The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Diagnosing and Improving Topic Models by Analyzing Posterior Variability

Linzi Xing
Department of Computer Science
University of Colorado, Boulder, CO 80309
linzi.xing@colorado.edu

Michael J. Paul
Department of Information Science
University of Colorado, Boulder, CO 80309
mpaul@colorado.edu

Abstract

Bayesian inference methods for probabilistic topic models can quantify uncertainty in the parameters, which has primarily been used to increase the robustness of parameter estimates. In this work, we explore other rich information that can be obtained by analyzing the posterior distributions in topic models. Experimenting with latent Dirichlet allocation on two datasets, we propose ideas incorporating information about the posterior distributions at the topic level and at the word level. At the topic level, we propose a metric called topic stability that measures the variability of the topic parameters under the posterior. We show that this metric is correlated with human judgments of topic quality as well as with the consistency of topics appearing across multiple models. At the word level, we experiment with different methods for adjusting individual word probabilities within topics based on their uncertainty. Humans prefer words ranked by our adjusted estimates nearly twice as often when compared to the traditional approach. Finally, we describe how the ideas presented in this work could potentially applied to other predictive or exploratory models in future work.

Introduction

Topic models, which extract themes from text datasets, have been widely used for large-scale corpus analysis with diverse applications including the study of topics in scientific articles (Hall, Jurafsky, and Manning 2008), finding patterns in classic literature (Jockers and Mimno 2013), understanding news media coverage (Roberts et al. 2013), and detecting population activities in online data (Paul and Dredze 2014). Of course, using topic models in these ways requires the belief that the topics meaningfully correspond to real concepts in a dataset. When topic models produce nonsensical topics or inconsistent outputs, it is difficult for a user to reliably use these models as a method of scientific inquiry. These challenges have motivated work on evaluating and understanding what it is that topic models discover (Chang et al. 2009; Mimno et al. 2011; Chuang et al. 2015).

In some sense, the lack of certainty about topics is already built into many commonly used models and inference algorithms. As a Bayesian model, the popular latent Dirichlet allocation (LDA) (Blei, Ng, and Jordan 2003) affords the possibility of inferring a posterior distribution over its parameters, which can provide an (approximate) measure of confidence of different parameter estimates. However, while Bayesian inference methods have been successful at obtaining more robust model estimates, they have mostly not been used for evaluating and understanding topic models.[1]

In this work, we investigate what kinds of characteristics about topics can be learned from the posterior of the parameters, focusing on variability in the posterior. Specifically, we look at how LDA topic parameters change across different posterior samples during Gibbs sampling. We find that the level of fluctuation in the parameter estimates can provide insights into the quality, consistency, and salience of the topics and their word probabilities.

This paper is divided into two main sections:

- **Topic-level analysis:** We explore how the variability in a topic's word distribution can be indicative of the quality and consistency of the topic. We propose a metric called topic stability that we show is correlated with human judgments of topic quality as well as with the consistency of the topic across multiple models.
- **Word-level analysis:** Within a topic, we examine how the probabilities of individual words vary, and find that words whose posteriors have high variability tend to be less salient and representative of the topic. We propose modifications to the ranking of words within a topic to adjust for this characteristic, and we show that people consistency prefer our modified rankings in experiments.

Our proposed methods are quite different from existing approaches. This leaves open a variety other possibilities for applying these ideas beyond what we investigated here, which we discuss after presenting our findings.

Topic Modeling

We use latent Dirichlet allocation (LDA) (Blei, Ng, and Jordan 2003) as our topic model. LDA has two types of parameters. Each document d has a probability distribution over topics, θ_d. Each topic k is associated with a distribution over

[1] An exception is work on posterior predictive checking of topic models (Mimno and Blei 2011), which diagnoses models by looking at discrepancies in how the model fits the data, but not variability of the posterior as proposed here.

words, ϕ_k. Additionally, the parameters ϕ and θ have priors defined by the Dirichlet distribution. The number of topics K must be specified, while the topic variables and parameters are unknown and must be inferred from the data.

Many algorithms existing for maximizing or inferring a posterior distribution over the latent variables and parameters (Blei 2012). We use Gibbs sampling (Geman and Geman 1984) as our posterior inference algorithm (Griffiths and Steyvers 2004) in this work. A Gibbs sampler generates samples of variable configurations from the posterior distribution. Each sample can provide a snapshot of the parameters, and our experiments explore how the topic model parameters, specifically the word distributions ϕ, vary across the different samples.

While topics are defined by an entire distribution over the vocabulary, they are usually presented to humans by displaying the most probable words, usually a fixed number of words. We represent topics as a ranked list of 10 words in this work. Many of our experiments will focus on how these 10 words are perceived by people, and when we refer to a word being "in" a topic, we mean that the word is in the set of 10 most probable words.

Evaluating Topic Models

A variety of work has developed methods to evaluate and characterize the quality and behavior of topic models (Boyd-Graber, Mimno, and Newman 2014). While it is possible to evaluate a topic model in a quantifiable predictive task (Wallach et al. 2009), it is non-trivial to evaluate the intrinsic quality of topics will be perceived by humans in exploratory tasks. Human feedback can be collected in the form of numeric ratings of quality (Newman et al. 2010) or by voting on different topic variants (Li and McCallum 2006). "Intruder" tasks conduct experiments to more objectively judge topic coherence by requiring people to identify out of place words or topics (Chang et al. 2009). Some "human-in-the-loop" systems are designed to interactively bring people into the modeling process, which is one way to help people diagnosis and improve their models (Hu et al. 2014; Chuang et al. 2015).

Due to the difficultly of obtaining human feedback, particularly when a large number of models are being considered and tuned, a number of automated alternatives to evaluation have been proposed. Most automated evaluations of quality focus on the semantic coherence of a topic—do the words form a cohesive group of related words? This is done by measuring the semantic similarity of the pairs of top words in the topic, usually using various co-occurrence statistics to estimate semantic similarity (Lau, Newman, and Baldwin. 2014; Roder, Both, and Hinneburg 2015). We will use two such metrics in this work:

The semantic coherence metric proposed by Mimno et al. (2011) is related to the sum of each conditional probability of each word in the topic given all other words, defined as:

$$C(t; V^{(t)}) = \sum_{m=2}^{M} \sum_{l=1}^{m-1} log \frac{D(v_m^{(t)}, v_l^{(t)}) + 1}{D(v_l^{(t)})} \qquad (1)$$

where t represents topic t and $V^{(t)} = (v_1^{(t)}, \cdots, v_M^{(t)})$ represents the M most probable words assigned to this topic. $D(v_l^{(t)})$ is the document frequency of word $v_l^{(t)}$ and $D(v_m^{(t)}, v_l^{(t)})$ is the co-document frequency of both words.

The NPMI metric used by Lau et al. (2014) uses the normalized pointwise mutual information (Bouma 2009) of unique word pairs:

$$NPMI(t; V^{(t)}) = \sum_{m=1}^{M-1} \sum_{l=m+1}^{M} \frac{log \frac{P(v_m^{(t)}, v_l^{(t)})}{P(v_m^{(t)}) P(v_l^{(t)})}}{-log P(v_m^{(t)}, v_l^{(t)})} \qquad (2)$$

where $P(v_m^{(t)})$ is marginal probability of word $v_m^{(t)}$ and $P(v_m^{(t)}, v_l^{(t)})$ represents joint probability of both words.

Dataset	# Docs	# Vocab	# Tokens
News	2,243	24,578	436,252
Wiki	10,000	80,011	7,383,116

Table 1: Statistics for the two document collections used in our experiments after pre-processing.

Experimental Setting

We now describe the datasets and experimental details that are used in both our topic-level and word-level analyses.

We experiment with two datasets. The *News* corpus contains 2,243 articles from the Associated Press. The *Wiki* corpus contains 10,000 articles from Wikipedia. We removed stop words and low-frequency words (appearing in fewer than five documents) from both datasets. Additionally, we removed proper nouns from the *Wiki* articles, following Chang et al. (2009), so that the topic model discovers more general concepts across the corpus. Statistics for the pre-processed datasets are provided in Table 1.

We set the number of topics to 50 for *News* and 100 for *Wiki*. We ran the Gibbs samplers for a burn-in period of 1,000 iterations, during which we also optimized the hyper-parameters of the Dirichlet priors, before freezing the hyper-parameters and collecting 100 samples, each separated by a 10-sample lag, running for a total of 2,000 iterations.[2]

Topic-Level Analysis

In this section, we explore what the posterior variability of a topic's word distribution can tell us about the topic. We hypothesize that topics whose distributions fluctuate between samples are more likely to contain ambiguous words and less likely to be topics that consistently represent the corpus. We will test this in two ways: comparing the variability to quality ratings provided by humans, and comparing the variability to how consistently topics appear in multiple models.

To measure posterior variability, we define a metric called **topic stability** which measures the degree to which a topic's parameters change during sampling. Examples of topics with high and low stability are show in Table 2.

[2]We also experimented with a total of 6,000 and 11,000 iterations, which made little difference in the results.

1000		1200		1400		1600		1800		2000	
Topic 6 (News, Stability = 0.9334)											
housing	.027	store	.020	stores	.022	store	.023	store	.023	store	.023
stores	.021	stores	.019	store	.021	stores	.023	stores	.020	stores	.022
store	.019	homeless	.019	homeless	.019	homeless	.019	homeless	.017	homeless	.018
homeless	.018	food	.016	housing	.015	food	.014	food	.016	food	.013
home	.015	housing	.016	food	.012	christmas	.012	christmas	.011	christmas	.013
food	.012	christmas	.011	home	.010	market	.011	home	.009	animals	.010
christmas	.011	city	.009	christmas	.010	clothing	.008	shopping	.008	market	.009
animals	.010	animals	.009	animals	.009	animals	.008	owner	.008	video	.008
city	.009	owner	.009	shopping	.008	video	.008	animals	.008	bought	.008
shopping	.007	shopping	.008	video	.008	shopping	.008	table	.007	owner	.007
Topic 1 (News, Stability = 0.9603)											
said	.033	said	.025	new	.020	new	.018	new	.018	said	.020
people	.015	people	.017	years	.015	people	.016	people	.016	people	.016
new	.015	new	.016	million	.015	report	.013	program	.015	program	.016
program	.014	million	.013	said	.015	national	.013	report	.012	report	.014
years	.013	years	.013	program	.014	million	.011	national	.011	states	.012
national	.012	national	.012	people	.014	said	.011	said	.011	new	.012
report	.012	program	.012	report	.014	years	.011	years	.010	national	.009
million	.010	report	.012	percent	.013	percent	.010	percent	.010	says	.009
problems	.009	percent	.010	national	.011	program	.01	states	.010	study	.008
percent	.009	says	.008	says	.009	says	.009	says	.009	federal	.007
Topic 11 (News, Stability = 0.9960)											
said	.060	said	.065	said	.071	said	.069	said	.070	said	.072
police	.053	police	.053	police	.052	police	.051	police	.049	police	.050
killed	.016	people	.016	people	.016	people	.016	people	.017	killed	.015
people	.014	killed	.014	killed	.013	killed	.015	killed	.016	people	.013
man	.009	city	.010	city	.010	man	.009	shot	.009	city	.009
shot	.009	arrested	.009	man	.009	shot	.008	man	.009	man	.008
city	.009	man	.009	shot	.009	authorities	.008	city	.008	shot	.008
arrested	.008	shot	.009	arrested	.008	arrested	.008	authorities	.008	died	.007
night	.007	night	.007	death	.007	city	.008	arrested	.007	arrested	.007
authorities	.007	men	.007	authorities	.007	night	.007	injured	.007	death	.007
Topic 52 (Wiki, Stability = 0.9999)											
age	.058	age	.058	age	.058	age	.059	age	.059	age	.059
population	.037	population	.037	population	.037	population	.037	population	.037	population	.037
median	.029	median	.029	median	.029	median	.029	median	.029	median	.029
income	.028	income	.028	income	.028	income	.028	income	.028	income	.028
census	.027	census	.027	census	.028	census	.027	census	.027	census	.027
living	.025	living	.025	living	.025	living	.025	living	.025	living	.025
households	.025	households	.024	households	.025	households	.025	households	.025	households	.025
average	.024	average	.024	average	.024	average	.024	average	.024	average	.024
years	.023	years	.024	years	.024	years	.024	years	.024	years	.024
families	.023	families	.023	families	.023	families	.023	families	.023	families	.023

Table 2: Examples of different topic samples. The columns correspond to different Gibbs sampling iterations (from 1000 to 2000), with the 10 most probable words shown based on the estimate from that specific sample. We show two topics with relatively low stability and two with relatively high stability. The high stability topics do not vary as much across samples.

The parameters for a topic k are its word distribution vector, denoted ϕ_k. Each sample from the Gibbs sampler can be used to obtain an estimate of ϕ_k. Let $\mathbf{\Phi_k}$ denote the set of estimates of ϕ_k from different samples, and $\bar{\phi}_k$ is the mean of those samples. We define the stability of the set of sample estimates as:

$$stability(\mathbf{\Phi_k}) = \frac{1}{|\mathbf{\Phi_k}|} \sum_{\phi_k \in \Phi_k} sim(\phi_k, \bar{\phi_k}) \qquad (3)$$

for a vector similarity function sim. We initially experimented with four similarity (or distance) metrics: cosine similarity, Euclidean distance, KL-divergence, and Jaccard similarity. For Jaccard similarity, we took the 10 most probable words in a sample ϕ_k as the set for comparison, as this is the set of words shown to humans in our experiments. Table 3 shows the performance of these four metrics on our two tasks (described in the two subsections below). We found that cosine similarity worked substantially better than the others in all cases, so this is what we use as our stability in the rest of this section.

Baselines In the experiments in this section, we will compare our proposed stability metric to two commonly used measurements of topic quality: the **coherence** metric of (Mimno et al. 2011) and the related **NPMI** metric (Lau, Newman, and Baldwin. 2014), both defined in the "Evaluating Topic Models" section above.

Comparison with Quality Rated by Humans

We explore whether our topic stability metric can indicate if a topic will be perceived as a high or low quality topic

Metric	Quality	Consistency
Cosine similarity	**0.249**	**0.315**
KL-divergence	0.016	0.254
Euclidean distance	0.013	0.152
Jaccard similarity	0.108	0.058

Table 3: The rank correlation between each potential metric of posterior variability compared to two topic-level metrics: mean human rating (quality) and model alignment cluster size (consistency).

	Quality		Consistency	
Metrics	*News*	*Wiki*	*News*	*Wiki*
Stability	0.248	0.253	**0.627**	**0.354**
Coherence	0.198	0.040	0.456	0.298
NPMI	**0.553**	**0.462**	0.340	0.142

Table 4: The rank correlation between each potential topic-level metric and the quality ratings or consistency.

to humans. We collected quality judgments from humans by having people rate topics (the 10 most probable words in the topic) on a 4-point Likert scale, with a 4 meaning that all words in the topic are related to each other, and a 1 meaning that most of the words are unrelated.

We collected ratings through Amazon Mechanical Turk. We collected ratings from seven different workers for each topic, in order to construct more robust estimates given the variability in human judgments. We removed and re-collected ratings from workers who completed the ratings were outliers in time to complete the tasks or in similarity to the ratings from other workers. We took the average of the seven ratings to produce a final rating for each topic. The average score across all topics is 3.07.

Figure 1 shows the topic ratings along with each of the three metrics (topic stability, coherence, NPMI), with correlations (Spearman's rank correlation ρ) in Table 4. On both datasets, topic stability is more correlated with quality than topic coherence, but NPMI has a higher correlation than either. Even though stability does not have the highest correlation, it is still noteworthy that it has a significant correlation with quality ratings. While coherence and NPMI both attempt to directly measure the relatedness of the words in the topic, the stability metric uses no information about the words that are in the topic, only the certainty of the parameters under the posterior. That topics with less certain parameters have a tendency to be judged as lower quality topics is an interesting finding that can be explored more in future work.

Comparison with Consistency across Models

The notion of "stability" in topic models was previously described by Chuang et al. (2015) in the context of comparing multiple models. Their study investigated variations of different trials of LDA, as different Gibbs sampling runs will result in different output each time due to randomness in the inference procedure. This behavior can be problematic for potential users of topic models, such as social scientists, who are not sure how to interpret a topic that only sometimes appears in topic model output.

Chuang et al. (2015) studied this by *aligning* topics from different modeling runs and quantifying how consistently a topic is discovered by LDA. Some topics, stable topics, are always inferred by LDA, while others may be one off topics that cannot be replicated. This work used an up-to-one alignment algorithm for topical alignment. Pairs of topics from differently trained topic models are merged together if they meet a similarity threshold. The number of topics that are cluster indicates how stable or consistent the topic is.

Compared to quality judgments, we consider this an orthogonal approach to understanding topic models, as low-quality topics may consistently appear across models, while high-quality topics may appear inconsistently.[3] However, we hypothesize that the consistency of a topic across models may be related to the consistency of a topic across its posterior distribution, and so we separately experiment to see if automated metrics applied to one model, including our proposed topic stability metric, can predict consistency across multiple models.

We implemented this approach and applied it to our two datasets. We ran LDA four times on each corpus and then applied the up-to-one topical alignment process, using a cosine similarity threshold of 0.2.

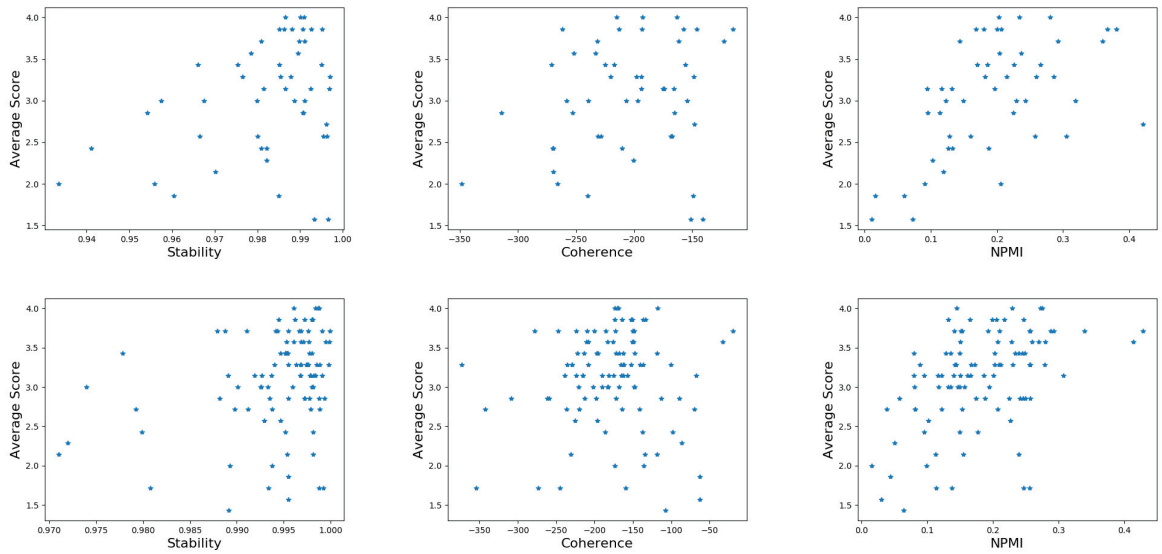
After getting topic clusters, we calculate the average value of each metric (topic stability, coherence, NPMI) in each cluster. Figure 2 shows the distribution of average cluster values for different sizes of clusters, with correlations shown in Table 4. Stability is most correlated with cluster size on both *News* and *Wiki* datasets. NPMI, which has the best correlation with topic quality, has a poor correlation with cluster size, suggesting that topic consistency is not necessarily related with topic quality, although there appears to at least be a small relation.

Overall, topic stability measured through posterior variability appears to be a good indicator of the consistency of topics across multiple models, although this finding is stronger in the *News* corpus than the larger *Wiki* corpus.

Word-Level Analysis

When focusing on the words within an individual topic, we also investigate how the variability of the posterior of individual word probabilities can be informative. Anecdotally, we find that words with high posterior variance tend to be less strongly associated with the topic, often common words like "said" and "new" that might be considered stop words, but were not in our stop word list during preprocessing. See Figure 3 for an example of this phenomenon.

[3] Indeed, the two measurements do not appear strongly related. The alignment cluster size has a low rank correlation with human ratings: .129 (*News*) and .110 (*Wiki*).



(a) Stability (b) Coherence (c) NPMI

Figure 1: Human-provided topic ratings along with stability, coherence and NPMI scores for the *News* (top, 50 topics) and *Wiki* (bottom, 100 topics) datasets.

In this section, we propose two methods for taking the posterior variability into consideration when ranking the top words in a topic, based on the hypothesis that words whose topic probabilities fluctuate highly are less likely to be salient representations of the topic. We obtain human judgments to evaluate different methods, and find that our adjustments result in higher quality topic representations.

Adjusted Word Scores

The conditional probability of word v in topic k is denoted ϕ_{kv}. We propose two methods of adjusting the values of ϕ_{kv} based on the set of posterior samples, $\mathbf{\Phi_k}$.

In the first method (referred to as **Mean/SD**), we divide the mean word probability $\bar{\phi}_{kv}$ by the standard deviation of ϕ_{kv} across the samples (the inverse coefficient of variation). This has the effect of downweighting the score of words whose posterior value of ϕ_{kv} fluctuates more, i.e., the value is less certain.

In the second method (referred to as **Min**), we take the lowest value of ϕ_{kv} that was observed in any of the samples. In other words, this is the empirical 0th percentile of the distribution of values.

Note that both of these modifications result in values that no longer form a valid probability distribution. For our setting, this does not matter because we only use these values to rank the words to select the 10 words that represent the topic. If one required probabilities, the values could be renormalized, but we did not experiment with this here.

Version	*News*	*Wiki*
Mean	3.04 (.10)	3.09 (.06)
Mean/SD	3.12 (.10)	**3.25** (.06)
Min	**3.19** (.10)	3.18 (.06)

Table 5: The average topic ratings (standard error in parentheses) on each dataset when the top 10 words are ranked by the mean probability (the baseline method), as well as our two proposed adjustments.

Baseline Our baseline comparison is the sample **mean** of ϕ_{kv}, typically used as an estimate of $\mathbb{E}[\phi_{kv}]$.

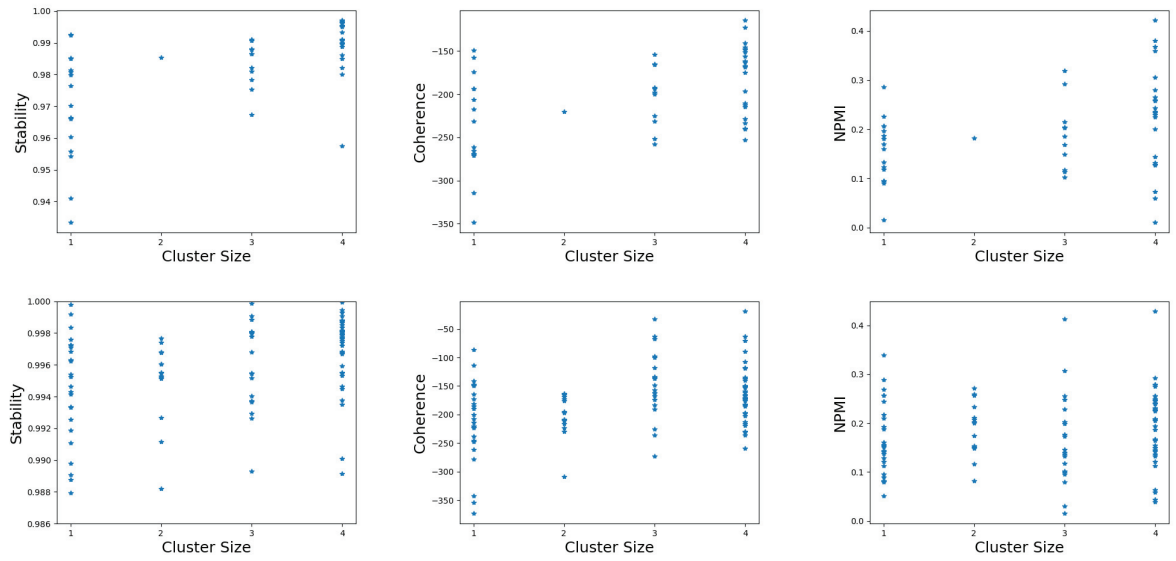
Experiments and Results

We conducted two experiments with human feedback collected through Amazon Mechanical Turk.

In the first experiment, we obtained quality ratings on the same 4-point Likert scale used in the previous section, but on topic representations where the top 10 words were ranked by our adjusted metrics. As before, we obtained ratings from seven different workers per topic.

The average topic ratings are shown in Table 5. Ratings under both adjusted scores are better than the baseline on both datasets. The difference in average ratings between the two adjusted methods (Mean/SD vs Min) is smaller, with inconsistent results between datasets.

In the second experiment, we did a direct pairwise comparison of topics where were ranked by two of the three scoring methods (the mean of ϕ_{kv}, the mean divided by standard deviation, and the minimum sample value). Workers were asked to choose which list of words was higher quality,



(a) Stability (b) Coherence (c) NPMI

Figure 2: Distribution of topic stability, coherence and NPMI scores within different sized clusters on the topic alignment task for the *News* (top) and *Wiki* (bottom) datasets. The topics in larger cluster sizes indicate they are easier to be replicated across multiple runs of LDA. On both datasets, topic stability is aligned relatively well with the cluster size. The number of topics within each cluster size (1, 2, 3, and 4, respectively) are 16, 1, 11, and 22 on *News* and 26, 15, 20, and 39 on *Wiki*.

	Mean	Mean/SD	Mean	Min	Mean/SD	Min
	News					
3/5	16	**34**	19	**30**	24	**26**
4/5	10	**21**	6	**13**	7	**9**
5/5	0	**1**	1	**3**	0	0
	Wiki					
3/5	38	**62**	39	**52**	**56**	44
4/5	16	**35**	17	**23**	**23**	15
5/5	1	**9**	0	**7**	**7**	3

Table 6: The number of times each word ranking method won a majority of votes in pairwise comparisons of different representations (Mean vs Mean/SD, Mean vs Min, Mean/SD vs Min), when the majority vote had at least 3, 4, or 5 votes.

where more words were related and representative of some concept. Workers were also told to consider the ranking order, where higher-ranked words should be more representative. There was not any sort of "don't know" or "they are tied" option; workers had to choose one, even if they were similar. In this experiment, we obtained results from five different workers per comparison. We excluded a small number of topics (10) whose ranked word list was identical between the two methods being compared.

Table 6 shows the results of these pairwise comparisons. We again find that both adjusted representations (Mean/SD and Min) are preferred over the baseline method (Mean), receiving a majority vote 1.59 times as often as the baseline. When counting only comparisons where the majority vote received at least four of the five votes, this ratio increases to 1.88. When all five voters agreed on the best representation, they voted for the adjusted representations over the baseline by a factor of 10.00.

Comparing the two new methods (Mean/SD and Min), *Min* is slightly favored on *News*, while *Mean/SD* is preferred by a larger amount on *Wiki*. There is not a clear conclusion in favor of either method, but we observed that *Mean/SD* produces larger changes to the ranking than *Min*. To quantify this, we measured the average Jaccard similarity between each topic's set of top 10 words ranked by *Mean* and the set of 10 words produced by one of the two new methods. The *Min* words were more similar to the baseline ranking (similarity of 0.73 on *News* and 0.84 on *Wiki*) than *Mean/SD* (similarity of 0.62 on *News* and 0.63 on *Wiki*).

Discussion and Future Directions

We have presented a new way of characterizing Bayesian topic models based on the variability of the posterior distri-

Topic	Method	Top 10 topic words
Topic 8 (*News*)	Mean	said ship water coast river boat sea guard island species
	Mean/SD	ship species coast water birds boat sea fish guard ships
	Min	ship water coast boat river sea species island ships fish
Topic 22 (*News*)	Mean	television network cbs nbc news tv abc million broadcast rating
	Mean/SD	cbs nbc network abc rating radio television cable cnn broadcast
	Min	network television cbs nbc tv abc news broadcast rating cable
Topic 74 (*Wiki*)	Mean	house building built castle th tower buildings city hall garden
	Mean/SD	building house built tower buildings garden castle designed hall design
	Min	building house built tower buildings garden castle hall houses site

Table 7: Example of 10-word topic representations using three different methods, where *Mean* is the baseline method of using the average sample probability. Highlighted words indicate words that only appear in the set for that particular method.

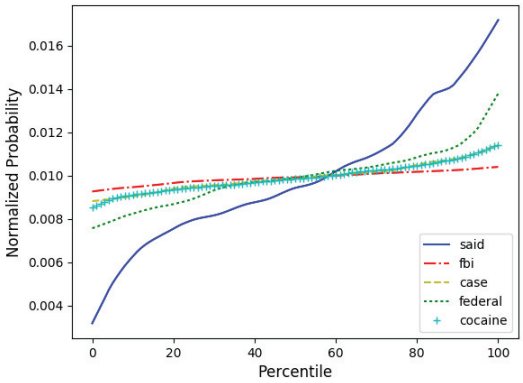


Figure 3: Different percentiles of probabilities throughout sampling of five words within the same topic (top 10 words: drug, attorney, office, investigation, said, case, charges, department, fbi, documents). From the word list, we might infer this topic is about justice and law enforcement. The word "said," which is not particularly related to this theme, fluctuates highly across different samples. The word "fbi," which is strongly related, fluctuates the least.

bution. We explored these ideas in two ways.

At the topic level, we introduced a metric called topic stability and showed that it is correlated with the consistency of topics across models and with the quality of topics rated by humans. Even though our proposed topic stability metric did not achieve state-of-the-art performance in the majority of cases, it always outperformed one of the two baseline methods. We argue that it is a strong and surprising result that topic stability is as highly correlated with other topic quality measures, given that it does not use any information about the words in the topic. This is a novel way of diagnosing and characterizing topics, which may end up being complementary to other types of metrics, and can likely be improved in future work.

At the word level, we found that words with high variability tend to be unrelated to the topic, and we proposed two ways of adjusting the weighting of words based on this information. In direct comparisons, people preferred our proposed modifications by a factors 1.6, 1.8, and 10.0 times compared to the standard way of ranking words in topics. These ideas can thus potentially provide a way to identify more salient representations of topics.

Because this work explored ideas that were quite different from prior approaches, there are many potential directions for future research. With respect to the tasks in this paper, future work could search for better definitions of topic stability and better methods for adjusting word scores. Additionally, this work only focused on the word distributions of topics, and not the topic distributions in documents. The latter is also important, and automatic evaluations have recently been proposed for these parameters as well (Bhatia, Lau, and Baldwin. 2017). It would likely be beneficial to explore the posterior variability at the document level. Beyond topic models, the idea of stability could be applied to other predictive models that would benefit from more interpretable parameters (Paul 2016). New metrics and methods using posterior variability could motivate performing Bayesian inference in models where this is not commonly done.

References

Bhatia, S.; Lau, J. H.; and Baldwin., T. 2017. An automatic approach for document-level topic model evaluation. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, 206–215.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research* 3:9931022.

Blei, D. M. 2012. Probabilistic topic models. *Communications of the ACM* 55(4):7784.

Bouma, G. 2009. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL* 31–40.

Boyd-Graber, J.; Mimno, D.; and Newman, D. 2014. *Care and Feeding of Topic Models: Problems, Diagnostics, and Improvements*. CRC Press.

Chang, J.; Boyd-Graber, J.; Wang, C.; Gerrish, S.; and Blei., D. M. 2009. Reading tea leaves: How humans interpret topic models. In *Neural Information Processing Systems*, 288–296.

Chuang, J.; Roberts, M. E.; Stewart, B. M.; Weiss, R.; Tingley, D.; Grimmer, J.; and Heer., J. 2015. TopicCheck: In-

teractive alignment for assessing topic model stability. In *HLT-NAACL*, 175–184.

Geman, S., and Geman, D. 1984. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence* 721–741.

Griffiths, T. L., and Steyvers, M. 2004. Finding scientific topics. In *Proceedings of the National academy of Sciences*, 5228–5235.

Hall, D.; Jurafsky, D.; and Manning, C. D. 2008. Studying the history of ideas using topic models. In *Proceedings of the conference on empirical methods in natural language processing*, 363–371.

Hu, Y.; Boyd-Graber, J.; Satinoff, B.; and Smith, A. 2014. Interactive topic modeling. *Machine Learning* 95(3):423–469.

Jockers, M. L., and Mimno, D. 2013. Significant themes in 19th-century literature. *Poetics* 750–769.

Lau, J. H.; Newman, D.; and Baldwin., T. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of EACL 2014*, 530–539.

Li, W., and McCallum, A. 2006. Pachinko allocation: DAG-structured mixture models of topic correlations. In *Proceedings of the 23rd International Conference on Machine Learning*, 577–584.

Mimno, D., and Blei, D. 2011. Bayesian checking for topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 227–237.

Mimno, D.; Wallach, H. M.; Talley, E.; Leenders, M.; and McCallum., A. 2011. Optimizing semantic coherence in topic models. In *EMNLP '11 Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 262–272.

Newman, D.; Lau, J. H.; Grieser, K.; and Baldwin, T. 2010. Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 100–108.

Paul, M. J., and Dredze, M. 2014. Discovering health topics in social media using topic models. *PLOS ONE* 9(8):e103408.

Paul, M. J. 2016. Interpretable machine learning: Lessons from topic modeling. In *CHI Workshop on Human-Centered Machine Learning*.

Roberts, M. E.; Stewart, B. M.; Tingley, D.; and Airoldi, E. M. 2013. The structural topic model and applied social science. In *Advances in Neural Information Processing Systems Workshop on Topic Models: Computation, Application, and Evaluation*.

Roder, M.; Both, A.; and Hinneburg, A. 2015. Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*.

Wallach, H. M.; Murray, I.; Salakhutdinov, R.; and Mimno, D. 2009. Evaluation methods for topic models. In *Proceedings of the 26th annual international conference on machine learning*, 1105–1112.