Topic Models

Advanced Machine Learning for NLP

Jordan Boyd-Graber

SLIDES ADAPTED FROM DAVID MIMNO

Learning the Hidden Space

- Two major tools:
  - Gibbs Sampling: Easier to implement, easier to understand
  - Variational Inference: faster, harder to implement
- Variational shows the connections to “deep” models better, so it’s the focus
- However, would be injustice to not at least discuss Gibbs sampling

Inference

- We are interested in posterior distribution

$$p(Z|X,\Theta) \tag{1}$$

Inference

- We are interested in posterior distribution

$$p(Z|X,\Theta) \tag{1}$$

- Here, latent variables are topic assignments z and topics θ. X is the words (divided into documents), and Θ are hyperparameters to Dirichlet distributions: α for topic proportion, λ for topics.

$$p(\boldsymbol{z},\boldsymbol{\beta},\boldsymbol{\theta}|\boldsymbol{w},\alpha,\lambda) \tag{2}$$

Inference

- We are interested in posterior distribution

$$p(Z|X,\Theta) \tag{1}$$

- Here, latent variables are topic assignments z and topics θ. X is the words (divided into documents), and Θ are hyperparameters to Dirichlet distributions: α for topic proportion, λ for topics.

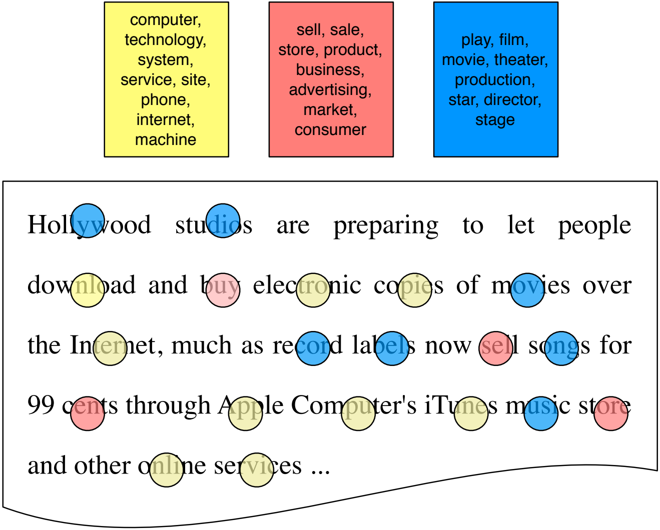
$$p(\boldsymbol{z},\boldsymbol{\beta},\boldsymbol{\theta}|\boldsymbol{w},\alpha,\lambda) \tag{2}$$

$$p(\boldsymbol{w},\boldsymbol{z},\boldsymbol{\theta},\boldsymbol{\beta}|\alpha,\lambda) = \prod_k p(\beta_k|\lambda)\prod_d p(\theta_d|\alpha)\prod_n p(z_{d,n}|\theta_d)p(w_{d,n}|\beta_{z_{d,n}})$$

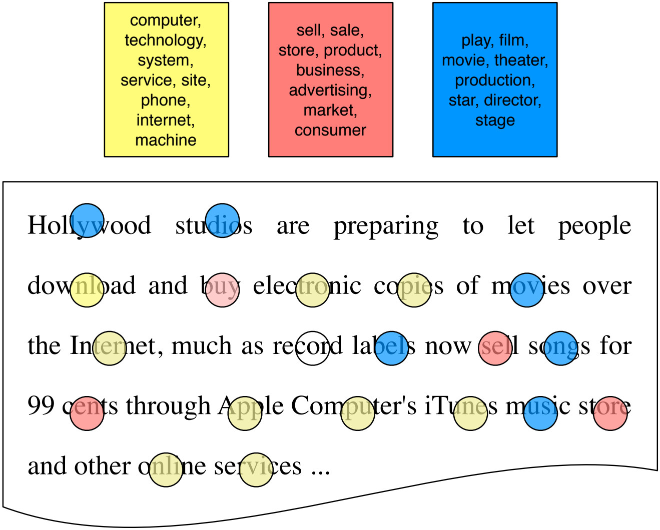
Gibbs Sampling

- A form of Markov Chain Monte Carlo
- Chain is a sequence of random variable states
- Given a state $\{z_1, \dots z_N\}$ given certain technical conditions, drawing $z_k \sim p(z_1, \dots z_{k-1}, z_{k+1}, \dots z_N | X, \Theta)$ for all k (repeatedly) results in a Markov Chain whose stationary distribution *is* the posterior.
- For notational convenience, call $\boldsymbol{z}$ with $z_{d,n}$ removed $\boldsymbol{z}_{-d,n}$

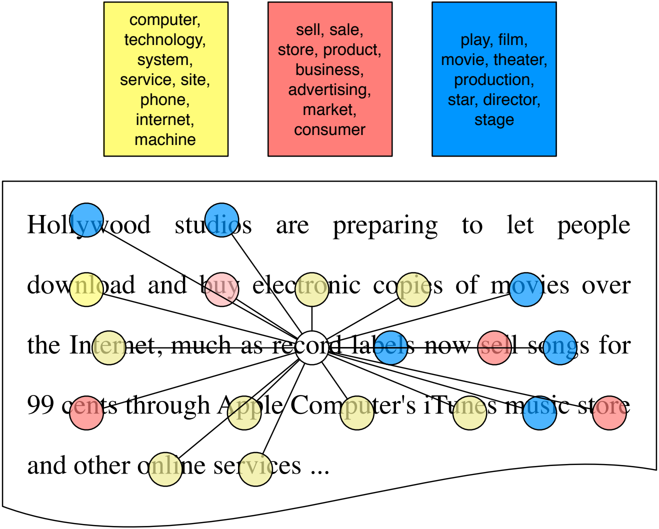
Inference


computer, technology, system, service, site, phone, internet, machine
sell, sale, store, product, business, advertising, market, consumer
play, film, movie, theater, production, star, director, stage
Hollywood studios are preparing to let people download and buy electronic copies of movies over the Internet, much as record labels now sell songs for 99 cents through Apple Computer's iTunes music store and other online services ...

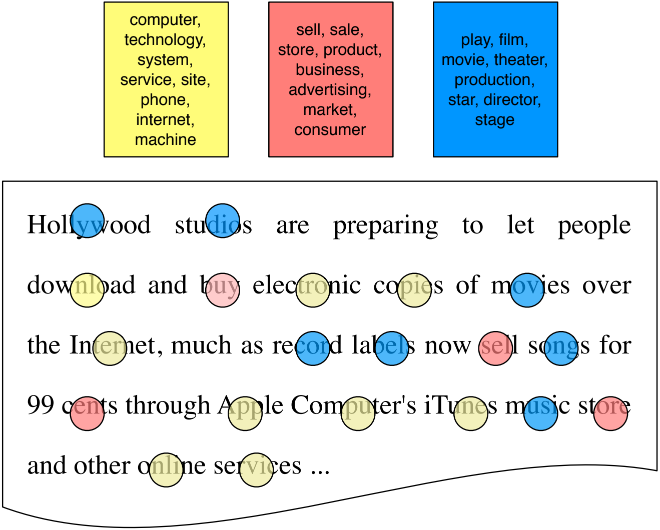
Inference


computer, technology, system, service, site, phone, internet, machine
sell, sale, store, product, business, advertising, market, consumer
play, film, movie, theater, production, star, director, stage
Hollywood studios are preparing to let people download and buy electronic copies of movies over the Internet, much as record labels now sell songs for 99 cents through Apple Computer's iTunes music store and other online services ...

Inference


computer, technology, system, service, site, phone, internet, machine
sell, sale, store, product, business, advertising, market, consumer
play, film, movie, theater, production, star, director, stage
Hollywood studios are preparing to let people download and buy electronic copies of movies over the Internet, much as record labels now sell songs for 99 cents through Apple Computer's iTunes music store and other online services ...

Inference


computer, technology, system, service, site, phone, internet, machine
sell, sale, store, product, business, advertising, market, consumer
play, film, movie, theater, production, star, director, stage
Hollywood studios are preparing to let people download and buy electronic copies of movies over the Internet, much as record labels now sell songs for 99 cents through Apple Computer's iTunes music store and other online services ...

Inference

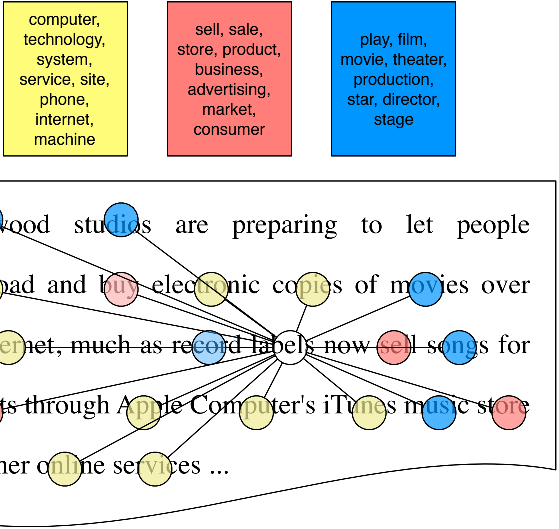
computer, technology, system, service, site, phone, internet, machine

sell, sale, store, product, business, advertising, market, consumer

play, film, movie, theater, production, star, director, stage

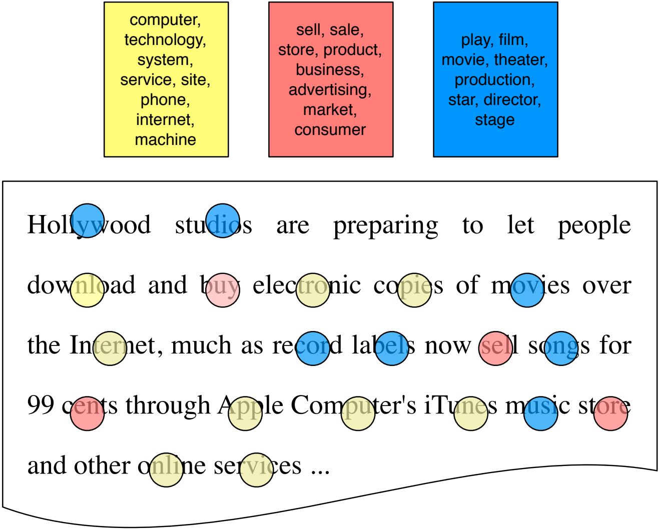
Hollywood studios are preparing to let people download and buy electronic copies of movies over the Internet, much as record labels now sell songs for 99 cents through Apple Computer's iTunes music store and other online services ...

Inference



computer, technology, system, service, site, phone, internet, machine
sell, sale, store, product, business, advertising, market, consumer
play, film, movie, theater, production, star, director, stage
Hollywood studios are preparing to let people download and buy electronic copies of movies over the Internet, much as record labels now sell songs for 99 cents through Apple Computer's iTunes music store and other online services ...

Inference


computer, technology, system, service, site, phone, internet, machine
sell, sale, store, product, business, advertising, market, consumer
play, film, movie, theater, production, star, director, stage
Hollywood studios are preparing to let people download and buy electronic copies of movies over the Internet, much as record labels now sell songs for 99 cents through Apple Computer's iTunes music store and other online services ...

Gibbs Sampling

- For LDA, we will sample the topic assignments
- Thus, we want:

$$p(z_{d,n} = k | \boldsymbol{z}_{-d,n}, \boldsymbol{w}, \alpha, \lambda) = \frac{p(z_{d,n} = k, \boldsymbol{z}_{-d,n} | \boldsymbol{w}, \alpha, \lambda)}{p(\boldsymbol{z}_{-d,n} | \boldsymbol{w}, \alpha, \lambda)}$$

Gibbs Sampling

- For LDA, we will sample the topic assignments
- Thus, we want:

$$p(z_{d,n} = k | \boldsymbol{z}_{-d,n}, \boldsymbol{w}, \alpha, \lambda) = \frac{p(z_{d,n} = k, \boldsymbol{z}_{-d,n} | \boldsymbol{w}, \alpha, \lambda)}{p(\boldsymbol{z}_{-d,n} | \boldsymbol{w}, \alpha, \lambda)}$$

- The topics and per-document topic proportions are integrated out / marginalized
- Let $n_{d,i}$ be the number of words taking topic i in document d. Let $v_{k,w}$ be the number of times word w is used in topic k.

$$= \frac{\int_{\theta_d} \left(\prod_{i \neq k} \theta_d^{\alpha_i + n_{d,i} - 1}\right) \theta_d^{\alpha_k + n_{d,k}} d\theta_d \int_{\beta_k} \left(\prod_{i \neq w_{d,n}} \beta_{k,i}^{\lambda_i + v_{k,i} - 1}\right) \beta_{k,w_{d,n}}^{\lambda_i + v_{k,w_{d,n}}} d\beta_k}{\int_{\theta_d} \left(\prod_i \theta_d^{\alpha_i + n_{d,i} - 1}\right) d\theta_d \int_{\beta_k} \left(\prod_i \beta_{k,i}^{\lambda_i + v_{k,i} - 1}\right) d\beta_k}$$

Gibbs Sampling

- Integral is normalizer of Dirichlet distribution

$$\int_{\beta_k} \left(\prod_i \beta_{k,i}^{\lambda_i + v_{k,i} - 1} \right) d\beta_k = \frac{\prod_i^V \Gamma\left(\beta_i + v_{k,i}\right)}{\Gamma\left(\sum_i^V \beta_i + v_{k,i}\right)}$$

Gibbs Sampling

- Integral is normalizer of Dirichlet distribution

$$\int_{\beta_k} \left(\prod_i \beta_{k,i}^{\lambda_i + v_{k,i} - 1} \right) d\beta_k = \frac{\prod_i^V \Gamma\left(\beta_i + v_{k,i}\right)}{\Gamma\left(\sum_i^V \beta_i + v_{k,i}\right)}$$

- So we can simplify

$$\frac{\int_{\theta_d} \left(\prod_{i \neq k} \theta_d^{\alpha_i + n_{d,i} - 1}\right) \theta_d^{\alpha_k + n_{d,k}} d\theta_d \int_{\beta_k} \left(\prod_{i \neq w_{d,n}} \beta_{k,i}^{\lambda_i + v_{k,i} - 1}\right) \beta_{k,w_{d,n}}^{\lambda_i + v_{k,w_{d,n}}} d\beta_k}{\int_{\theta_d} \left(\prod_i \theta_d^{\alpha_i + n_{d,i} - 1}\right) d\theta_d \int_{\beta_k} \left(\prod_i \beta_{k,i}^{\lambda_i + v_{k,i} - 1}\right) d\beta_k} =$$

$$\frac{\frac{\Gamma(\alpha_k + n_{d,k} + 1)}{\Gamma\left(\sum_i^K \alpha_i + n_{d,i} + 1\right)} \prod_{i \neq k}^K \Gamma\left(\alpha_k + n_{d,k}\right)}{\frac{\prod_i^K \Gamma(\alpha_i + n_{d,i})}{\Gamma\left(\sum_i^K \alpha_i + n_{d,i}\right)}} \frac{\frac{\Gamma\left(\lambda_{w_{d,n}} + v_{k,w_{d,n}} + 1\right)}{\Gamma\left(\sum_i^V \lambda_i + v_{k,i} + 1\right)} \prod_{i \neq w_{d,n}}^V \Gamma\left(\lambda_k + v_{k,w_{d,n}}\right)}{\frac{\prod_i^V \Gamma\left(\lambda_i + v_{k,i}\right)}{\Gamma\left(\sum_i^V \lambda_i + v_{k,i}\right)}}$$

Gibbs Sampling

- Integral is normalizer of Dirichlet distribution

$$\int_{\beta_k} \left(\prod_i \beta_{k,i}^{\lambda_i + v_{k,i} - 1} \right) d\beta_k = \frac{\prod_i^V \Gamma\left(\beta_i + v_{k,i}\right)}{\Gamma\left(\sum_i^V \beta_i + v_{k,i}\right)}$$

- So we can simplify

$$\frac{\int_{\theta_d} \left(\prod_{i \neq k} \theta_d^{\alpha_i + n_{d,i} - 1} \right) \theta_d^{\alpha_k + n_{d,k}} d\theta_d \int_{\beta_k} \left(\prod_{i \neq w_{d,n}} \beta_{k,i}^{\lambda_i + v_{k,i} - 1} \right) \beta_{k,w_{d,n}}^{\lambda_i + v_{k,w_{d,n}}} d\beta_k}{\int_{\theta_d} \left(\prod_i \theta_d^{\alpha_i + n_{d,i} - 1} \right) d\theta_d \int_{\beta_k} \left(\prod_i \beta_{k,i}^{\lambda_i + v_{k,i} - 1} \right) d\beta_k} =$$

$$\frac{\frac{\Gamma(\alpha_k + n_{d,k} + 1)}{\Gamma\left(\sum_i^K \alpha_i + n_{d,i} + 1\right)} \prod_{i \neq k}^K \Gamma\left(\alpha_k + n_{d,k}\right)}{\frac{\prod_i^K \Gamma(\alpha_i + n_{d,i})}{\Gamma\left(\sum_i^K \alpha_i + n_{d,i}\right)}} \frac{\frac{\Gamma\left(\lambda_{w_{d,n}} + v_{k,w_{d,n}} + 1\right)}{\Gamma\left(\sum_i^V \lambda_i + v_{k,i} + 1\right)} \prod_{i \neq w_{d,n}}^V \Gamma\left(\lambda_k + v_{k,w_{d,n}}\right)}{\frac{\prod_i^V \Gamma\left(\lambda_i + v_{k,i}\right)}{\Gamma\left(\sum_i^V \lambda_i + v_{k,i}\right)}}$$

Gamma Function Identity

$$z = \frac{\Gamma(z+1)}{\Gamma(z)} \tag{3}$$

$$\frac{\frac{\Gamma(\alpha_k + n_{d,k} + 1)}{\Gamma\left(\sum_i^K \alpha_i + n_{d,i} + 1\right)} \prod_{i \neq k}^K \Gamma\left(\alpha_k + n_{d,k}\right)}{\frac{\prod_i^K \Gamma(\alpha_i + n_{d,i})}{\Gamma\left(\sum_i^K \alpha_i + n_{d,i}\right)}} \frac{\frac{\Gamma\left(\lambda_{w_{d,n}} + v_{k,w_{d,n}} + 1\right)}{\Gamma\left(\sum_i^V \lambda_i + v_{k,i} + 1\right)} \prod_{i \neq w_{d,n}}^V \Gamma\left(\lambda_k + v_{k,w_{d,n}}\right)}{\frac{\prod_i^V \Gamma\left(\lambda_i + v_{k,i}\right)}{\Gamma\left(\sum_i^V \lambda_i + v_{k,i}\right)}}$$

$$= \frac{n_{d,k} + \alpha_k}{\sum_i^K n_{d,i} + \alpha_i} \frac{v_{k,w_{d,n}} + \lambda_{w_{d,n}}}{\sum_i v_{k,i} + \lambda_i}$$

Gibbs Sampling Equation

$$\frac{n_{d,k} + \alpha_k}{\sum_i^K n_{d,i} + \alpha_i} \frac{v_{k,w_{d,n}} + \lambda_{w_{d,n}}}{\sum_i v_{k,i} + \lambda_i} \tag{4}$$

- Number of times document d uses topic k
- Number of times topic k uses word type $w_{d,n}$
- Dirichlet parameter for document to topic distribution
- Dirichlet parameter for topic to word distribution
- How much this document likes topic k
- How much this topic likes word $w_{d,n}$

Gibbs Sampling Equation

$$\frac{n_{d,k} + \alpha_k}{\sum_i^K n_{d,i} + \alpha_i} \frac{v_{k,w_{d,n}} + \lambda_{w_{d,n}}}{\sum_i v_{k,i} + \lambda_i} \tag{4}$$

- Number of times document d uses topic k
- Number of times topic k uses word type $w_{d,n}$
- Dirichlet parameter for document to topic distribution
- Dirichlet parameter for topic to word distribution
- How much this document likes topic k
- How much this topic likes word $w_{d,n}$

Gibbs Sampling Equation

$$\frac{n_{d,k} + \alpha_k}{\sum_i^K n_{d,i} + \alpha_i} \frac{v_{k,w_{d,n}} + \lambda_{w_{d,n}}}{\sum_i v_{k,i} + \lambda_i} \tag{4}$$

- Number of times document d uses topic k
- Number of times topic k uses word type $w_{d,n}$
- Dirichlet parameter for document to topic distribution
- Dirichlet parameter for topic to word distribution
- How much this document likes topic k
- How much this topic likes word $w_{d,n}$

Gibbs Sampling Equation

$$\frac{n_{d,k} + \alpha_k}{\sum_i^K n_{d,i} + \alpha_i} \frac{v_{k,w_{d,n}} + \lambda_{w_{d,n}}}{\sum_i v_{k,i} + \lambda_i} \tag{4}$$

- Number of times document d uses topic k
- Number of times topic k uses word type $w_{d,n}$
- Dirichlet parameter for document to topic distribution
- Dirichlet parameter for topic to word distribution
- How much this document likes topic k
- How much this topic likes word $w_{d,n}$

Gibbs Sampling Equation

$$\frac{n_{d,k} + \alpha_k}{\sum_i^K n_{d,i} + \alpha_i} \frac{v_{k,w_{d,n}} + \lambda_{w_{d,n}}}{\sum_i v_{k,i} + \lambda_i} \tag{4}$$

- Number of times document d uses topic k
- Number of times topic k uses word type $w_{d,n}$
- Dirichlet parameter for document to topic distribution
- Dirichlet parameter for topic to word distribution
- How much this document likes topic k
- How much this topic likes word $w_{d,n}$

Gibbs Sampling Equation

$$\frac{n_{d,k} + \alpha_k}{\sum_i^K n_{d,i} + \alpha_i} \frac{v_{k,w_{d,n}} + \lambda_{w_{d,n}}}{\sum_i v_{k,i} + \lambda_i} \tag{4}$$

- Number of times document d uses topic k
- Number of times topic k uses word type $w_{d,n}$
- Dirichlet parameter for document to topic distribution
- Dirichlet parameter for topic to word distribution
- How much this document likes topic k
- How much this topic likes word $w_{d,n}$

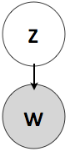
Sample Document

Etruscan	trade	price	temple	market

Sample Document

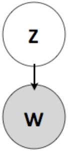
Etruscan	trade	price	temple	market

Randomly Assign Topics

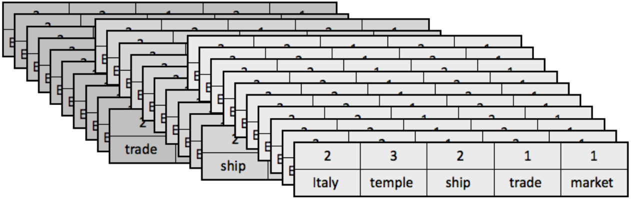


3	2	1	3	1
Etruscan	trade	price	temple	market

Randomly Assign Topics



3	2	1	3	1
Etruscan	trade	price	temple	market



Total Topic Counts

3	2	1	3	1
Etruscan	trade	price	temple	market

Total counts from **all** docs

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

Total Topic Counts

3	2	1	3	1
Etruscan	trade	price	temple	market

Total

	1	2	3
Etruscan	1	0	35
market	50	0	1
...			

Sampling Equation

$$\frac{n_{d,k} + \alpha_k}{\sum_i^K n_{d,i} + \alpha_i} \frac{v_{k,w_{d,n}} + \lambda_{w_{d,n}}}{\sum_i v_{k,i} + \lambda_i}$$

Total Topic Counts

3	2	1	3	1
Etruscan	trade	price	temple	market

Total

	1	2	3
Etruscan	1	0	35
market	50	0	1
...			

Sampling Equation

$$\frac{n_{d,k} + \alpha_k}{\sum_i^K n_{d,i} + \alpha_i} \frac{v_{k,w_{d,n}} + \lambda_{w_{d,n}}}{\sum_i v_{k,i} + \lambda_i}$$

We want to sample this word . . .

3	**2**	1	3	1
Etruscan	trade	price	temple	market

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

We want to sample this word . . .

3	**2**	1	3	1
Etruscan	trade	price	temple	market

	1	2	3
Etruscan	1	0	35
market	50	0	1
price	42	1	0
temple	0	0	20
trade	10	**8**	1
...			

Decrement its count

3	?	1	3	1
Etruscan	trade	price	temple	market

	1	2	3
Etruscan	1	0	35
market	50	0	1
price	42	1	0
temple	0	0	20
trade	10	**7**	1
...			

What is the conditional distribution for this topic?

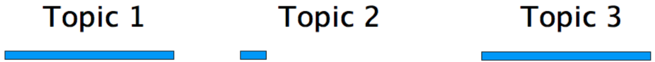
3	?	1	3	1
Etruscan	trade	price	temple	market

Part 1: How much does this document like each topic?

3	?	1	3	1
Etruscan	trade	price	temple	market

Part 1: How much does this document like each topic?

3	?	1	3	1
Etruscan	trade	price	temple	market



Part 1: How much does this document like each topic?

3	?	1	3	1
Etruscan	trade	price	temple	market

Topic 1 Topic 2 Topic 3

Sampling Equation

$$\frac{n_{d,k} + \alpha_k}{\sum_i^K n_{d,i} + \alpha_i} \frac{v_{k,w_{d,n}} + \lambda_{w_{d,n}}}{\sum_i v_{k,i} + \lambda_i}$$

Part 1: How much does this document like each topic?

3	?	1	3	1
Etruscan	trade	price	temple	market

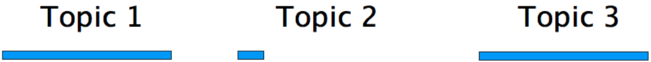
Topic 1 Topic 2 Topic 3

Sampling Equation

$$\frac{n_{d,k} + \alpha_k}{\sum_i^K n_{d,i} + \alpha_i} \frac{v_{k,w_{d,n}} + \lambda_{w_{d,n}}}{\sum_i v_{k,i} + \lambda_i}$$

Part 2: How much does each topic like the word?

3	?	1	3	1
Etruscan	trade	price	temple	market



	1	2	3
trade	10	7	1

Part 2: How much does each topic like the word?

3	?	1	3	1
Etruscan	trade	price	temple	market

Topic 1 Topic 2 Topic 3

Sampling Equation

$$\frac{n_{d,k}+\alpha_k}{\sum_i^K n_{d,i}+\alpha_i}\frac{v_{k,w_{d,n}}+\lambda_{w_{d,n}}}{\sum_i v_{k,i}+\lambda_i}$$

trade	10	7	1

Part 2: How much does each topic like the word?

3	?	1	3	1
Etruscan	trade	price	temple	market

	Topic 1	Topic 2	Topic 3
trade	10	7	1

Sampling Equation

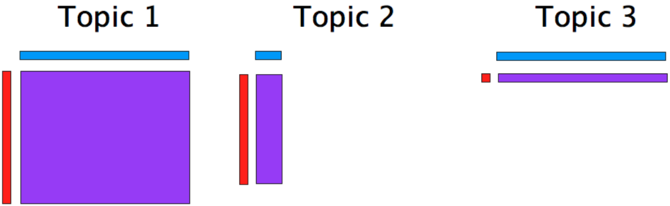
$$\frac{n_{d,k} + \alpha_k}{\sum_i^K n_{d,i} + \alpha_i} \frac{v_{k,w_{d,n}} + \lambda_{w_{d,n}}}{\sum_i v_{k,i} + \lambda_i}$$

Geometric interpretation

3	?	1	3	1
Etruscan	trade	price	temple	market

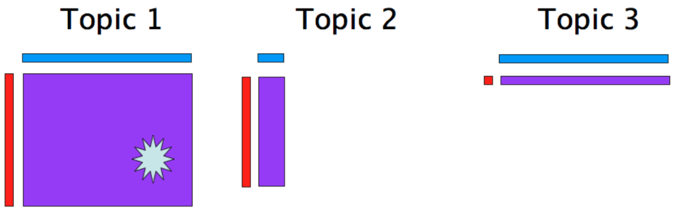
Geometric interpretation

3	?	1	3	1
Etruscan	trade	price	temple	market



Geometric interpretation

3	?	1	3	1
Etruscan	trade	price	temple	market



Update counts

3	?	1	3	1
Etruscan	trade	price	temple	market

	1	2	3
Etruscan	1	0	35
market	50	0	1
price	42	1	0
temple	0	0	20
trade	**10**	7	1
...			

Update counts

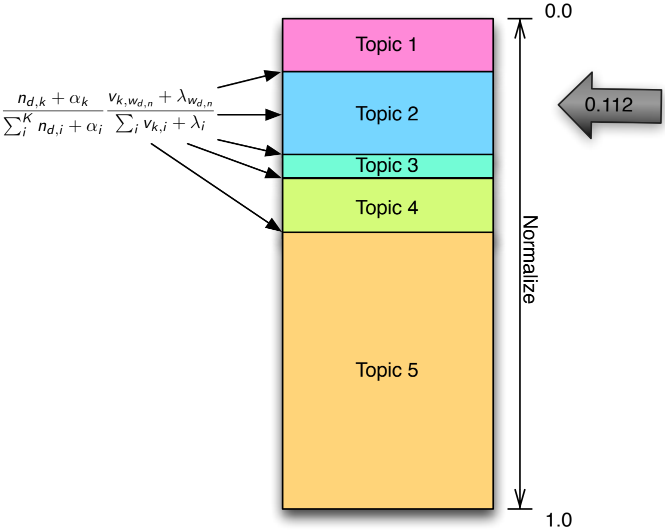
3	**1**	1	3	1
Etruscan	trade	price	temple	market

	1	2	3
Etruscan	1	0	35
market	50	0	1
price	42	1	0
temple	0	0	20
trade	**11**	7	1
...			

Update counts

3	1	1	3	1
Etruscan	trade	price	temple	market

Details: how to sample from a distribution



Algorithm

1. For each iteration i:
   1. For each document d and word n currently assigned to z_{old}:
      1. Decrement $n_{d,z_{old}}$ and $v_{z_{old},w_{d,n}}$
      2. Sample $z_{new} = k$ with probability proportional to $\frac{n_{d,k}+\alpha_k}{\sum_i^K n_{d,i}+\alpha_i} \frac{v_{k,w_{d,n}}+\lambda_{w_{d,n}}}{\sum_i v_{k,i}+\lambda_i}$
      3. Increment $n_{d,z_{new}}$ and $v_{z_{new},w_{d,n}}$

Algorithm

1. For each iteration i:
   1. For each document d and word n currently assigned to z_{old}:
      1. Decrement $n_{d,z_{old}}$ and $v_{z_{old},w_{d,n}}$
      2. Sample $z_{new} = k$ with probability proportional to $\frac{n_{d,k}+\alpha_k}{\sum_i^K n_{d,i}+\alpha_i} \frac{v_{k,w_{d,n}}+\lambda_{w_{d,n}}}{\sum_i v_{k,i}+\lambda_i}$
      3. Increment $n_{d,z_{new}}$ and $v_{z_{new},w_{d,n}}$

Desiderata

- Hyperparameters: Sample them too (slice sampling)
- Initialization: Random
- Sampling: Until likelihood converges
- Lag / burn-in: Difference of opinion on this
- Number of chains: Should do more than one

Available implementations

- Mallet (http://mallet.cs.umass.edu)
- LDAC (http://www.cs.princeton.edu/ blei/lda-c)
- Topicmod (http://code.google.com/p/topicmod)