Case Study 5: Mixed Membership Modeling

Latent Dirichlet Allocation Collapsed Gibbs Sampler, Variational Methods

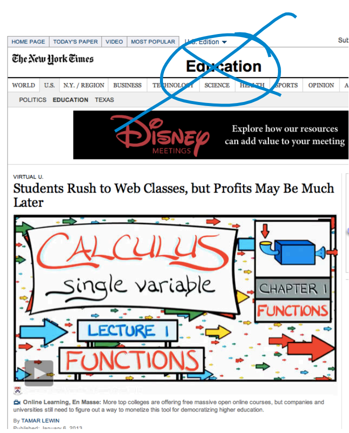
Machine Learning for Big Data
CSE547/STAT548, University of Washington

Emily Fox

March 11th, 2014

Task 3: Mixed Membership Models

- **Now:** Document may belong to multiple clusters



EDUCATION

FINANCE

TECHNOLOGY

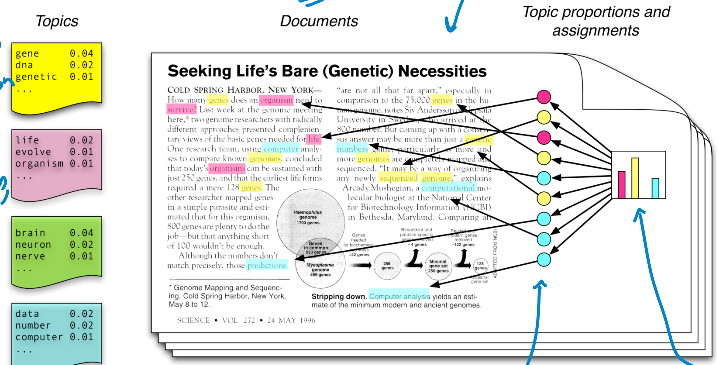
mixture of topics

Latent Dirichlet Allocation (LDA)

each topic k is a distribution over words in vocab β_k, just as before

previously, each doc had one topic

now, each is a mixture of topics



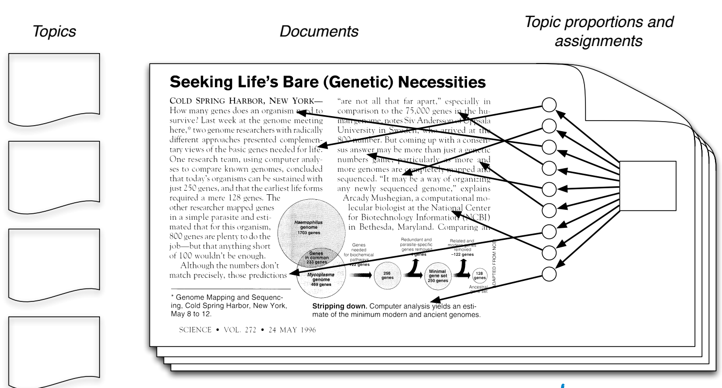
β_k

1 2 ... V

every word is assigned to a topic

each doc has its own prevalence of topics in that doc

Latent Dirichlet Allocation (LDA)



All we see are words

Want: posterior p(topics (β's), doc prop. of topics, assign. vars. | words in docs)

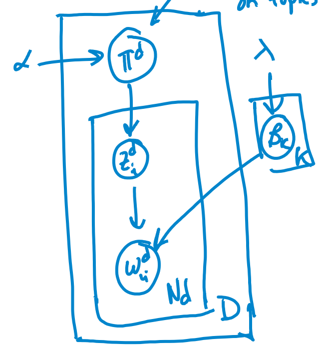
LDA Generative Model

- Observations: $w_1^d, \dots, w_{N_d}^d$
- Associated topics: $z_1^d, \dots, z_{N_d}^d$ ← topic per word in doc d
- Parameters: $\theta = \{\{\pi^d\}, \{\beta_k\}\}$
- Generative model:

$z_i^d \sim \pi^d$ $\quad d=1,\dots,D$, $\quad i=1,\dots,N_d$

$w_i^d \mid z_i^d \sim \beta_{z_i^d}$

doc-specific weighting on topics

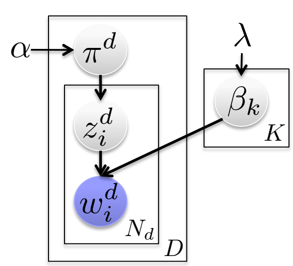


Priors:

$\pi^d \sim \mathrm{Dir}(\alpha_1, \dots, \alpha_K)$ $\quad d=1,\dots,D$

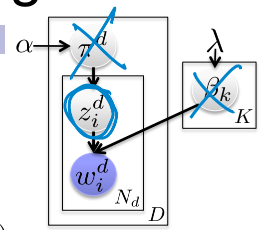
$\beta_k \sim \mathrm{Dir}(\lambda_1, \dots, \lambda_V)$ $\quad k=1,\dots,K$

LDA Joint Probability



$$p(\cdot) = \prod_{k=1}^{K} p(\beta_k \mid \lambda) \prod_{d=1}^{D} p(\pi^d \mid \alpha) \left(\prod_{i=1}^{N_d} p(z_i^d \mid \pi^d) p(w_i^d \mid z_i^d, \beta) \right)$$

Collapsed LDA Sampling

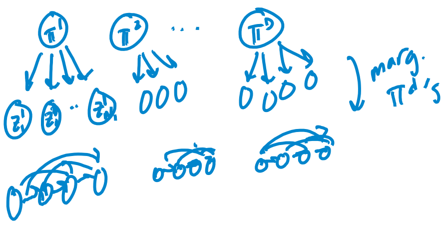


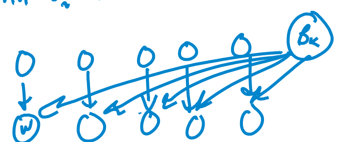
- Marginalize parameters
  - Document-specific topic weights
  - Corpus-wide topic-specific word distributions

$$p(z_i^d = k \mid z_{\setminus id}, \{w_i^d\}, \alpha, \lambda)$$
$$\propto p(z_i^d = k \mid z_{\setminus id}, \alpha) p(w_i^d | z_i^d = k, z_{\setminus id}, w_{\setminus id}, \lambda)$$

- Unplate to see dependencies induced

All $z_i^d = k$: (in all of corpus)

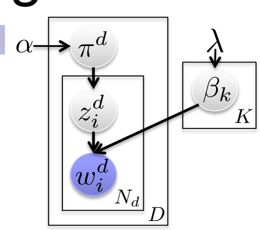




⇓ marg. β_k

introducing coupling in all of above

Collapsed LDA Sampling



- Sample topic indicators for each word
  - Algorithm:

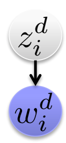
$$p(z_i^d = k \mid z_{\setminus id}, \{w_i^d\}, \alpha, \lambda)$$

$$\propto p(z_i^d = k \mid \{z_j^d, j \neq i\}, \alpha) p(w_i^d | \{w_j^c : z_j^c = d, (j,c) \neq (i,d)\}, \lambda)$$

Select a Document

Etruscan	trade	price	temple	market

Randomly Assign Topics

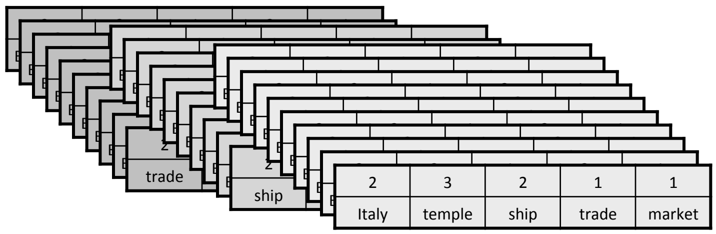


3	2	1	3	1
Etruscan	trade	price	temple	market

Randomly Assign Topics

$z_i^d \rightarrow w_i^d$

3	2	1	3	1
Etruscan	trade	price	temple	market



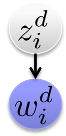
Maintain Local Statistics

$z_i^d \rightarrow w_i^d$

3	2	1	3	1
Etruscan	trade	price	temple	market

	Topic 1	Topic 2	Topic 3
Doc d			

Maintain Global Statistics



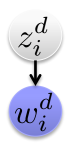
3	2	1	3	1
Etruscan	trade	price	temple	market

	Topic 1	Topic 2	Topic 3
Doc d	2	1	2

	Topic 1	Topic 2	Topic 3
Etruscan	1	0	35
market	50	0	1
price	42	1	0
temple	0	0	20
trade	10	8	1
…			

Total counts from **all** docs

Resample Assignments



3	2	1	3	1
Etruscan	trade	price	temple	market

	Topic 1	Topic 2	Topic 3
Doc d	2	1	2

	Topic 1	Topic 2	Topic 3
Etruscan	1	0	35
market	50	0	1
price	42	1	0
temple	0	0	20
trade	10	8	1
…			

What is the conditional distribution for this topic?

$z_i^d \rightarrow w_i^d$

3	?	1	3	1
Etruscan	trade	price	temple	market

What is the conditional distribution for this topic?

- Part I: How much does this document like each topic?

$z_i^d \rightarrow w_i^d$

3	?	1	3	1
Etruscan	trade	price	temple	market

Topic 1 Topic 2 Topic 3

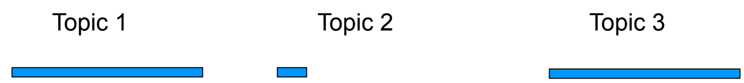
	Topic 1	Topic 2	Topic 3
Doc d	2	0	2

What is the conditional distribution for this topic?

- Part I: How much does this document like each topic?
- Part II: How much does each topic like this word?

$z_i^d \rightarrow w_i^d$

3	?	1	3	1
Etruscan	trade	price	temple	market



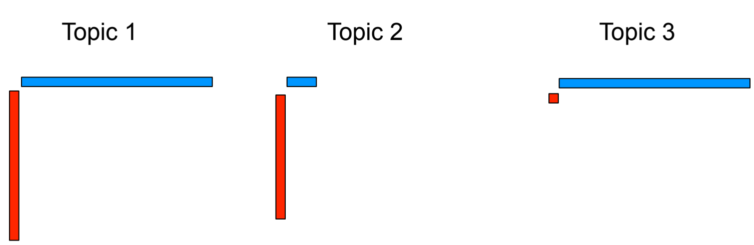
	Topic 1	Topic 2	Topic 3
trade	10	7	1

What is the conditional distribution for this topic?

- Part I: How much does this document like each topic?
- Part II: How much does each topic like this word?

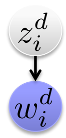
$z_i^d \rightarrow w_i^d$

3	?	1	3	1
Etruscan	trade	price	temple	market

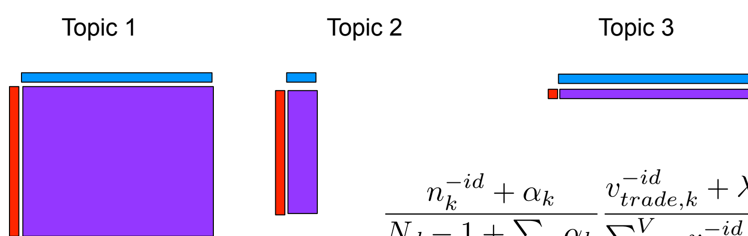


What is the conditional distribution for this topic?

- Part I: How much does this document like each topic?
- Part II: How much does each topic like this word?

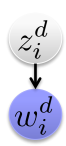


3	?	1	3	1
Etruscan	trade	price	temple	market

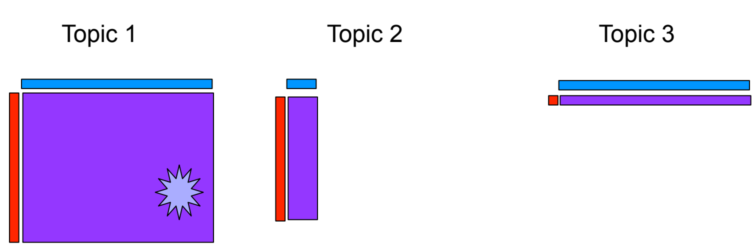


$$\frac{n_k^{-id}+\alpha_k}{N_d - 1 + \sum_k \alpha_k}\;\frac{v_{trade,k}^{-id}+\lambda_{trade}}{\sum_{\gamma=1}^{V} v_{\gamma,k}^{-id}+\lambda_\gamma}$$

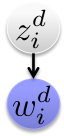
Sample a New Topic Indicator



3	?	1	3	1
Etruscan	trade	price	temple	market



Update Counts

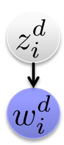


3	?	1	3	1
Etruscan	trade	price	temple	market

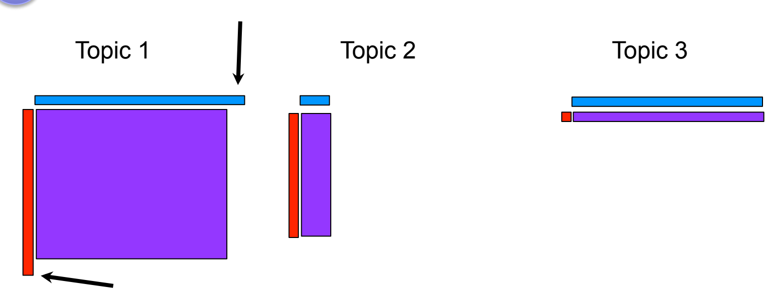
	Topic 1	Topic 2	Topic 3
Doc d	2	0	2

	Topic 1	Topic 2	Topic 3
Etruscan	1	0	35
market	50	0	1
price	42	1	0
temple	0	0	20
trade	10	7	1
…			

Geometrically…



3	**1**	1	3	1
Etruscan	trade	price	temple	market



Issues with Generic LDA Sampling

- Slow mixing rates → Need many iterations
- Each iteration cycles through sampling topic assignments for *all* words in *all* documents
- Modern approaches include:
  - Large-scale LDA. For example, Mimno, David, Matthew D. Hoffman and David M. Blei. "Sparse stochastic inference for latent Dirichlet allocation." International Conference on Machine Learning, 2012.
  - Distributed LDA. For example, Ahmed, Amr, et al. "Scalable inference in latent variable models." Proceedings of the fifth ACM international conference on Web search and data mining (2012): 123-132
  - And many, many more!
- Alternative: Variational methods instead of sampling
  - Approximate posterior with an optimized variational distribution

Case Study 5: Mixed Membership Modeling

Variational Methods

Machine Learning for Big Data
CSE547/STAT548, University of Washington
Emily Fox
March 11th, 2014

Variational Methods Goal

- Recall task: Characterize the posterior
- Turn posterior inference into an optimization task
- Introduce a "tractable" family of distributions over parameters and latent variables
  - Family is indexed by a set of "free parameters"
  - Find member of the family closest to:

Variational Methods Cartoon

- Cartoon of goal:
- Questions:
  - How do we measure "closeness"?
  - If the posterior is intractable, how can we approximate something we do not have to begin with?

A Measure of Closeness

- Kullback-Leibler (KL) divergence
  - Measures "distance" between two distributions *p* and *q*
- If *p* = *q* for all *θ*
- Otherwise,

A Measure of Closeness

$$\mathrm{KL}(p||q) \triangleq D(p||q) = \int_\theta p(\theta) \log \frac{p(\theta)}{q(\theta)} d\theta$$

- Not symmetric
- *p* determines where the difference is important:
  - $p(\theta)=0$ and $q(\theta)\neq 0$
  - $p(\theta)\neq 0$ and $q(\theta)=0$
- Want
- Just as hard as the original problem!

Reverse Divergence

- Divergence D(*p* || *q*)
  - true distribution *p* defines support of diff.
  - the “correct” direction
  - will be intractable to compute
- Reverse divergence D(*q* || *p*)
  - approximate distribution defines support
  - tends to give overconfident results
  - will be tractable

Interpretations of Minimizing Reverse KL

$$D(q||p) = E_q\left[\log \frac{q}{p}\right]$$

- Similarity measure:

- Evidence lower bound (ELBO)

Interpretations of Minimizing Reverse KL

- Evidence lower bound (ELBO)

$$\log p(x) = D(q(z,\theta)||p(z,\theta|x)) + \mathcal{L}(q) \geq \mathcal{L}(q)$$

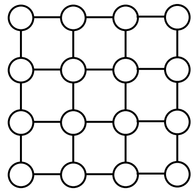
- Therefore,
  - ELBO provides a lower bound on marginal likelihood
  - Maximizing ELBO is equivalent to minimizing KL

Mean Field

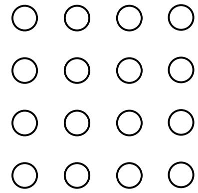
$$\mathcal{L}(q) = E_q[\log p(z,\theta,x)] - E_q[\log q(z,\theta)]$$

- How do we choose a Q such that the following is tractable?

- Simplest case = mean field approximation
  - Assume each parameter and latent variable is conditionally independent given the set of free parameters



Original graph

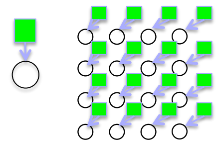


Naïve mean field

Mean Field $\mathcal{L}(q) = E_q[\log p(z, \theta, x)] - E_q[\log q(z, \theta)]$

- Naïve mean field decomposition:

$$q(z, \theta) = q(\theta \mid \gamma) \prod_{i=1}^{N} q(z^i \mid \phi^i)$$



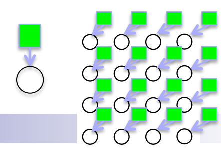
- Under this approximation, entropy term decomposes as

- Can (always) rewrite joint term as

$$E_q[\log p(\theta, z, x)] = E_q[\log p(\theta \mid z, x)] + E_q[\log p(z, x)]$$

$$E_q[\log p(\theta, z, x)] = E_q[\log p(z^i \mid z_{\setminus i}, \theta, x)] + E_q[\log p(z_{\setminus i}, \theta, x)]$$

Mean Field – Optimize γ



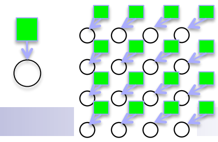
- Examine one free parameter, *e.g.*, γ

$$\mathcal{L}(q) = E_q[\log p(\theta \mid z, x)] + E_q[\log p(z, x)] - E_q[\log q(\theta \mid \gamma)] - \sum_i E_q[\log q(z^i \mid \phi^i)]$$

  - Look at terms of ELBO just depending on γ

$$\mathcal{L}^{\gamma} =$$

Mean Field – Optimize ϕ^i



- Examine another free parameter, *e.g.*, ϕ^i

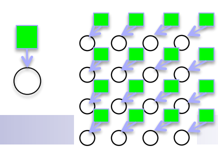
$$\mathcal{L}(q) = E_q[\log p(z^i \mid z_{\setminus i}, \theta, x)] + E_q[\log p(z_{\setminus i}, \theta, x) - E_q[\log q(\theta \mid \gamma)] - \sum_i E_q[\log q(z^i \mid \phi^i)]$$

  - Look at terms of ELBO just depending on ϕ^i

$$\mathcal{L}^{\phi^i} =$$

- This motivates using a coordinate ascent algorithm for optimization
  - Iteratively optimize each free parameter holding all others fixed

Algorithm Outline



- **Initialization:** Randomly select starting distribution $q_\theta^{(0)}$
- **E-Step:** Given parameters, find posterior of hidden data

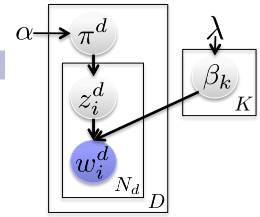
$$q_z^{(t)} = \arg\max_{q_z} \mathcal{L}(q_z, q_\theta^{(t-1)})$$

- **M-Step:** Given posterior distributions, find likely parameters

$$q_\theta^{(t)} = \arg\max_{q_\theta} \mathcal{L}(q_z^{(t)}, q_\theta)$$

- **Iteration:** Alternate E-step & M-step until convergence

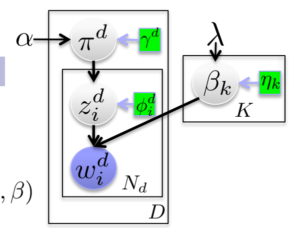
Mean Field for LDA



- In LDA, our parameters are $\theta = \{\pi^d\}, \{\beta_k\}$
  $z = \{z_i^d\}$
- The variational distribution factorizes as
- The joint distribution factorizes as

$$p(\pi, \beta, z, w) = \prod_{k=1}^{K} p(\beta_k \mid \lambda) \prod_{d=1}^{D} p(\pi^d \mid \alpha) \prod_{i=1}^{N_d} p(z_i^d \mid \pi^d) p(w_i^d \mid z_i^d, \beta)$$

Mean Field for LDA



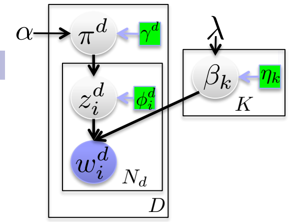
$$q(\pi, \beta, z) = \prod_{k=1}^{K} q(\beta_k \mid \eta_k) \prod_{d=1}^{D} q(\pi^d \mid \gamma^d) \prod_{i=1}^{N_d} q(z_i^d \mid \phi_i^d)$$

$$p(\pi, \beta, z, w) = \prod_{k=1}^{K} p(\beta_k \mid \lambda) \prod_{d=1}^{D} p(\pi^d \mid \alpha) \prod_{i=1}^{N_d} p(z_i^d \mid \pi^d) p(w_i^d \mid z_i^d, \beta)$$

- Examine the ELBO

$$\mathcal{L}(q) = \sum_{k=1}^{K} E_q[\log p(\beta_k \mid \lambda)] + \sum_{d=1}^{D} E_q[\log p(\pi^d \mid \alpha)]$$
$$+ \sum_{d=1}^{d} \sum_{i=1}^{N_d} E_q[\log p(z_i^d \mid \pi^d)] + E_q[\log p(w_i^d \mid z_i^d, \beta)]$$
$$- \sum_{k=1}^{K} E_q[\log q(\beta_k \mid \eta_k)] - \sum_{d=1}^{D} E_q[\log q(\pi^d \mid \gamma^d)] - \sum_{d=1}^{d} \sum_{i=1}^{N_d} E_q[\log q(z_i^d \mid \phi_i^d)]$$

Mean Field for LDA



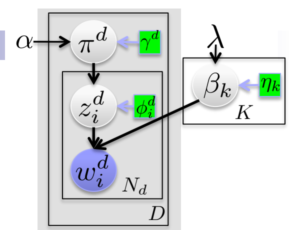
- Let's look at some of these terms

$E_q[\log p(z_i^d \mid \pi^d)]$

$E_q[\log q(z_i^d \mid \phi_i^d)]$

- Other terms follow similarly

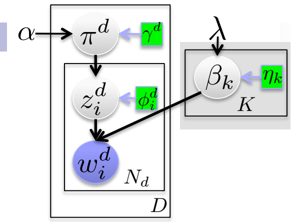
Optimize via Coordinate Ascent



- Algorithm:

Optimize via Coordinate Ascent

- Algorithm:



Generalizing… $\log p(x) \geq \int q_z(z) q_\theta(\theta) \log \frac{p(x,z,\theta)}{q_z(z) q_\theta} dz d\theta$

- Condition 1: Complete data likelihood is in exponential family
- Condition 2: Parameter prior is conjugate to complete data likelihood

EM for MAP estimation	**Variational Bayesian EM**
Goal: maximise $p(\boldsymbol{\theta}\,\vert\, x)$ w.r.t. $\boldsymbol{\theta}$	**Goal:** lower bound $p(x)$
E Step: compute	**VB-E Step:** compute $\bar{\phi}^{(t)} = \mathbb{E}_{q_\theta^{(t)}}[\phi(\theta)]$
$q_z^{(t+1)}(z) = p(z\,\vert\, x, \boldsymbol{\theta}^{(t)})$	$q_z^{(t+1)}(z) = p(z\,\vert\, x, \overline{\boldsymbol{\phi}}^{(t)})$
M Step:	**VB-M Step:**
$\boldsymbol{\theta}^{(t+1)} = \arg\max_{\boldsymbol{\theta}} \int q_z^{(t+1)}(z) \ln p(z, x, \boldsymbol{\theta})\, dz$	$q_{\boldsymbol{\theta}}^{(t+1)}(\boldsymbol{\theta}) \propto \exp\left[\int q_z^{(t+1)}(z) \ln p(z, x, \boldsymbol{\theta})\, dz\right]$

What you need to know…

- Latent Dirichlet allocation (LDA)
  - Motivation and generative model specification
  - Collapsed Gibbs sampler
- Variational methods
  - Overall goal
  - Interpretation in terms of minimizing (reverse) KL
  - Mean field approximation
  - Mean field for LDA

Reading

- **Mixed Membership Models: KM Sec. 27.3**
  - Basic LDA:
    Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." the Journal of machine Learning research 3 (2003): 993-1022.
  - Introduction:
    Blei, David M. "Probabilistic topic models." Communications of the ACM, vol. 55, no. 4 (2012): 77-84.
  - Sampling:
    Griffith, Thomas L. and Mark Steyvers. "Finding scientific topics." Proceedings of the National Academy of Sciences of the United States of America, Volume: 101, Supplement: 1 (2004): Pages: 5228-5235

Acknowledgements

- Thanks to Dave Blei, David Mimno, and Jordan Boyd-Graber for some material in this lecture relating to LDA