Visualization Techniques for Topic Model Checking

Jaimie Murdock and **Colin Allen**
Program in Cognitive Science
Indiana University
{jammurdo,colallen}@indiana.edu

Abstract

Topic models remain a black box both for modelers and for end users in many respects. From the modelers' perspective, many decisions must be made which lack clear rationales and whose interactions are unclear – for example, how many topics the algorithms should find (K), which words to ignore (aka the "stop list"), and whether it is adequate to run the modeling process once or multiple times, producing different results due to the algorithms that approximate the Bayesian priors. Furthermore, the results of different parameter settings are hard to analyze, summarize, and visualize, making model comparison difficult. From the end users' perspective, it is hard to understand why the models perform as they do, and information-theoretic similarity measures do not fully align with humanistic interpretation of the topics. We present the Topic Explorer, which advances the state-of-the-art in topic model visualization for document-document and topic-document relations. It brings topic models to life in a way that fosters deep understanding of both corpus and models, allowing users to generate interpretive hypotheses and to suggest further experiments. Such tools are an essential step toward assessing whether topic modeling is a suitable technique for AI and cognitive modeling applications.

1 Introduction

Topic modeling using Latent Dirichlet Allocation (LDA – Blei, Jordan, and Ng (2003)) represents the current state of the art for extraction of meaningful data from digitized texts. LDA topic models are intriguing because they don't simply "count words", but they treat all the documents in a corpus as different mixtures of some number of topics. Each computed "topic" is an unlabeled probability distribution over the words in the corpus. However, the question of determining the interpretability of the models – which Blei (2012) labels as the model checking problem – is among the most significant open issues facing topic modelers.

One way of addressing the model checking problem is through interactive visualization supporting rapid experimentation for interpretive hypotheses. LDA visualizations manage interactions among three entity types: *topics*, *documents*, and *words*. We have developed the Topic Explorer to navigate topic-document and document-document relations (Figure 1), with words exposed as elements of the topics.

The Topic Explorer contrasts with earlier approaches. Termite (Chuang, Manning, and Heer 2012) is a heatmap for corpus-wide topic-word distributions, without document interactivity. The topic navigator of Chaney and Blei (2012) enables document interactivity, but does not show comparative topic distribution among documents. TopicNets (Gretarsson et al. 2012) uses dimensionality reduction on topic composition to plot documents in a 2D space, but does not show topic or document composition. LDAvis (Sievert and Shirley 2014) directly addresses the model checking problem by aiding topic interpretation through a relevance method for ranking terms within topics for display to the end user. It expands upon the 2D space of TopicNets, displaying topic-word and topic-topic relationships alongside composition information. In contrast to earlier approaches, the Topic Explorer allows users to interact with topic-document and document-document space, while keeping comparative topic distribution and document composition visible.

2 Topic Model Checking

Indirect Comparison — By using multiple windows of the Topic Explorer with different numbers of topics, the similarity space can be compared across different models. For example, by examining the list of similar documents in the 20 topic model and the 40 topic model (Figure 1), one can investigate how coarse-graining affects the topic space.

"Junk" Topics — The search for so-called "junk" topics has been a focus for some topic model explorations (Snyder et al. 2013). This search is driven by the frequent misrepresentation of topics by their top N words rather than as a distribution of words. While the Topic Explorer also previews

This work is funded by a Seed Funding Grant from the Indiana University Office of the Vice Provost for Research (IU OVPR). Special thanks to Robert Rose, Jun Otsuka, and Doori Lee for development of the InPhO VSM framework used to generate the LDA topic models visualized here. VSM framework development funded by the National Endowment for the Humanities (NEH) Office of Digital Humanities (ODH), award HJ-50092-12.

All software described in this document is published under an open-source license at GitHub. The Topic Explorer is available at http://github.com/inpho/topic-explorer. The VSM framework is available at http://github.com/inpho/vsm. Live demos trained on a digital encyclopedia, newspaper stories, and a selection of digitized books are available at http://inphodata.cogs.indiana.edu.

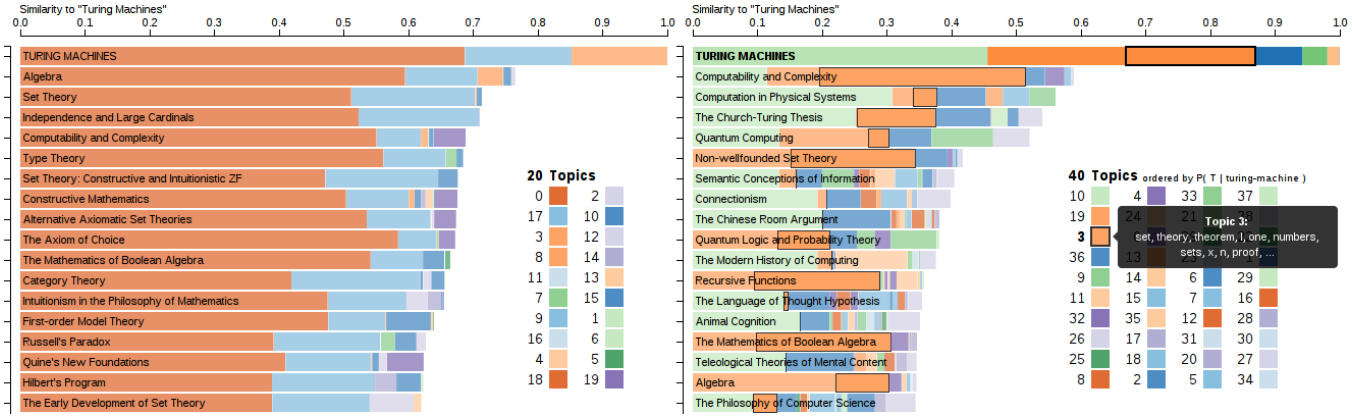

Figure 1: Screenshot of the Topic Explorer showing a 20-topic model (left) and a 40-topic model (right) centered on the Stanford Encyclopedia of Philosophy (SEP) article on Turing Machines. The color bands within each article's row show the topic distribution within that article, and the size of each band indicates the weight of that topic in the article. The combined width of each row indicates the similarity to the focal topic or document, measured by the quantity $\text{SIM}(doc) = 1 - \text{JSD}(doc, focal)$, where JSD is the Jensen-Shannon distance (Lin 1991) between the word probability distributions of each item. Hovering over a topic shows the top 10 words in that topic and highlights the distribution of that topic across selected documents. By clicking a topic, the documents will reorder according to that topic's weight and topic bars will reorder according to the topic weights in the highest weighted document. When a topic is selected, clicking "Top Documents for [Topic]" will navigate to a new page showing the most similar documents to that topic's word distribution. By normalizing topics, the combined width of each bar expands so that topic weights per document can be compared. Additionally, users may select among different topic models using different values for K, the number of topics. Each topic's label and color are arbitrarily assigned, but are consistent across articles in the browser for each topic model.

the top N words, it can suggest hypotheses about the deeper structure of the topic through the topic-document similarity view. For instance, topics that seem initially uninterpretable are sometimes diagnostic of a sub-genre or style of writing (Hughes et al. 2012); they may also select for specialized sections of the documents, such as bibliographies; or highlight digitization errors, such as the inclusion of page numbers or section headings, or common misspellings introduced by optical character recognition (OCR). The Topic Explorer can make these salient, rendering one person's "junk" another person's treasure.

3 Conclusion

In this paper, we described the Topic Explorer, a new visualization that exposes the topic-document and document-document space of LDA topic models. It addresses the model checking problem – that is, the humanistic interpretation of topics – rather than formal topic model evaluation. Given that multiple models yield different but interpretable results, future attempts to exploit LDA for AI or cognitive modeling purposes would be well advised to consider simultaneously modeling multiple levels of K.

References

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3:993–1022.

Blei, D. M. 2012. Probabilistic Topic Models. *Communications of the ACM* 55(4):77–84.

Chaney, A. J.-B., and Blei, D. M. 2012. Visualizing Topic Models. In *International AAAI Conference on Social Media and Weblogs*.

Chuang, J.; Manning, C. D.; and Heer, J. 2012. Termite: Visualization Techniques for Assessing Textual Topic Models. In *Proceedings of the International Working Conference on Advanced Visual Interfaces*, 74–77. ACM.

Gretarsson, B.; Odonovan, J.; Bostandjiev, S.; Höllerer, T.; Asuncion, A.; Newman, D.; and Smyth, P. 2012. TopicNets: Visual Analysis of Large Text Corpora with Topic Modeling. *ACM Transactions on Intelligent Systems and Technology (TIST)* 3(2):23.

Hughes, J. M.; Foti, N. J.; Krakauer, D. C.; and Rockmore, D. N. 2012. Quantitative Patterns of Stylistic Influence in the Evolution of Literature. *Proceedings of the National Academy of Sciences* 109(20):7682–7686.

Lin, J. 1991. Divergence Measures Based on the Shannon Entropy. *IEEE Transactions on Information Theory* 37(1):145–151.

Sievert, C., and Shirley, K. E. 2014. LDAvis: A Method for Visualizing and Interpreting Topics. *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces* 63–70.

Snyder, J.; Knowles, R.; Dredze, M.; Gormley, M.; and Wolfe, T. 2013. Topic Models and Metadata for Visualizing Text Corpora. In *Proceedings of the 2013 NAACL HLT Demonstration Session*, 5–9. Atlanta, Georgia: Association for Computational Linguistics.