Towards Interactive Visualizations for Selecting Topic Model Parameters

Helping Latent Dirichlet Allocation Users Pick the Number of Topics

Elena Glassman
MIT CSAIL
elg@mit.edu

Manasi Vartak
MIT CSAIL
mvartak@mit.edu

ABSTRACT

Latent Dirichlet Allocation (LDA) is a popular strategy for inferring the underlying structure of unlabeled data. A key hyperparameter for LDA is k, the number of topics fit to the data. Empirically we find that the quality of results can be impacted significantly by the choice of k. In addition, research shows that objective metrics, e.g. perplexity, are not guaranteed to produce better or more interpretable topics; this tuning must be performed through manual inspection. While there are various tools available for understanding the results of LDA via visualization, no existing interactive tools can help the user tune k. In this paper, we describe ongoing work to build an interactive visualization tool for choosing k. Similar to existing tools, our tool visualizes topic and document distribution for a given k, but more importantly, it visualizes *changes in model structure* induced by different k. The tool can communicate the evolution of the model with varying k at the document and topic level, thus enabling the user to choose a k best suited to the task and corpus.

Keywords

topic models, user interface design, model selection

1. INTRODUCTION

Topic modeling is a Bayesian inference technique for learning the underlying structure of unlabeled data. Since its publication just over a decade ago, one of the most popular methods, Latent Dirichlet Allocation (LDA) [1], has been cited approximately 14,000 times and been applied in a variety of contexts, including medicine [13], social media analysis [10], object classification [3], digital humanities [2], and source code analysis [8]. Many toolboxes offer LDA implementations that can be used by anyone with sufficient programming knowledge and little to no knowledge of Bayesian inference.

However, LDA, and other parametric Bayesian methods like it, require the user to pre-set the number of topics to learn about the corpus. Given that the data given to LDA is often both unlabeled and too large to be reviewed in its entirety by a human, it is not clear how LDA users should go about setting the parameter that controls the number of topics. This parameter can significantly affect the subjective quality of the results. We will refer to this parameter as K. LDA_k will refer to the LDA topic model with the parameter K set to the value k.

Mathematical metrics for evaluating the quality of LDA_k are an active area of research, in part because existing measures do not necessarily correlate with metrics of human interpretability. For example, Chang et al. [4] found that a common mathematical measure of topic model quality, perplexity, was anti-correlated with a metric of topic quality based on human judgments.

Inspection, i.e., examining LDA_k by hand for the (subjectively) best k, is difficult without some vehicle for visualization of the results and how they differ as k is changed. Therefore, even if the user writes a script to sweep across different values of k and fit each LDA_k to the data, it is not clear how to select the best model at the end of the sweep.

More mathematically advance methods, i.e., non-parametric Bayesian inference methods like Hierarchical Dirichlet Processes, only partially address this limitation of parametric methods. Despite the 'non-parametric' label, they still require the user to pre-set a hyperparameter that (indirectly) controls the number of topics. The challenges of evaluating which parameter value to choose still apply.

We present on-going work on an interactive visualization tool for choosing k in the context of LDA. Similar to existing tools, our tool visualizes topic and document distribution for a given k, but more importantly, it visualizes *changes in model structure* induced by different k. The tool can communicate the evolution of the model with varying k at the document and topic level, thus enabling the user to choose a k best suited to the task and corpus.

2. RELATED WORK

While topic modeling is a powerful and versatile technique, its effectiveness is highly dependent on hyperparameters, particularly, k, the number of topics chosen to model the corpus. Given their popularity, researchers from machine learning, human-computer interaction, and visualization have begun to propose and develop systems that aid end-user understanding and manipulation of topic models.

Existing visualizations, like LDAVis [15] and Termite [7], help users inspect a single fitted LDA model. LDAVis allows users to explore topic-term relationships and gives users a global view of topics in the form of an Intertopic Distance

IDEA workshop, KDD '16 August, 2016, San Francisco, CA, USA
© 2016 ACM. ISBN 978-1-4503-2138-9.
DOI: 10.1145/1235

Map. Termite helps users assess topic model quality with a tabular layout that promotes comparison of terms both within and across latent topics. It visualizes words associated with a given topic via a heat-map-like visualization. The associated paper also proposes new heuristics to determine what topics and words should be visualized. Chuang et al. [5] propose new techniques to align inferred sets of topics. Their topic comparison framework that is general enough to support comparing topic models generated with different parameters and inference algorithms. Chuang et al. [6] outline how these visualizations and methods could be made more iterative, interactive, and and directly refined by end-users.

Frameworks and tools for how to build interactive topic models are continuing to be proposed and evaluated. For example, [11] proposes a method to allow users to iteratively refine topics by adding constraints on sets of words. Another tool, ITMViz [14], presents a specialized user interface for performing interactive topic modeling on source code with an emphasis on using experts to refine topics.

In an abstract for the recent Human Centered Machine Learning workshop at the ACM CHI conference, Smith et. al. [16] argue that these existing tools and algorithms for helping end-users tune topic models are not sufficiently rooted in user needs. They enumerate several challenges and associated desiderata for future tools, including reducing algorithm latency, communicating model changes effectively, and actively supporting the user via recommendations. Our work specifically addresses the second requirement, i.e., *effectively communicating model changes to the user in the face of varying model parameters*.

Our work draws inspiration from recent work on TopicFlow [17] a visual analysis tool to trace the evolution of Twitter (and other) topics over time. TopicFlow uses a cosine similarity metric to identify similarities in topics that are then visualized via a Sankey diagram.

3. DESIGN GOALS

We have three main design goals with this tool: (i) effectively visualize topics for a given k, (ii) communicate model evolution as the number of topics changes, and (iii) highlight topic- and document-level changes that occur as k changes. The first goal is necessary for the user to understand the results of LDA and has been explored in the literature as described above. However, the two remaining problems of communicating model evolution and highlighting changes in topics have received limited attention, and these are the problems that motivate our work.

We find that there are primarily two ways to visualize changes to the LDA model structure with changing k. The first is a *document-centric* visualization that emphasizes the document as the key abstraction and visualizes the changes in topics per document. The second visualization, in contrast, is *topic-centric* and emphasizes changes in topics as a means to understand model evolution. As we will see in the following sections, both types of visualizations convey distinct views of the results to the user and can help build a more complete picture of the LDA model.

Since this work is ongoing, we present UI and interaction designs for our tool in the following sections. We ensure that our designs reflect real use cases by using results obtained by applying LDA to a real dataset. Specifically, we obtained results for LDA on a small sample (100) of research paper titles from NIPS, a top machine learning conference. We swept the values of k between 2 and 10, and for each k produced results including the set of topics, distribution of words per topic, and distribution of topics per document. These results inform the UI designs presented in the next two sections.

4. DOCUMENT-CENTRIC VIEW

While LDA does not strictly classify documents, the distribution of topics inferred to be present in each document is a kind of classification that can be very useful, in addition to the topics themselves. The document-centric visualization, shown in Figure 1, is intended to communicate how the topics associated with each document change, given the topics inferred by LDA_k.

The top-most subfigure in the document-centric view shows an automated evaluation metric of topic quality called ”topic coherence” [12], calculated for each set of topics generated by LDA_k for $k = 1$ through 10. Topic coherence is defined for each topic in each set of topics, so for each k, there is a distribution of topic coherence measurements represented in a box-plot. For the NIPS dataset, the average topic coherence increases nearly monotonically with the number of topics over the range of k shown. This is one of a variety of possible metrics of topic quality, which may or may not be a reasonable substitute for the end-user's judgment for a given corpus or task. This subfigure is intended to help the user get some intuition about the relationship between an automated measure of topic quality and their own judgment about the quality of topic model fit.

The second part of the document-centric view is composed of a series of LDA models visualized, one for each k. Each LDA_k representation includes (1) topics represented by a distribution over words (2) documents represented by columns of stacked bars (3) example documents suggested by the system or selected by the user. The topics can be ordered from left to right by topic coherence, and mapped to grayscale values between white and black. The length of each bar that stacks together to represent an individual document represents how strongly the topic with that bar's color is associated with that document. In this representation, documents are ordered to to minimize the difference in topic proportions between adjacent documents.

Some documents' topic distributions will change significantly across different LDA_k while others will stay largely the same. Using the topic alignment methods in [5], the system could call out to the user, by highlighting the same document across all LDA_k models, documents that expose what is changing as k changes.

5. TOPIC-CENTRIC VIEW

As compared to the document-centric view that visualizes results at the granularity of individual documents, the topic-centric view, shown in Figure 2, provides an overview of topics and visualizes their evolution as a function of k. Similar to [17], we use a Sankey diagram to capture the evolution of topics. Nodes represent topics; edges represent the *flow* of documents as they transition from being strongly associated with one topic to another. As is typical in Sankey diagrams, the breadth of connections between different nodes is proportional to the flow between the topics, i.e., the number of documents strongly associated with both topics. By default, topics are ordered to minimize crossing of flows; however, for

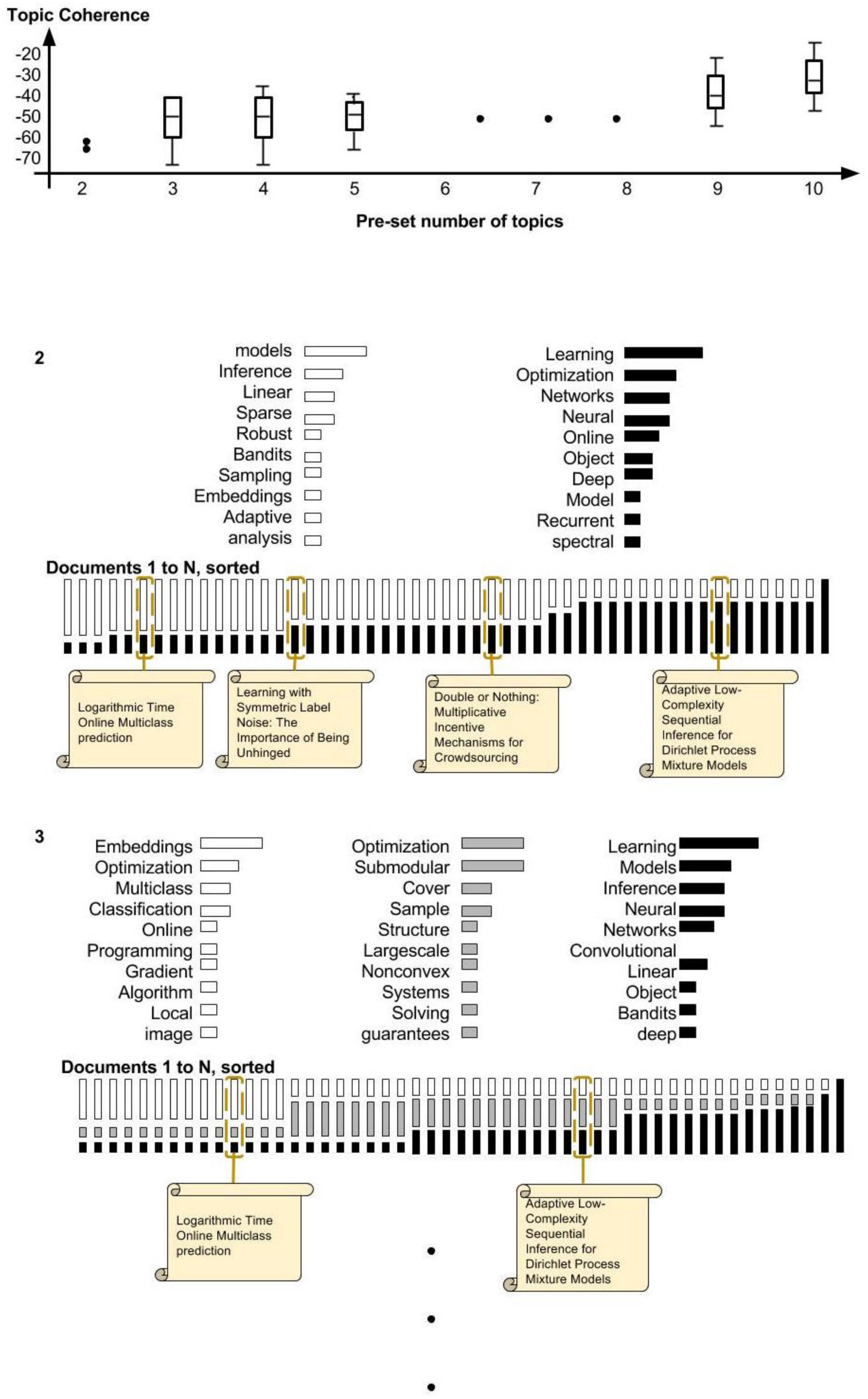

Figure 1: Representing documents and their associated topics as a function of pre-set number of topics.

ease of understanding, the topics can also be ordered by the number of documents strongly associated with a topic.

Clicking on a node belonging to a particular topic, say Topic 1 (k=2), brings up the list of top words for the topic as well as a set of representative documents for it (see Figure 2.b). For example, clicking Topic 1 (k=2) produces the words including ”learning”, ”optimization”, ”neural” and ”networks”. One can infer that this topic refers broadly to papers related to neural networks and optimization. To further aid understanding of this topic, the tool shows representative documents that have a high proportion of Topic 1 (e.g. ”Deep learning with Elastic Averaging SGD”, ”Bidirectional Recurrent Neural Networks as Generative Models”).

To understand the evolution of a particular topic between two values of k, the user can click on the edge connecting the two topic nodes. For example, consider the edge connecting Topic 1 (k=2) to Topic 1 (k=3). This action brings up two types of information about the evolution of Topic 1 between k=2 and 3 (see Figure 2). First, the visualization shows differences in the two topics at the *word level* by visualizing changes to coherence of different words. For example, we find that the coherence of the word ”deep” reduces from k=2 to k=3 while that of ”optimization” increases, and ”robust” is a new word introduced in k=3. Second, it shows documents that moved from the old topic to the new one (once again, these are documents that had a high proportion of the respective topics).

We envision the topic-centric visualization to be used as follows: the user runs LDA for varying values of k and feeds the results to our tool. They then use the topic-centric view to examine topics, and the documents most strongly associated with them, to find the best number of topics for their application. Note that the ability to provide feedback on topics (e.g., that a particular word doesn't belong to a particular topic) is important but orthogonal to our primary goal. We leave the question of incorporating feedback for future work.

6. DISCUSSION

The document-centric and topic-centric modes of visualization represent two distinct means for the user to consume the results of LDA. The document-centric visualization allows the user to understand how the topics impact individual examples in the dataset and how the composition of individual examples changes with changing topics. In contrast, the topic-centric model provides a coarse-grained view of the documents and helps the user understand the abstract topics present in the corpus.

While each offers distinct advantages, they also have certain drawbacks. The document-centric model can be easier for the user to consume in the case where individual topics are too difficult to coherently identify. For example, when identifying design patterns in student-written Python programs by running topic modeling on features of their execution on teachers' test cases [9], the topics can be difficult to directly represent or visualize. However, the document-centric visualization cannot easily scale to thousands of documents. The tool must perform downsampling or support search in order to support large corpora. In addition, visualizing the composition of a document in the presence of many topics can be challenging.

The topic-centric visualization on the other hand, scales better with large corpora since the number of topics is typically much smaller than the number of documents. Moreover, the representative words and documents can aid the user in understanding topics. The challenge with the topcentric view arises from the fact that it is difficult for a user to a Sankey diagram with hundreds of nodes. Moreover, since a document is composed of many different topics, assigning a document to a topic (in the Sankey diagram) has the potential to confuse the reader.

We believe that both visualizations can provide the user mutually independent insights about the impact of k on the resulting topics, and that a hybrid system capturing the strengths of both views would best aid the user in choosing an appropriate k. In ongoing work, we plan to conduct a controlled user study to compare the efficacy of both visualizations and design an effective hybrid visualization. We also hope to explore how many other topic models' parameters and hyperparameters, such as the concentration parameter of the non-parametric Hierarchical Dirichlet Process (HDP) model, can be tuned with this visualization.

7. SUMMARY

Topic models, like LDA, are a popular statistical technique for inferring structure within unlabeled data. Empirically, model fit is greatly dependent on the value of parameter values, such as LDA's parameter for the number of topics, which we have referred to as k. Given that existing measures of model fit do not necessarily correlate with human interpretability of topics, it can be helpful to choose k via direct human inspection of models with different k values. There are currently no tools that directly help the user tune k by communicating changes induced by different k. In this work, we describe ongoing work to build a tool specifically tailored for visualizing evolution of models with changing k, thus allowing the user to pick a k most appropriate for the task.

8. REFERENCES

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[2] C. Blevins. Topic modeling historical sources: Analyzing the diary of martha ballard. *Proceedings of Digital Humanities. Stanford, CA: Digital Humanities*, 2011.

[3] L. Cao and L. Fei-Fei. Spatially coherent latent topic model for concurrent segmentation and classification of objects and scenes. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE, 2007.

[4] J. Chang, S. Gerrish, C. Wang, J. L. Boyd-Graber, and D. M. Blei. Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems*, pages 288–296, 2009.

[5] J. Chuang, S. Gupta, C. Manning, and J. Heer. Topic model diagnostics: Assessing domain relevance via topical alignment. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pages 612–620, 2013.

[6] J. Chuang, Y. Hu, A. Jin, J. D. Wilkerson, D. A. McFarland, C. D. Manning, and J. Heer. Document exploration with topic modeling: Designing interactive visualizations to support effective analysis workflows.

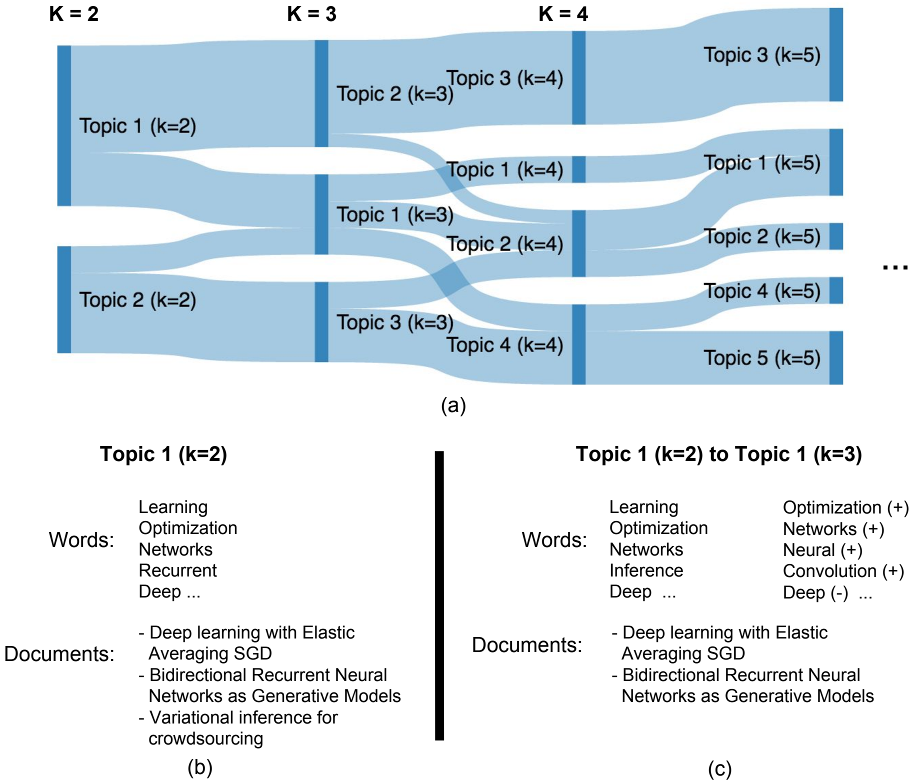

Figure 2: Topic-centric visualization.

[7] J. Chuang, C. D. Manning, and J. Heer. Termite: Visualization techniques for assessing textual topic models. In *Proceedings of the International Working Conference on Advanced Visual Interfaces*, pages 74–77. ACM, 2012.

[8] A. De Lucia, M. D. Penta, R. Oliveto, A. Panichella, and S. Panichella. Using ir methods for labeling source code artifacts: Is it worthwhile? In *Program Comprehension (ICPC), 2012 IEEE 20th International Conference on*, pages 193–202. IEEE, 2012.

[9] E. L. Glassman. Learning latent student design decisions in python programming classes, 2016.

[10] L. Hong and B. D. Davison. Empirical study of topic modeling in twitter. In *Proceedings of the first workshop on social media analytics*, pages 80–88. ACM, 2010.

[11] Y. Hu, J. Boyd-Graber, and B. Satinoff. Interactive topic modeling. In *Association for Computational Linguistics*, 2011.

[12] D. Mimno, H. M. Wallach, E. Talley, M. Leenders, and A. McCallum. Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 262–272. Association for Computational Linguistics, 2011.

[13] A. J. Perotte, F. Wood, N. Elhadad, and N. Bartlett. Hierarchically supervised latent dirichlet allocation. In *Advances in Neural Information Processing Systems*, pages 2609–2617, 2011.

[14] A. M. Saeidi, J. Hage, R. Khadka, and S. Jansen. Itmviz: interactive topic modeling for source code analysis. In *Program Comprehension (ICPC), 2015 IEEE 23rd International Conference on*, pages 295–298. IEEE, 2015.

[15] C. Sievert and K. E. Shirley. Ldavis: A method for visualizing and interpreting topics. 2014.

[16] A. Smith, N. Elmqvist, T. Y. Lee, K. Seppi, F. Poursabzi-Sangdeh, J. Boyd-Graber, and L. Findlater. Human-centered and interactive: Expanding the impact of topic models. In *Human-Centered Machine Learning workshop*. CHI, 2016.

[17] A. Smith, S. Malik, and B. Shneiderman. Visual analysis of topical evolution in unstructured text: Design and evaluation of topicflow. In *Applications of Social Media and Social Network Analysis*, pages 159–175. Springer, 2015.