# Document Exploration with Topic Modeling: Designing Interactive Visualizations to Support Effective Analysis Workflows

**Jason Chuang**
Computer Science and Engineering
University of Washington
jcchuang@cs.washington.edu

**Yuening Hu**
Department of Computer Science
University of Maryland, College Park
ynhu@cs.umd.edu

**Ashley Jin**
Department of Computer Science
Stanford University
ashpjin@cs.stanford.edu

**John D. Wilkerson**
Department of Political Science
University of Washington
jwilker@u.washington.edu

**Daniel A. McFarland**
Graduate School of Education
Stanford University
mcfarland@stanford.edu

**Christopher D. Manning**
Department of Computer Science
Stanford University
manning@cs.stanford.edu

**Jeffrey Heer**
Computer Science and Engineering
University of Washington
jheer@uw.edu

## Abstract

We present a series of published and ongoing projects, on developing interactive visualizations for document exploration using statistical topic models. Taking a user-centric approach, we engage analysts working with real-world datasets, identify bottlenecks in their analysis workflow, and respond by developing appropriate visual presentations, interactions, and algorithms to address the users' needs.

We first review visual analysis tools produced by the authors of this paper. While these tools have been previously published in various venues, here we summarize their collective contributions, report on user feedbacks received since the deployment of the tools, and identity potential areas of improvement. We then present various visualization components currently under development, and report on informal feedbacks received from early prototypes. Effective visual designs often need to be paired with appropriate algorithmic development. We highlight potential areas of collaboration between machine learning and visualization researchers, to address the present barriers to analysis. Finally, we outline how a completed system may enable evaluations of topic modeling based on task performance.

## 1 Introduction

The growing amount of digitalized text data provides us with an opportunity to study human languages and communications at an unprecedented scale. Statistical topic models can aid the exploration of massive corpora by uncovering large-scale patterns in the documents. For many end users, textual context analysis is one component within a larger scientific investigation. Any observations

derived from a topic model, therefore, must stand rigorous testing to ensure the validity and stability of the underlying model. In these cases, building high-quality topic models can be time consuming and involve a significant amount of manual verification and model refinement [6, 10].

Our long-term research goal is to develop visual analysis tools that can support an efficient workflow for constructing accurate, interpretable, and verifiable topic models. We engage domain experts who examine thematic patterns in real-world text corpora using statistical topic models. We identify modeling and analysis tasks that are difficult to accomplish, and respond by selecting or developing appropriate visual presentations, interactions, and algorithms. We would like to take the opportunity at this workshop to present our vision of a completed tool, elicit feedback from other participants on analysis needs that they encounter in their respective domains, and seek feedback on the ongoing and planned developments for our tool.

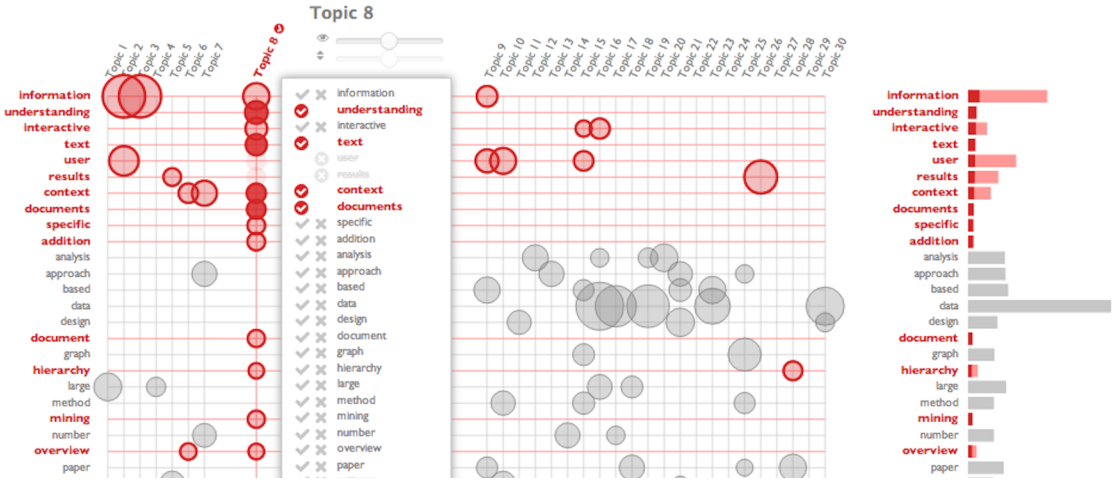


Figure 1: Interactive topic modeling user interface

## 2 Analysis Workflow and Interactive Visualization Components

In our experiences [4], experts often build a sequence of topic models in order to improve upon and determine the quality of their models. Subtasks in their workflow include assessing model outputs, comparing the quality of multiple models, and documenting their work history. Iterative, interactive, and direct refinement of an existing model is strongly preferred over indirect options such as text preprocessing (e.g., exclusion of stopwords and inclusion of phrases), parameter and hyper-parameter search. However, as model quality is often critical to the success of an analysis, domain experts will spend a significant amount of efforts applying indirect model modification if direct options are exhausted — and they often do. Maintaining a work history enables the users to experiment with the effects of potential model modification and backtrack if necessary. A complete documentation of the modeling process also allows external parties to scrutinize the choice of the final model and may help demonstrate the validity of an analysis.

### 2.1 Assessing the Quality of a Statistical Topic Model

We previously designed Termite [3], a visual analysis tool for inspecting the term-topic associations for a single topic model. The tool incorporates a matrix view to increase the information density that can be presented on a given computer display, so that users can visually inspect a greater number of topics and terms at once. Two technical contributions in the tool include a *term saliency* measure for identifying distinct terms and a *term seriation* algorithm that arrange terms to reveal grouping of related words and preserve phrases to aid interpretation. After the release of the software, the visualization has since been incorporated as a component into other text analysis tools [8].

In informal observation sessions, users working with Termite suggested that term seriation was the key component that helped them perceive patterns among the latent topics. The current seriation algorithm, however, takes into account word associations across the whole corpus. As users drill down to individual topics, they often re-arrange terms based on word frequency within the topic of

interest and lose the advantages afforded by the seriation algorithm (grouping and reading ordering). Presently, we are looking into modifications to the algorithm to enable topic-specific word ordering.

A unresolved research question regarding the seriation algorithm concerns *why* users find the grouping patterns useful. Our seriation algorithm analyzes the co-occurrence of words within the same documents, within the same sentence, within a sliding window of $K$ adjacent words, among trigrams and/or bigrams. It comes across strange that a visualization displaying the output of a topic model should need to separately and independently analyze the content of the corpus at the word level.

Our hypothesis is that the seriation algorithm captures word associations at varying levels of granularity, and that such patterns can help users recognize potential connections between words at multiple resolutions. If so, a principled approach to improving Termite would be to incorporate *previews* of models that can account for word associations under different conditions. Fast inference algorithms (e.g., Arora et al. [1]) that can be tailored toward a specific subset of terms (e.g., by appropriate choice of anchor words) might be suitable for generating multiple topic-specific previews.

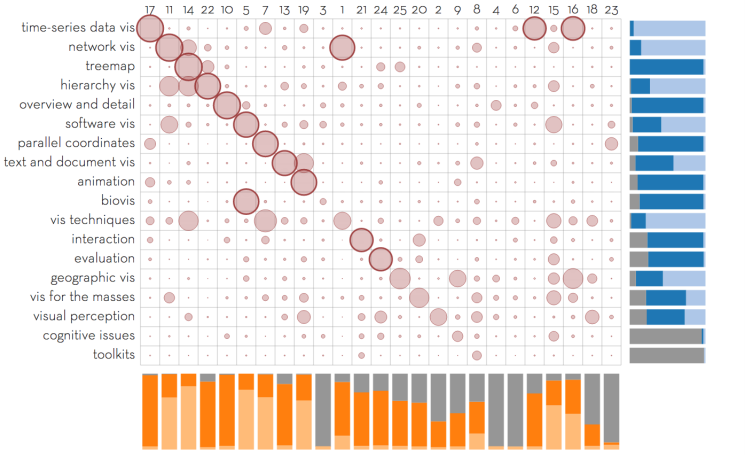


Figure 2: Correspondence chart: Alignment between two sets of latent topics

### 2.2 Model Diagnostics and Comparisons across the Modeling Process

We previously devised a model diagnostics framework [2] for examining the alignment among two set of topics (Figure 2). While the original paper demonstrate the framework by comparing statistical topic model outputs to a set of expert-generated concepts, the framework is sufficiently general to allow for model-to-model comparisons. Our intent is to help users evaluate changes to their models between successive iterations in the modeling process. In one ongoing project, our collaborating domain experts needed to align models comprised of over 4,000 latent topics trained using different parameters; they responded positively to the use of the framework in the task.

We are currently developing ways to help experts record a history of all models constructed during their analysis. Such a *lab notebook* provides the experts with interactive tools to quickly review their work and gain insights about the effects of model refinement. In addition, while the domain experts value of insights gained from statistical topic models, they also express concerns that results obtained from qualitative analysis methods can be difficult to replicate, a necessary criteria for scientific investigations and academic publications alike. A detailed notebook allows their work to be scrutinized in greater details and may help validate their choice of a topic model.

We also observe that experts often drilled down to individual topics as a means to make sense of multiple models. In particular, an often requested feature is to track a "good" topic across all models to identify whether they persist across the models.

### 2.3 Iterative, Interactive, and Direct Model Refinement

More recently, we started incorporating interactive topic modeling developed by Hu et al. [7] into our analysis tools. The tree-based inference algorithm by Hu et al. can directly incorporate user

inputs expressed as word level constraints in terms of both positive (e.g., association of words) and negative (e.g., disassociation of words) affinities. The technique can refine an existing model instead of training a new set of topics. Users interact with the model by directly clicking on words in the visualization to promote or demote a word's usage in a topic (Figure 1).

Analysts working with our earlier prototypes find the iterative modeling process intuitive and response time acceptable (typically under 30 seconds for a corpus ranging from 300 to 20,000 documents). While the analysts find word-level constraints useful for expressing and refining some concepts, we receive the feedback for wanting more *topic*-level operations consistent with earlier observations that experts consider "topics" as meaningful units of analysis for evaluating a model.

A solution may be reached from either visualization or algorithm development. For example, one might design visual operations that translate a user's intentions (e.g., combining two topics) into word level constraints for the underlying inference algorithm. An open research question concerns operations that cannot be expressed (e.g., splitting a topic into two, increasing/decreasing the number of latent topics as a means to reduce/increase the level of overall word aggregation).

In addition, it is unclear what the *minimum set of operations* are needed in order for experts to express the necessary domain knowledge to construct a high-quality model. As pointed out earlier, if direct manipulation of a model does not provide sufficient degrees of freedom to express domain knowledge or result in sufficiently good models, users will bypass the interactive interface and resort to indirect means to *try* to alter the output of a model.

## 3 Evaluations

Human judgment of the quality of a topic can be subjective, as people's interpretations of topical concepts can vary according to expertise and the context of analysis. Our goal is to develop a tool that can support the entire modeling process, so that users can produce a high-quality model for use in subsequent analysis tasks. Having such a tool may enable new means of evaluating the value of statistical topic modeling. As a point of comparison, social scientists often apply a grounded approach to codify large text corpora [5, 9] in order to identify recurring thematic patterns. We may be able to measure the cost of performing the final analysis task and the quality of the analysis results, based on codes identified through established grounded theory vs. those based on topics belonging to a high-quality model constructed using our tool.

## References


[1] Sanjeev Arora, Rong Ge, Yoni Halpern, David M. Mimno, Ankur Moitra, David Sontag, Yichen Wu, and Michael Zhu. A practical algorithm for topic modeling with provable guarantees. In *ICML*, 2013.

[2] Jason Chuang, Sonal Gupta, Christopher D. Manning, and Jeffrey Heer. Topic model diagnostics: Assessing domain relevance via topical alignment. In *ICML*, 2013.

[3] Jason Chuang, Christopher D. Manning, and Jeffrey Heer. Termite: Visualization techniques for assessing textual topic models. In *AVI*, 2012.

[4] Jason Chuang, Daniel Ramage, Daniel A. McFarland, Christopher D. Manning, and Jeffrey Heer. Large-scale examination of academic publications using statistical models. In *AVI Workshop*, 2012.

[5] Laura A. Dabbish, H. Colleen Stuart, Jason Tsay, and James D. Herbsleb. Social coding in github: transparency and collaboration in an open software repository. In *CSCW*, 2012.

[6] David Hall, Daniel Jurafsky, and Christopher D Manning. Studying the history of ideas using topic models. In *EMNLP*, 2008.

[7] Yuening Hu, Jordan Boyd-Graber, Brianna Satinoff, and Alison Smith. Interactive topic modeling. *Machine Learning*, 2013.

[8] David Kale, Samuel Di, Yan Liu, and Yolanda Gil. Capturing data analysis expertise with visualization in workflows. In *AAAI Fall Symposium*, 2013.

[9] Taylor Jackson Scott, Katie Kuksenok, Daniel Perry, Michael Brooks, Ona Anicello, and Cecilia Aragon. Adapting grounded theory to construct a taxonomy of affect in collaborative online chat. In *SIGDOC*.

[10] Edmund M. Talley, David Newman, David Mimno, Bruce William Herr II, Hanna M. Wallach, Gully A P C Burns, A G Miriam Leenders, and Andrew McCallum. Database of NIH grants using machine-learned categories and graphical clustering. *Nature Methods*, 2011.