In: In T. Landauer, D McNamara, S. Dennis, and W. Kintsch (eds), *Latent Semantic Analysis: A Road to Meaning*. Laurence Erlbaum

Probabilistic Topic Models

Mark Steyvers
University of California, Irvine

Tom Griffiths
Brown University

Send Correspondence to:

Mark Steyvers
Department of Cognitive Sciences
3151 Social Sciences Plaza
University of California, Irvine
Irvine, CA 92697-5100
Email: msteyver@uci.edu

1. Introduction

Many chapters in this book illustrate that applying a statistical method such as Latent Semantic Analysis (LSA; Landauer & Dumais, 1997; Landauer, Foltz, & Laham, 1998) to large databases can yield insight into human cognition. The LSA approach makes three claims: that semantic information can be derived from a word-document co-occurrence matrix; that dimensionality reduction is an essential part of this derivation; and that words and documents can be represented as points in Euclidean space. In this chapter, we pursue an approach that is consistent with the first two of these claims, but differs in the third, describing a class of statistical models in which the semantic properties of words and documents are expressed in terms of probabilistic topics.

Topic models (e.g., Blei, Ng, & Jordan, 2003; Griffiths & Steyvers, 2002; 2003; 2004; Hofmann, 1999; 2001) are based upon the idea that documents are mixtures of topics, where a topic is a probability distribution over words. A topic model is a *generative model* for documents: it specifies a simple probabilistic procedure by which documents can be generated. To make a new document, one chooses a distribution over topics. Then, for each word in that document, one chooses a topic at random according to this distribution, and draws a word from that topic. Standard statistical techniques can be used to invert this process, inferring the set of topics that were responsible for generating a collection of documents. Figure 1 shows four example topics that were derived from the TASA corpus, a collection of over 37,000 text passages from educational materials (e.g., language & arts, social studies, health, sciences) collected by Touchstone Applied Science Associates (see Landauer, Foltz, & Laham, 1998). The figure shows the sixteen words that have the highest probability under each topic. The words in these topics relate to drug use, colors, memory and the mind, and doctor visits. Documents with different content can be generated by choosing different distributions over topics. For example, by giving equal probability to the first two topics, one could construct a document about a person that has taken too many drugs, and how that affected color perception. By giving equal probability to the last two topics, one could construct a document about a person who experienced a loss of memory, which required a visit to the doctor.

Topic 247

word	prob.
DRUGS	.069
DRUG	.060
MEDICINE	.027
EFFECTS	.026
BODY	.023
MEDICINES	.019
PAIN	.016
PERSON	.016
MARIJUANA	.014
LABEL	.012
ALCOHOL	.012
DANGEROUS	.011
ABUSE	.009
EFFECT	.009
KNOWN	.008
PILLS	.008

Topic 5

word	prob.
RED	.202
BLUE	.099
GREEN	.096
YELLOW	.073
WHITE	.048
COLOR	.048
BRIGHT	.030
COLORS	.029
ORANGE	.027
BROWN	.027
PINK	.017
LOOK	.017
BLACK	.016
PURPLE	.015
CROSS	.011
COLORED	.009

Topic 43

word	prob.
MIND	.081
THOUGHT	.066
REMEMBER	.064
MEMORY	.037
THINKING	.030
PROFESSOR	.028
FELT	.025
REMEMBERED	.022
THOUGHTS	.020
FORGOTTEN	.020
MOMENT	.020
THINK	.019
THING	.016
WONDER	.014
FORGET	.012
RECALL	.012

Topic 56

word	prob.
DOCTOR	.074
DR.	.063
PATIENT	.061
HOSPITAL	.049
CARE	.046
MEDICAL	.042
NURSE	.031
PATIENTS	.029
DOCTORS	.028
HEALTH	.025
MEDICINE	.017
NURSING	.017
DENTAL	.015
NURSES	.013
PHYSICIAN	.012
HOSPITALS	.011

Figure 1. An illustration of four (out of 300) topics extracted from the TASA corpus.

Representing the content of words and documents with probabilistic topics has one distinct advantage over a purely spatial representation. Each topic is individually interpretable, providing a probability distribution over words that picks out a coherent cluster of correlated terms. While Figure 1 shows only four out of 300 topics that were derived, the topics are typically as interpretable as the ones shown here. This contrasts with the arbitrary axes of a spatial representation, and can be extremely useful in many applications (e.g., Griffiths & Steyvers, 2004; Rosen-Zvi, Griffiths, Steyvers, & Smyth, 2004; Steyvers, Smyth, Rosen-Zvi, & Griffiths, 2004).

The plan of this chapter is as follows. First, we describe the key ideas behind topic models in more detail, and outline how it is possible to identify the topics that appear in a set of documents. We then discuss methods for

answering two kinds of similarities: assessing the similarity between two documents, and assessing the associative similarity between two words. We close by considering how generative models have the potential to provide further insight into human cognition.

2. Generative Models

A generative model for documents is based on simple probabilistic sampling rules that describe how words in documents might be generated on the basis of latent (random) variables. When fitting a generative model, the goal is to find the best set of latent variables that can explain the observed data (i.e., observed words in documents), assuming that the model actually generated the data. Figure 2 illustrates the topic modeling approach in two distinct ways: as a generative model and as a problem of statistical inference. On the left, the generative process is illustrated with two topics. Topics 1 and 2 are thematically related to money and rivers and are illustrated as bags containing different distributions over words. Different documents can be produced by picking words from a topic depending on the weight given to the topic. For example, documents 1 and 3 were generated by sampling only from topic 1 and 2 respectively while document 2 was generated by an equal mixture of the two topics. Note that the superscript numbers associated with the words in documents indicate which topic was used to sample the word. The way that the model is defined, there is no notion of mutual exclusivity that restricts words to be part of one topic only. This allows topic models to capture polysemy, where the same word has multiple meanings. For example, both the money and river topic can give high probability to the word BANK, which is sensible given the polysemous nature of the word.

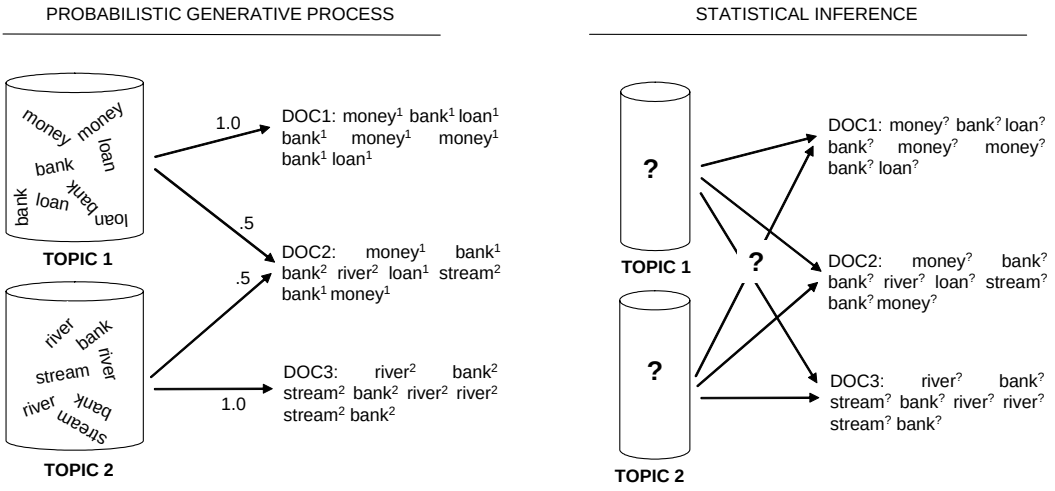


Figure 2. Illustration of the generative process and the problem of statistical inference underlying topic models

The generative process described here does not make any assumptions about the order of words as they appear in documents. The only information relevant to the model is the number of times words are produced. This is known as the *bag-of-words assumption*, and is common to many statistical models of language including LSA. Of course, word-order information might contain important cues to the content of a document and this information is not utilized by the model. Griffiths, Steyvers, Blei, and Tenenbaum (2005) present an extension of the topic model that is sensitive to word-order and automatically learns the syntactic as well as semantic factors that guide word choice (see also Dennis, this book for a different approach to this problem).

The right panel of Figure 2 illustrates the problem of statistical inference. Given the observed words in a set of documents, we would like to know what topic model is most likely to have generated the data. This involves inferring the probability distribution over words associated with each topic, the distribution over topics for each document, and, often, the topic responsible for generating each word.

3. Probabilistic Topic Models

A variety of probabilistic topic models have been used to analyze the content of documents and the meaning of words (Blei et al., 2003; Griffiths and Steyvers, 2002; 2003; 2004; Hofmann, 1999; 2001). These models all use the same fundamental idea – that a document is a mixture of topics – but make slightly different statistical assumptions. To introduce notation, we will write $P(z)$ for the distribution over topics z in a particular document and $P(w \mid z)$ for the probability distribution over words w given topic z. Several topic-word distributions $P(w \mid z)$ were illustrated in Figures 1 and 2, each giving different weight to thematically related words. Each word w_i in a document (where the index refers to the ith word token) is generated by first sampling a topic from the topic distribution, then choosing a word from the topic-word distribution. We write $P(z_i = j)$ as the probability that the jth topic was sampled for the ith word token and $P(w_i \mid z_i = j)$ as the probability of word w_i under topic j. The model specifies the following distribution over words within a document:

$$P(w_i) = \sum_{j=1}^{T} P(w_i \mid z_i = j) P(z_i = j) \tag{1}$$

where T is the number of topics. To simplify notation, let $\phi^{(j)} = P(w \mid z=j)$ refer to the multinomial distribution over words for topic j and $\theta^{(d)} = P(z)$ refer to the multinomial distribution over topics for document d. Furthermore, assume that the text collection consists of D documents and each document d consists of N_d word tokens. Let N be the total number of word tokens (i.e., $N = \Sigma N_d$). The parameters ϕ and θ indicate which words are important for which topic and which topics are important for a particular document, respectively.

Hofmann (1999; 2001) introduced the probabilistic topic approach to document modeling in his Probabilistic Latent Semantic Indexing method (pLSI; also known as the aspect model). The pLSI model does not make any assumptions about how the mixture weights θ are generated, making it difficult to test the generalizability of the model to new documents. Blei et al. (2003) extended this model by introducing a Dirichlet prior on θ, calling the resulting generative model Latent Dirichlet Allocation (LDA). As a conjugate prior for the multinomial, the Dirichlet distribution is a convenient choice as prior, simplifying the problem of statistical inference. The probability density of a T dimensional Dirichlet distribution over the multinomial distribution $p=(p_1, \ldots, p_T)$ is defined by:

$$\mathrm{Dir}(\alpha_1, \ldots, \alpha_T) = \frac{\Gamma\left(\sum_j \alpha_j\right)}{\prod_j \Gamma(\alpha_j)} \prod_{j=1}^{T} p_j^{\alpha_j - 1} \tag{2}$$

The parameters of this distribution are specified by $\alpha_1 \ldots \alpha_T$. Each hyperparameter α_j can be interpreted as a prior observation count for the number of times topic j is sampled in a document, before having observed any actual words from that document. It is convenient to use a symmetric Dirichlet distribution with a single hyperparameter α such that $\alpha_1 = \alpha_2 = \ldots = \alpha_T = \alpha$. By placing a Dirichlet prior on the topic distribution θ, the result is a smoothed topic distribution, with the amount of smoothing determined by the α parameter. Figure 3 illustrates the Dirichlet distribution for three topics in a two-dimensional simplex. The simplex is a convenient coordinate system to express all possible probability distributions -- for any point $p = (p_1, \ldots, p_T)$ in the simplex, we have $\sum_j p_j = 1$. The Dirichlet prior on the topic distributions can be interpreted as forces on the topic combinations with higher α moving the topics away from the corners of the simplex, leading to more smoothing (compare the left and right panel). For $\alpha < 1$, the modes of the Dirichlet distribution are located at the corners of the simplex. In this regime (often used in practice), there is a bias towards sparsity, and the pressure is to pick topic distributions favoring just a few topics.

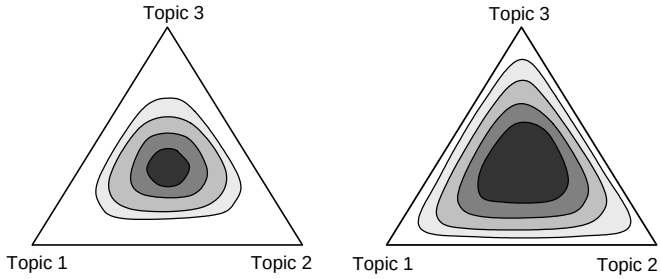


Figure 3. Illustrating the symmetric Dirichlet distribution for three topics on a two-dimensional simplex. Darker colors indicate higher probability. Left: $\alpha = 4$. Right: $\alpha = 2$.

Griffiths and Steyvers (2002; 2003; 2004) explored a variant of this model, discussed by Blei et al. (2003), by placing a symmetric Dirichlet(β) prior on ϕ as well. The hyperparameter β can be interpreted as the prior observation count on the number of times words are sampled from a topic before any word from the corpus is observed. This smoothes the word distribution in every topic, with the amount of smoothing determined by β. Good choices for the hyperparameters α and β will depend on number of topics and vocabulary size. From previous research, we have found α=50/T and β = 0.01 to work well with many different text collections.

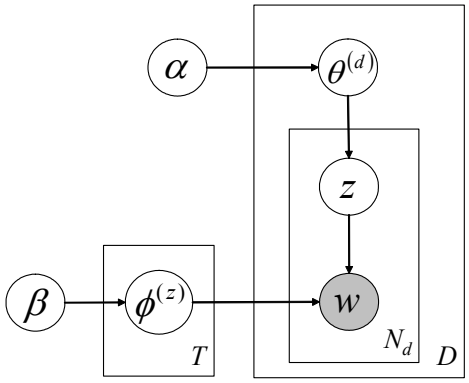


Figure 4. The graphical model for the topic model using plate notation.

Graphical Model. Probabilistic generative models with repeated sampling steps can be conveniently illustrated using plate notation (see Buntine, 1994, for an introduction). In this graphical notation, shaded and unshaded variables indicate observed and latent (i.e., unobserved) variables respectively. The variables ϕ and θ, as well as z (the assignment of word tokens to topics) are the three sets of latent variables that we would like to infer. As discussed earlier, we treat the hyperparameters α and β as constants in the model. Figure 4 shows the graphical model of the topic model used in Griffiths & Steyvers (2002; 2003; 2004). Arrows indicate conditional dependencies between variables while plates (the boxes in the figure) refer to repetitions of sampling steps with the variable in the lower right corner referring to the number of samples. For example, the inner plate over z and w illustrates the repeated sampling of topics and words until N_d words have been generated for document d. The plate surrounding $\theta^{(d)}$ illustrates the sampling of a distribution over topics for each document d for a total of D documents. The plate surrounding $\phi^{(z)}$ illustrates the repeated sampling of word distributions for each topic z until T topics have been generated.

Geometric Interpretation. The probabilistic topic model has an elegant geometric interpretation as shown in Figure 5 (following Hofmann, 1999). With a vocabulary containing W distinct word types, a W dimensional space can be constructed where each axis represents the probability of observing a particular word type. The W-1 dimensional simplex represents all probability distributions over words. In Figure 5, the shaded region is the two-dimensional simplex that represents all probability distributions over three words. As a probability distribution over words, each

document in the text collection can be represented as a point on the simplex. Similarly, each topic can also be represented as a point on the simplex. Each document that is generated by the model is a convex combination of the T topics which not only places all word distributions generated by the model as points on the W-1 dimensional simplex, but also as points on the T-1 dimensional simplex spanned by the topics. For example, in Figure 5, the two topics span a one-dimensional simplex and each generated document lies on the line-segment between the two topic locations. The Dirichlet prior on the topic-word distributions can be interpreted as forces on the topic locations with higher β moving the topic locations away from the corners of the simplex.

When the number of topics is much smaller than the number of word types (i.e., $T << W$), the topics span a low-dimensional subsimplex and the projection of each document onto the low-dimensional subsimplex can be thought of as dimensionality reduction. This formulation of the model is similar to Latent Semantic Analysis. Buntine (2002) has pointed out formal correspondences between topic models and principal component analysis, a procedure closely related to LSA.

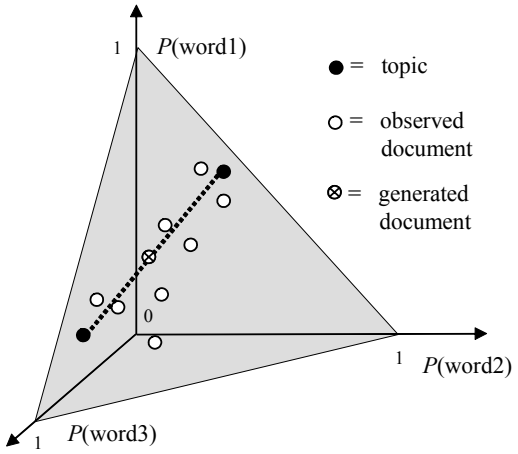


Figure 5. A geometric interpretation of the topic model.

Matrix Factorization Interpretation. In LSA, a word document co-occurrence matrix can be decomposed by singular value decomposition into three matrices (see other chapter in this book by Martin & Berry): a matrix of word vectors, a diagonal matrix with singular values and a matrix with document vectors. Figure 6 illustrates this decomposition. The topic model can also be interpreted as matrix factorization, as pointed out by Hofmann (1999). In the model described above, the word-document co-occurrence matrix is split into two parts: a topic matrix Φ and a document matrix Θ. Note that the diagonal matrix D in LSA can be absorbed in the matrix U or V, making the similarity between the two representations even clearer.

This factorization highlights a conceptual similarity between LSA and topic models, both of which find a low-dimensional representation for the content of a set of documents. However, it also shows several important differences between the two approaches. In topic models, the word and document vectors of the two decomposed matrices are probability distributions with the accompanying constraint that the feature values are non-negative and sum up to one. In the LDA model, additional a priori constraints are placed on the word and topic distributions. There is no such constraint on LSA vectors, although there are other matrix factorization techniques that require non-negative feature values (Lee & Seung, 2001). Second, the LSA decomposition provides an orthonormal basis which is computationally convenient because one decomposition for T dimensions will simultaneously give all lower dimensional approximations as well. In the topic model, the topic-word distributions are independent but not orthogonal; model inference needs to be done separately for each dimensionality.

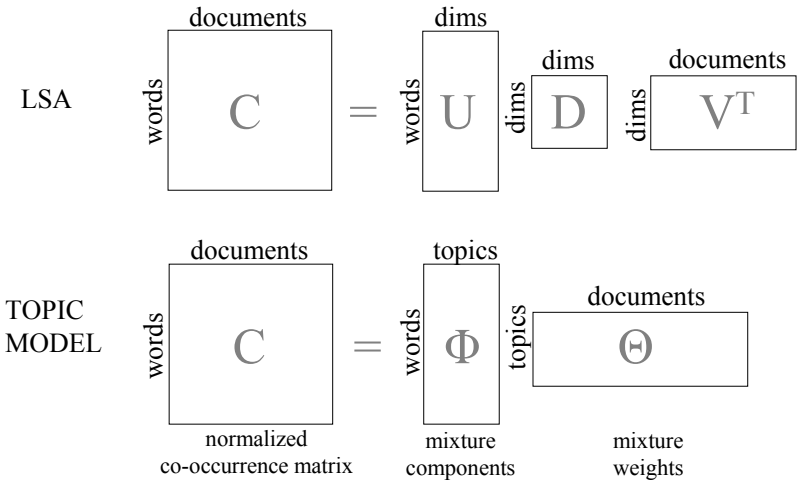


Figure 6. The matrix factorization of the LSA model compared to the matrix factorization of the topic model

Other applications. The statistical model underlying the topic modeling approach has been extended to include other sources of information about documents. For example, Cohn and Hofmann (2001) extended the pLSI model by integrating content and link information. In their model, the topics are associated not only with a probability distribution over terms, but also over hyperlinks or citations between documents. Recently, Steyvers, Smyth, Rosen-Zvi, and Griffiths (2004) and Rosen-Zvi, Griffiths, Steyvers, and Smyth (2004) proposed the author-topic model, an extension of the LDA model that integrates authorship information with content. Instead of associating each document with a distribution over topics, the author-topic model associates each author with a distribution over topics and assumes each multi-authored document expresses a mixture of the authors' topic mixtures. The statistical model underlying the topic model has also been applied to data other than text. The grade-of-membership (GoM) models developed by statisticians in the 1970s are of a similar form (Manton, Woodbury, & Tolley, 1994), and Erosheva (2002) considers a GoM model equivalent to a topic model. The same model has been used for data analysis in genetics (Pritchard, Stephens, & Donnelly, 2000).

4. Algorithm for Extracting Topics

The main variables of interest in the model are the topic-word distributions ϕ and the topic distributions θ for each document. Hofmann (1999) used the expectation-maximization (EM) algorithm to obtain direct estimates of ϕ and θ. This approach suffers from problems involving local maxima of the likelihood function, which has motivated a search for better estimation algorithms (Blei et al., 2003; Buntine, 2002; Minka, 2002). Instead of directly estimating the topic-word distributions ϕ and the topic distributions θ for each document, another approach is to directly estimate the posterior distribution over z (the assignment of word tokens to topics), given the observed words $\boldsymbol{w}$, while marginalizing out ϕ and θ. Each z_i gives an integer value [$1..T$] for the topic that word token i is assigned to. Because many text collections contain millions of word token, the estimation of the posterior over $\boldsymbol{z}$ requires efficient estimation procedures. We will describe an algorithm that uses Gibbs sampling, a form of Markov chain Monte Carlo, which is easy to implement and provides a relatively efficient method of extracting a set of topics from a large corpus (Griffiths & Steyvers, 2004; see also Buntine, 2004, Erosheva 2002 and Pritchard et al., 2000). More information about other algorithms for extracting topics from a corpus can be obtained in the references given above.

Markov chain Monte Carlo (MCMC) refers to a set of approximate iterative techniques designed to sample values from complex (often high-dimensional) distributions (Gilks, Richardson, & Spiegelhalter, 1996). Gibbs sampling (also known as alternating conditional sampling), a specific form of MCMC, simulates a high-dimensional distribution by sampling on lower-dimensional subsets of variables where each subset is conditioned on the value of all others. The sampling is done sequentially and proceeds until the sampled values approximate the target distribution. While the Gibbs procedure we will describe does not provide direct estimates of ϕ and θ, we will show how ϕ and θ can be approximated using posterior estimates of $\boldsymbol{z}$.

The Gibbs Sampling algorithm. We represent the collection of documents by a set of word indices w_i and document indices d_i, for each word token i. The Gibbs sampling procedure considers each word token in the text collection in turn, and estimates the probability of assigning the current word token to each topic, conditioned on the topic assignments to all other word tokens. From this conditional distribution, a topic is sampled and stored as the new topic assignment for this word token. We write this conditional distribution as $P(z_i = j \mid \mathbf{z}_{-i}, w_i, d_i, \cdot)$, where $z_i = j$ represents the topic assignment of token i to topic j, $\mathbf{z}_{-i}$ refers to the topic assignments of all other word tokens, and "·" refers to all other known or observed information such as all other word and document indices $\boldsymbol{w}_{-i}$ and $\boldsymbol{d}_{-i}$, and hyperparameters α, and β. Griffiths and Steyvers (2004) showed how this can be calculated by:

$$P(z_i = j \mid \mathbf{z}_{-i}, w_i, d_i, \cdot) \propto \frac{C^{WT}_{w_i j} + \beta}{\sum_{w=1}^{W} C^{WT}_{wj} + W\beta} \frac{C^{DT}_{d_i j} + \alpha}{\sum_{t=1}^{T} C^{DT}_{d_i t} + T\alpha} \tag{3}$$

where $\mathbf{C}^{WT}$ and $\mathbf{C}^{DT}$ are matrices of counts with dimensions W x T and D x T respectively; C^{WT}_{wj} contains the number of times word w is assigned to topic j, not including the current instance i and C^{DT}_{dj} contains the number of times topic j is assigned to some word token in document d, not including the current instance i. Note that Equation 3 gives the unnormalized probability. The actual probability of assigning a word token to topic j is calculated by dividing the quantity in Equation 3 for topic t by the sum over all topics T.

The factors affecting topic assignments for a particular word token can be understood by examining the two parts of Equation 3. The left part is the probability of word w under topic j whereas the right part is the probability that topic j has under the current topic distribution for document d. Once many tokens of a word have been assigned to topic j (across documents), it will increase the probability of assigning any particular token of that word to topic j. At the same time, if topic j has been used multiple times in one document, it will increase the probability that any word from that document will be assigned to topic j. Therefore, words are assigned to topics depending on how likely the word is for a topic, as well as how dominant a topic is in a document.

The Gibbs sampling algorithm starts by assigning each word token to a random topic in [1..T]. For each word token, the count matrices $\mathbf{C}^{WT}$ and $\mathbf{C}^{DT}$ are first decremented by one for the entries that correspond to the current topic assignment. Then, a new topic is sampled from the distribution in Equation 3 and the count matrices $\mathbf{C}^{WT}$ and $\mathbf{C}^{DT}$ are incremented with the new topic assignment. Each Gibbs sample consists the set of topic assignments to all N word tokens in the corpus, achieved by a single pass through all documents. During the initial stage of the sampling process (also known as the burnin period), the Gibbs samples have to be discarded because they are poor estimates of the posterior. After the burnin period, the successive Gibbs samples start to approximate the target distribution (i.e., the posterior distribution over topic assignments). At this point, to get a representative set of samples from this distribution, a number of Gibbs samples are saved at regularly spaced intervals, to prevent correlations between samples (see Gilks et al. 1996).

Estimating ϕ and θ. The sampling algorithm gives direct estimates of z for every word. However, many applications of the model require estimates ϕ' and θ' of the word-topic distributions and topic-document distributions respectively. These can be obtained from the count matrices as follows:

$$\phi'^{(j)}_i = \frac{C^{WT}_{ij} + \beta}{\sum_{k=1}^{W} C^{WT}_{kj} + W\beta} \qquad \theta'^{(d)}_j = \frac{C^{DT}_{dj} + \alpha}{\sum_{k=1}^{T} C^{DT}_{dk} + T\alpha} \tag{4}$$

These values correspond to the predictive distributions of sampling a new token of word i from topic j, and sampling a new token (as of yet unobserved) in document d from topic j, and are also the posterior means of these quantities conditioned on a particular sample $\mathbf{z}$.

An example. The Gibbs sampling algorithm can be illustrated by generating artificial data from a known topic model and applying the algorithm to check whether it is able to infer the original generative structure. We illustrate this by expanding on the example that was given in Figure 2. Suppose topic 1 gives equal probability to words MONEY, LOAN, and BANK, i.e., $\phi^{(1)}_{MONEY} = \phi^{(1)}_{LOAN} = \phi^{(1)}_{BANK} = 1/3$, while topic 2 gives equal probability to words RIVER, STREAM, and BANK, i.e., $\phi^{(2)}_{RIVER} = \phi^{(2)}_{STREAM} = \phi^{(2)}_{BANK} = 1/3$. Figure 7, top panel, shows how 16 documents can be

generated by arbitrarily mixing the two topics. Each circle corresponds to a single word token and each row to a document (for example, document 1 contains 4 times the word BANK). In Figure 7, the color of the circles indicate the topic assignments (black = topic 1; white = topic 2). At the start of sampling (top panel), the assignments show no structure yet; these just reflect the random assignments to topics. The lower panel shows the state of the Gibbs sampler after 64 iterations. Based on these assignments, Equation 4 gives the following estimates for the distributions over words for topic 1 and 2: $\phi'^{(1)}_{MONEY} = .32$, $\phi'^{(1)}_{LOAN} = .29$, $\phi'^{(1)}_{BANK} = .39$ and $\phi'^{(2)}_{RIVER} = .25$, $\phi'^{(2)}_{STREAM} = .4$, $\phi'^{(2)}_{BANK} = .35$. Given the size of the dataset, these estimates are reasonable reconstructions of the parameters used to generate the data.

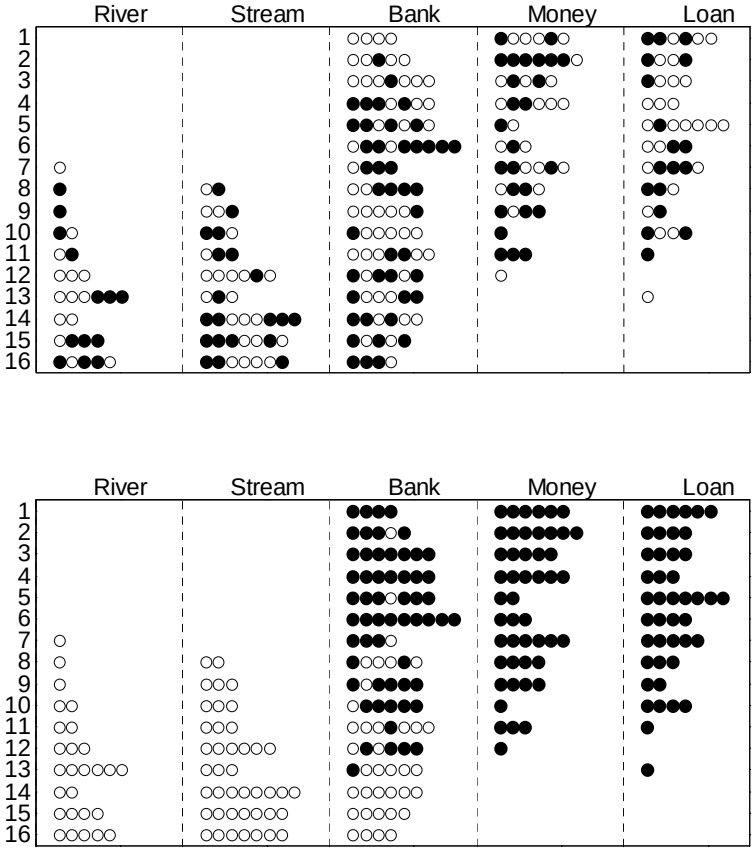


Figure 7. An example of the Gibbs sampling procedure.

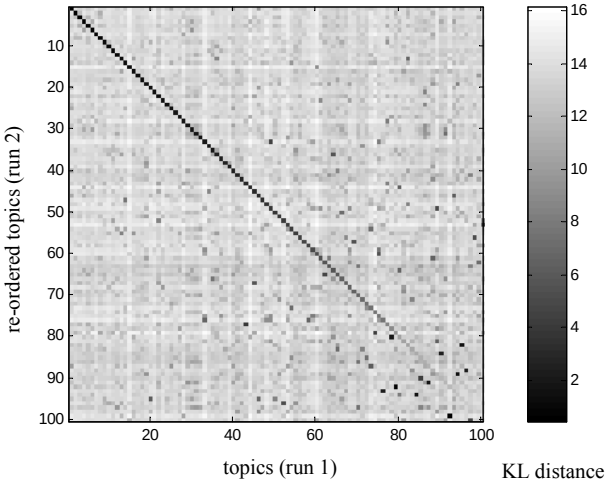
Exchangeability of topics. There is no a priori ordering on the topics that will make the topics identifiable between or even within runs of the algorithm. Topic *j* in one Gibbs sample is theoretically not constrained to be similar to topic *j* in another sample regardless of whether the samples come from the same or different Markov chains (i.e., samples spaced apart that started with the same random assignment or samples from different random assignments). Therefore, the different samples *cannot* be averaged at the level of topics. However, when topics are used to calculate a statistic which is invariant to the ordering of the topics, it becomes possible and even important to average over different Gibbs samples (see Griffiths and Steyvers, 2004). Model averaging is likely to improve results because it allows sampling from multiple local modes of the posterior.

Stability of Topics. In some applications, it is desirable to focus on a single topic solution in order to interpret each individual topic. In that situation, it is important to know which topics are stable and will reappear across samples and which topics are idiosyncratic for a particular solution. In Figure 8, an analysis is shown of the degree to which two topic solutions can be aligned between samples from different Markov chains. The TASA corpus was taken as

input (W=26,414; D=37,651; N=5,628,867; T=100; α=50/T=.5; β=.01) and a single Gibbs sample was taken after 2000 iterations for two different random initializations. The left panel shows a similarity matrix of the two topic solutions. Dissimilarity between topics j_1 and j_2 was measured by the symmetrized Kullback Liebler (KL) distance between topic distributions:

$$KL(j_1, j_2) = \frac{1}{2}\sum_{k=1}^{W} \phi'^{(j_1)}_k \log_2 \phi'^{(j_1)}_k \Big/ \phi''^{(j_2)}_k + \frac{1}{2}\sum_{k=1}^{W} \phi''^{(j_2)}_k \log_2 \phi''^{(j_2)}_k \Big/ \phi'^{(j_1)}_k \quad (5)$$

where ϕ' and ϕ'' correspond to the estimated topic-word distributions from two different runs. The topics of the second run were re-ordered to correspond as best as possible (using a greedy algorithm) with the topics in the first run. Correspondence was measured by the (inverse) sum of KL distances on the diagonal. The similarity matrix in Figure 8 suggests that a large percentage of topics contain similar distributions over words. The right panel shows the *worst* pair of aligned topics with a *KL* distance of 9.4. Both topics seem related to money but stress different themes. Overall, these results suggest that in practice, the solutions from different samples will give different results but that many topics are stable across runs.



Run 1		Run 2	
MONEY	.094	MONEY	.086
GOLD	.044	PAY	.033
POOR	.034	BANK	.027
FOUND	.023	INCOME	.027
RICH	.021	INTEREST	.022
SILVER	.020	TAX	.021
HARD	.019	PAID	.016
DOLLARS	.018	TAXES	.016
GIVE	.016	BANKS	.015
WORTH	.016	INSURANCE	.015
BUY	.015	AMOUNT	.011
WORKED	.014	CREDIT	.010
LOST	.013	DOLLARS	.010
SOON	.013	COST	.008
PAY	.013	FUNDS	.008

Figure 8. Stability of topics between different runs.

Determining the Number of Topics. The choice of the number of topics can affect the interpretability of the results. A solution with too few topics will generally result in very broad topics whereas a solution with too many topics will result in uninterpretable topics that pick out idiosyncratic word combinations. There are a number of objective methods to choose the number of topics. Griffiths and Steyvers (2004) discussed a Bayesian model selection approach. The idea is to estimate the posterior probability of the model while integrating over all possible parameter settings (i.e., all ways to assign words to topics). The number of topics is then based on the model that leads to the highest posterior probability. Another approach is to choose the number of topics that lead to best generalization performance to new tasks. For example, a topic model estimated on a subset of documents should be able to predict word choice in the remaining set of documents. In computational linguistics, the measure of perplexity has been proposed to assess generalizability of text models across subsets of documents (e.g., see Blei et al. 2003; Rosen-Zvi et al., 2004). Recently, researchers have used methods from non-parametric Bayesian statistics to define models that automatically select the appropriate number of topics (Blei, Griffiths, Jordan, & Tenenbaum, 2004; Teh, Jordan, Beal, & Blei, 2004).

5. Polysemy with Topics

Many words in natural language are polysemous, having multiple senses; their semantic ambiguity can only be resolved by other words in the context. Probabilistic topic models represent semantic ambiguity through uncertainty over topics. For example, Figure 9 shows 3 topics selected from a 300 topic solution for the TASA corpus (Figure 1

showed four other topics from this solution). In each of these topics, the word PLAY is given relatively high probability related to the different senses of the word (*playing* music, theater *play*, *playing* games).

Topic 77

word	prob.
MUSIC	.090
DANCE	.034
SONG	.033
PLAY	.030
SING	.026
SINGING	.026
BAND	.026
PLAYED	.023
SANG	.022
SONGS	.021
DANCING	.020
PIANO	.017
PLAYING	.016
RHYTHM	.015
ALBERT	.013
MUSICAL	.013

Topic 82

word	prob.
LITERATURE	.031
POEM	.028
POETRY	.027
POET	.020
PLAYS	.019
POEMS	.019
PLAY	.015
LITERARY	.013
WRITERS	.013
DRAMA	.012
WROTE	.012
POETS	.011
WRITER	.011
SHAKESPEARE	.010
WRITTEN	.009
STAGE	.009

Topic 166

word	prob.
PLAY	.136
BALL	.129
GAME	.065
PLAYING	.042
HIT	.032
PLAYED	.031
BASEBALL	.027
GAMES	.025
BAT	.019
RUN	.019
THROW	.016
BALLS	.015
TENNIS	.011
HOME	.010
CATCH	.010
FIELD	.010

Figure 9. Three topics related to the word PLAY.

Document #29795

Bix beiderbecke, at age[060] fifteen[207], sat[174] on the slope[071] of a bluff[055] overlooking[027] the mississippi[137] river[137]. He was listening[077] to music[077] coming[009] from a passing[043] riverboat. The music[077] had already captured[006] his heart[157] as well as his ear[119]. It was jazz[077]. Bix beiderbecke had already had music[077] lessons[077]. He showed[002] promise[134] on the piano[077], and his parents[035] hoped[268] he might consider[118] becoming a concert[077] pianist[077]. But bix was interested[268] in another kind[050] of music[077]. He wanted[268] to play[077] the cornet. And he wanted[268] to play[077] jazz[077]...

Document #1883

There is a simple[050] reason[106] why there are so few periods[078] of really great theater[082] in our whole western[046] world. Too many things[300] have to come right at the very same time. The dramatists must have the right actors[082], the actors[082] must have the right playhouses, the playhouses must have the right audiences[082]. We must remember[288] that plays[082] exist[143] to be performed[077], not merely[050] to be read[254]. (even when you read[254] a play[082] to yourself, try[288] to perform[062] it, to put[174] it on a stage[078], as you go along.) as soon[028] as a play[082] has to be performed[082], then some kind[126] of theatrical[082]...

Document #21359

Jim[296] has a game[166] book[254]. Jim[296] reads[254] the book[254]. Jim[296] sees[081] a game[166] for one. Jim[296] plays[166] the game[166]. Jim[296] likes[081] the game[166] for one. The game[166] book[254] helps[081] jim[296]. Don[180] comes[040] into the house[038]. Don[180] and jim[296] read[254] the game[166] book[254]. The boys[020] see a game[166] for two. The two boys[020] play[166] the game[166]. The boys[020] play[166] the game[166] for two. The boys[020] like the game[166]. Meg[282] comes[040] into the house[282]. Meg[282] and don[180] and jim[296] read[254] the book[254]. They see a game[166] for three. Meg[282] and don[180] and jim[296] play[166] the game[166]. They play[166]...

Figure 10. Three TASA documents with the word *play*.

In a new context, having only observed a single word PLAY, there would be uncertainty over which of these topics could have generated this word. This uncertainty can be reduced by observing other less ambiguous words in context. The disambiguation process can be described by the process of iterative sampling as described in the previous section (Equation 4), where the assignment of each word token to a topic depends on the assignments of the other words in the context. In Figure 10, fragments of three documents are shown from TASA that use PLAY in

three different senses. The superscript numbers show the topic assignments for each word token. The gray words are stop words or very low frequency words that were not used in the analysis. The sampling process assigns the word PLAY to topics 77, 82, and 166 in the three document contexts. The presence of other less ambiguous words (e.g., MUSIC in the first document) builds up evidence for a particular topic in the document. When a word has uncertainty over topics, the topic distribution developed for the document context is the primary factor for disambiguating the word.

6. Computing Similarities

The set of topics derived from a corpus can be used to answer questions about the similarity of words and documents: two words are similar to the extent that they appear in the same topics, and two documents are similar to the extent that the same topics appear in those documents.

Similarity between documents. The similarity between documents d_1 and d_2 can be measured by the similarity between their corresponding topic distributions $\theta^{(d_1)}$ and $\theta^{(d_2)}$. There are many choices for similarity functions between probability distributions (Lin, 1991). A standard function to measure the difference or *divergence* between two distributions p and q is the Kullback Leibler (KL) divergence,

$$D(p,q) = \sum_{j=1}^{T} p_j \log_2 \frac{p_j}{q_j} \tag{6}$$

This non-negative function is equal to zero when for all j, $p_j = q_j$. The KL divergence is asymmetric and in many applications, it is convenient to apply a symmetric measure based on KL divergence:

$$KL(p,q) = \frac{1}{2}\left[D(p,q) + D(q,p) \right] \tag{7}$$

Another option is to apply the symmetrized Jensen-Shannon (JS) divergence:

$$JS(p,q) = \frac{1}{2}\left[D\left(p,(p+q)/2\right) + D\left(q,(p+q)/2\right) \right] \tag{8}$$

which measures similarity between p and q through the average of p and q -- two distributions p and q will be similar if they are similar to their average $(p+q)/2$. Both the symmetrized KL and JS divergence functions seem to work well in practice. In addition, it is also possible to consider the topic distributions as vectors and apply geometrically motivated functions such as Euclidian distance, dot product or cosine.

For information retrieval applications, document comparison is necessary to retrieve the most relevant documents to a query. The query can be a (new) set of words produced by a user or it can be an existing document from the collection. In the latter case, the task is to find similar documents to the given document. One approach to finding relevant documents is to assess the similarity between the topic distributions corresponding to the query and each candidate documents d_i, using one of the distributional similarity functions as discussed earlier. Another approach (e.g. Buntine et al., 2004) is to model information retrieval as a probabilistic query to the topic model -- the most relevant documents are the ones that maximize the conditional probability of the query, given the candidate document. We write this as $P(\, q \mid d_i \,)$ where q is the set of words contained in the query. Using the assumptions of the topic model, this can be calculated by:

$$\begin{aligned} P(q \mid d_i) &= \prod\nolimits_{w_k \in q} P(w_k \mid d_i) \\ &= \prod\nolimits_{w_k \in q} \sum_{j=1}^{T} P\left(w_k \mid z = j\right) P\left(z = j \mid d_i\right) \end{aligned} \tag{9}$$

Note that this approach also emphasizes similarity through topics, with relevant documents having topic distributions that are likely to have generated the set of words associated with the query.

Whatever similarity or relevance function is used, it is important to obtain stable estimates for the topic distributions. This is especially important for short documents. With a single Gibbs sample, the topic distribution might be

influenced by idiosyncratic topic assignments to the few word tokens available. In that case, it becomes important to average the similarity function over multiple Gibbs samples.

Similarity between two words. The similarity between two words w_1 and w_2 can be measured by the extent that they share the same topics. Using a probabilistic approach, the similarity between two words can be calculated based on the similarity between $\theta^{(1)}$ and $\theta^{(2)}$, the conditional topic distributions for words w_1 and w_2 where $\theta^{(1)} = P(z \mid w_i = w_1)$ and $\theta^{(2)} = P(z \mid w_i = w_2)$. Either the symmetrized KL or JS divergence would be appropriate to measure the distributional similarity between these distributions.

There is an alternative approach to express similarity between two words, emphasizing the associative relations between words. The association between two words can be expressed as a conditional distribution over potential response words w_2 for cue word w_1, i.e. $P(w_2 \mid w_1)$ -- what are likely words that are generated as an associative response to another word? Much data has been collected on human word association. Typically, a cue word is presented and the subject writes down the first word that comes to mind. Nelson, McEvoy, and Schreiber (1998) have developed word association norms for over 5000 words using hundreds of subjects per cue word. In Figure 9, left panel, the distribution of human responses is shown for the cue word PLAY. The responses reveal that different subjects associate with the cue in the different senses of the word (e.g., PLAY→BALL and PLAY→ACTOR). In the topic model, word association corresponds to having observed a single word in a new context, and trying to predict new words that might appear in the same context, based on the topic interpretation for the observed word. For a particular subject who activates a single topic j, the predicted distribution for w_2 is just $P(w_2 \mid z = j)$. If it assumed that each subject activates only a single topic sampled from the distribution $P(z = j \mid w_1)$, the predictive conditional distributions can be calculated by:

$$P(w_2 \mid w_1) = \sum_{j=1}^{T} P(w_2 \mid z = j) P(z = j \mid w_1) \tag{10}$$

In human word association, high frequency words are more likely to be used as response words than low frequency words. The model captures this pattern because the left term $P(w_2 \mid z = j)$ will be influenced by the word frequency of w_2 – high frequency words (on average) have high probability conditioned on a topic.

The right panel of Figure 9 shows the predictions of the topic model for the cue word PLAY using a 300 topic solution from the TASA corpus. Griffiths and Steyvers (2002; 2003) compared the topic model with LSA in predicting word association, finding that the balance between the influence of word frequency and semantic relatedness found by the topic model can result in better performance than LSA on this task.

HUMANS		TOPICS	
FUN	.141	BALL	.036
BALL	.134	GAME	.024
GAME	.074	CHILDREN	.016
WORK	.067	TEAM	.011
GROUND	.060	WANT	.010
MATE	.027	MUSIC	.010
CHILD	.020	SHOW	.009
ENJOY	.020	HIT	.009
WIN	.020	CHILD	.008
ACTOR	.013	BASEBALL	.008
FIGHT	.013	GAMES	.007
HORSE	.013	FUN	.007
KID	.013	STAGE	.007
MUSIC	.013	FIELD	.006

Figure 9. Observed and predicted response distributions for the word PLAY.

7. Conclusion

Generative models for text, such as the topic model, have the potential to make important contributions to the statistical analysis of large document collections, and the development of a deeper understanding of human language

learning and processing. These models make explicit assumptions about the causal process responsible for generating a document, and enable the use of sophisticated statistical methods to identify the latent structure that underlies a set of words. Consequently, it is easy to explore different representations of words and documents, and to develop richer models capable of capturing more of the content of language. Topic models illustrate how using a different representation can provide new insights into the statistical modeling of language, incorporating many of the key assumptions behind LSA but making it possible to identify a set of interpretable probabilistic topics rather than a semantic space. Topic models have also been extended to capture some interesting properties of language, such as the hierarchical semantic relations between words (Blei et al., 2004), and the interaction between syntax and semantics (Griffiths et al., 2004). The vast majority of generative models are yet to be defined, and investigating these models provides the opportunity to expand both the practical benefits and the theoretical understanding of statistical language learning.

Author Note

We would like to thank Simon Dennis and Sue Dumais for thoughtful comments that improved this chapter. Matlab implementations of a variety of probabilistic topic models are available at:
http://psiexp.ss.uci.edu/research/programs_data/toolbox.htm

8. References

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research, 3,* 993-1022.

Blei, D. M., Griffiths, T. L., Jordan, M. I., & Tenenbaum, J. B. (2004). Hierarchical topic models and the nested Chinese restaurant process. In *Advances in Neural Information Processing Systems 16*. Cambridge, MA: MIT Press.

Buntine, w. (2002). Variational Extensions to EM and Multinomial PCA. In: T. Elomaa et al. (Eds.): ECML, LNAI 2430, 23–34. Springer-Verlag, Berlin.

Buntine, W.L. (1994). Operations for learning with graphical models, Journal of Artificial Intelligence Research 2, 159-225.

Buntine, W., Löfström, J., Perkiö, J., Perttu, S., Poroshin, V., Silander, T., Tirri, H., Tuominen, A., & Tuulos, V. (2004). A Scalable Topic-Based Open Source Search Engine. In: Proceedings of the IEEE/WIC/ACM Conference on Web Intelligence, 228-234.

Cohn, D. & Hofmann, T. (2001). The missing link: A probabilistic model of document content and hypertext connectivity. Neural Information Processing Systems 13, 430-436.

Erosheva, E. A. (2002). Grade of membership and latent structure models with applications to disability survey data. *Unpublished doctoral dissertation, Department of Statistics, Carnegie Mellon University.*

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (1996). *Markov chain Monte Carlo in practice*. London: Chapman & Hall.

Griffiths, T. L., & Steyvers, M. (2002). A probabilistic approach to semantic representation. In *Proceedings of the 24th Annual Conference of the Cognitive Science Society.*

Griffiths, T. L., & Steyvers, M. (2003). Prediction and semantic association. In *Neural information processing systems 15.* Cambridge, MA: MIT Press.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Science, 101,* 5228-5235.

Griffiths, T. L., Steyvers, M., Blei, D. M., & Tenenbaum, J. B. (2005) Integrating topics and syntax. In *Advances in Neural Information Processing 17*. Cambridge, MA: MIT Press.

Hofmann, T. (1999). Probabilistic Latent Semantic Analysis. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence.*

Hofmann, T. (2001). Unsupervised Learning by Probabilistic Latent Semantic Analysis. *Machine Learning Journal, 42(1)*, 177-196.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: the Latent Semantic Analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*, 211-240.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to latent semantic analysis. *Discourse Processes, 25,* 259–284.

Lee, D.D., & Seung, H.S. (2001). Algorithms for Non-negative Matrix Factorization. In: *Neural information processing systems 13*. Cambridge, MA: MIT Press.

Lin, J. (1991). Divergence measures based on Shannon entropy. *IEEE Transactions on Information Theory, 37*(14), 145–51.

Manton, K.G., Woodbury, M.A., & Tolley, H.D. (1994). *Statistical Applications Using Fuzzy Sets*. Wiley, New York.

Minka, T. & Lafferty, J. (2002). Expectation-propagation for the generative aspect model. In: *Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence*. Elsevier, New York.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). The university of south Florida word association, rhyme, and word fragment norms. (http://www.usf.edu/FreeAssociation/)

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, *155*, 945-955.

Rosen-Zvi, M., Griffiths T., Steyvers, M., & Smyth, P. (2004). The Author-Topic Model for Authors and Documents. In *20th Conference on Uncertainty in Artificial Intelligence.* Banff, Canada

Steyvers, M., Smyth, P., Rosen-Zvi, M., & Griffiths, T. (2004). Probabilistic Author-Topic Models for Information Discovery. The *Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Seattle, Washington.

Teh, Y. W., Jordan, M. I., Beal, M.J. & Blei, D. M. (2004). Hierarchical Dirichlet Processes. Technical Report 653, UC Berkeley Statistics, 2004.

Ueda, N., & Saito, K. (2003). Parametric mixture models for multi-labeled text. In *Advances in Neural Information Processing Systems 15*. Cambridge, MA: MIT Press.