# Supplementary Material for the manuscript: "A new evaluation framework for topic modeling algorithms based on synthetic corpora"

**Hanyu Shi** [1,†], **Martin Gerlach**[1,†], **Isabel Diersen**[1], **Doug Downey**[2], **Luís A. N. Amaral**[1,*]
[1]Department of Chemical and Biological Engineering,
[2]Department of Electrical Engineering and Computer Science
Northwestern University, Evanston, Illinois 60208, USA
[*]amaral@northwestern.edu, [†]Both authors contributed equally to this manuscript.

## S1 Real world corpora

We use 4 datasets and apply different filtering strategies yielding 8 different real world corpora each with $D$ documents, $N$ tokens, and $C$ category labels, as shown in Table S1.

## S2 Topic model algorithms

We use the following topic modeling algorithms (shown in Table S2) with default parameter settings unless stated otherwise.

LDA requires hyperparameter values for the topic-document distribution, a $K_a$-dimensional vector $\vec{\alpha} = (\alpha_j)_{j=1,\ldots,K_a}$ (where $K_a$ is the assumed number of topics), and the word-topic distribution, a $V$-dimensional vector $\vec{\beta} = (\beta_w)_{w=1,\ldots,V}$ (where $V$ is the size of the vocabulary). We assume symmetric priors, i.e. $\alpha_j = \alpha$ and $\beta_w = \beta$, such that the hyperparameters are fully determined by the scalar parameters $\alpha$ and $\beta$. For LDAVB we use the default values of the gensim implementations. For LDAGS we use the default values of the gensim-wrapper of the mallet implementation.

## S3 Usage of synthetic corpora in previous studies

Different types of synthetic corpora in previous studies are given in Table S3. As we can see, in previous research a large portion of synthetic corpora are generated directly from LDA.

## S4 Document classification

In practical applications, topic models are often used to find documents of similar topical content in an unsupervised fashion. In this spirit we quantify the performance of topic models by checking how much the inferred topic distributions reflect the assignment of human-labeled categories in real world corpora.

More specifically, we fit a topic model to the entire corpus and obtain the inferred topic-distribution of each document, $P(t|d)$. Identifying each topic with a category, we predict the category-membership of each document, $s_d$, from the topic with maximum probability (Xie and Xing, 2013)

$$s_d = \arg\max_t P(t|d) \tag{S1}$$

Comparing this with the given metadata-category, $r_d$, we can construct a confusion matrix

$$p_{s,r} \equiv \frac{1}{D}\sum_d \delta_{s,s_d}\cdot\delta_{r,r_d}, \tag{S2}$$

which yields the fraction of documents that have metadata-category $r$ and predicted category $s$. With confusion matrix $p_{s,r}$, we can quantify the performance of the topic model in the classification task using the normalized mutual information.

## S5 Generation of synthetic benchmark corpora

With the distributions $P(w|t)$ and $P(t|d)$ described in the main text Sec. 3.1, we can generate the synthetic benchmark corpora from distributions $P(w|t)$ and $P(t|d)$ according to the generative process. One major advantage of our work is that our approach allows for the inclusion of many realistic features, such as Zipfian distribution, stopword, and burstiness, as described below.

**Zipfian word-frequency distribution.** One of the most well-known statistical laws in language is the so-called Zipf's law (Zipf, 1936), which states that the

Table S1: **Details for Real-world Corpora.**

| Dataset | Variation | Filtering | Characteristics |
|---|---|---|---|
| Reuters-21578 (Hettich and Bay, 1999) | 1 | Only documents from the 10 largest categories | $D$=7,518; $N$=775,771; $C$=10 |
| | 2 | Only documents from categories with more than 10 documents | $D$=8,559; $N$=936,004; $C$=41 |
| RCV1 (Lewis et al., 2004) | 1 | Only documents with one category label; subsample 10% of documents from the largest category | $D$=3,070; $N$=574,249; $C$=4 |
| | 2 | Only documents with two category labels; subsample 10% of documents | $D$=20,474; $N$=2,917,939; $C$=54 |
| Web of Science (Lancichinetti, 2016) | 1 | None | $D$=40,526; $N$=3,828,735; $C$=7 |
| | 2 | Only keep the first 20 tokens of each document | $D$=40,526; $N$=808,672; $C$=7 |
| 20 News Group (Cachopo, 2007) | 1 | Remove all words with less than 3 characters | $D$=18,803; $N$=3,831,559; $C$=20 |
| | 2 | Same as in variation 1 and remove all stopwords from list given in Ref. (McCallum, 2002) | $D$=18,799; $N$=2,654,710, $C$=20; |

frequency $f$ of the $r$-th most frequent word is given by a power-law with exponent $\gamma > 1$:

$$f(r) \propto r^{-\gamma} \tag{S3}$$

We incorporate a Zipfian distribution by accommodating any global word-frequency distribution $P(w)$ as an average over all topics, i.e. $P(w) = \sum_t P(w,t) = \sum_t P(w|t)P(t)$ where $P(t)$ is the size of topic $t$.

**Stopwords.** An important statistical property of real texts is the generic appearance of stopwords. While there is no general agreed-upon definition, in the context of topic modeling this usually refers to very common words (such as "the," "and," etc.) which are considered not informative in inferring topical structure related to semantics. In practice, these words are typically removed from a corpus using a pre-specified list of stopwords, however, there exist differing opinions on the effect of stopwords in the result of the quality of inferred topic models (Zaman et al., 2011; Schofield et al., 2017).

We model stopwords as words which have the same probability of appearance in any topic, i.e. $P(w|t) = P(w)$. Varying the fraction of stopwords (of unique words in the vocabulary) by a parameter $P_s \in [0,1]$ allows us to investigate the robustness of a topic model with respect to these non-informative words.

**Burstiness.** The phenomenon of burstiness refers to non-stationarity in the usage of words (Katz, 1996; Altmann et al., 2009), that is a word is more likely to occur in a text after its first occurrence.

We incorporate burstiness following approaches proposed in Refs. (Madsen et al., 2005; Doyle and Elkan, 2009) using Dirichlet-distributions. Given a topic $t$, instead of drawing from a fixed word-topic distribution $P(w|t)$, we obtain a different word-topic distribution in each document which is drawn from a $V$-dimensional Dirichlet distribution with concentration parameter $a_c$, i.e. $P_d(w|t) \sim Dir_V(a_c \cdot P(w|t))$. This means that the smaller $a_c$ the more "bursty" the synthetic corpora. For example, in the limiting case $a_c \to 0$ ($a_c \to \infty$) the word-topic distribution in each document will contain only one word with non-zero probability (the original global word-topic distribution from the non-bursty case).

## S6 Supplementary figures

As an example in Fig. S1, consider two planted classes in the synthetic benchmark, where 50% of the tokens belong to each planted class, we obtain different values $\hat{I}$ depending on the inferred structure: (*1*) If all tokens are correctly assigned into two inferred classes yielding a perfectly diagonal confusion matrix $p_{t,t'}$, this leads

Table S2: **Details for topic modeling algorithms.**

| Topic Model | Implementation | Default hyperparameter values |
|---|---|---|
| Gibbs Sampling LDA (LDAGS) (Griffiths and Steyvers, 2004) | Mallet (McCallum, 2002) | $\alpha = 5/K_a$, $\beta = 0.01$ |
| Variational Bayes LDA (LDAVB) (Blei et al., 2003) | gensim (Řehůřek and Sojka, 2010) | $\alpha = 1/K_a$, $\beta = 1/K_a$ |
| Hierarchical Dirichlet Processes (HDP) (Teh et al., 2006) | From work of Wang (2010) | n.a. |
| TopicMapping (TM) (Lancichinetti et al., 2015) | From work of Lancichinetti (2016) | n.a. |

to $I = log(2)$ and $\hat{I} = 1$; (*2*) In case one of the inferred classes gets split into two equal-sized classes, we get the same $I$, but a smaller value for $\hat{I}$; (*3*) If one of the smaller inferred classes is uninformative with respect to the planted classes, this leads to a further reduction of $I$ and $\hat{I}$; (*4*) If the tokens are just randomly assigned to two inferred classes, this yields $I = 0$ and $\hat{I} = 0$.

In Fig. S2 we show the resulting synthetic corpora for the random ($c = 0$), mixed ($c = 0.5$), and ordered ($c = 1$) case. While for $c = 0$ the topics are not distinguishable in the ground truth topic distributions, $c > 0$ yields a block-diagonal structure (Fig. S2A). Looking at the empirically observed corpus in the form of the counts $n(d, w)$, i.e., the number of times word $w$ appears in document $d$, words are distributed randomly across all documents for $c = 0$, while increasing $c$ leads to a higher concentration of words in certain documents reflecting the increasing degree of structure (Fig. S2B).

For algorithms such as LDA one needs to specify the number of topics for fitting the topic model. While in the synthetic corpus we know the true number of topics, in practice, this value is unknown. Therefore, we investigate the effect of over- and under-fitting by varying the assumed number of topics, $K_a$, for LDA in a synthetic corpus with 10 planted topics (Fig. S3).

Fig. S4 compares the reproducibility of HDP and TM. There are 10 repetitions for each data points. In each repetition, only one synthetic benchmark corpus is generated. A topic model will be run on this corpus twice. The inferred token topics form the two experiments of the topic model will be compared.

In Fig. S5, we show the planted and inferred $P(t|d)$ and $P(w|t)$ by the implementation of different algorithms for LDA, using $K_a = 100$ as the assumed number of topics.

In Fig. S6 and Fig. S7, we confirm that the solutions for LDA obtained in Fig. 4 and Fig. S5 have converged with respect to the number of iterations in each topic model.

In Fig. S8 we compare the planted and inferred topic distributions $P(t|d)$ and $P(w|t)$ and show how they provide a more detailed view on the performance of a topic model than obtained from document classification tasks.

In Fig. S9 we show that the results from Fig. 5 (B, C) investigating the effect of stopwords on the performance of topic models in synthetic corpora remain qualitatively similar when varying the parameter of the degree of structure, $c$.

In Fig. S10 we show differences between the planted and inferred topic distributions $P(t|d)$ and $P(w|t)$ for synthetic corpora in the case of stopwords.

In Fig. S11 we show that the results from Fig. 5 (E, F) investigating the effect of document length on the performance of topic models in synthetic corpora remain qualitatively similar when varying the parameter of the degree of structure, $c$.

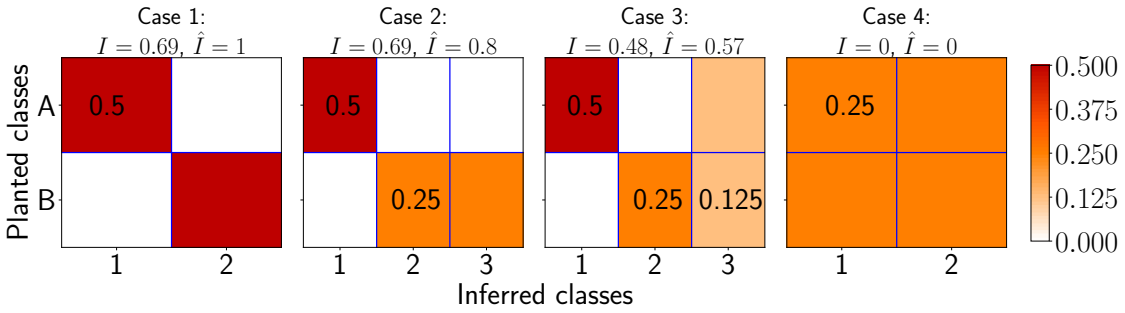


Figure S1: **Quantifying overlap using the normalized mutual information.** Unnormalized, $I$, and normalized mutual information, $\hat{I}$, for different examples of confusion matrices $p_{t,t'}$ (cases 1-4). Number indicate the values of the confusion matrix according to color.

Table S3: **Usage of Synthetic Corpora in Previous studies.**

| Reference | Synthetic Corpora | Corresponding evaluation metric |
|---|---|---|
| Mukherjee and Blei (2009) | Generated from LDA | Likelihood; Variational free energy |
| Newman et al. (2009) | Generated from LDA | L1-norm between true and inferred word-topic distribution |
| Wallach et al. (2009) | Generated from LDA | Held-out likelihood |
| Mimno and Blei (2011) | Generated from LDA | Check hypothesis that words and documents are independent given the topic using mutual information |
| Taddy (2012) | Generated from LDA | Compare entries (and residuals) in $\theta_k$ (defined as the distribution over words for each topic) between true topics and inferred topics |
| Tang et al. (2014) | Generated from LDA | Posterior contraction analysis of the topic polytope |
| Hsu and Poupart (2016) | Generated from LDA | Use the synthetic corpora to test inferred number of topics |
| Minka and Lafferty (2002) | Multinomial with 5 equiprobable words | Likelihood; Classification |
| Griffiths and Steyvers (2004) | Bar-data (5x5 grid) | Visual comparison |
| Andrzejewski et al. (2009) | Small size synthetic corpora based on their proposed topic model (LDA with Dirichlet Forest Priors) | Visual inspection of the word-document matrix |
| AlSumait et al. (2009) | Small size synthetic corpora: 6 samples of 16 documents from three static equally weighted topic distributions. On average, the document size was 16 words. | Topic signficance score (similar to topic coherence) |
| Arora et al. (2013, 2016) | Semi-synthetic data (train parameters of a model on a real corpus, then use the model to generate synthetic data) | Training time; L1-error between true and inferred matrix A (defined as the word-topic matrix). |
| Lancichinetti et al. (2015) | Language data with non-overlapping topics | L1-norm between true and inferred $p(d\|t)$ |
| This work | A flexible framework that could include a range of topic structure and realistic features | Measure the overlap between the planted and the inferred topic labels on the token level |

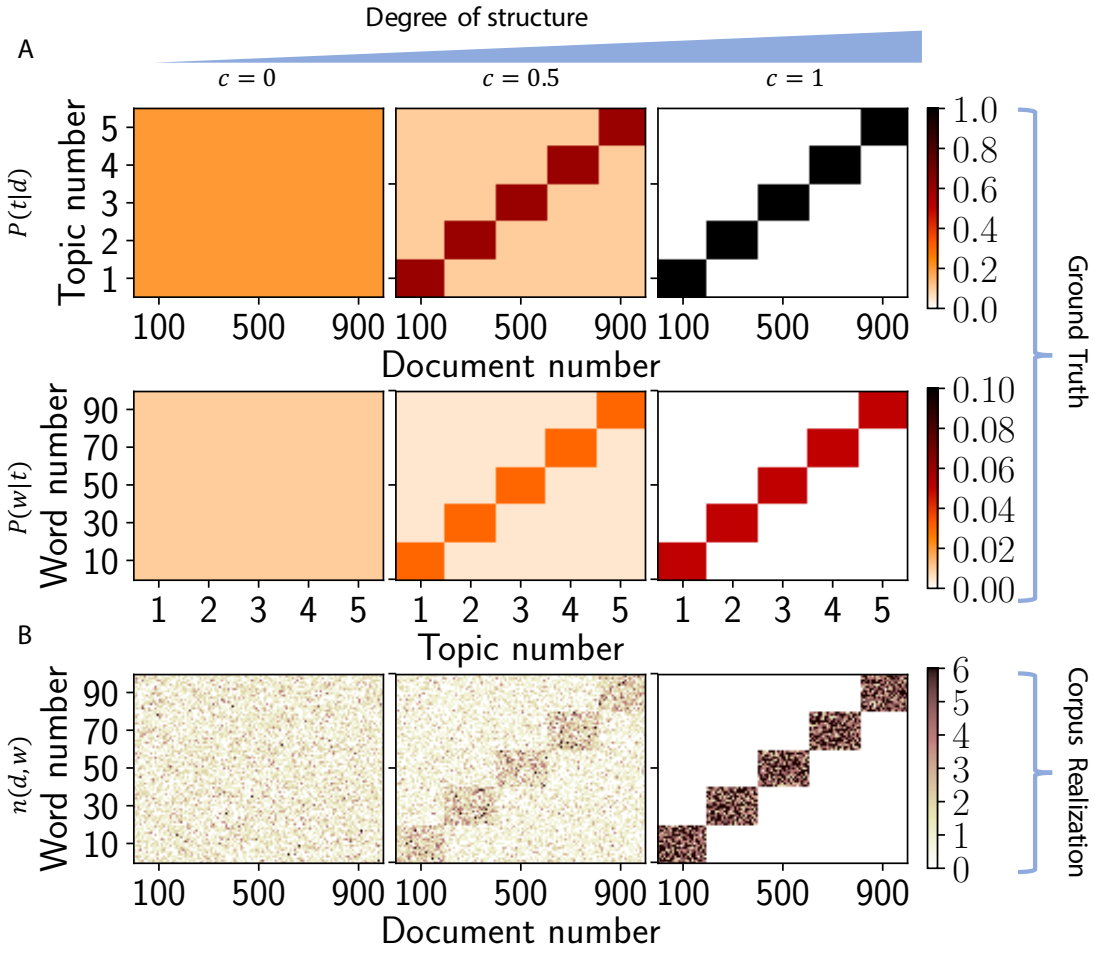


Figure S2: **Synthetic benchmarks with known ground truth from the generative process of topic models.** Three synthetic benchmark corpora with equal size but different degrees of structure $c \in \{0, 0.5, 1\}$ (left, middle, right column). **(A)** Ground truth topic distributions $P(t|d)$ and $P(w|t)$. **(B)** Resulting observable corpus showing the number of times word $w$ appears in document $d$, $n(d, w)$. For all panels, the number of topics is $K = 5$; there are $V = 100$ words in the vocabulary; and each corpus contains $D = 1,000$ document with length of $m_d = 100$.

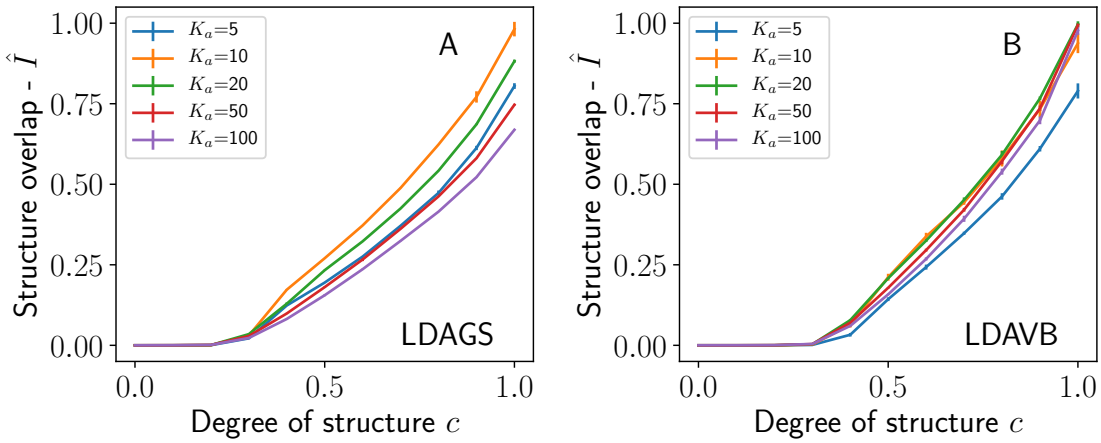


Figure S3: **Both overfitting and underfitting reduce the performance of LDA models.** Normalized mutual information, $\hat{I}$, between planted and inferred structure as a function of the structure parameter $c$ varying the assumed number of topics $K_a \in \{5, 10, 20, 50, 100\}$. **(A)** Gibbs Sampling LDA. **(B)** Variational Bayes LDA. For the synthetic benchmark corpora, we set the planted number of topics as $K = 10$, the vocabulary as $V = 1,000$, and the number of documents as $D = 10,000$ each with length $m_d = 100$. In the experiment we use different assumed numbers of topics (5, 10, 20, 50, 100) for LDA methods. The curves denote averages (and $\pm$ one standard deviation) over 10 realizations.

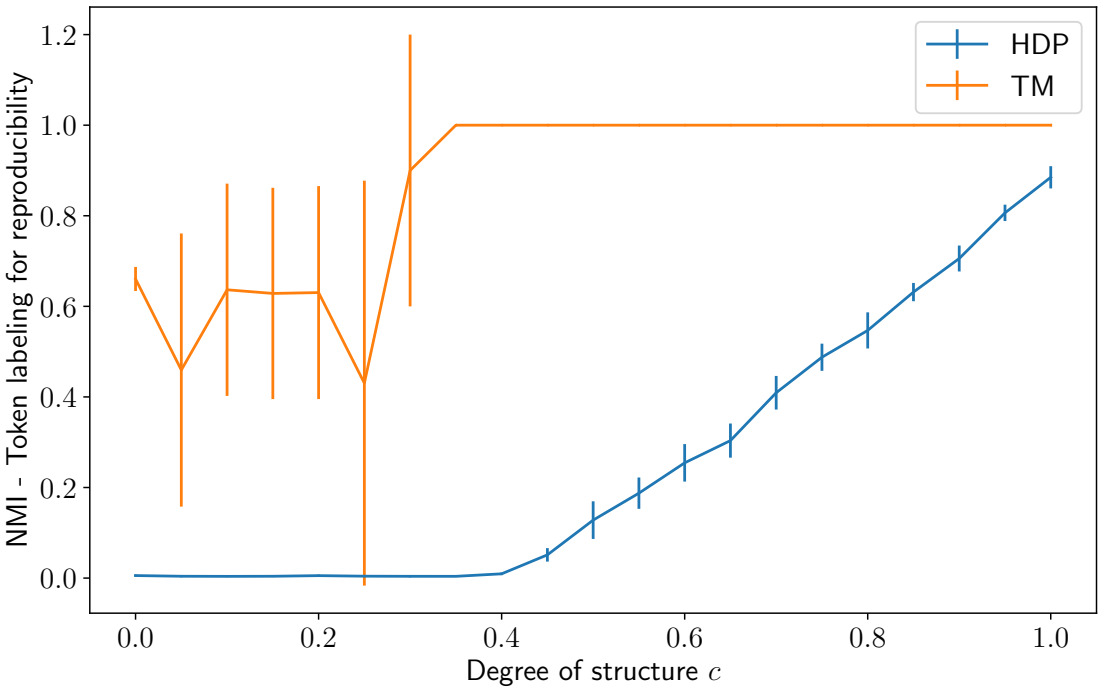


Figure S4: **Compare the reproducibility of two nonparametric topic modeling algorithms, HDP and TM, based on token labeling comparison.** Synthetic corpora were generated with $K = 10$ topics, $D = 10^4$ documents, document length $m = 100$, and vocabulary size $V = 10^3$. The lines (error bars) denote averages (one standard deviation) estimated from 10 realizations.

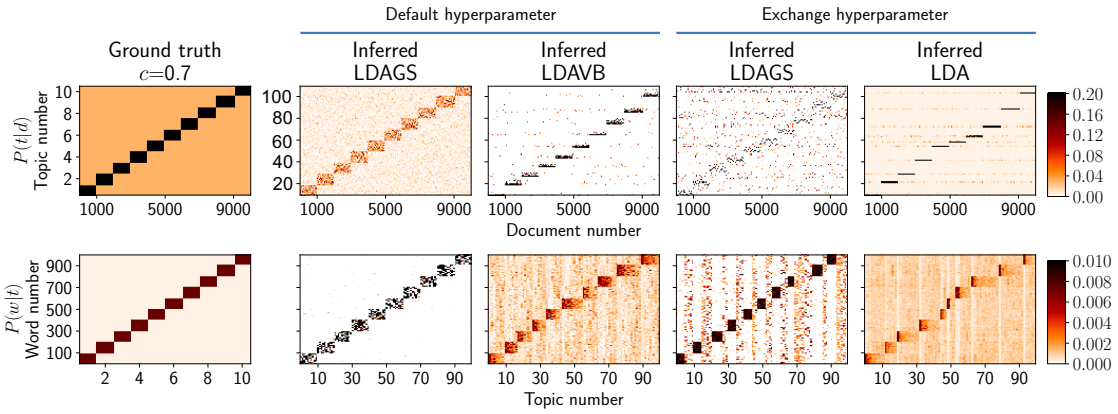


Figure S5: **Hyperparameters bias the inferred topic structure of different algorithms of LDA models.** Comparison of topic distributions $P(t|d)$ (top row) and $P(w|t)$ (bottom row) from the planted and inferred structure from LDAGS and LDAVB using two different sets of hyperparameters: original defaults as defined in each implementation (middle panels) and defaults from the other implementation, respectively (right panels). Same parameters as in Fig. 3 fixing $c = 0.7$ and using $K_a = 100$.

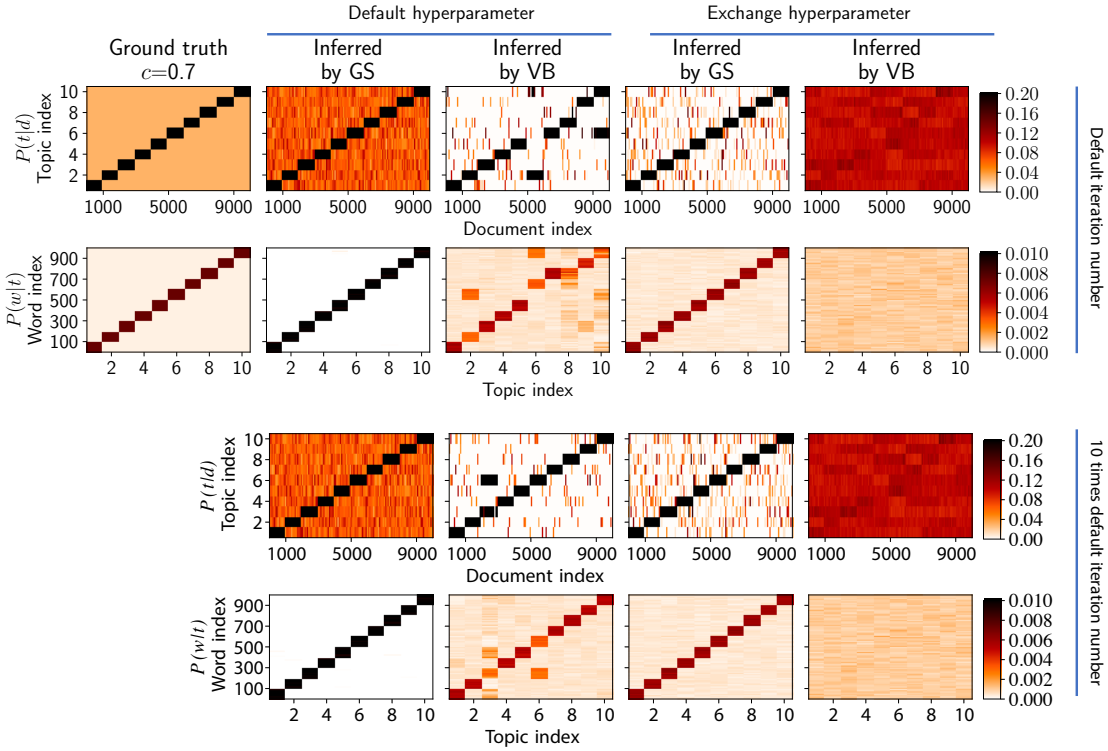


Figure S6: **Topic models converge with the default iteration setting for $K_a = 10$.** Inferred topic distributions $P(t|d)$ and $P(w|t)$ as in Fig. 4 for Gibbs Sampling LDA and Variational Bayes LDA with different hyperparameter settings comparing the case where we use the default number of iterations (top two columns) with the case where we increase the number of iterations 10-fold (bottom two columns).

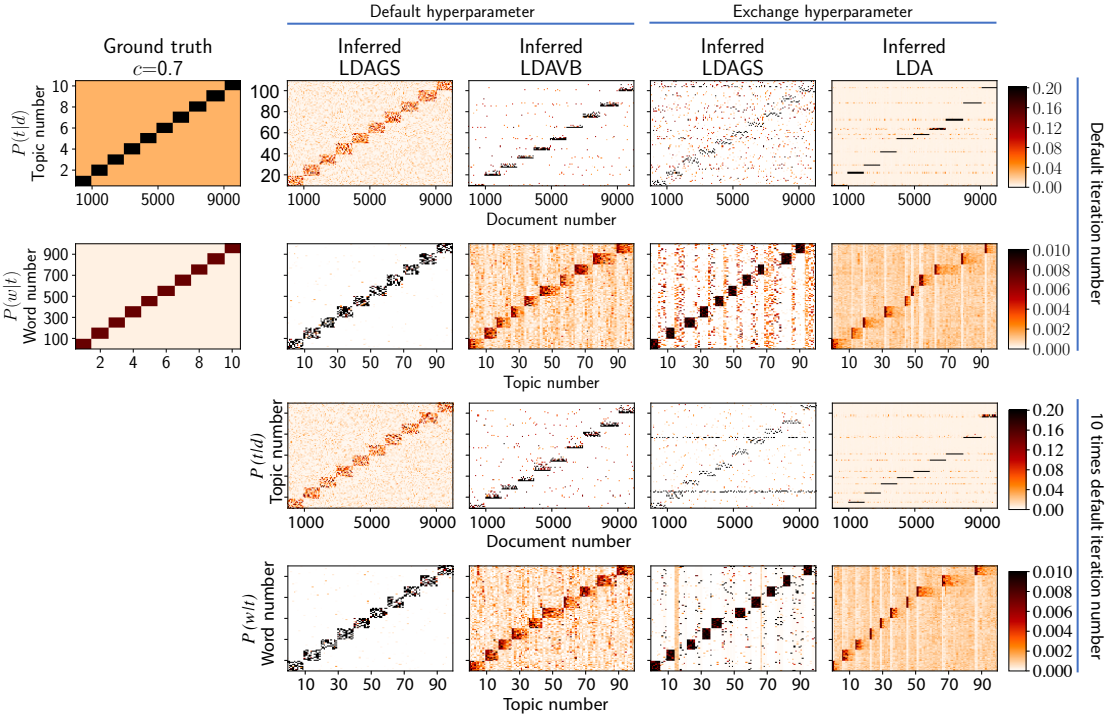


Figure S7: **Topic models converge with the default iteration setting for $K_a = 100$.** Inferred topic distributions $P(t|d)$ and $P(w|t)$ as in Fig. 4 for Gibbs Sampling LDA and Variational Bayes LDA with different hyperparameter settings comparing the case where we use the default number of iterations (top two columns) with the case where we increase the number of iterations 10-fold (bottom two columns).

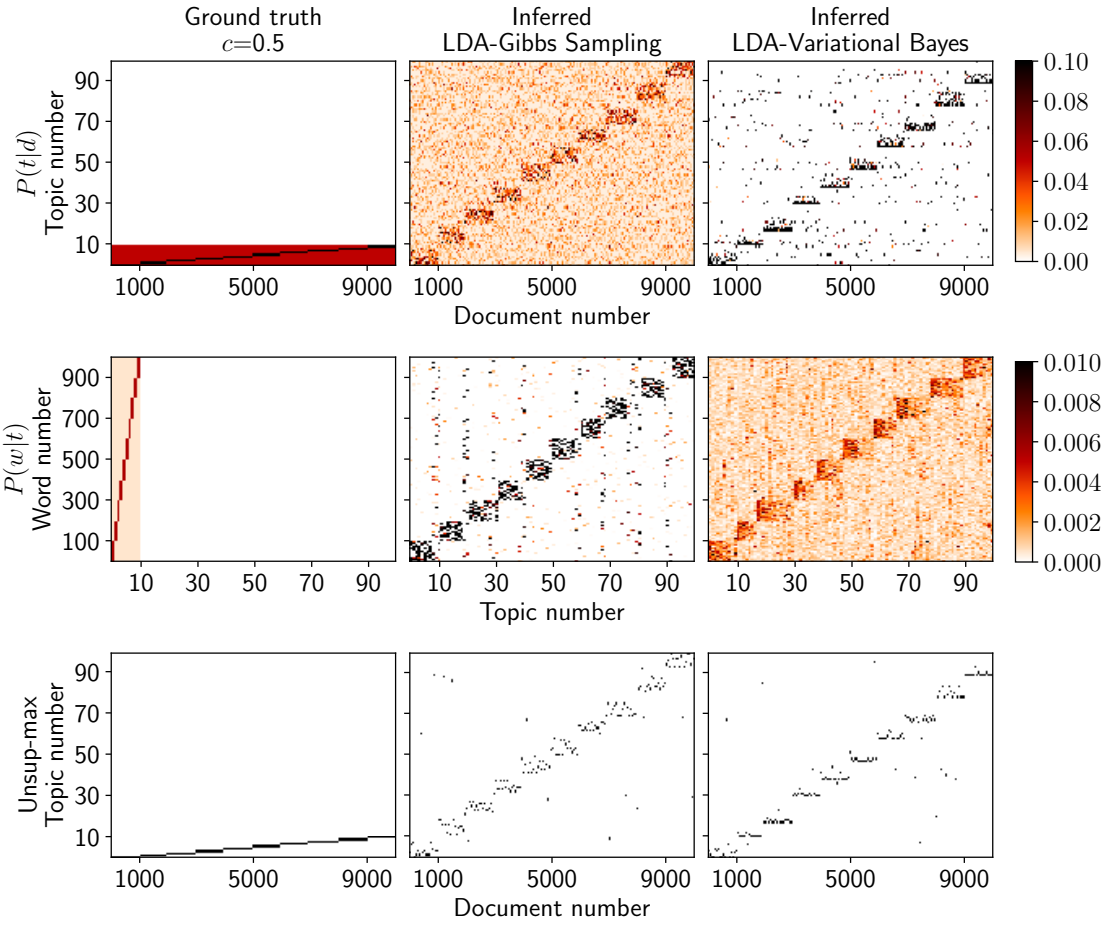


Figure S8: **Document classification overlooks information of the inferred topic structure.** Comparison of the topic-document distribution $P(t|d)$ (top row), the word-topic distribution $P(w|t)$ (middle row), and the predicted topic in unsupervised document classification $\arg\max_t P(t|d)$ (bottom row) for three cases: Ground truth as planted in the synthetic corpus (left column), inferred from Gibbs Sampling LDA (middle column), and inferred from Variational Bayes LDA (right column). For the synthetic benchmark corpora, we set parameters as $K = 10$ topics, $D = 10,000$ documents each of length $m_d = 100$, $V = 10^3$ as vocabulary size, and $c = 0.5$ for the degree of structure. For LDA models, we use $K_a = 100$ as the assumed number of topics.

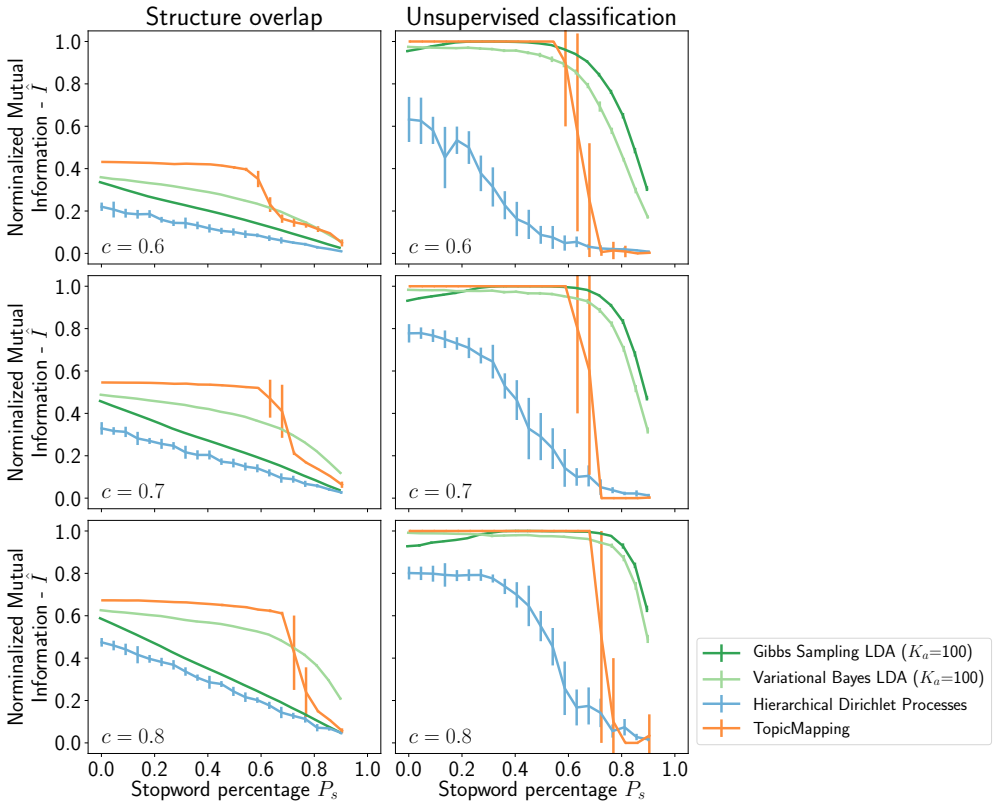


Figure S9: **Varying the degree of structure does not effect results on stopword dependency in synthetic corpora.** Normalized mutual information, $\hat{I}$, as measured by structure overlap (left column) and unsupervised document classification (right column) as in Fig. 5 (E, F) varying the degree of structure: $c = 0.6$ (top row), $c = 0.7$ (middle row), and $c = 0.8$ (bottom row).

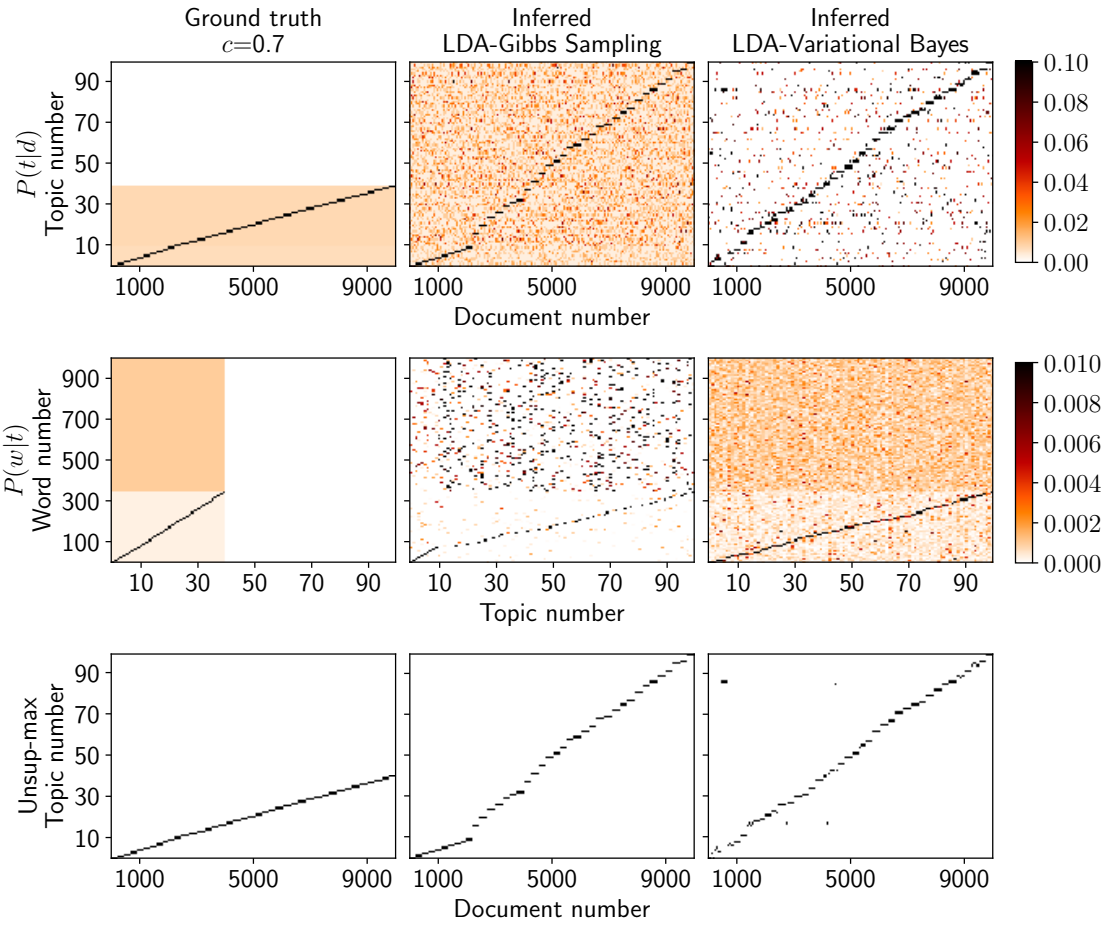


Figure S10: **Different LDA models lead to qualitatively different solutions in the case of stopwords.** Comparison of the topic-document distribution (top row), $P(t|d)$ , word-topic distribution (middle row), $P(w|t)$, and predicted topic in unsupervised document classification (bottom row), $\arg\max_t P(t|d)$ for three cases: Ground truth as planted in the synthetic corpus (left column), inferred from Gibbs Sampling LDA (middle column), and inferred from Variational Bayes LDA (right column). Same parameters as in Fig. 5 (E, F) setting the fraction of stopwords $P_s = 0.65$ and the degree of structure $c = 0.7$.

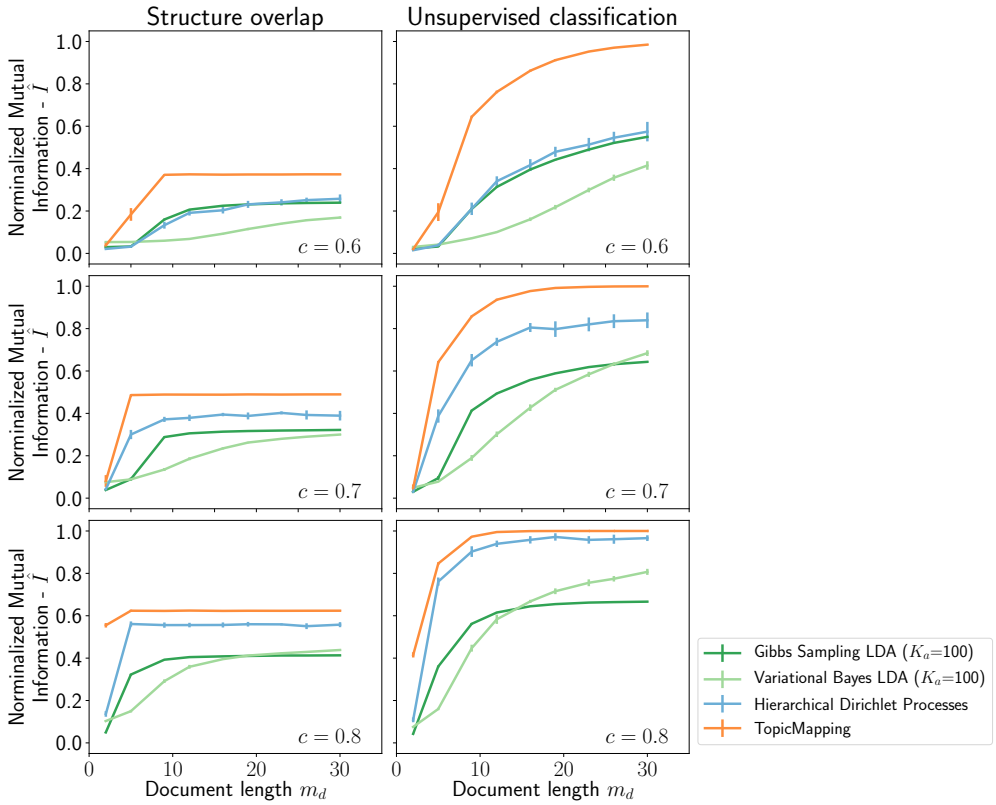


Figure S11: **Varying the degree of structure does not effect results on document length dependency in synthetic corpora.** Normalized mutual information, $\hat{I}$, as measured by structure overlap (left column) and unsupervised document classification (right column) as in Fig. 5 (B, C) varying the degree of structure: $c = 0.6$ (top row), $c = 0.7$ (middle row), and $c = 0.8$ (bottom row).

## References


L. AlSumait, D. Barbará, J. Gentle, and C. Domeniconi. Topic significance ranking of LDA generative models. In *Machine Learning and Knowledge Discovery in Databases*, pages 67–82. Springer Berlin Heidelberg, 2009.

E. G. Altmann, J. B. Pierrehumbert, and A. E. Motter. Beyond word frequency: Bursts, lulls, and scaling in the temporal distributions of words. *PloS one*, 4(11):e7678, 2009.

D. Andrzejewski, X. Zhu, and M. Craven. Incorporating domain knowledge into topic modeling via dirichlet forest priors. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 25–32. ACM, 2009.

S. Arora, R. Ge, Y. Halpern, D. Mimno, A. Moitra, D. Sontag, Y. Wu, and M. Zhu. A practical algorithm for topic modeling with provable guarantees. In *Proceedings of the 30th International Conference on Machine Learning*, pages 280–288. PMLR, 2013.

S. Arora, R. Ge, F. Koehler, T. Ma, and A. Moitra. Provable algorithms for inference in topic models. In *Proceedings of the 33rd International Conference on Machine Learning*, pages 2859–2867. PMLR, 2016.

D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

A. M. d. J. C. Cachopo. Improving methods for single-label text categorization. *Instituto Superior Técnico, Portugal*, 2007.

G. Doyle and C. Elkan. Accounting for burstiness in topic models. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 281–288. ACM Press, 2009.

T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101:5228–5235, 2004.

S. Hettich and S. Bay. The UCI KDD archive. *http://kdd. ics. uci. edu, Irvine, CA: University of California, Department of Information and Computer Science*, 1999.

W.-S. Hsu and P. Poupart. Online Bayesian Moment Matching for Topic Modeling with Unknown Number of Topics. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 4536–4544. Curran Associates, Inc., 2016.

S. M. Katz. Distribution of content words and phrases in text and language modelling. *Natural Language Engineering*, 2(1):15–59, 1996.

A. Lancichinetti. Topicmapping. *https://bitbucket.org/andrealanci/topicmapping*, 2016.

A. Lancichinetti, M. Irmak Sirer, J. X. Wang, D. Acuna, K. Kording, and L. A. N. Amaral. High-reproducibility and high-accuracy method for automated topic classification. *Physical Review X*, 5(1): 011007, 2015.

D. D. Lewis, Y. Yang, T. G. Rose, and F. Li. RCV1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5:361–397, 2004. ISSN 15337928. doi: 10.1145/122860.122861.

R. E. Madsen, D. Kauchak, and C. Elkan. Modeling word burstiness using the Dirichlet distribution. In *Proceedings of the 22nd Annual International Conference on Machine Learning*, pages 545–552. ACM Press, 2005.

A. K. McCallum. Mallet: A machine learning for language toolkit. *http://mallet.cs.umass.edu*, 2002.

D. Mimno and D. Blei. Bayesian checking for topic models. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 227–237, 2011.

T. Minka and J. Lafferty. Expectation-Propagation for the Generative Aspect Model. *Uncertainty in Artificial Intelligence*, pages 352–359, 2002.

I. Mukherjee and D. M. Blei. Relative performance guarantees for approximate inference in latent Dirichlet allocation. In *Advances in Neural Information Processing Systems*, pages 1129–1136. Curran Associates, Inc., 2009.

D. Newman, A. Asuncion, P. Smyth, and M. Welling. Distributed Algorithms for Topic Models. *J. Mach. Learn. Res.*, 10:1801–1828, dec 2009. ISSN 1532-4435.

R. Řehůřek and P. Sojka. Software framework for topic modelling with large corpora. In *Proceedings of the*

*LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50. ELRA, 2010.

A. Schofield, M. Magnusson, and D. Mimno. Pulling out the stops: Rethinking stopword removal for topic models. In *Proceedings of Conference of the European Chapter of the Association for Computational Linguistics*, pages 432–436. Association for Computational Linguistics, 2017.

M. Taddy. On estimation and selection for topic models. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, pages 1184–1193. PMLR, 2012.

J. Tang, Z. Meng, X. Nguyen, Q. Mei, and M. Zhang. Understanding the limiting factors of topic modeling via posterior contraction analysis. In *Proceedings of the 31st International Conference on Machine Learning*, pages 190–198. PMLR, 2014.

Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476): 1566–1581, 2006.

H. M. Wallach, I. Murray, R. Salakhutdinov, and D. Mimno. Evaluation methods for topic models. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1105–1112. ACM Press, 2009.

C. Wang. Hierarchical Dirichlet process. *https://github.com/blei-lab/hdp*, 2010.

P. Xie and E. P. Xing. Integrating document clustering and topic modeling. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, pages 694–703. AUAI Press, 2013.

A. N. K. Zaman, P. Matsakis, and C. Brown. Evaluation of stop word lists in text retrieval using Latent Semantic Indexing. In *International Conference on Digital Information Management*, pages 133–136. IEEE, 2011.

G. K. Zipf. *The Psychobiology of Language.* Routledge, London, 1936.