Distributional Semantics

Lecture 3. Topic Modeling

Daniil Vodolazsky

March 9th, 2019

Lecture Plan

Lecture Plan

1. What Is Topic Modeling
2. Graphical Models
3. EM-algorithm
4. Probabilistic Generative Topic Model
5. Probabilistic LSA (PLSA)
6. Latent Dirichlet Allocation (LDA)
7. Additive Regularization of Topic Models (ARTM)
8. Seminar

What Is Topic Modeling

What Is Topic Modeling

President of Russia | Events | Structure | Videos and Photos | Documents | Contacts | Search

Meeting on the development of the Russia-Belarus cultural and humanitarian ties

Vladimir Putin and President of Belarus Alexander Lukashenko had a meeting at the Sirius Education Centre, where they discussed the prospects of developing bilateral cooperation in culture, education and sports.

February 15, 2019   14:00   Sochi

What is this text about?

What Is Topic Modeling

President of Russia | Events | Structure | Videos and Photos | Documents | Contacts | Search

Meeting on the development of the Russia-Belarus cultural and humanitarian ties

Vladimir Putin and President of Belarus Alexander Lukashenko had a meeting at the Sirius Education Centre, where they discussed the prospects of developing bilateral cooperation in culture, education and sports.

February 15, 2019 14:00 Sochi

What is this text about? **politics** 60%,

What Is Topic Modeling

President of Russia | Events | Structure | Videos and Photos | Documents | Contacts | Search

Meeting on the development of the Russia-Belarus cultural and humanitarian ties

Vladimir Putin and President of Belarus Alexander Lukashenko had a meeting at the Sirius Education Centre, where they discussed the prospects of developing bilateral cooperation in culture, education and sports.

February 15, 2019 14:00 Sochi

What is this text about? **politics** 60%, **education** 20%, President of Russia | Events | Structure | Videos and Photos | Documents | Contacts | Search

Meeting on the development of the Russia-Belarus cultural and humanitarian ties

Vladimir Putin and President of Belarus Alexander Lukashenko had a meeting at the Sirius Education Centre, where they discussed the prospects of developing bilateral cooperation in culture, education and sports.

February 15, 2019 14:00 Sochi

What is this text about? **politics** 60%, **education** 20%, **culture** 10%,

What Is Topic Modeling

President of Russia | Events Structure Videos and Photos Documents Contacts Search

Meeting on the development of the Russia-Belarus cultural and humanitarian ties

Vladimir Putin and President of Belarus Alexander Lukashenko had a meeting at the Sirius Education Centre, where they discussed the prospects of developing bilateral cooperation in culture, education and sports.

February 15, 2019 14:00 Sochi

What is this text about? **politics** 60%, **education** 20%, **culture** 10%, **sports** 10%, President of Russia | Events | Structure | Videos and Photos | Documents | Contacts | Search

Meeting on the development of the Russia-Belarus cultural and humanitarian ties

Vladimir Putin and President of Belarus Alexander Lukashenko had a meeting at the Sirius Education Centre, where they discussed the prospects of developing bilateral cooperation in culture, education and sports.

February 15, 2019 14:00 Sochi

What is this text about? **politics** 60%, **education** 20%, **culture** 10%, **sports** 10%, **economics** 0%.

What Is Topic Modeling

What is the difference between topic modeling and classification?

What Is Topic Modeling

What is the difference between topic modeling and classification?

Unspecified in advance topics!

Topic Modeling in Social Media

Актуальные темы: Россия · Изменить

- **Доренко**
- **Сталина**
  Твитов: 8 256
- **Через 10**
  Твитов: 2 008
- **Дудя**
- **На Кубани**
- **Самарской**
  Твитов: 1 480
- **Международному**
- **запад**
  Твитов: 1 413
- **В Кремле**
- **Сергей**
  Твитов: 4 158

Актуальные темы: США · Изменить

- **#FatTuesday**
  Твитов: 6 886
- **#TuesdayThoughts**
  Твитов: 43 тыс.
- **King Kong Bundy**
  Wrestling legend King Kong Bundy has died, aged 61
- **#PancakeDay**
  Твитов: 39,4 тыс.
- **#TuesdayMotivation**
  Твитов: 18,9 тыс.
- **#PaczkiDay**
- **Ty Cobb**
- **Good Tuesday**
  Твитов: 11,1 тыс.
- **McCarthyism**
  Твитов: 25 тыс.
- **Former Trump White House**
  Твитов: 2 317

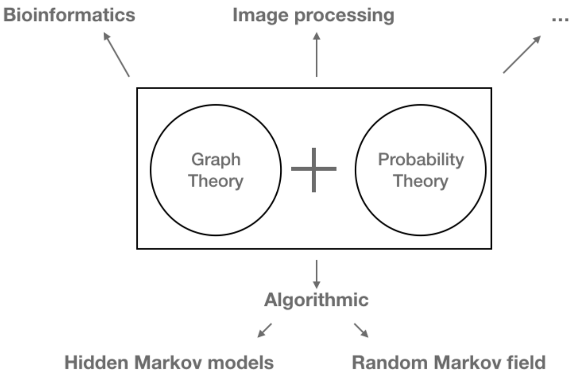
Актуальные темы: Франция · Изменить

- **#MardiGras**
  Твитов: 14,6 тыс.
- **#MardiConseil**
  Твитов: 1 883
- **#lesplanetes**
  Твитов: 4 012
- **Hugo Clément**
  Твитов: 2 220
- **#GILyon2019**
  Твитов: 1 504
- **Vieilles Charrues**
- **Condé-sur-Sarthe**
  Твитов: 1 836
- **Sonic**
  Твитов: 114 тыс.
- **Black Eyed Peas**
  Твитов: 2 013
- **Bigard**
  Твитов: 11,4 тыс.

Graphical Models

Introduction to Graphical Models

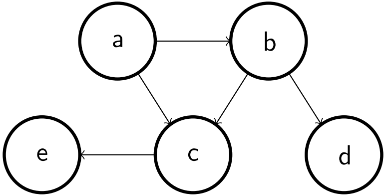
Graphical models bring together graph theory and probability theory in a powerful formalism for multivariate statistical modeling.



Directed Models

A **directed model** contains one factor for every random variable x_i in the distribution, and that factor consists of the conditional distribution over x_i given the parents of x_i, denoted $Pa_{\mathcal{G}}(x_i)$:

$$p(\mathbf{x}) = \prod_i p(x_i \mid Pa_{\mathcal{G}}(x_i)).$$

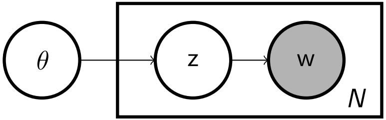


This graph corresponds to probability distributions that can be factored as follows:

$$p(a, b, c, d, e) = p(a)p(b \mid a)p(c \mid a, b)p(d \mid b)p(e \mid c).$$

Observable and Latent Variables

Observable variables are variables that can be observed and directly measured.
Latent (hidden) variables are variables that are not directly observed but are rather inferred from other variables that are observed (directly measured).



θ is a parameter, $z_1, \ldots, z_N$ are hidden (latent) variables, and $w_1, \ldots, w_N$ are observable variables.

Given θ, the joint probability for such model can be computes as follows:

$$p(\mathbf{w}, \mathbf{z} \mid \theta) = \prod_{i=1}^{N} p(z_i \mid \theta) p(w_i \mid z_i).$$

EM-algorithm

EM-algorithm

EM-algorithm is a powerful method that is widely used in probabilistic models. We will use $\mathbf{X}$ to denote the observable variables, $\mathbf{Z}$ to denote the latent variables, and $\mathbf{\Theta}$ to denote the model parameters (that we want to estimate).

Assume that we know how to compute a joint distribution over $\mathbf{X}$ and $\mathbf{Z}$ given $\mathbf{\Theta}$: $p(\mathbf{X}, \mathbf{Z} \mid \mathbf{\Theta})$. Let's write a log-likelihood function as an expectation over $\mathbf{Z}$:

$$\log p(\mathbf{X} \mid \mathbf{\Theta}) = \int q(\mathbf{Z}) \log p(\mathbf{X} \mid \mathbf{\Theta}) d\mathbf{Z}$$

where $q(\mathbf{Z})$ is an arbitrary probability density for $\mathbf{Z}$.

EM-algorithm

$$\log p(\mathbf{X} \mid \boldsymbol{\Theta}) = \int q(\mathbf{Z}) \log p(\mathbf{X} \mid \boldsymbol{\Theta}) d\mathbf{Z} =$$

EM-algorithm

$$\log p(\mathbf{X} \mid \boldsymbol{\Theta}) = \int q(\mathbf{Z}) \log p(\mathbf{X} \mid \boldsymbol{\Theta}) d\mathbf{Z} = \int q(\mathbf{Z}) \log \frac{p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\Theta})}{p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\Theta})} \frac{q(\mathbf{Z})}{q(\mathbf{Z})} d\mathbf{Z}$$

EM-algorithm

$$\log p(\mathbf{X} \mid \boldsymbol{\Theta}) = \int q(\mathbf{Z}) \log p(\mathbf{X} \mid \boldsymbol{\Theta}) d\mathbf{Z} = \int q(\mathbf{Z}) \log \frac{p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\Theta})}{p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\Theta})} \frac{q(\mathbf{Z})}{q(\mathbf{Z})} d\mathbf{Z}$$

$$= \int q(\mathbf{Z}) \log \frac{p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\Theta})}{q(\mathbf{Z})} d\mathbf{Z} + \int q(\mathbf{Z}) \log \frac{q(\mathbf{Z})}{p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\Theta})} d\mathbf{Z}$$

$$
\begin{aligned}
\log p(\mathbf{X} \mid \mathbf{\Theta}) &= \int q(\mathbf{Z}) \log p(\mathbf{X} \mid \mathbf{\Theta}) d\mathbf{Z} = \int q(\mathbf{Z}) \log \frac{p(\mathbf{X}, \mathbf{Z} \mid \mathbf{\Theta})}{p(\mathbf{Z} \mid \mathbf{X}, \mathbf{\Theta})} \frac{q(\mathbf{Z})}{q(\mathbf{Z})} d\mathbf{Z} \\
&= \int q(\mathbf{Z}) \log \frac{p(\mathbf{X}, \mathbf{Z} \mid \mathbf{\Theta})}{q(\mathbf{Z})} d\mathbf{Z} + \int q(\mathbf{Z}) \log \frac{q(\mathbf{Z})}{p(\mathbf{Z} \mid \mathbf{X}, \mathbf{\Theta})} d\mathbf{Z} \\
&= \mathscr{L}(q, \mathbf{\Theta}) + D_{\mathrm{KL}}(q(\mathbf{Z}) \| p(\mathbf{Z} \mid \mathbf{X}, \mathbf{\Theta})) \geq \mathscr{L}(q, \mathbf{\Theta}).
\end{aligned}
$$

EM-algorithm

$$
\begin{aligned}
\log p(\mathbf{X} \mid \boldsymbol{\Theta}) &= \int q(\mathbf{Z}) \log p(\mathbf{X} \mid \boldsymbol{\Theta}) d\mathbf{Z} = \int q(\mathbf{Z}) \log \frac{p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\Theta})}{p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\Theta})} \frac{q(\mathbf{Z})}{q(\mathbf{Z})} d\mathbf{Z} \\
&= \int q(\mathbf{Z}) \log \frac{p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\Theta})}{q(\mathbf{Z})} d\mathbf{Z} + \int q(\mathbf{Z}) \log \frac{q(\mathbf{Z})}{p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\Theta})} d\mathbf{Z} \\
&= \mathscr{L}(q, \boldsymbol{\Theta}) + D_{\mathrm{KL}}(q(\mathbf{Z}) \| p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\Theta})) \geq \mathscr{L}(q, \boldsymbol{\Theta}).
\end{aligned}
$$

A full log-likelihood function is hard to optimize. Instead of maximizing $\log p(\mathbf{X} \mid \boldsymbol{\Theta})$ by $\boldsymbol{\Theta}$, we are going to maximize $\mathscr{L}(q, \boldsymbol{\Theta})$ by $q, \boldsymbol{\Theta}$. Notice that $\log p(\mathbf{X} \mid \boldsymbol{\Theta})$ does not depend on q so we are not limited to choose q. Thus we can turn the last inequality into equality by putting $q(\mathbf{Z}) = p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\Theta})$ so that $D_{\mathrm{KL}}(q(\mathbf{Z}) \| p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\Theta})) = 0$.

EM-algorithm

- **E-step** creates a function for the expectation of the log-likelihood evaluated using the current estimate for the parameters:

$$q(\mathbf{Z})^{(n+1)} = \arg\max_{q'} \mathscr{L}(q', \boldsymbol{\Theta}^{(n)}) = p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\Theta}^{(n)}).$$

- **M-step** computes parameters maximizing the expected log-likelihood found on the E-step:

$$\begin{aligned}
\boldsymbol{\Theta}^{(n+1)} &= \arg\max_{\boldsymbol{\Theta}'} \mathscr{L}(q^{(n+1)}, \boldsymbol{\Theta}') \\
&= \arg\max_{\boldsymbol{\Theta}'} \int q(\mathbf{Z})^{(n+1)} \log \frac{p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\Theta}')}{q(\mathbf{Z})^{(n+1)}} d\mathbf{Z} \\
&= \arg\max_{\boldsymbol{\Theta}'} \int q(\mathbf{Z})^{(n+1)} \log p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\Theta}') d\mathbf{Z} \\
&= \arg\max_{\boldsymbol{\Theta}'} \mathbb{E}_{\mathbf{Z} \sim q(\mathbf{Z})^{(n+1)}} [\log p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\Theta}')].
\end{aligned}$$

Probabilistic Generative Topic Model

Probabilistic Generative Topic Model

The joint probability of a document d and a word w can be expressed with conditional probability formula as

$$P(d, w) = P(d)P(w \mid d).$$

Now we will define a model with hidden variables $\mathcal{Z} = \{z_1, \ldots, z_T\}$. We assume that each document has its on distribution over (latent) topics, and that each topic has its on distribution over (observable) words. So our goal is to find the matrices $\mathbf{\Phi}$ and $\mathbf{\Theta}$ such that

$$P(w \mid d) = \sum_{z \in \mathcal{Z}} P(w \mid z, d)P(z \mid d) = \sum_{z \in \mathcal{Z}} P(w \mid z)P(z \mid d) = \sum_{z \in \mathcal{Z}} \Phi_{w,z}\Theta_{z,d}.$$

In other words, we want to get a low-rank matrix factorization.

Probabilistic Generative Topic Model

A log-likelihood function for our model:

$$\log P(\mathcal{D}, \mathcal{W} \mid \boldsymbol{\Phi}, \boldsymbol{\Theta}) = \log \prod_{d \in \mathcal{D}} \prod_{w \in \mathcal{W}} P(d, w)^{\#(w,d)}$$

Probabilistic Generative Topic Model

A log-likelihood function for our model:

$$\begin{aligned}
\log P(\mathcal{D}, \mathcal{W} \mid \mathbf{\Phi}, \mathbf{\Theta}) &= \log \prod_{d \in \mathcal{D}} \prod_{w \in \mathcal{W}} P(d, w)^{\#(w,d)} \\
&= \sum_{d \in \mathcal{D}} \sum_{w \in \mathcal{W}} \#(w, d) \log P(d, w)
\end{aligned}$$

Probabilistic Generative Topic Model

A log-likelihood function for our model:

$$\begin{aligned}\log P(\mathcal{D}, \mathcal{W} \mid \mathbf{\Phi}, \mathbf{\Theta}) &= \log \prod_{d \in \mathcal{D}} \prod_{w \in \mathcal{W}} P(d, w)^{\#(w,d)} \\ &= \sum_{d \in \mathcal{D}} \sum_{w \in \mathcal{W}} \#(w, d) \log P(d, w) \\ &= \sum_{d \in \mathcal{D}} \sum_{w \in \mathcal{W}} \#(w, d) \log(P(d) P(w \mid d))\end{aligned}$$

Probabilistic Generative Topic Model

A log-likelihood function for our model:

$$
\begin{aligned}
\log P(\mathcal{D}, \mathcal{W} \mid \mathbf{\Phi}, \mathbf{\Theta}) &= \log \prod_{d \in \mathcal{D}} \prod_{w \in \mathcal{W}} P(d, w)^{\#(w,d)} \\
&= \sum_{d \in \mathcal{D}} \sum_{w \in \mathcal{W}} \#(w, d) \log P(d, w) \\
&= \sum_{d \in \mathcal{D}} \sum_{w \in \mathcal{W}} \#(w, d) \log(P(d) P(w \mid d)) \\
&= \sum_{d \in \mathcal{D}} \sum_{w \in \mathcal{W}} \#(w, d) \log \left(P(d) \sum_{z \in \mathcal{Z}} \Phi_{w,z} \Theta_{z,d} \right) \longrightarrow \max_{\mathbf{\Phi}, \mathbf{\Theta}},
\end{aligned}
$$

which is an objective function with restrictions:

$$
\sum_{w \in \mathcal{W}} \Phi_{w,z} = 1,\ \Phi_{w,z} \geq 0,\ \sum_{z \in \mathcal{Z}} \Theta_{z,d} = 1,\ \Theta_{z,d} \geq 0.
$$

Probabilistic Generative Topic Model

If $P(d)$ does not depend on $\mathbf{\Phi}$ or $\mathbf{\Theta}$, then the problem is equivalent to

$$\sum_{d \in \mathcal{D}} \sum_{w \in \mathcal{W}} \#(w, d) \log \sum_{z \in \mathcal{Z}} \Phi_{w,z} \Theta_{z,d} \longrightarrow \max_{\mathbf{\Phi}, \mathbf{\Theta}} .$$

Probabilistic Generative Topic Model

If $P(d)$ does not depend on $\mathbf{\Phi}$ or $\mathbf{\Theta}$, then the problem is equivalent to

$$\sum_{d \in \mathcal{D}} \sum_{w \in \mathcal{W}} \#(w, d) \log \sum_{z \in \mathcal{Z}} \Phi_{w,z} \Theta_{z,d} \longrightarrow \max_{\mathbf{\Phi}, \mathbf{\Theta}}.$$

A point $(\mathbf{\Phi}, \mathbf{\Theta})$ of a local extremum can be found as a solution of the system:

$$\begin{aligned} P(z \mid d, w) &\propto \Phi_{w,z} \Theta_{z,d}, \\ \Phi_{w,z} &\propto \sum_{d \in \mathcal{D}} \#(w, d) P(z \mid d, w), \\ \Theta_{z,d} &\propto \sum_{w \in \mathcal{W}} \#(w, d) P(z \mid d, w). \end{aligned}$$

It means that on an E-step we should compute $P(z \mid d, w)$, and on an M-step we should compute $P(w \mid z) = \Phi_{w,z}$ and $P(z \mid d) = \Theta_{z,d}$ according to the formulas obtained.

PLSA

PLSA

Probabilistic latent semantic analysis (PLSA) (1999) was the first technique in history for probabilistic topic modeling. The PLSA model is a latent variable model for co-occurrence data which associates an unobserved class variable $z \in \mathcal{Z} = \{z_1, \dots, z_T\}$ with each observation. A joint probability model over $\mathcal{D} \times \mathcal{W}$ is defined by the mixture

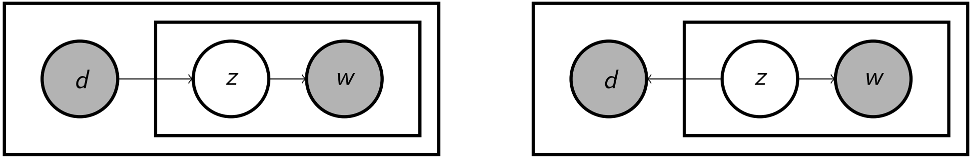
$$P(d, w) = P(d)P(w \mid d), \; P(w \mid d) = \sum_{z \in \mathcal{Z}} P(w \mid z)P(z \mid d).$$

PLSA

Since the cardinality of z is smaller than the number of documents/words in the collection, z acts as a bottleneck variable in predicting words. It is worth noticing that the model can be equivalently parameterized by:

$$P(d, w) = \sum_{z \in \mathcal{Z}} P(z)P(d \mid z)P(w \mid z)$$

which is perfectly symmetric in both entities, documents and words.



(a) non-symmetrical (b) symmetrical

Two versions of the PLSA model.

PLSA

PLSA model uses EM algorithm for maximum likelihood estimation.

- **E-step**. For each $z \in \mathcal{Z}$, $d \in \mathcal{D}$, $w \in \mathcal{W}$ compute

$$P(z \mid d, w) = \frac{P(d, z, w)}{P(d, w)} = \frac{P(z)P(d \mid z)P(w \mid z)}{\sum\limits_{z' \in \mathcal{Z}} P(z')P(d \mid z')P(w \mid z')}.$$

- **M-step** For each $z \in \mathcal{Z}$, $d \in \mathcal{D}$, $w \in \mathcal{W}$ compute

$$P(w \mid z) \propto \sum_{d \in \mathcal{D}} \#(w, d)P(z \mid d, w),$$

$$P(d \mid z) \propto \sum_{w \in \mathcal{W}} \#(w, d)P(z \mid d, w),$$

$$P(z) \propto \sum_{d \in \mathcal{D}} \sum_{w \in \mathcal{W}} \#(w, d)P(z \mid d, w).$$

PLSA: Relation to LSA

Define matrices by $\hat{\boldsymbol{U}} : \hat{U}_{i,k} = P(w^{(i)} \mid z_k)$, $\hat{\boldsymbol{\Sigma}} : \hat{\Sigma}_{k,k} = P(z_k)$, $\hat{\boldsymbol{V}}^\top : \hat{V}^\top_{k,j} = P(d^{(j)} \mid z_k)$.

PLSA: Relation to LSA

Define matrices by $\hat{\boldsymbol{U}} : \hat{U}_{i,k} = P(w^{(i)} \mid z_k)$, $\hat{\boldsymbol{\Sigma}} : \hat{\Sigma}_{k,k} = P(z_k)$, $\hat{\boldsymbol{V}}^\top : \hat{V}^\top_{k,j} = P(d^{(j)} \mid z_k)$. The joint probability model $\boldsymbol{P}$ can then be written as a matrix product $\boldsymbol{P} = \hat{\boldsymbol{U}}\hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{V}}^\top$.

PLSA: Relation to LSA

Define matrices by $\hat{\boldsymbol{U}} : \hat{U}_{i,k} = P(w^{(i)} \mid z_k)$, $\hat{\boldsymbol{\Sigma}} : \hat{\Sigma}_{k,k} = P(z_k)$, $\hat{\boldsymbol{V}}^\top : \hat{V}^\top_{k,j} = P(d^{(j)} \mid z_k)$. The joint probability model $\boldsymbol{P}$ can then be written as a matrix product $\boldsymbol{P} = \hat{\boldsymbol{U}}\hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{V}}^\top$.

The crucial difference between PLSA and LSA, however, is the objective function utilized to determine the optimal decomposition/approximation.

- In LSA, this is the L_2 or Frobenius norm, which corresponds to an implicit additive Gaussian noise assumption on counts.
- In contrast, PLSA relies on the likelihood function of multinomial sampling and aims at an explicit maximization of the predictive power of the model.

LDA

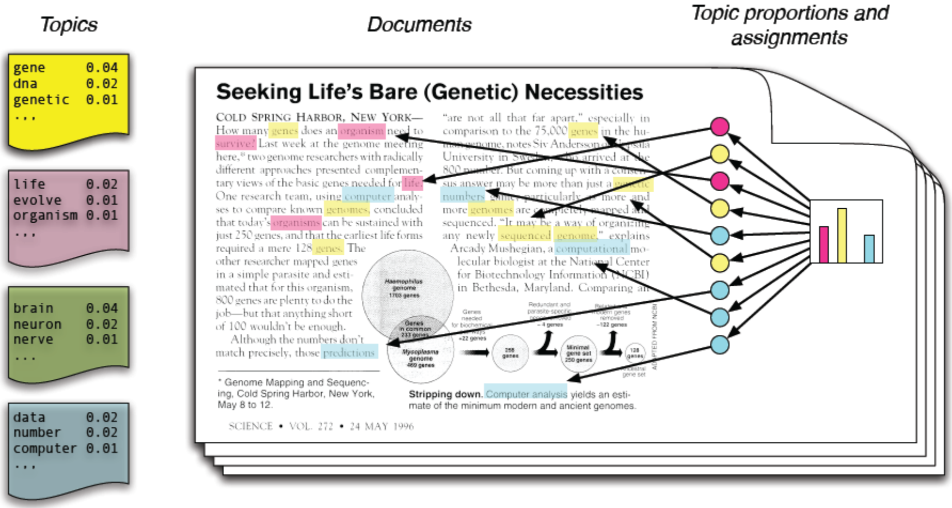
LDA

Latent Dirichlet allocation (LDA) (2003) is a topic modeling algorithm based on a generative graphical model. The concept of LDA is to consider all documents as composed of set of topics or clusters of words and to observe the distribution of them.

LDA

- We have a document collection $\mathcal{D}$ and a vocabulary $\mathcal{W}$.
- Each document $d \in \mathcal{D}$ consists of N_d words $w_{d,1}, \ldots, w_{d,N_d}$.
- The probability distributions over topics $\mathcal{Z}$ are sampled for each document $d \in \mathcal{D}$ from a Dirichlet distribution with parameter $\boldsymbol{\alpha}$.
- For each topic $z \in \mathcal{Z}$ we generate a probability distribution over all words $\mathcal{W}$ $\Phi_{:,z}$.
- After that, for each document d we generate N_d topics $z_{d,1}, \ldots, z_{d,N_d}$ from distribution $\Theta_{:,d}$, and for each of these topics $z_{d,k}$ generate per a single word from distribution $\Phi_{:,z_{d,k}}$.

LDA



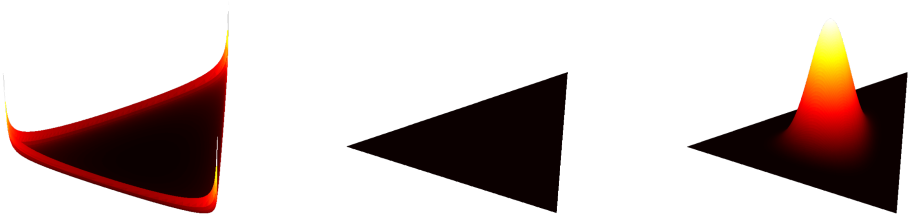
Topics
Documents
Topic proportions and assignments
gene 0.04
dna 0.02
genetic 0.01
...
life 0.02
evolve 0.01
organism 0.01
...
brain 0.04
neuron 0.02
nerve 0.01
...
data 0.02
number 0.02
computer 0.01
...
Seeking Life's Bare (Genetic) Necessities
COLD SPRING HARBOR, NEW YORK—
How many genes does an organism need to
survive? Last week at the genome meeting
here,* two genome researchers with radically
different approaches presented complemen-
tary views of the basic genes needed for life.
One research team, using computer analy-
ses to compare known genomes, concluded
that today's organisms can be sustained with
just 250 genes, and that the earliest life forms
required a mere 128 genes. The
other researcher mapped genes
in a simple parasite and esti-
mated that for this organism,
800 genes are plenty to do the
job—but that anything short
of 100 wouldn't be enough.
Although the numbers don't
match precisely, those predictions
"are not all that far apart," especially in
comparison to the 75,000 genes in the hu-
Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 469 genes
Minimal gene set 250 genes
* Genome Mapping and Sequenc-
ing, Cold Spring Harbor, New York,
May 8 to 12.
Stripping down. Computer analysis yields an esti-
mate of the minimum modern and ancient genomes.
SCIENCE • VOL. 272 • 24 MAY 1996

Dirichlet Distribution

A K-dimensional Dirichlet random variable $\boldsymbol{\theta}$ can take values in the $(K-1)$-simplex (a K-vector $\boldsymbol{\theta}$ lies in the $(K-1)$-simplex if $\theta_k \geq 0$ and $\sum_{k=1}^{K} \theta_k = 1$), and has the following probability density on this simplex:

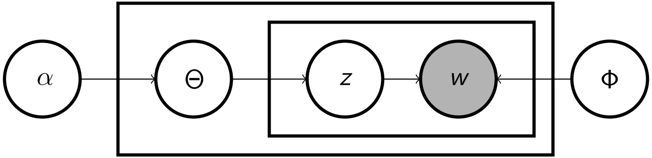
$$p(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) = \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}$$

where the parameter $\boldsymbol{\alpha}$ is a K-vector with components $\alpha_k > 0$, and where $\Gamma(x)$ is the Gamma function.



Dirichlet distribution plots for $\boldsymbol{\alpha} = (0.75, 0.75, 0.75), (1, 1, 1), (10, 10, 10)$

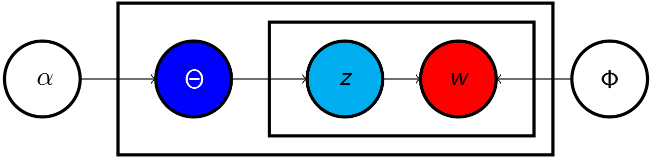
LDA Model



A graphical model for LDA.

In this model, $\boldsymbol{\alpha}$ and $\boldsymbol{\Phi}$ are parameters. $\boldsymbol{\Phi}$ is a matrix of size $|\mathcal{W}| \times |\mathcal{Z}|$, $\boldsymbol{\Theta}$ is a matrix of size $|\mathcal{Z}| \times |\mathcal{D}|$, $\boldsymbol{\alpha}$ is a vector of size $|\mathcal{Z}|$. Moreover, $\boldsymbol{\Theta}_{:,d} \sim Dir(\boldsymbol{\alpha})$.

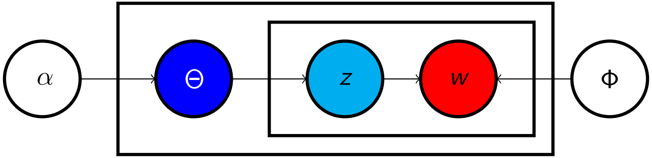
LDA Model



A graphical model for LDA.

$$P(\mathcal{D}, \mathcal{W}, \mathcal{Z}, \boldsymbol{\Theta} \mid \boldsymbol{\Phi}, \boldsymbol{\alpha}) =$$

LDA Model

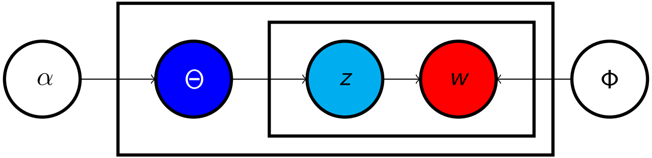


A graphical model for LDA.

$$P(\mathcal{D}, \mathcal{W}, \mathcal{Z}, \boldsymbol{\Theta} \mid \boldsymbol{\Phi}, \boldsymbol{\alpha}) = \underbrace{\prod_{d=1}^{|\mathcal{D}|}}_{1}$$

- 1. for each document

LDA Model

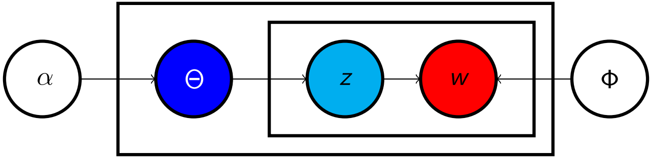


A graphical model for LDA.

$$P(\mathcal{D}, \mathcal{W}, \mathcal{Z}, \boldsymbol{\Theta} \mid \boldsymbol{\Phi}, \boldsymbol{\alpha}) = \underbrace{\prod_{d=1}^{|\mathcal{D}|}}_{1} \underbrace{P(\boldsymbol{\Theta}_{:,d} \mid \boldsymbol{\alpha})}_{2}$$

- 1. for each document
- 2. generate topic probabilities

LDA Model

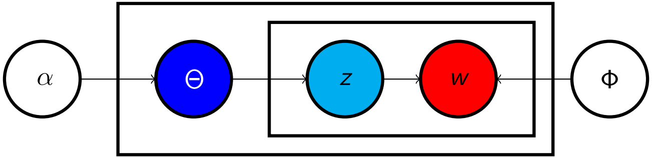


A graphical model for LDA.

$$P(\mathcal{D}, \mathcal{W}, \mathcal{Z}, \boldsymbol{\Theta} \mid \boldsymbol{\Phi}, \boldsymbol{\alpha}) = \underbrace{\prod_{d=1}^{|\mathcal{D}|}}_{1} \underbrace{P(\boldsymbol{\Theta}_{:,d} \mid \boldsymbol{\alpha})}_{2} \underbrace{\prod_{n=1}^{N_d}}_{3}$$

- 1. for each document
- 2. generate topic probabilities
- 3. for each word

LDA Model

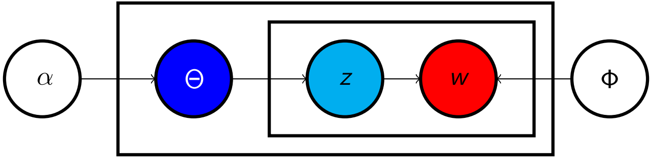


A graphical model for LDA.

$$P(\mathcal{D}, \mathcal{W}, \mathcal{Z}, \boldsymbol{\Theta} \mid \boldsymbol{\Phi}, \boldsymbol{\alpha}) = \underbrace{\prod_{d=1}^{|\mathcal{D}|}}_{1} \underbrace{P(\boldsymbol{\Theta}_{:,d} \mid \boldsymbol{\alpha})}_{2} \underbrace{\prod_{n=1}^{N_d}}_{3} \underbrace{P(z_{d,n} \mid \boldsymbol{\Theta}_{:,d})}_{4}$$

- 1. for each document
- 2. generate topic probabilities
- 3. for each word
- 4. select topic

LDA Model



A graphical model for LDA.

$$P(\mathcal{D}, \mathcal{W}, \mathcal{Z}, \boldsymbol{\Theta} \mid \boldsymbol{\Phi}, \boldsymbol{\alpha}) = \underbrace{\prod_{d=1}^{|\mathcal{D}|}}_{1} \underbrace{P(\boldsymbol{\Theta}_{:,d} \mid \boldsymbol{\alpha})}_{2} \underbrace{\prod_{n=1}^{N_d}}_{3} \underbrace{P(z_{d,n} \mid \boldsymbol{\Theta}_{:,d})}_{4} \underbrace{P(w_{d,n} \mid z_{d,n}, \boldsymbol{\Phi})}_{5} \longrightarrow \max_{\boldsymbol{\Phi}, \cancel{\boldsymbol{\alpha}}}$$

- 1. for each document
- 2. generate topic probabilities
- 3. for each word
- 4. select topic
- 5. select word from topic

LDA Model

$$P(\mathcal{D}, \mathcal{W}, \mathcal{Z}, \boldsymbol{\Theta} \mid \boldsymbol{\Phi}, \boldsymbol{\alpha}) = \prod_{d=1}^{|\mathcal{D}|} P(\boldsymbol{\Theta}_{:,d} \mid \boldsymbol{\alpha}) \prod_{n=1}^{N_d} P(z_{d,n} \mid \boldsymbol{\Theta}_{:,d}) P(w_{d,n} \mid z_{d,n}, \boldsymbol{\Phi})$$

LDA Model

$$
\begin{aligned}
P(\mathcal{D}, \mathcal{W}, \mathcal{Z}, \boldsymbol{\Theta} \mid \boldsymbol{\Phi}, \boldsymbol{\alpha}) &= \prod_{d=1}^{|\mathcal{D}|} P(\boldsymbol{\Theta}_{:,d} \mid \boldsymbol{\alpha}) \prod_{n=1}^{N_d} P(z_{d,n} \mid \boldsymbol{\Theta}_{:,d}) P(w_{d,n} \mid z_{d,n}, \boldsymbol{\Phi}) \\
&= \prod_{d=1}^{|\mathcal{D}|} \frac{\Gamma(\sum_{t=1}^{|\mathcal{Z}|} \alpha_t)}{\prod_{t=1}^{|\mathcal{Z}|} \Gamma(\alpha_t)} \prod_{t=1}^{|\mathcal{Z}|} \Theta_{t,d}^{\alpha_t - 1} \prod_{n=1}^{N_d} \Theta_{z_{d,n},d} \Phi_{w_{d,n}, z_{d,n}}
\end{aligned}
$$

LDA Model

$$
\begin{aligned}
P(\mathcal{D}, \mathcal{W}, \mathcal{Z}, \boldsymbol{\Theta} \mid \boldsymbol{\Phi}, \boldsymbol{\alpha}) &= \prod_{d=1}^{|\mathcal{D}|} P(\boldsymbol{\Theta}_{:,d} \mid \boldsymbol{\alpha}) \prod_{n=1}^{N_d} P(z_{d,n} \mid \boldsymbol{\Theta}_{:,d}) P(w_{d,n} \mid z_{d,n}, \boldsymbol{\Phi}) \\
&= \prod_{d=1}^{|\mathcal{D}|} \frac{\Gamma(\sum_{t=1}^{|\mathcal{Z}|} \alpha_t)}{\prod_{t=1}^{|\mathcal{Z}|} \Gamma(\alpha_t)} \prod_{t=1}^{|\mathcal{Z}|} \Theta_{t,d}^{\alpha_t - 1} \prod_{n=1}^{N_d} \Theta_{z_{d,n},d} \Phi_{w_{d,n},z_{d,n}} \\
&= \prod_{d=1}^{|\mathcal{D}|} \frac{\Gamma(\sum_{t=1}^{|\mathcal{Z}|} \alpha_t)}{\prod_{t=1}^{|\mathcal{Z}|} \Gamma(\alpha_t)} \prod_{t=1}^{|\mathcal{Z}|} \Theta_{t,d}^{\alpha_t - 1} \prod_{n=1}^{N_d} \prod_{t=1}^{|\mathcal{Z}|} \Theta_{t,d}^{[z_{d,n}=t]} \Phi_{w_{d,n},t}^{[z_{d,n}=t]} .
\end{aligned}
$$

EM-algorithm for LDA

- **E-step**.

$$P(\boldsymbol{\Theta}, \mathcal{Z} \mid \mathcal{D}, \mathcal{W}, \boldsymbol{\Phi}, \boldsymbol{\alpha}) \approx q(\boldsymbol{\Theta})q(\mathcal{Z}) = \arg \min_{q(\boldsymbol{\Theta}), q(\mathcal{Z})} D_{\mathrm{KL}}(q(\boldsymbol{\Theta})q(\mathcal{Z}) || P(\boldsymbol{\Theta}, \mathcal{Z} \mid \mathcal{D}, \mathcal{W}, \boldsymbol{\Phi}, \boldsymbol{\alpha})).$$

- **M-step**.

$$\boldsymbol{\Phi} = \arg \max_{\boldsymbol{\Phi}} \mathbb{E}_{\boldsymbol{\Theta} \sim q(\boldsymbol{\Theta}), \mathcal{Z} \sim q(\mathcal{Z})} \log P(\boldsymbol{\Theta}, \mathcal{Z} \mid \mathcal{D}, \mathcal{W}, \boldsymbol{\Phi}, \boldsymbol{\alpha}).$$

Questions?

Seminar

Topic Modeling on Hillary Clinton Emails

1. Reading data
2. Preprocessing
3. LSA
4. LDA
5. Homework: PLSA

https://www.kaggle.com/s231644/topic-modeling-on-hillary-clinton-s-emails

▸ Link