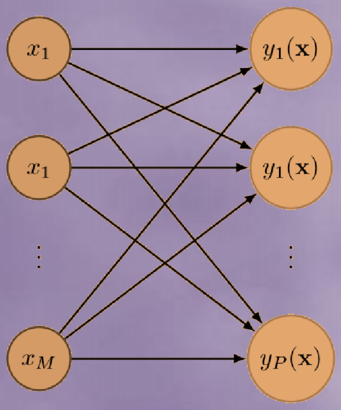


Topic Modelling

Daniel Khashabi [1]
KHASHAB2@ILLINOIS.EDU

0.1 Topic Modelling

Since the introduction of Latent Dirichlet Allocation[Blei et al.(2003)Blei, Ng, and Jordan], for latent clustering of documents, due to the flexibility of the mode, there has been numerous works on extending this work to many interesting problems, with different models, and inference methods. These applications ranged from many NLP applications, like text-author modelling, to many other applications, like in computer vision, for modelling image clustering.

Though very flexible, and easy to work in nature, for many applications, the major problem is the inclusion of semantic informations into the problem; while in many cases, one could exploit the structural differences to create clustering, there are many cases that are not easy to capture, and the differences lie in semantics. Thus, there has been great deal of efforts to make the clustering in LDA-like models more coherent and meaningful.

In topic models the goal is to develop tools for statistical analysis of document collections. There has been a lot of works mainly initiated by [Blei et al.(2003)Blei, Ng, and Jordan]; let's say we have a bunch of documents, each of which have bunch of words. The connection between these documents has some interesting properties. To create a model that could capture the mutual connection between the documents, it is assumed that set of latent variables Z which is set of topics. Each document is comprised of several topics, with some proportions. Some documents might share topics. Two documents are semantically more closer to each other if they share more common topics with similar proportions. To find the topic proportion of each document, one needs to look into the contents of the documents, i.e. words. Thus for each word there is a probability of membership to each topic. To model this, we consider a multinomial θ_i representing topic distribution for the i-th document and a multinomial ϕ_k

[1] This is part of my notes; to find the complete list of notes visit `http://web.engr.illinois.edu/~khashab2/learn.html`. This document is updated on May 11, 2013.

representing the word distribution for topic k. To make the model robust to overfitting, we put Dirichlet priors on each of the variables. Now the model could be learnt using mean-field variational approximation of the likelihood.

This document includes a comprehensive review of the related works, and their properties. In Section 2, we are summarizing the LDA-like models for relation extraction. In section, I am reviewing semi-supervision in topic modelling. Secion 3, reviews some of the related applications of topic modelling. Section 4 is about making topic modelling more semantically meaningful and coherent. And Section 5 explains the proposed model and the progress in that model.

0.2 Latent Dirichlet Allocation(LDA)

In topic models [2] the goal is to develop tools for statistical analysis of a set of documents. There has been a lot of works mainly initiated by [Blei et al.(2003)Blei, Ng, and Jordan] and some related works on probabilistic document modelling specially Latent Semantic Indexing (LSI) [Hofmann(1999)]. The graphical model for the model is shown in Figure 1 and the parameters are in Table 1. Let's say we have D documents, each of them comprised of N words. The connection between these documents has some interesting properties. To create a model that could capture the mutual connection between the documents, it is assumed that set of latent variables Z which acts like a switch which assigns topics to each words. Each documents is comprised of several topics, with some proportions. Some documents might share topics. Two documents are semantically more closer to each other if they share more common topics with similar proportion. To find the topic proportion of each document, one needs to look into the contents of the documents, i.e. words. Thus for each word there is a probability of membership to each topic. To model this, we consider a multinomial θ_i representing topic distribution for the i-th document and a multinomial ϕ_k representing the word distribution for topic k. To make the model robust to over-fitting, we put Dirichlet priors on each of the multinomials.

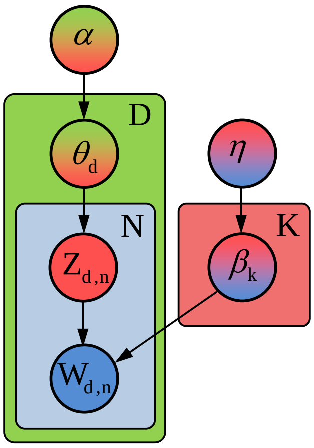


Figure 1: Graphical model for Latent Dirichlet Allocation.

As it could be seen from the model, each topic is defined over set of words. The prior over the distribution of topic in documents, is shared over all of the documents. This imposes a similarity of topic distribution among documents, but not restricted. Now we could sample this generative model (generate set of documents, topics for each document, and words generated based on the topic distributions) as follows

[2] To make explanations more clear I am using green to denote document, red for topic, and blue for word.

D	# of documents.
N	# of words.
K	# of topics.
V	size of vocabulary.
α	A positive K-vector, topic/document Dirichlet parameter.
η	A scalar positive value, word/topic Dirichlet parameter(a symmetric distribution).
θ_d	distribution of topics for the i-th document, $\theta_d \sim \text{Dir}(\alpha)$.
β_k	distribution of words given k-th topic, $\beta_k \sim \text{Dir}(\eta)$.
$Z_{d,n}$	topic assignment of words, e.g. if $Z_{i,j} = k$, the j-th word of i-th document has k-th topic. specifically we have $Z_{d,n} \sim \text{Mult}(\theta_d), Z_{d,n} \in \{1, \dots, K\}$.
$W_{d,n}$	generated word, for example $W_{i,j}$ is the j -th word from the i-th document. specifically $W_{d,n} \sim \text{Mult}(\beta_{Z_{d,n}}), W_{d,n} \in \{1, \dots, V\}$

Table 1: Parameters of LDA.

1. For each topic k,
   (a) Select a random topic proportion over words , $\beta_k \sim \text{Dir}(\eta)$.
2. For each document d,
   (a) Select a random topic proportion per documents, $\theta_d \sim \text{Dir}(\alpha)$
   (b) For each word,
      i. Select a random topic, $Z_{d,n} \sim \text{Mult}(\theta_d), Z_{d,n} \in \{1, \dots, K\}$.
      ii. Select a random word, $W_{d,n} \sim \text{Mult}(\beta_{Z_{d,n}}), W_{d,n} \in \{1, \dots, V\}$

Now having the model given, we can write the posterior over variables, given observations and hyper-parameters, could be written as following:

$$p(\theta_{1:D}, Z_{1:D,1:N}, \beta_{1:K} | W_{1:D,1:N}, \alpha, \eta) = \frac{p(\theta_{1:D}, Z_{1:D}, \beta_{1:K} | W_{1:D}, \alpha, \eta)}{\int_{\theta_{1:D}} \int_{\beta_{1:K}} \sum_{Z_{1:D}} p(\theta_{1:D}, Z_{1:D}, \beta_{1:K} | W_{1:D}, \alpha, \eta)} \tag{1}$$

The complex integration+summation at the denominator of the above formula makes it intractable to directly use the above formula for estimation of model parameters. Thus we try to devise other tricks to solve this problem.

0.2.1 Variational Bayes for LDA

In variational inference, we approximate the posterior over variables inside model, by assuming a set of independence assumptions, and decompose the posterior into smaller distributions. One decomposition for the variables inside our model could be as follows:

$$q(\theta_{1:D}, z_{1:D,1:N}, \beta_{1:K}) = \prod_{k=1:K} q(\beta_k | \lambda_k) \prod_{d=1:D} \left\{ q(\theta_d | \gamma_d) \prod_{n=1:N} q(z_{d,n} | \phi_{d,n}) \right\}$$

This decomposition decouples the dependence between variables; the decomposed graphical model is shown in Figure 2. We use the decomposed version of the posterior and minimize its and the original posterior's difference, to find the variational parameters of the approximated distribution:

$$\begin{aligned} &(\lambda^*_{1:K}, \gamma^*_{1:D}, \phi^*_{1:D,1:N}) = \\ &\qquad \arg \min_{\lambda_{1:K}, \gamma_{1:D}, \phi_{1:D,1:N}} \text{KL}\left(q(\theta_{1:D}, z_{1:D,1:N}, \beta_{1:K}) || p(\theta_{1:D}, z_{1:D,1:N}, \beta_{1:K} | w_{1:D,1:N}, \alpha, \eta) \right) \end{aligned}$$

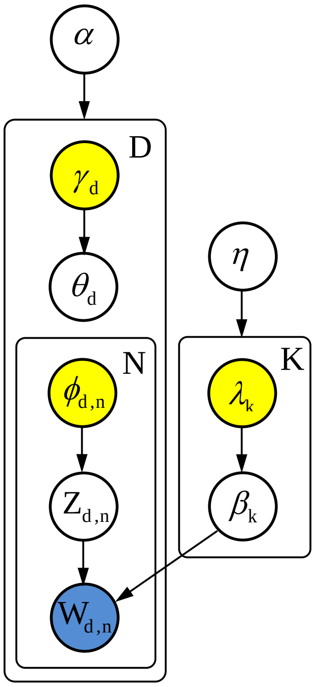


Figure 2: Decomposition of LDA; the newly added parameters are shown in yellow. The blue variable is observation.

For simplicity we drop the index of variables.

$$
\begin{aligned}
\mathcal{L} = \ln p(w|\alpha,\eta) &= \ln \int_\theta \int_\beta \sum_z p(w,\theta,\beta,z|\alpha,\eta)\mathrm{d}\theta\mathrm{d}\beta \\
&= \ln \int_\theta \int_\beta \sum_z q(\theta,\beta,z|\gamma,\phi,\lambda)\frac{p(w,\theta,\beta,z|\alpha,\eta)}{q(\theta,\beta,z|\gamma,\phi,\lambda)}\mathrm{d}\theta\mathrm{d}\beta \\
&\geq \int_\theta \int_\beta \sum_z q(\theta,\beta,z|\gamma,\phi,\lambda) \ln \frac{p(w,\theta,\beta,z|\alpha,\eta)}{q(\theta,\beta,z|\gamma,\phi,\lambda)}\mathrm{d}\theta\mathrm{d}\beta \\
&= \mathbb{E}_q\left[\ln \frac{p(w,\theta,\beta,z|\alpha,\eta)}{q(\theta,\beta,z|\gamma,\phi,\lambda)}\right] = \mathcal{L}(\gamma,\phi,\lambda;\alpha,\eta)
\end{aligned}
$$

$$
\mathcal{L}(\gamma,\phi,\lambda;\alpha,\eta) = \underbrace{\mathbb{E}_q\left[\sum_d^D \ln p(\theta_d|\alpha)\right]}_{\text{T1}} + \underbrace{\mathbb{E}_q\left[\sum_d^D\sum_n^N \ln p(z_{d,n}|\theta_d)\right]}_{\text{T2}} + \underbrace{\mathbb{E}_q\left[\sum_k^K \mathrm{ln}p(\beta_k|\eta)\right]}_{\text{T3}} +
$$

$$
+\underbrace{\mathbb{E}_q\left[\sum_d^D\sum_n^N \ln p(w_{d,n}|z_{d,n},\beta_{z_{d,n}})\right]}_{\text{T4}} - \underbrace{\mathbb{E}_q\left[\sum_d^D \mathrm{ln}q(\theta_d|\gamma_d)\right]}_{\text{T5}} - \underbrace{\mathbb{E}_q\left[\sum_k^K \mathrm{ln}q(\beta_k|\lambda_k)\right]}_{\text{T6}}
$$

$$
-\underbrace{\mathbb{E}_q\left[\sum_d^D\sum_n^N \mathrm{ln}q(z_{d,n}|\phi_{d,n})\right]}_{\text{T7}}
$$

For a Dirichlet distribution, we can find an analytic expression for the expectation of the logarithm of its variable:

$$\mathbb{E}_{\theta\sim\text{Dir}(\alpha)}[\ln\theta_k] = \psi(\alpha_k) - \psi(\sum_i \alpha_i)$$

Where $\psi(x)$ is *digamma* function,

$$\psi(x) = \frac{d}{dx}\ln\Gamma(x) = \frac{\Gamma'(x)}{\Gamma(x)}$$

And similarly

$$\mathbb{E}_{\theta\sim\text{Dir}(\alpha)}\left[\ln p(\theta|\alpha)\right] = \ln\Gamma(\sum_i \alpha_i) - \sum_i \ln\Gamma(\alpha_i) + \sum_i (\alpha_i - 1)(\psi(\alpha_i) - \psi(\sum_{\hat{i}} \alpha_{\hat{i}}))$$

Using the above identities we simplify each of the terms in the lower-bound of the likelihood:

$$\text{T1} = \sum_d^D \left[\ln\Gamma(\sum_k^K \alpha_k) - \sum_k^K \ln\Gamma(\alpha_k) + \sum_k^K (\alpha_k - 1)(\psi(\gamma_{d,k}) - \psi(\sum_{k'}^K \gamma_{d,k'}))\right]$$

$$\text{T2} = \sum_d^D \sum_n^N \sum_k^K \phi_{d,n,k}(\psi(\gamma_{d,k}) - \psi(\sum_{k'} \gamma_{d,k'}))$$

$$\text{T3} = \sum_k^K \left[\ln\Gamma(\sum_v^V \eta_v) - \sum_v^V \ln\Gamma(\eta_v) + \sum_v^V (\eta_v - 1)(\psi(\lambda_{z,v}) - \psi(\sum_{v'}^V \lambda_{z,v'}))\right]$$

$$\text{T4} = \sum_d^D \sum_n^N \sum_k^K \phi_{d,n,k}(\psi(\lambda_{k,w_{d,n}}) - \psi(\sum_{v'}^V \lambda_{k,v'}))$$

$$\text{T5} = \sum_d^D \left[\ln\Gamma(\sum_k^K \gamma_{d,k}) - \sum_k^K \ln\Gamma(\gamma_{d,k}) + \sum_k^K (\gamma_{d,k} - 1)(\psi(\gamma_{d,k}) - \psi(\sum_{k'}^K \gamma_{d,k'}))\right]$$

$$\text{T6} = \sum_k^K \left[\ln\Gamma(\sum_v^V \lambda_{k,v}) - \sum_v^V \ln\Gamma(\lambda_{k,v}) + \sum_v^V (\lambda_{k,v} - 1)(\psi(\lambda_{k,v}) - \psi(\sum_{v'}^V \lambda_{k,v'}))\right]$$

$$\text{T7} = \sum_d^D \sum_n^N \sum_k^K \phi_{d,n,k} \ln\phi_{d,n,k}$$

E-step: (equivalent to minimizing KL(.||.))

$$\frac{\partial\mathcal{L}}{\gamma_{d*,k*}} = 0 \Rightarrow \gamma_{d*,k*} = \alpha_{k*} + \sum_n^N \phi_{d*,n,k*}$$

$$\frac{\partial\mathcal{L}}{\partial\lambda_{k*,v*}} = 0 \Rightarrow \lambda_{k*,v*} = \eta_{v*} + \sum_d^D \sum_n^N \phi_{d,n,k*}\text{I}_{w_{d,n}=v*}$$

$$\frac{\partial}{\partial \phi_{d*,n*,k*}}\{\mathcal{L} - \lambda_{d*,n*}(\sum_{k} \phi_{d*,n*,k} - 1)\} = 0 \Rightarrow$$

$$\phi_{d*,n*,k*} \propto \exp\{\psi(\gamma_{d*,k*}) - \psi(\sum_{k'} \gamma_{d*,k'}) + \psi(\lambda_{k*,w_{d*,n*}}) - \psi(\sum_{v'}^{V} \lambda_{k*,v'})\}$$

M-step: Updates for the gamma distribution. The updating could be done using Newton-Raphson optimization:

$$\frac{\partial \mathcal{L}}{\partial \alpha_{k*}} = \sum_{d}^{D} [\psi(\sum_{k}^{K} \alpha_k) - \psi(\alpha_{k*}) + \psi(\gamma_{d,k*}) - \psi(\sum_{k'}^{K} \gamma_{d,k'})]$$

$$\frac{\partial^2 \mathcal{L}}{\partial \alpha_{k_1} \partial \alpha_{k_2}} = \sum_{d}^{D} [\psi'(\sum_{k}^{K} \alpha_k) - \psi'(\alpha_{k_1}) \mathrm{I}_{k_1 = k_2}]$$

Inference using Gibbs sampling

Gibbs sampling is considered as a special variant of Metropolis-Hastings algorithm, also a Monte-Carlo Markov Chain (MCMC) sampling method. We can show that,

$$p(z_i = j | \mathbf{z}_{-i}, \mathbf{w}, \alpha, \eta) \propto p(z_i = j, \mathbf{z}_{-i}, \mathbf{w} | \alpha, \eta) = p(\mathbf{z}, \mathbf{w} | \alpha, \eta)$$

The above distribution also shows the exchangeablity property for each of the random variables in $\mathbf{z}$. Now we can simplify $p(z_i = j | \mathbf{z}_{-i}, \mathbf{w}, \alpha, \eta)$ based on the words counts for topics and documents. To do so, we can expand the above distribution,

$$\begin{aligned} p(\mathbf{z}, \mathbf{w}, \alpha, \eta) &= \int \int p(\mathbf{z}, \mathbf{w}, \theta, \beta | \alpha, \eta) d\theta d\beta \\ &= \int \int p(\mathbf{z}|\theta) p(\mathbf{w}|\beta) p(\theta|\alpha) p(\beta|\eta) d\theta d\beta \\ &= \int p(\mathbf{z}|\theta) p(\theta|\alpha) d\theta \int p(\mathbf{w}|\beta) p(\beta|\eta) d\beta \\ &= \prod_{d} \frac{B(n^{d}_{.,j} + \alpha)}{B(n^{d}_{-i,j} + \alpha)} \prod_{w} \frac{B(n^{w}_{.,j} + \eta)}{B(n^{w}_{-i,j} + \eta)} \end{aligned}$$

Using a few simplification we could show that,

$$p(z_i = j | \mathbf{z}_{-i}, \mathbf{w}, \alpha, \eta) = \underbrace{\frac{n^{w_i}_{-i,j} + \eta}{\sum_{w_i} n^{w_i}_{-i,j} + W\eta}}_{\text{Probability of } w_i \text{ under topic } j} \quad \overbrace{\frac{n^{d_i}_{-i,j} + \alpha}{\sum_{d_i} n^{d_i}_{-i,j} + T\alpha}}^{\text{Probability of } z_i \text{ in document containing } w_i} \tag{2}$$

Using the above probability we can do the sampling as follows

Comparing variational procedure with Gibbs sampling, Variational procedure is faster, but Gibbs sampling is guaranteed to find the global optimum answer, and if used appropriately it is more accurate. Is also has an easy setup. But the downside with the Gibbs sampling is that, there is no concrete results on its convergence rate. In practice it might take infinite long time.

Algorithm 1: The sampling procedure for LDA

```
Data: documents and words.
Result: topic assignments $Z_{d,n}$
Randomly initialize the topic assignments $Z_{d,n}$;
while not-converged do
    for each document, $d \in \{1, \dots, D\}$ do
        for each word, $n, \in \{1, \dots, N\}$ do
            $w \leftarrow W_{d,n}$
            $z \leftarrow Z_{d,n}$
            $n_{d,z} \leftarrow 1; n_{w,z} \leftarrow 1; n_z \leftarrow 1$
            for  each topic, $K \in \{1, \dots, K\}$ do
                $p(z_i = j|\mathbf{z}_{-i}, \mathbf{w})$ using equation 2
            end
            $Z_{d,n} \leftarrow$ Sample( $p(z_i = j|\mathbf{z}_{-i}, \mathbf{w})$ )
            $z \leftarrow Z_{d,n}$
            $n_{d,z} \leftarrow n_{d,z} + 1; n_{w,z} \leftarrow n_{w,z} + 1; n_z \leftarrow n_z + 1$
        end
    end
end
```

0.3 Correlated Topic Modelling

One deficiency in the previous model defined for topic modelling, it the independence assumption between topics. The independence assumption comes from using Dirichlet prior over topics, in which the correlation between topics is not taken into account. In [Blei and Lafferty(2006)] the *Correlated Topic Models* is suggested in which, it replaces Dirichlet with *Logistic Normal Distribution*, which is also a distribution over a simplex for a richer class of distributions, and unlike Dirichlet captures better inter-component correlations [Huang and Malisiewicz(2009)]. Now it just needs to train this model similar to LDA which are extensively discussed in [Blei and Lafferty(2006)

0.4 Comparing Topic Models

In [Boyd-Graber et al.(2009)Boyd-Graber, Chang, Gerrish, Wang, and Blei] it is trying to do so, i.e. bring a good interpretation and causality behind latent variables. In this paer, using human experiments, tries to analyze the performance of each of the models for LDA, CTM and pLSI, relevance of topic models. The first is *word intrusion* which “measures how semantically cohesive the topics inferred by a model"are" and “tests whether topics correspond to natural groupings for humans". The second one is *topic intrusion* is measures “how well a topic model’s decomposition of a document as a mixture of topics agrees with human associations of topics with a document". They demonstrated that traditional metrics do not capture whether topics are coherent or not. They argue that their human-oriented experimental evaluations give better results, since the semantic aggregation of topics must be closer to human cognition, not necessarily what mathematics demands.

0.5 Supervised Topic Models

Topic Models, though being a good method, doesn’t suffice! These models are not able to give us predictions with respect to contents that they cluster. For examlpe, if we want to predict movie rates based on the a bunch of reviews. So in [Blei and McAuliffe(2008)] they assume to have a response variable for each of the contents and the goal is to infer the latent topics predictive of the response variable. However we know that in the previous works, unsupervised topic models like LDA had been used for feature selection, hence for supervised learning, but the clustering in that case work disregard of the output variable. So it might be a good idea to create a joint model for response variable and the LDA clustering. The model consists of the conventional LDA plus a response variable on words and a multivariate normal distribution. The generation from the normal distribution models the correlation between different values. To train the model, a maximum likelihood is found and is simplified using mean-field variational approximation similar to LDA.

0.6 Correspondence LDA

In the paper [Blei and Jordan(2003)], it aims to model annotated data, model the underlying correlation between samples and annotations by joint distribution of types and conditional distribution of annotated data given types. The paper gives three models: (1) A Gaussian-multinomial mixture model which assumes set of binary latent variables that are used for allocation of clusters. The model is assumed to first generate the binary latent variables and then sample the labels for each data. The second model is based on a variation of LDA [Blei et al.(2003)Blei, Ng, and Jordan], which is called Gaussian-Multinomial LDA. Unlike the previous model, this gives the ability to allocate the label each part of the given sample(like image, or a document) with different labels, with various proportions. In Correspondence Topic Models it combines the flexibility of GM-LDA with GM-Mixture. The model is so much similar to that of GM-LDA but difference is that the image regions and caption words can be considered as conditional on two disjoint sets of factors.

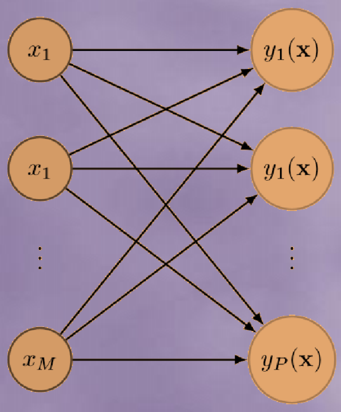
0.7 Inference for Topic Models

There has been various methods suggested for topic modelling, e.g. variational Bayes, Gibbs sampling, ML estimation, MAP estimation, etc. Some of the papers proposing these methods, claim having superior results over the other papers; while [Asuncion et al.(2009)Asuncion, Welling, Smyth, an shows that all of the methods, in fact give the same structure of answer, having only slight differences on the smoothing which is caused by model hyperparameters, i.e. by careful selection of hyperparameters all of the inference methods almost give a similar answer.

In the rest of the paper they derive the update equations for the models and show that all of the them are equivalent to each other with different hyperparameters.

0.8 Bibliographical notes

In Gibbs sampling I have used [Darling(2011)]. Thanks to Xiaolong Wang's kind helps; I used some parts of his slides.



Bibliography

[Asuncion et al.(2009)Asuncion, Welling, Smyth, and Teh] Arthur Asuncion, Max Welling, Padhraic Smyth, and Yee Whye Teh. On smoothing and inference for topic models. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 27–34. AUAI Press, 2009.

[Blei and Lafferty(2006)] D. Blei and J. Lafferty. Correlated topic models. *Advances in neural information processing systems*, 18:147, 2006.

[Blei and McAuliffe(2008)] David Blei and Jon McAuliffe. Supervised topic models. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 121–128. MIT Press, Cambridge, MA, 2008.

[Blei and Jordan(2003)] David M Blei and Michael I Jordan. Modeling annotated data. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 127–134. ACM, 2003.

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

[Boyd-Graber et al.(2009)Boyd-Graber, Chang, Gerrish, Wang, and Blei] Jonathan Boyd-Graber, Jordan Chang, Sean Gerrish, Chong Wang, and David Blei. Reading tea leaves: How humans interpret topic models. In *Proceedings of the 23rd Annual Conference on Neural Information Processing Systems*, 2009.

[Darling(2011)] William M Darling. A theoretical and practical implementation tutorial on topic modeling and gibbs sampling. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 642–647, 2011.

[Hofmann(1999)] Thomas Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57. ACM, 1999.

[Huang and Malisiewicz(2009)] JONATHAN Huang and TOMASZ Malisiewicz. Fitting a hierarchical logistic normal distribution, 2009.