Improving Topic Coherence with Latent Feature Word Representations in MAP Estimation for Topic Modeling

Dat Quoc Nguyen, Kairit Sirts and **Mark Johnson**
Department of Computing
Macquarie University, Australia
dat.nguyen@students.mq.edu.au, {kairit.sirts, mark.johnson}@mq.edu.au

Abstract

Probabilistic topic models are widely used to discover latent topics in document collections, while latent feature word vectors have been used to obtain high performance in many natural language processing (NLP) tasks. In this paper, we present a new approach by incorporating word vectors to directly optimize the maximum a posteriori (MAP) estimation in a topic model. Preliminary results show that the word vectors induced from the experimental corpus can be used to improve the assignments of topics to words.

Keywords: MAP estimation, LDA, Topic model, Word vectors, Topic coherence

1 Introduction

Topic modeling algorithms, such as Latent Dirichlet Allocation (LDA) (Blei et al., 2003) and related methods (Blei, 2012), are often used to learn a set of latent topics for a corpus of documents and to infer document-to-topic and topic-to-word distributions from the co-occurrence of words within the documents (Wallach, 2006; Blei and McAuliffe, 2008; Wang et al., 2007; Johnson, 2010; Yan et al., 2013; Xie et al., 2015; Yang et al., 2015). With enough training data there is sufficient information in the corpus to accurately estimate the distributions. However, most topic models consider each document as a bag-of-words, i.e. the word order or the window-based local context information is not taken into account.

Topic models have also been constructed using latent features (Salakhutdinov and Hinton, 2009; Srivastava et al., 2013; Cao et al., 2015). Latent feature vectors have been recently successfully exploited for a wide range of NLP tasks (Glorot et al., 2011; Socher et al., 2013; Pennington et al., 2014). Rather than relying solely on word count information as the standard multinomial LDA does, or using only distributed feature representations, as in Salakhutdinov and Hinton (2009), Srivastava et al. (2013) and Cao et al. (2015), Nguyen et al. (2015) integrated pre-trained latent feature word representations containing external information from very large corpora into existing topic models and obtained significant improvements on small document collections and short text datasets. However, their implementation is computationally quite expensive because they have to compute a MAP estimate in each Gibbs sampling iteration.

In this paper, we experiment with MAP estimation using word vectors for LDA. Instead of mixing the Gibbs sampling and MAP estimation, we propose to optimize the MAP estimation of the full model directly. In addition, instead of using the pre-trained word vectors learned on external large corpora, we propose to learn the internal word vectors from the same topic-modeling corpus that we induce the document-to-topic and topic-to-word distributions from. In this manner, we can also handle the words that are not found in the list of the pre-trained word vectors. Furthermore, the internal word vectors can capture various aspects including word order information or local context information in the topic-modeling corpus. Preliminary results show that the internal word vectors can also help to significantly improve the topic-to-word assignments.

2 Related work

LDA (Blei et al., 2003) represents each document d in the document collection D as a mixture $\boldsymbol{\theta}_d$ over T topics, where each topic z is modeled by a probability distribution $\boldsymbol{\phi}_z$ over words in a vocab-

Dat Quoc Nguyen, Kairit Sirts and Mark Johnson. 2015. Improving Topic Coherence with Latent Feature Word Representations in MAP Estimation for Topic Modeling . In *Proceedings of Australasian Language Technology Association Workshop*, pages 116−121.

ulary W. As presented in Figure 1, where α and β are hyper-parameters, the generative process for LDA is described as follows:

$$\boldsymbol{\theta}_d \sim \mathrm{Dir}(\alpha) \qquad z_{d_i} \sim \mathrm{Cat}(\boldsymbol{\theta}_d)$$
$$\boldsymbol{\phi}_z \sim \mathrm{Dir}(\beta) \qquad w_{d_i} \sim \mathrm{Cat}(\boldsymbol{\phi}_{z_{d_i}})$$

where Dir and Cat stand for a Dirichlet distribution and a categorical distribution, and z_{d_i} is the topic indicator for the i^{th} word w_{d_i} in document d. Inference for LDA is typically performed by variational inference or Gibbs sampling (Blei et al., 2003; Griffiths and Steyvers, 2004; Teh et al., 2006; Porteous et al., 2008; Yao et al., 2009; Foulds et al., 2013).

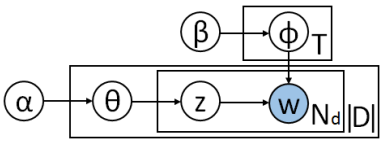

Figure 1: Graphical representation of LDA

When we ignore the Dirichlet priors and apply the Expectation Maximization (EM) algorithm to optimize the likelihood over the document-to-topic and topic-to-word parameters $\theta_{d,z}$ and $\phi_{z,w}$, we obtain the Probabilistic Latent Semantic Analysis model (Hofmann, 1999; Girolami and Kabán, 2003). Optimizing the MAP estimation for the LDA model has been suggested before. Chien and Wu (2008), Asuncion et al. (2009) and Taddy (2012) proposed EM algorithms for estimating $\theta_{d,z}$ and $\phi_{z,w}$, while we use direct gradient-based optimization methods. Sontag and Roy (2011) optimized the MAP estimates of $\phi_{z,w}$ and $\theta_{d,z}$ in turn by integrating out $\theta_{d,z}$ and $\phi_{z,w}$ respectively. We, on the other hand, estimate all parameters jointly in a single optimization step.

In addition to Taddy (2012)'s approach, applying MAP estimation to learn log-linear models for topic models is also found in Eisenstein et al. (2011) and Paul and Dredze (2015). Our MAP model is also defined in log-linear representation. However, unlike our MAP approach, those approaches do not use latent feature word vectors to characterize the topic-to-word distributions.

Furthermore, Berg-Kirkpatrick et al. (2010) proposed a direct optimization approach of the objective function for Hidden Markov Model-like generative models. However, they applied the approach to various unsupervised NLP tasks, such as part-of-speech induction, grammar induction, word alignment, and word segmentation, but not to topic models.

3 Direct MAP estimation approach

In this section, we describe our new direct MAP estimation approach using word vectors for LDA.

Following the likelihood principle, the document-to-topic and topic-to-word distributions $\boldsymbol{\theta}_d$ and $\boldsymbol{\phi}_z$ are determined by maximizing the log likelihood function:

$$\mathcal{L} = \sum_{d \in D} \sum_{w \in W} n_{d,w} \log \sum_z \theta_{d,z} \phi_{z,w} \tag{1}$$

where $n_{d,w}$ is the number of times the word type w appears in document d.

Estimating the parameters $\theta_{d,z}$ and $\phi_{z,w}$ in the original simplex space requires constraints: $\theta_{d,z} \geq 0, \phi_{z,w} \geq 0, \sum_z \theta_{d,z} = 1$ and $\sum_w \phi_{z,w} = 1$. In order to avoid those constraints and to improve estimation efficiency, we transfer the parameters into the natural exponential family parameterization. So we define $\theta_{d,z}$ and $\phi_{z,w}$ as follows:

$$\begin{aligned} \theta_{d,z} &= \frac{\exp(\xi_{d,z})}{\sum_{z'} \exp(\xi_{d,z'})} \\ \phi_{z,w} &= \frac{\exp(\boldsymbol{v}_w \cdot \boldsymbol{\mu}_z + \psi_{z,w})}{\sum_{w' \in W} \exp(\boldsymbol{v}_{w'} \cdot \boldsymbol{\mu}_z + \psi_{z,w'})} \end{aligned} \tag{2}$$

where $\boldsymbol{v}_w$ is the m-dimensional vector associated with word w, while $\boldsymbol{\mu}_z$ is the m-dimensional topic vector associated with topic z. Here $\boldsymbol{v}$ is fixed, and we will learn $\boldsymbol{\mu}$ together with $\boldsymbol{\xi}$ and $\boldsymbol{\psi}$.

With L_2 and L_1 regularizers, we have a new objective function as follows:

$$\begin{aligned} \mathcal{L} = & \sum_{d \in D} \sum_{w \in W} n_{d,w} \log \sum_z \left(\frac{\exp(\xi_{d,z})}{\sum_{z'} \exp(\xi_{d,z'})} \times \frac{\exp(\boldsymbol{v}_w \cdot \boldsymbol{\mu}_z + \psi_{z,w})}{\sum_{w' \in W} \exp(\boldsymbol{v}_{w'} \cdot \boldsymbol{\mu}_z + \psi_{z,w'})} \right) \\ & - \sum_{d \in D} \left(\lambda_2 \|\boldsymbol{\xi}_d\|_2^2 + \lambda_1 \|\boldsymbol{\xi}_d\|_1 \right) \\ & - \sum_z \left(\pi_2 \|\boldsymbol{\mu}_z\|_2^2 + \pi_1 \|\boldsymbol{\mu}_z\|_1 \right) \\ & - \sum_z \left(\epsilon_2 \|\boldsymbol{\psi}_z\|_2^2 + \epsilon_1 \|\boldsymbol{\psi}_z\|_1 \right) \end{aligned} \tag{3}$$

The MAP estimate of the model parameters is obtained by maximizing the regularized log likelihood $\mathcal{L}$. The derivatives with respect to the parameters $\xi_{d,z}$ and $\psi_{z,w}$ are:

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \xi_{d,z}} = & \sum_{w \in W} n_{d,w} \mathrm{P}(z \mid w, d) - n_d \theta_{d,z} \\ & - 2\lambda_2 \xi_{d,z} - \lambda_1 \mathrm{sign}(\xi_{d,z}) \end{aligned} \tag{4}$$

where $\mathrm{P}(z \mid w, d) = \frac{\theta_{d,z}\phi_{z,w}}{\sum_{z'}\theta_{d,z'}\phi_{z',w}}$, and n_d is the total number of word tokens in the document d.

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \psi_{z,w}} &= \sum_{d \in D} n_{d,w} \mathrm{P}(z \mid w, d) \\ &- \phi_{z,w} \sum_{d \in D} \sum_{w' \in W} n_{d,w'} \mathrm{P}(z \mid w', d) \\ &- 2\epsilon_2 \psi_{z,w} - \epsilon_1 \mathrm{sign}(\psi_{z,w})\end{aligned} \quad (5)$$

And the derivative with respect to the j^{th} element of the vector for each topic z is:

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \mu_{z,j}} =& \sum_{d \in D} \sum_{w \in W} n_{d,w} \mathrm{P}(z \mid w, d) \Big(v_{w,j} - \sum_{w' \in W} v_{w',j} \phi_{z,w'} \Big) \\ &- 2\pi_2 \mu_{z,j} - \pi_1 \mathrm{sign}(\mu_{z,j})\end{aligned} \quad (6)$$

We used OWL-QN[1] (Andrew and Gao, 2007) to find the topic vector $\boldsymbol{\mu}_z$ and the parameters $\xi_{d,z}$ and $\psi_{z,w}$ that maximize $\mathcal{L}$.

4 Experiments

To investigate the performance of our new approach, we compared it with two baselines on topic coherence: 1) variational inference LDA (Blei et al., 2003); and 2) Gibbs sampling LDA (Griffiths and Steyvers, 2004). The topic coherence evaluation measures the coherence of the topic-to-word associations, i.e. it directly evaluates how the high-probability words in each topic are semantically coherent (Chang et al., 2009; Newman et al., 2010; Mimno et al., 2011; Stevens et al., 2012; Lau et al., 2014; Röder et al., 2015).

4.1 Experimental setup

We conducted experiments on the standard benchmark 20-Newsgroups dataset.[2]

In addition to converting into lowercase and removing non-alphabetic characters, we removed stop-words found in the stop-word list in the Mallet toolkit (McCallum, 2002). We then removed words shorter than 3 characters or words appearing less than 10 times. Table 1 presents details of the experimental dataset.

As pointed out in Levy and Goldberg (2014) and Pennington et al. (2014), the prediction-based methods and count-based methods for learning word vectors are not qualitatively different on a range of semantic evaluation tasks. Thus, we simply use the Word2Vec toolkit[3] (Mikolov et al., 2013) to learn 25-dimensional word vectors on the experimental dataset, using a local 10-word window context.[4]

Dataset	#docs	#w/d	|W|
20-Newsgroups	18,820	105	20,940

Table 1: Details of the experimental dataset. #docs: number of documents; #w/d: the average number of words per document; |W|: the number of word types.

The numbers of topics is set to 20. For variational inference LDA, we use Blei's implementation.[5] For Gibbs sampling LDA, we use the `jLDADMM` package[6] (Nguyen, 2015) with common hyper-parameters $\beta = 0.01$ and $\alpha = 0.1$ (Newman et al., 2009; Hu et al., 2011; Xie and Xing, 2013). We ran Gibbs sampling LDA for 2000 iterations and evaluated the topics assigned to words in the last sample. We then used the document-to-topic and topic-to-word distributions from the last sample of Gibbs sampling LDA to initialize the parameters $\xi_{d,z}$ and $\psi_{z,w}$ while topic vectors $\boldsymbol{\mu}_z$ are initialized as zero vectors in our MAP learner. For our MAP approach, we set[7] $\lambda_2 = \pi_2 = 0.01$, $\lambda_1 = \pi_1 = 1.0e{-}6$, $\epsilon_2 = 0.1$ and $\epsilon_1 = 0.01$. We report the mean and standard deviation of the results of ten repetitions of each experiment.

4.2 Quantitative analysis

For a quantitative analysis on topic coherence, we use the normalized pointwise mutual information (NPMI) score. Lau et al. (2014) showed that human scores on a word intrusion task are strongly correlated with NPMI. A higher NPMI score indicates that the topic distributions are semantically more coherent.

Given a topic t represented by its top-N topic words $w_1, w_2, ..., w_N$, the NPMI score for t is: $\mathrm{NPMI}(t) = \sum_{1 \leqslant i < j \leqslant N} \frac{\log \frac{\mathrm{P}(w_i, w_j)}{\mathrm{P}(w_i)\mathrm{P}(w_j)}}{-\log \mathrm{P}(w_i, w_j)}$, where the

[1] We employed the OWL-QN implementation from the Mallet toolkit (McCallum, 2002).

[2] We used the "all-terms" version of the 20-Newsgroups dataset available at http://web.ist.utl.pt/acardoso/datasets/ (Cardoso-Cachopo, 2007).

[3] https://code.google.com/p/word2vec/

[4] The parameters of Word2Vec are set to "-cbow 0 -size 25 -window 10 -negative 0 -hs 1."

[5] http://www.cs.princeton.edu/~blei/lda-c/. We used initial value $\alpha = 0.1$ and settings of "var max iter 20, var convergence 1.0e$-$12, em convergence 1.0e$-$8, em max iter 500, alpha estimate".

[6] http://jldadmm.sourceforge.net/

[7] We simply fixed the values of λ_2, π_2, λ_1, π_1, and then varied the values of ϵ_2 and ϵ_1 in $\{0.01, 0.05, 0.1\}$.

Topic 1		Topic 2		Topic 12		Topic 18		Topic 19	
G-LDA	MAP+V	G-LDA	MAP+V	G-LDA	MAP+V	G-LDA	MAP+V	G-LDA	MAP+V
car	car	power	sale	game	game	space	space	medical	medical
writes	cars	sale	power	team	team	nasa	nasa	disease	disease
article	engine	work	**shipping**	year	games	gov	earth	article	health
cars	oil	battery	**offer**	games	year	earth	gov	health	food
engine	speed	radio	battery	hockey	play	writes	launch	drug	drug
good	miles	good	radio	writes	hockey	article	moon	food	cancer
oil	price	high	ground	play	players	launch	orbit	cancer	**doctor**
price	**dealer**	sound	sound	players	season	moon	shuttle	msg	drugs
speed	**ford**	ground	high	season	**win**	orbit	**mission**	drugs	msg
miles	**drive**	writes	**cable**	article	**baseball**	shuttle	**henry**	writes	**patients**

Table 3: Examples of the 10 most probable topical words on the 20-Newsgroups dataset. G-LDA → Gibbs sampling LDA; MAP+V → Our MAP approach using internal word vectors. The words found by G-LDA and not by MAP+V are underlined. The words found by MAP+V but not by G-LDA are in **bold**.

Method	Top-10	Top-15	Top-20
V-LDA	-4.2 ± 0.4	-12.2 ± 0.6	-24.1 ± 0.6
G-LDA	-4.2 ± 0.4	-11.7 ± 0.7	-22.9 ± 0.9
MAP-O	-3.8 ± 0.5	-10.8 ± 0.6	-22.1 ± 1.2
MAP+V	**-3.4 ± 0.3**	**-10.1 ± 0.7**	**-20.6 ± 1.0**
Improve.	0.8	1.6	2.3

Table 2: NPMI scores (mean and standard deviation) on the 20-Newsgroups dataset with different numbers of top topical words; V-LDA → Variational inference LDA; G-LDA → Gibbs sampling LDA; MAP-O → Our MAP learner where we fix topic vectors $\boldsymbol{\mu}$ as zero vectors and only learn parameters $\boldsymbol{\xi}$ and ψ; MAP+V → Our MAP learner where we learn $\boldsymbol{\mu}$ together with $\boldsymbol{\xi}$ and ψ. The *Improve.* row denotes the absolute improvement accounted for MAP+V over the best result produced by the baselines V-LDA and G-LDA.

probabilities are derived from a 10-word sliding window over an external corpus.[8] The NPMI score for a topic model is the average score for all topics.

Table 2 shows that our approach using internal word vectors MAP+V produces significantly higher[9] NPMI scores than the baseline variational inference LDA and Gibbs sampling LDA models. So this indicates that the word vectors containing internal context information from experimental dataset can help to improve topic coherence.

4.3 Qualitative analysis

This section provides an example of how our approach improves topic coherence. Table 3 compares the top-10 words produced by the baseline Gibbs sampling LDA and our MAP+V approach on the 20-Newsgroups dataset. It is clear that all top-10 words learned with our MAP+V model are qualitatively more coherent. For example, topic 19 of the Gibbs sampling LDA model consists of words related to "medicine" together with other unrelated words, whereas our MAP+V approach produced a purer topic 19 only about "medicine."

On 20-Newsgroups dataset, it is common that the baseline variational inference LDA and Gibbs sampling LDA models include the frequent words such as "writes" and "article" as top topical words in many topics. However, our MAP+V model using the internal word vectors is able to exclude these words out of the top words in these topics.

5 Conclusions and future work

In this paper, we proposed a new approach of fully direct MAP estimation for the LDA topic model inference, incorporating latent feature representations of words. Preliminary results show that the latent feature representations trained from the experimental topic-modeling corpus can improve the topic-to-word mapping.

In future work, we plan to investigate the effects of the context window size as well as the size of the word vectors further. In addition, we plan to test our approach on a range of different datasets. We also plan to compare the presented results with Nguyen et al. (2015)'s model using internal word vectors. Even though we learn the internal word vectors from the experimental dataset, we believe that it is worth trying to initialize them from vectors learned from an external corpus, thus also incorporating generalizations from that corpus.

[8] We use the English Wikipedia dump of July 8, 2014, containing 4.6 million articles as our external corpus.

[9] Using the two sample Wilcoxon test, the improvement is significant ($p < 0.01$).

Acknowledgments

The first author is supported by an International Postgraduate Research Scholarship and a NICTA NRPA Top-Up Scholarship.

References

Galen Andrew and Jianfeng Gao. 2007. Scalable Training of L1-regularized Log-linear Models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.

Arthur Asuncion, Max Welling, Padhraic Smyth, and Yee Whye Teh. 2009. On Smoothing and Inference for Topic Models. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 27–34.

Taylor Berg-Kirkpatrick, Alexandre Bouchard-Côté, John DeNero, and Dan Klein. 2010. Painless Unsupervised Learning with Features. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 582–590.

David M. Blei and Jon D. McAuliffe. 2008. Supervised Topic Models. In *Advances in Neural Information Processing Systems 20*, pages 121–128. MIT Press.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022.

David M. Blei. 2012. Probabilistic Topic Models. *Communications of the ACM*, 55(4):77–84.

Ziqiang Cao, Sujian Li, Yang Liu, Wenjie Li, and Heng Ji. 2015. A Novel Neural Topic Model and Its Supervised Extension. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 2210–2216.

Ana Cardoso-Cachopo. 2007. Improving Methods for Single-label Text Categorization. PhD Thesis, Instituto Superior Tecnico, Universidade Tecnica de Lisboa.

Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L. Boyd-graber, and David M. Blei. 2009. Reading Tea Leaves: How Humans Interpret Topic Models. In *Advances in Neural Information Processing Systems 22*, pages 288–296.

Jen-Tzung Chien and Meng-Sung Wu. 2008. Adaptive Bayesian Latent Semantic Analysis. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(1):198–207.

Jacob Eisenstein, Amr Ahmed, and Eric Xing. 2011. Sparse Additive Generative Models of Text. In *Proceedings of the 28th International Conference on Machine Learning*, pages 1041–1048.

James Foulds, Levi Boyles, Christopher DuBois, Padhraic Smyth, and Max Welling. 2013. Stochastic Collapsed Variational Bayesian Inference for Latent Dirichlet Allocation. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 446–454.

Mark Girolami and Ata Kabán. 2003. On an Equivalence Between PLSI and LDA. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, pages 433–434.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Domain Adaptation for Large-Scale Sentiment Classification: A Deep Learning Approach. In *Proceedings of the 28th International Conference on Machine Learning*, pages 513–520.

Thomas L. Griffiths and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101(Suppl 1):5228–5235.

Thomas Hofmann. 1999. Probabilistic Latent Semantic Indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57.

Yuening Hu, Jordan Boyd-Graber, and Brianna Satinoff. 2011. Interactive Topic Modeling. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, pages 248–257.

Mark Johnson. 2010. PCFGs, Topic Models, Adaptor Grammars and Learning Topical Collocations and the Structure of Proper Names. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1148–1157.

Han Jey Lau, David Newman, and Timothy Baldwin. 2014. Machine Reading Tea Leaves: Automatically Evaluating Topic Coherence and Topic Model Quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 530–539.

Omer Levy and Yoav Goldberg. 2014. Neural Word Embedding as Implicit Matrix Factorization. In *Advances in Neural Information Processing Systems 27*, pages 2177–2185.

Andrew McCallum. 2002. MALLET: A Machine Learning for Language Toolkit.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems 26*, pages 3111–3119.

David Mimno, Hanna M. Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing Semantic Coherence in Topic Models.

In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 262–272.

David Newman, Arthur Asuncion, Padhraic Smyth, and Max Welling. 2009. Distributed Algorithms for Topic Models. *The Journal of Machine Learning Research*, 10:1801–1828.

David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. Automatic Evaluation of Topic Coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 100–108.

Dat Quoc Nguyen, Richard Billingsley, Lan Du, and Mark Johnson. 2015. Improving Topic Models with Latent Feature Word Representations. *Transactions of the Association for Computational Linguistics*, 3:299–313.

Dat Quoc Nguyen. 2015. jLDADMM: A Java package for the LDA and DMM topic models. http://jldadmm.sourceforge.net/.

Michael Paul and Mark Dredze. 2015. SPRITE: Generalizing Topic Models with Structured Priors. *Transactions of the Association for Computational Linguistics*, 3:43–57.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543.

Ian Porteous, David Newman, Alexander Ihler, Arthur Asuncion, Padhraic Smyth, and Max Welling. 2008. Fast Collapsed Gibbs Sampling for Latent Dirichlet Allocation. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 569–577.

Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the Space of Topic Coherence Measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pages 399–408.

Ruslan Salakhutdinov and Geoffrey Hinton. 2009. Replicated Softmax: an Undirected Topic Model. In *Advances in Neural Information Processing Systems 22*, pages 1607–1614.

Richard Socher, John Bauer, Christopher D. Manning, and Ng Andrew Y. 2013. Parsing with Compositional Vector Grammars. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 455–465.

David Sontag and Dan Roy. 2011. Complexity of Inference in Latent Dirichlet Allocation. In *Advances in Neural Information Processing Systems 24*, pages 1008–1016.

Nitish Srivastava, Ruslan Salakhutdinov, and Geoffrey Hinton. 2013. Modeling Documents with a Deep Boltzmann Machine. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, pages 616–624.

Keith Stevens, Philip Kegelmeyer, David Andrzejewski, and David Buttler. 2012. Exploring Topic Coherence over Many Models and Many Topics. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 952–961.

Matthew A. Taddy. 2012. On Estimation and Selection for Topic Models. In *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics*.

Yee W Teh, David Newman, and Max Welling. 2006. A Collapsed Variational Bayesian Inference Algorithm for Latent Dirichlet Allocation. In *Advances in Neural Information Processing Systems 19*, pages 1353–1360.

Hanna M Wallach. 2006. Topic Modeling: Beyond Bag-of-Words. In *Proceedings of the 23rd international conference on Machine learning*, pages 977–984.

Xuerui Wang, Andrew McCallum, and Xing Wei. 2007. Topical N-Grams: Phrase and Topic Discovery, with an Application to Information Retrieval. In *Proceedings of the 2007 Seventh IEEE International Conference on Data Mining*, pages 697–702.

Pengtao Xie and Eric P. Xing. 2013. Integrating Document Clustering and Topic Modeling. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, pages 694–703.

Pengtao Xie, Diyi Yang, and Eric Xing. 2015. Incorporating Word Correlation Knowledge into Topic Modeling. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 725–734.

Xiaohui Yan, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng. 2013. A Biterm Topic Model for Short Texts. In *Proceedings of the 22Nd International Conference on World Wide Web*, pages 1445–1456.

Yi Yang, Doug Downey, and Jordan Boyd-Graber. 2015. Efficient Methods for Incorporating Knowledge into Topic Models. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 308–317.

Limin Yao, David Mimno, and Andrew McCallum. 2009. Efficient Methods for Topic Model Inference on Streaming Document Collections. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 937–946.