Latent Dirichlet Allocation

David M. Blei, Andrew Y. Ng, Michael I. Jordan

Motivation

- “Find short descriptions of the members of a collection”
- and at the same time “enable efficient processing of large collections while preserving the essential statistical relationships.”

Tasks LDA tries to solve

- Document classification
- Novelty detection.
- Summarization.
- Similarity and relevance judgments

Intuition - What does this text talk about?

Lets assume that

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

Figure 8: An example article from the AP corpus. Each color codes a different factor from which the word is putatively generated.

Intuition - What does this text talk about?

Lets assume that

and conditional multinomial parameters for a 100-topic LDA model. The top words from some of the resulting multinomial distributions $p(w|z)$ are illustrated in Figure 8 (top). As we have hoped, these distributions seem to capture some of the underlying topics in the corpus (and we have named

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

Figure 8: An example article from the AP corpus. Each color codes a different factor from which the word is putatively generated.

Previous Works

Notation

N- Number of words in document

M-Number of documents in corpus

θ_m is the topic distribution for document *m*,

z_{mn} is the topic for the *n*-th word in document *m*, and

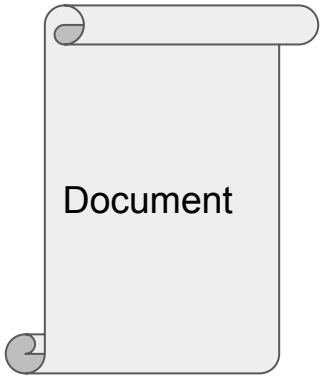
w_{mn} is the specific word.

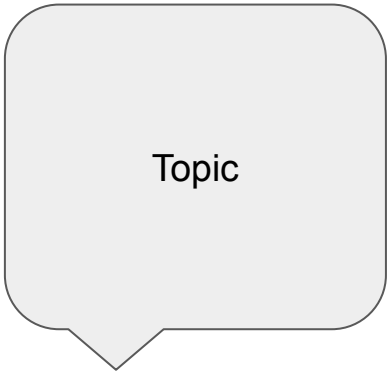
α is the parameter of the Dirichlet prior on the per-document topic distributions,

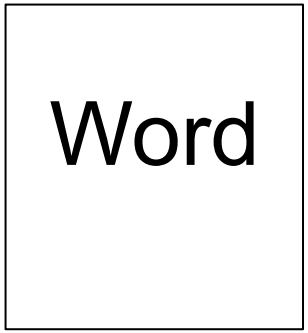
β is the parameter of the Dirichlet prior on the per-topic word distribution,

Previous works - pLSI

Mixture model







Document(d) ~ Multi(D1,...,Dn)

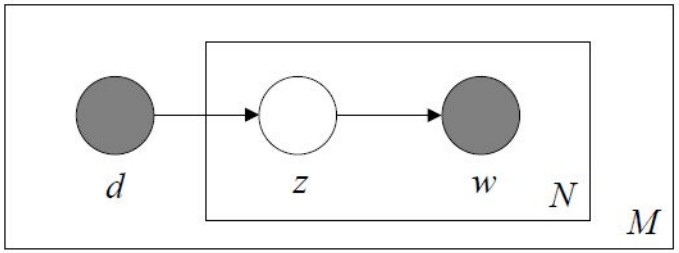
Each document as Multinomial distribution defining the topics.
p(z|d)~ Multi(z1,...,zn|D???)

Each topic as a multinomial distribution defining the words.
p(wn|z) ~ Multi(w1,...,wn|z)

Previous works - pLSI

“Document label d and a word wn are conditionally independent given an unobserved topic z:”

$$p(d, w_n) = p(d) \sum_z p(w_n | z) p(z | d).$$



(c) pLSI/aspect model

“(...) A document may contain multiple topics since p(z|d) serves as the mixture weights of the topics for a particular document d.”

Although d is a dummy variable -> generated by a Multinomial.

Model learns “inside” each document

Previous works - pLSI

Disadvantages:

- Nr of parameters = kV+ kM (k-nr of topics,V-vocabulary size, M- Nr of Docs.)
- High Nr of parameters makes the model prone to overfitting.
- Cannot be used in unseen documents.

Previous works - Dirichlet clustering

Two-level Model:

- A Dirichlet is sampled once for a corpus
  - Then a multinomial clustering variable is defined for each document.

Restriction:

- One topic per document.

It is important to distinguish LDA from a simple Dirichlet-multinomial clustering model. A classical clustering model would involve a two-level model in which a Dirichlet is sampled once for a corpus, a multinomial clustering variable is selected once for each document in the corpus, and a set of words are selected for the document conditional on the cluster variable. As with many clustering models, such a model restricts a document to being associated with a single topic. LDA, on the other hand, involves three levels, and notably the topic node is sampled *repeatedly* within the document. Under this model, documents can be associated with multiple topics.

To solve this

Latent Dirichlet Allocation is a three-level model

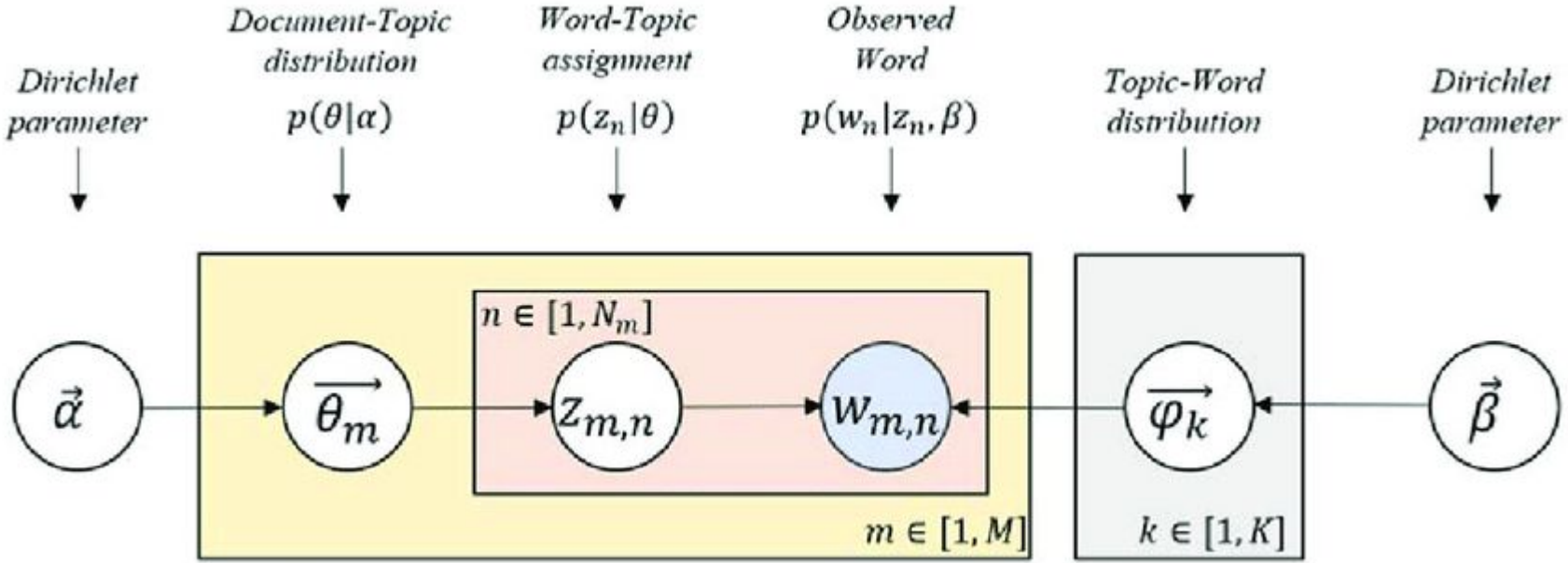
Latent Dirichlet Allocation (LDA)

Formulation

LDA assumes the following generative process for each document $\mathbf{w}$ in a corpus D:

1. Choose $N \sim \text{Poisson}(\xi)$.
2. Choose $\theta \sim \text{Dir}(\alpha)$.
3. For each of the N words w_n:
   (a) Choose a topic $z_n \sim \text{Multinomial}(\theta)$.
   (b) Choose a word w_n from $p(w_n \mid z_n, \beta)$, a multinomial probability conditioned on the topic z_n.

Graphical model



α is the parameter of the Dirichlet prior on the per-document topic distributions,

β is the parameter of the Dirichlet prior on the per-topic word distribution,

θ_m is the topic distribution for document m,

z_{mn} is the topic for the n-th word in document m, and

w_{mn} is the specific word.

Formulation

Probability of a word, topic and a prior distribution parameter theta.

Given the parameters α and β, the joint distribution of a topic mixture θ, a set of N topics $\mathbf{z}$, and a set of N words $\mathbf{w}$ is given by:

$$p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta) = p(\theta \mid \alpha) \prod_{n=1}^{N} p(z_n \mid \theta) p(w_n \mid z_n, \beta), \tag{2}$$

Go through the topics and the distributions that generate those topics

where $p(z_n|\theta)$ is simply θ_i for the unique i such that $z_n^i = 1$. Integrating over θ and summing over z, we obtain the marginal distribution of a document:

$$p(\mathbf{w}|\alpha,\beta) = \int p(\theta|\alpha)\left(\prod_{n=1}^{N}\sum_{z_n} p(z_n|\theta)p(w_n|z_n,\beta)\right)d\theta. \quad (3)$$

Margin distribution:

Joint distribution: $P(H, L)$

H \ L	Red	Yellow	Green	Marginal probability P(*H*)
Not Hit	0.198	0.09	0.14	0.428
Hit	0.002	0.01	0.56	0.572
Total	0.2	0.1	0.7	1

Go through the topics and the parameters of the multinomial distributions that generate those topics

where $p(z_n|\theta)$ is simply θ_i for the unique i such that $z_n^i = 1$. Integrating over θ and summing over z, we obtain the marginal distribution of a document:

$$p(\mathbf{w}|\alpha,\beta) = \int p(\theta|\alpha)\left(\prod_{n=1}^{N}\sum_{z_n} p(z_n|\theta)p(w_n|z_n,\beta)\right)d\theta. \quad (3)$$

Margin distribution:

Joint distribution: $P(H, L)$

H \ L	Red	Yellow	Green	Marginal probability P(*H*)
Not Hit	0.198	0.09	0.14	0.428
Hit	0.002	0.01	0.56	0.572
Total	0.2	0.1	0.7	1

Go through all the documents

Finally, taking the product of the marginal probabilities of single documents, we obtain the probability of a corpus:

$$p(D|\alpha,\beta) = \prod_{d=1}^{M} \int p(\theta_d|\alpha) \left(\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn}|\theta_d) p(w_{dn}|z_{dn},\beta) \right) d\theta_d.$$

Margin distribution:

Joint distribution: $P(H, L)$

H \ L	Red	Yellow	Green	Marginal probability P(*H*)
Not Hit	0.198	0.09	0.14	0.428
Hit	0.002	0.01	0.56	0.572
Total	0.2	0.1	0.7	1

Due to exchangeability (Finetti's Theorem)

3.1 LDA and exchangeability

A finite set of random variables $\{z_1, \ldots, z_N\}$ is said to be *exchangeable* if the joint distribution is invariant to permutation. If π is a permutation of the integers from 1 to N:

$$p(z_1, \ldots, z_N) = p(z_{\pi(1)}, \ldots, z_{\pi(N)}).$$

An infinite sequence of random variables is *infinitely exchangeable* if every finite subsequence is exchangeable.

De Finetti's representation theorem states that the joint distribution of an infinitely exchangeable sequence of random variables is as if a random parameter were drawn from some distribution and then the random variables in question were *independent* and *identically distributed*, conditioned on that parameter.

In LDA, we assume that words are generated by topics (by fixed conditional distributions) and that those topics are infinitely exchangeable within a document. By de Finetti's theorem, the probability of a sequence of words and topics must therefore have the form:

$$p(\mathbf{w}, \mathbf{z}) = \int p(\theta) \left(\prod_{n=1}^{N} p(z_n \mid \theta) p(w_n \mid z_n) \right) d\theta,$$

where θ is the random parameter of a multinomial over topics. We obtain the LDA distribution on documents in Eq. (3) by marginalizing out the topic variables and endowing θ with a Dirichlet distribution.

LDA - Inference

Inferential problem:

Compute posterior distribution of the hidden variables given a document:

$$p(\theta, \mathbf{z} \mid \mathbf{w}, \alpha, \beta) = \frac{p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta)}{p(\mathbf{w} \mid \alpha, \beta)}.$$

LDA - Inference

Inferential problem:

Compute posterior distribution of the hidden variables given a document:

$$p(\theta, \mathbf{z} \mid \mathbf{w}, \alpha, \beta) = \frac{p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta)}{p(\mathbf{w} \mid \alpha, \beta)}.$$

It is impossible to define a analytical result!

LDA - inference

ing the resulting simplified graphical model with free variational parameters, we obtain a family of distributions on the latent variables. This family is characterized by the following variational distribution:

$$q(\theta, \mathbf{z} \mid \gamma, \phi) = q(\theta \mid \gamma) \prod_{n=1}^{N} q(z_n \mid \phi_n), \tag{4}$$

So what can we do?

So what can we do?

Variational inference

Define a model that can be computed and at the same time can approximate to the original model

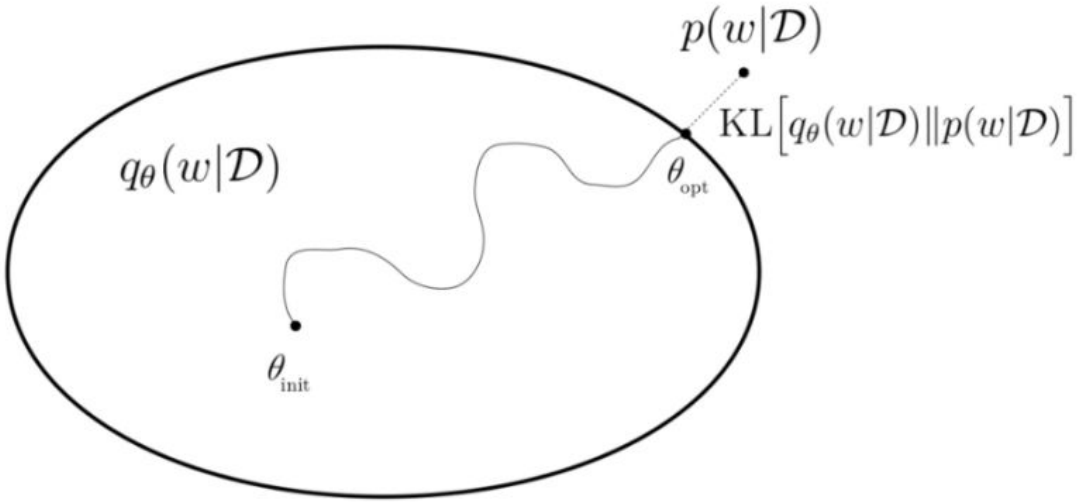


Fig. 1: Intuition of variational learning.

Source:https://medium.com/neuralspace/probabilistic-deep-learning-bayes-by-backprop-c4a3de0d9743

Variational inference

Define a model that can be computed to approximate the posterior distribution formula

Old Model

Model to approximate the Old Model

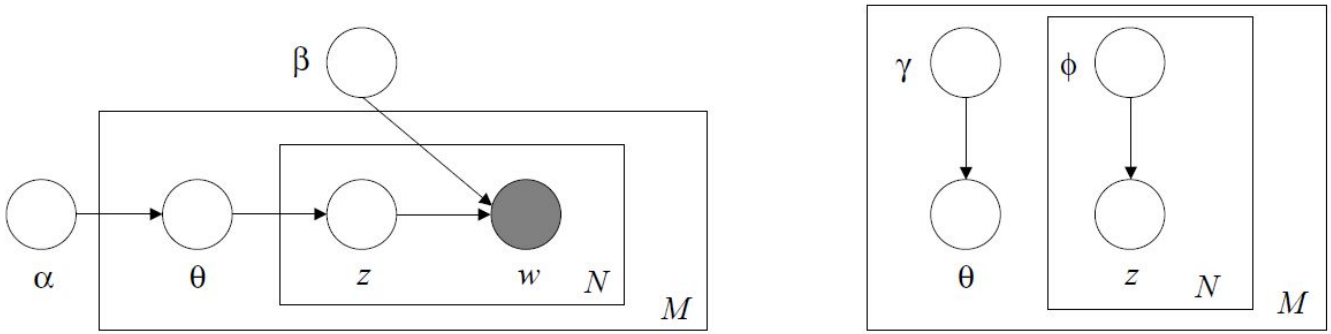
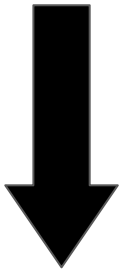


Figure 5: (Left) Graphical model representation of LDA. (Right) Graphical model representation of the variational distribution used to approximate the posterior in LDA.

Variational inference

$$p(\theta, \mathbf{z} \mid \mathbf{w}, \alpha, \beta) = \frac{p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta)}{p(\mathbf{w} \mid \alpha, \beta)}.$$

Posterior distribution



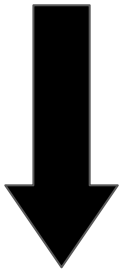
$$q(\theta, \mathbf{z} \mid \gamma, \phi) = q(\theta \mid \gamma) \prod_{n=1}^{N} q(z_n \mid \phi_n), \quad (4)$$

Approximated Model

Variational inference

$$p(\theta, \mathbf{z} \mid \mathbf{w}, \alpha, \beta) = \frac{p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta)}{p(\mathbf{w} \mid \alpha, \beta)}.$$

Posterior distribution



Independent

$$q(\theta, \mathbf{z} \mid \gamma, \phi) = q(\theta \mid \gamma) \prod_{n=1}^{N} q(z_n \mid \phi_n), \quad (4)$$

Approximated Model

Variational Inference

$$(\gamma^*, \phi^*) = \arg\min_{(\gamma,\phi)} \mathrm{D}(q(\theta, \mathbf{z} \mid \gamma, \phi) \parallel p(\theta, \mathbf{z} \mid \mathbf{w}, \alpha, \beta)). \tag{5}$$

Parameters update:

$$\phi_{ni} \propto \beta_{iw_n} \exp\{\mathrm{E}_q[\log(\theta_i) \mid \gamma]\} \tag{6}$$

$$\gamma_i = \alpha_i + \sum_{n=1}^{N} \phi_{ni}. \tag{7}$$

Variational Inference

$$(\gamma^*, \phi^*) = \arg\min_{(\gamma,\phi)} \mathrm{D}(q(\theta, \mathbf{z} \mid \gamma, \phi) \parallel p(\theta, \mathbf{z} \mid \mathbf{w}, \alpha, \beta)). \tag{5}$$

Parameters update:

$$\phi_{ni} \propto \beta_{iw_n} \exp\{\mathrm{E}_q[\log(\theta_i) \mid \gamma]\} \tag{6}$$

$$\gamma_i = \alpha_i + \sum_{n=1}^{N} \phi_{ni}. \tag{7}$$

Multinomial Update

$$\mathrm{E}_q[\log(\theta_i) \mid \gamma] = \Psi(\gamma_i) - \Psi\left(\sum_{j=1}^{k} \gamma_j\right),$$

Ψ is the first derivative of the $\log\Gamma$ function

Variational Inference

Multinomial Update

$$\phi_{ni} \propto \beta_{iw_n} \exp\{E_q[\log(\theta_i) \mid \gamma]\} \quad (6)$$

multinomial update is akin to using Bayes' theorem, $p(z_n \mid w_n) \propto p(w_n \mid z_n)p(z_n)$, where $p(z_n)$ is approximated by the exponential of the expected value of its logarithm under the variational distribution.

Variational Inference

(1) initialize $\phi_{ni}^0 := 1/k$ for all i and n
(2) initialize $\gamma_i := \alpha_i + N/k$ for all i
(3) **repeat**
(4) **for** $n = 1$ **to** N
(5) **for** $i = 1$ **to** k
(6) $\phi_{ni}^{t+1} := \beta_{iw_n} \exp(\Psi(\gamma_i^t))$ update the topic distribution in each word
(7) normalize ϕ_n^{t+1} to sum to 1.
(8) $\gamma^{t+1} := \alpha + \sum_{n=1}^{N} \phi_n^{t+1}$
(9) **until** convergence

Figure 6: A variational inference algorithm for LDA.

Variational Inference

(1) initialize $\phi_{ni}^0 := 1/k$ for all i and n
(2) initialize $\gamma_i := \alpha_i + N/k$ for all i
(3) **repeat**
(4) **for** $n = 1$ **to** N
(5) **for** $i = 1$ **to** k
(6) $\phi_{ni}^{t+1} := \beta_{iw_n} \exp(\Psi(\gamma_i^t))$
(7) normalize ϕ_n^{t+1} to sum to 1.
(8) $\gamma^{t+1} := \alpha + \sum_{n=1}^{N} \phi_n^{t+1}$
(9) **until** convergence

fi is a matrix and works like as a representation of BETA(matrix that defines topic distribution in each word)
Gama is a representation of topics distribution in a document

Figure 6: A variational inference algorithm for LDA.