M., Streit, M., 2014. Opening the Black Box: Strategies for Increased User Involvement in Existing Algorithm Implementations. IEEE Trans. Vis. Comput. Graph. 20, 1643–1652. doi:10.1109/TVCG.2014.2346578

Newman, D., Lau, J., Grieser, K., Baldwin, T., 2010. Automatic evaluation of topic coherence. … Lang. Technol. … 100–108.

Ramage, D., Hall, D., Nallapati, R., Manning, C.D., 2009. Labeled LDA: A Supervised Topic Model for Credit Attribution in Multi-labeled Corpora, in: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1, EMNLP '09. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 248–256.

Sandhaus, E., 2008. The New York Times Annotated Corpus. Linguistic Data Consortium, Philadelphia.

Šilić, A., Bašić, B.D., 2010. Visualization of Text Streams: A Survey, in: Setchi, R., Jordanov, I., Howlett, R.J., Jain, L.C. (Eds.), Knowledge-Based and Intelligent Information and Engineering Systems, Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 31–43.

Smith, A., Lee, T., Poursabzi-Sangdeh, F., Boyd-Graber, J., Elmqvist, N., Findlater, L., 2017. Evaluating Visual Representations for Topic Understanding and Their Effects on Manually Generated Labels. Trans. Assoc. Comput. Linguist. 5, 1–15.

Smith, A., Lee, T.Y., Poursabzi-Sangdeh, F., Findlater, L., Boyd-Graber, J., Elmqvist, N., 2016. Evaluating Visual Representations for Topic Understanding and Their Effects on Manually Generated Labels. Trans. Assoc. Comput. Linguist.

Stevens, K., Kegelmeyer, P., Andrzejewski, D., Buttler, D., 2012. Exploring Topic Coherence over Many Models and Many Topics, in: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLP-CoNLL '12. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 952–961.

Stumpf, S., Rajaram, V., Li, L., Burnett, M., Dietterich, T., Sullivan, E., Drummond, R., Herlocker, J., 2007. Toward harnessing user feedback for machine learning. Presented at the Proceedings of the 12th international conference on Intelligent user interfaces, ACM, pp. 82–91. doi:10.1145/1216295.1216316

Titov, I., McDonald, R., 2008. A joint model of text and aspect ratings for sentiment summarization. ACL 8, 308--316.

Wallach, H.M., 2006. Topic Modeling: Beyond Bag-of-words, in: Proceedings of the 23rd International Conference on Machine Learning, ICML '06. ACM, New York, NY, USA, pp. 977–984. doi:10.1145/1143844.1143967

Wallach, H.M., Murray, I., Salakhutdinov, R., Mimno, D., 2009. Evaluation Methods for Topic Models, in: Proceedings of the 26th Annual International Conference on Machine Learning, ICML '09. ACM, New York, NY, USA, pp. 1105–1112. doi:10.1145/1553374.1553515

Zhai, H., Guo, J., Wu, Q., Cheng, X., Sheng, H., Zhang, J., 2009. Query Classification Based on Regularized Correlated Topic Model, in: IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technologies, 2009. WI-IAT '09. Presented at the IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technologies, 2009. WI-IAT '09, pp. 552–555. doi:10.1109/WI-IAT.2009.91