LDAvis: A method for visualizing and interpreting topics

Carson Sievert
Iowa State University
3414 Snedecor Hall
Ames, IA 50014, USA
cpsievert1@gmail.com

Kenneth E. Shirley
AT&T Labs Research
33 Thomas Street, 26th Floor
New York, NY 10007, USA
kshirley@research.att.com

Abstract

We present LDAvis, a web-based interactive visualization of topics estimated using Latent Dirichlet Allocation that is built using a combination of R and D3. Our visualization provides a global view of the topics (and how they differ from each other), while at the same time allowing for a deep inspection of the terms most highly associated with each individual topic. First, we propose a novel method for choosing which terms to present to a user to aid in the task of topic interpretation, in which we define the *relevance* of a term to a topic. Second, we present results from a user study that suggest that ranking terms purely by their probability under a topic is suboptimal for topic interpretation. Last, we describe LDAvis, our visualization system that allows users to flexibly explore topic-term relationships using relevance to better understand a fitted LDA model.

1 Introduction

Recently much attention has been paid to visualizing the output of topic models fit using Latent Dirichlet Allocation (LDA) (Gardner et al., 2010; Chaney and Blei, 2012; Chuang et al., 2012b; Gretarsson et al., 2011). Such visualizations are challenging to create because of the high dimensionality of the fitted model – LDA is typically applied to many thousands of documents, which are modeled as mixtures of dozens (or hundreds) of topics, which themselves are modeled as distributions over thousands of terms (Blei et al., 2003; Griffiths and Steyvers, 2004). The most promising basic technique for creating LDA visualizations that are both compact and thorough is *interactivity*.

We introduce an interactive visualization system that we call LDAvis that attempts to answer a few basic questions about a fitted topic model: (1) What is the meaning of each topic?, (2) How prevalent is each topic?, and (3) How do the topics relate to each other? Different visual components answer each of these questions, some of which are original, and some of which are borrowed from existing tools.

Our visualization (illustrated in Figure 1) has two basic pieces. First, the left panel of our visualization presents a global view of the topic model, and answers questions 2 and 3. In this view, we plot the topics as circles in the two-dimensional plane whose centers are determined by computing the distance between topics, and then by using multidimensional scaling to project the inter-topic distances onto two dimensions, as is done in (Chuang et al., 2012a). We encode each topic's overall prevalence using the areas of the circles, where we sort the topics in decreasing order of prevalence.

Second, the right panel of our visualization depicts a horizontal barchart whose bars represent the individual terms that are the most useful for interpreting the currently selected topic on the left, and allows users to answer question 1, "What is the meaning of each topic?". A pair of overlaid bars represent both the corpus-wide frequency of a given term as well as the topic-specific frequency of the term, as in (Chuang et al., 2012b).

The left and right panels of our visualization are linked such that selecting a topic (on the left) reveals the most useful terms (on the right) for interpreting the selected topic. In addition, selecting a term (on the right) reveals the conditional distribution over topics (on the left) for the selected term. This kind of linked selection allows users to examine a large number of topic-term relationships in a compact manner.

A key innovation of our system is how we determine the most useful terms for interpreting a given topic, and how we allow users to interactively ad-

Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces, pages 63–70,
Baltimore, Maryland, USA, June 27, 2014.

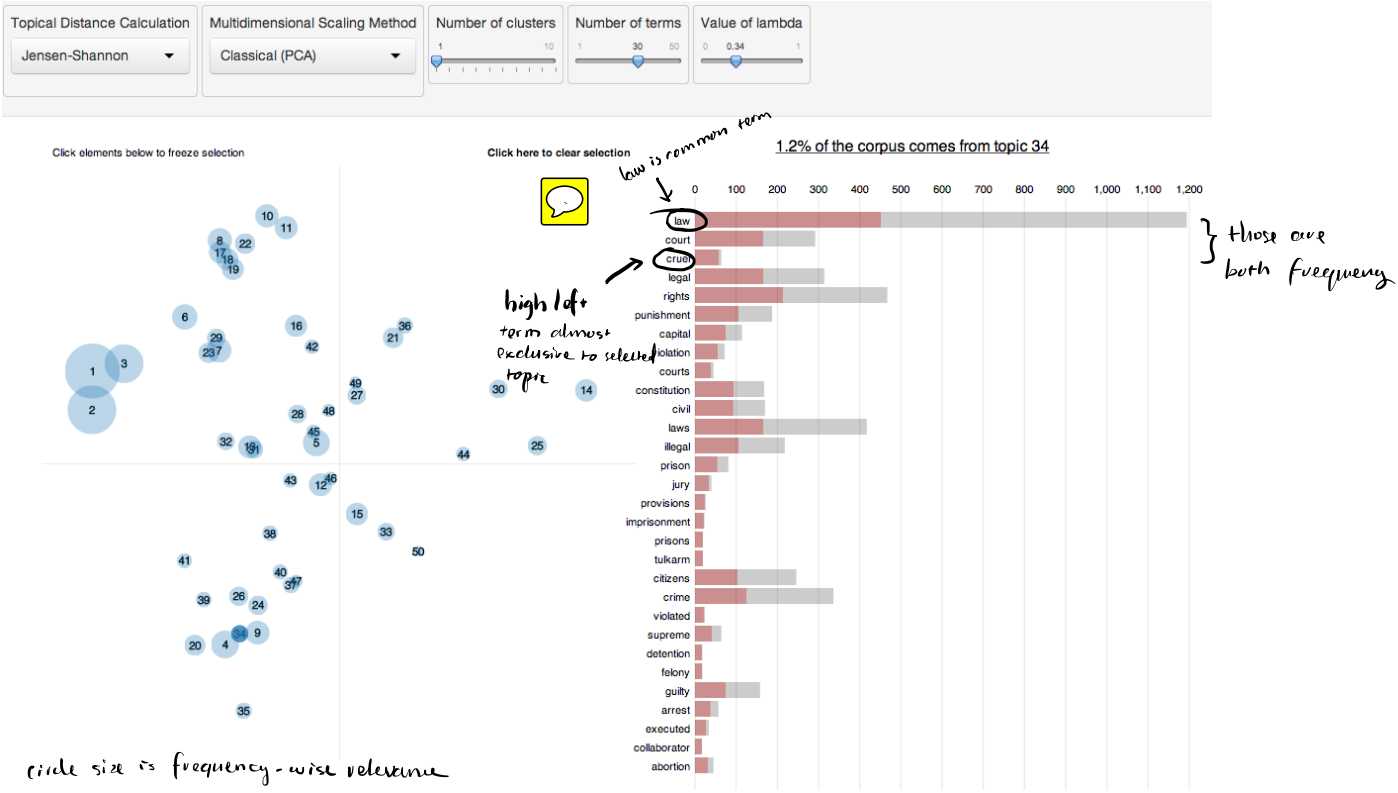

Figure 1: The layout of LDAvis, with the global topic view on the left, and the term barcharts (with Topic 34 selected) on the right. Linked selections allow users to reveal aspects of the topic-term relationships compactly.

just this determination. A topic in LDA is a multinomial distribution over the (typically thousands of) terms in the vocabulary of the corpus. To interpret a topic, one typically examines a ranked list of the most probable terms in that topic, using anywhere from three to thirty terms in the list. The problem with interpreting topics this way is that common terms in the corpus often appear near the top of such lists for multiple topics, making it hard to differentiate the meanings of these topics.

Bischof and Airoldi (2012) propose ranking terms for a given topic in terms of both the *frequency* of the term under that topic as well as the term's *exclusivity* to the topic, which accounts for the degree to which it appears in that particular topic to the exclusion of others. We propose a similar measure that we call the *relevance* of a term to a topic that allows users to flexibly rank terms in order of usefulness for interpreting topics. We discuss our definition of relevance, and its graphical interpretation, in detail in Section 3.1. We also present the results of a user study conducted to determine the optimal tuning parameter in the definition of relevance to aid the task of topic interpretation in Section 3.2, and we describe how we incorporate relevance into our interactive visualization in Section 4.

2 Related Work

Much work has been done recently regarding the interpretation of topics (i.e. measuring topic "coherence") as well as visualization of topic models.

2.1 Topic Interpretation and Coherence

It is well-known that the topics inferred by LDA are not always easily interpretable by humans. Chang et al. (2009) established via a large user study that standard quantitative measures of fit, such as those summarized by Wallach et al. (2009), do not necessarily agree with measures of topic interpretability by humans. Ramage et al. (2009) assert that "characterizing topics is hard" and describe how using the top-k terms for a given topic might not always be best, but offer few concrete alternatives.

AlSumait et al. (2009), Mimno et al. (2011), and Chuang et al. (2013b) develop quantitative methods for measuring the interpretability of top-

ics based on experiments with data sets that come with some notion of topical ground truth, such as document metadata or expert-created topic labels. These methods are useful for understanding, in a global sense, which topics are interpretable (and why), but they don't specifically attempt to aid the user in interpreting *individual* topics.

Blei and Lafferty (2009) developed "Turbo Topics", a method of identifying n-grams within LDA-inferred topics that, when listed in decreasing order of probability, provide users with extra information about the usage of terms within topics. This two-stage process yields good results on experimental data, although the resulting output is still simply a ranked list containing a mixture of terms and n-grams, and the usefulness of the method for topic interpretation was not tested in a user study.

Newman et al. (2010) describe a method for ranking terms within topics to aid interpretability called Pointwise Mutual Information (PMI) ranking. Under PMI ranking of terms, each of the ten most probable terms within a topic are ranked in decreasing order of approximately how often they occur in close proximity to the nine other most probable terms from that topic in some large, external "reference" corpus, such as Wikipedia or Google n-grams. Although this method correlated highly with human judgments of term importance within topics, it does not easily generalize to topic models fit to corpora that don't have a readily available external source of word co-occurrences.

In contrast, Taddy (2011) uses an intrinsic measure to rank terms within topics: a quantity called *lift*, defined as the ratio of a term's probability within a topic to its marginal probability across the corpus. This generally decreases the rankings of globally frequent terms, which can be helpful. We find that it can be noisy, however, by giving high rankings to very rare terms that occur in only a single topic, for instance. While such terms may contain useful topical content, if they are very rare the topic may remain difficult to interpret.

Finally, Bischof and Airoldi (2012) develop and implement a new statistical topic model that infers both a term's frequency as well as its *exclusivity* – the degree to which its occurrences are limited to only a few topics. They introduce a univariate measure called a FREX score ("**FR**equency and **EX**clusivity") which is a weighted harmonic mean of a term's rank within a given topic with respect to frequency and exclusivity, and they recommend it as a way to rank terms to aid topic interpretation. We propose a similar method that is a weighted average of the logarithms of a term's probability and its lift, and we justify it with a user study and incorporate it into our interactive visualization.

2.2 Topic Model Visualization Systems

A number of visualization systems for topic models have been developed in recent years. Several of them focus on allowing users to browse documents, topics, and terms to learn about the relationships between these three canonical topic model units (Gardner et al., 2010; Chaney and Blei, 2012; Snyder et al., 2013). These browsers typically use lists of the most probable terms within topics to summarize the topics, and the visualization elements are limited to barcharts or word clouds of term probabilities for each topic, pie charts of topic probabilities for each document, and/or various barcharts or scatterplots related to document metadata. Although these tools can be useful for browsing a corpus, we seek a more compact visualization, with the more narrow focus of quickly and easily understanding the individual topics themselves (without necessarily visualizing documents).

Chuang et al. (2012b) develop such a tool, called "Termite", which visualizes the set of topic-term distributions estimated in LDA using a matrix layout. The authors introduce two measures of the usefulness of terms for understanding a topic model: *distinctiveness* and *saliency*. These quantities measure how much information a term conveys about topics by computing the Kullback-Liebler divergence between the distribution of topics given the term and the marginal distribution of topics (distinctiveness), optionally weighted by the term's overall frequency (saliency). The authors recommend saliency as a thresholding method for selecting which terms are included in the visualization, and they further use a seriation method for ordering the most salient terms to highlight differences between topics.

Termite is a compact, intuitive interactive visualization of the topics in a topic model, but by only including terms that rank high in saliency or distinctiveness, which are *global* properties of terms, it is restricted to providing a *global* view of the model, rather than allowing a user to deeply in-

spect individual topics by visualizing a potentially different set of terms for every single topic. In fact, Chuang et al. (2013a) describe the use of a "topic-specific word ordering" as potentially useful future work.

3 Relevance of terms to topics

Here we define *relevance*, our method for ranking terms within topics, and we describe the results of a user study to learn an optimal tuning parameter in the computation of relevance.

3.1 Definition of Relevance

Let ϕ_{kw} denote the probability of term $w \in \{1, ..., V\}$ for topic $k \in \{1, ..., K\}$, where V denotes the number of terms in the vocabulary, and let p_w denote the marginal probability of term w in the corpus. One typically estimates ϕ in LDA using Variational Bayes methods or Collapsed Gibbs Sampling, and p_w from the empirical distribution of the corpus (optionally smoothed by including prior weights as pseudo-counts).

We define the *relevance* of term w to topic k given a weight parameter λ (where $0 \leq \lambda \leq 1$) as:

$$r(w, k \mid \lambda) = \lambda \log(\phi_{kw}) + (1 - \lambda) \log\left(\frac{\phi_{kw}}{p_w}\right),$$

where λ determines the weight given to the probability of term w under topic k relative to its lift (measuring both on the log scale). Setting $\lambda = 1$ results in the familiar ranking of terms in decreasing order of their topic-specific probability, and setting $\lambda = 0$ ranks terms solely by their lift. We wish to learn an "optimal" value of λ for topic interpretation from our user study.

First, though, to see how different values of λ result in different ranked term lists, consider the plot in Figure 2. We fit a 50-topic model to the 20 Newsgroups data (details are described in Section 3.2) and plotted $\log(\text{lift})$ on the y-axis vs. $\log(\phi_{kw})$ on the x-axis for each term in the vocabulary (which has size $V = 22,524$) for a given topic. Figure 2 shows this plot for Topic 29, which occurred mostly in documents posted to the "Motorcycles" Newsgroup, but also from documents posted to the "Automobiles" Newsgroup and the "Electronics" Newsgroup. Graphically, the line separating the most relevant terms for this topic, given λ, has slope $-\lambda/(1 - \lambda)$ (see Figure 2).

For this topic, the top-5 most relevant terms given $\lambda = 1$ (ranking solely by probability) are {out, #emailaddress, #twodigitnumer, up, #onedigitnumber}, where a "#" symbol denotes a term that is an entity representing a class of things. In contrast to this list, which contains globally common terms and which provides very little meaning regarding motorcycles, automobiles, or electronics, the top-5 most relevant terms given $\lambda = 1/3$ are {oil, plastic, pipes, fluid, and lights}. The second set of terms is much more descriptive of the topic being discussed than the first.

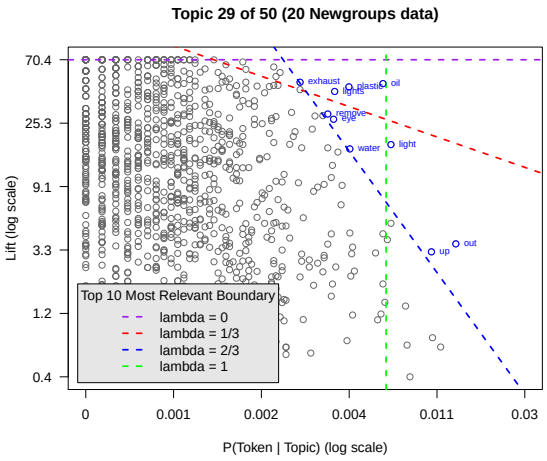

Figure 2: Dotted lines separating the top-10 most relevant terms for different values of λ, with the most relevant terms for $\lambda = 2/3$ displayed and highlighted in blue.

3.2 User Study

We conducted a user study to determine whether there was an optimal value of λ in the definition of relevance to aid topic interpretation. First, we fit a 50-topic model to the $D = 13,695$ documents in the 20 Newsgroups data which were posted to a single Newsgroup (rather than two or more Newsgroups). We used the Collapsed Gibbs Sampler algorithm (Griffiths and Steyvers, 2004) to sample the latent topics for each of the $N = 1,590,376$ tokens in the data, and we saved their topic assignments from the last iteration (after convergence). We then computed the 20 by 50 table, T, which contains, in cell T_{gk}, the count of the number of times a token from topic $k \in \{1, ..., 50\}$ was assigned to Newsgroup $g \in \{1, ..., 20\}$, where we defined the Newsgroup of a token to be the Newsgroup to which the document containing that token was posted. Some of the LDA-inferred topics occurred almost exclusively ($> 90\%$ of occur-

rences) in documents from a single Newsgroup, such as Topic 38, which was the estimated topic for 15,705 tokens in the corpus, 14,233 of which came from documents posted to the "Medicine" (or "sci.med") Newsgroup. Other topics occurred in a wide variety of Newsgroups. One would expect these "spread-out" topics to be harder to interpret than the "pure" topics like Topic 38.

In the study we recruited 29 subjects among our colleagues (research scientists at AT&T Labs with moderate familiarity with text mining techniques and topic models), and each subject completed an online experiment consisting of 50 tasks, one for each topic in the fitted LDA model. Task k (for $k \in \{1, ..., 50\}$) was to read a list of five terms, ranked from 1-5 in order of relevance to topic k, where $\lambda \in (0, 1)$ was randomly sampled to compute relevance. The user was instructed to identify which "topic" the list of terms discussed from a list of three possible "topics", where their choices were names of the Newsgroups. The correct answer for task k (i.e. our "ground truth") was defined as the Newsgroup that contributed the most tokens to topic k (i.e. the Newsgroup with the largest count in the kth column of the table T), and the two alternative choices were the Newsgroups that contributed the second and third-most tokens to topic k.

We anticipated that the effect of λ on the probability of a user making the correct choice could be different across topics. In particular, for "spread-out" topics that were inherently difficult to interpret, because their tokens were drawn from a wide variety of Newsgroups (similar to a "fused" topic in Chuang et al. (2013b)), we expected the proportion of correct responses to be roughly 1/3 no matter the value of λ used to compute relevance. Similarly, for very "pure" topics, whose tokens were drawn almost exclusively from one Newsgroup, we expected the task to be easy for any value of λ. To account for this, we analyzed the experimental data by fitting a varying-intercepts logistic regression model to allow each of the fifty topics to have its own baseline difficulty level, where the effect of λ is shared across topics. We used a quadratic function of λ in the model (linear, cubic and quartic functions were explored and rejected).

As expected, the baseline difficulty of each topic varied widely. In fact, seven of the topics were correctly identified by all 29 users,[1] and one topic was incorrectly identified by all 29 users.[2] For the remaining 42 topics we estimated a topic-specific intercept term to control for the inherent difficulty of identifying the topic (not just due to its tokens being spread among multiple Newsgroups, but also to account for the inherent familiarity of each topic to our subject pool – subjects, on average, were more familiar with "Cars" than "The X Window System", for example).

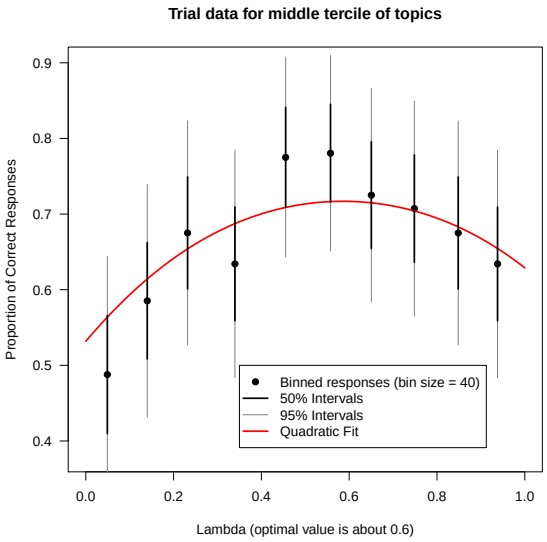

Figure 3: A plot of the proportion of correct responses in a user study vs. the value of λ used to compute the most relevant terms for each topic.

The estimated effects of λ and λ^2 were 2.74 and -2.34, with standard errors 1.03 and 1.00. Taken together, their joint effect was statistically significant (χ^2 p-value = 0.018). To see the estimated effect of λ on the probability of correctly identifying a topic, consider Figure 3. We plot binned proportions of correct responses (on the y-axis) vs. λ (on the x-axis) for the 14 topics whose estimated topic-specific intercepts fell into the middle tercile among the 42 topics that weren't trivial or impossible to identify. Among these topics there was roughly a 67% baseline probability of correct identification. As Figure 3 shows, for these topics, the "optimal" value of λ was about 0.6, and it resulted in an estimated 70% probability of correct identification, whereas for values of λ near 0 and

[1] Whose ground truth labels were Medicine (twice), Miscellaneous Politics, Christianity, Gun Politics, Space (Astronomy), and Middle East Politics.

[2] The ground truth label for this topic was "Christianity", but the presence of the term "islam" or "quran" among the top-5 for every value of λ led each subject to choose "Miscellaneous Religion".

1, the estimated proportions of correct responses were closer to 53% and 63%, respectively. We view this as evidence that ranking terms according to relevance, where $\lambda < 1$ (i.e. not strictly in decreasing order of probability), can improve topic interpretability.

Note that in our experiment, we used the collection of single-posted 20 Newsgroups documents to define our "ground truth" data. An alternative method for collecting "ground truth" data would have been to recruit experts to label topics from an LDA model. We chose against this option because doing so would present a classic "chicken-or-egg" problem: If we use expert-labeled topics in an experiment to learn how to summarize topics so that they can be interpreted (i.e. "labeled"), we would only re-learn the way that our experts were instructed, or allowed, to label the topics in the first place! If, for instance, the experts were presented with a ranked list of the most probable terms for each topic, this would influence the interpretations and labels they give to the topics, and the experimental result would be the circular conclusion that ranking terms by probability allows users to recover the "expert" labels most easily. To avoid this, we felt strongly that we should use data in which documents have metadata associated with them. The 20 Newsgroups data provides an externally validated source of topic labels, in the sense that the labels were presented to users (in the form of Newsgroup names), and users subsequently filled in the content. It represents, essentially, a crowd-sourced collection of tokens, or content, for a certain set of topic labels.

4 The `LDAvis` System

Our interactive, web-based visualization system, `LDAvis`, has two core functionalities that enable users to understand the topic-term relationships in a fitted LDA model, and a number of extra features that provide additional perspectives on the model.

First and foremost, `LDAvis` allows one to select a topic to reveal the most relevant terms for that topic. In Figure 1, Topic 34 is selected, and its 30 most relevant terms (given $\lambda = 0.34$, in this case) populate the barchart to the right (ranked in order of relevance from top to bottom). The widths of the gray bars represent the corpus-wide frequencies of each term, and the widths of the red bars represent the topic-specific frequencies of each term. A slider allows users to change the value of λ, which can alter the rankings of terms to aid topic interpretation. By default, λ is set to 0.6, as suggested by our user study in Section 3.2. If $\lambda = 1$, terms are ranked solely by ϕ_{kw}, which implies the red bars would be sorted from widest (at the top) to narrowest (at the bottom). By comparing the widths of the red and gray bars for a given term, users can quickly understand whether a term is highly relevant to the selected topic because of its lift (a high ratio of red to gray), or its probability (absolute width of red). The top 3 most relevant terms in Figure 1 are "law", "court", and "cruel". Note that "law" is a common term which is generated by Topic 34 in about 40% of its corpus-wide occurrences, whereas "cruel" is a relatively rare term with very high lift in Topic 34 – it occurs almost exclusively in this topic. Such properties of the topic-term relationships are readily visible in `LDAvis` for every topic.

On the left panel, two visual features provide a global perspective of the topics. First, the areas of the circles are proportional to the relative prevalences of the topics in the corpus. In the 50-topic model fit to the 20 Newsgroups data, the first three topics comprise 12%, 9%, and 6% of the corpus, and all contain common, non-specific terms (although there are interesting differences: Topic 2 contains formal debate-related language such as "conclusion", "evidence", and "argument", whereas Topic 3 contains slang conversational language such as "kinda", "like", and "yeah"). In addition to visualizing topic prevalence, the left pane shows inter-topic differences. The default for computing inter-topic distances is Jensen-Shannon divergence, although other metrics are enabled. The default for scaling the set of inter-topic distances defaults to Principal Components, but other algorithms are also enabled.

The second core feature of `LDAvis` is the ability to select a term (by hovering over it) to reveal its conditional distribution over topics. This distribution is visualized by altering the areas of the topic circles such that they are proportional to the term-specific frequencies across the corpus. This allows the user to verify, as discussed in Chuang et al. (2012a), whether the multidimensional scaling of topics has faithfully clustered similar topics in two-dimensional space. For example, in Figure 4, the term "file" is selected. In the majority of this term's occurrences, it is drawn from one of several topics located in the upper left-hand region of the

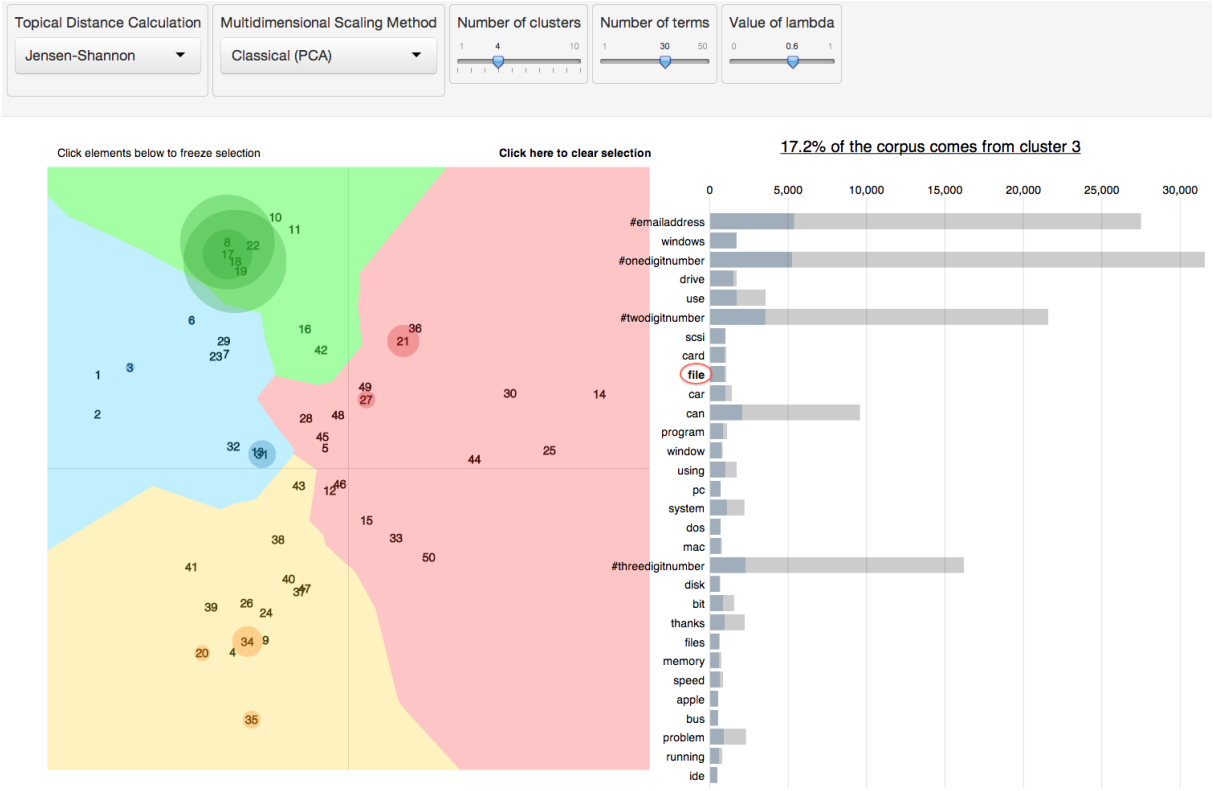

Figure 4: The user has chosen to segment the fifty topics into four clusters, and has selected the green cluster to populate the barchart with the most relevant terms for that cluster. Then, the user hovered over the ninth bar from the top, "file", to display the conditional distribution over topics for this term.

global topic view. Upon inspection, this group of topics can be interpreted broadly as a discussion of computer hardware and software. This verifies, to some extent, their placement, via multidimensional scaling, into the same two-dimensional region. It also suggests that the term "file" used in this context refers to a computer file. However, there is also conditional probability mass for the term "file" on Topic 34. As shown in Figure 1, Topic 34 can be interpreted as discussing the criminal punishment system where "file" refers to court filings. Similar discoveries can be made for any term that exhibits polysemy (such as "drive" appearing in computer- and automobile-related topics, for example).

Beyond its within-browser interaction capability using D3 (Bostock et al., 2011), `LDAvis` leverages the `R` language (R Core Team, 2014) and specifically, the `shiny` package (Rstudio, 2014), to allow users to easily alter the topical distance measurement as well as the multidimensional scaling algorithm to produce the global topic view. In addition, there is an option to apply k-means clustering to the topics (as a function of their two-dimensional locations in the global topic view). This is merely an effort to facilitate semantic zooming in an LDA model with many topics where 'after-the-fact' clustering may be an easier way to estimate clusters of topics, rather than fitting a hierarchical topic model (Blei et al., 2003), for example. Selecting a cluster of topics (by clicking the Voronoi region corresponding to the cluster) reveals the most relevant terms for that cluster of topics, where the term distribution of a cluster of topics is defined as the weighted average of the term distributions of the individual topics in the cluster. In Figure 4, the green cluster of topics is selected, and the most relevant terms, displayed in the barchart on the right, are predominantly related to computer hardware and software.

5 Discussion

We have described a web-based, interactive visualization system, `LDAvis`, that enables deep inspection of topic-term relationships in an LDA model, while simultaneously providing a global view of the topics, via their prevalences and similarities to each other, in a compact space. We

also propose a novel measure, *relevance*, by which to rank terms within topics to aid in the task of topic interpretation, and we present results from a user study that show that ranking terms in decreasing order of probability is suboptimal for topic interpretation. The `LDAvis` visualization system (including the user study data) is currently available as an `R` package on GitHub: https://github.com/cpsievert/LDAvis.

For future work, we anticipate performing a larger user study to further understand how to facilitate topic interpretation in fitted LDA models, including a comparison of multiple methods, such as ranking by Turbo Topics (Blei and Lafferty, 2009) or FREX scores (Bischof and Airoldi, 2012), in addition to relevance. We also note the need to visualize correlations between topics, as this can provide insight into what is happening on the document level without actually displaying entire documents. Last, we seek a solution to the problem of visualizing a large number of topics (say, from 100 - 500 topics) in a compact way.

References

Loulwah AlSumait, Daniel Barbara, James Gentle, and Carlotta Domeniconi. 2009. *Topic Significance Ranking of LDA Generative Models*. ECML.

Jonathan M. Bischof and Edoardo M. Airoldi. 2012. *Summarizing topical content with word frequency and exclusivity*. ICML.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2012. *Latent Dirichlet Allocation*. JMLR.

David M. Blei and John Lafferty. 2009. Visualizing Topics with Multi-Word Expressions. arXiv:0907.1013v1 [stat.ML], 2009

David M. Blei, Thomas L. Griffiths, Michael I. Jordan, and Joshua B. Tenenbaum. 2003. *Hierarchical Topic Models and the Nested Chinese Restaurant Process*. NIPS.

Michael Bostock, Vadim Ogievetsky, Jeffrey Heer 2011. *D3: Data-Driven Documents*. InfoVis.

Allison J.B. Chaney and David M. Blei. 2012. *Visualizing topic models*. ICWSM.

Jonathan Chang, Jordan Boyd-Graber, Sean Gerrish, Chong Wang, and David M. Blei. 2009. *Reading Tea Leaves: How Humans Interpret Topic Models*. NIPS.

Jason Chuang, Daniel Ramage, Christopher D. Manning and Jeffrey Heer. 2012a. *Interpretation and Trust: Designing Model-Driven Visualizations for Text Analysis*. CHI.

Jason Chuang, Christopher D. Manning and Jeffrey Heer. 2012b. *Termite: Visualization Techniques for Assessing Textual Topic Models*. AVI.

Jason Chuang, Yuening Hu, Ashley Jin, John D. Wilkerson, Daniel A. McFarland, Christopher D. Manning and Jeffrey Heer. 2013a. *Document Exploration with Topic Modeling: Designing Interactive Visualizations to Support Effective Analysis Workflows*. NIPS Workshop on Topic Models: Computation, Application, and Evaluation.

Jason Chuang, Sonal Gupta, Christopher D. Manning and Jeffrey Heer. 2013b. *Topic Model Diagnostics: Assessing Domain Relevance via Topical Alignment*. ICML.

Matthew J. Gardner, Joshua Lutes, Jeff Lund, Josh Hansen, Dan Walker, Eric Ringger, and Kevin Seppi. 2010. *The topic browser: An interactive tool for browsing topic models*. NIPS Workshop on Challenges of Data Visualization.

Brynjar Gretarsson, John O'Donovan, Svetlin Bostandjieb, Tobias Hollerer, Arthur Asuncion, David Newman, and Padhraic Smyth. 2011. *TopicNets: Visual Analysis of Large Text Corpora with Topic Modeling*. ACM Transactions on Intelligent Systems and Technology, pp 1-26.

Thomas L. Griffiths and Mark Steyvers. 2004. *Finding scientific topics*. PNAS.

David Mimno, Hanna M. Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. *Optimizing Semantic Coherence in Topic Models*. EMNLP.

David Newman, Youn Noh, Edmund Talley, Sarvnaz Karimi, and Timothy Baldwin 2010. *Evaluating Topic Models for Digital Libraries*. JCDL.

R Core Team 2014. *R: A Language and Environment for Statistical Computing*. http://www.R-project.org.

R Studio, Inc. 2014. *shiny: Web Application Framework for R; package version 0.9.1*. http://CRAN.R-project.org/package=shiny.

Daniel Ramage, Evan Rosen and Jason Chuang and Christopher D. Manning, and Daniel A. McFarland. 2009. *Topic Modeling for the Social Sciences*. NIPS Workshop on Applications for Topic Models: Text and Beyond.

Justin Snyder, Rebecca Knowles, Mark Dredze, Matthew Gormley, and Travis Wolfe. 2013. *Topic Models and Metadata for Visualizing Text Corpora*. Proceedings of the 2013 NAACL HLT Demonstration Session.

Matthew A. Taddy 2011. *On Estimation and Selection for Topic Models*. AISTATS.

Hanna M. Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. 2009. *Evaluation Methods for Topic Models*. ICML.