

Statistical Learning for Text Data Analytics

Latent Dirichlet Allocation

Yangqiu Song

Hong Kong University of Science and Technology

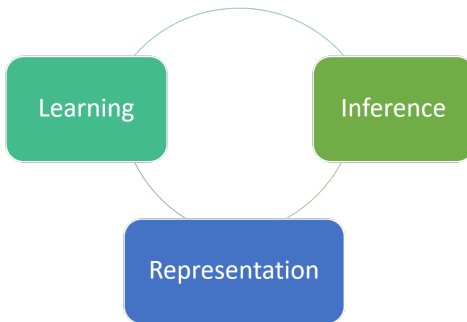
yqsong@cse.ust.hk

Spring 2018

*Contents are based on materials created by Noah Smith, Xiaojin (Jerry) Zhu, Chengxiang Zhai, David Mackay, Yoav Goldberg

- Noah Smith. CSE 517: Natural Language Processing
<https://courses.cs.washington.edu/courses/cse517/16wi/>
- Xiaojin (Jerry) Zhu. CS 769: Advanced Natural Language Processing.
<http://pages.cs.wisc.edu/~jerryzhu/cs769.html>
- Yoav Goldberg. Introduction to Natural Language Processing.
<http://u.cs.biu.ac.il/~89-680/>

Course Topics



- Representation: language models, word embeddings, **topic models**
- Learning: supervised learning, semi-supervised learning, sequence models, deep learning, optimization techniques
- Inference: constraint modeling, joint inference, search algorithms

NLP applications: tasks introduced in Lecture 1

Overview

- 1 Language Models: Recap
- 2 Topic Models
- 3 Probabilistic Latent Semantic Analysis (PLSA)
- 4 Latent Dirichlet Allocation (LDA)**
 - Motivation: Bayesian Modeling
 - **Background of Monte Carlo Methods**
 - Important Sampling
 - Rejection Sampling
 - Metropolis Methods
 - **Gibbs Sampling**
 - Sampling for EM Algorithm
 - Collapsed Gibbs Sampling for LDA

Gibbs Sampling

- In the general case of a system with K variables, a single iteration involves sampling one parameter at a time:

- $x_1^{(t+1)} \sim P(x_2^{(t)}, x_3^{(t)}, \dots, x_K^{(t)})$
- $x_2^{(t+1)} \sim P(x_1^{(t+1)}, x_3^{(t)}, \dots, x_K^{(t)})$
- $x_3^{(t+1)} \sim P(x_1^{(t+1)}, x_2^{(t+1)}, \dots, x_K^{(t)})$
- ...
- $x_K^{(t+1)} \sim P(x_1^{(t+1)}, x_2^{(t+1)}, \dots, x_{K-1}^{(t+1)})$

- Denote $\mathbf{x}_{\setminus k}^{(t)} = P(x_1^{(t+1)}, x_2^{(t+1)}, \dots, x_{k-1}^{(t+1)}, x_{k+1}^{(t)}, \dots, x_K^{(t)})$

- Gibbs sampling can be viewed as a Metropolis method

$$\begin{aligned} a_G &= \frac{P^*(\mathbf{x}')Q(\mathbf{x}^{(t)}|\mathbf{x}')}{P^*(\mathbf{x}^{(t)})Q(\mathbf{x}'|\mathbf{x}^{(t)})} = \frac{P(\mathbf{x}')P(x_k^{(t)}|\mathbf{x}'_{\setminus k})}{P(\mathbf{x}^{(t)})P(x'_k|\mathbf{x}_{\setminus k}^{(t)})} \\ &= \frac{P(x'_k|\mathbf{x}'_{\setminus k})P(\mathbf{x}'_{\setminus k})P(x_k^{(t)}|\mathbf{x}'_{\setminus k})}{P(x_k^{(t)}|\mathbf{x}_{\setminus k}^{(t)})P(\mathbf{x}_{\setminus k}^{(t)})P(x'_k|\mathbf{x}_{\setminus k}^{(t)})} \stackrel{\mathbf{x}'_{\setminus k} = \mathbf{x}_{\setminus k}^{(t)}}{=} \frac{P(x'_k|\mathbf{x}'_{\setminus k})P(\mathbf{x}'_{\setminus k})P(x_k^{(t)}|\mathbf{x}'_{\setminus k})}{P(x_k^{(t)}|\mathbf{x}'_{\setminus k})P(\mathbf{x}'_{\setminus k})P(x'_k|\mathbf{x}'_{\setminus k})} = 1 \end{aligned}$$

- The samples are always accepted

Example of Gibbs Sampling

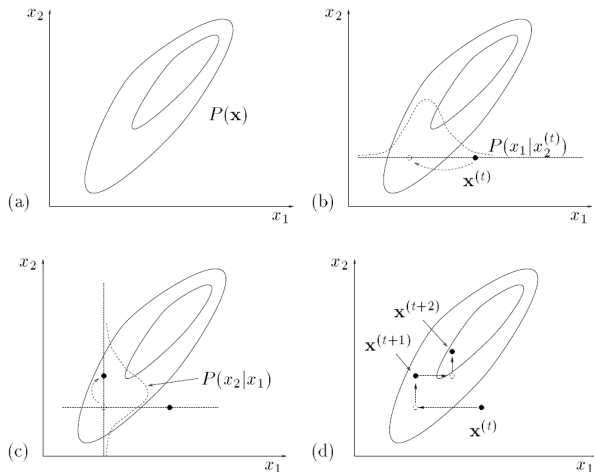


Figure 9. Gibbs sampling. (a) The joint density $P(\mathbf{x})$ from which samples are required. (b) Starting from a state $\mathbf{x}^{(t)}$, x_1 is sampled from the conditional density $P(x_1|x_2^{(t)})$. (c) A sample is then made from the conditional density $P(x_2|x_1)$. (d) A couple of iterations of Gibbs sampling.

Overview

- 1 Language Models: Recap
- 2 Topic Models
- 3 Probabilistic Latent Semantic Analysis (PLSA)
- 4 Latent Dirichlet Allocation (LDA)**
 - Motivation: Bayesian Modeling
 - **Background of Monte Carlo Methods**
 - Important Sampling
 - Rejection Sampling
 - Metropolis Methods
 - Gibbs Sampling
 - **Sampling for EM Algorithm**
 - Collapsed Gibbs Sampling for LDA

Mixture Models

$$\mathcal{J}(\Theta^t) = \sum_{m=1}^M \log \sum_{z_m} P(\mathbf{x}_m, z_m | \Theta^t)$$

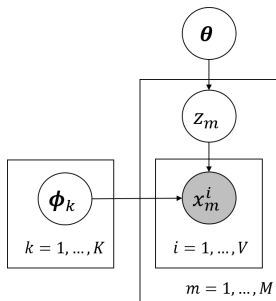


Figure: Mixture Models

EM Algorithm and Sampling

- Change Sum to Integral (to be general and better illustrate the idea)

$$\begin{aligned}\mathcal{J}(\Theta^t) &= \sum_{m=1}^M \log \int_{\mathbf{z}} P(\mathbf{x}_m, \mathbf{z} | \Theta^t) \\ &= \sum_{m=1}^M \log \int_{\mathbf{z}} q_{\mathbf{x}_m, \mathbf{z}}(\Theta) \frac{P(\mathbf{x}_m, \mathbf{z} | \Theta^t)}{q_{\mathbf{x}_m, \mathbf{z}}(\Theta)} \\ &\geq \sum_{m=1}^M \int_{\mathbf{z}} q_{\mathbf{x}_m, \mathbf{z}}(\Theta) \log \frac{P(\mathbf{x}_m, \mathbf{z} | \Theta^t)}{q_{\mathbf{x}_m, \mathbf{z}}(\Theta)} \\ &\doteq Q(\Theta, \Theta^t)\end{aligned}$$

where $\int_{\mathbf{z}} q_{\mathbf{x}_m, \mathbf{z}}(\Theta) = 1$ is some distribution

- In E-step, we solve $q_{\mathbf{x}_m, \mathbf{z}}(\Theta) = P(\mathbf{z} | \mathbf{x}_m, \Theta^t)$
- In M-step, we optimize $Q(\Theta^t, \Theta) = \sum_{m=1}^M \int_{\mathbf{z}} P(\mathbf{z} | \mathbf{x}_m, \Theta^t) \log P(\mathbf{x}_m, \mathbf{z} | \Theta) + \text{Const}$ w.r.t. Θ
- With sampling methods, we can approximate this M-step by a finite sum over samples \mathbf{z}^r from $P(\mathbf{z}^r | \mathbf{x}_m, \Theta^t)$

$$Q(\Theta^t, \Theta) \approx \sum_{m=1}^M \frac{1}{R} \sum_{\mathbf{z}^r \sim P(\mathbf{z}^r | \mathbf{x}_m, \Theta^t)} \log P(\mathbf{x}_m, \mathbf{z}^r | \Theta) + \text{Const}$$

- This procedure is called **Monte Carlo EM Algorithm**

EM Algorithm and Sampling: Variants

- Monte Carlo EM Algorithm

$$Q(\Theta^t, \Theta) \approx \sum_{m=1}^M \frac{1}{R} \sum_{\mathbf{z}^r \sim P(\mathbf{z}^r | \mathbf{x}_m, \Theta^t)} \log P(\mathbf{x}_m, \mathbf{z}^r | \Theta) + \text{Const}$$

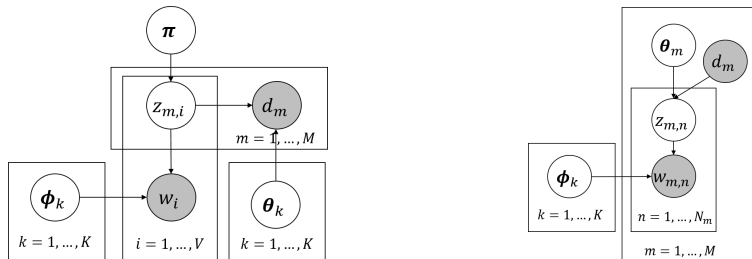
- When we consider a finite mixture model, and draw just one sample at each E-step
 - This is called **stochastic EM**
 - Here the latent variable \mathbf{z} characterizes which of the K components of the mixture is responsible for generating each data point
 - In the E-step, a sample of \mathbf{z} is taken from the posterior distribution $P(\mathbf{z} | \mathbf{X}, \Theta^t)$ where \mathbf{X} is the data set
 - This effectively makes a hard assignment of each data point to one of the components in the mixture
- If Gibbs sampling is used
 - Instead of drawing a sample from the corresponding conditional distribution, we make a point estimate of the variable given by the maximum of the conditional distribution
 - Then we obtain the **iterated conditional modes (ICM)** algorithm
 - For finite mixture models, it's similar to **K-means**

Overview

- 1 Language Models: Recap
- 2 Topic Models
- 3 Probabilistic Latent Semantic Analysis (PLSA)
- 4 Latent Dirichlet Allocation (LDA)**
 - Motivation: Bayesian Modeling
 - Background of Monte Carlo Methods
 - Important Sampling
 - Rejection Sampling
 - Metropolis Methods
 - Gibbs Sampling
 - Sampling for EM Algorithm
 - Collapsed Gibbs Sampling for LDA

Alternative Way for PLSA to Generate Texts

$$\begin{aligned}
 P(\mathcal{D}, \mathcal{W}) &= \prod_{m=1}^M \prod_{i=1}^{N_m} \sum_{k=1}^K P(z_{m,i} = k) P(d_m | \theta_k) P(w_i | \phi_k) \\
 &= \prod_{m=1}^M \prod_{i=1}^V \left(\sum_{k=1}^K P(z_{m,i} = k) P(d_m | \theta_k) P(w_i | \phi_k) \right)^{c_{d_m}(w_i)}
 \end{aligned}$$



$$P(\mathcal{D}, \mathcal{W}) = \prod_{m=1}^M \prod_{i=1}^V P(d_m) \left(\sum_{k=1}^K P(z_{m,i} = k | \theta_m) P(w_i | \phi_k) \right)^{c_{d_m}(w_i)}$$

Bayesian Modeling: Topic Models

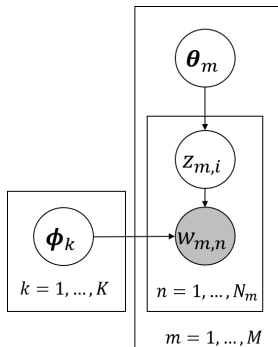


Figure: PLSA

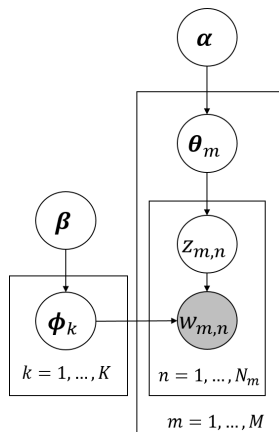


Figure: LDA

Generative Process of Latent Dirichlet Allocation

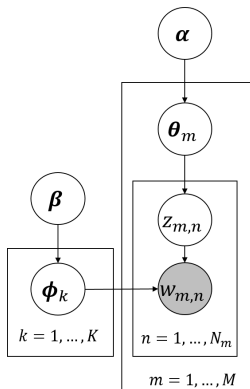


Figure: LDA

- For all clusters/components $k \in [1, K]$:
 - Choose mixture components $\phi_k \sim \text{Dir}(\phi|\beta)$
- For all documents $m \in [1, M]$:
 - Choose $N_m \sim \text{Poisson}(\xi)$
 - Choose mixture probability $\theta_m \sim \text{Dir}(\theta|\alpha)$
 - For all words $n \in [1, N_m]$ in document d_m :
 - Choose a component index
 $z_{m,n} \sim \text{Mult}(z|\theta_m)$
 - Choose a word $w_{m,n} \sim \text{Mult}(w|\phi_{z_{m,n}})$

A More Detailed Look of LDA

- The probability distribution of the k th latent topic that generates a word is a multinomial distribution

$$\begin{aligned} P(w|z = k, \phi_k) &\sim \text{Mult}(w|\phi_k) \\ &= \text{Mult}(w|\phi_{k,1}, \phi_{k,2}, \dots, \phi_{k,V}) = \prod_{i=1}^V \phi_{k,i}^{\delta_{w=v_i}} \end{aligned}$$

where

- $\phi_k = (\phi_{k,1}, \phi_{k,2}, \dots, \phi_{k,V})^T \in \mathbb{R}^V$
- $P(w = v_i|z = k) = P(v_i|z_k) = \phi_{k,i}$
- The delta function is $\delta_{w=v_i} = 1$ if $w = v_i$; and 0 otherwise
- We also denote the parameter for the topic mixture probabilities as $\Phi = (\phi_1, \phi_2, \dots, \phi_K)^T \in \mathbb{R}^{K \times V}$ where we have K topics

A More Detailed Look of LDA

- The probability distribution that a document generates a topic is:

$$P(z|\boldsymbol{\theta}_m) \sim \text{Mult}(z|\boldsymbol{\theta}_m) = \text{Mult}(w|\theta_{m,1}, \theta_{m,2}, \dots, \theta_{m,K}) = \prod_{k=1}^K \theta_{m,k}^{\delta_{z=k}}$$

where

- $\boldsymbol{\theta}_m = (\theta_{m,1}, \theta_{m,2}, \dots, \theta_{m,K})^T \in \mathbb{R}^K$
- $P(z = k|d = m) = P(z_k|d_m) = \theta_{m,k}$
- Here we omit the document id in $P(z|\boldsymbol{\theta}_m) = P(z|d_m, \boldsymbol{\theta}_m)$ since $\boldsymbol{\theta}_m$ has the document index m
- We also use $P(z_k|d_m)$ for short rather than the complete form $P(z = k|d = m, \boldsymbol{\theta}_m)$ sometimes
- The delta function is $\delta_{z=k} = 1$ if $z = k$; and 0 otherwise
- We also denote the parameter for the document mixture probabilities as $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \dots, \boldsymbol{\theta}_M)^T \in \mathbb{R}^{M \times K}$ where we have M documents

A More Detailed Look of LDA

- For a full Bayesian view of this mixture model, we add the conjugate Dirichlet priors to both multinomial distributions

$$P(\boldsymbol{\theta}|\boldsymbol{\alpha}) = \text{Dir}(\boldsymbol{\theta}|\boldsymbol{\alpha})$$

where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \dots, \alpha_K) \in \mathbb{R}^K$ and

$$P(\boldsymbol{\phi}|\boldsymbol{\beta}) = \text{Dir}(\boldsymbol{\phi}|\boldsymbol{\beta})$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_V) \in \mathbb{R}^V$

A More Detailed Look of LDA

- We formulate the conditional probability of a word $w_{m,n}$ in document d_m given θ_m and Φ as:

$$\begin{aligned} P(w_{m,n}|\theta_m, \Phi) &= \sum_{k=1}^K P(w_{m,n}|z_{m,n} = k, \Phi)P(z_{m,n} = k|\theta_m) \\ &= \sum_{k=1}^K P(w_{m,n}|\phi_k)P(z_{m,n} = k|\theta_m) \end{aligned}$$

- This means for each document, we generate a set of topics and each topic generate a word
- The probability of a word given a document and parameters is also a multinomial distribution

A More Detailed Look of LDA

- Now we can show the data likelihood given a document condition on hyper-parameters:

$$P(\mathcal{W}_m, \mathcal{Z}_m, \theta_m, \Phi | \alpha, \beta) = \underbrace{\prod_{n=1}^{N_m} P(w_{m,n} | \phi_k) P(z_{m,n} | \theta_m) P(\theta_m | \alpha)}_{\text{word plate}} \underbrace{P(\Phi | \beta)}_{\text{topic plate}}$$

document plate

where $\mathcal{Z}_m = \{z_{m,1}, z_{m,2}, \dots, z_{m,N_m}\}$ associated with word sequence \mathcal{W}_m .

A More Detailed Look of LDA

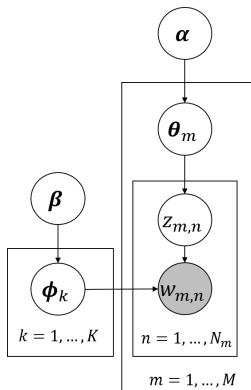


Figure: LDA

- Therefore, the complete likelihood for all documents are given by:

$$P(\mathcal{W}|\alpha, \beta) = \prod_{m=1}^M \int_{\boldsymbol{\Phi}} P(\boldsymbol{\Phi}|\beta) \int_{\boldsymbol{\theta}_m} P(\boldsymbol{\theta}_m|\alpha)$$

$$\left(\prod_{n=1}^{N_m} \sum_{k=1}^K P(w_{m,n}|\phi_k) P(z_{m,n} = k|\boldsymbol{\theta}_m) \right) d\boldsymbol{\theta}_m d\boldsymbol{\Phi}$$

- Inference a topic model given a set of training documents involves estimation of document-topic distribution θ 's and topic-word distribution ϕ 's
- MAP estimation is intractable due to the interaction between both parameters and also the hyper-parameters
- Thus, approximated methods can be used, such as MCMC (Griffiths and Steyvers (2004)) and variational techniques (Blei et al. (2003))
- Both methods finally produce the estimation of θ 's and ϕ 's

Collapsed Gibbs Sampling for LDA

- The collapsed sampling integrate out the parameters of θ 's and ϕ 's and only sample the latent topic variables by assigning topics to words
- The central idea of Gibbs sampling is to recover the joint marginal (integrating out the parameters) distribution given hyper-parameters:

$$\begin{aligned} P(\mathcal{Z}|\mathcal{W}, \alpha, \beta) &= \frac{P(\mathcal{W}, \mathcal{Z}|\alpha, \beta)}{P(\mathcal{W}|\alpha, \beta)} \\ &= \frac{\prod_{m=1}^M \prod_{n=1}^{N_m} P(w_{m,n}, z_{m,n}|\alpha, \beta)}{\prod_{m=1}^M \prod_{n=1}^{N_m} \sum_{k=1}^K P(w_{m,n}|\alpha, \beta)} \\ &= \frac{P(\mathcal{W}|\mathcal{Z}, \beta)P(\mathcal{Z}|\alpha)}{P(\mathcal{W}|\alpha, \beta)} \end{aligned}$$

- Gibbs sampling uses the procedure that samples one variable conditioned on all the other to approximate this distribution

$$P(z_{m,n}|\mathcal{Z}_{\setminus z_{m,n}}, \mathcal{W}, \alpha, \beta)$$

to sample a topic associated with a word. The notation $\mathcal{Z}_{\setminus z_{m,n}}$ means the topic assignment set without $z_{m,n}$

Dirichlet Distribution

- Recall the **Dirichlet distribution**:

$$P(\boldsymbol{\theta}|\boldsymbol{\alpha}) = \text{Dir}(\boldsymbol{\theta}|\boldsymbol{\alpha}) \triangleq \frac{\Gamma(\sum_{i=1}^V \alpha_i)}{\prod_{i=1}^V \Gamma(\alpha_i)} \prod_{i=1}^V \theta_i^{\alpha_i-1} \triangleq \frac{1}{\Delta(\boldsymbol{\alpha})} \prod_{i=1}^V \theta_i^{\alpha_i-1}$$

- The “Dirichlet Delta function” $\Delta(\boldsymbol{\alpha})$ is introduced for convenience
- $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \dots, \alpha_V)^\top \in \mathbb{R}^V$
- The Gamma function satisfies $\Gamma(x+1) = x\Gamma(x)$
 - For integer variable, Gamma function is $\Gamma(x) = (x-1)!$
 - For real numbers, it is $\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt$
- Note that $\int_{\boldsymbol{\theta}} d\boldsymbol{\theta} \prod_{i=1}^V \theta_i^{\alpha_i-1} = \Delta(\boldsymbol{\alpha})$ because
$$\int_{\boldsymbol{\theta}} d\boldsymbol{\theta} P(\boldsymbol{\theta}|\boldsymbol{\alpha}) = \int_{\boldsymbol{\theta}} d\boldsymbol{\theta} \frac{1}{\Delta(\boldsymbol{\alpha})} \prod_{i=1}^V \theta_i^{\alpha_i-1} = 1$$

- We introduce
 - u_{k,v_i} to represent the count for the word v_i being observed as topic k
- The multinomial distribution of words given topics is

$$\begin{aligned}
 P(\mathcal{W}|\mathcal{Z}, \Phi) &= \prod_{m=1}^M \prod_{n=1}^{N_m} P(w_{m,n}|z_{m,n}, \Phi) \\
 &= \prod_{m=1}^M \prod_{n=1}^{N_m} \phi_{z_{m,n}, w_{m,n}}^{u_{k,v_i}} \\
 &= \prod_{k=1}^K \prod_{i=1}^V \phi_{k,i}^{u_{k,v_i}}
 \end{aligned}$$

- By integrating out the parameters $\phi_{k,i}$, we can obtain the target distribution $P(\mathcal{W}|\mathcal{Z}, \beta)$

$$\begin{aligned}
 P(\mathcal{W}|\mathcal{Z}, \beta) &= \int_{\Phi} P(\mathcal{W}|\mathcal{Z}, \Phi) P(\Phi|\beta) d\Phi \\
 &= \int_{\Phi} \prod_{k=1}^K \frac{1}{\Delta(\beta)} \prod_{i=1}^V \phi_{k,i}^{\beta_i + u_{k,v_i} - 1} d\phi_k \\
 &= \prod_{k=1}^K \frac{\Delta(\mathbf{u}_k + \beta)}{\Delta(\beta)}
 \end{aligned}$$

where we denote $\mathbf{u}_k = (u_{k,v_1}, u_{k,v_2}, \dots, u_{k,v_V})^T \in \mathbb{R}^V$

- We introduce
 - $u_{d_m,k}$ represent the count for the topic k for a word being observed in document d_m
- Similarly, we can formulate the multinomial topic distributions given document parameters.

$$\begin{aligned} P(\mathcal{Z}|\Theta) &= \prod_{m=1}^M \prod_{n=1}^{N_m} P(z_{m,n}|d_m, \theta_m) = \prod_{m=1}^M \prod_{n=1}^{N_m} \theta_{m,z_{m,n}} \\ &= \prod_{m=1}^M \prod_{k=1}^K \theta_{m,k}^{u_{d_m,k}} \end{aligned}$$

- By integrating out the parameters $\theta_{m,k}$, we can obtain the other target distribution $P(\mathcal{Z}|\alpha)$

$$\begin{aligned} P(\mathcal{Z}|\alpha) &= \int_{\Theta} P(\mathcal{Z}|\Theta) P(\Theta|\alpha) d\Phi \\ &= \int_{\Theta} \prod_{m=1}^M \frac{1}{\Delta(\alpha)} \prod_{k=1}^K \theta_{m,k}^{\alpha_k + u_{d_m,k} - 1} d\phi_k \\ &= \prod_{m=1}^M \frac{\Delta(\mathbf{u}_{d_m} + \alpha)}{\Delta(\alpha)} \end{aligned}$$

where we denote $\mathbf{u}_{d_m} = (u_{d_m,1}, u_{d_m,2}, \dots, u_{d_m,K})^T \in \mathbb{R}^K$.

- Given

$$P(\mathcal{W}, \mathcal{Z} | \alpha, \beta) = P(\mathcal{W} | \mathcal{Z}, \beta) P(\mathcal{Z} | \alpha)$$

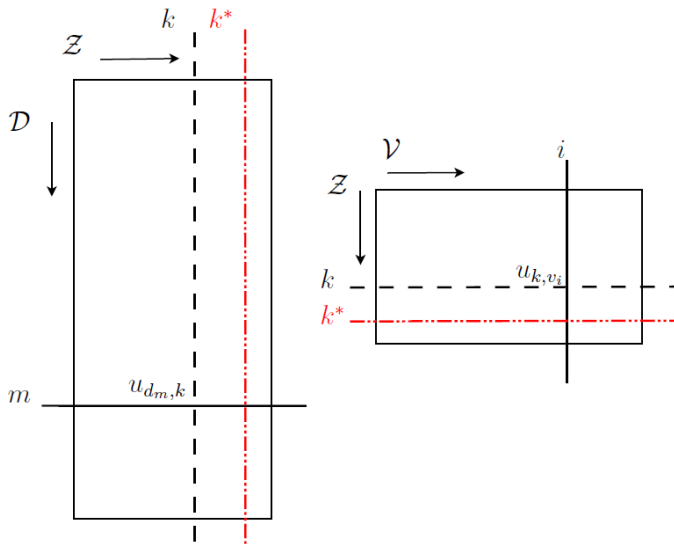
- The joint distribution is

$$P(\mathcal{W}, \mathcal{Z} | \alpha, \beta) = \prod_{k=1}^K \frac{\Delta(\mathbf{u}_k + \beta)}{\Delta(\beta)} \cdot \prod_{m=1}^M \frac{\Delta(\mathbf{u}_{d_m} + \alpha)}{\Delta(\alpha)}$$

Conditional Distribution

$$\begin{aligned}
 & P(z_{m,n} = k | \mathcal{Z}_{\setminus z_{m,n}}, \mathcal{W}, \alpha, \beta) \\
 = & P(z_{m,n} = k | w_{m,n} = v_i, \mathcal{Z}_{\setminus z_{m,n}}, \mathcal{W}_{\setminus w_{m,n}}, \alpha, \beta) = \frac{P(\mathcal{Z}, \mathcal{W} | \alpha, \beta)}{P(\mathcal{Z}_{\setminus z_{m,n}}, \mathcal{W} | \alpha, \beta)} \\
 \text{(Using the fact } w_{m,n} \perp \mathcal{W}_{\setminus w_{m,n}} | \mathcal{Z}_{\setminus z_{m,n}} \text{ and } P(w_{m,n} | \beta) = \sum_{i=1}^K P(w_{m,n}, z_{m,n} | \beta) \text{ is irrelevant to } z_{m,n}) \\
 = & \frac{P(\mathcal{W} | \mathcal{Z}, \beta)}{P(\mathcal{W}_{\setminus w_{m,n}} | \mathcal{Z}_{\setminus z_{m,n}}, \beta) P(w_{m,n} | \beta)} \cdot \frac{P(\mathcal{Z} | \alpha)}{P(\mathcal{Z}_{\setminus z_{m,n}} | \alpha)} \propto \frac{\Delta(u_k + \beta)}{\Delta(u_{k, \setminus z_{m,n}} + \beta)} \cdot \frac{\Delta(u_{d_m} + \alpha)}{\Delta(u_{d_m, \setminus z_{m,n}} + \alpha)} \\
 \text{(For } w_{m,n} = v_i \text{ and current corresponding topic is } z_{m,n} = k^*) \\
 \propto & \frac{\Gamma(u_{k, v_i} + \beta_i + (1 - \delta_{k=k^*}))}{\Gamma(\sum_{i=1}^V (u_{k, v_i} + \beta_i) + (1 - \delta_{k=k^*}))} \cdot \frac{\Gamma(\sum_{i=1}^V (u_{k, v_i} + \beta_i) - \delta_{k=k^*})}{\Gamma(u_{k, v_i} + \beta_i - \delta_{k=k^*})} \\
 & \frac{\Gamma(u_{d_m, k} + \alpha_k + (1 - \delta_{k=k^*}))}{\Gamma(\sum_{k=1}^K (u_{d_m, k} + \alpha_k))} \cdot \frac{\Gamma(\sum_{k=1}^K (u_{d_m, k} + \alpha_k) - 1)}{\Gamma(u_{d_m, k} + \alpha_k - \delta_{k=k^*})} \quad \left(\text{given } \frac{\Gamma(\sum_{i=1}^V \alpha_i)}{\prod_{i=1}^V \Gamma(\alpha_i)} = \frac{1}{\Delta(\alpha)} \right) \\
 \text{(Using } \Gamma(x+1) = x\Gamma(x)) \\
 \propto & \frac{u_{k, v_i} + \beta_i - \delta_{k=k^*}}{\sum_{i=1}^V (u_{k, v_i} + \beta_i) - \delta_{k=k^*}} \cdot \frac{u_{d_m, k} + \alpha_k - \delta_{k=k^*}}{\sum_{k=1}^K (u_{d_m, k} + \alpha_k) - 1} \\
 \text{(\sum_{k=1}^K (u_{d_m, k} + \alpha_k) - 1 \text{ is constant for all } k\text{'s})} \\
 \propto & \frac{u_{k, v_i} + \beta_i - \delta_{k=k^*}}{\sum_{i=1}^V (u_{k, v_i} + \beta_i) - \delta_{k=k^*}} \cdot (u_{d_m, k} + \alpha_k - \delta_{k=k^*})
 \end{aligned}$$

Matrix Illustration



Sampling Algorithm

Input: Document data set \mathcal{W}

repeat

for all documents $m = 1$ **to** M **do**

for all words $w_{m,n} = v_i$ where $n = 1$ **to** N_m **do**

 ◇ for the current assignment topic k^* to word $w_{m,n} = v_i$:

 decrement counts: $u_{d_m, k^*} - 1$ and $u_{k^*, v_i} - 1$

 ◇ multinomial sampling topic

$z_{m,n} = k^{new} \sim p(z_{m,n} | \mathcal{Z}_{\setminus z_{m,n}}, \mathcal{W}, \alpha, \beta)$ according to

$$\frac{u_{k, v_i} + \beta_i - \delta_{k=k^*}}{\sum_{i=1}^V (u_{k, v_i} + \beta_i) - \delta_{k=k^*}} \cdot (u_{d_m, k} + \alpha_k - \delta_{k=k^*})$$

 ◇ use the new assignment of $z_{m,n}$ to $w_{m,n} = v_i$:

 increment counts: $u_{d_m, k^{new}} + 1$ and $u_{k^{new}, v_i} + 1$

end for

end for

until Convergence

- Having the sampling counts, we can estimate the posterior of multinomial parameters Θ and Φ according to the state of the Markov Chain $\mathcal{M} = \{\mathcal{W}, \mathcal{Z}\}$ (MAP estimation)

$$\begin{aligned} & p(\theta_m | \mathcal{M}, \alpha) \\ &= \frac{1}{Z_{\theta_m}} \prod_{n=1}^{N_m} p(z_{m,n} | \theta_m) p(\theta_m | \alpha) \\ &= \text{Dir}(\theta_m | \mathbf{u}_{d_m} + \alpha) \end{aligned}$$

and

$$\begin{aligned} & p(\phi_k | \mathcal{M}, \beta) \\ &= \frac{1}{Z_{\phi_m}} \prod_{m=1}^M \prod_{n=1}^{N_m} p(w_{m,n} | z_{m,n} = k, \phi_k) p(\phi_k | \beta) \\ &= \text{Dir}(\phi_k | \mathbf{u}_k + \beta) \end{aligned}$$

Parameter Estimation (Cont'd)

- Based on the expectation formulation of Dirichlet distribution $\langle \text{Dir}(\boldsymbol{\alpha}) \rangle = (\alpha_i / \sum_i \alpha_i)_i$, we have:

$$\hat{\theta}_{m,k} = \frac{u_{d_m,k} + \alpha_k}{\sum_{k=1}^K (u_{d_m,k} + \alpha_k)}$$

and

$$\hat{\phi}_{k,i} = \frac{u_{k,vi} + \beta_i}{\sum_{i=1}^V (u_{k,vi} + \beta_i)}$$

Inference for New Coming Documents

- For a new coming document data set $\tilde{\mathcal{W}}$, we assume that the assigned topic set is $\tilde{\mathcal{Z}}$
- Each word $\tilde{w}_{m,n}$ will be assigned with a topic index $\tilde{z}_{m,n}$ also via Gibbs sampling procedure
- By fixing the training data and parameters Θ and Φ , we first randomly assign a topic to new coming word
- Then, perform sampling based on the following conditional probability:

$$\propto \frac{p(\tilde{z}_{m,n} = k | \tilde{w}_{m,n} = v_i, \tilde{\mathcal{Z}}_{\setminus \tilde{z}_{m,n}}, \tilde{\mathcal{W}}_{\setminus \tilde{w}_{m,n}}, \mathcal{M}, \alpha, \beta)}{\sum_{i=1}^V (u_{k,v_i} + \tilde{u}_{k,v_i} + \beta_i - \delta_{k=k^*})} \cdot \frac{\tilde{u}_{\tilde{d}_m,k} + \alpha_k - \delta_{k=k^*}}{\sum_{k=1}^K (\tilde{u}_{\tilde{d}_m,k} + \alpha_k) - 1}$$

Inference for New Coming Documents

- If the new coming documents size are small, u_{k,v_i} dominates the first term compared with \tilde{u}_{k,v_i} , which are randomly assigned
- Thus, repeatedly sampling from this distribution $p(\tilde{z}_{m,n} = k|\cdot)$ and updating $\tilde{u}_{\tilde{d}_m,k}$, topic-word associations are propagated into document-topic association
- For simplicity, we can even omit the topic-word term (Heinrich (2008)):

$$p(\tilde{z}_{m,n} = k | \tilde{w}_{m,n} = v_i, \tilde{\mathcal{Z}}_{\setminus \tilde{z}_{m,n}}, \tilde{\mathcal{W}}_{\setminus \tilde{w}_{m,n}}, \mathcal{M}, \alpha, \beta) \\ \propto \hat{\phi}_{k,i} \cdot \frac{\tilde{u}_{\tilde{d}_m,k} + \alpha_k - \delta_{k=k^*}}{\sum_{k=1}^K (\tilde{u}_{\tilde{d}_m,k} + \alpha_k) - 1}$$

Inference for New Coming Documents

- The topic distribution posterior for new coming documents are:

$$\hat{\theta}_{m,k} = \frac{\tilde{u}_{d_m,k} + \alpha_k}{\sum_{k=1}^K (\tilde{u}_{d_m,k} + \alpha_k)}$$

- In practice, we often assume the set of new coming data are much smaller than the training data: $|\tilde{\mathcal{W}}| \ll |\mathcal{W}|$.
- Otherwise, the new data will make the topic-word count distortion as $u_{k,v_i} + \tilde{u}_{k,v_i}$.

- We can also do hyperparameter estimation using maximum likelihood estimation
 - Refer to (Heinrich (2008))
 - Detailed Dirichlet distribution analysis can be found from (Minka (2000))
`https://tminka.github.io/papers/dirichlet/minka-dirichlet.pdf`

- There are many variants of LDA
 - Online and Incremental Learning
 - Distributed Computing
 - Dynamic Topic Model
 - Author Topic (AT) and Author Recipient Topic (ART) Model
 - Hierarchical Dirichlet Processes

Use of Topic Models

Topics

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
...

Documents

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 complementary views of the basic genes needed for **life**. One research team, using **computer** analyses 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 estimated 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 human genome, notes Siv Andersson, a geneticist at the University in Sweden. But coming up with a **concrete** answer may be more than just a **simple** numbers game, particularly if more and more **genomes** are **carefully** mapped and sequenced. "It may be a way of organizing any newly **sequenced genome**," explains Arcady Mushegian, a **computational** molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

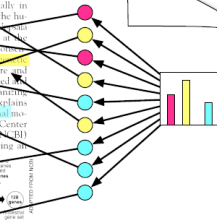


* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 6 to 12.

Stripping down. **Computer analysis** yields an estimate of the minimum modern and ancient genomes.

SCIENCE • VOL. 272 • 24 MAY 1996

Topic proportions and assignments

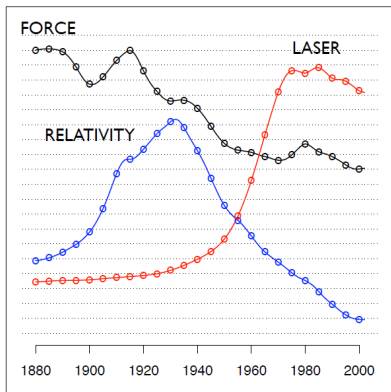


Discover topics from a corpus

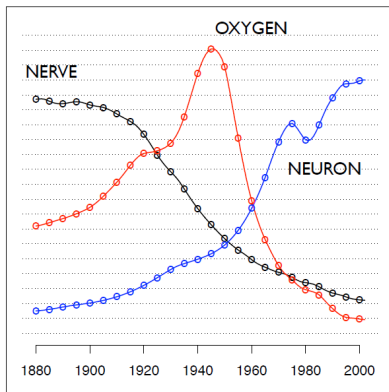
human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

Model the evolution of topics over time

"Theoretical Physics"



"Neuroscience"



Annotate images



SKY WATER TREE
MOUNTAIN PEOPLE



SCOTLAND WATER
FLOWER HILLS TREE



SKY WATER BUILDING
PEOPLE WATER



FISH WATER OCEAN
TREE CORAL

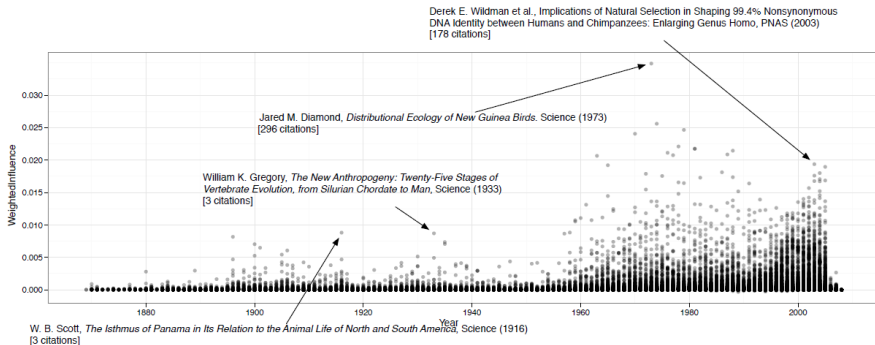


PEOPLE MARKET PATTERN
TEXTILE DISPLAY

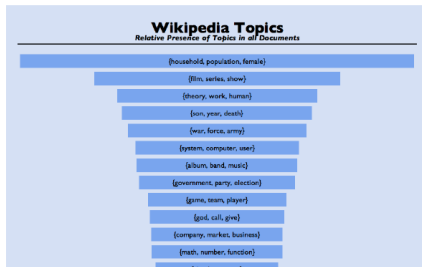


BIRDS NEST TREE
BRANCH LEAVES

Discover influential articles



Organize and browse large corpora



{film, series, show}

words	related documents	related topics
film	The X-Files	{son, year, death}
series	Orson Welles	{work, book, publish}
show	Stanley Kubrick	{album, band, music}
character	B movie	{woman, child, man}
play	Mystery Science Theater 3000	{law, state, case}
make	Monty Python	{black, white, people}
episode	Doctor Who	{theory, work, human}
movie	Sam Peckinpah	{@card@, make, design}
good	Married... with Children	{war, force, army}
release	History of film	{god, call, give}
feature	The A-Team	{game, team, player}
television	Pulp Fiction (film)	{day, year, event}
star	Mad (magazine)	{company, market, business}

Stanley Kubrick

related topics

- {film, series, show}
- {theory, work, human}
- {son, year, death}
- {black, white, people}
- {god, call, give}
- {math, energy, light}

Stanley Kubrick (July 26, 1928 – March 7, 1999) was an American film director, writer, producer, and photographer who lived in England during most of the last four decades of his career. Kubrick was noted for the scrupulous care with which he chose his subjects, his slow method of working, the variety of genres he worked in, his technical perfectionism, and his reluctance about his films and personal life. He worked far beyond the confines of the Hollywood system, maintaining almost complete artistic control and making movies according to his own whims and time constraints, but with the rare advantage of big-studio financial support for all his endeavors.

Kubrick's films are characterized by a formal visual style and meticulous attention to detail—his later films often have elements of surrealism and expressionism that eschews structured linear narrative. His films are repeatedly described as slow and methodical, and are often perceived as a reflection of his obsessive and perfectionist nature.^[1] A recurring theme in his films is man's inhumanity to man. While often viewed as

related documents

- Orson Welles
- B movie
- Mystery Science Theater 3000
- Monty Python
- Doctor Who
- Sam Peckinpah
- The A-Team
- Pulp Fiction (film)
- Buffy the Vampire Slayer (TV series)
- The X-Files
- Sunset Boulevard (film)
- Jack Benny

{theory, work, human}

words	related documents	related topics
theory	Meme	{work, book, publish}
work	Intelligent design	{law, state, case}
human	Immanuel Kant	{son, year, death}
idea	Philosophy of mathematics	{woman, child, man}
term	History of science	{god, call, give}
study	Free will	{black, white, people}
view	Truth	{film, series, show}
science	Psychoanalysis	{war, force, army}
concept	Charles Peirce	{language, word, form}
form	Existentialism	{@card@, make, design}
world	Deconstruction	{church, century, christian}
argue	Social sciences	{rate, high, increase}
social	Idealism	{company, market, business}

References I

- Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.
- Griffiths, T. L. and Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101:5228–5235.
- Heinrich, G. (2008). Parameter estimation for text analysis. Technical Report Version 2.4, vsonix GmbH + University of Leipzig, Germany.
- Minka, T. P. (2000). Estimating a dirichlet distribution. Technical report.