Not-So-Latent Dirichlet Allocation: Collapsed Gibbs Sampling Using Human Judgments

Jonathan Chang
Facebook
1601 S. California Ave.
Palo Alto, CA 94304
jonchang@facebook.com

Abstract

Probabilistic topic models are a popular tool for the unsupervised analysis of text, providing both a predictive model of future text and a latent topic representation of the corpus. Recent studies have found that while there are suggestive connections between topic models and the way humans interpret data, these two often disagree. In this paper, we explore this disagreement from the perspective of the learning process rather than the output. We present a novel task, *tag-and-cluster*, which asks subjects to simultaneously annotate documents and cluster those annotations. We use these annotations as a novel approach for constructing a topic model, grounded in human interpretations of documents. We demonstrate that these topic models have features which distinguish them from traditional topic models.

1 Introduction

Probabilistic topic models have become popular tools for the unsupervised analysis of large document collections (Deerwester et al., 1990; Griffiths and Steyvers, 2002; Blei and Lafferty, 2009). These models posit a set of latent *topics*, multinomial distributions over words, and assume that each document can be described as a mixture of these topics. With algorithms for fast approximate posterior inference, we can use topic models to discover both the topics and an assignment of topics to documents from a collection of documents. (See Figure 1.)

These modeling assumptions are useful in the sense that, empirically, they lead to good models of documents (Wallach et al., 2009). However, recent work has explored how these assumptions correspond to humans' understanding of language (Chang et al., 2009; Griffiths and Steyvers, 2006; Mei et al., 2007). Focusing on the latent space, i.e., the inferred mappings between topics and words and between documents and topics, this work has discovered that although there are some suggestive correspondences between human semantics and topic models, they are often discordant.

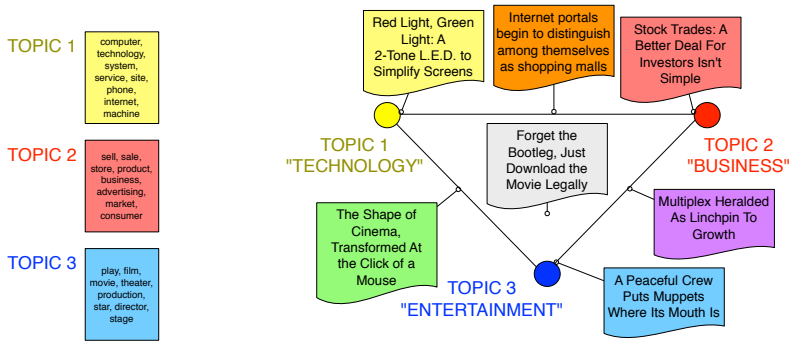
In this paper we build on this work to further explore how humans relate to topic models. But whereas previous work has focused on the results of topic models, here we focus on the process by which these models are learned. Topic models lend themselves to sequential procedures through which the latent space is inferred; these procedures are in effect programmatic encodings of the modeling assumptions. By substituting key steps in this program with human judgments, we obtain insights into the semantic model conceived by humans.

Here we present a novel task, *tag-and-cluster*, which asks subjects to simultaneously annotate a document and cluster that annotation. This task simulates the sampling step of the collapsed Gibbs sampler (described in the next section), except that the posterior defined by the model has been replaced by human judgments. The task is quick to complete and is robust against noise. We report the results of a large-scale human study of this task, and show that humans are indeed able to construct a topic model in this fashion, and that the learned topic model has semantic properties distinct from existing topic models. We also demonstrate that the judgments can be used to guide computer-learned topic models towards models which are more concordant with human intuitions.

2 Topic models and inference

Topic models posit that each document is expressed as a mixture of topics. These topic proportions are drawn once per document, and the topics are shared across the corpus. In this paper we focus on Latent Dirichlet allocation (LDA) (Blei et al., 2003) a topic model which treats each document's topic assignment as a multinomial random variable drawn from a symmetric Dirichlet prior. LDA, when applied to a collection of documents, will build a latent space: a collection of topics for the corpus and a collection of topic proportions for each of its documents.

Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, pages 131–138,
Los Angeles, California, June 2010.



(a) Topics (b) Document Assignments to Topics

Figure 1: The latent space of a topic model consists of topics, which are distributions over words, and a distribution over these topics for each document. On the left are three topics from a fifty topic LDA model trained on articles from the New York Times. On the right is a simplex depicting the distribution over topics associated with seven documents. The line from each document's title shows the document's position in the topic space.

LDA can be described by the following generative process:

1. For each topic k,
   (a) Draw topic $\beta_k \sim \text{Dir}(\eta)$
2. For each document d,
   (a) Draw topic proportions $\theta_d \sim \text{Dir}(\alpha)$
   (b) For each word $w_{d,n}$,
      i. Draw topic assignment $z_{d,n} \sim \text{Mult}(\theta_d)$
      ii. Draw word $w_{d,n} \sim \text{Mult}(\beta_{z_{d,n}})$

This process is depicted graphically in Figure 2. The parameters of the model are the number of topics, K, as well as the Dirichlet priors on the topic-word distributions and document-topic distributions, α and η. The only observed variables of the model are the words, $w_{d,n}$. The remaining variables must be learned.

There are several techniques for performing posterior inference, i.e., inferring the distribution over hidden variables given a collection of documents, including variational inference (Blei et al., 2003) and Gibbs sampling (Griffiths and Steyvers, 2006). In the sequel, we focus on the latter approach.

Collapsed Gibbs sampling for LDA treats the topic-word and document-topic distributions, θ_d and β_k, as nuisance variables to be marginalized out. The posterior distribution over the remaining latent variables,

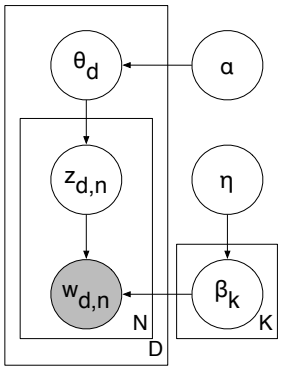


Figure 2: A graphical model depiction of latent Dirichlet allocation (LDA). Plates denote replication. The shaded circle denotes an observed variable and unshaded circles denote hidden variables.

the topic assignments $z_{d,n}$, can be expressed as

$$p(\boldsymbol{z}|\alpha,\eta,\boldsymbol{w}) \propto \prod_d \prod_k \left[\Gamma(n_{d,k}+\alpha_k)\frac{\Gamma(\eta_{w_{d,n}}+n_{w_{d,n},k})}{\Gamma(\sum_w n_{w,k}+\eta_w)}\right],$$

where $n_{d,k}$ denotes the number of words in document d assigned to topic k and $n_{w,k}$ the number of times word w is assigned to topic k. This leads to the sampling equations,

$$p(z_{d,i}=k|\alpha,\eta,\boldsymbol{w},\boldsymbol{z}y) \propto (n_{d,k}^{\neg d,i}+\alpha_k)\frac{\eta_{w_{d,i}}+n_{w_{d,i},k}^{\neg d,i}}{\sum_w n_{w,k}^{\neg d,i}+\eta_w}, \quad (1)$$

where the superscript $\neg d,i$ indicates that these statis-

tics should exclude the current variable under consideration, $z_{d,i}$.

In essence, the model performs inference by looking at each word in succession, and probabilistically assigning it to a topic according to Equation 1. Equation 1 is derived through the modeling assumptions and choice of parameters. By replacing Equation 1 with a different equation or with empirical observations, we may construct new models which reflect different assumptions about the underlying data.

3 Constructing topics using human judgments

In this section we propose a task which creates a formal setting where humans can create a latent space representation of the corpus. Our task, *tag-and-cluster*, replaces the collapsed Gibbs sampling step of Equation 1 with a human judgment. In essence, we are constructing a gold-standard series of samples from the posterior.[1]

Figure 3 shows how *tag-and-cluster* is presented to users. The user is shown a document along with its title; the document is randomly selected from a pool of available documents. The user is asked to select a word from the document which is discriminative, i.e, a word which would help someone looking for the document find it. Once the word is selected, the user is then asked to assign the word to the topic which best suits the sense of the word used in the document. Users are specifically instructed to focus on the meanings of words, not their syntactic usage or orthography.

The user assigns a word to a topic by selecting an entry out of a menu of topics. Each topic is represented by the five words occuring most frequently in that topic. The order of the topics presented to the user is determined by the number of words in that document already assigned to each topic. Once an instance of a word in a document has been assigned, it cannot be reassigned and will be marked in red when subsequent users encounter this document. In practice, we also prohibit users from selecting infrequently occurring words and stop words.

[1]Additionally, since Gibbs sampling is by nature stochastic, we believe that the task is robust against small perturbations in the quality of the assignments, so long as in aggregate they tend toward the mode.

4 Experimental results

We conducted our experiments using Amazon Mechanical Turk, which allows workers (our pool of prospective subjects) to perform small jobs for a fee through a Web interface. No specialized training or knowledge is typically expected of the workers. Amazon Mechanical Turk has been successfully used in the past to develop gold-standard data for natural language processing (Snow et al., 2008).

We prepare two randomly-chosen, 100-document subsets of English Wikipedia. For convenience, we denote these two sets of documents as *set1* and *set2*. For each document, we keep only the first 150 words for our experiments. Because of the encyclopedic nature of the corpus, the first 150 words typically provides a broad overview of the themes in the article. We also removed from the corpus stop words and words which occur infrequently[2], leading to a lexicon of 8263 words. After this pruning *set1* contained 11614 words and *set2* contained 11318 words.

Workers were asked to perform twenty of the taggings described in Section 3 for each task; workers were paid $0.25 for each such task. The number of latent topics, K, is a free parameter. Here we explore two values of this parameter, $K = 10$ and $K = 15$, leading to a total of four experiments — two for each set of documents and two for each value of K.

4.1 Tagging behavior

For each experiment we issued 100 HITs, leading to a total of 2000 tags per experiment. Figure 4 shows the number of HITs performed per person in each experiment. Between 12 and 30 distinct workers participated in each experiment. The number of HITs performed per person is heavily skewed, with the most active participants completing an order of magnitude more HITs than other particpants.

Figure 5 shows the amount of time taken per tag, in log seconds. Each color represents a different experiment. The bulk of the tags took less than a minute to perform and more than a few seconds.

4.2 Comparison with LDA

Learned topics As described in Section 3, the tag-and-cluster task is a way of allowing humans to con-

[2]Infrequently occurring words were identified as those appearing fewer than eight times on a larger collection of 7726 articles.

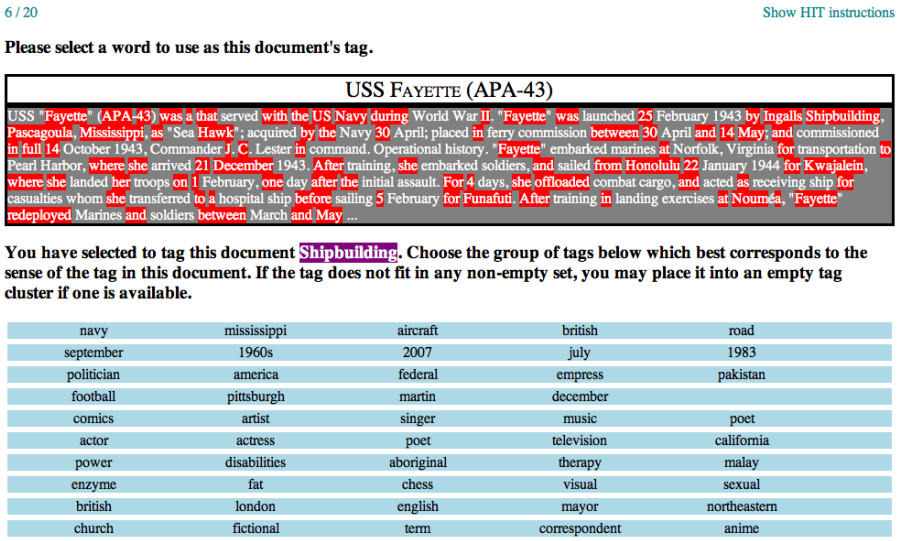
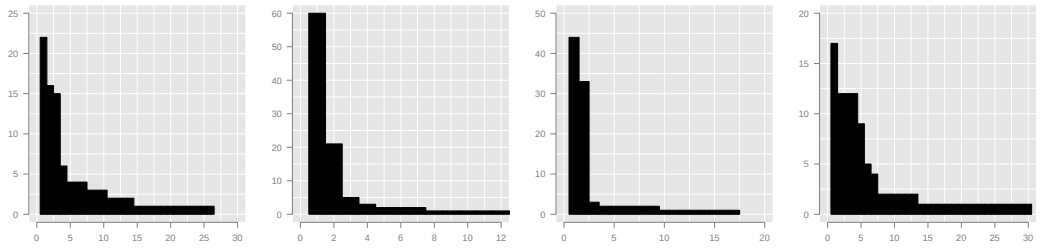
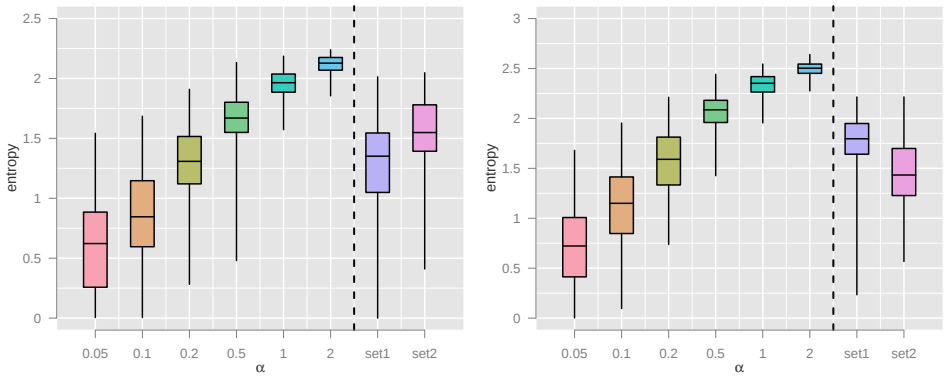


Figure 3: Screenshots of our task. In the center, the document along with its title is shown. Words which cannot be selected, e.g., distractors and words previously selected, are shown in red. Once a word is selected, the user is asked to find a topic in which to place the word. The user selects a topic by clicking on an entry in a menu of topics, where each topic is expressed by the five words which occur most frequently in that topic.



(a) *set1*, $K = 10$ (b) *set1*, $K = 15$ (c) *set2*, $K = 10$ (d) *set2*, $K = 15$

Figure 4: The number of HITs performed (y-axis) by each participant (x-axis). Between 12 and 30 people participated in each experiment.



(a) $K = 10$ (b) $K = 15$

Figure 6: A comparison of the entropy of distributions drawn from a Dirichlet distribution versus the entropy of the topic proportions inferred by workers. Each column of the boxplot shows the distribution of entropies for 100 draws from a Dirichlet distribution with parameter α. The two rightmost columns show the distribution of the entropy of the topic proportions inferred by workers on *set1* and *set2*. The α of workers typically falls between 0.2 and 0.5.

Table 1: The five words with the highest probability mass in each topic inferred by humans using the task described in Section 3. Each subtable shows the results for a particular experimental setup. Each row is a topic; the most probable words are ordered from left to right.

(a) *set1*, $K = 10$

railway	lighthouse	rail	huddersfield	station
school	college	education	history	conference
catholic	church	film	music	actor
runners	team	championships	match	racing
engine	company	power	dwight	engines
university	london	british	college	county
food	novel	book	series	superman
november	february	april	august	december
paint	photographs	american	austin	black
war	history	army	american	battle

(b) *set2*, $K = 10$

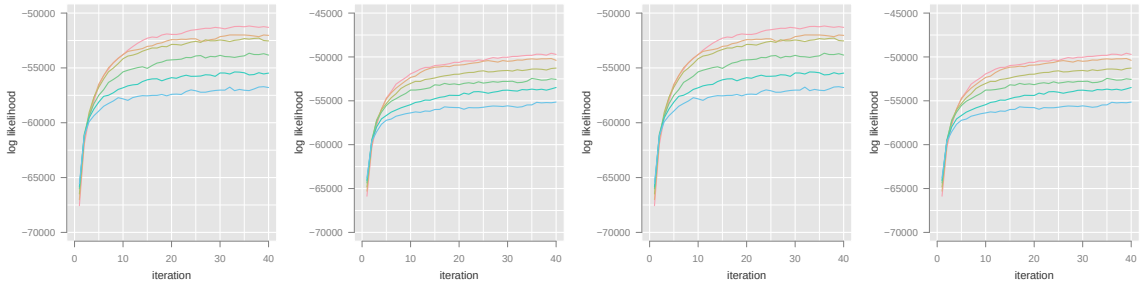
president	emperor	politician	election	government
american	players	swedish	team	zealand
war	world	navy	road	torpedo
system	pop	microsoft	music	singer
september	2007	october	december	1999
television	dog	name	george	film
people	malay	town	tribes	cliff
diet	chest	enzyme	hair	therapy
british	city	london	english	county
school	university	college	church	center

(c) *set1*, $K = 15$

australia	knee	british	israel	set
catholic	roman	island	village	columbia
john	devon	michael	austin	charles
school	university	class	community	district
november	february	2007	2009	2005
lighthouse	period	architects	construction	design
railway	rail	huddersfield	ownership	services
cyprus	archdiocese	diocese	king	miss
carson	gordon	hugo	ward	whitney
significant	application	campaign	comic	considered
born	london	american	england	black
war	defense	history	military	artillery
actor	film	actress	band	designer
york	michigan	florida	north	photographs
church	catholic	county	2001	agricultural

(d) *set2*, $K = 15$

music	pop	records	singer	artist
film	paintings	movie	painting	art
school	university	english	students	british
drama	headquarters	chess	poet	stories
family	church	sea	christmas	emperor
dog	broadcast	television	bbc	breed
champagne	regular	character	characteristic	common
election	government	parliament	minister	politician
enzyme	diet	protein	hair	oxygen
war	navy	weapons	aircraft	military
september	october	december	2008	1967
district	town	marin	america	american
car	power	system	device	devices
hockey	players	football	therapy	champions
california	zealand	georgia	india	kolkata



(a) *set1*, $K = 10$ (b) *set2*, $K = 10$ (c) *set1*, $K = 15$ (d) *set2*, $K = 15$

Figure 7: The log-likelihood achieved by LDA as a function of iteration. There is one series for each value of $\alpha \in \{0.05, 0.1, 0.2, 0.5, 1.0, 2.0\}$ from top to bottom.

Table 2: The five words with the highest probability mass in each topic inferred by LDA, with $\alpha = 0.2$. Each subtable shows the results for a particular experimental setup. Each row is a topic; the most probable words are ordered from left to right.

(a) *set1*, $K = 10$

born	2004	team	award	sydney
regiment	army	artillery	served	scouting
line	station	main	island	railway
region	street	located	site	knee
food	february	conference	day	2009
pride	greek	knowledge	portland	study
catholic	church	roman	black	time
class	series	film	actor	engine
travel	human	office	management	defense
school	born	war	world	university

(b) *set2*, $K = 10$

september	english	edit	nord	hockey
black	hole	current	england	model
training	program	war	election	navy
school	university	district	city	college
family	word	international	road	japan
publication	time	day	india	bridge
born	pop	world	released	march
won	video	microsoft	project	hungary
film	hair	bank	national	town
people	name	french	therapy	artist

(c) *set1*, $K = 15$

time	michael	written	experience	match
line	station	railway	branch	knowledge
film	land	pass	set	battle
william	florida	carson	virginia	newfoundland
war	regiment	british	army	south
reaction	terminal	copper	running	complex
born	school	world	college	black
food	conference	flight	medium	rail
township	scouting	census	square	county
travel	defense	training	management	edges
series	actor	engine	november	award
pride	portland	band	northwest	god
team	knee	2004	sydney	israel
catholic	located	site	region	church
class	february	time	public	king

(d) *set2*, $K = 15$

family	protein	enzyme	acting	oxygen
england	producer	popular	canadian	sea
system	death	artist	running	car
character	series	dark	main	village
english	word	publication	stream	day
training	program	hair	students	electrical
district	town	city	local	kolkata
september	edit	music	records	recorded
black	pop	bank	usually	hole
people	choir	road	diet	related
war	built	navy	british	service
center	million	cut	champagne	players
born	television	current	drama	won
school	university	college	election	born
film	nord	played	league	hockey

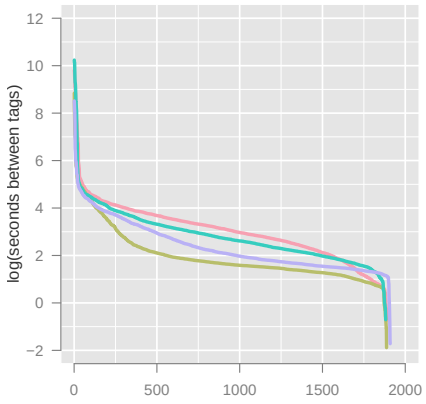


Figure 5: The distribution of times taken per HIT. Each series represents a different experiment. The bulk of the tags took less than one minute and more than a few seconds.

struct a topic model. One way of visualizing a learned topic model is by examining its topics. Table 1 shows the topics constructed by human judgments. Each subtable shows a different experimental setup and each row shows an individual topic. The five most frequently occurring words in each topic are shown, ordered from left to right.

Many of the topics inferred by humans have straightforward interpretations. For example, the {`november, february, april, august, december`} topic for the *set1* corpus with $K = 10$ is simply a collection of months. Similar topics (with years and months combined) can be found in the other experimental configurations. Other topics also cover specific semantic domains, such as {`president, emperor, politican, election, government`} or {`music, pop, records, singer, artist`}. Several of the topics are combinations of distinct concepts, such as {`catholic, church, film, music, actor`}, which is often indicative of the number of clusters, K, being too low.

Table 2 shows the topics learned by LDA under the same experimental conditions, with the Dirichlet hyperparameter $\alpha = 0.2$ (we justify this choice in the following section). These topics are more difficult to interpret than the ones created by humans. Some topics seem to largely make sense except for some anomalous words, such as {district, town, city, local, kolkata} or {school, university, college, election, born}. But the small amount of data means that it is difficult for a model which does not leverage prior knowledge to infer meaningful topic. In contrast, several humans, even working independently, can leverage prior knowledge to construct meaningful topics with little data.

There is another qualitative difference between the topics found by the tag-and-cluster task and LDA. Whereas LDA must rely on co-occurrence, humans can use ontological information. Thus, a topic which has ontological meaning, such as a list of months, may rarely be discovered by LDA since the co-occurrence patterns of months do not form a strong pattern. But users in every experimental configuration constructed this topic, suggesting that the users were consistently leveraging information that would not be available to LDA, even with a larger corpus.

Hyperparameter values A persistent question among practitioners of topic models is how to set or learn the value of the hyperparameter α. α is a Dirichlet parameter which acts as a control on sparsity — smaller values of α lead to sparser document-topic distributions. By comparing the sparsity patterns of human judgments to those of LDA for different settings of α, we can infer the value of α that would best match human judgments.

Figure 6 shows a boxplot comparison of the entropy of draws from a Dirichlet distribution (the generative process in LDA), versus the observed entropy of the models learned by humans. The first six columns show the distributions for the Dirichlet draws for various values of α; the last two columns show the observed entropy distributions on the two corpora, *set1* and *set2*.

The empirical entropy distributions across the corpora are comparable to those of a Dirichlet distribution with α between approximately 0.2 and 0.5.

Table 3: The five words with the highest probability mass in each topic inferred by LDA on *set1* with $\alpha = 0.2$, $K = 10$, and initialized using human judgments. Each row is a topic; the most probable words are ordered from left to right.

line	station	lighthouse	local	main
school	history	greek	knowledge	university
catholic	church	city	roman	york
team	club	2004	scouting	career
engine	knee	series	medium	reaction
located	south	site	land	region
food	film	conference	north	little
february	class	born	august	2009
pride	portland	time	northwest	june
war	regiment	army	civil	black

These settings of α are slightly higher than, but still in line with a common rule-of-thumb of $\alpha = 1/K$. Figure 7 shows the log-likelihood, a measure of model fit, achieved by LDA for each value of α. Higher log-likelihoods indicate better fits. Commensurate with the rule of thumb, using log-likelihoods to select α would encourage values smaller than human judgments.

However, while the entropy of the Dirichlet draws increases significantly when the number of clusters is increased from $K = 10$ to $K = 15$, the entropies assigned by humans does not vary as dramatically. This suggests that for any given document, humans are likely to pull words from only a small number of topics, regardless of how many topics are available, whereas a model will continue to spread probability mass across all topics even as the number of topics increases.

Improving LDA using human judgments The results of the previous sections suggests that human behavior differs from that of LDA, and that humans conceptualize documents in ways LDA does not. This motivates using the human judgments to augment the information available to LDA. To do so, we initialize the topic assignments used by LDA's Gibbs sampler to those made by humans. We then run the LDA sampler till convergence. This provides a method to weakly incorporate human knowledge into the model.

Table 3 shows the topics inferred by LDA when initialized with human judgments. These topics resemble those directly inferred by humans, although as we predicted in the previous sections, the topic consisting of months has largely disappeared. Other semantically coherent topics, such as {located,

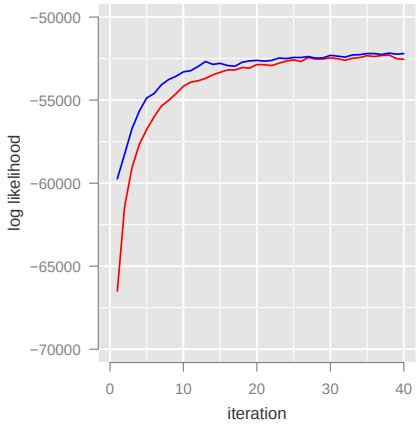


Figure 8: The log likelihood achieved by LDA on *set1* with $\alpha = 0.2$, $K = 10$, and initialized using human judgments (blue). The red line shows the log likelihood without incorporating human judgments. LDA with human judgments dominates LDA without human judgments and helps the model converge more quickly.

`south, site, land, region`}, have appeared in its place.

Figure 8 shows the log-likelihood course of LDA when initialized by human judgments (blue), versus LDA without human judgments (red). Adding human judgments strictly helps the model converge to a higher likelihood and converge more quickly. In short, incorporating human judgments shows promise at improving both the interpretability and convergence of LDA.

5 Discussion

We presented a new method for constructing topic models using human judgments. Our approach relies on a novel task, *tag-and-cluster*, which asks users to simultaneously annotate a document with one of its words and to cluster those annotations. We demonstrate using experiments on Amazon Mechanical Turk that our method constructs topic models quickly and robustly. We also show that while our topic models bear many similarities to traditionally constructed topic models, our human-learned topic models have unique features such as fixed sparsity and a tendency for topics to be constructed around concepts which models such as LDA typically fail to find.

We also underscore that the collapsed Gibbs sampling framework is expressive enough to use as the basis for human-guided topic model inference. This may motivate, as future work, the construction of different modeling assumptions which lead to sampling equations which more closely match the empirically observed sampling performed by humans. In effect, our method constructs a series of samples from the posterior, a gold standard which future topic models can aim to emulate.

Acknowledgments

The author would like to thank Eytan Bakshy and Professor Jordan Boyd-Graber-Ying for their helpful comments and discussions.

References

David Blei and John Lafferty, 2009. *Text Mining: Theory and Applications*, chapter Topic Models. Taylor and Francis.

David Blei, Andrew Ng, and Michael Jordan. 2003. Latent Dirichlet allocation. *JMLR*, 3:993–1022.

Jonathan Chang, Jordan Boyd-Graber, Sean Gerrish, Chong Wang, and David M. Blei. 2009. Reading tea leaves: How humans interpret topic models. In *NIPS*.

Scott Deerwester, Susan Dumais, Thomas Landauer, George Furnas, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6):391–407.

Thomas L. Griffiths and Mark Steyvers. 2002. A probabilistic approach to semantic representation. In *Proceedings of the 24th Annual Conference of the Cognitive Science Society*.

T. Griffiths and M. Steyvers. 2006. Probabilistic topic models. In T. Landauer, D. McNamara, S. Dennis, and W. Kintsch, editors, *Latent Semantic Analysis: A Road to Meaning*. Laurence Erlbaum.

Qiaozhu Mei, Xuehua Shen, and ChengXiang Zhai. 2007. Automatic labeling of multinomial topic models. In *KDD*.

R. Snow, Brendan O'Connor, D. Jurafsky, and A. Ng. 2008. Cheap and fast—but is it good? Evaluating non-expert annotations for natural language tasks. In *EMNLP*.

Hanna M. Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. 2009. Evaluation methods for topic models. In *ICML*.