Reading Tea Leaves: How Humans Interpret Topic Models

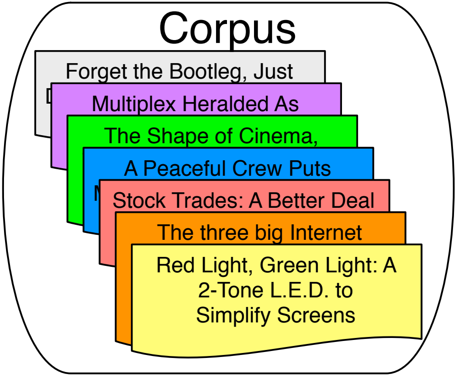
Jonathan Chang Jordan Boyd-Graber
Sean Gerrish Chong Wang David M. Blei

NIPS 2009
Dec 9th, 2009

PRINCETON UNIVERSITY

Topic Models in a Nutshell

From an **input corpus** $\rightarrow$ words to topics

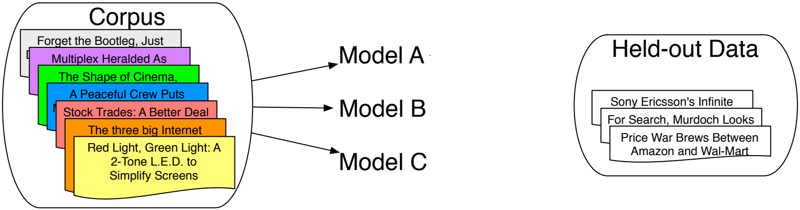


Topic Models in a Nutshell

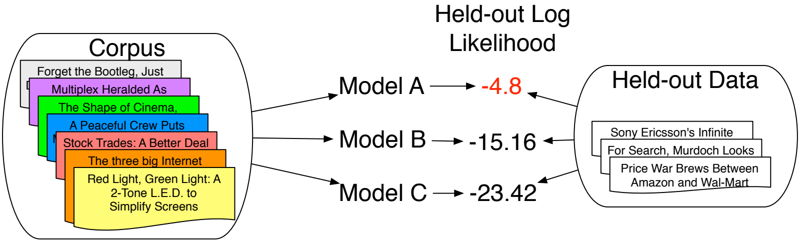
From an input corpus → **words to topics**

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

Evaluation


Corpus
Forget the Bootleg, Just
Multiplex Heralded As
The Shape of Cinema,
A Peaceful Crew Puts
Stock Trades: A Better Deal
The three big Internet
Red Light, Green Light: A 2-Tone L.E.D. to Simplify Screens
Model A
Model B
Model C
Held-out Data
Sony Ericsson's Infinite
For Search, Murdoch Looks
Price War Brews Between Amazon and Wal-Mart

Evaluation



Measures predictive power, not latent structure

Qualitative Evaluation of the Latent Space

"segment 1"	"segment 2"	"matrix 1"	"matrix 2"	"line 1"	"line 2"	"power 1"	power 2"
imag	speaker	robust	manufactur	constraint	alpha	POWER	load
SEGMENT	speech	MATRIX	cell	LINE	redshift	spectrum	memori
texture	recogni	eigenvalu	part	match	LINE	omega	vlsi
color	signal	uncertainti	MATRIX	locat	galaxi	mpc	POWER
tissue	train	plane	cellular	imag	quasar	hsup	systolic
brain	hmm	linear	famili	geometr	absorp	larg	input
slice	source	condition	design	impos	high	redshift	complex
cluster	speakerind.	perturb	machinepart	segment	ssup	galaxi	arrai
mri	SEGMENT	root	format	fundament	densiti	standard	present
volume	sound	suffici	group	recogn	veloc	model	implement

Figure 3: Eight selected factors from a 128 factor decomposition. The displayed word stems are the 10 most probable words in the class-conditional distribution $P(w|z)$, from top to bottom in descending order.

[Hofmann, 1999]

Qualitative Evaluation of the Latent Space

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

[Blei et al., 2003]

Qualitative Evaluation of the Latent Space

DA	centralbank europæiske ecb s lån centralbanks
DE	zentralbank ezb bank europäischen investitionsbank darlehen
EL	τράπεζα τράπεζας κεντρική εκτ κεντρικής τράπεζες
EN	**bank central ecb banks european monetary**
ES	banco central europeo bce bancos centrales
FI	keskuspankin ekp n euroopan keskuspankki eip
FR	banque centrale bce européenne banques monétaire
IT	banca centrale bce europea banche prestiti
NL	bank centrale ecb europese banken leningen
PT	banco central europeu bce bancos empréstimos
SV	centralbanken europeiska ecb centralbankens s lån

[Mimno et al., 2009]

Qualitative Evaluation of the Latent Space

(a) Topic labeled as SSL

Keyword	Probability
ssl	0.373722
expr	0.042501
init	0.033207
engine	0.026447
var	0.022222
ctx	0.023067
ptemp	0.017153
mctx	0.013773
lookup	0.012083
modssl	0.011238
ca	0.009548

(b) Topic labeled as Logging

Keyword	Probability
log	0.141733
request	.036017
mod	0.0311
config	0.029871
name	0.023725
headers	0.021266
autoindex	0.020037
format	0.017578
cmd	0.01512
header	0.013891
add	0.012661

Table 2: Sample Topics extracted from Apache source code

[Maskeri et al., 2008]

Qualitative Evaluation of the Latent Space

Probabilistic Models	model word probability set data number algorithm language corpus method
Prosody	prosodic speech pitch boundary prosody phrase boundaries accent repairs intonation
Semantic Roles*	semantic verb frame argument verbs role roles predicate arguments
Yale School Semantics	knowledge system semantic language concept representation information network concepts base
Sentiment	subjective opinion sentiment negative polarity positive wiebe reviews sentence opinions
Speech Recognition	speech recognition word system language data speaker error test spoken
Spell Correction	errors error correction spelling ocr correct corrections checker basque corrected detection
Statistical MT	english word alignment language source target sentence machine bilingual mt
Statistical Parsing	dependency parsing treebank parser tree parse head model al np
Summarization	sentence text evaluation document topic summary summarization human summaries score
Syntactic Structure	verb noun syntactic sentence phrase np subject structure case clause
TAG Grammars*	tree node trees nodes derivation tag root figure adjoining grammar
Unification	feature structure grammar lexical constraints unification constraint type structures rule
WSD*	word senses wordnet disambiguation lexical semantic context similarity dictionary
Word Segmentation	chinese word character segmentation corpus dictionary korean language table system
WordNet*	synset wordnet synsets hypernym ili wordnets hypernyms eurowordnet hyponym ewn wn

Table 2: Top 10 words for 43 of the topics. Starred topics are hand-seeded.

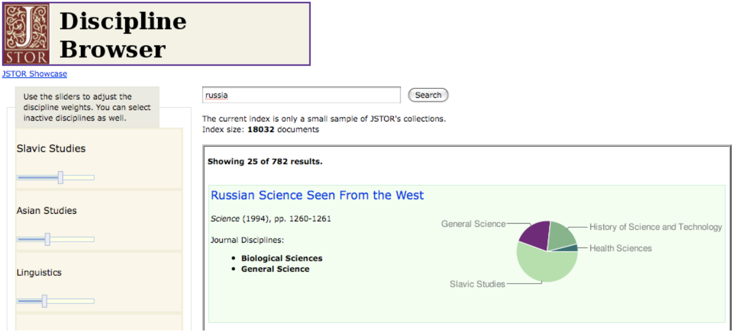
[Hall et al., 2008]

Rexa.info

Research • People × Connections

H-Index	Topic terms
20	book, book review, chapter, ed, edition, eds, handbook, introduction, nd, nd ed, nd edition, nd edn, pp, rd ed, rd edition, rd edn, review, revised edition, second edition, th ed, th edition, theory, third edition, vol
29	advances, current, current status, current trends, developments, future, future directions, new, new developments, new directions, past, past future, recent, recent advances, recent developments, recent progress, recent trends, research, review, state art, status, status quo, survey, trends
21	approaches, comparative, comparative statics, comparing, comparison, comparison approaches, comparison different, comparisons, differences, different, empirical comparison, experimental comparison, monozygotic twins, performance comparison, qualitative, qualitative quantitative, quantitative, quantitative comparison, quantitative qualitative, similarities differences, sp, sp sp, versus, vs
16	annotated bibliography, bibliography, book review, brief announcement, brief review, comments, discussion, introduction, literature, literature review, notes, overview, panel, panel session, peer review, review, review literature, review see, see, see comments, session, systematic review, unpublished, unpublished manuscript

Topics are shown to users during web search.



Users can refine queries through topics.

Key Points

1. "Reading Tea Leaves" alternative: measuring **interpretability**
2. Direct, quantitative human evaluation of latent space
3. Testing interpretability on different models and corpora
4. Disconnect with likelihood

Key Points

1. "Reading Tea Leaves" alternative: measuring **interpretability**
2. Direct, quantitative human evaluation of latent space
3. Testing interpretability on different models and corpora
4. Disconnect with likelihood

Evaluating Topic Interpretability

- Interpretability is a human judgement
- We will ask people directly
- Experiment Goals
  - Quick
  - Fun
  - Consistent

Evaluating Topic Interpretability

- Interpretability is a human judgement
- We will ask people directly
- Experiment Goals
  - Quick
  - Fun
  - Consistent
- We turn to Amazon Mechanical Turk
- Two tasks: Word Intrusion and Topic Intrusion

Task One: 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

Task One: Word Intrusion

1. Take the highest probability words from a topic

Original Topic

dog, cat, horse, pig, cow

Task One: 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

Task One: 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 Turkers to find the word that doesn't belong

Hypothesis

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

Task One: 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

Task One: Word Intrusion

1 / 10	Reveal additional response				
crash	accident	board	agency	tibetan	safety

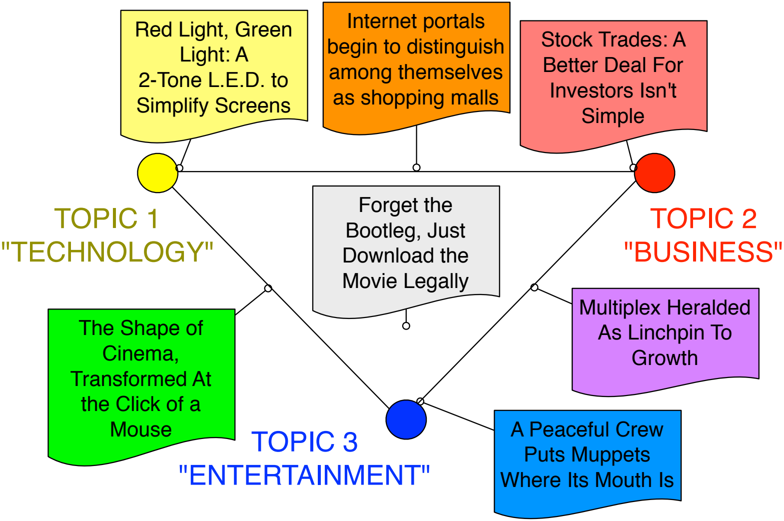
2 / 10					
commercial	network	television	advertising	viewer	layoff

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

Task Two: Topic Intrusion



Task Two: Topic Intrusion

1. Display document title and first 500 characters to Turkers
2. Show the three topics with highest probability and one topic chosen randomly
3. Have the user click on the the set of words that is out of place

Task Two: Topic Intrusion

1. Display document title and first 500 characters to Turkers
2. Show the three topics with highest probability and one topic chosen randomly
3. Have the user click on the the set of words that is out of place

Hypothesis

If the association of topics to a document is interpretable, users will consistently choose true intruding topic

Task Two: Topic Intrusion

1 / 10

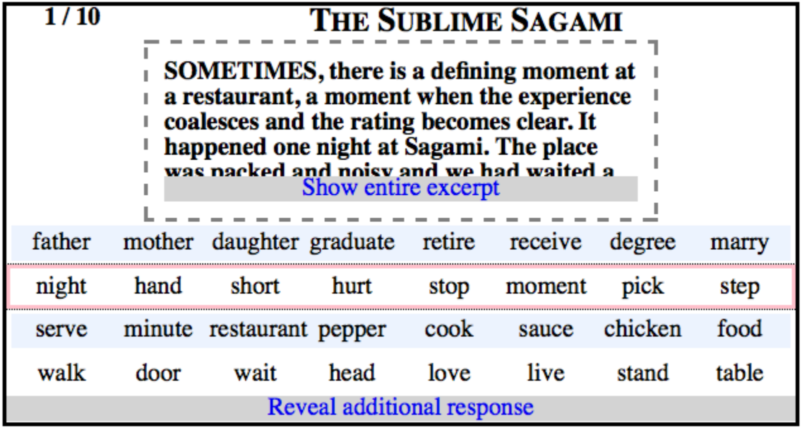
THE SUBLIME SAGAMI

SOMETIMES, there is a defining moment at a restaurant, a moment when the experience coalesces and the rating becomes clear. It happened one night at Sagami. The place was packed and noisy and we had waited a

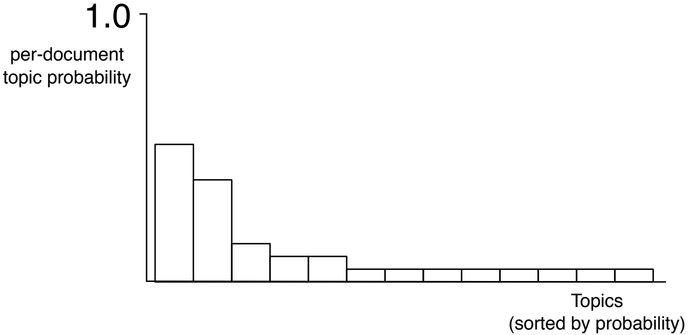
Show entire excerpt

father	mother	daughter	graduate	retire	receive	degree	marry
night	hand	short	hurt	stop	moment	pick	step
serve	minute	restaurant	pepper	cook	sauce	chicken	food
walk	door	wait	head	love	live	stand	table

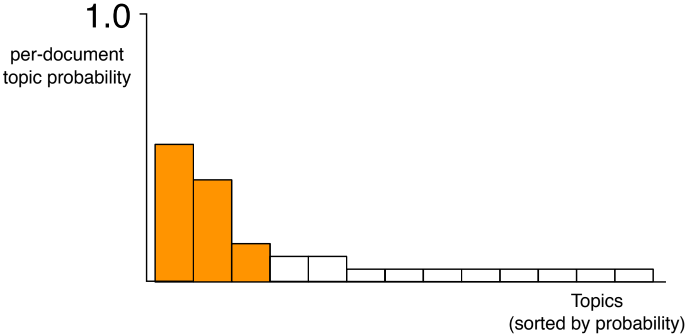
Task Two: Topic Intrusion



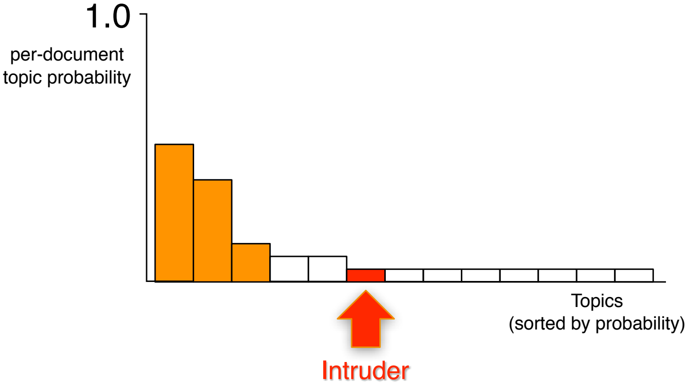
Task Two: Topic Intrusion



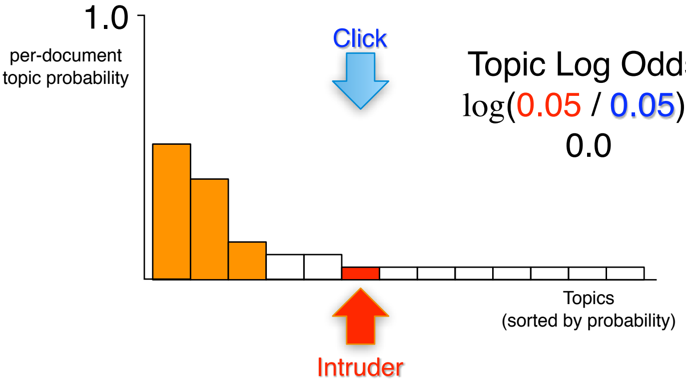
Task Two: Topic Intrusion



Task Two: Topic Intrusion



Task Two: Topic Intrusion

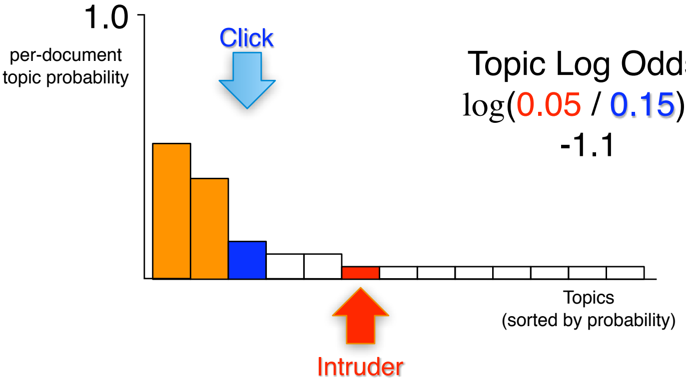


Topic Log Odds:

$\log(0.05\ /\ 0.05) =$

0.0

Task Two: Topic Intrusion

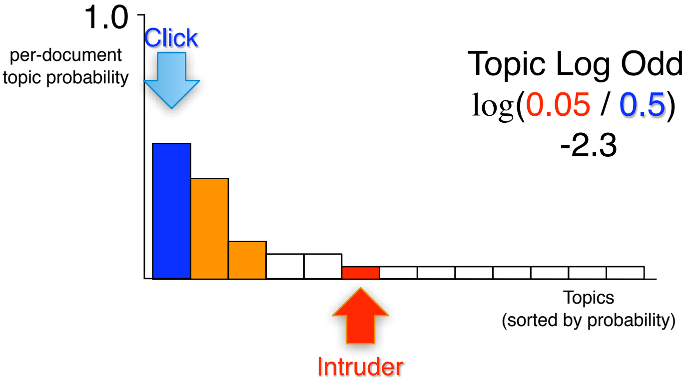


Topic Log Odds:

$\log($0.05 / 0.15$) =$

-1.1

Task Two: Topic Intrusion



Topic Log Odds:

log(0.05 / 0.5) =

-2.3

Three Topic Models

Different assumptions lead to different topic models

- Free parameter fit with smoothed EM (pLSI variant) [Hofmann, 1999]
- Dirichlet: latent Dirichlet allocation (LDA) [Blei et al., 2003]
- Normal with covariance: correlated topic model (CTM) [Blei and Lafferty, 2005]

Corpora

- 8477 articles
- 8269 types
- 1M tokens

- Sample of 10000 articles
- 15273 types
- 3M tokens

Corpora

The New York Times

- 8477 articles
- 8269 types
- 1M tokens

WIKIPEDIA
The Free Encyclopedia

- Sample of 10000 articles
- 15273 types
- 3M tokens

Corpora properties

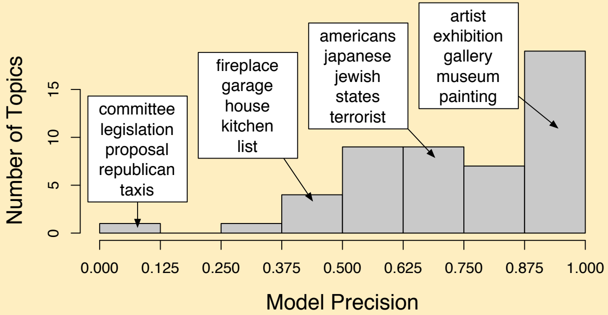
- Well structured (should begin with summary paragraph)
- Real-world
- Many different themes

Experiments

1. Fit pLSI, LDA, and CTM to both corpora
2. Each model had 50, 100, or 150 topics
3. 50 topics from each condition presented to 8 workers
4. 100 documents form each condition presented to 8 workers

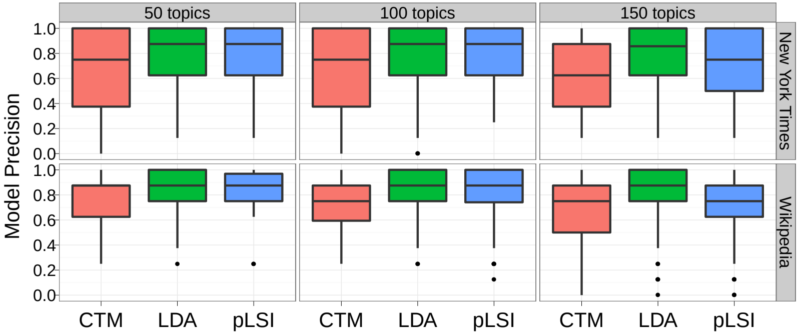
Word Intrusion: Which Topics are Interpretable?

New York Times, 50 LDA Topics



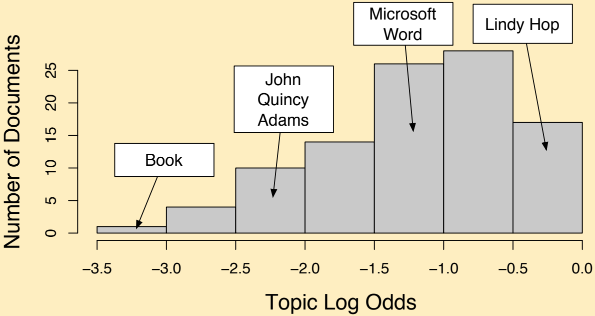
Model Precision: percentage of correct intruders found

Word intrusion: Models with Interpretable Topics

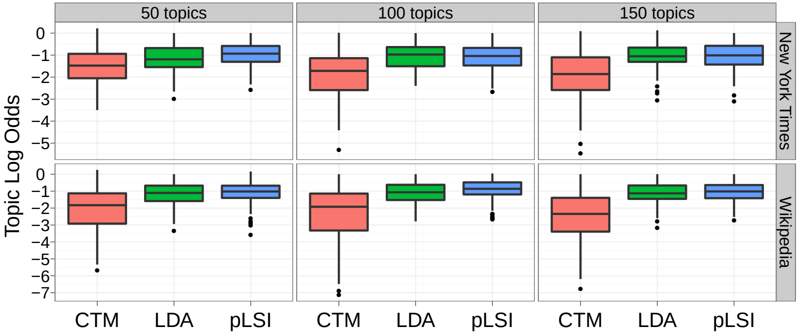


Which documents have clear topic associations?

Wikipedia, 50 LDA Topics



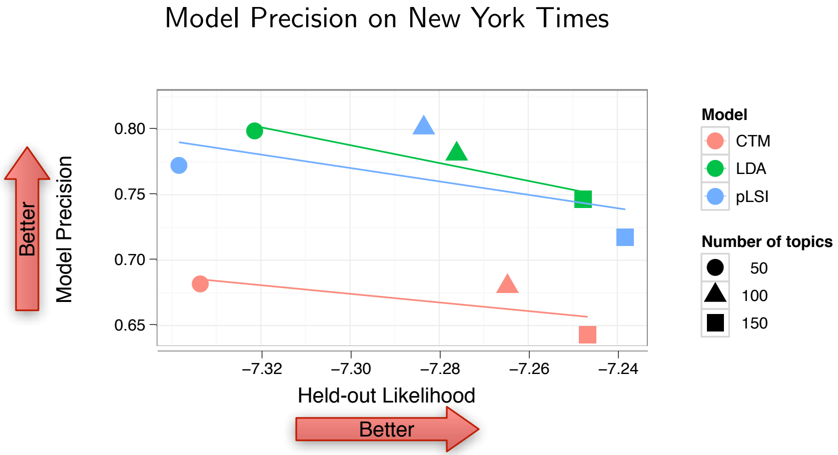
Which Models Produce Interpretable Topics



Held-out Likelihood

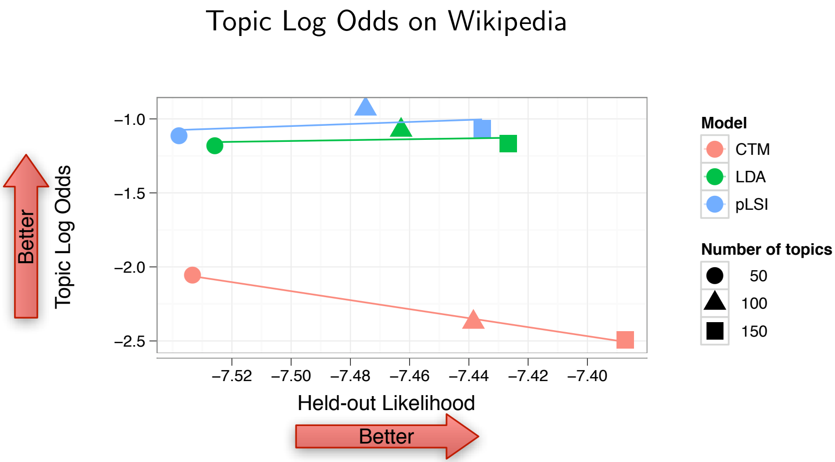
Corpus	Topics	pLSI	LDA	CTM
New York Times	50	-7.3384	**-7.3214**	-7.3335
	100	-7.2834	-7.2761	**-7.2647**
	150	**-7.2382**	-7.2477	-7.2467
Wikipedia	50	-7.5378	**-7.5257**	-7.5332
	100	-7.4748	-7.4629	**-7.4385**
	150	-7.4355	-7.4266	**-7.3872**

Interpretability and Likelihood



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

Interpretability and Likelihood



across models, higher likelihood $\neq$ higher interpretability

Conclusion

- Disconnect between evaluation and use
- Means of evaluating an *unsupervised* method
- For topic models, direct measurement of interpretability
- Surprising relationship between interpretability and likelihood
- Measure what you care about

Future Work

- Influence of inference techniques and hyperparmeters
- Investigate shape of likelihood / interpretability curve
- Model human intuition

Applications for Topic Models: Text and Beyond

7:30am - 6:30pm Friday
Westin: Callaghan

Blei, D., Ng, A., and Jordan, M. (2003).
Latent Dirichlet allocation.
JMLR, 3:993–1022.

Blei, D. M. and Lafferty, J. D. (2005).
Correlated topic models.
In *NIPS*.

Hall, D., Jurafsky, D., and Manning, C. D. (2008).
Studying the history of ideas using topic models.
In *EMNLP*.

Hofmann, T. (1999).
Probabilistic latent semantic analysis.
In *UAI*.

Maskeri, G., Sarkar, S., and Heafield, K. (2008).
Mining business topics in source code using latent dirichlet allocation.
In *ISEC '08: Proceedings of the 1st conference on India software engineering conference*, pages 113–120, New York, NY, USA. ACM.

Mimno, D., Wallach, H., Yao, L., Naradowsky, J., and McCallum, A. (2009).
Polylingual topic models.
In *Snowbird Learning Workshop*. Clearwater, FL.