Topic Models

Computational Linguistics: Jordan Boyd-Graber
University of Maryland
INTRODUCTION

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?

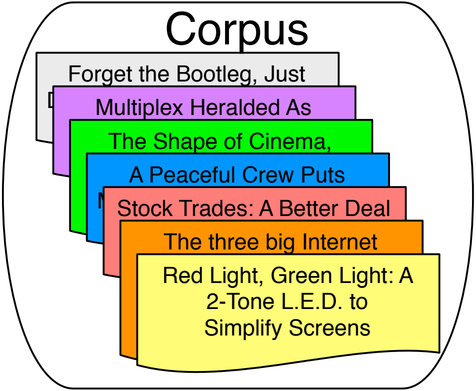
- Suppose you have a huge number of documents
- Want to know what's going on
- Can't read them all (e.g. every New York Times article from the 90's)
- Topic models offer a way to get a corpus-level view of major themes
- Unsupervised

Roadmap

- What are topic models
- How to know if you have good topic model
- How to go from raw data to topics

Conceptual Approach

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



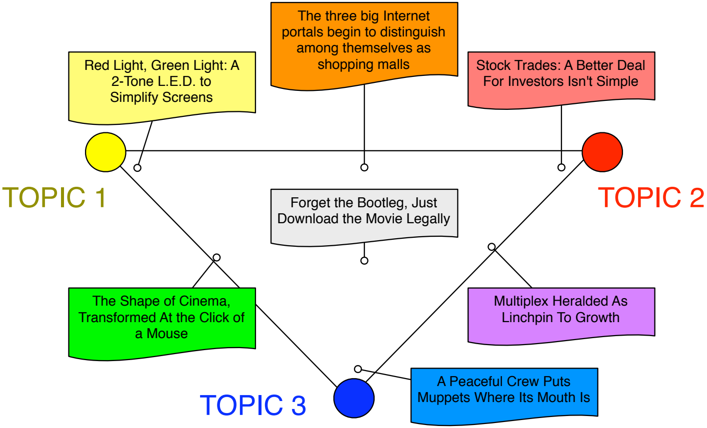
Conceptual Approach

From an input corpus and number of topics $K \longrightarrow$ **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
  - Word Sense Disambiguation
  - Discourse Segmentation
  - Machine Translation
- Psychology: word meaning, polysemy
- Inference is (relatively) simple

Matrix Factorization Approach

$$\underbrace{\begin{bmatrix} M \times K \end{bmatrix}}_{\text{Topic Assignment}} \times \underbrace{\begin{bmatrix} K \times V \end{bmatrix}}_{\text{Topics}} \approx \underbrace{\begin{bmatrix} M \times V \end{bmatrix}}_{\text{Dataset}}$$

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

Matrix Factorization Approach

$$\underbrace{\begin{bmatrix} M \times K \end{bmatrix}}_{\text{Topic Assignment}} \times \underbrace{\begin{bmatrix} \quad K \times V \quad \end{bmatrix}}_{\text{Topics}} \approx \underbrace{\begin{bmatrix} \quad M \times V \quad \end{bmatrix}}_{\text{Dataset}}$$

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

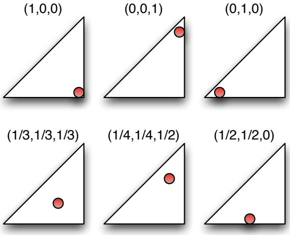
- How your data came to be
- Sequence of Probabilistic Steps
- Posterior Inference

Alternative: Generative Model

- How your data came to be
- Sequence of Probabilistic Steps
- Posterior Inference
- Blei, Ng, Jordan. Latent **Dirichlet** Allocation. JMLR, 2003.

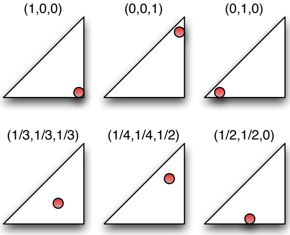
Multinomial Distribution

- 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



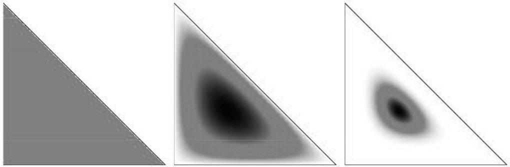
- Come from a Dirichlet distribution

Dirichlet Distribution

$$P(\boldsymbol{p} \mid \alpha \boldsymbol{m}) = \frac{\Gamma(\sum_k \alpha m_k)}{\prod_k \Gamma(\alpha m_k)} \prod_k p_k^{\alpha m_k - 1}$$

Dirichlet Distribution

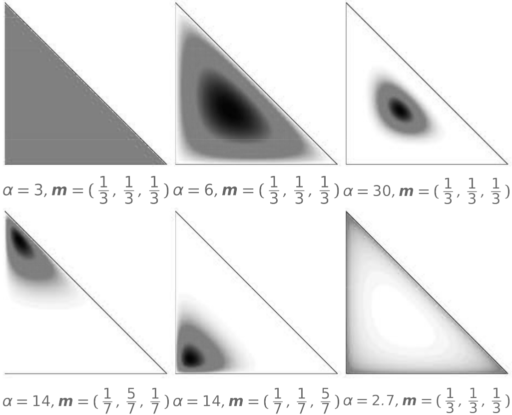
$$P(\boldsymbol{p} \,|\, \alpha \boldsymbol{m}) = \frac{\Gamma(\sum_k \alpha m_k)}{\prod_k \Gamma(\alpha m_k)} \prod_k p_k^{\alpha m_k - 1}$$



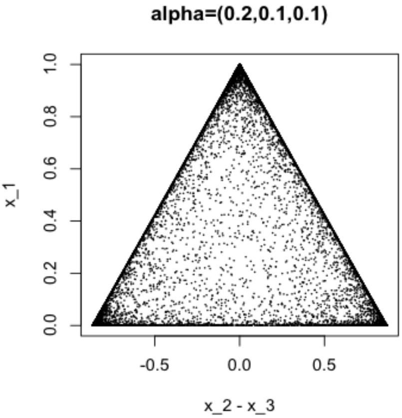
$\alpha = 3, \boldsymbol{m} = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ $\alpha = 6, \boldsymbol{m} = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ $\alpha = 30, \boldsymbol{m} = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$

Dirichlet Distribution

$$P(\boldsymbol{p} \mid \alpha \boldsymbol{m}) = \frac{\Gamma(\sum_k \alpha m_k)}{\prod_k \Gamma(\alpha m_k)} \prod_k p_k^{\alpha m_k - 1}$$



Dirichlet Distribution



Dirichlet Distribution

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

$$
\begin{aligned}
p(\phi|\alpha, \vec{w}) &\propto p(\vec{w}|\phi)p(\phi|\alpha) && (1)\\
&\propto \prod_k \phi^{n_k} \prod_k \phi^{\alpha_k - 1} && (2)\\
&\propto \prod_k \phi^{\alpha_k + n_k - 1} && (3)
\end{aligned}
$$

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

Dirichlet Distribution

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

$$p(\phi|\alpha, \vec{w}) \propto p(\vec{w}|\phi)p(\phi|\alpha) \quad (1)$$

$$\propto \prod_k \phi^{n_k} \prod_k \phi^{\alpha_k - 1} \quad (2)$$

$$\propto \prod_k \phi^{\alpha_k + n_k - 1} \quad (3)$$

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

Generative Model

TOPIC 1

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

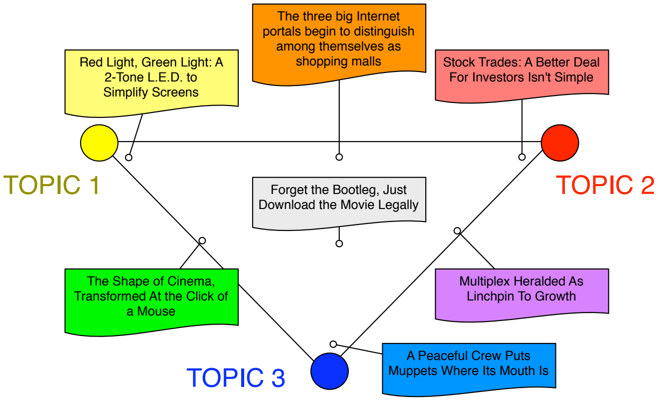
TOPIC 2

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

TOPIC 3

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

Generative Model



Generative Model

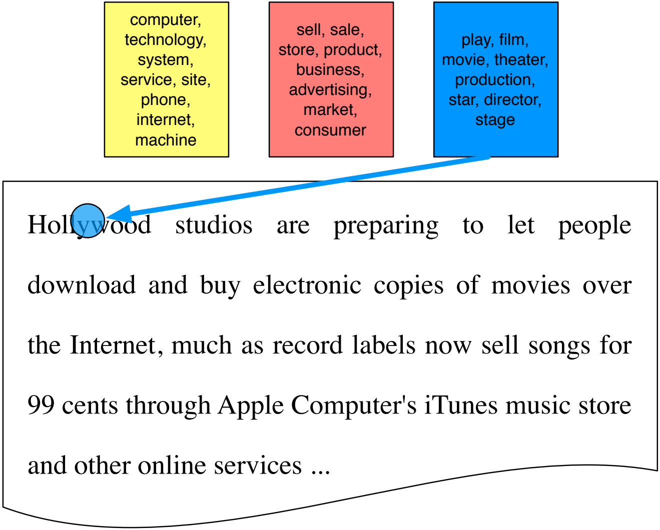
computer, technology, system, service, site, phone, internet, machine

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

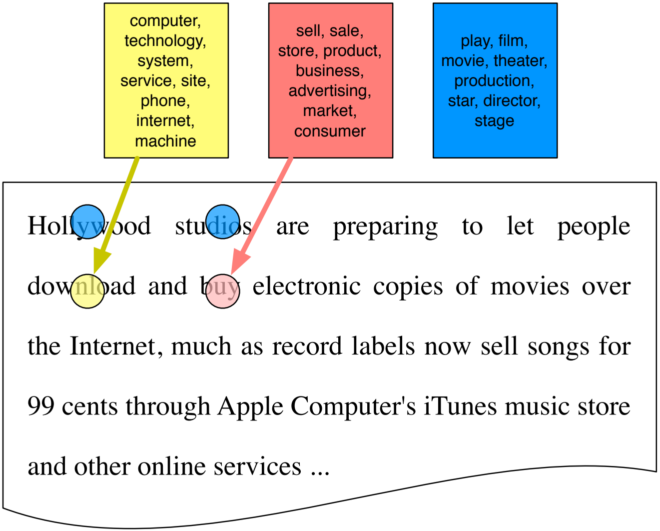
play, film, movie, theater, production, star, director, stage

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 ...

Generative Model



Generative Model

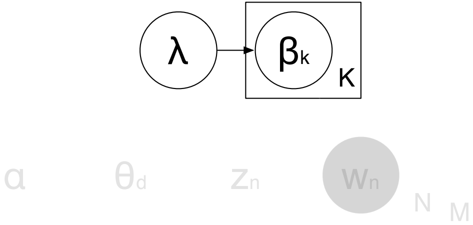


Generative Model

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

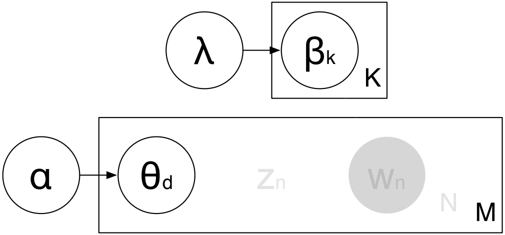
Hollywood studios are preparing to let people download and buy electronic copies of movies over the Internet, much as record labels now sell songs for 99 cents through Apple Computer's iTunes music store and other online services ...

Generative Model Approach



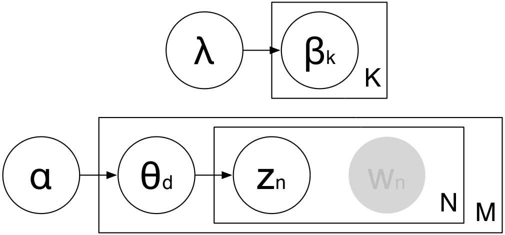
- For each topic $k \in \{1, \dots, K\}$, draw a multinomial distribution β_k from a Dirichlet distribution with parameter λ

Generative Model Approach



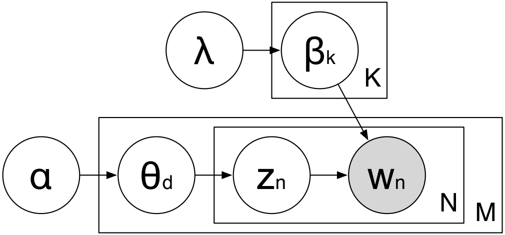
- For each topic $k \in \{1, \dots, K\}$, draw a multinomial distribution β_k from a Dirichlet distribution with parameter λ
- For each document $d \in \{1, \dots, M\}$, draw a multinomial distribution θ_d from a Dirichlet distribution with parameter α

Generative Model Approach



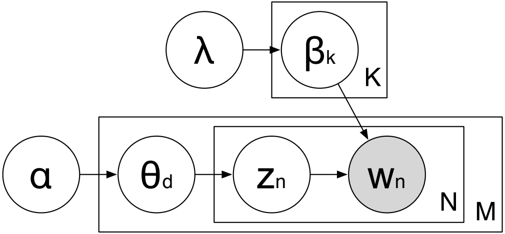
- For each topic $k \in \{1, \ldots, K\}$, draw a multinomial distribution β_k from a Dirichlet distribution with parameter λ
- For each document $d \in \{1, \ldots, M\}$, draw a multinomial distribution θ_d from a Dirichlet distribution with parameter α
- For each word position $n \in \{1, \ldots, N\}$,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

Generative Model Approach



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

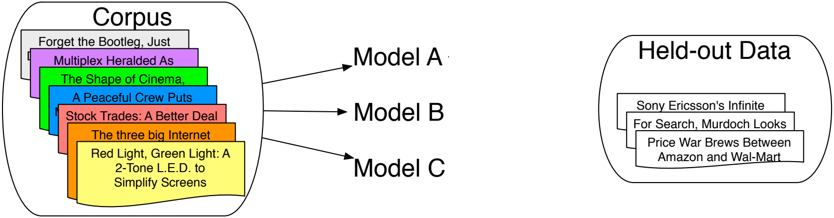
Topic Models: What's Important

- Topic models
  - Topics to word types
  - Documents to topics
  - Topics to word types—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), a fully Bayesian version of pLSI, probabilistic version of LSA

Topic Models: What's Important

- Topic models (latent variables)
  - Topics to word types
  - Documents to topics
  - Topics to word types—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), a fully Bayesian version of pLSI, probabilistic version of LSA

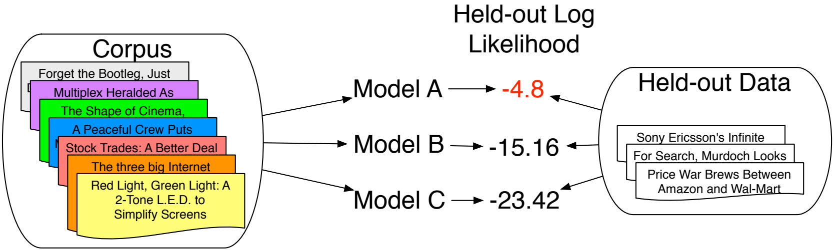
Evaluation



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

How you compute it is important too (Wallach et al. 2009)

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 (Wallach et al. 2009)

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

1 / 10	crash	accident	board	agency	tibetan	safety
2 / 10	commercial	network	television	advertising	viewer	layoff
3 / 10	arrest	crime	inmate	pitcher	prison	death
4 / 10	hospital	doctor	health	care	medical	tradition

Word Intrusion

1 / 10 Reveal additional response

crash	accident	board	agency	tibetan	safety

2 / 10

commercial	network	television	advertising	viewer	layoff

3 / 10

arrest	crime	inmate	pitcher	prison	death

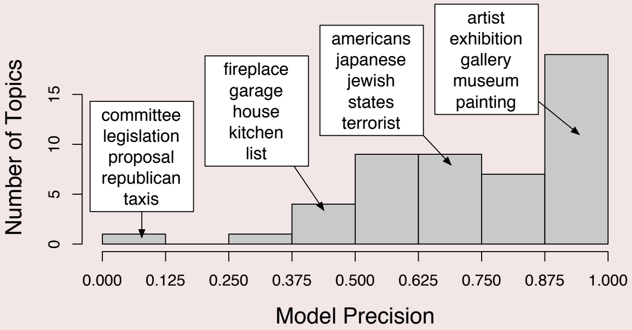
4 / 10

hospital	doctor	health	care	medical	tradition

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

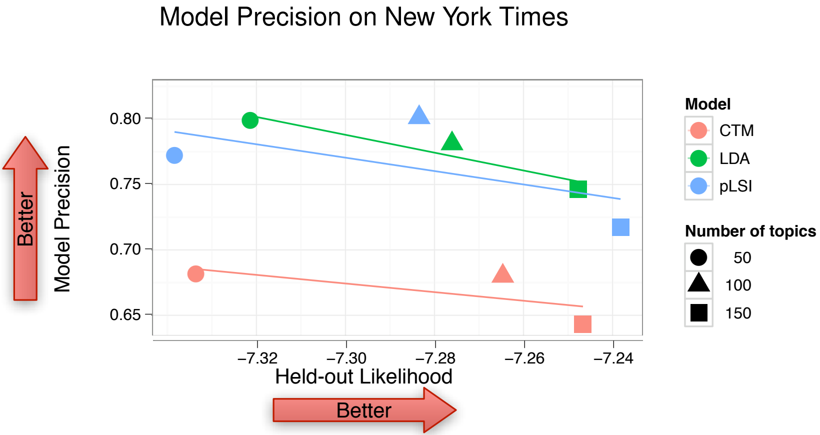
Word Intrusion: Which Topics are Interpretable?

New York Times, 50 LDA Topics



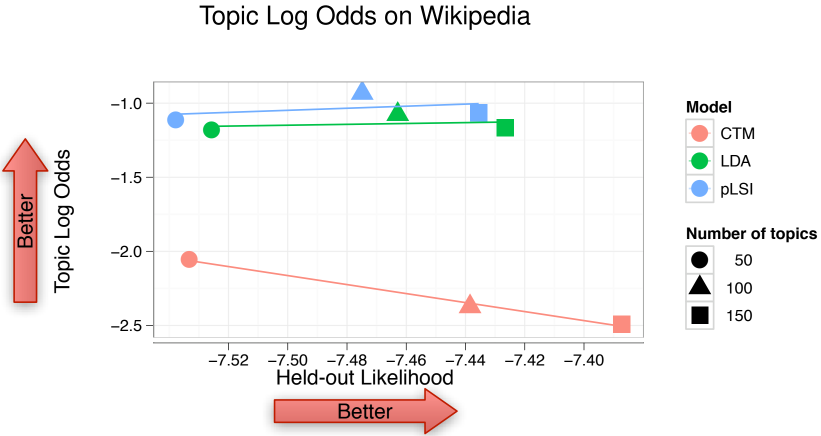
Model Precision: percentage of correct intruders found

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
- If you care about prediction, likelihood is good
- If you care about a particular task, measure that

Evaluation