Topic Models

Jordan Boyd-Graber

1. April 2014

Outline

Why topic models?

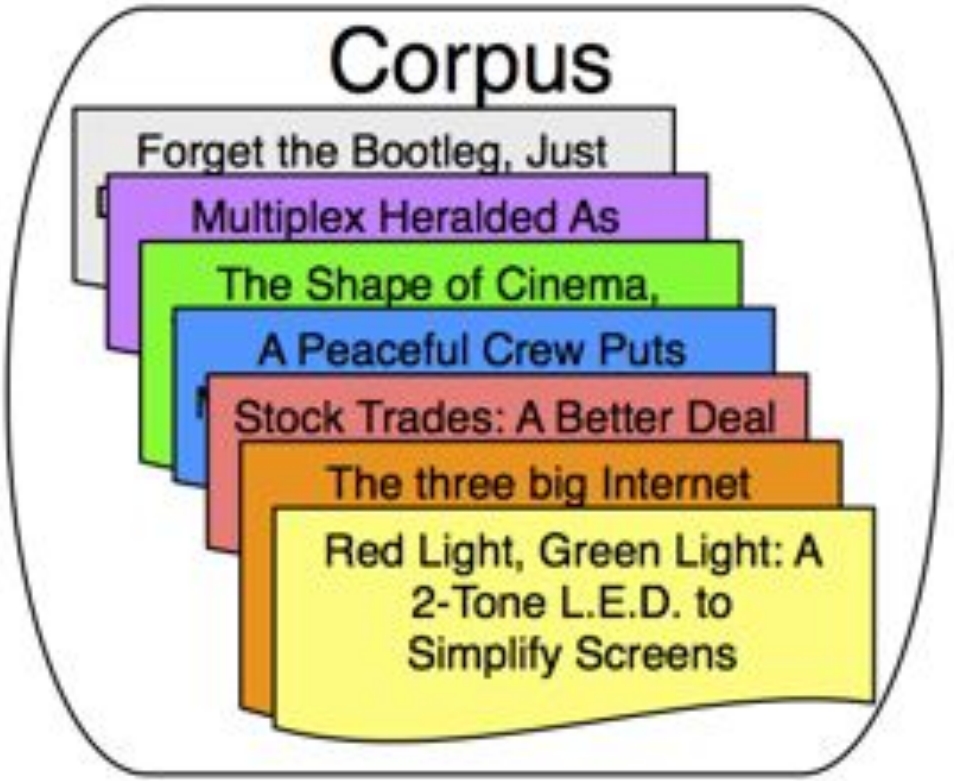
- Suppose you have a huge number of documents
- Want to know what's going on
- Can't read them all (e.g. every New York Times article from the 90's)
- Topic models offer a way to get a corpus-level view of major themes

Why topic models?

- Suppose you have a huge number of documents
- Want to know what's going on
- Can't read them all (e.g. every New York Times article from the 90's)
- Topic models offer a way to get a corpus-level view of major themes
- Unsupervised

Conceptual Approach

From an **input corpus** and number of topics $K \rightarrow$ words to topics



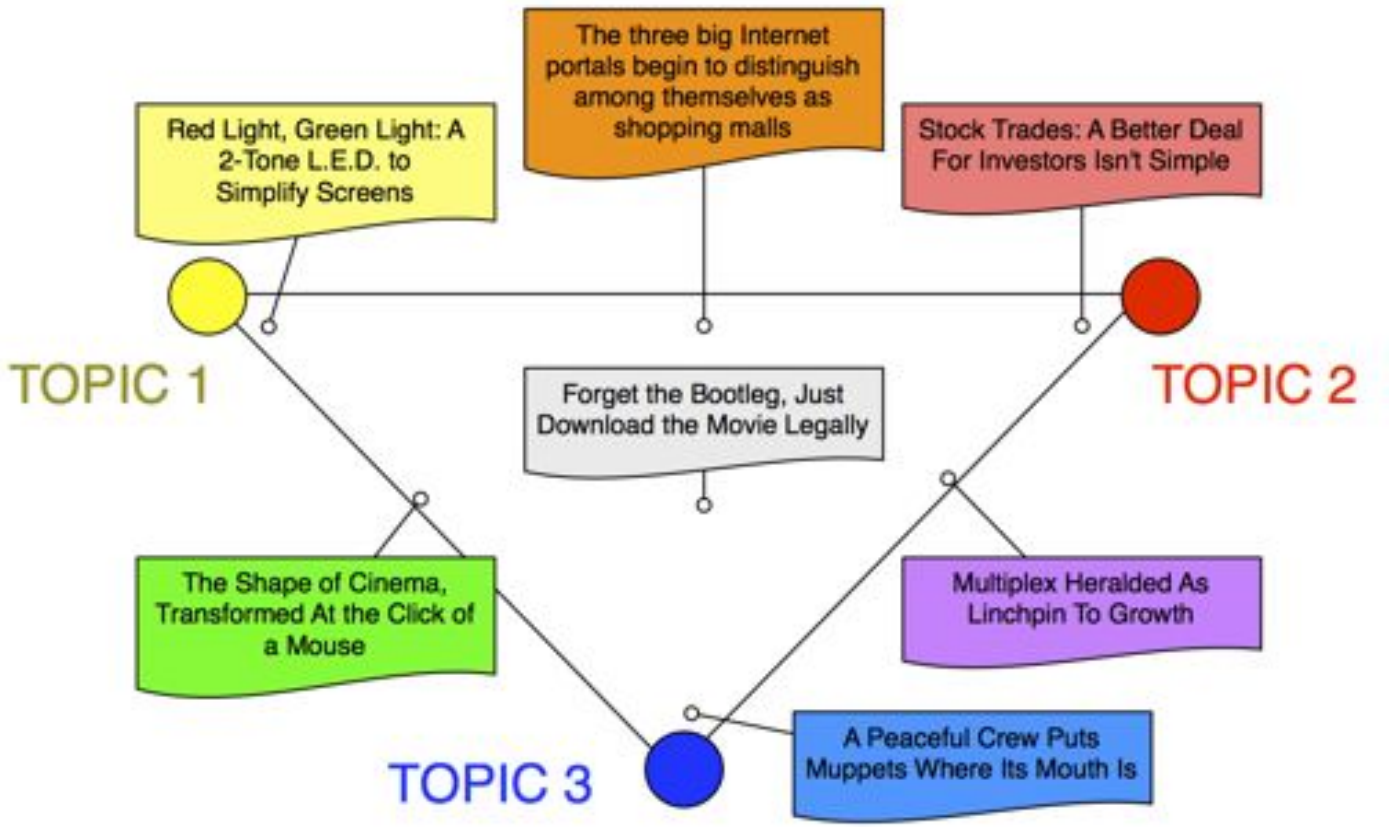
Conceptual Approach

From an input corpus and number of topics $K \rightarrow$ **words to topics**

TOPIC 1	TOPIC 2	TOPIC 3
computer, technology, system, service, site, phone, internet, machine	sell, sale, store, product, business, advertising, market, consumer	play, film, movie, theater, production, star, director, stage

Conceptual Approach

- For each document, what topics are expressed by that document?



Topics from *Science*

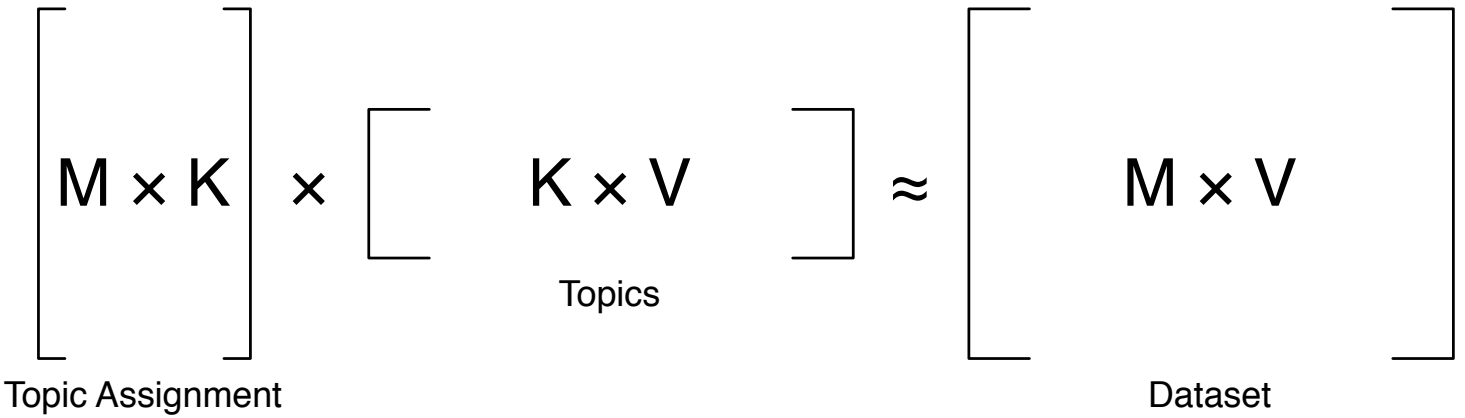
human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living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

Why should you care?

- Neat way to explore / understand corpus collections
  - E-discovery
  - Social media
  - Scientific data
- NLP Applications
  - POS Tagging [11]
  - Word Sense Disambiguation [3]
  - Word Sense Induction [4]
  - Discourse Segmentation [10]
- Psychology [6]: word meaning, polysemy
- Inference is (relatively) simple

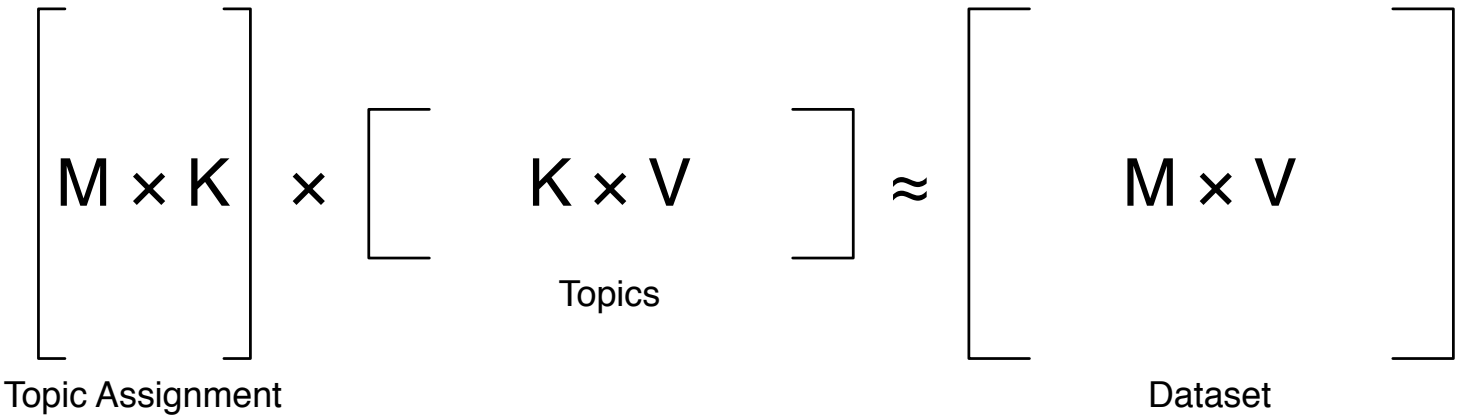
Outline

Matrix Factorization Approach



- K Number of topics
- M Number of documents
- V Size of vocabulary

Matrix Factorization Approach



K Number of topics

M Number of documents

V Size of vocabulary

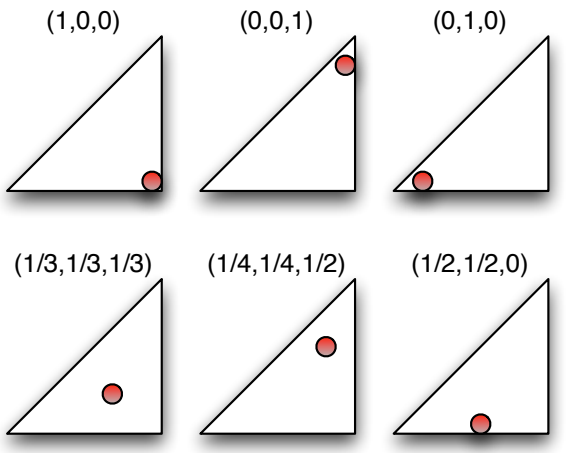
- If you use singular value decomposition (SVD), this technique is called latent semantic analysis.
- Popular in information retrieval.

Alternative: Generative Model

- How your data came to be
- Sequence of Probabilistic Steps
- Posterior Inference

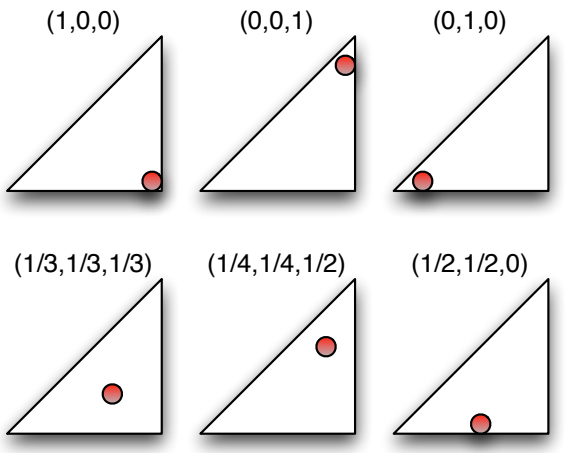
Multinomial Distribution

- Distribution over discrete outcomes
- Represented by non-negative vector that sums to one
- Picture representation



Multinomial Distribution

- Distribution over discrete outcomes
- Represented by non-negative vector that sums to one
- Picture representation

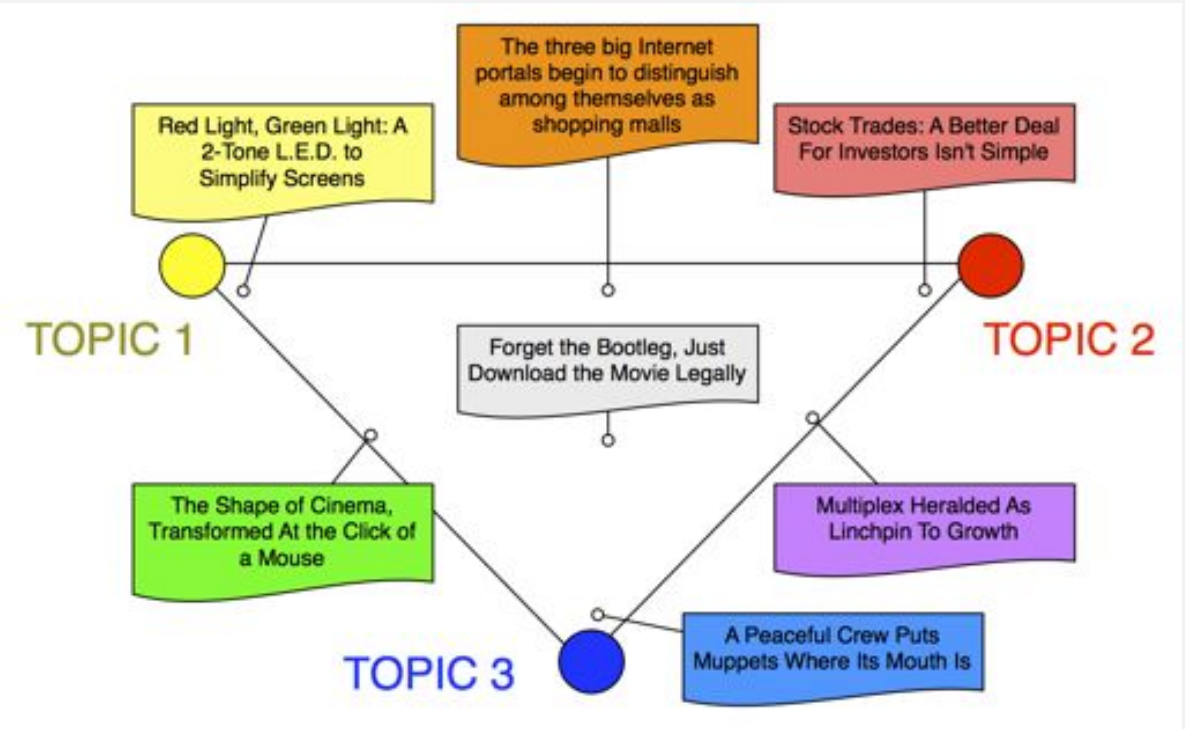


- Come from a Dirichlet distribution

Multinomial Distribution

- Distribution over discrete outcomes
- Represented by non-negative vector that sums to one

Look familiar?

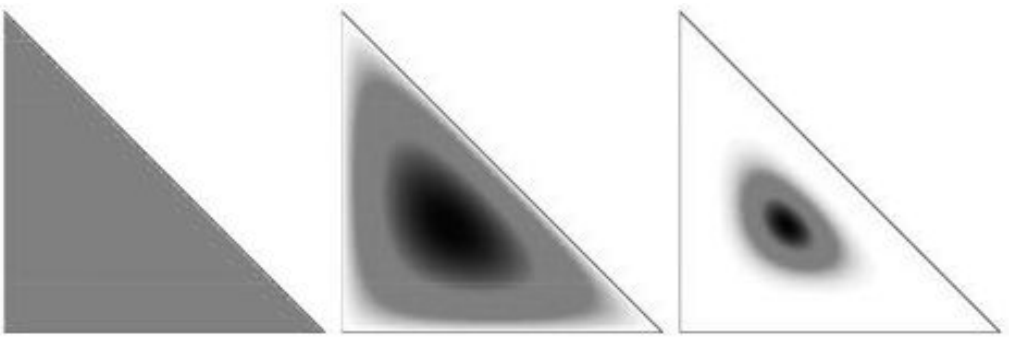


Dirichlet Distribution

$$p(\theta|\vec{\alpha}) = \frac{\Gamma\left(\sum_k \alpha_k\right)}{\prod_k \Gamma(\alpha_k)} \prod_k \theta_k^{\alpha_k - 1} \tag{1}$$

Dirichlet Distribution

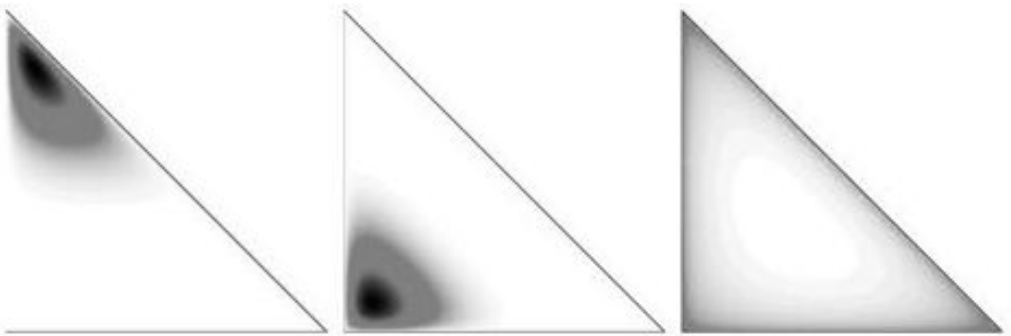
$$p(\theta|\vec{\alpha}) = \frac{\Gamma\left(\sum_k \alpha_k\right)}{\prod_k \Gamma(\alpha_k)} \prod_k \theta_k^{\alpha_k - 1} \tag{1}$$



$\vec{\alpha} = (1, 1, 1)$ $\vec{\alpha} = (2, 2, 2)$ $\vec{\alpha} = (10, 10, 10)$

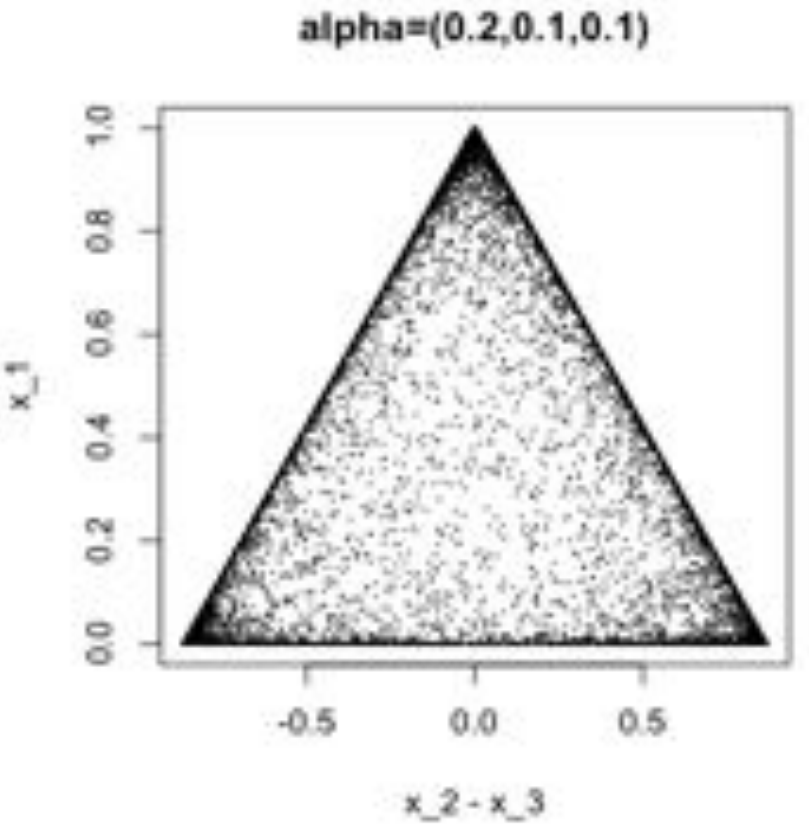
Dirichlet Distribution

$$p(\theta|\vec{\alpha}) = \frac{\Gamma\left(\sum_k \alpha_k\right)}{\prod_k \Gamma(\alpha_k)} \prod_k \theta_k^{\alpha_k - 1} \tag{1}$$



$\vec{\alpha} = (2, 10, 2)$ $\vec{\alpha} = (2, 2, 10)$ $\vec{\alpha} = \left(\frac{4}{5}, \frac{4}{5}, \frac{4}{5}\right)$

Dirichlet Distribution



Dirichlet Distribution

- If $\boldsymbol{\phi} \sim \mathrm{Dir}(\alpha)$, $\mathbf{w} \sim \mathrm{Mult}(\phi)$, and $n_k = |\{w_i : w_i = k\}|$ then

$$p(\phi|\alpha, \mathbf{w}) \propto p(\mathbf{w}|\phi)p(\phi|\alpha) \tag{2}$$

$$\propto \prod_k \phi^{n_k} \prod_k \phi^{\alpha_k - 1} \tag{3}$$

$$\propto \prod_k \phi^{\alpha_k + n_k - 1} \tag{4}$$

- Conjugacy: this **posterior** has the same form as the **prior**

Dirichlet Distribution

- If $\boldsymbol{\phi} \sim \mathrm{Dir}(\alpha)$, $\mathbf{w} \sim \mathrm{Mult}(\phi)$, and $n_k = |\{w_i : w_i = k\}|$ then

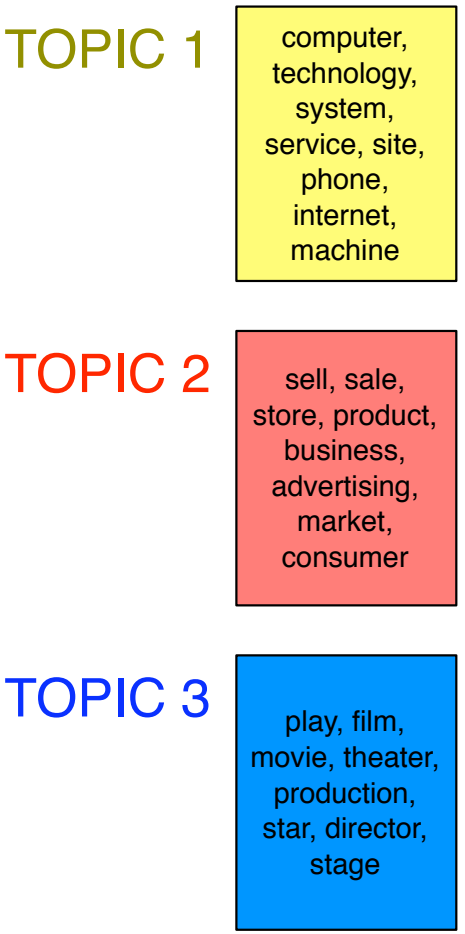
$$p(\phi|\alpha, \mathbf{w}) \propto p(\mathbf{w}|\phi)p(\phi|\alpha) \tag{2}$$

$$\propto \prod_k \phi^{n_k} \prod_k \phi^{\alpha_k - 1} \tag{3}$$

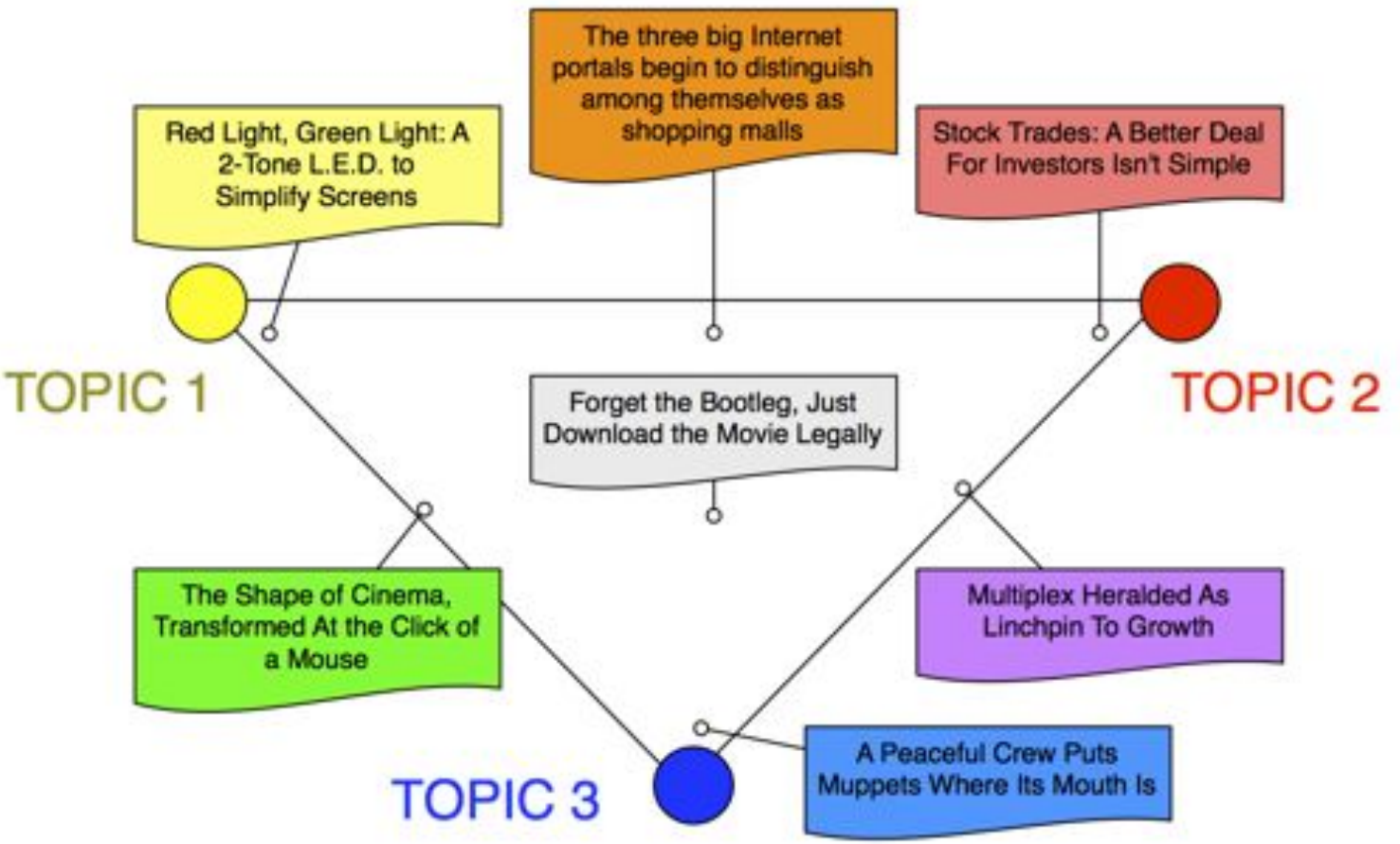
$$\propto \prod_k \phi^{\alpha_k + n_k - 1} \tag{4}$$

- Conjugacy: this **posterior** has the same form as the **prior**

Generative Model



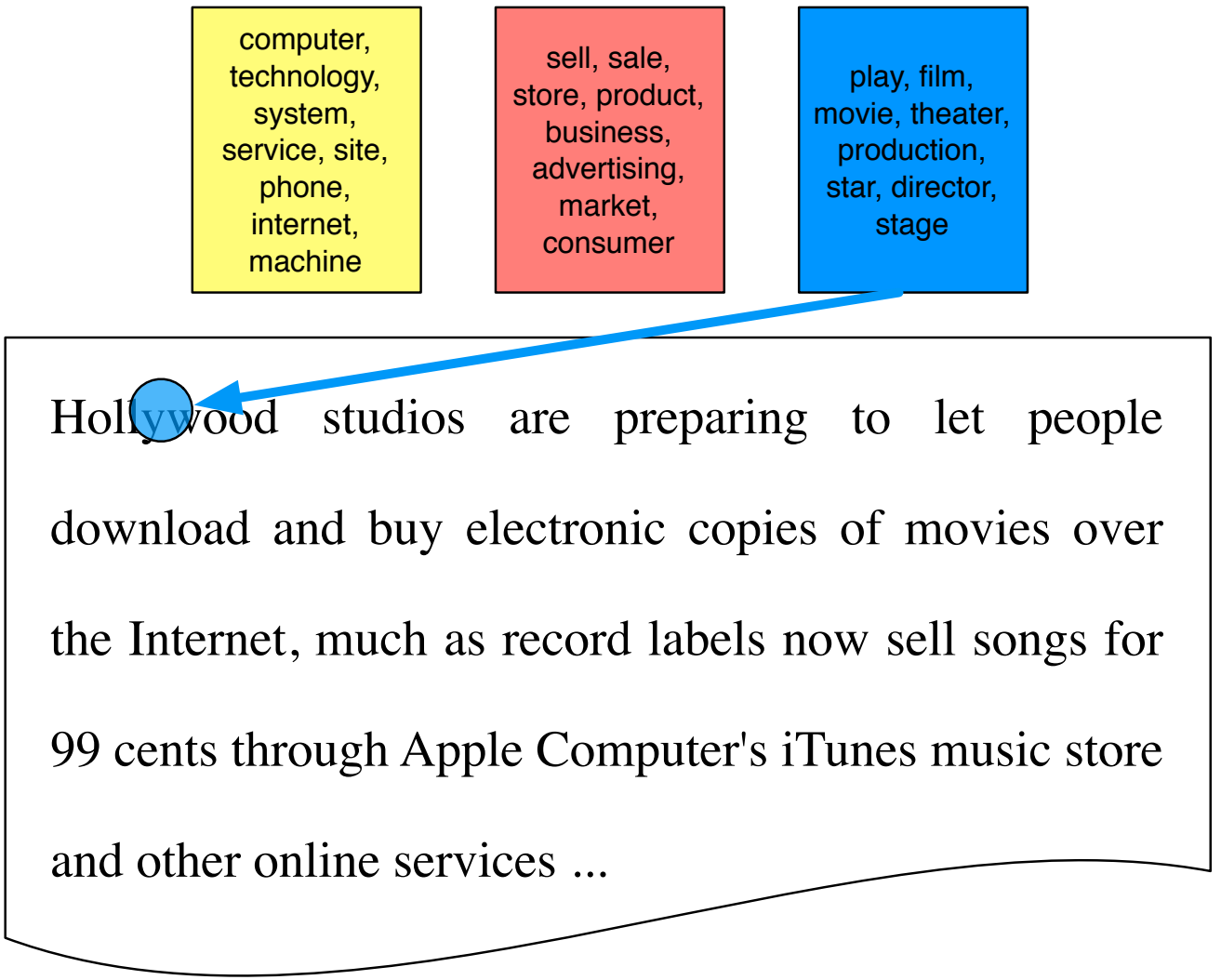
Generative Model



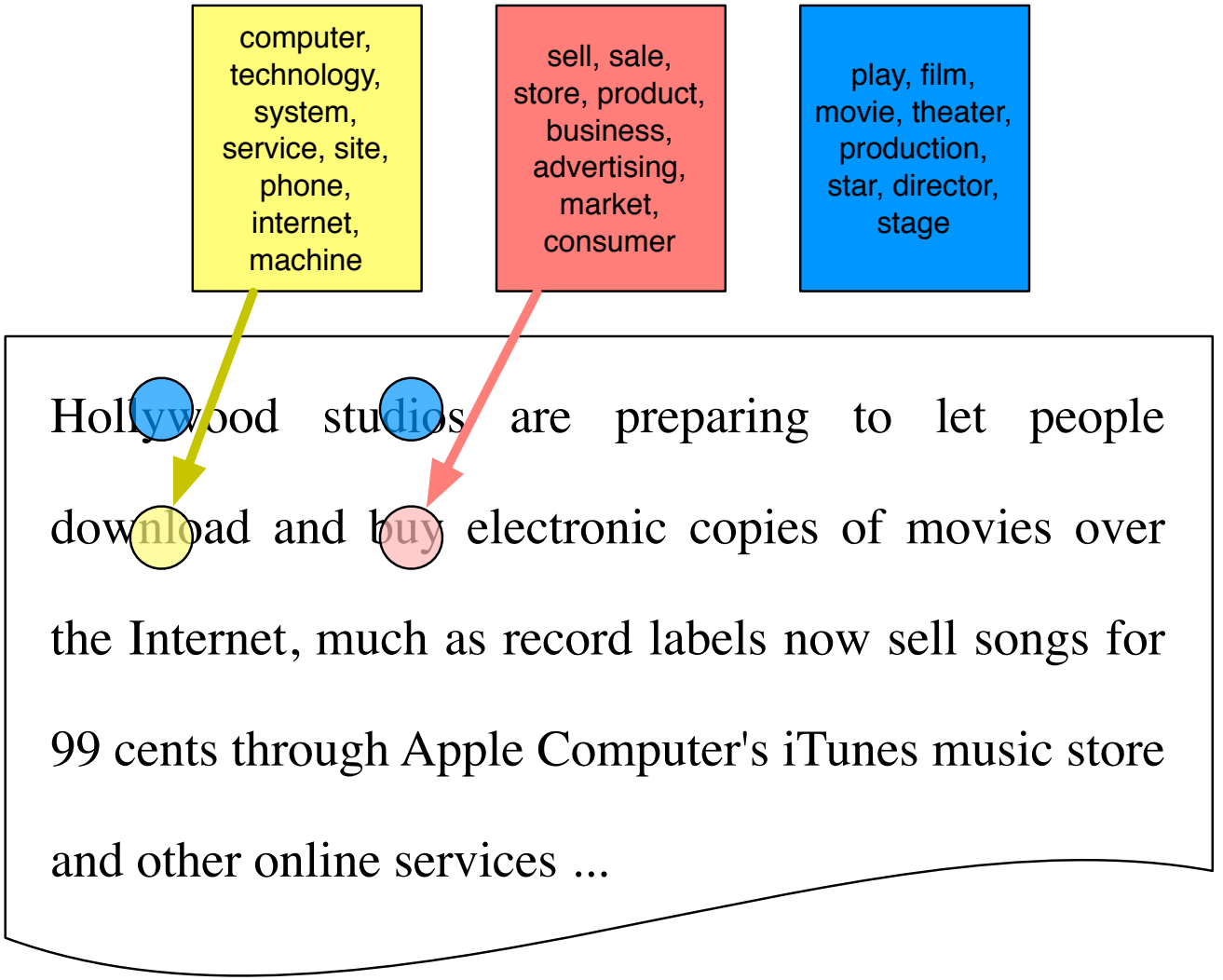
Generative Model

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 ...

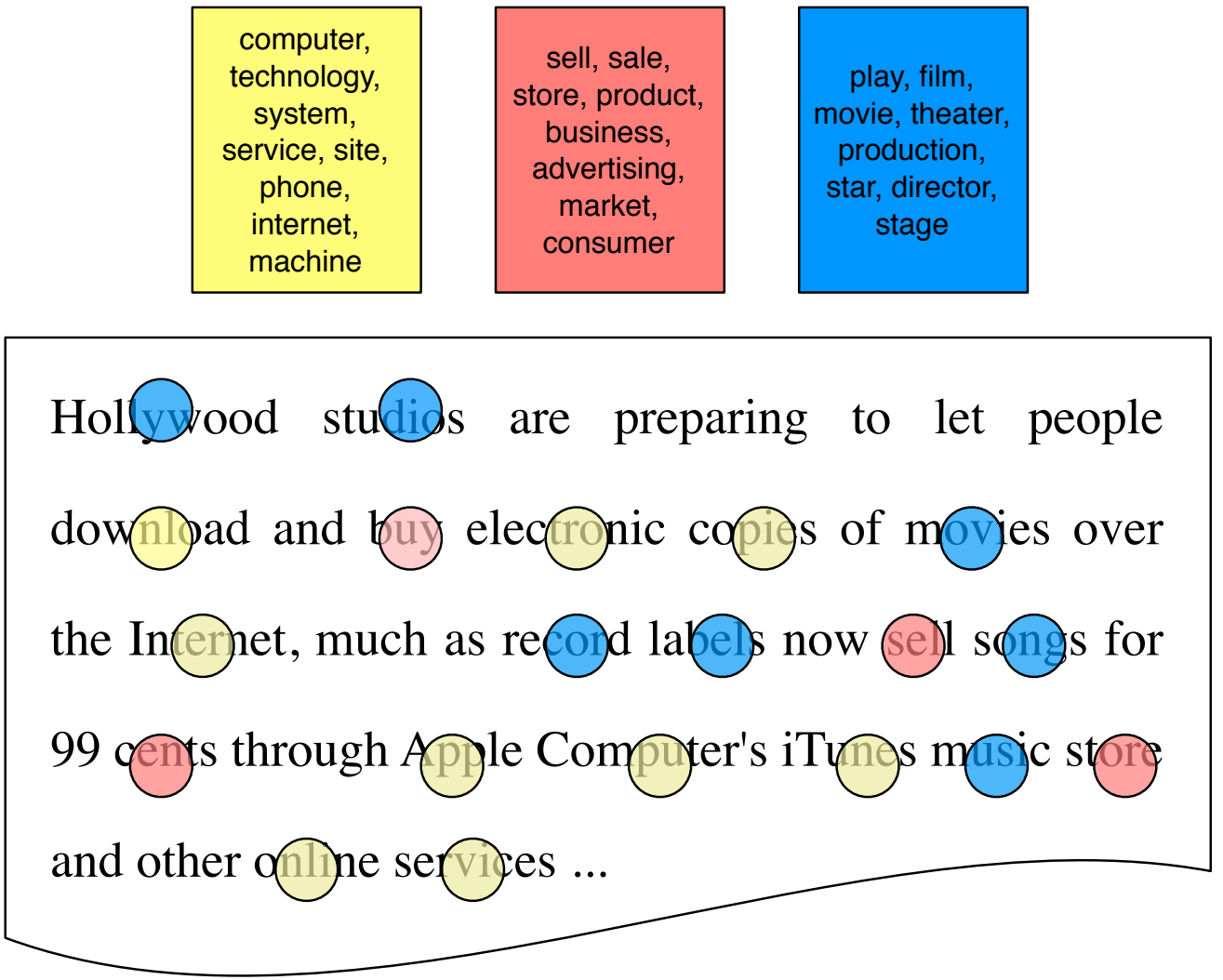
Generative Model



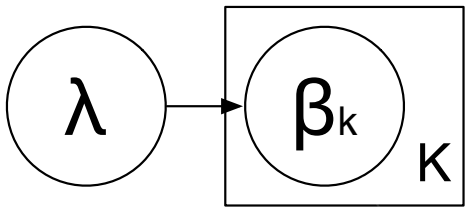
Generative Model



Generative Model

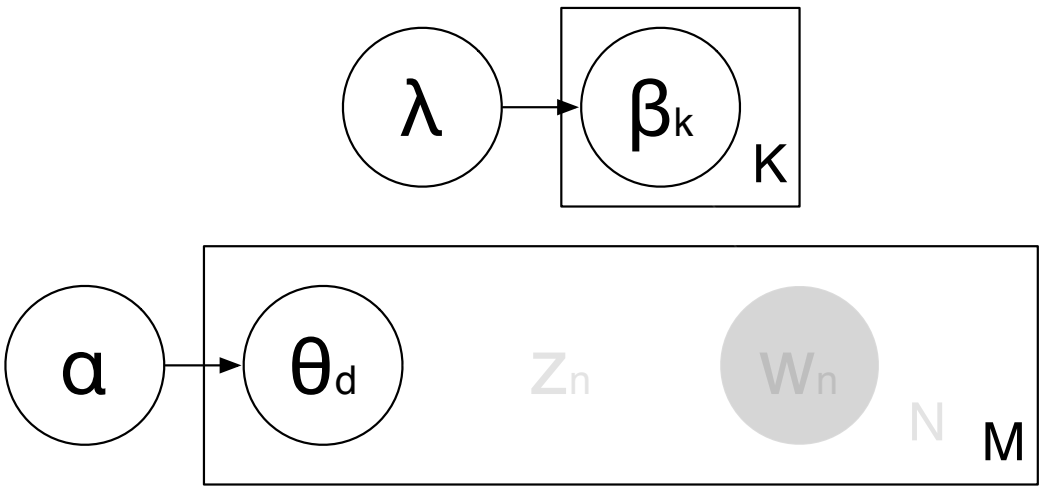


Generative Model Approach



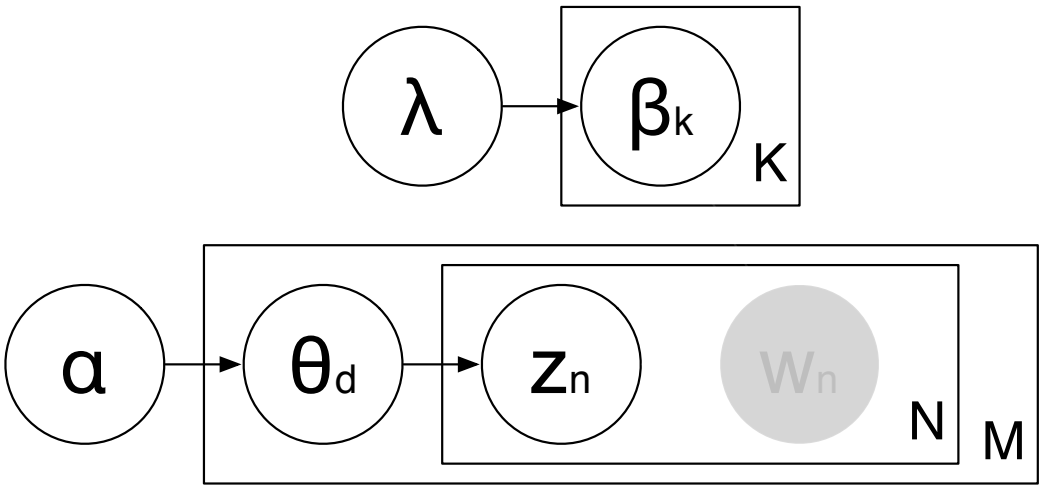
- For each topic $k \in \{1, \ldots, K\}$, draw a multinomial distribution β_k from a Dirichlet distribution with parameter λ

Generative Model Approach



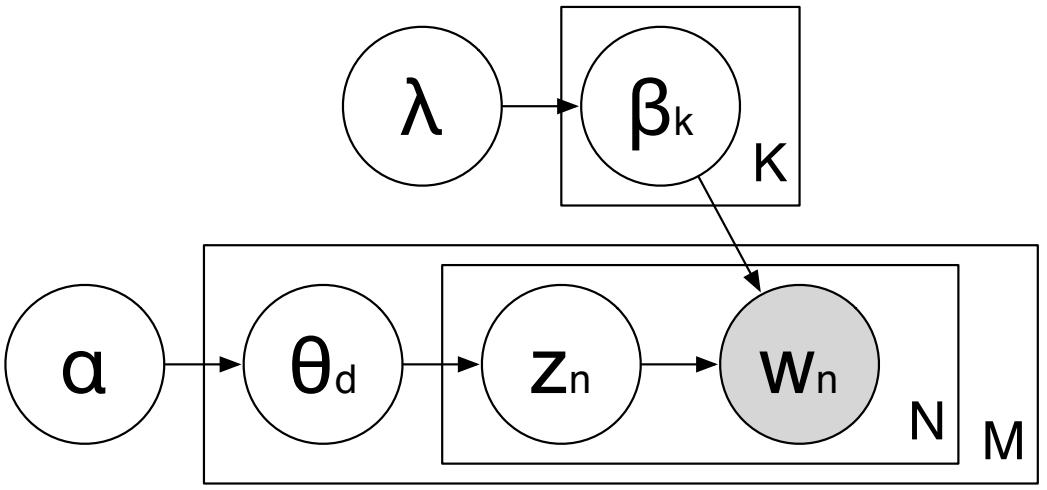
- For each topic $k \in \{1, \ldots, K\}$, draw a multinomial distribution β_k from a Dirichlet distribution with parameter λ
- For each document $d \in \{1, \ldots, M\}$, draw a multinomial distribution θ_d from a Dirichlet distribution with parameter α

Generative Model Approach



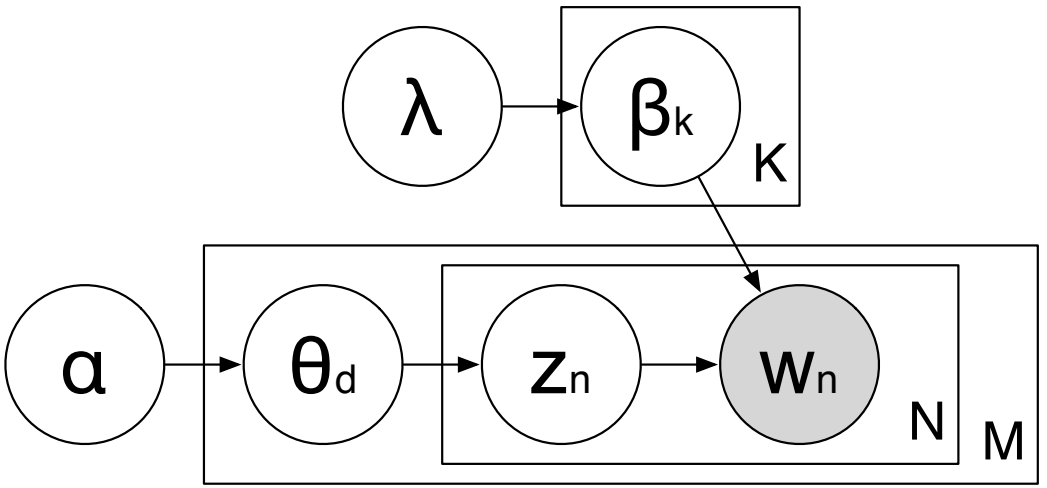
- For each topic $k \in \{1, \ldots, K\}$, draw a multinomial distribution β_k from a Dirichlet distribution with parameter λ
- For each document $d \in \{1, \ldots, M\}$, draw a multinomial distribution θ_d from a Dirichlet distribution with parameter α
- For each word position $n \in \{1, \ldots, N\}$, select a hidden topic z_n from the multinomial distribution parameterized by θ.

Generative Model Approach



- For each topic $k \in \{1, \ldots, K\}$, draw a multinomial distribution β_k from a Dirichlet distribution with parameter λ
- For each document $d \in \{1, \ldots, M\}$, draw a multinomial distribution θ_d from a Dirichlet distribution with parameter α
- For each word position $n \in \{1, \ldots, N\}$, select a hidden topic z_n from the multinomial distribution parameterized by θ.
- Choose the observed word w_n from the distribution β_{z_n}.

Generative Model Approach



- For each topic $k \in \{1, \ldots, K\}$, draw a multinomial distribution β_k from a Dirichlet distribution with parameter λ
- For each document $d \in \{1, \ldots, M\}$, draw a multinomial distribution θ_d from a Dirichlet distribution with parameter α
- For each word position $n \in \{1, \ldots, N\}$, select a hidden topic z_n from the multinomial distribution parameterized by θ.
- Choose the observed word w_n from the distribution β_{z_n}.

We use statistical inference to uncover the most likely unobserved

Topic Models: What's Important

- Topic models
  - Topics to words—multinomial distribution
  - Documents to topics—multinomial distribution
- Focus in this talk: statistical methods
  - Model: story of how your data came to be
  - Latent variables: missing pieces of your story
  - Statistical inference: filling in those missing pieces
- We use latent Dirichlet allocation (LDA) [2], a fully Bayesian version of pLSI [7], probabilistic version of LSA [9]

Topic Models: What's Important

- Topic models (latent variables)
  - Topics to words—multinomial distribution
  - Documents to topics—multinomial distribution
- Focus in this talk: statistical methods
  - Model: story of how your data came to be
  - Latent variables: missing pieces of your story
  - Statistical inference: filling in those missing pieces
- We use latent Dirichlet allocation (LDA) [2], a fully Bayesian version of pLSI [7], probabilistic version of LSA [9]

Outline

Inference

- We are interested in posterior distribution

$$p(Z|X, \Theta) \tag{5}$$

Inference

- We are interested in posterior distribution

$$p(Z|X, \Theta) \tag{5}$$

- 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.

$$p(\mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\theta}|\mathbf{w}, \alpha, \lambda) \tag{6}$$

Inference

- We are interested in posterior distribution

$$p(Z|X, \Theta) \tag{5}$$

- 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.

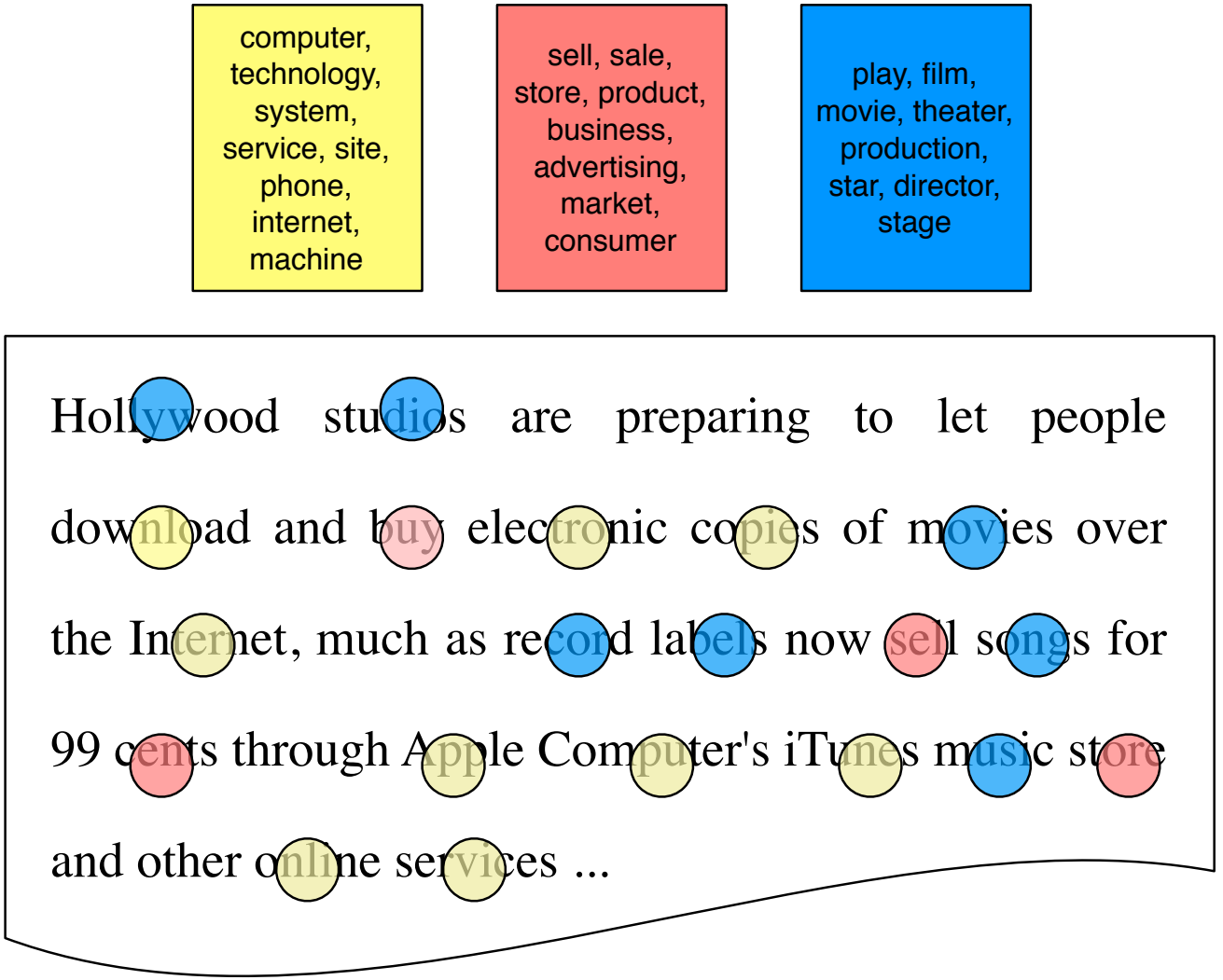
$$p(\mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\theta}|\mathbf{w}, \alpha, \lambda) \tag{6}$$

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

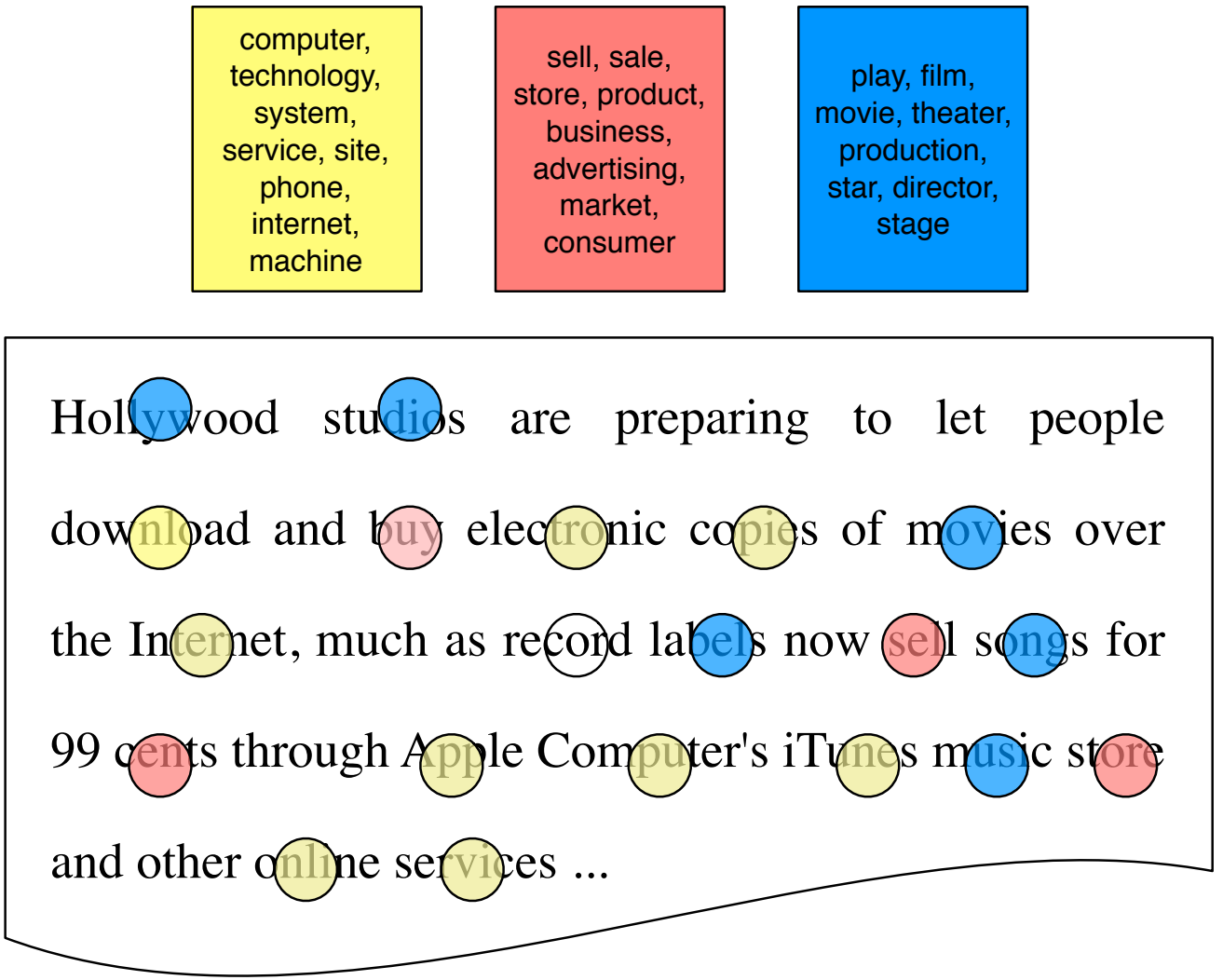
Gibbs Sampling

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

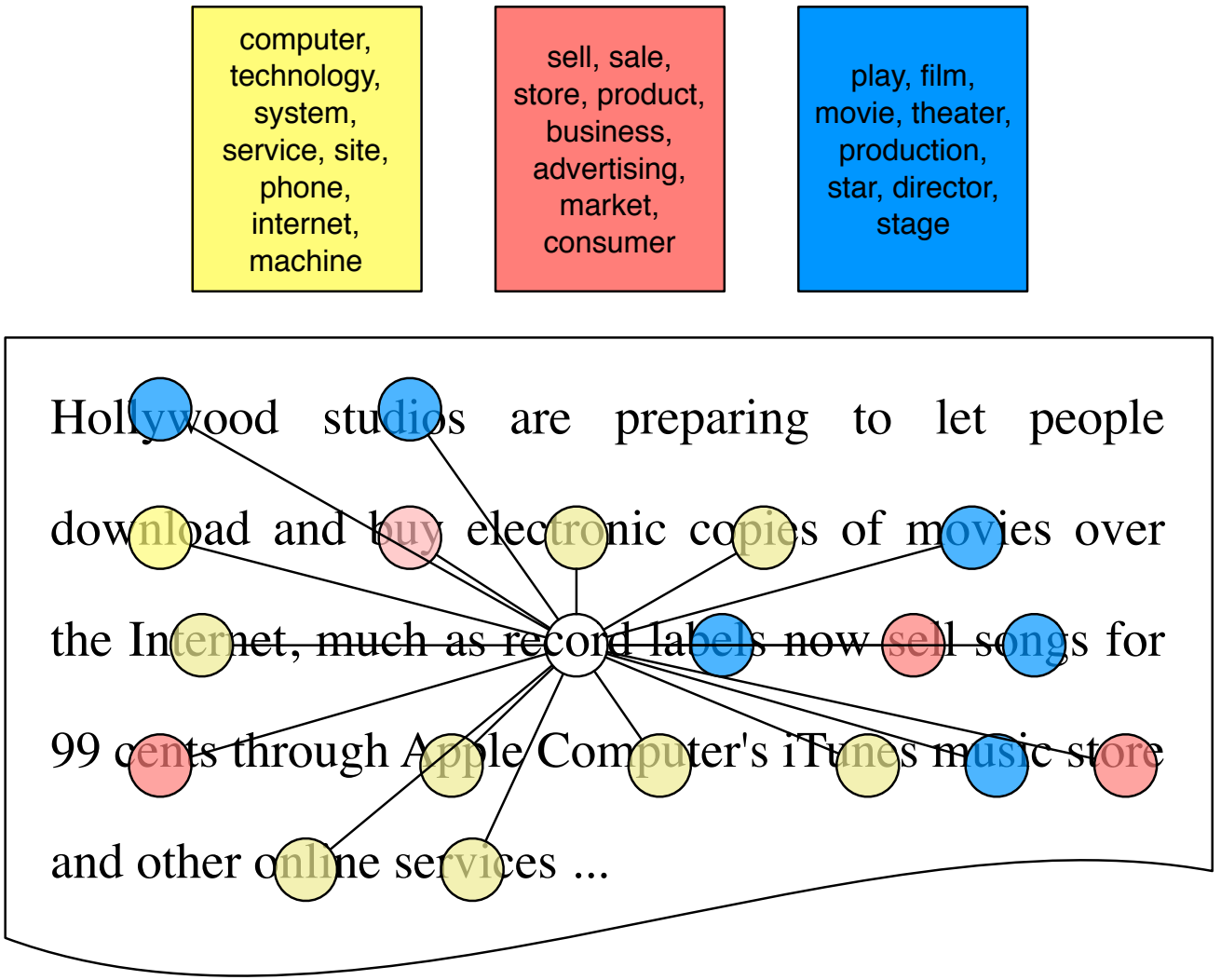
Inference


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 ...

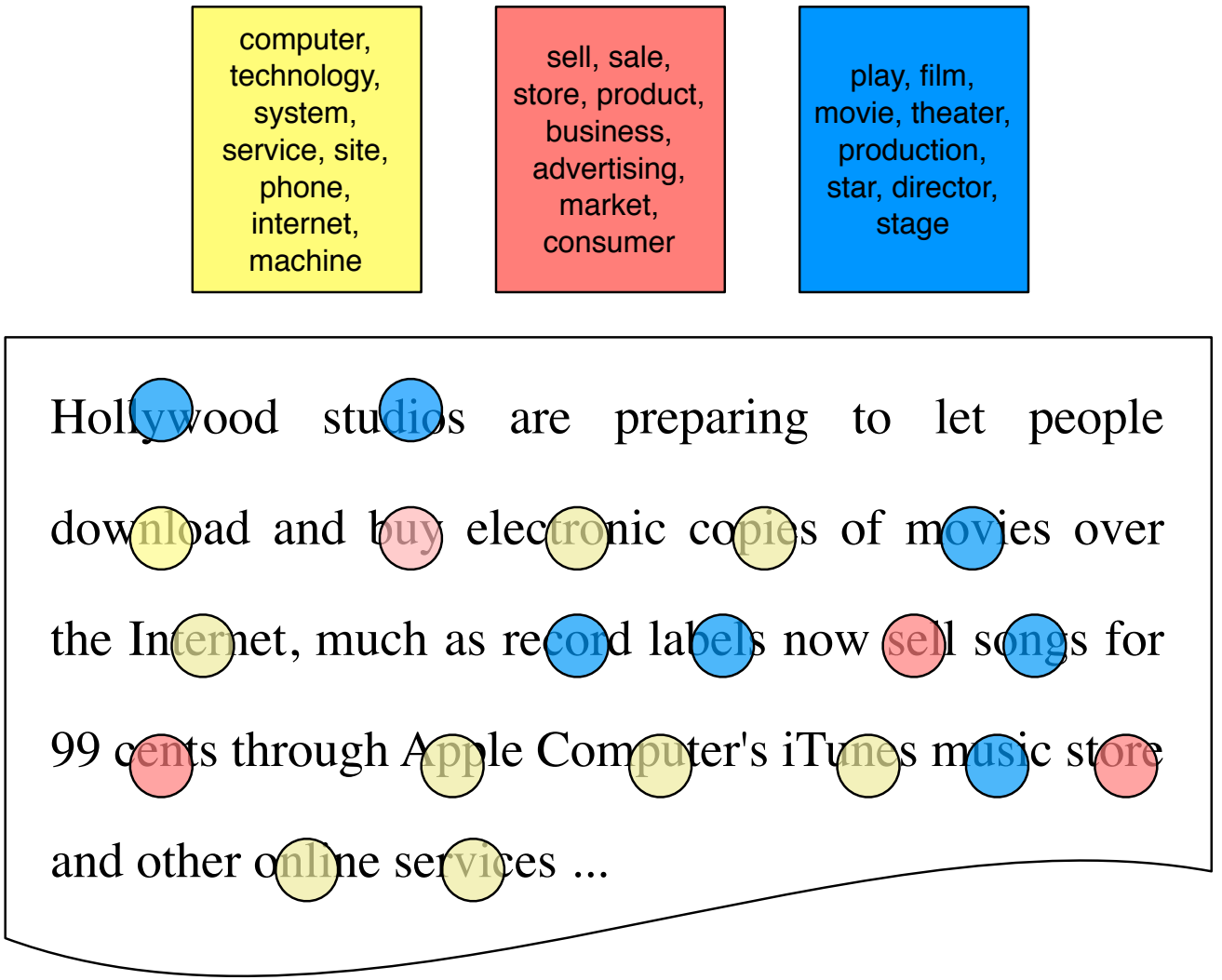
Inference


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 ...

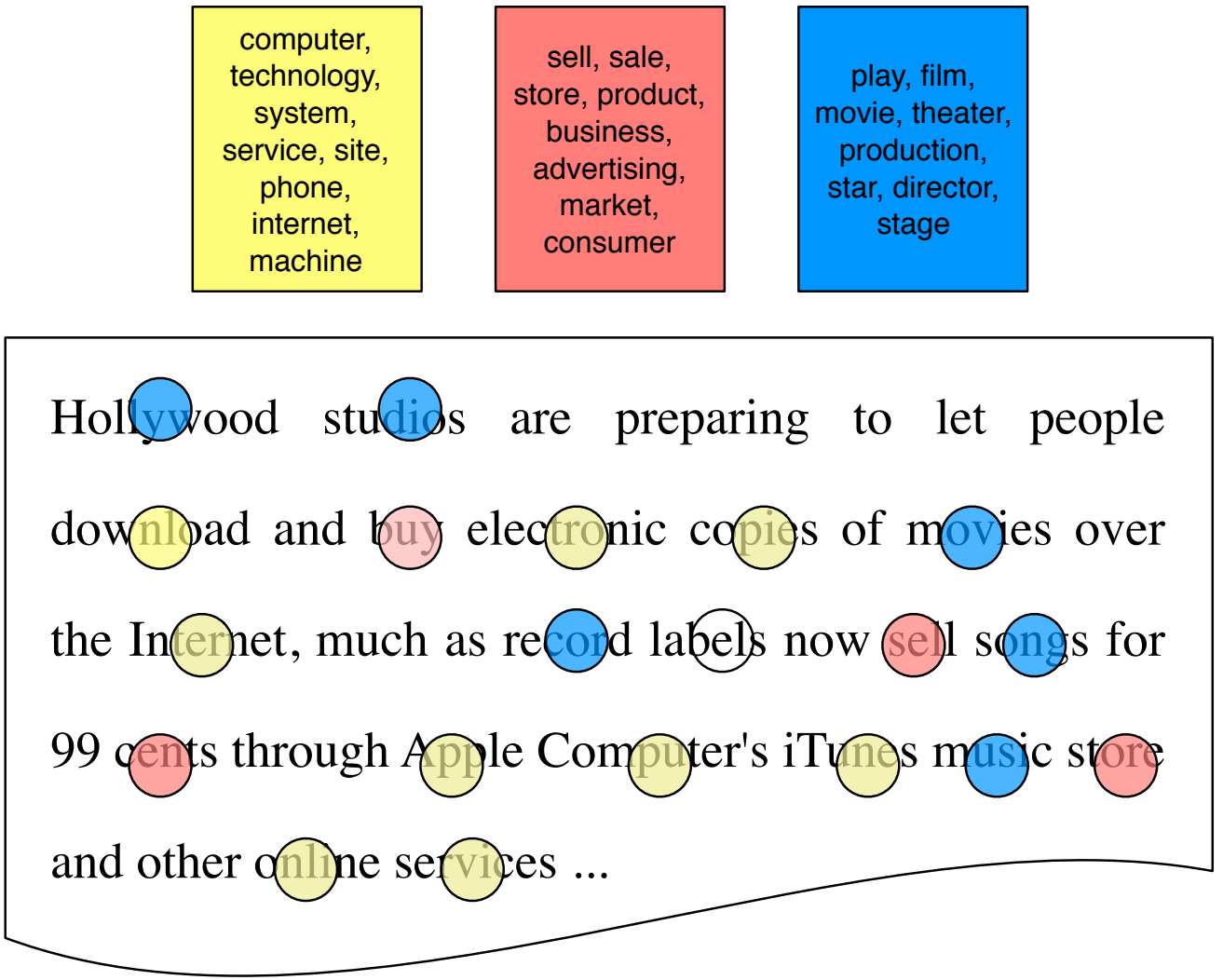
Inference


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 ...

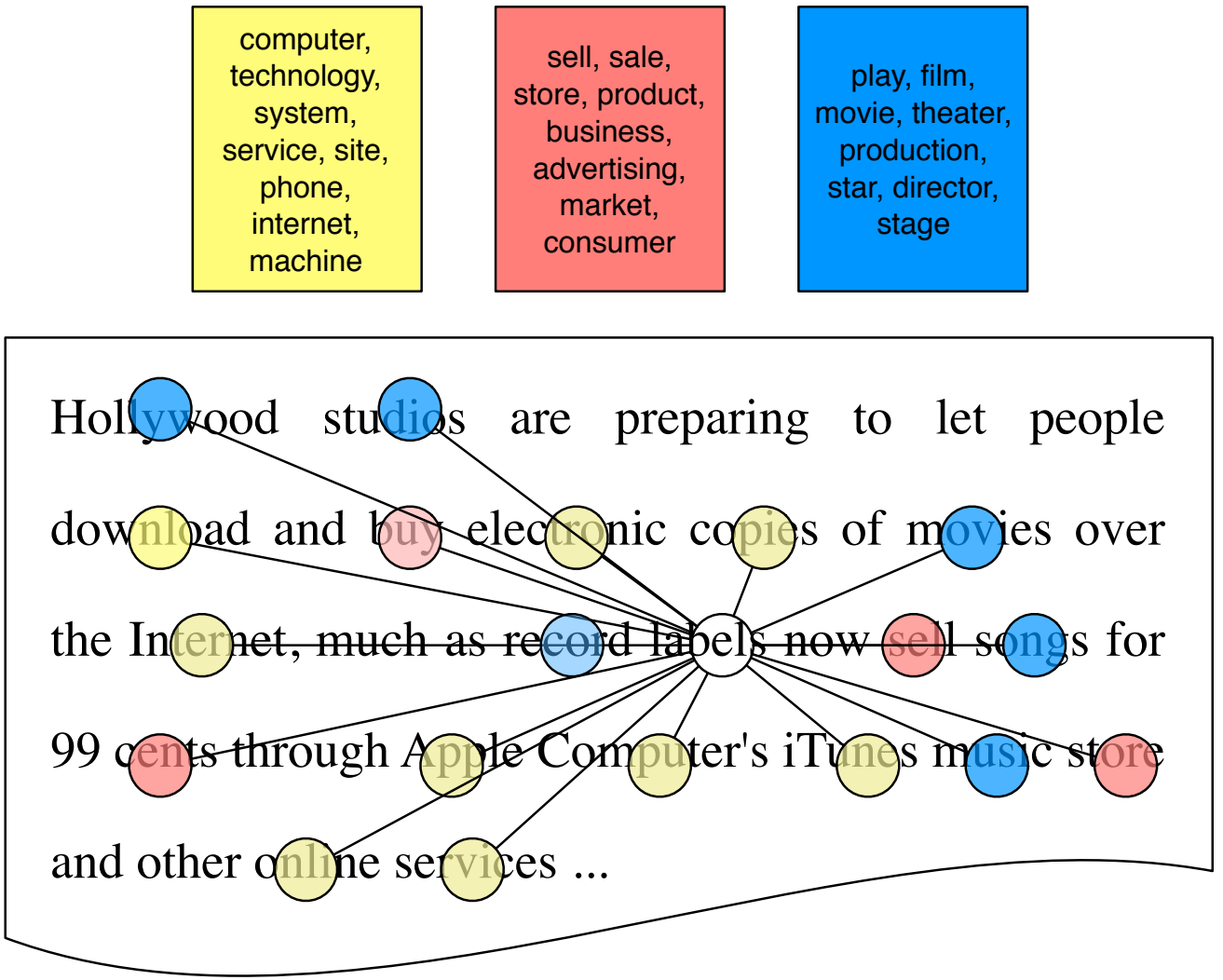
Inference


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 ...

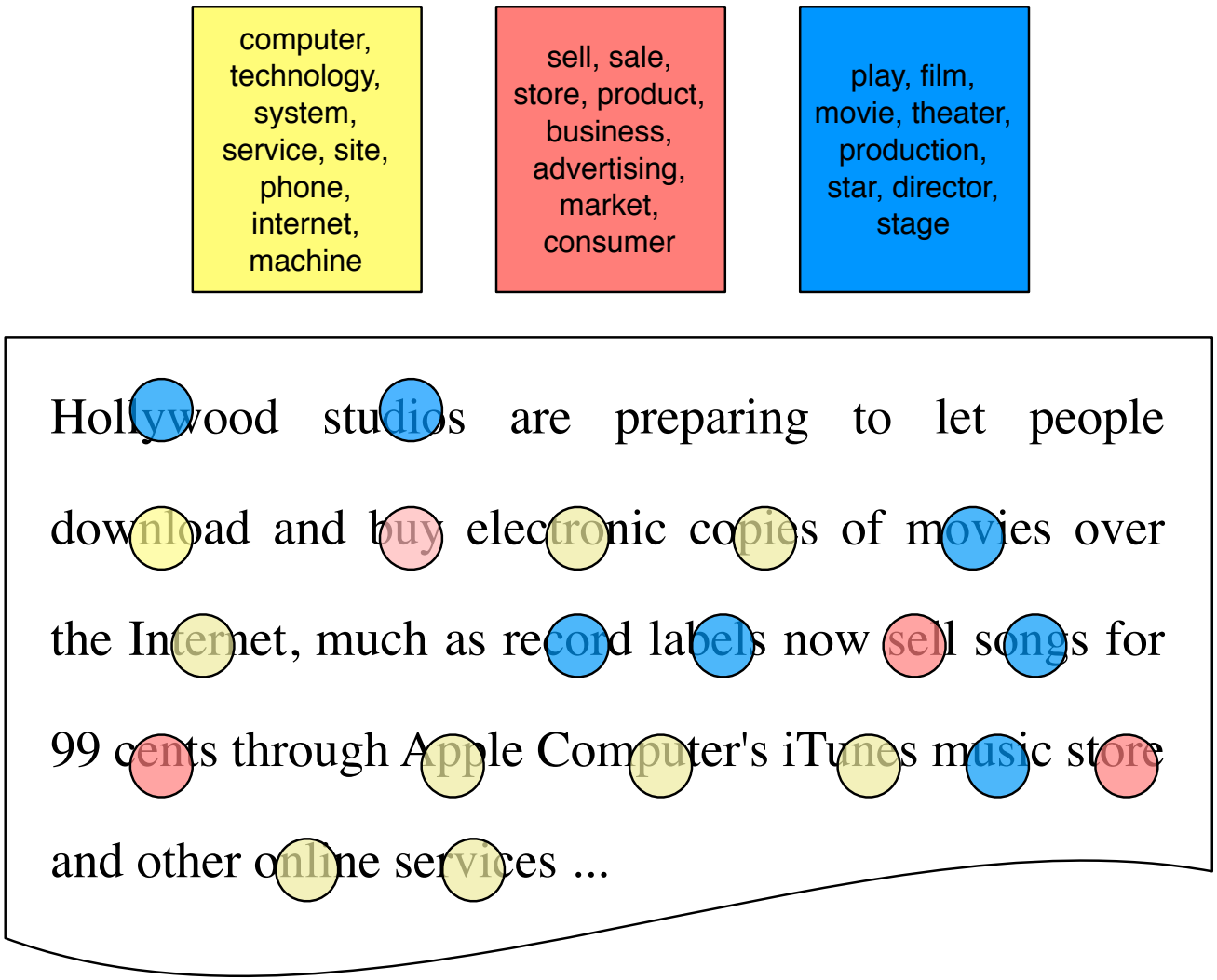
Inference


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:

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

Gibbs Sampling

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

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

- 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.

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

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

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

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

Gamma Function Identity

$$z = \frac{\Gamma(z+1)}{\Gamma(z)} \tag{7}$$

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

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

Gibbs Sampling Equation

$$\frac{n_{d,k} + \alpha_k}{\sum_i^K n_{d,i} + \alpha_i} \frac{v_{k,w_{d,n}} + \lambda_{w_{d,n}}}{\sum_i v_{k,i} + \lambda_i} \tag{8}$$

- 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

$$\frac{n_{d,k} + \alpha_k}{\sum_i^K n_{d,i} + \alpha_i} \frac{v_{k,w_{d,n}} + \lambda_{w_{d,n}}}{\sum_i v_{k,i} + \lambda_i} \tag{8}$$

- 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

$$\frac{n_{d,k} + \alpha_k}{\sum_i^K n_{d,i} + \alpha_i} \frac{v_{k,w_{d,n}} + \lambda_{w_{d,n}}}{\sum_i v_{k,i} + \lambda_i} \tag{8}$$

- 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

$$\frac{n_{d,k} + \alpha_k}{\sum_i^K n_{d,i} + \alpha_i} \frac{v_{k,w_{d,n}} + \lambda_{w_{d,n}}}{\sum_i v_{k,i} + \lambda_i} \tag{8}$$

- 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

$$\frac{n_{d,k} + \alpha_k}{\sum_i^K n_{d,i} + \alpha_i} \frac{v_{k,w_{d,n}} + \lambda_{w_{d,n}}}{\sum_i v_{k,i} + \lambda_i} \tag{8}$$

- 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

$$\frac{n_{d,k} + \alpha_k}{\sum_i^K n_{d,i} + \alpha_i} \frac{v_{k,w_{d,n}} + \lambda_{w_{d,n}}}{\sum_i v_{k,i} + \lambda_i} \tag{8}$$

- 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 $w_{d,n}$

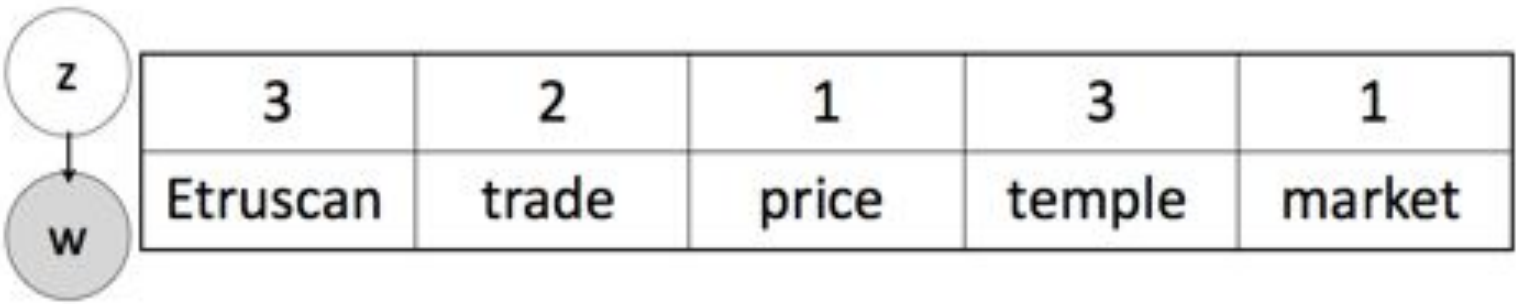
Sample Document

Etruscan	trade	price	temple	market

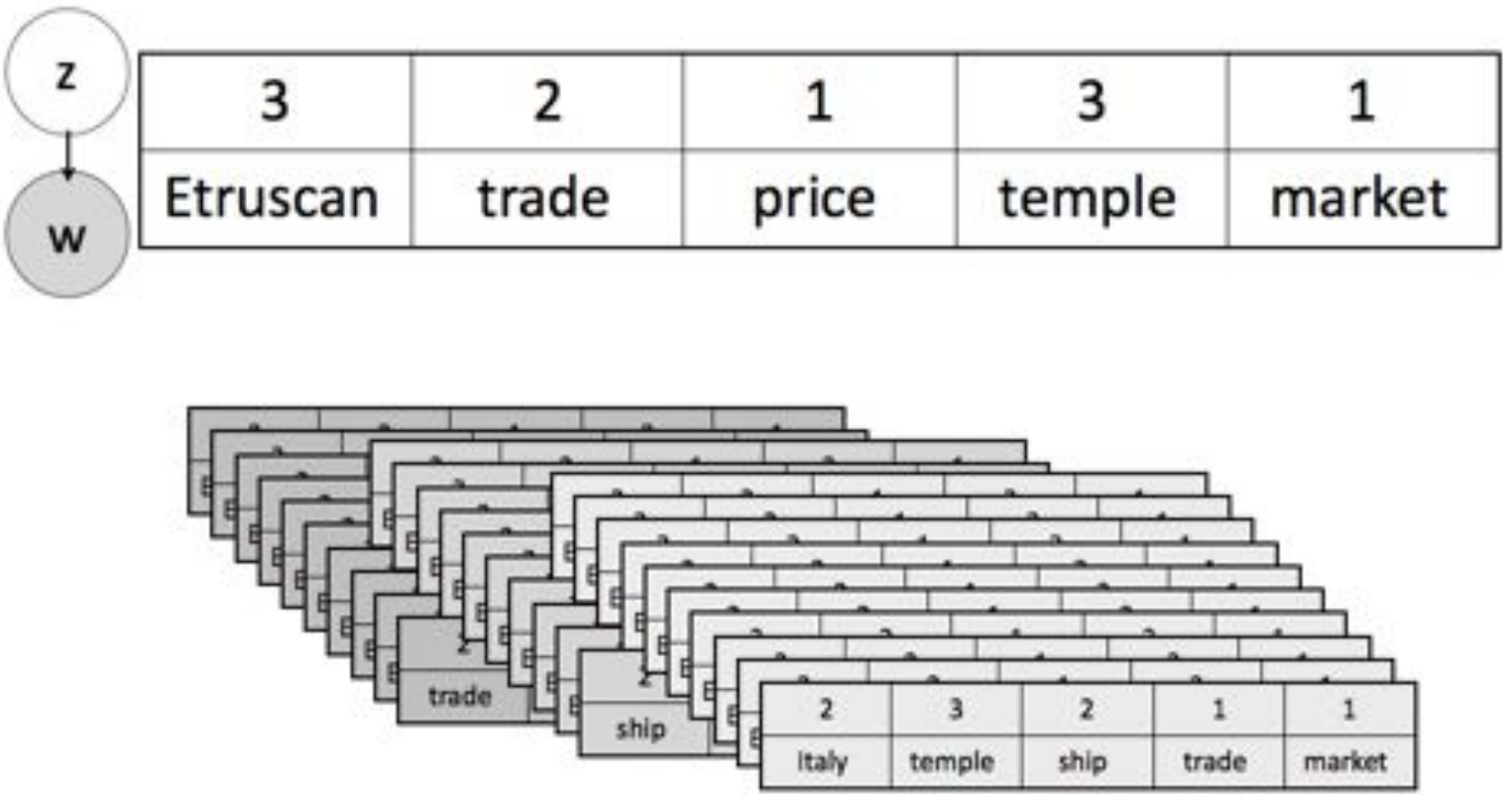
Sample Document

Etruscan	trade	price	temple	market

Randomly Assign Topics



Randomly Assign Topics



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 counts from all

	1	2	3
Etruscan	1	0	35
market	50	0	1
price	42	1	0

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 counts from all

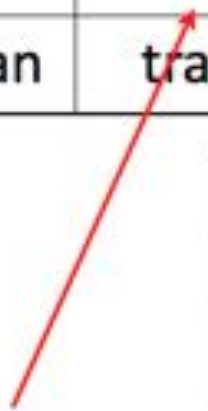
	1	2	3
Etruscan	1	0	35
market	50	0	1
price	42	1	0

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

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?

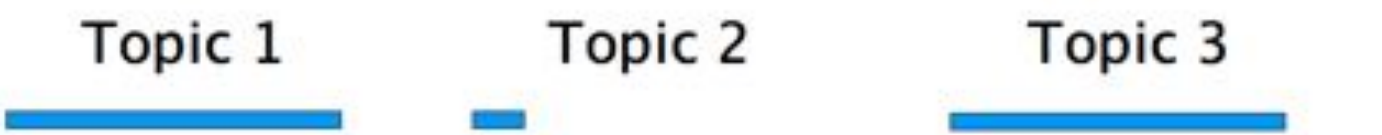
3	?	1	3	1
Etruscan	trade	price	temple	market

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

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

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

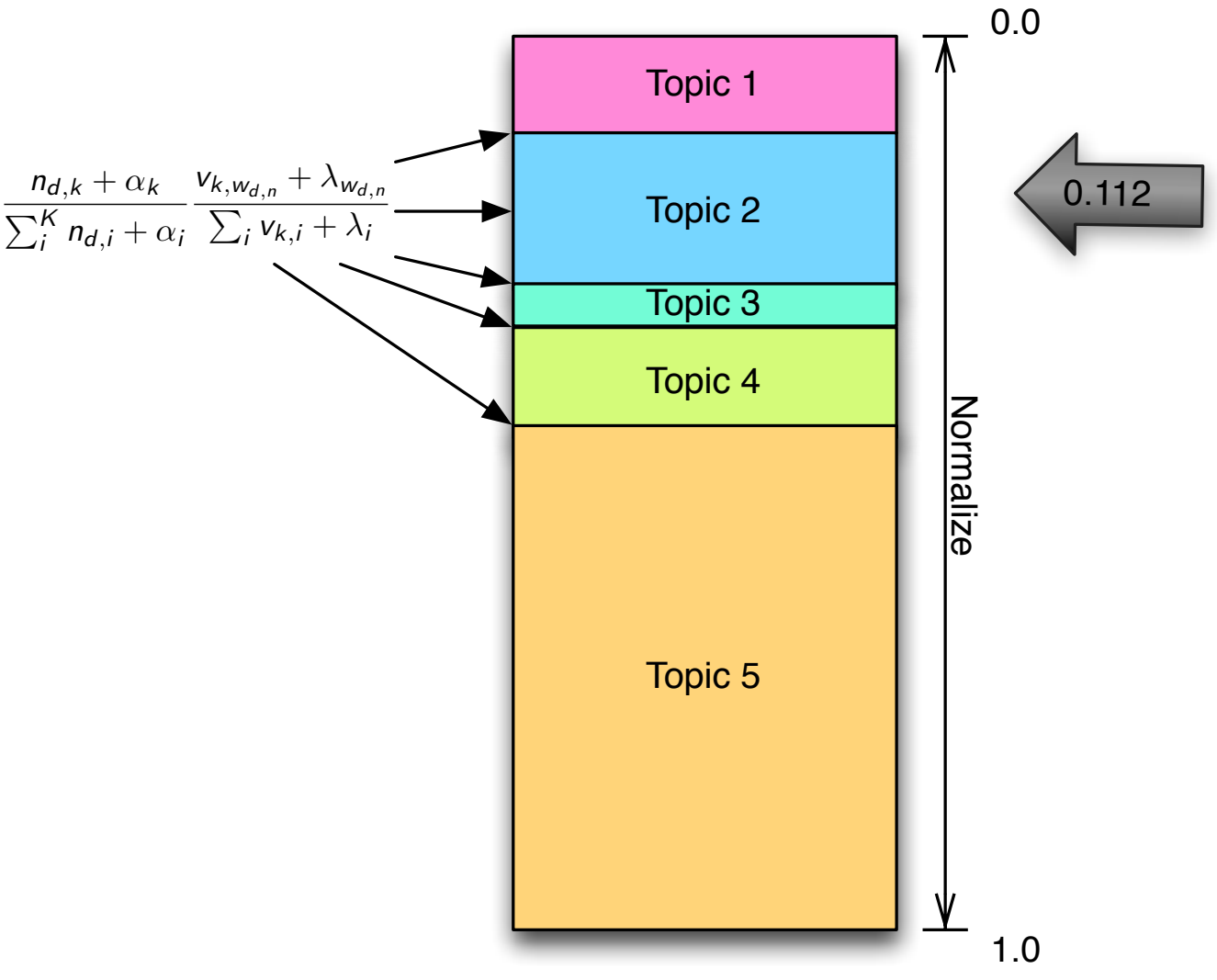
3	1	1	3	1
Etruscan	trade	price	temple	market

	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 $v_{z_{new},w_{d,n}}$

Implementation

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)

Outline

Mallet

- Developed at UMass Amherst by Andrew McCallum and David Mimno (among others)
- Very fast implementation of Gibbs sampling for topic modeling
- (Somewhat) friendly interface
- Easiest on UNIX-derived operating systems, but also works on Windows
- Requires Java

Download Location

http://mallet.cs.umass.edu/

Scenario

- Learn a (small-sized) topic model on NSF data
- Apply those topics to NRC data
- Discover the priorities of NSF
- Connects NRC grants to that model
- Walk through the commands to do everything

Getting Your Data

- Text file separated by columns
- First column: doc id
- Second column: language
- Third column: text

doc lang text

Download

`http://umiacs.umd.edu/~jbg/lda_demo/`

Getting Your Data

- Text file separated by columns
- First column: doc id
- Second column: language
- Third column: text

doc lang text

Download

`http://umiacs.umd.edu/~jbg/lda_demo/`

Getting Your Data

- Text file separated by columns
- First column: doc id
- Second column: language
- Third column: text

doc lang text

Download

`http://umiacs.umd.edu/~jbg/lda_demo/`

Getting Your Data

- Text file separated by columns
- First column: doc id
- Second column: language
- Third column: text

doc lang text

Download

`http://umiacs.umd.edu/~jbg/lda_demo/`

Getting Your Data

- Text file separated by columns
- First column: doc id
- Second column: language
- Third column: text

doc lang text

Download

`http://umiacs.umd.edu/~jbg/lda_demo/`

```
doc0    none    It is proposed to grow and characterize ternary alloys of ...
doc1    none    This project will focus on development of new cutting tool designs ...
doc2    none    The purpose of the proposed work is to design a novel cooling device ...
doc3    none    The objective of this research project is to develop a wireless ...
```

Preparing NSF Data

Mallet Command

mallet import-file –remove-stopwords –keep-sequence –input nsf-30k.txt –output nsf.mallet

Preparing NSF Data

Mallet Command

mallet import-file –remove-stopwords –keep-sequence –input nsf-30k.txt –output nsf.mallet

- Tell Mallet what to do
- Remove words like “the”, “and”, “of” (otherwise, they’d be at the top of every topic)
- Remember the order of words (required for Gibbs sampling)
- The input text file
- Where it writes the binary file

Preparing NSF Data

Mallet Command

mallet import-file –remove-stopwords –keep-sequence –input nsf-30k.txt –output nsf.mallet

- Tell Mallet what to do
- Remove words like “the”, “and”, “of” (otherwise, they’d be at the top of every topic)
- Remember the order of words (required for Gibbs sampling)
- The input text file
- Where it writes the binary file

Preparing NSF Data

Mallet Command

mallet import-file –remove-stopwords –keep-sequence –input nsf-30k.txt –output nsf.mallet

- Tell Mallet what to do
- Remove words like “the”, “and”, “of” (otherwise, they’d be at the top of every topic)
- Remember the order of words (required for Gibbs sampling)
- The input text file
- Where it writes the binary file

Preparing NSF Data

Mallet Command

mallet import-file –remove-stopwords –keep-sequence –input nsf-30k.txt –output nsf.mallet

- Tell Mallet what to do
- Remove words like “the”, “and”, “of” (otherwise, they’d be at the top of every topic)
- Remember the order of words (required for Gibbs sampling)
- The input text file
- Where it writes the binary file

Preparing NSF Data

Mallet Command

mallet import-file –remove-stopwords –keep-sequence –input nsf-30k.txt –output nsf.mallet

- Tell Mallet what to do
- Remove words like “the”, “and”, “of” (otherwise, they’d be at the top of every topic)
- Remember the order of words (required for Gibbs sampling)
- The input text file
- Where it writes the binary file

Preparing NRC data

Mallet Command

mallet import-file –remove-stopwords –keep-sequence –input merged.txt –output nrc.mallet –use-pipe-from nsf.mallet

- Nearly identical to previous command

Preparing NRC data

Mallet Command

mallet import-file –remove-stopwords –keep-sequence –input merged.txt –output nrc.mallet –use-pipe-from nsf.mallet

- Nearly identical to previous command
- Main difference: use NSF vocabulary to encode the data
- LDA doesn't know what words mean
- Internally, these algorithms map words to numbers: oxygen (2134), neutrino (33), Weber (1701)
- This ensures that this mapping is consistent between datasets

Fitting a topic model

Mallet Command

mallet train-topics –input nsf.mallet –num-topics 10 –num-iterations 100 –output-model nsf_10.model –output-state nsf_10.state.gz –output-doc-topics nsf_10.doc –output-topic-keys nsf_10.topics –inferencer-filename nsf_10.inf

Fitting a topic model

Mallet Command

mallet train-topics –input nsf.mallet –num-topics 10 –num-iterations 100 –output-model nsf_10.model –output-state nsf_10.state.gz –output-doc-topics nsf_10.doc –output-topic-keys nsf_10.topics –inferencer-filename nsf_10.inf

- The documents we learn topics from
- The number of topics we'll learn
- The number of sweeps through data
- Save the resulting model
- Save Gibbs sampling states
- Save document-topic associations
- Save word-topic associations
- Save inferencer

Fitting a topic model

Mallet Command

```
mallet train-topics –input nsf.mallet –num-topics 10 –num-iterations 100
–output-model nsf_10.model –output-state nsf_10.state.gz
–output-doc-topics nsf_10.doc –output-topic-keys nsf_10.topics
–inferencer-filename nsf_10.inf
```

- The documents we learn topics from
- The number of topics we'll learn
- The number of sweeps through data
- Save the resulting model
- Save Gibbs sampling states
- Save document-topic associations
- Save word-topic associations
- Save inferencer

Fitting a topic model

Mallet Command

```
mallet train-topics –input nsf.mallet –num-topics 10 –num-iterations 100
–output-model nsf_10.model –output-state nsf_10.state.gz
–output-doc-topics nsf_10.doc –output-topic-keys nsf_10.topics
–inferencer-filename nsf_10.inf
```

- The documents we learn topics from
- The number of topics we'll learn
- The number of sweeps through data
- Save the resulting model
- Save Gibbs sampling states
- Save document-topic associations
- Save word-topic associations
- Save inferencer

Fitting a topic model

Mallet Command

mallet train-topics –input nsf.mallet –num-topics 10 –num-iterations 100 –output-model nsf_10.model –output-state nsf_10.state.gz –output-doc-topics nsf_10.doc –output-topic-keys nsf_10.topics –inferencer-filename nsf_10.inf

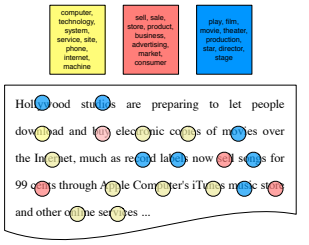
- The documents we learn topics from
- The number of topics we'll learn
- The number of sweeps through data
- Save the resulting model
- Save Gibbs sampling states
- Save document-topic associations
- Save word-topic associations
- Save inferencer

Fitting a topic model

Mallet Command

mallet train-topics –input nsf.mallet –num-topics 10 –num-iterations 100 –output-model nsf_10.model –output-state nsf_10.state.gz –output-doc-topics nsf_10.doc –output-topic-keys nsf_10.topics –inferencer-filename nsf_10.inf

- The documents we learn topics from
- The number of topics we'll learn
- The number of sweeps through data
- Save the resulting model
- Save Gibbs sampling states
- Save document-topic associations
- Save word-topic associations
- Save inferencer

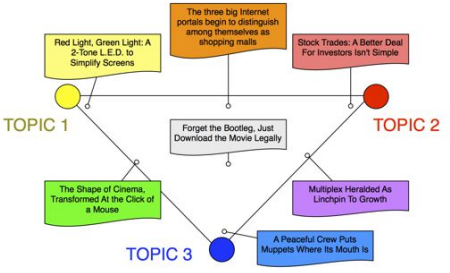


Fitting a topic model

Mallet Command

mallet train-topics –input nsf.mallet –num-topics 10 –num-iterations 100 –output-model nsf_10.model –output-state nsf_10.state.gz –output-doc-topics nsf_10.doc –output-topic-keys nsf_10.topics –inferencer-filename nsf_10.inf

- The documents we learn topics from
- The number of topics we'll learn
- The number of sweeps through data
- Save the resulting model
- Save Gibbs sampling states
- Save document-topic associations
- Save word-topic associations
- Save inferencer



Fitting a topic model

Mallet Command

mallet train-topics –input nsf.mallet –num-topics 10 –num-iterations 100 –output-model nsf_10.model –output-state nsf_10.state.gz –output-doc-topics nsf_10.doc –output-topic-keys nsf_10.topics –inferencer-filename nsf_10.inf

- The documents we learn topics from
- The number of topics we'll learn
- The number of sweeps through data
- Save the resulting model
- Save Gibbs sampling states
- Save document-topic associations
- Save word-topic associations
- Save inferencer

Fitting a topic model

Mallet Command

mallet train-topics –input nsf.mallet –num-topics 10 –num-iterations 100 –output-model nsf_10.model –output-state nsf_10.state.gz –output-doc-topics nsf_10.doc –output-topic-keys nsf_10.topics –inferencer-filename nsf_10.inf

- The documents we learn topics from
- The number of topics we'll learn
- The number of sweeps through data
- Save the resulting model
- Save Gibbs sampling states
- Save document-topic associations
- Save word-topic associations
- Save inferencer

Inferencer

Allows us to apply these topics to another dataset

As Mallet Runs . . .

```
<10> LL/token:  -10,01271
<20> LL/token:  -9,17157
<30> LL/token:  -9,01933
<40> LL/token:  -8,96692
```

- we want to discover best collection of topics that describes our data
- this is defined in terms of probability
- stochastic search
  - stop when probability levels off
  - requires thousands of iterations

As Mallet Runs . . .

```
0 5 species environmental water natural understanding study research processes climate c
1 5 materials research chemistry properties chemical high magnetic optical surface elect
2 5 systems system design research data control performance network based time applicati
```

- ID of the topic
- Weight of the topic (start the same)
- The most probable words in the topic

Word-Topic Association

```
0       5       species environmental water natural understanding study research process
1       5       materials research chemistry properties chemical high magnetic optical s
```

- Same information as displayed as Mallet runs
- Shows the most probable words in each topic

What topics did we discover?

0. **Computer Science**: data performance network software algorithms
1. **Students**: students education undergraduate school learning faculty
2. **Chemistry**: chemistry chemical organic molecular reactions metal
3. **Physics**: high observations waves energy time physics stars flow
4. **Math**: equations models geometry analysis number mathematics
5. **Social Science**: social economic information policy human political
6. **Earth Science**: water climate ice ocean carbon high models soil sea
7. **Training**: research support award workshop conference program
8. **Biology**: species cell genes plant gene molecular cells dna proteins
9. **Materials Science**: materials research high properties project optical magnetic

What documents are associated with each topic?

```
8724     doc8724 8        0.5295448158189615       0        0.46418192697602284      7
        0.0011271253410239611    ...
8725     doc8725 9        0.8864657651849538       7        0.05057323989491986      2
        0.0453216052686360 66      ...
```

- Document 8724 is associated with topic 8 and topic 0
- These are the Biology and Computer Science topics

What documents are associated with each topic?

doc8724

Humans and other animals use visual looming of a stimulus to detect change in distance of a stimulus in the depth of the visual field. It is unclear how such visual cues drive neural signals that guide appropriate behavioral responses such as approach or avoidance for such a stimulus. Results will be important beyond insect vision, for understanding depth detection and obstacle avoidance by visual mechanisms in general, and for developing useful machine vision and guidance systems in robotics for computational neuroscience.

Applying topics to NRC data

- Assume that we are satisfied with this topic analysis
- NSF is convenient example
  - EU-wide research
  - Wikipedia
  - Publications
- Associate new documents with some standard
- Compare funding levels for comparable topics (regardless of internal classification)

Applying topics to NRC data

- Assume that we are satisfied with this topic analysis
- NSF is convenient example
  - EU-wide research
  - Wikipedia
  - Publications
- Associate new documents with some standard
- Compare funding levels for comparable topics (regardless of internal classification)

Applying topics to NRC data

- Assume that we are satisfied with this topic analysis
- NSF is convenient example
  - EU-wide research
  - Wikipedia
  - Publications
- Associate new documents with some standard
- Compare funding levels for comparable topics (regardless of internal classification)

Applying Topics

Mallet Command

```
mallet infer-topics –input nrc.mallet –inferencer nsf_10.inf
–output-doc-topics nrc.doc
```

- Document to apply model to
- Inferencer created from model
- Output file

Applying Topics

Mallet Command

mallet infer-topics –input nrc.mallet –inferencer nsf_10.inf –output-doc-topics nrc.doc

- Document to apply model to
- Inferencer created from model
- Output file

Applying Topics

Mallet Command

mallet infer-topics –input nrc.mallet –inferencer nsf_10.inf –output-doc-topics nrc.doc

- Document to apply model to
- Inferencer created from model
- Output file

Applying Topics

Mallet Command

mallet infer-topics –input nrc.mallet –inferencer nsf_10.inf –output-doc-topics nrc.doc

- Document to apply model to
- Inferencer created from model
- Output file

Finding Similar Articles

- We found a computational biology grant from NSF
- How do we find similar grants in NRC?
- Look for grants with high probability for Topic 0 and Topic 8

```
187     pages/1253988731023     8       0.4546086322042478      0       0.3405459548
4236727     5       0.13293471300185633     ...
186     pages/1253988731014     8       0.41662982704829793     9       0.1797146162
3145737     ...
0981687477E-4
148     pages/1253984734617     8       0.502404238103382       0       0.4069265225
375647      6       0.050378383174511626    ...
```

Finding Similar Articles

- We found a computational biology grant from NSF
- How do we find similar grants in NRC?

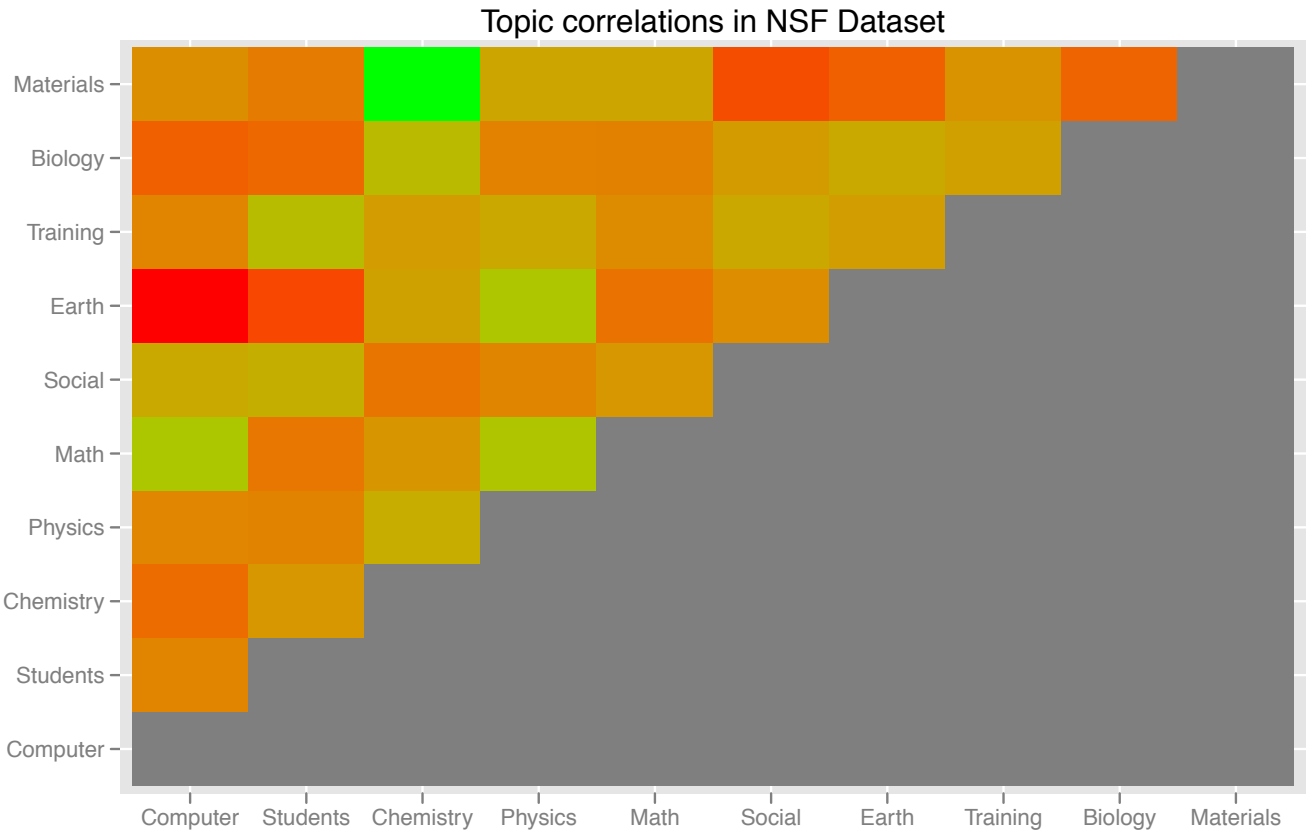
Look it up . . .

548

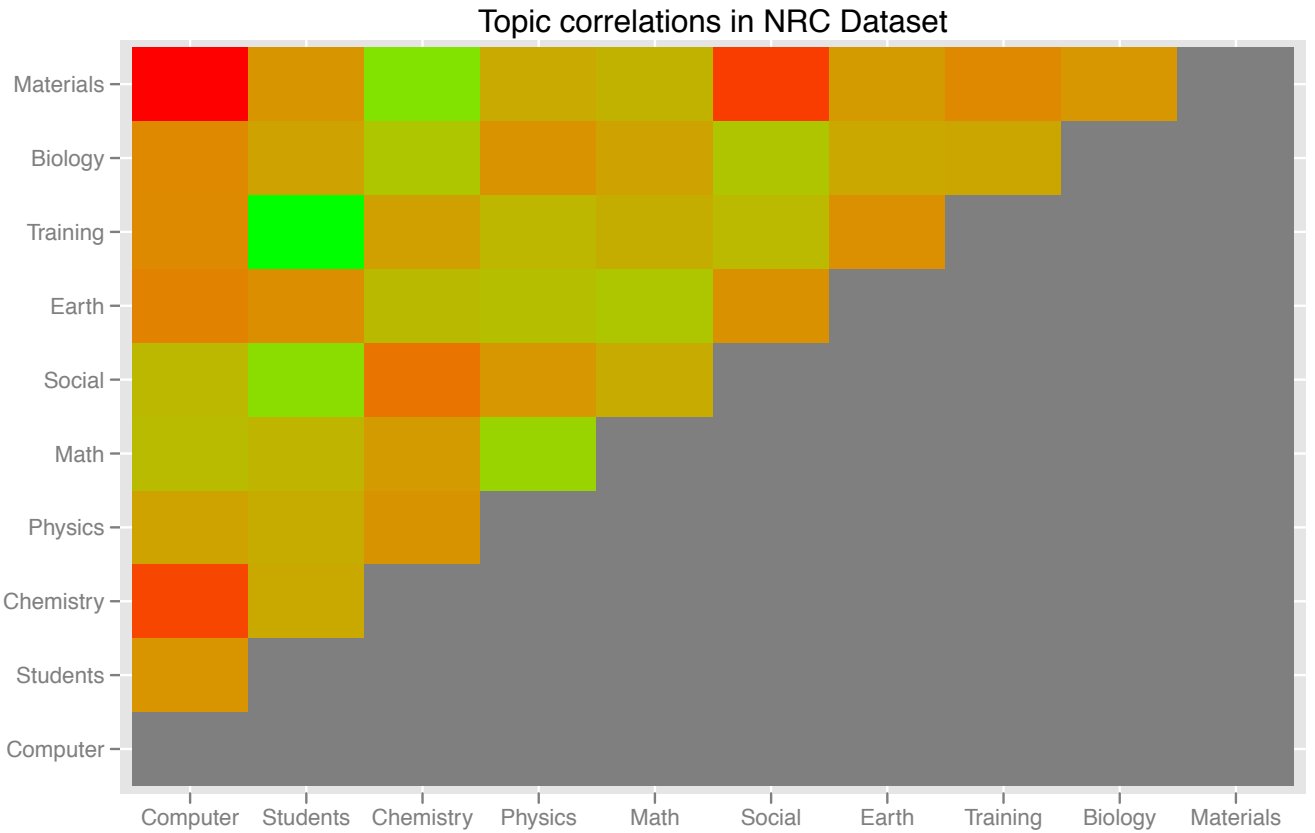
162

225

Which topics appear together?



Which topics appear together?



Much more to be done!

- Bigrams: not all strings separated by spaces are words
  - high energy, nano materials, seismic models, undergraduate students
  - need to be discovered along with topics
- Choosing correct granularity
- Refining stopword list: investigator, study, research
- Adding constraints
- Multiple languages

Much more to be done!

- Bigrams: not all strings separated by spaces are words
  - high energy, nano materials, seismic models, undergraduate students
  - need to be discovered along with topics
- Choosing correct granularity
- Refining stopword list: investigator, study, research
- Adding constraints
- Multiple languages

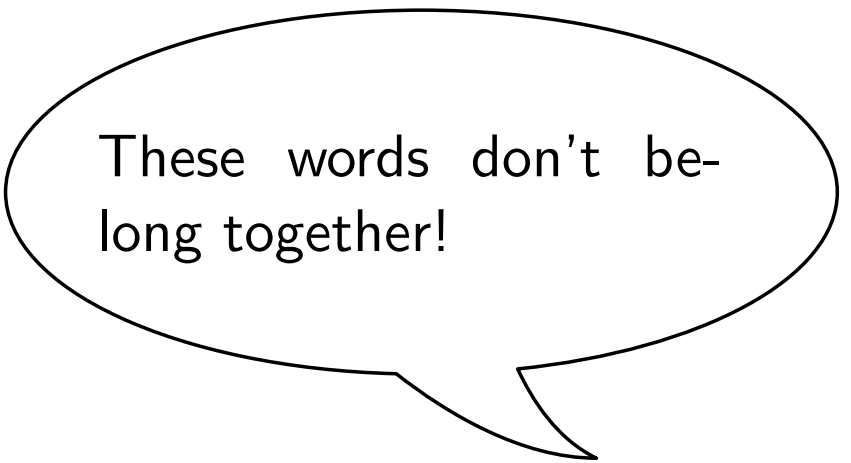
Outline

The Problem: User Perspective

bladder
spinal_cord
sci
spinal_cord_injury
spinal
urinary
urothelial
cervical
injury
recovery
urinary_tract
locomotor
lumbar

The Problem: User Perspective

bladder
spinal_cord
sci
spinal_cord_injury
spinal
urinary
urothelial
cervical
injury
recovery
urinary_tract
locomotor
lumbar



The Problem: User Perspective

bladder

spinal_cord

sci

spinal_cord_injury

spinal

urinary

urothelial

cervical

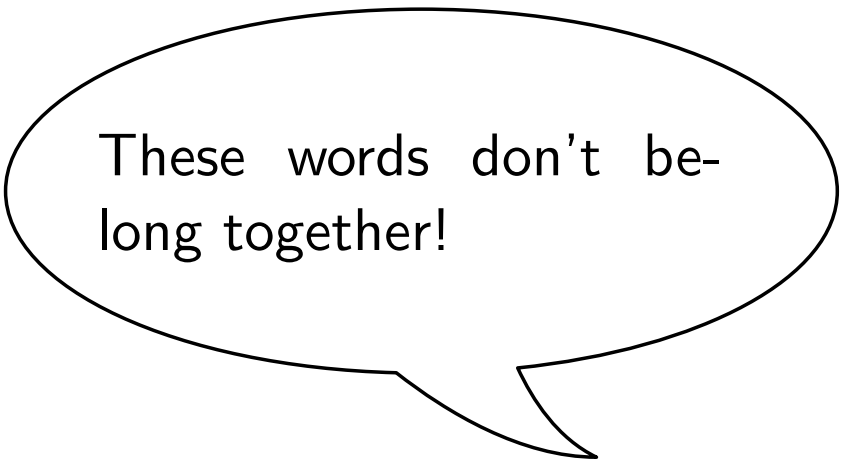
injury

recovery

urinary_tract

locomotor

lumbar



This is serious business!

- Decision makers see problems
- No easy way to correct the problem
- Result: entire approach is abandoned

This is serious business!

- Decision makers see problems
- No easy way to correct the problem
- Result: entire approach is abandoned
- Offer support for two kinds of feedback [3, 1]
  - Positive correlations: words that **should** appear together
  - Negative correlations: words that **should not** appear together

How to incorporate feedback?

TOPIC 1

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

TOPIC 2

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

TOPIC 3

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

1. Fit initial topic model

How to incorporate feedback?

TOPIC 1

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

TOPIC 2

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

TOPIC 3

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

1. Fit initial topic model
2. Get feedback from user

How to incorporate feedback?

TOPIC 1

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

TOPIC 2

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

TOPIC 3

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

1. Fit initial topic model
2. Get feedback from user
3. Incrementally relearn model
   - Forget your mistakes
   - Replace the model with a correlated one
   - Continue inference

How to incorporate feedback?

TOPIC 1

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

TOPIC 2

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

TOPIC 3

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

1. Fit initial topic model
2. Get feedback from user
3. Incrementally relearn model
   - Forget your mistakes
   - Replace the model with a correlated one
   - Continue inference

Keep computation fast and consistent [8]

Topic	Before
1	election, yeltsin, russian, political, party, democratic, russia, president, democracy, boris, country, south, years, month, government, vote, since, leader, presidential, military
2	new, york, city, state, mayor, budget, giuliani, council, cuomo, gov, plan, year, rudolph, dinkins, lead, need, governor, legislature, pataki, david
3	nuclear, arms, weapon, defense, treaty, missile, world, unite, yet, soviet, lead, secretary, would, control, korea, intelligence, test, nation, country, testing
4	president, bush, administration, clinton, american, force, reagan, war, unite, lead, economic, iraq, congress, america, iraqi, policy, aid, international, military, see
⋮	⋮
20	soviet, lead, gorbachev, union, west, mikhail, reform, change, europe, leaders, poland, communist, know, old, right, human, washington, western, bring, party

Topic	Before
1	election, yeltsin, russian, political, party, democratic, russia, president, democracy, boris, country, south, years, month, government, vote, since, leader, presidential, military
2	new, york, city, state, mayor, budget, giuliani, council, cuomo, gov, plan, year, rudolph, dinkins, lead, need, governor, legislature, pataki, david
3	nuclear, arms, weapon, defense, treaty, missile, world, unite, yet, soviet, lead, secretary, would, control, korea, intelligence, test, nation, country, testing
4	president, bush, administration, clinton, american, force, reagan, war, unite, lead, economic, iraq, congress, america, iraqi, policy, aid, international, military, see
⋮	⋮
20	soviet, lead, gorbachev, union, west, mikhail, reform, change, europe, leaders, poland, communist, know, old, right, human, washington, western, bring, party

Topic	Before
1	election, yeltsin, russian, political, party, democratic, russia, president, democracy, boris, country, south, years, month, government, vote, since, leader, presidential, military
2	new, york, city, state, mayor, budget, giuliani, council, cuomo, gov, plan, year, rudolph, dinkins, lead, need, governor, legislature, pataki, david
3	nuclear, arms, weapon, defense, treaty, missile, world, unite, yet, soviet, lead, secretary, would, control, korea, intelligence, test, nation, country, testing
4	president, bush, administration, clinton, american, force, reagan, war, unite, lead, economic, iraq, congress, america, iraqi, policy, aid, international, military, see
⋮	⋮
20	soviet, lead, gorbachev, union, west, mikhail, reform, change, europe, leaders, poland, communist, know, old, right, human, washington, western, bring, party

Suggestion

boris, communist, gorbachev, mikhail, russia, russian, soviet, union, yeltsin

Topic	Before	Topic	After
1	election, yeltsin, russian, political, party, democratic, russia, president, democracy, boris, country, south, years, month, government, vote, since, leader, presidential, military	**1**	election, democratic, south, country, president, party, africa, lead, even, democracy, leader, presidential, week, politics, minister, percent, voter, last, month, years
2	new, york, city, state, mayor, budget, giuliani, council, cuomo, gov, plan, year, rudolph, dinkins, lead, need, governor, legislature, pataki, david	2	new, york, city, state, mayor, budget, council, giuliani, gov, cuomo, year, rudolph, dinkins, legislature, plan, david, governor, pataki, need, cut
3	nuclear, arms, weapon, defense, treaty, missile, world, unite, yet, soviet, lead, secretary, would, control, korea, intelligence, test, nation, country, testing	3	nuclear, arms, weapon, treaty, defense, war, missile, may, come, test, american, world, would, need, lead, get, join, yet, clinton, nation
4	president, bush, administration, clinton, american, force, reagan, war, unite, lead, economic, iraq, congress, america, iraqi, policy, aid, international, military, see	4	president, administration, bush, clinton, war, unite, force, reagan, american, america, make, nation, military, iraq, iraqi, troops, international, country, yesterday, plan
⋮		⋮	
20	soviet, lead, gorbachev, union, west, mikhail, reform, change, europe, leaders, poland, communist, know, old, right, human, washington, western, bring, party	**20**	soviet, union, economic, reform, yeltsin, russian, lead, russia, gorbachev, leaders, west, president, boris, moscow, europe, poland, mikhail, communist, power, relations

Topic	Before	Topic	After
1	election, yeltsin, russian, political, party, democratic, russia, president, democracy, boris, country, south, years, month, government, vote, since, leader, presidential, military	**1**	election, democratic, south, country, president, party, africa, lead, even, democracy, leader, presidential, week, politics, minister, percent, voter, last, month, years
2	new, york, city, state, mayor, budget, giuliani, council, cuomo, gov, plan, year, rudolph, dinkins, lead, need, governor, legislature, pataki, david	2	new, york, city, state, mayor, budget, council, giuliani, gov, cuomo, year, rudolph, dinkins, legislature, plan, david, governor, pataki, need, cut
3	nuclear, arms, weapon, defense, treaty, missile, world, unite, yet, soviet, lead, secretary, would, control, korea, intelligence, test, nation, country, testing	3	nuclear, arms, weapon, treaty, defense, war, missile, may, come, test, american, world, would, need, lead, get, join, yet, clinton, nation
4	president, bush, administration, clinton, american, force, reagan, war, unite, lead, economic, iraq, congress, america, iraqi, policy, aid, international, military, see	4	president, administration, bush, clinton, war, unite, force, reagan, american, america, make, nation, military, iraq, iraqi, troops, international, country, yesterday, plan
⋮		⋮	
20	soviet, lead, gorbachev, union, west, mikhail, reform, change, europe, leaders, poland, communist, know, old, right, human, washington, western, bring, party	**20**	soviet, union, economic, reform, yeltsin, russian, lead, russia, gorbachev, leaders, west, president, boris, moscow, europe, poland, mikhail, communist, power, relations

Topic	Before	Topic	After
1	election, yeltsin, russian, political, party, democratic, russia, president, democracy, boris, country, south, years, month, government, vote, since, leader, presidential, military	**1**	election, democratic, south, country, president, party, africa, lead, even, democracy, leader, presidential, week, politics, minister, percent, voter, last, month, years
2	new, york, city, state, mayor, budget, giuliani, council, cuomo, gov, plan, year, rudolph, dinkins, lead, need, governor, legislature, pataki, david	2	new, york, city, state, mayor, budget, council, giuliani, gov, cuomo, year, rudolph, dinkins, legislature, plan, david, governor, pataki, need, cut
3	nuclear, arms, weapon, defense, treaty, missile, world, unite, yet, soviet, lead, secretary, would, control, korea, intelligence, test, nation, country, testing	3	nuclear, arms, weapon, treaty, defense, war, missile, may, come, test, american, world, would, need, lead, get, join, yet, clinton, nation
4	president, bush, administration, clinton, american, force, reagan, war, unite, lead, economic, iraq, congress, america, iraqi, policy, aid, international, military, see	4	president, administration, bush, clinton, war, unite, force, reagan, american, america, make, nation, military, iraq, iraqi, troops, international, country, yesterday, plan
⋮		⋮	
20	soviet, lead, gorbachev, union, west, mikhail, reform, change, europe, leaders, poland, communist, know, old, right, human, washington, western, bring, party	**20**	soviet, union, economic, reform, yeltsin, russian, lead, russia, gorbachev, leaders, west, president, boris, moscow, europe, poland, mikhail, communist, power, relations

Example: Negative Constraint

Topic	Words
318	bladder, sci, spinal_cord, spinal_cord_injury, spinal, urinary, urinary_tract, urothelial,injury, motor, recovery, reflex, cervical, urothelium, functional_recovery

Example: Negative Constraint

Topic	Words
318	bladder, sci, spinal_cord, spinal_cord_injury, spinal, urinary, urinary_tract, urothelial,injury, motor, recovery, reflex, cervical, urothelium, functional_recovery

Negative Constraint

spinal_cord, bladder

Example: Negative Constraint

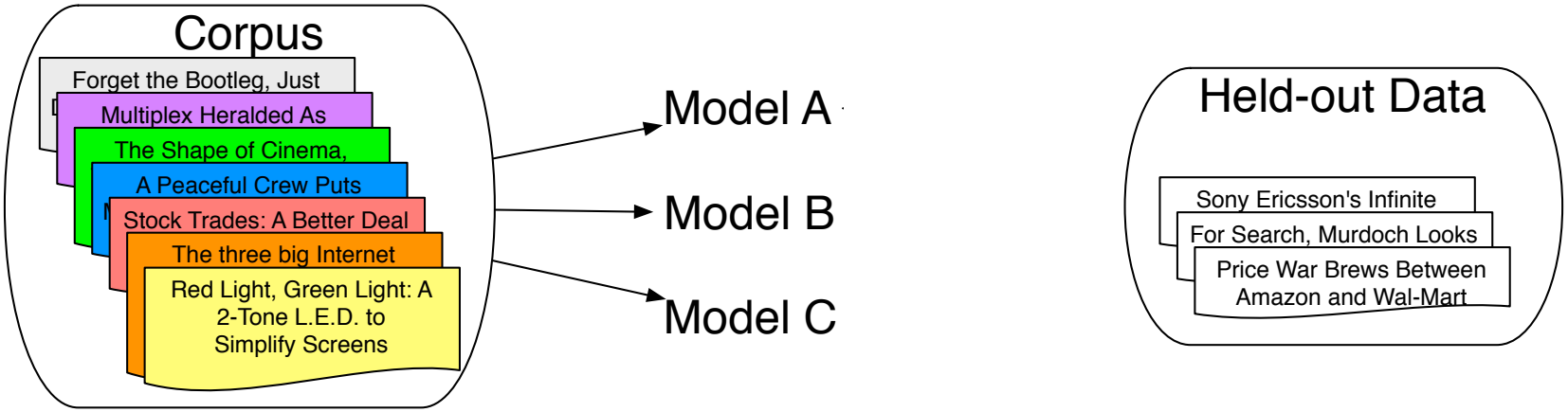
Topic	Words
318	bladder, sci, spinal_cord, spinal_cord_injury, spinal, urinary, urinary_tract, urothelial,injury, motor, recovery, reflex, cervical, urothelium, functional_recovery

Topic	Words
318	sci, spinal_cord, spinal_cord_injury, spinal, injury, recovery, motor, reflex, urothelial, injured, functional_recovery, plasticity, locomotor, cervical, locomotion

Negative Constraint

spinal_cord, bladder

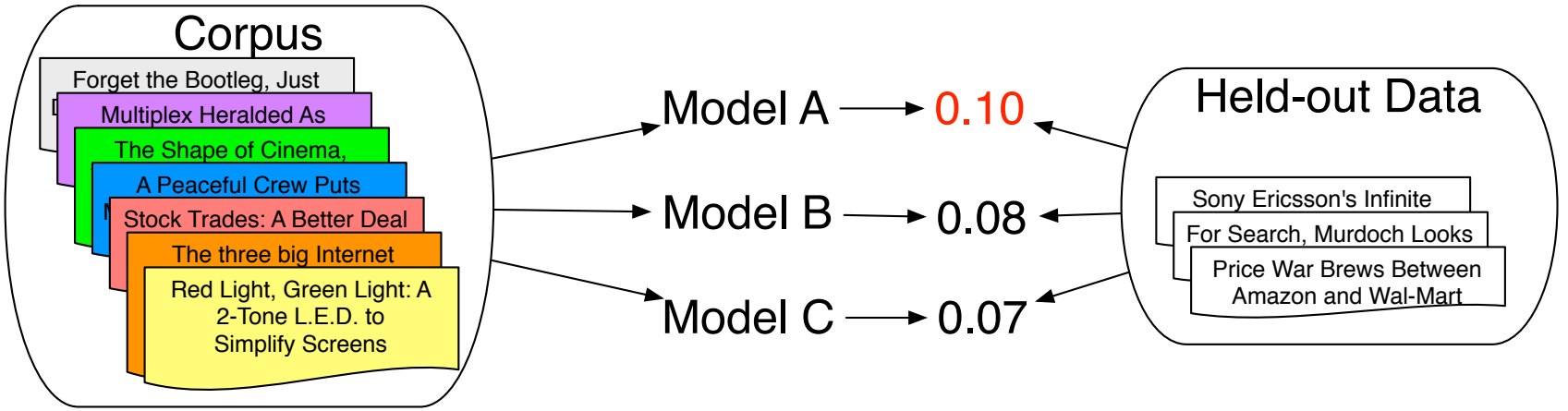
Evaluation



$$P(\boldsymbol{w} \mid \boldsymbol{w}', \boldsymbol{z}', \alpha\boldsymbol{m}, \beta\boldsymbol{u}) = \sum_{\boldsymbol{z}} P(\boldsymbol{w}, \boldsymbol{z} \mid \boldsymbol{w}', \boldsymbol{z}', \alpha\boldsymbol{m}, \beta\boldsymbol{u})$$

How you compute it is important too [12]

Evaluation



Measures predictive power, not what the topics are

$$P(\boldsymbol{w} \mid \boldsymbol{w}', \boldsymbol{z}', \alpha\boldsymbol{m}, \beta\boldsymbol{u}) = \sum_{\boldsymbol{z}} P(\boldsymbol{w}, \boldsymbol{z} \mid \boldsymbol{w}', \boldsymbol{z}', \alpha\boldsymbol{m}, \beta\boldsymbol{u})$$

How you compute it is important too [12]

Word Intrusion

TOPIC 1	TOPIC 2	TOPIC 3
computer, technology, system, service, site, phone, internet, machine	sell, sale, store, product, business, advertising, market, consumer	play, film, movie, theater, production, star, director, stage

Word Intrusion

1. Take the highest probability words from a topic

Original Topic

dog, cat, horse, pig, cow

Word Intrusion

1. Take the highest probability words from a topic

Original Topic

dog, cat, horse, pig, cow

2. Take a high-probability word from another topic and add it

Topic with Intruder

dog, cat, apple, horse, pig, cow

Word Intrusion

1. Take the highest probability words from a topic

Original Topic

dog, cat, horse, pig, cow

2. Take a high-probability word from another topic and add it

Topic with Intruder

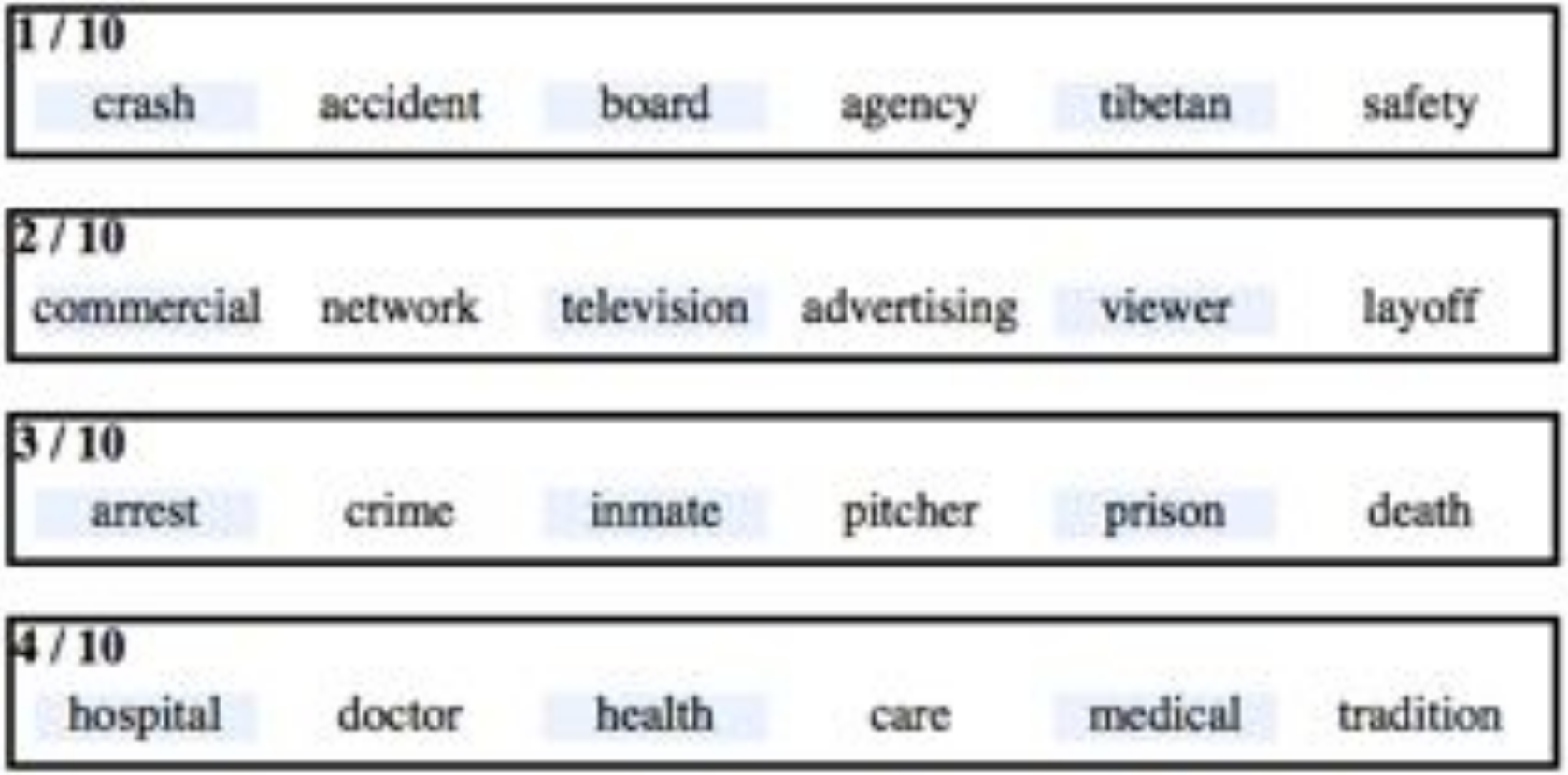
dog, cat, apple, horse, pig, cow

3. We ask users to find the word that doesn't belong

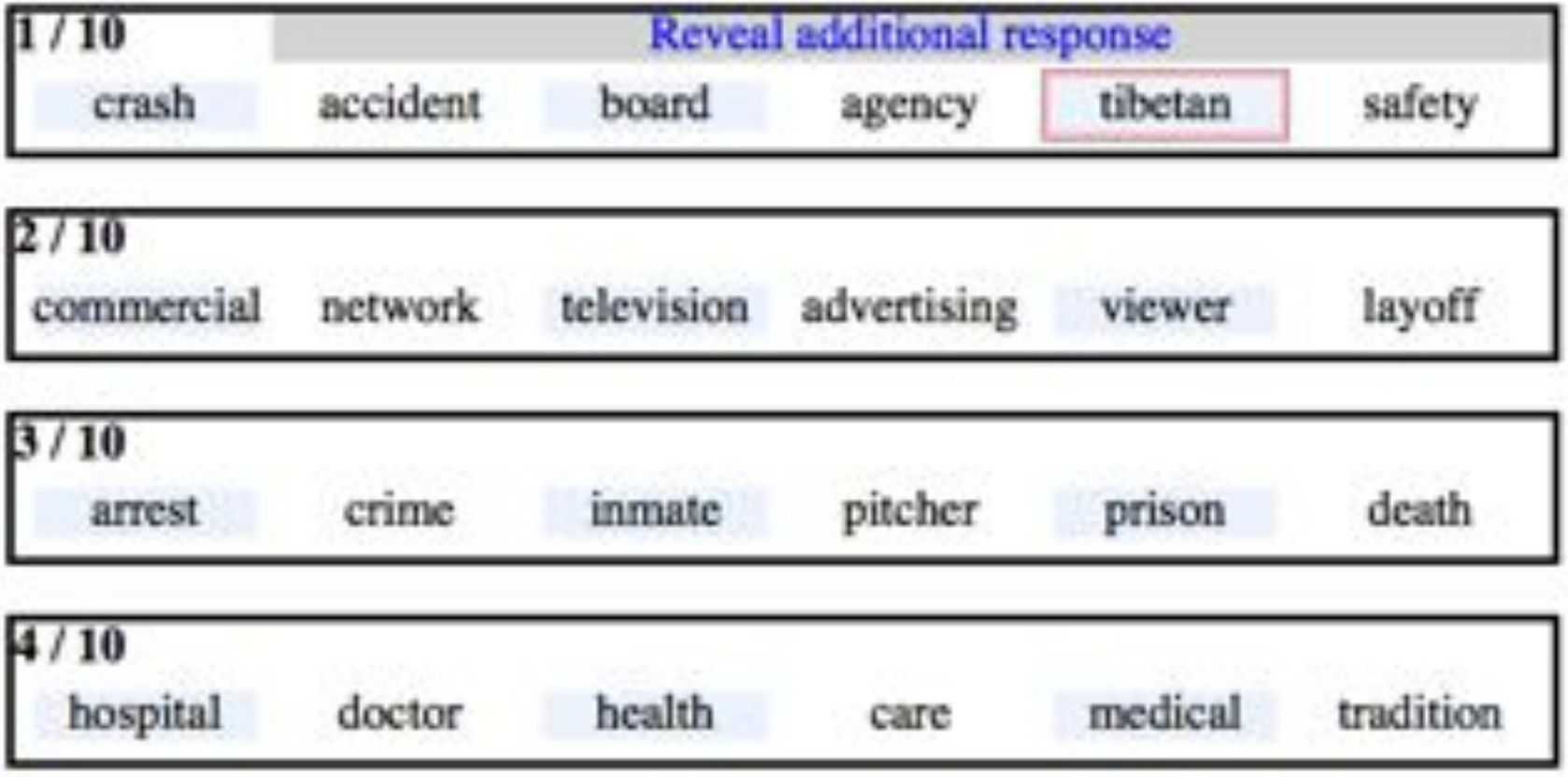
Hypothesis

If the topics are interpretable, users will consistently choose true intruder

Word Intrusion



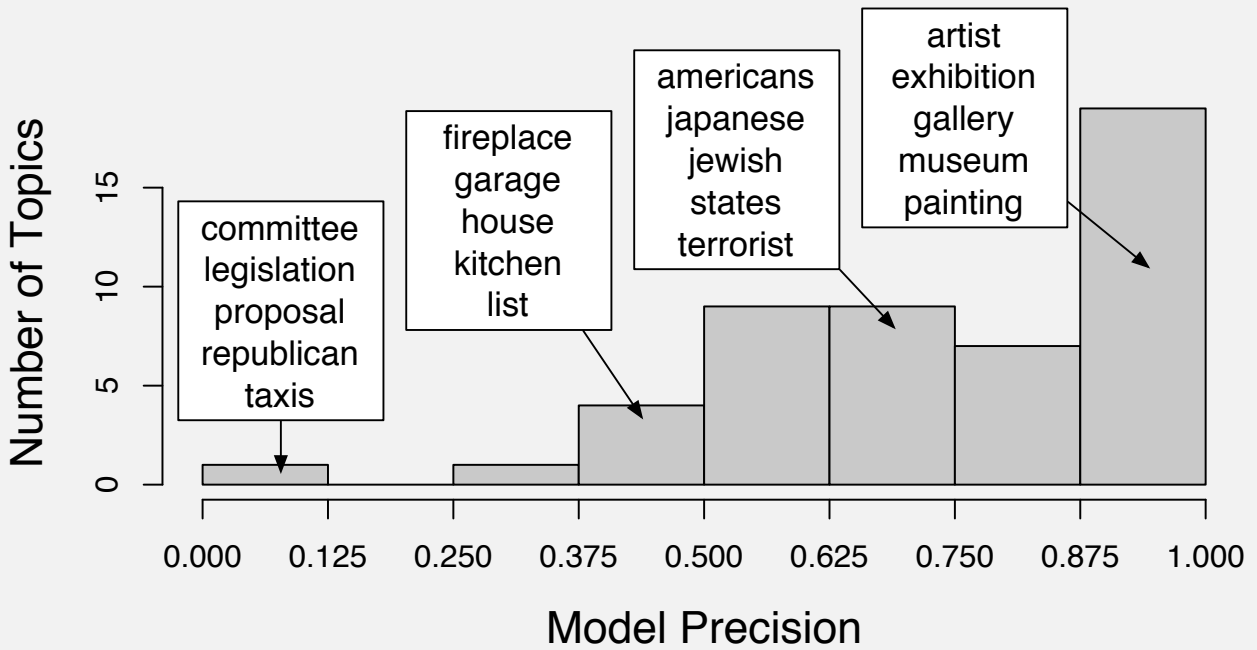
Word Intrusion



- Order of words was shuffled
- Which intruder was selected varied
- Model precision: percentage of users who clicked on intruder

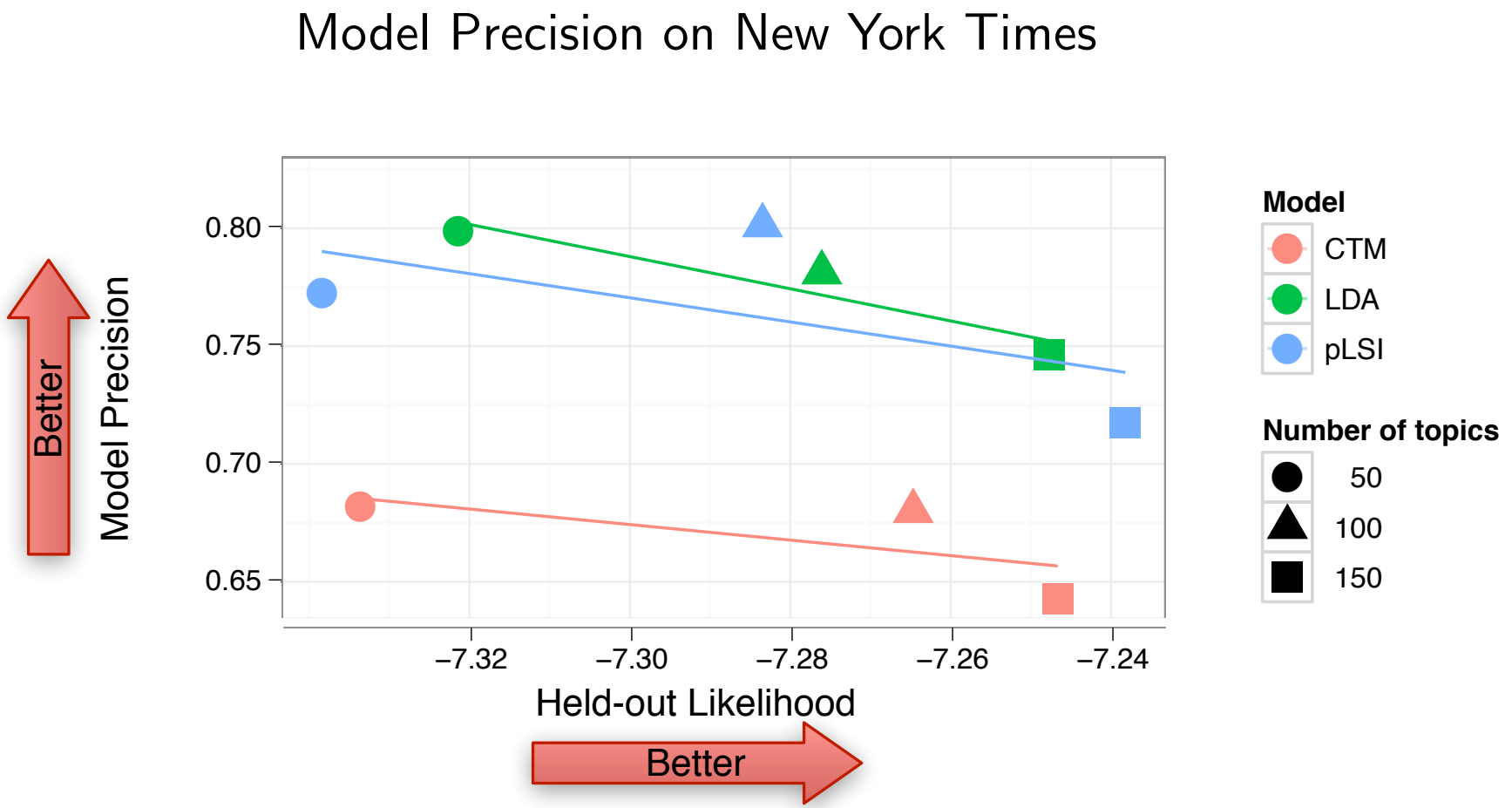
Word Intrusion: Which Topics are Interpretable?

New York Times, 50 LDA Topics



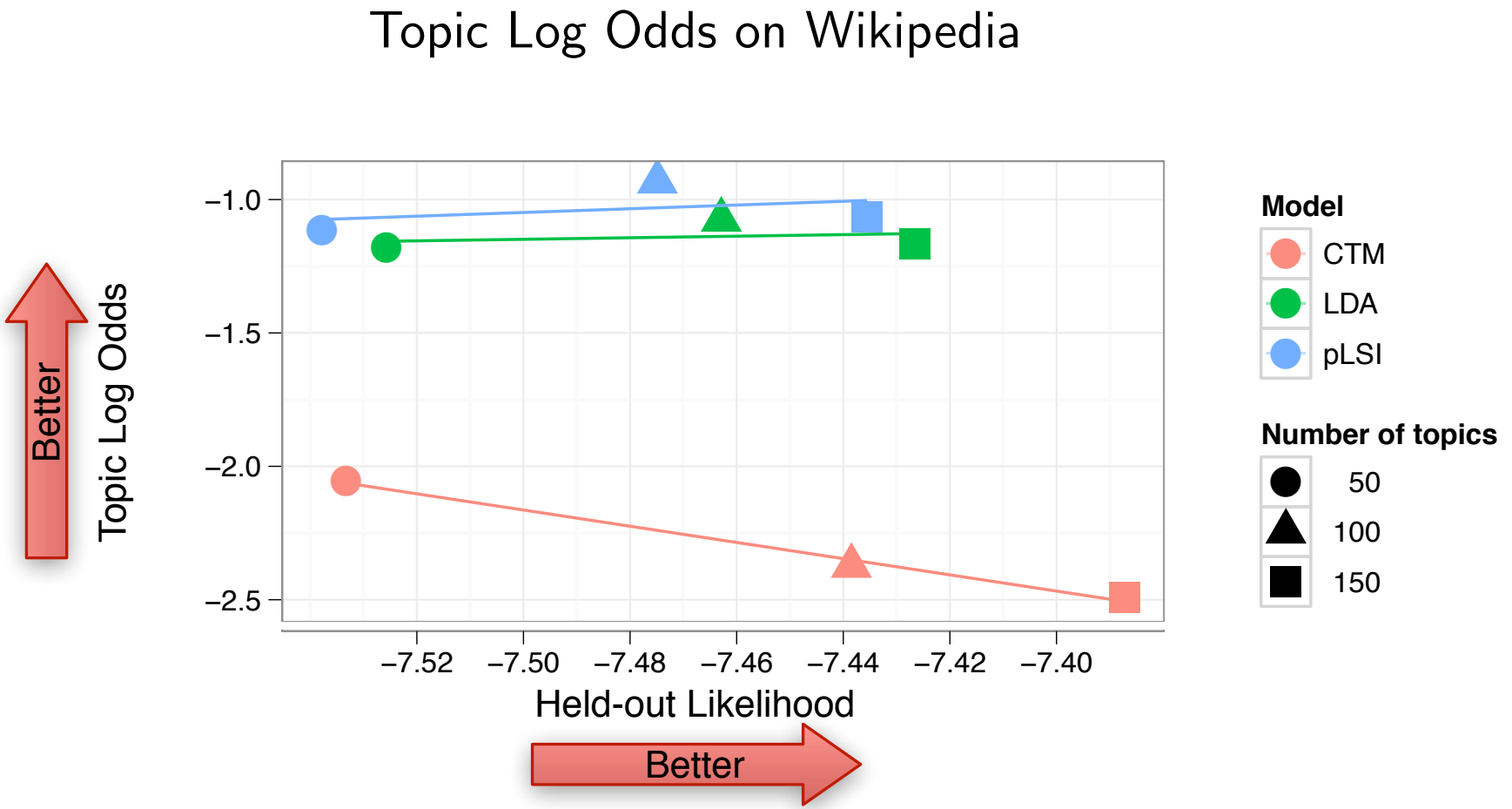
Model Precision: percentage of correct intruders found

Interpretability and Likelihood



within a model, higher likelihood $\neq$ higher interpretability

Interpretability and Likelihood

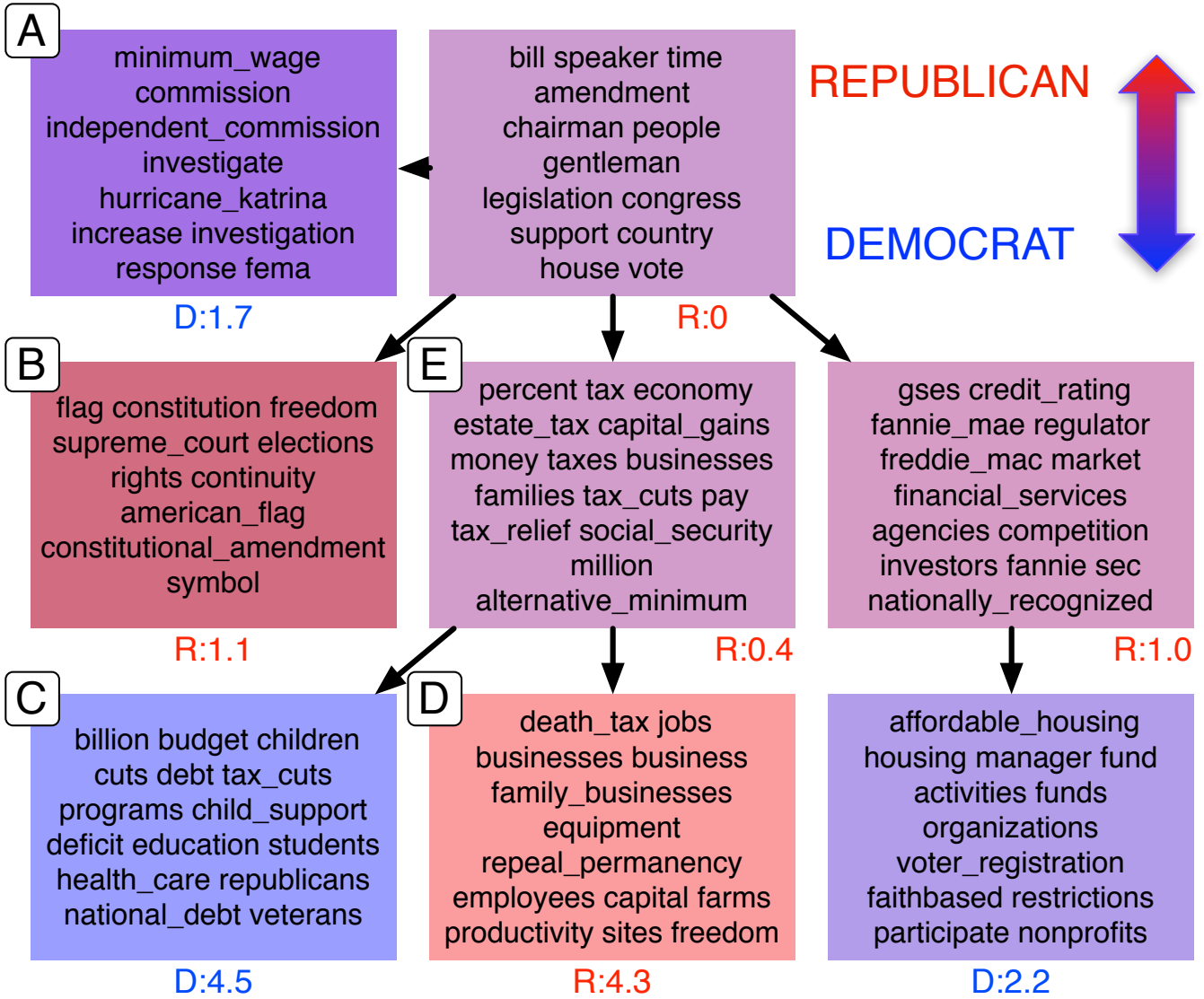


across models, higher likelihood $\neq$ higher interpretability

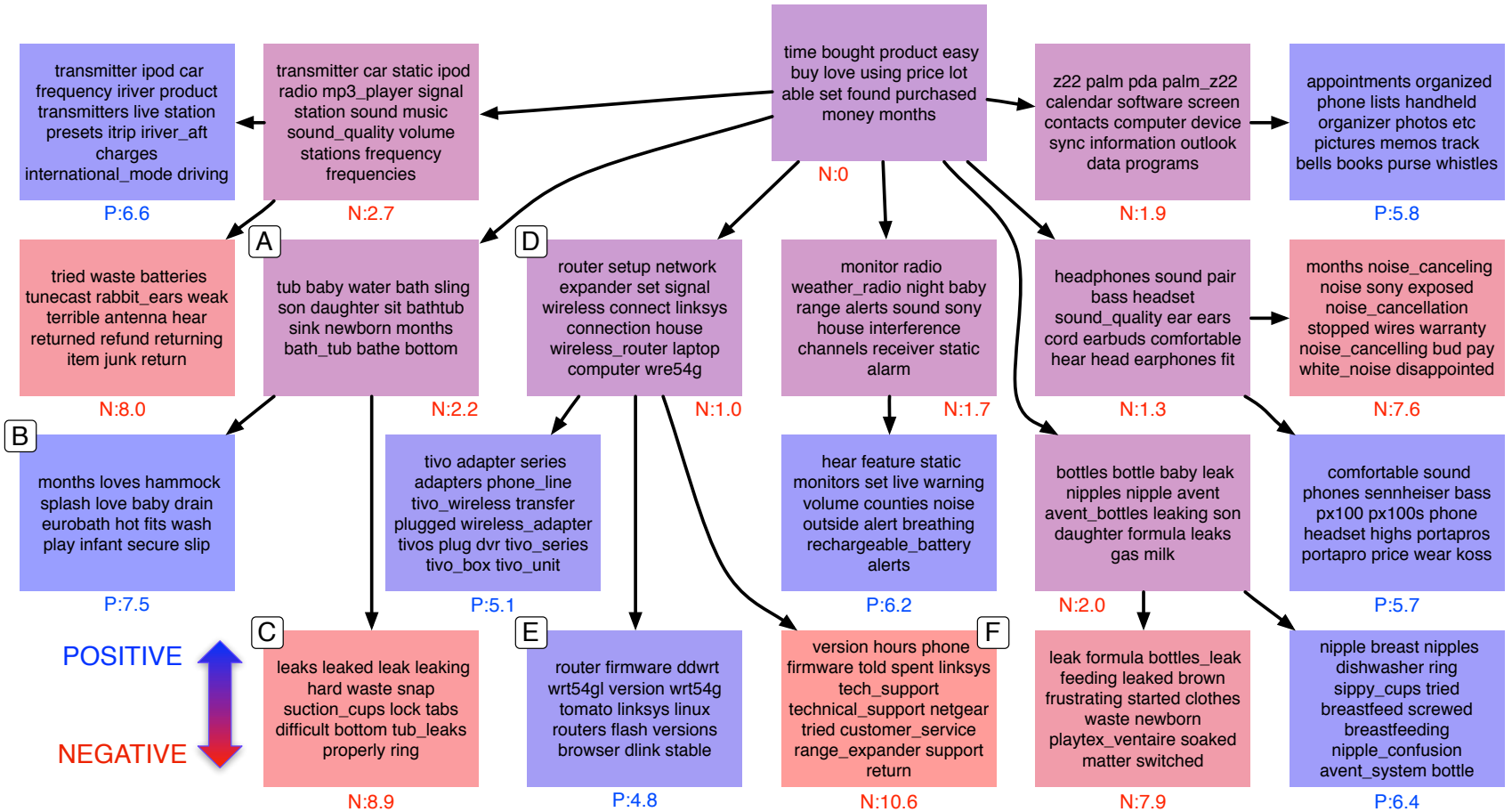
Evaluation Takeaway

- Measure what you care about [5]
- If you care about prediction, likelihood is good
- If you care about a particular task, measure that

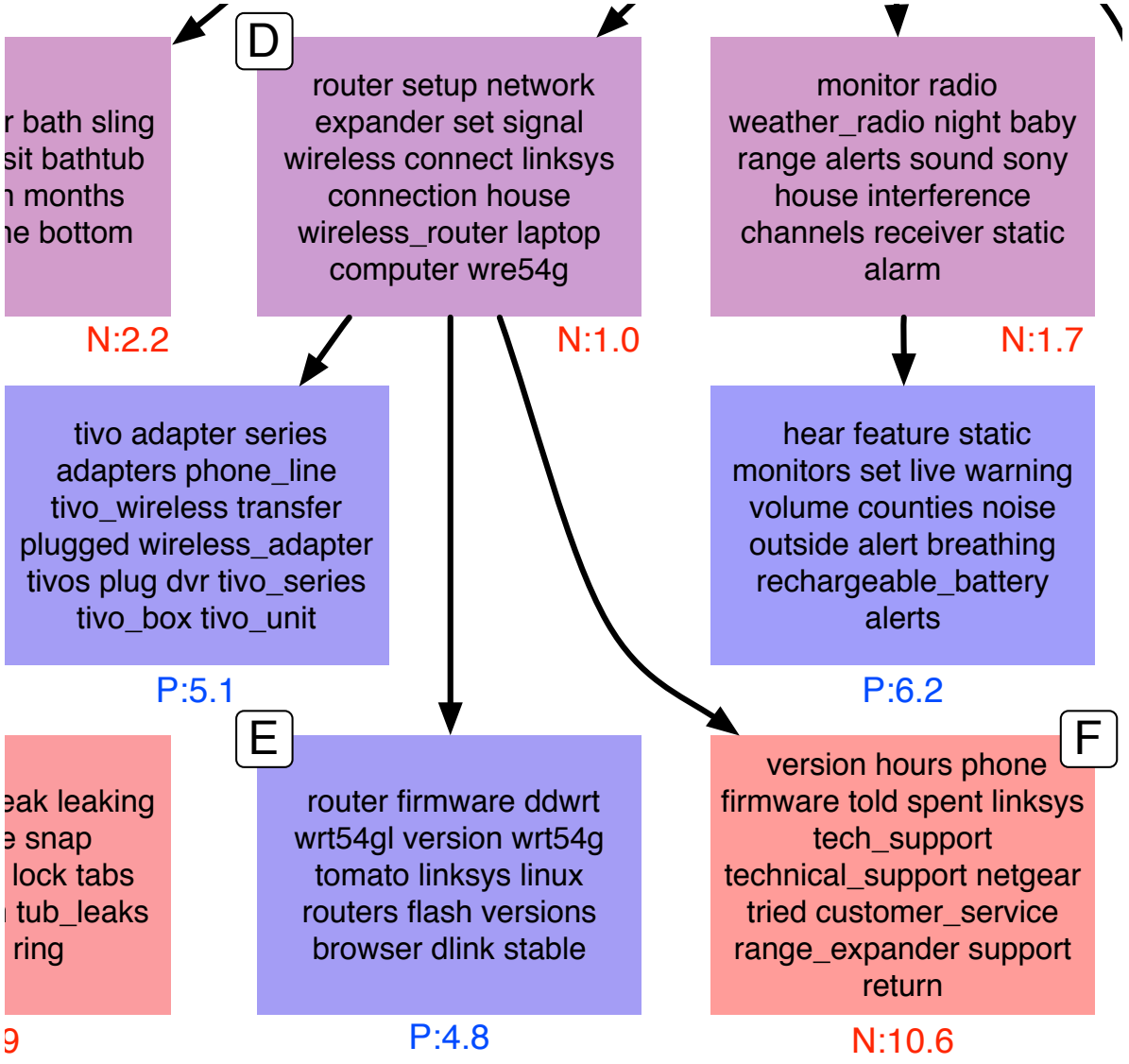
Other Research: Hierarchical Models



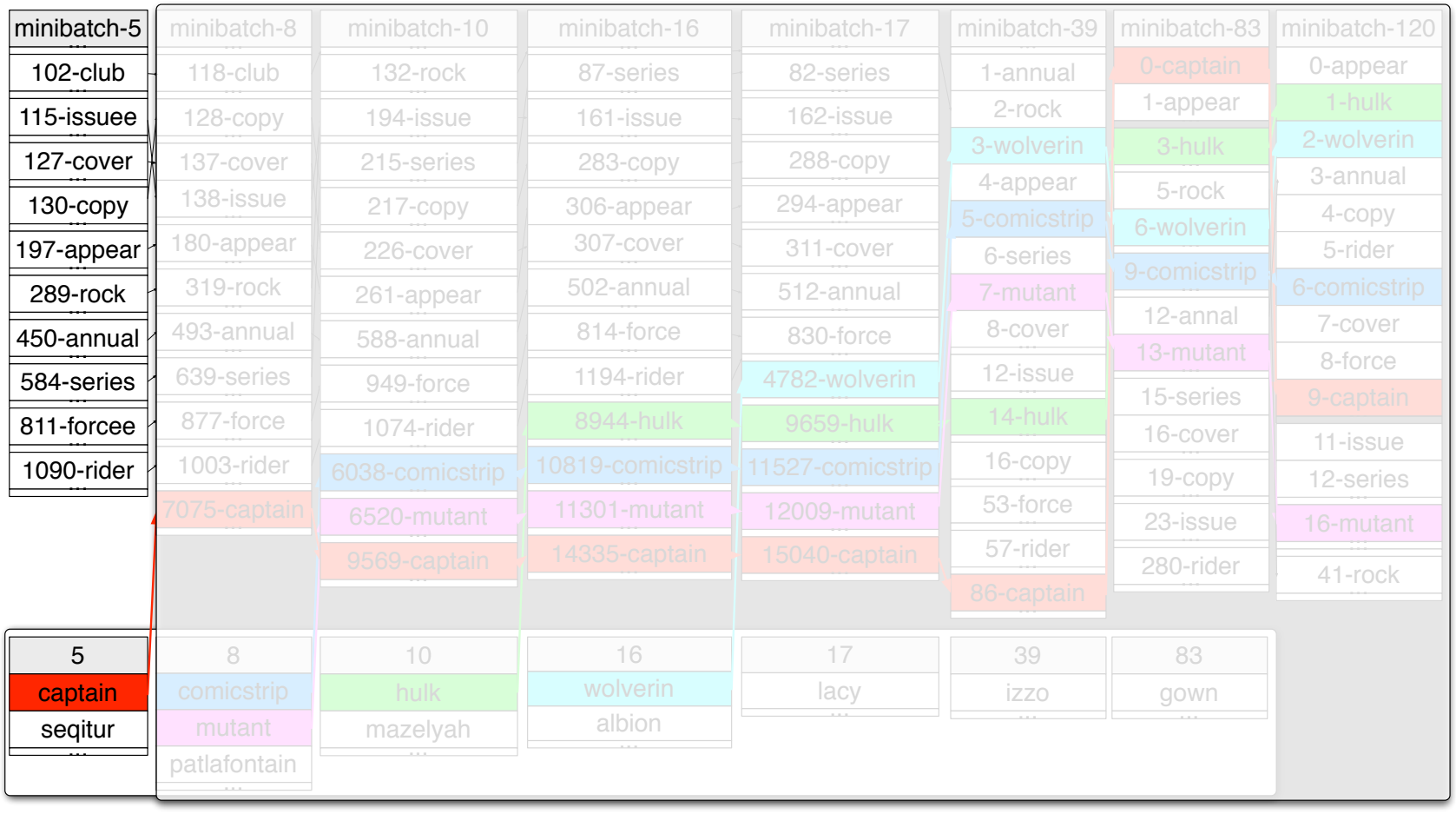
Other Research: Hierarchical Models



Other Research: Hierarchical Models

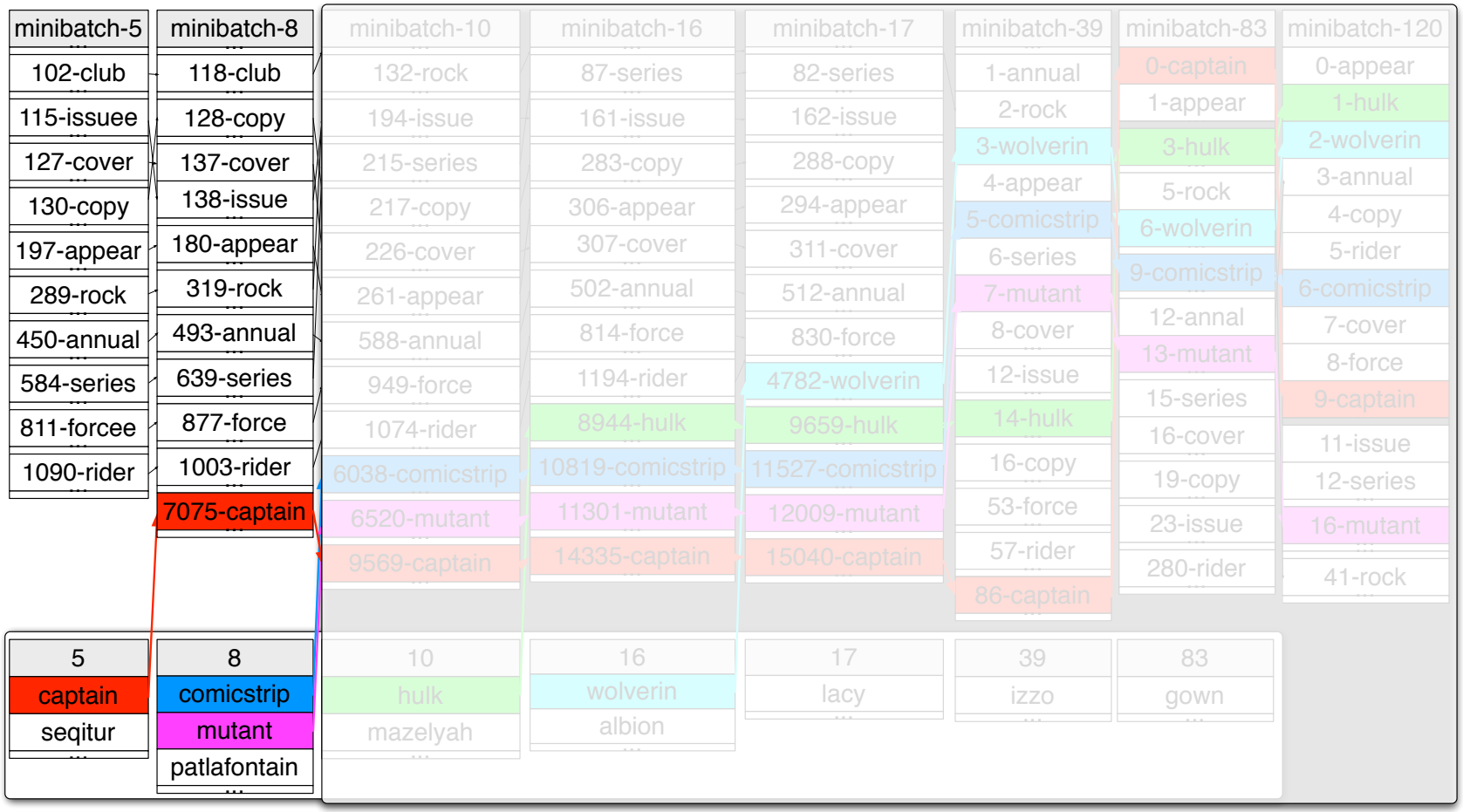


Other Research: Incorporating New Words



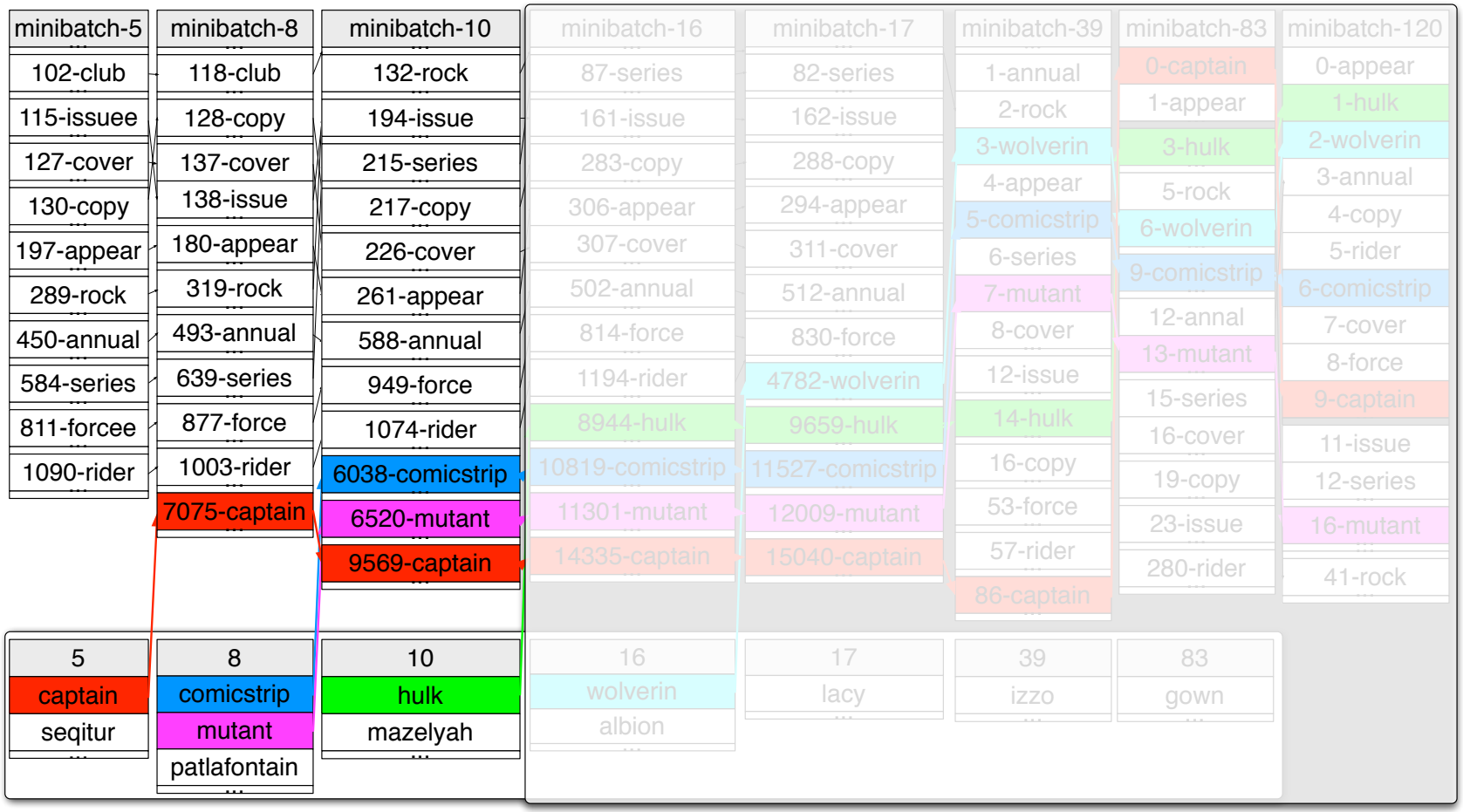
An example topic from *20 newsgroups* under our model. Numbers preceeding words are ranks in topic.

Other Research: Incorporating New Words



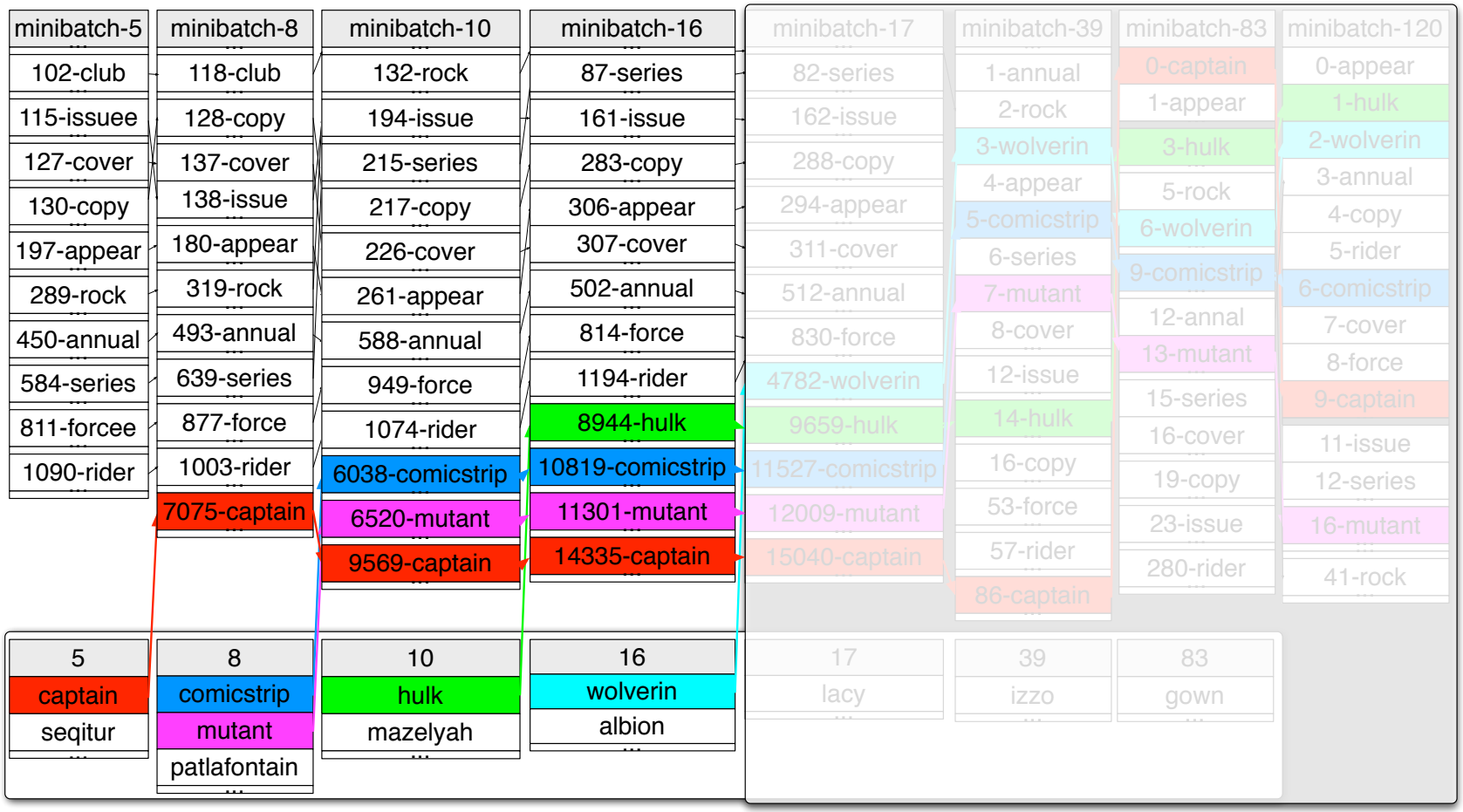
An example topic from *20 newsgroups* under our model. Numbers preceeding words are ranks in topic.

Other Research: Incorporating New Words



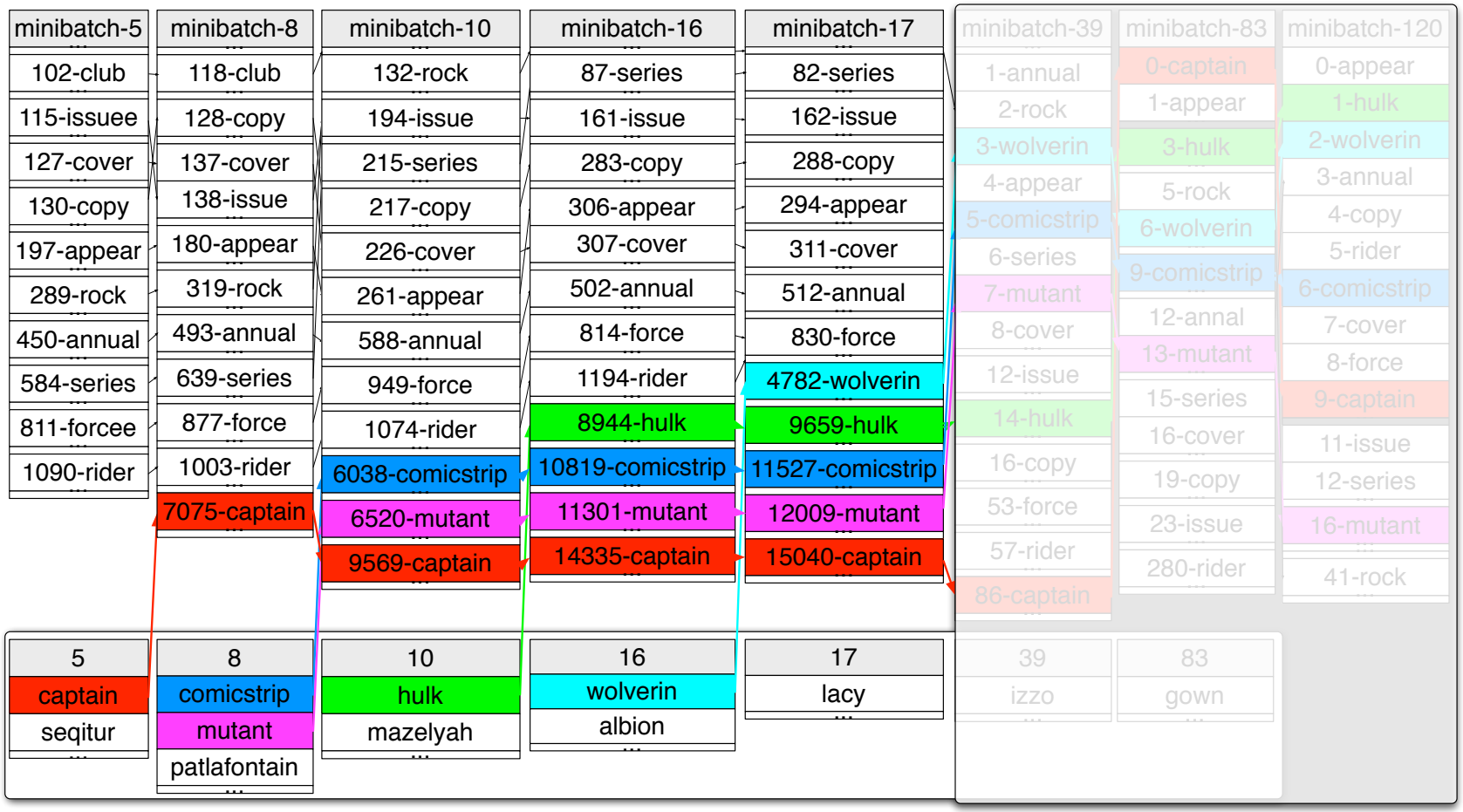
An example topic from *20 newsgroups* under our model. Numbers preceeding words are ranks in topic.

Other Research: Incorporating New Words



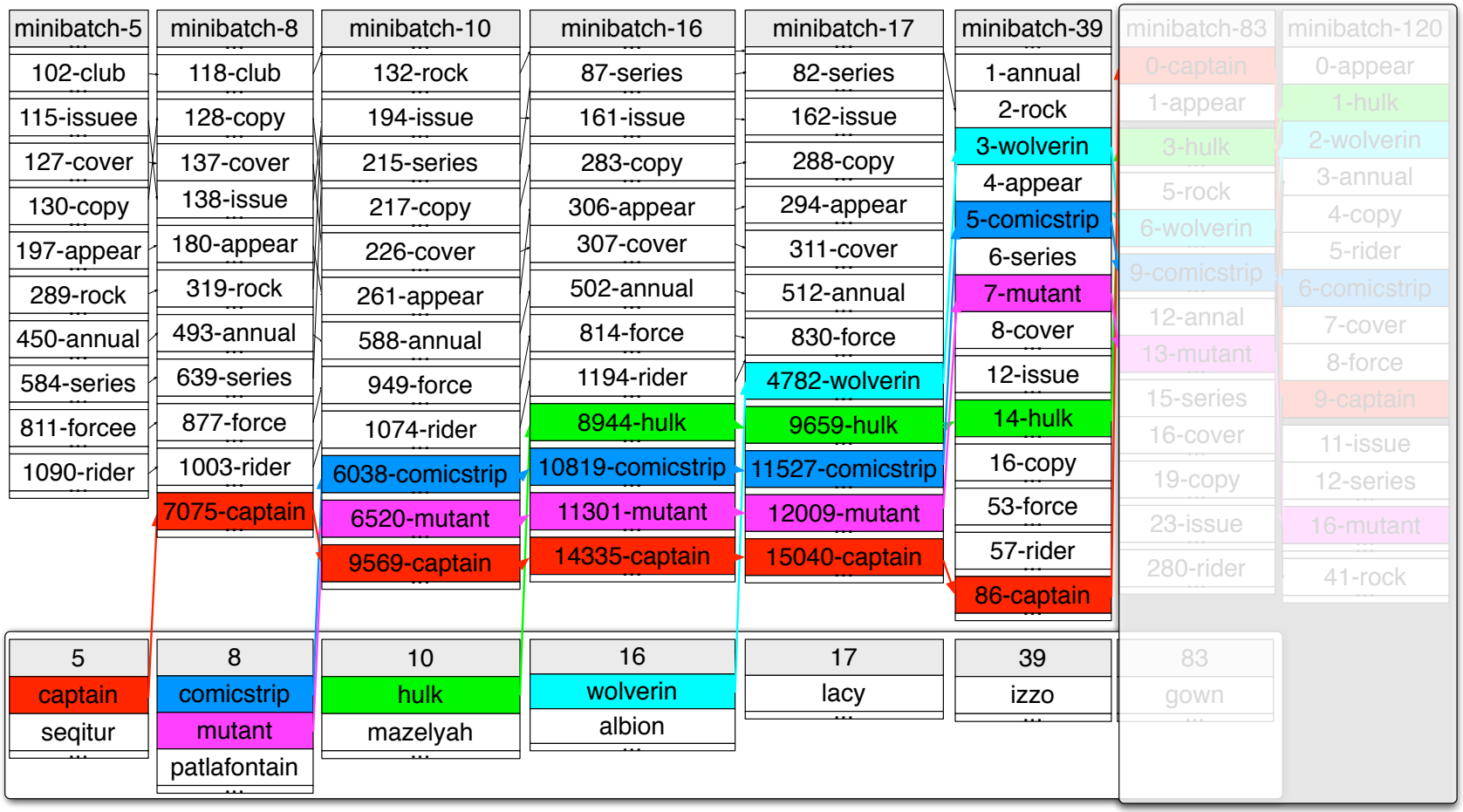
An example topic from *20 newsgroups* under our model. Numbers preceeding words are ranks in topic.

Other Research: Incorporating New Words



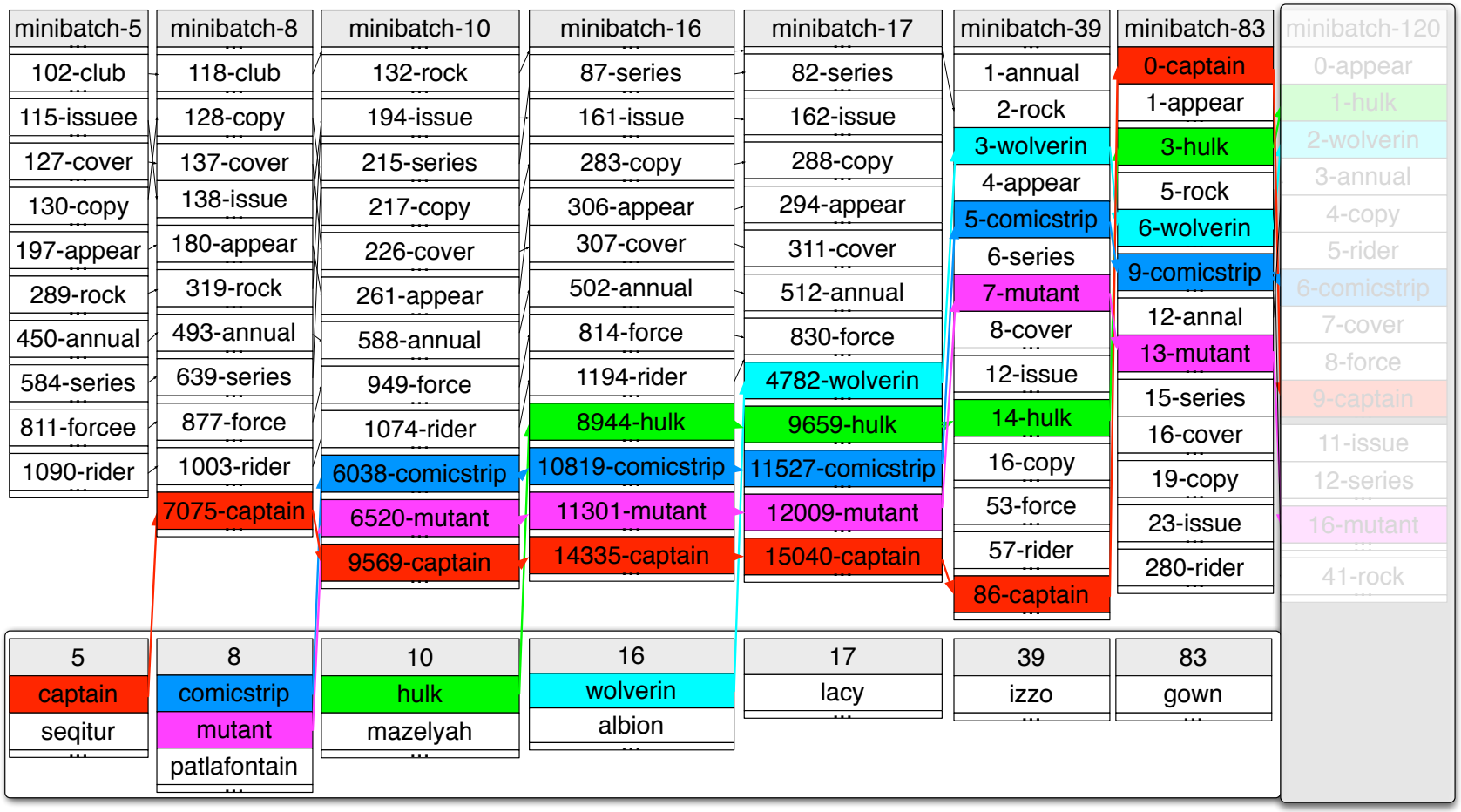
An example topic from *20 newsgroups* under our model. Numbers preceeding words are ranks in topic.

Other Research: Incorporating New Words



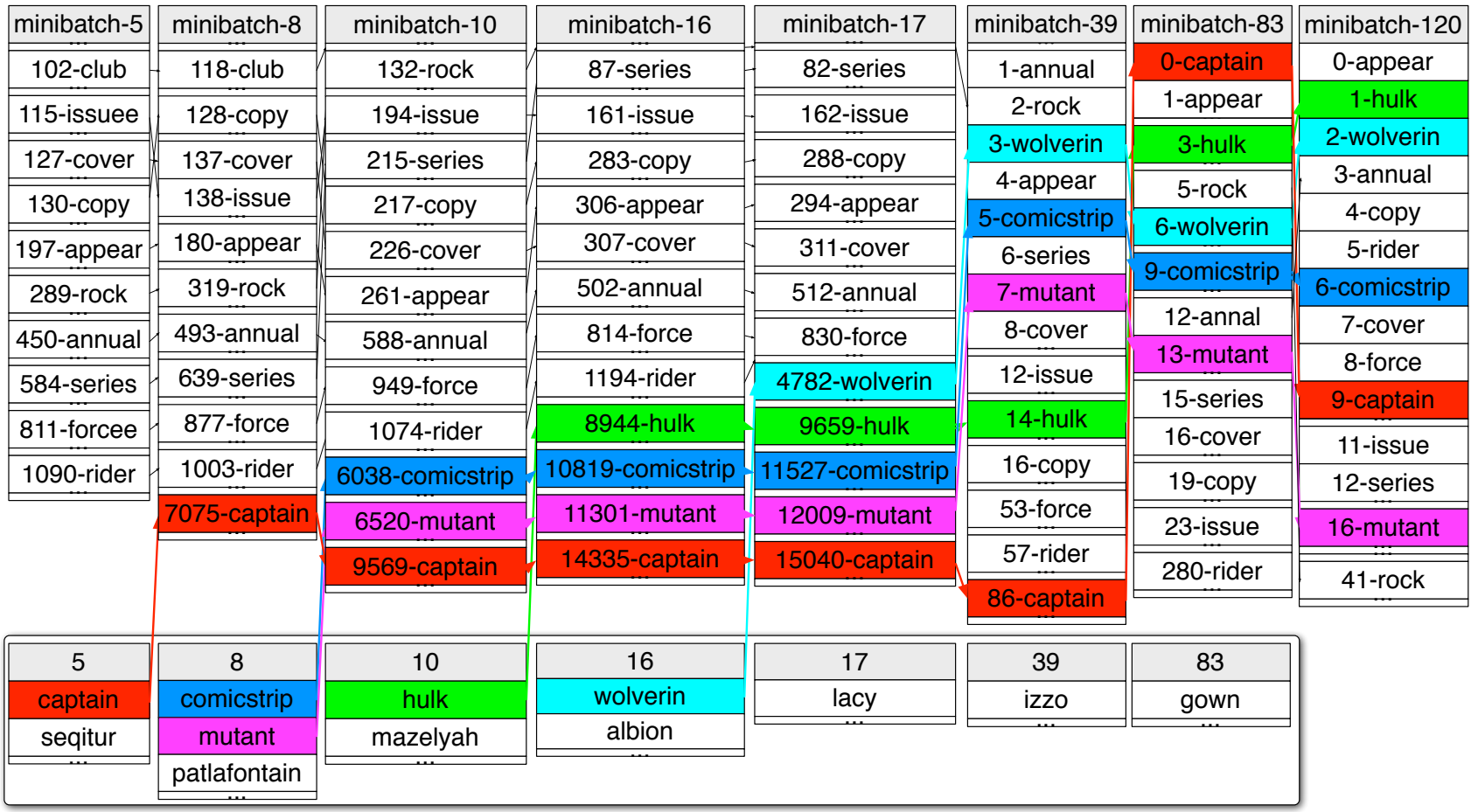
An example topic from *20 newsgroups* under our model. Numbers preceeding words are ranks in topic.

Other Research: Incorporating New Words



An example topic from *20 newsgroups* under our model. Numbers preceeding words are ranks in topic.

Other Research: Incorporating New Words



An example topic from *20 newsgroups* under our model. Numbers preceeding words are ranks in topic.

Outline

Topic Models for Large Text Collections

- Unsupervised tool for understanding large text collections
- Statistical foundation
- Open source tools for learning these models
- Example application to NSF and NRC data

Thanks

Collaborators

Yuening Hu (UMD), Ke Zhai (UMD), Viet-An Nguyen (UMD), Dave Blei (Princeton), Jonathan Chang (Facebook), Philip Resnik (UMD), Christiane Fellbaum (Princeton), Jerry Zhu (Wisconsin), Sean Gerrish (Sift), Chong Wang (CMU), Dan Osherson (Princeton), Sinead Williamson (CMU)

Funders



David Andrzejewski, Xiaojin Zhu, and Mark Craven.
Incorporating domain knowledge into topic modeling via Dirichlet forest priors.
In *Proceedings of the International Conference of Machine Learning*, 2009.


David M. Blei, Andrew Ng, and Michael Jordan.
Latent Dirichlet allocation.
Journal of Machine Learning Research, 2003.


Jordan Boyd-Graber, David M. Blei, and Xiaojin Zhu.
A topic model for word sense disambiguation.
In *Proceedings of Emperical Methods in Natural Language Processing*, 2007.


Samuel Brody and Mirella Lapata.
Bayesian word sense induction.
In *Proceedings of the European Chapter of the Association for Computational Linguistics*, Athens, Greece, 2009.


Jonathan Chang, Jordan Boyd-Graber, and David M. Blei.
Connections between the lines: Augmenting social networks with text.
In *Knowledge Discovery and Data Mining*, 2009.


Thomas L. Griffiths, Mark Steyvers, and Joshua Tenenbaum.
Topics in semantic representation.
Psychological Review, 114(2):211–244, 2007.


Thomas Hofmann.
Probabilistic latent semantic analysis.
In *Proceedings of Uncertainty in Artificial Intelligence*, 1999.


Yuening Hu and Jordan Boyd-Graber.
Efficient tree-based topic modeling.