Reading Tea Leaves: How Humans Interpret Topic Models

Simón Roca

UC3M

19 de marzo de 2019

Outline

- Introduction
- Topic Models
- Likelihood
- Human Evaluation
- Experiments
- Results
- Discussion

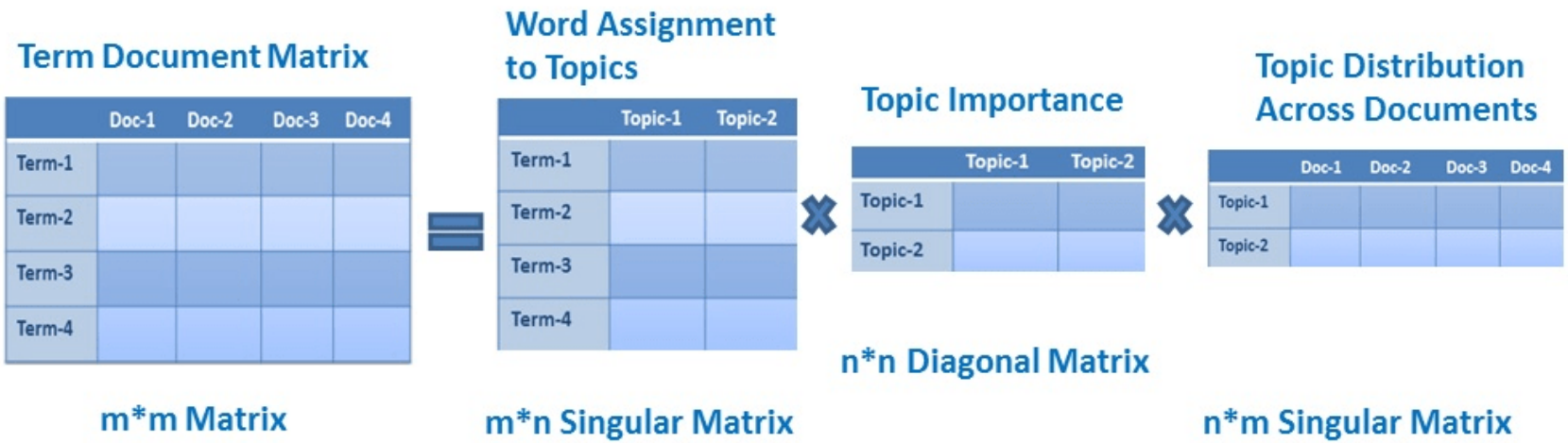
Introductiom

Written by
Chang, Boyd-Graber, Gerrish, Wang and **David Blei**
NIPS 2009.

- First human experiment on evaluating Topic Models.

- 'Good perplexity $\neq$ Good interpretability'

Topic Models: LSA



Documents: DxV matrix.
Singular Value Decomposition.

Topic Models: pLSA

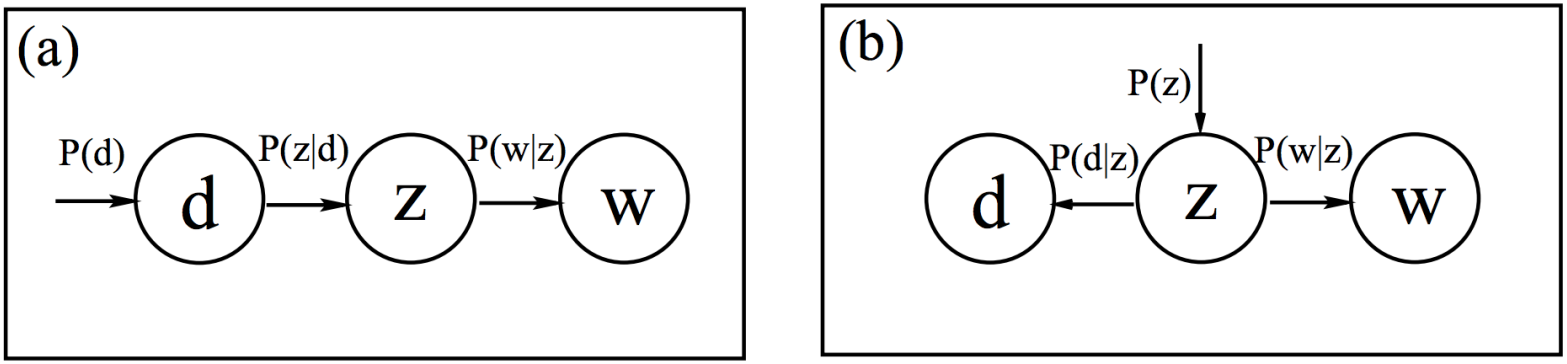
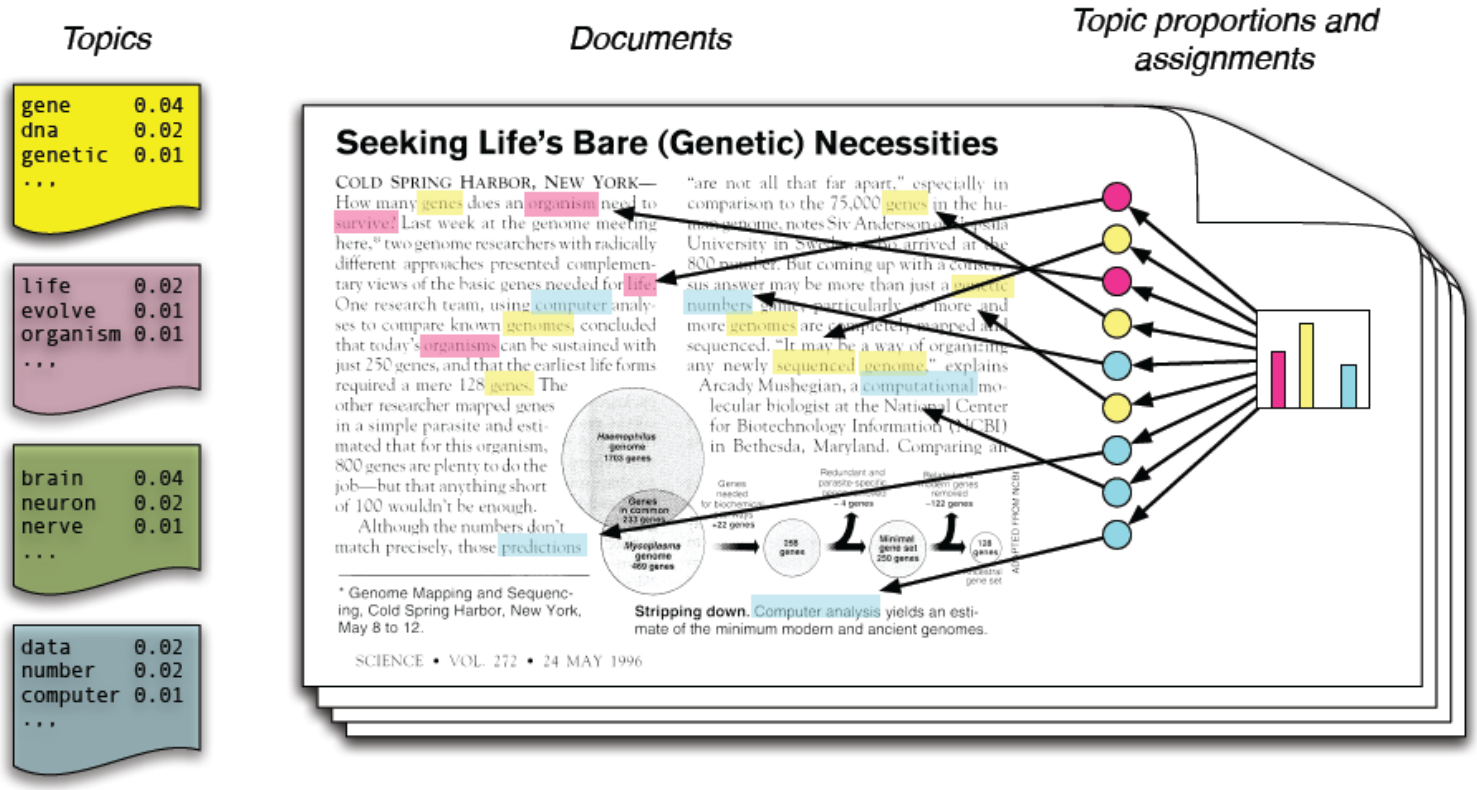


Figure 1: Graphical model representation of the aspect model in the asymmetric (a) and symmetric (b) parameterization.

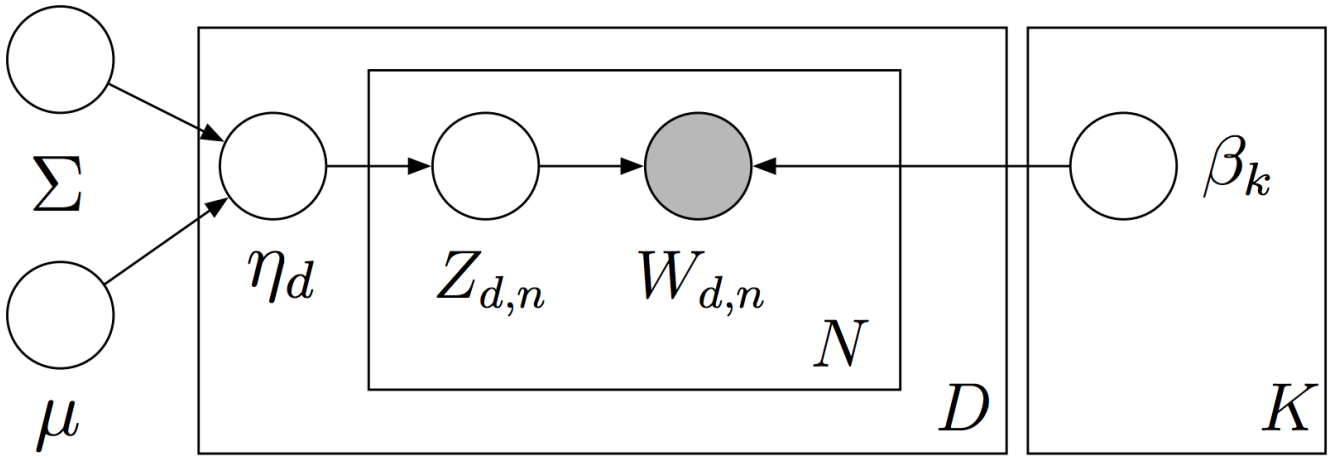
Mixture Model, solved by EM.

Topic Models: LDA



Dirichlet Prior on Documents-Topics and Topics-Words matrices.

Topic Models: CTM



Logistic-Normal prior on Topic proportions.

Likelihood

Convergence and evaluation.

$$L(w) = \log p(w|\Theta, \alpha) = \sum_d \log p(w_d|\Theta, \alpha) \tag{1}$$

$$Perplexity = exp(-\frac{L(w)}{\text{count of tokens}}) \tag{2}$$

For LDA, Perplexity is intractable, and approximations are needed.

Likelihood

Predictive probability

$$p(w^{test}) = \prod_{ij} \sum_{k} \bar{\theta}_{jk} \bar{\phi}_{k w_{ij}^{test}} \tag{3}$$

Human eval.: Word Intrusion

Q: What of the following words is the topic intruder?

- Floppy, Alphabet, Computer, Processor, Memory, Disk

- Linguistics, Actor, Film, Comedy, Director, Movie

Human eval.: Topic Intrusion

Douglas Richard Hofstadter (born February 15, 1945, in New York, New York) is an American academic whose research focuses on consciousness, thinking and creativity. He is best known for ", first published in...

Q: What of the following topics is the intruder?

Human eval.: Topic Intrusion

Q: What of the following topics is the intruder?

- Student, School, Study, Education, Research, University, Science, Learn
- Human, Life, Scientific, Science, Scientist, Experiment, Work, Idea
- Play, Role, Good, Actor, Star, Career, Show, Performance
- Write, Work, Book, Publish, Life, Friend, Influence, Father

Experiments

Corpus:

Corpus	N. Documents	Vocab	Tokens
New York Times	8447	8269	1000000
Wikipedia	10000	15273	3000000

Models (for all of them, (50,100,150) topics, β inferred, $\alpha = 1$):

- pLSA
- LDA
- CTM

Experiments

Evaluation Metrics:

- Predictive Log Likelihood/ Predictive rank.
- Model precision (word intrusion): $MP_k^m = \mathbb{1}(i_{k,s}^m = w_k^m)/S$
- Topic log odds (topic intrusion):
  $TLO_d^m = (\sum_s \log \hat{\theta}^m_{d,j^m_{d,*}} - \log \hat{\theta}^m_{d,j^m_{d,s}})/S$

Results

Table 1: Two predictive metrics: predictive log likelihood/predictive rank. Consistent with values reported in the literature, CTM generally performs the best, followed by LDA, then pLSI. The bold numbers indicate the best performance in each row.

CORPUS	TOPICS	LDA	CTM	PLSI
NEW YORK TIMES	50	**-7.3214 / 784.38**	-7.3335 / 788.58	-7.3384 / 796.43
	100	-7.2761 / 778.24	**-7.2647 / 762.16**	-7.2834 / 785.05
	150	-7.2477 / 777.32	-7.2467 / **755.55**	**-7.2382** / 770.36
WIKIPEDIA	50	**-7.5257** / 961.86	-7.5332 / **936.58**	-7.5378 / 975.88
	100	-7.4629 / 935.53	**-7.4385 / 880.30**	-7.4748 / 951.78
	150	-7.4266 / 929.76	**-7.3872 / 852.46**	-7.4355 / 945.29

Results

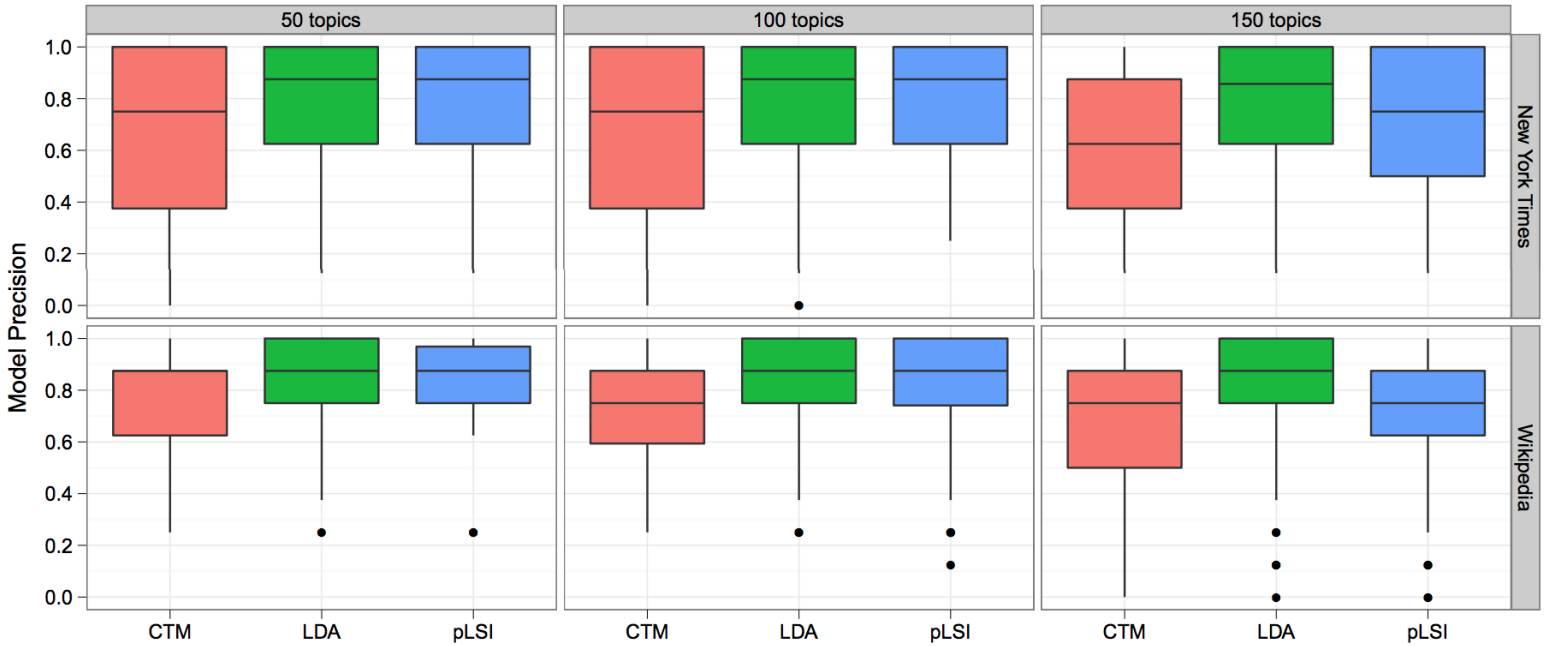


Figure 3: The model precision (Equation 1) for the three models on two corpora. Higher is better. Surprisingly, although CTM generally achieves a better predictive likelihood than the other models (Table 1), the topics it infers fare worst when evaluated against human judgments.

Results

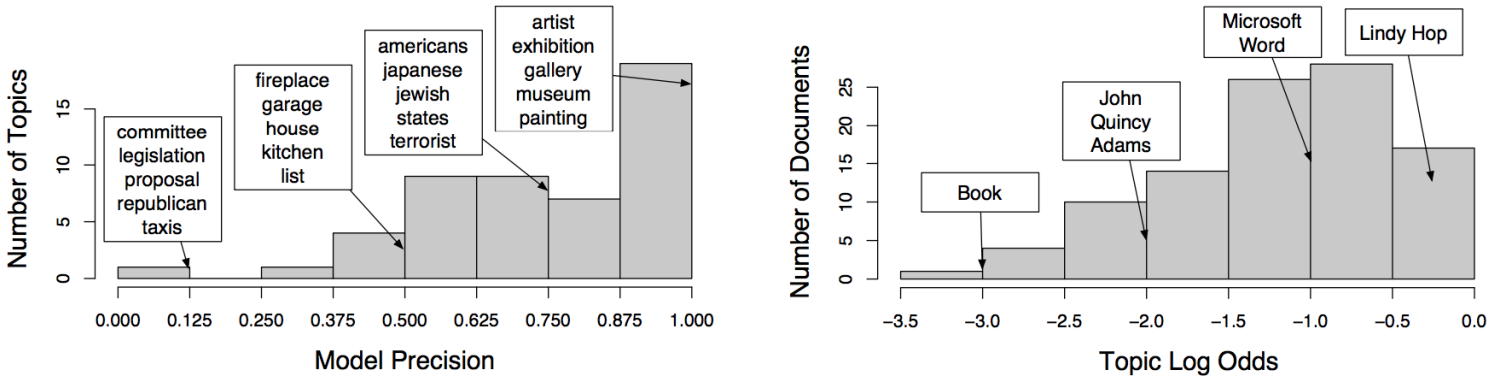


Figure 4: A histogram of the model precisions on the New York Times corpus (left) and topic log odds on the Wikipedia corpus (right) evaluated for the fifty topic LDA model. On the left, example topics are shown for several bins; the topics in bins with higher model precision evince a more coherent theme. On the right, example document titles are shown for several bins; documents with higher topic log odds can be more easily decomposed as a mixture of topics.

Results

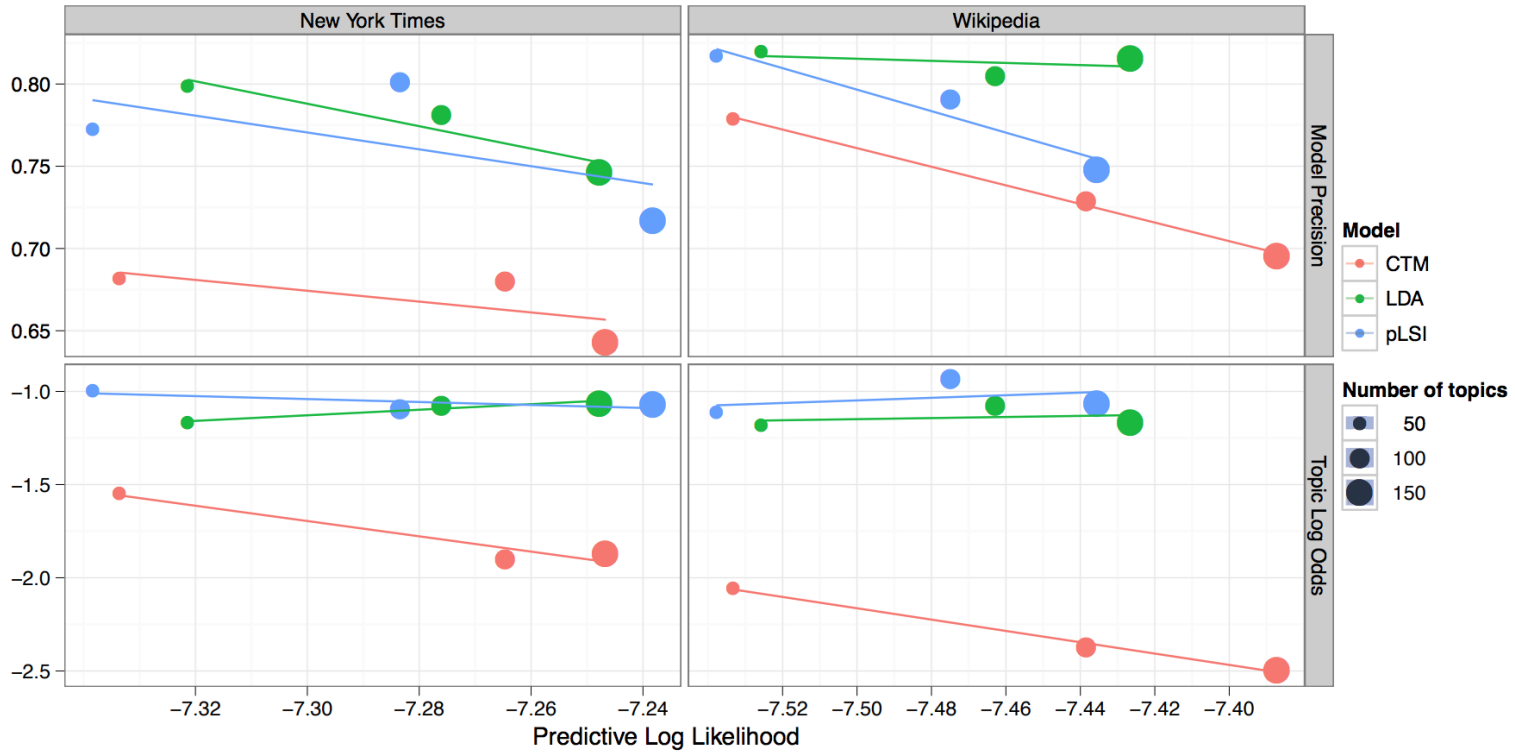


Figure 5: A scatter plot of model precision (top row) and topic log odds (bottom row) vs. predictive log likelihood. Each point is colored by model and sized according to the number of topics used to fit the model. Each model is accompanied by a regression line. Increasing likelihood does not increase the agreement between human subjects and the model for either task (as shown by the downward-sloping regression lines).

Results

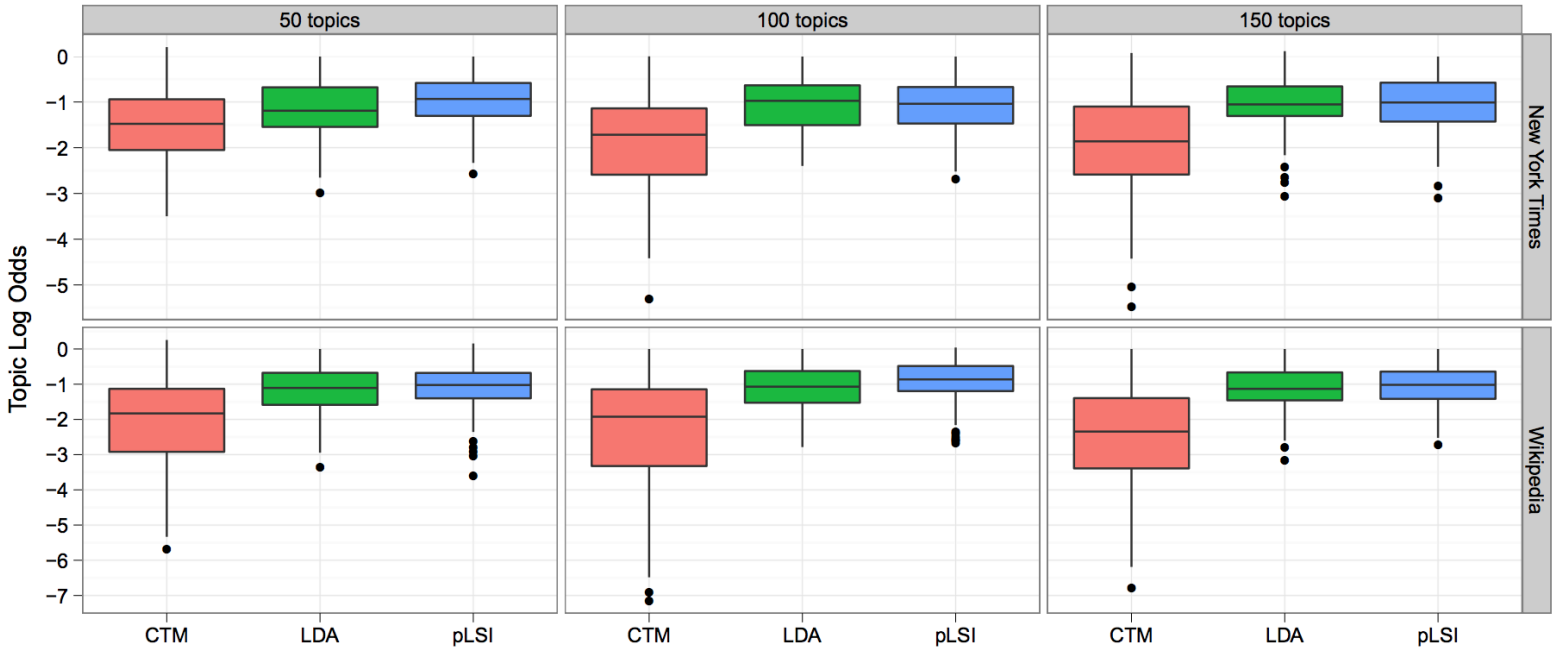


Figure 6: The topic log odds (Equation 2) for the three models on two corpora. Higher is better. Although CTM generally achieves a better predictive likelihood than the other models (Table 1), the topics it infers fare worst when evaluated against human judgments.

Discussion

- First human evaluation.
- New model selection methodology.
- Suggestion: look for real world check, instead of maximizing likelihoods.