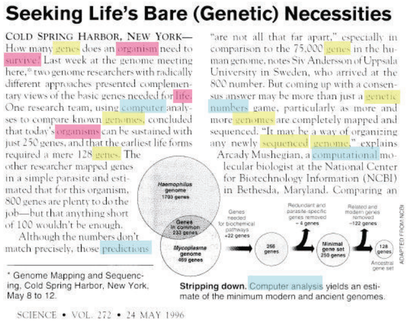
Latent Dirichlet Allocation

David Blei, Andrew Ng, Michael Jordan

27 April, 2010

presented by Zhaoyin Jia, Ainur Yessenalina

Intuition behind LDA



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 of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely 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

Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 469 genes
Genes needed for biochemical pathways +22 genes
256 genes
Redundant and parasite-specific genes removed −4 genes
Minimal gene set 250 genes
Related and modern genes removed −122 genes
128 genes
Ancestral gene set
ADAPTED FROM NCBI

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

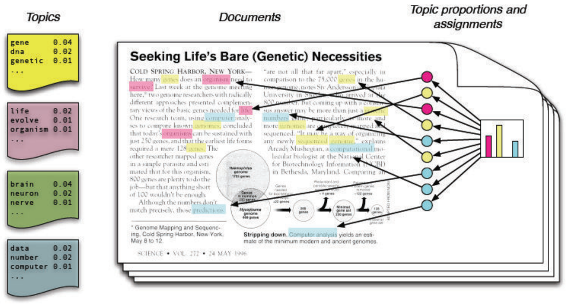
* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996

Simple intuition: Documents exhibit multiple topics.

(from David Blei)

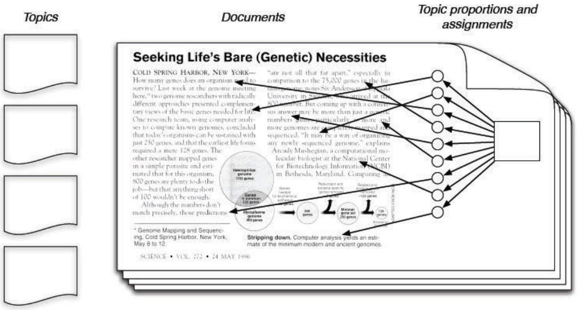
Probabilistic model



(from David Blei)

- Each document is a random mixture of corpus-wide topics
- Each word is drawn from one of those topics

Probabilistic model (2)


Topics
Documents
Topic proportions and assignments
Seeking Life's Bare (Genetic) Necessities

(from David Blei)

- We only observe the documents
- Our goal is to **infer** the underlying topic structure

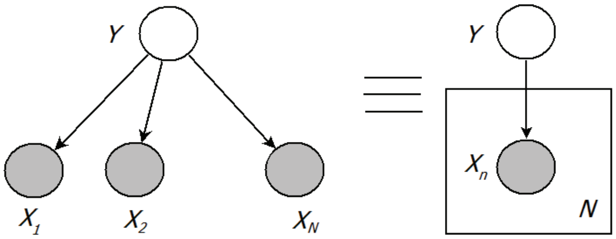
Probabilistic model (2)

- The observations are generated from a generative probabilistic process that includes hidden variables
- Infer the hidden structure using posterior inference. What are the topics that describe this collection?
- Situate new data into the estimated model.
  - How does this query or new document fit into the estimated topic structure?

Notation

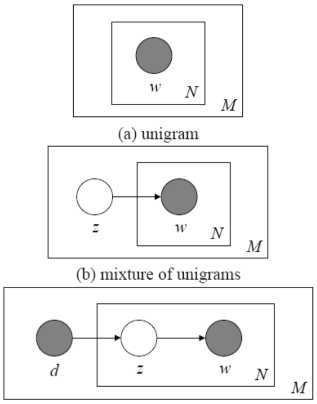
1. word: $1..V$
2. document: $\mathbf{w} = (w_1, w_2, ...w_N)$ sequence of N words
3. corpus: $D=\{\mathbf{w}_1, ..., \mathbf{w}_M\}$ collection of M documents

Graphical models notation



- Nodes are random variables
- Edges denote possible dependence
- Observed variables are shaded
- Plates denote replicated structure

Other models of the discrete data.

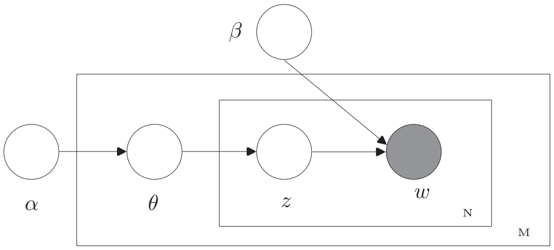


(a) unigram

(b) mixture of unigrams

(c) pLSI/aspect model

Latent Dirichlet allocation

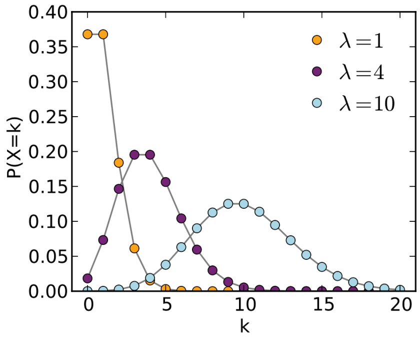


Latent Dirichlet allocation

LDA assumes the following generative process:

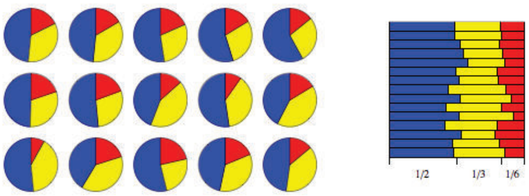
1. Choose $N \sim \text{Poisson}(\xi)$
2. Choose $\theta \sim \text{Dir}(\alpha)$
3. For each of N words w_n:
   (a) Choose topic $z_n \sim \text{Multinomial}(\theta)$
   (b) Choose word $w_n \sim$ from $P(w_n | z_n, \beta)$

Recap on distributions: Poisson



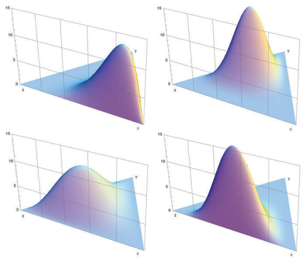
(from Wikipedia)

Recap on distributions: Dirichlet example



$Dir(\alpha)$; $\alpha = (3, 2, 1)$
Cut strings (each of initial length 1.0) into K pieces with different lengths
(from Wikipedia)

Recap on distributions: Dirichlet example (2)



Dirichlet distribution, K=3 for various parameter vectors α
Clockwise from top left:
$\alpha = (6,2,2), (3,7,5), (6,2,6), (2,3,4)$.
(from Wikipedia)

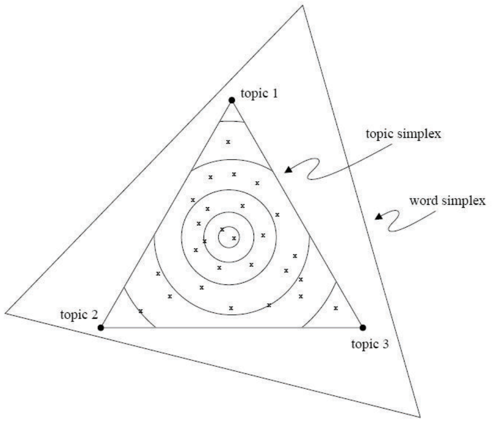
The Dirichlet distribution

- The Dirichlet distribution is an exponential family distribution over the simplex, i.e., positive vectors that sum to one

$$p(\theta \mid \vec{\alpha}) = \frac{\Gamma\left(\sum_i \alpha_i\right)}{\prod_i \Gamma(\alpha_i)} \prod_i \theta_i^{\alpha_i - 1}.$$

- The Dirichlet is **conjugate** to the multinomial. Given a multinomial observation, the posterior distribution of θ is a Dirichlet.
- The parameter α controls the mean shape and sparsity of θ.
- The topic proportions are a K dimensional Dirichlet.
  The topics are a V dimensional Dirichlet.

Geometric intuition

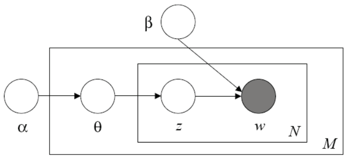


The Dirichlet distribution

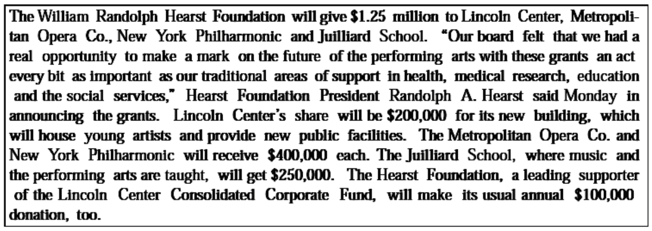
From a collection of documents, **infer**

- Per-word topic assignment $z_{d,n}$
- Per-document topic proportions θ_d
- Per-corpus topic distributions β_k

Inference

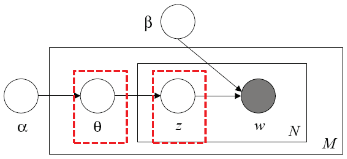


“Arts” “Budgets” “Children” “Education”



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.

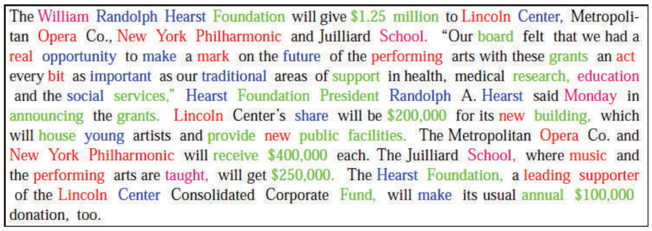
Inference



θ: Per-document topic proportions

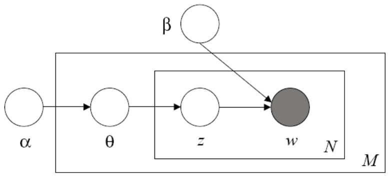
z: Per-word topic assignment

“Arts” “Budgets” “Children” “Education”



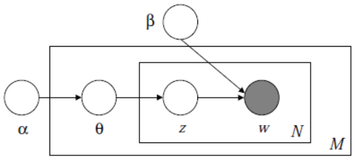
The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. “Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our 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.

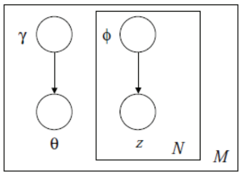
Inference



- Given corpus (w is observed), parameters (α, β), calculate p(θ,z| α, β, w)
- Intractable
  $$\frac{p(\theta \mid \alpha)\prod_{n=1}^{N} p(z_n \mid \theta)p(w_n \mid z_n, \beta_{1:K})}{\int_\theta p(\theta \mid \alpha)\prod_{n=1}^{N}\sum_{z=1}^{K} p(z_n \mid \theta)p(w_n \mid z_n, \beta_{1:K})}$$
  – Gibbs sampling
  – Variational inference

Variational Inference





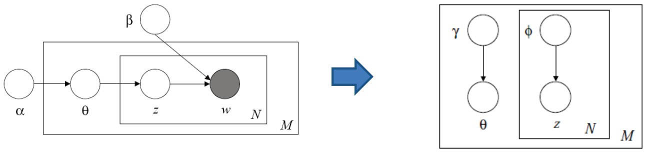
Choose ϒ, ϕ to approximate posterior distribution of θ,z

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

$$\phi_{ni} \propto \beta_{iv} \exp\left(\Psi(\gamma_i) - \Psi\left(\textstyle\sum_{j=1}^{k} \gamma_j\right)\right).$$

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

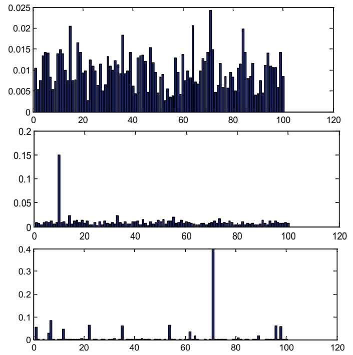
Variational Inference



```
initialize $\phi_{ni}^0 := 1/k$ for all $i$ and $n$
initialize $\gamma_i := \alpha_i + N/k$ for all $i$
repeat
    for $n = 1$ to $N$
        for $i = 1$ to $k$
            $\phi_{ni}^{t+1} := \beta_{iw_n} \exp(\Psi(\gamma_i^t))$
        normalize $\phi_n^{t+1}$ to sum to 1.
    $\gamma^{t+1} := \alpha + \sum_{n=1}^{N} \phi_n^{t+1}$
until convergence
```

Parameter estimation

- α controls proportion distribution of topics in one document.



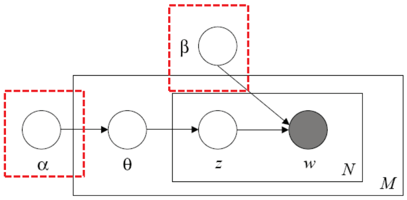
$\alpha=(\alpha_1, \alpha_2,\ldots, \alpha_K)$ equally large — Topics are almost equally likely

$\alpha=(\alpha_1, \alpha_2,\ldots, \alpha_K)$ equal, but α_{10} is larger — 10th topic is more likely to appear

$\alpha=(\alpha_1, \alpha_2,\ldots, \alpha_K)$ are equally small — Topics distribution is sparse (few topics in one document), with one random peak

- β is the probability matrix of topics and words

Parameter Estimation



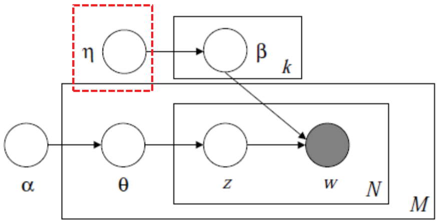
α : Dirichlet parameter

β : Topics

- Try to estimate parameters (α, β), given corpus {w}.
- EM algorithm:
  – E step: find the optimizing value of ϒ, ϕ
  – M step: maximize log likelihood w.r.t α and β.

Smoothing for unseen words

- For unseen word, MLE of β will assign zero probability during inference.
- Take β as Dirichlet distribution parameterized by η.



Parameter Estimation Example

- 16,000 documents of TREC AP corpus
- 100-topic LDA model

Top words of p(w|z)

"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

Inference example

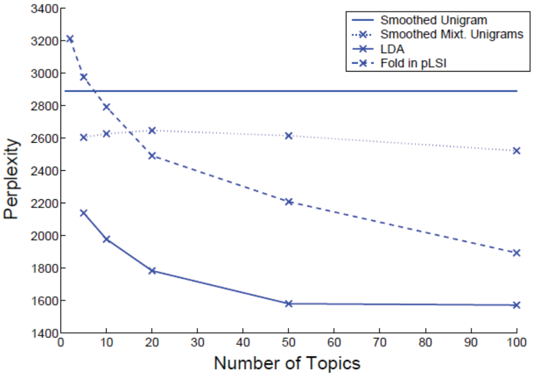
"Arts" "Budgets" "Children" "Education"

q(z|w)>0.9

Bag-of-words assumption

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.

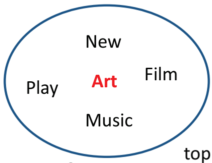
Application/Empirical Results



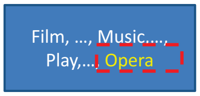
$$perplexity(D_{test}) = \exp\left\{-\frac{\sum_{d=1}^{M}\log p(\mathbf{w}_d)}{\sum_{d=1}^{M} N_d}\right\}.$$

Overfitting discussion

- Mixture of unigrams model:



topic



Never seen in art topic,
p(d) decreases a lot,
Perplexity explodes

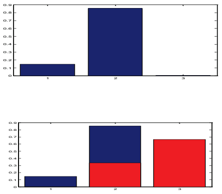
document

- pLSI:
  - Heuristic Inference:

$$p(\mathbf{w}) = \sum_{d}\prod_{n=1}^{N}\sum_{z} p(w_n\,|\,z)\,p(z\,|\,d)\,p(d).$$

  - Fold-in pLSI: refit p(z|d)

Topic distribution



Train document

Test document

Document classification

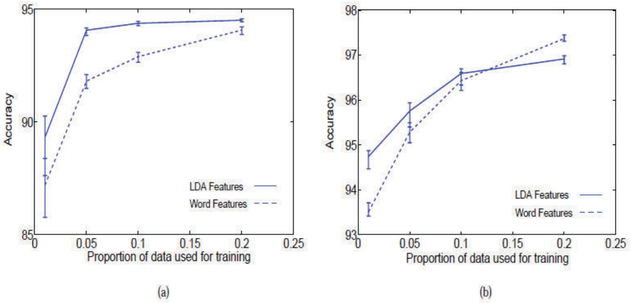
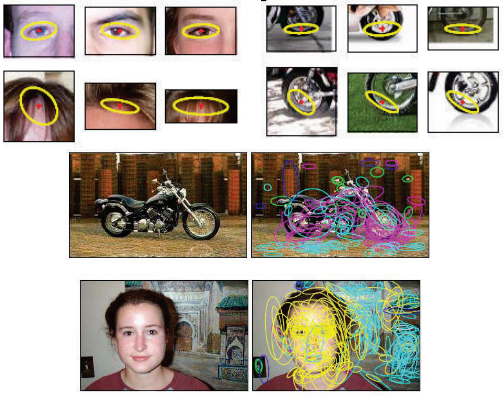


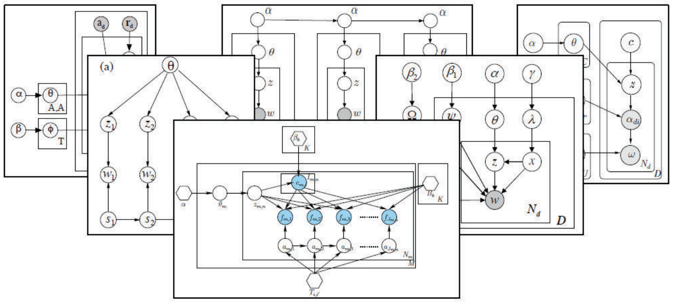
Figure 10: Classification results on two binary classification problems from the Reuters-21578 dataset for different proportions of training data. Graph (a) is EARN vs. NOT EARN. Graph (b) is GRAIN vs. NOT GRAIN.

Application in Vision



Discovering object categories in image collections. J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, W. T. Freeman. MIT AI Lab Memo AIM-2005-005, February, 2005.

LDA is modular, general, useful



LDA can be embedded in a more complicated model, embodying further intuition of structure of text

Slide from David Blei's lecture at Machine Learning Summer School 2009 - Cambridge

Summary

- Better graphic model
  - Compared to unigram, mixture of unigram, PLSI
- Approximate inference/Parameter estimation
- Applications:
  - generalizing documents/Images
  - Feature reduction
  - Other extensions

Thanks

Questions?