A Neural Framework for Generalized Topic Models

Dallas Card  
Machine Learning Department  
Carnegie Mellon University  
Pittsburgh, PA 15213, USA  
dcard@cmu.edu

Chenhao Tan  
Paul G. Allen School  
University of Washington  
Seattle, WA 98195, USA  
chenhao@chenhaot.com

Noah A. Smith  
Paul G. Allen School  
University of Washington  
Seattle, WA 98195, USA  
nasmith@cs.washington.edu

Abstract

Topic models for text corpora comprise a popular family of methods that have inspired many extensions to encode properties such as sparsity, interactions with covariates, and the gradual evolution of topics. In this paper, we combine certain motivating ideas behind variations on topic models with modern techniques for variational inference to produce a flexible framework for topic modeling that allows for rapid exploration of different models. We first discuss how our framework relates to existing models, and then demonstrate that it achieves strong performance, with the introduction of sparsity controlling the trade off between perplexity and topic coherence. We have released our code and preprocessing scripts to support easy future comparisons and exploration.

1 Introduction

Topic models have been one of the most successful avenues of research into unsupervised learning, particularly within the domain of document modeling, with the best known example being latent Dirichlet allocation (Blei et al., 2003). A variety of inference algorithms have been developed for LDA, based on both Markov chain Monte Carlo and variational inference techniques, as well as extensions for stochastic and streaming settings (Hoffman et al., 2010; Broderick et al., 2013). Simultaneously, a large number of variations on LDA have been proposed which allow us to take advantage of prior knowledge, such as sparsity in topic distributions, or the gradual variation in topics over time, or to incorporate other document covariates, such as author (Rosen-Zvi et al., 2004; Blei and Lafferty, 2006; Lafferty and Blei, 2006; Mcauliffe and Blei, 2008; Ahmed and Xing, 2010; Roberts et al., 2014). However, these more complex models also typically require custom inference algorithms, which makes it difficult to rapidly explore variations.

Here, we present a flexible approach to topic modeling based on neural networks. Our framework allows for rapid exploration of models and additional ways to incorporate prior information, without requiring any model-specific derivations. Drawing inspiration from sparse additive generative (SAGE) models (Eisenstein et al., 2011), we are able to provide inference for multi-faceted models and sparse topics, and obtain strong performance by making use of recent advances in variational inference.

The key idea introduced in SAGE was to replace topics with deviations from a background word distribution. This leads to three main advantages:

1. Sparsity: by placing a sparsity-inducing prior on the deviations from the background frequency, we obtain topics characterized using a small set of words, which is arguably helpful for interpretation and may protect against overfitting.

2. Multi-faceted models: because the deviations are additive in log space, it is straightforward to incorporate many facets for different labels or covariates (e.g., time the document was written, political ideology of the author) and their interactions with topics.
3. Efficiency: the combined result of the above effects means that there is no need to learn a separate parameter in each topic for common words that have approximately the same frequency across documents.

Although SAGE offers great flexibility, it involves an intractable posterior for which there are no closed form variational updates. Instead, the authors proposed using a form of variational EM with Newton optimization for both parameters and latent variables, which required a separate derivation for each parameter type and care in implementation. Moreover, although it achieved high levels of sparsity in topics, it did not always outperform LDA in terms of perplexity.

Using our framework, we present a model that is able to incorporate all the desiderata of SAGE, while being amenable to modern variational inference techniques. These advances in inference are made possible by past work on sampling-based (Monte Carlo) algorithms for variational inference. A line of work beginning with the variational autoencoder (VAE) demonstrated how variational inference could be applied to otherwise intractable posterior distributions without incurring the high variance associated with previous approaches to black-box variational inference (Ranganath et al., 2014; Mnih and Gregor, 2014; Kingma and Welling, 2014; Rezende et al., 2014).

The neural variational document model (NVDM) made use of this approach, improving upon previously stated perplexity results on a pair of benchmark datasets (Miao et al., 2016). The model we present here includes NVDM as a special case, and we demonstrate that we can achieve better results through modification to the model. Moreover, because the same inference framework can be used without modification for a wide range of models, this allows us to painlessly explore variations involving multiple facets, interactions, and custom regularizers, without having to derive new update equations. In combination with tools for automatic differentiation, this approach to inference allows for a user-driven, iterative approach to modeling and model fitting (Blei, 2014; Kucukelbir et al., 2015).

We have also made available the code for our model, as well as preprocessing scripts, to help ensure replicability and fair future comparisons.

2 A Neural Framework for Topic Models

We present a general framework for topic modeling based on neural networks and sampling-based variational inference using the reparameterization trick (Kingma and Welling, 2014).

**Generative story for our framework.** Consider a corpus of $D$ documents, where document $i$ is a list of $N_i$ words, $w_i$, with $V$ words in the vocabulary. For each document, we may have some observed covariate information (e.g., its authors and year of publication), $y_i \in \mathbb{R}^P$. The goal of topic modeling is typically to recover a matrix $W \in \mathbb{R}^{V \times T}$, where $T$ is the number of topics and each row represents the importance of each word within that topic. In LDA, each column represents a multinomial distribution (i.e., the word distribution for a document entirely about that topic) and thus is constrained on the simplex.

Analogous to the generative story in LDA, our framework follows the generative story below (see also Figure 1a). The key differences are twofold: 1) we replace the Dirichlet-multinomial process to get topic distributions for documents with a multivariate normal distribution and a generative neural network $f_g$; 2) we replace the multinomial distribution over words for each topic with a topic network, which in the simplest case is parameterized by a matrix $W \in \mathbb{R}^{V \times T}$.

For each document $i$ of length $N_i$:

(a) $z_i \sim \mathcal{N}(0, I)$

(b) $r_i = f_g(z_i)$

(c) For each word $j$ in document $i$, $j = 1, \ldots, N_i$:

$$w_{ij} \sim p(w_{ij} \mid W, r_i),$$

where $p(w_{ij} \mid W, r_i)$ refers to the probability distribution over words in the vocabulary for document $i$. Each word $j$ will be drawn from this distribution, parametrized by the topic-word network, i.e.,

$$p(w_{ij} \mid W, r_i) \propto \exp(W \cdot r_i).$$  \hspace{1cm} (1)
Figure 1: Figure 1a presents the generative story of our model. Figure 1b illustrates the inference network using the reparametrization trick to perform variational inference on our model. Shaded nodes are observed; double circles indicate deterministic transformations of parent nodes.

In the simplest case, \( f_g \) could be the identity function, as is used in NVDM (Miao et al., 2016). However, we can also use more complex functions, such as a parameterized multi-layer perceptron.

A SAGE-style model. Using the above framework, we can easily instantiate a model with the important properties of SAGE. Because we exponentiate the output of the topic-word network, we can freely add terms inside the exponential to represent the log background frequency as well as additional facets, including interactions. Furthermore, just as SAGE obtained sparse topic deviations using a Laplace prior on the weights, we can encourage sparse deviations using \( l_1 \) regularization on the weights in the topic-word network. Specifically, we extend the topic-word network as follows:

\[
p(w_{ij} | d, W, W_1, W_2, r_i, y_i) \propto \exp(d + W \cdot r_i + W_1 \cdot y_i + W_2 \cdot (r_i \otimes y_i)),
\]

where \( d \) is the \( V \)-dimensional background log-frequency distribution, \( r_i \) is a corresponding \( T \)-dimensional latent representation of document \( i \), \( y_i \) is a corresponding \( P \)-dimensional vector of observed covariates, \( r_i \otimes y_i \) is a vector of interactions (of length \( T \times P \)), and \( W, W_1 \) and \( W_2 \) are weight matrices. Obviously we can choose to ignore various parts of this model, for example if we don’t have any observed covariates, or we don’t wish to use interactions. Similarly, additional facets can easily be added as additional terms inside the exponential, such as interactions between covariates.

Collectively, the generative network \((f_g)\) and the topic-word network define a distribution over words conditional on \( z \) and \( y \), which we summarize as:

\[
p_\theta(w_{ij} | z_i, y_i) = p(w_{ij} | d, W, W_1, W_2, f_g(z_i), y_i),
\]

where \( \theta \) refers to \( W, W_1, W_2 \) and \( d \), as well as all the parameters of \( f_g \).

Inference. In order to infer an approximate posterior distribution over the latent representation of each document \((r_i)\), we adopt the sampling-based variational inference framework developed in previous work (Kingma and Welling, 2014; Rezende et al., 2014). The key idea is to compute a variational approximation to an intractable posterior using a highly-expressive but differentiable function, such as a multi-layer perceptron.

Similar to Miao et al. (2016) but incorporating covariates, we use the following inference network and transformations:

\[
\begin{align*}
\pi_i &= f_e(w_i, y_i) \\
\mu_i &= W_\mu \pi_i + b_\mu \\
\log \sigma_i^2 &= W_\sigma \pi_i + b_\sigma,
\end{align*}
\]
where \( f_c(w_i, y_i) \) is a multilayer perceptron acting on \( y_i \) and the word counts in document \( i \). These equations define a diagonal multivariate normal distribution, \( q_\phi(z_i \mid w_i, y_i) = N(\mu_i, \sigma_i^2) \), which will serve as our variational approximation to the intractable posterior \( p(z_i \mid w_i, y_i) \), where \( \phi \) refers to \( W_\mu, b_\mu, W_z, \) and \( b_z \), as well as all the parameters of \( f_c \) (see Figure 1b).

As in past work, we attempt to minimize an objective function equal to the negative log-likelihood of \( p(w_i, y_i) \), which will serve as our variational approximation to the intractable posterior \( p(z_i \mid w_i, y_i) \), where \( \phi \) refers to \( W_\mu, b_\mu, W_z, \) and \( b_z \). We present two versions of our model – a simpler model than NVDM which uses only a

As in typical variational inference, we want to minimize the KL divergence between the true posterior, \( p(z_i \mid w_i, y_i) \) and the variational approximation of it. After some manipulations, we obtain the lower bound (ELBO),

\[
L(w_i) = \mathbb{E}_{q_\phi(z_i \mid w_i, y_i)} \left[ \sum_{j=1}^{N_i} \log p_\theta(w_{ij} \mid z_i, y_i) \right] - D_{KL}[q_\phi(z_i \mid w_i, y_i) \| p(z_i)]. \tag{4}
\]

Using the reparameterization trick (Kingma and Welling, 2014), the combined inference-generator-topic network becomes fully differentiable, and we can update all parameters involved using stochastic gradient descent based on a single sample from the approximate posterior, as shown in Figure 1b. Specifically, a draw from \( q_\phi(z_i \mid w_i, y_i) \) is equivalent to a transformed sample from a standard multivariate normal, according to

\[
\epsilon^{(s)} \sim N(0, I)
\]

\[
g_\phi(w_i, y_i, \epsilon^{(s)}) = \mu_i + \sigma_i \cdot \epsilon^{(s)}
\]

Thus, we can replace \( p_\theta(w_{ij} \mid z_i, y_i) \) with \( p_\theta(w_{ij} \mid g_\phi(w_i, y_i, \epsilon^{(s)}), y_i) \). This replacement in turn allows us to estimate the bound with a Monte Carlo approximation using \( S \) independent samples of \( \epsilon \):

\[
L(w_i) \approx \frac{1}{S} \sum_{s=1}^{S} \left[ \sum_{j=1}^{N_i} \log p_\theta(w_{ij} \mid g_\phi(w_i, y_i, \epsilon^{(s)}), y_i) \right] - D_{KL}[q_\phi(z_i \mid w_i, y_i) \| p(z_i)], \tag{5}
\]

which will be used to compute an upper bound on perplexity in the experiments.

As shown in Kingma and Welling (2014), because we have placed a standard multivariate normal prior on \( z \), there is a closed form solution to the KL divergence term above. Furthermore, because we have defined the above networks in terms of differentiable functions, and we can easily compute the gradient of the conditional log likelihood term with respect to both \( \theta \) and \( \phi \).

Note that unlike LDA and SAGE, we do not constrain document representations to lie on the simplex. Rather, like NVDM and exponential-family PCA (Collins et al., 2001), our model assumes that document representations are unconstrained, although we could in principle use the generator network to apply such a transformation.

Adaptive sparsity. As in past work, we attempt to minimize an objective function equal to the negative of the bound (Equation 5). As described above, we can encourage sparse deviations from the background by adding an \( l_1 \) penalty on the weights in the topic network (i.e., \( W, W_1, \) and \( W_2 \)) to this objective function. We can determine the strength of this regularization by tuning a hyperparameter, \( \lambda \). Alternatively, we can use an adaptive penalty to push the model towards a desired level of sparsity, specified by the user. For the latter approach, on each epoch, we update the strength or regularization according to:

\[
\lambda^{(i+1)} = \lambda^{(i)} \cdot 2^{(t-c)}
\]

where \( t \) is the target sparsity level, and \( c \) is the current sparsity level (that is, the proportion of weights that are less than some threshold, which we take to be \( 10^{-3} \)).

### 3 Experiments

#### 3.1 Experiment Setup

To evaluate and demonstrate the potential of this model, we present a series of experiments below. We first test our model on benchmark data without observed covariates, in comparison to LDA, SAGE, and NVDM. We present two versions of our model – a simpler model than NVDM which uses only a
single layer encoder \((f_e)\) and no generator \((e_1 g_0)\), and a more expressive model that uses a single-layer encoder \((f_e)\) and a 4-layer generator \((f_g)\), both with “shortcut” connections \((e_1s g_4s)\); see Appendix for details).

To compare with NVDM, we adopt the configuration of their inference network, with 500-dimensional layers and ReLu non-linearities for all of our models. We report results from own implementation of NVDM, however, and obtain better perplexity results than were originally reported, perhaps due to differences in preprocessing. We optimize these models using Adagrad \((Duchi et al., 2010)\), with grid search to choose the base learning rate. For SAGE, we use the original Matlab implementation \(\), while for LDA we use Mallet \(\). We also investigate the effects of stopword removal and different levels of sparsity.

Our evaluation is based on two datasets:

- the classic 20 newsgroups dataset, for which we use the standard train/test split;
- the full collection of NIPS papers from 1987 to 2016, for which we randomly sample 20% of the papers as a test set \((Tan et al., 2017)\).

As in past work, we report the value of variational bound on test data as an upper bound on perplexity for our model and NVDM, using 20 samples per document to evaluate the bound for test data \((Miao et al., 2016; Srivastava and Sutton, 2017)\).

We then demonstrate the flexibility our model by applying multi-facted variations to the NIPS dataset, and to the CMU political blogs corpus – a collection of approximately 13,000 liberal and conservative blog posts from 2008–2009 \((Eisenstein and Xing, 2010)\). For the blogs data, we consider the interactions of topics and ideology, as was done with SAGE. For the NIPS data, we implement a customized model with an additional regularizer to encourage topic deviations that change slowly over time, similar to the motivating assumption of the dynamic topic model \((Blei and Lafferty, 2006)\). To make this modification, we use year of publication as an observed covariate (breaking the data into ten blocks of 3 years each), and then simply add an extra term to our objective that penalizes the \(l_1\) norm of the differences between neighboring time periods of topic-year interactions weights \((W_2)\). This model would have been difficult to implement in the original SAGE framework, as it would have required deriving and implementing modifications to the variational update equations, which may or may not have involved difficult-to-optimize parameters.

### 3.2 Perplexity Benchmarks

In this section we compare the performance of our model to that of SAGE, LDA, and NVDM. Results are given in Table \(\).

Using our more complex model (with rich encoder and generator networks, but without sparsity) we outperform NVDM and LDA in terms of perplexity in all cases, and our simple model also does nearly as well. This finding confirms that modeling deviations from a background distribution is useful in modeling texts, as \((Eisenstein and Xing, 2010)\) argued.

The results for SAGE are mixed, as was found in the original paper. Without removing stopwords, it obtains the lowest perplexity on 20 newsgroups, but it does much worse than LDA (and our model) when Mallet stopwords have been removed. The performance of NVDM is similar to our model for 50 topics, but considerably worse for 200 topics.

Using the approach to adaptive sparsity described above, we also experiment with different levels of sparsity in our topic network, and find that as we increase the level of sparsity, perplexity tends to suffer. As we will see below, however, sparsity tends to improve the cohesiveness of topics overall.

Last but not least, we note that stopword removal has a major effect on performance, as has been noted elsewhere \((Schofield et al., 2017)\). Furthermore, the results we obtain from LDA on the 20 newsgroup data are much stronger than previously published results \((Miao et al., 2016)\).

---

1. https://github.com/dallascard/neural_topic_models
2. https://github.com/jacobeisenstein/SAGE
3. http://mallet.cs.umass.edu/
Table 1: Perplexity results on held out data (lower is better) for two datasets with varying numbers of topics. Note that the results for LDA and SAGE were estimated using Monte Carlo methods (Wallach et al., 2009), whereas all other results are upper bounds on perplexity based on the ELBO. In NIPS data, since the texts are math-heavy, we remove words that have less than three letters or have non-alphabetical characters.

2016; Hinton and Salakhutdinov, 2009) although comparable to another more recent paper (Srivastava and Sutton, 2017). Overall, the performance gap between NVDM and LDA is much smaller than previously reported, suggesting that previous work has underestimated the performance of LDA, though small differences may relate to different preprocessing decisions and use of different test sets. By making our code and preprocessing scripts public, we hope to facilitate easy comparisons in the future.

3.3 Topic Coherence and Interpretability

The above results demonstrate that our model is able to better approximate the true distribution of held out data better than LDA and NVDM. However, it has been shown that perplexity does not necessarily correlate well with topic coherence (Chang et al., 2009; Srivastava and Sutton, 2017). To test this, we use an automated heuristic evaluation of coherence based on the normalized pointwise mutual information (NPMI) among the top 10 words in each topic, using the full data set (train and test) as a background corpus. This metric has been show to correlate well with human judgments (Newman et al., 2010; Lau et al., 2014).

In Table 2 we report mean NPMI results for the various models on the 20 newsgroups data with 50 topics and Mallet stopwords removed, along with some example topics. As can be seen, increasing sparsity in our model results in much better coherence, at least according to NPMI. This suggests that we can use the strength of regularization to trade off perplexity and coherence as desired. Although SAGE obtains the best overall coherence, our model with 80% sparsity produces more coherent topics than LDA and almost as coherent as SAGE.

4Unless we explicitly remove them, stopwords tend to dominate the top words of most topics returned by vanilla LDA with default settings.
Table 2: Mean coherence using NPMI (higher is better) and example topics returned by various models on 20 newsgroups with a 2000 word vocabulary, 50 topics, and Mallet stopwords removed.

3.4 Qualitative Results on SAGE-style models

Blogs data: The CMU political blogs corpus consists of posts representing two ideological perspectives (liberal and conservative). Table 3 shows examples that demonstrate how different perspectives emphasize different aspects of a similar topic. The first line is the base topic ($W$) and following lines show the liberal and conservative variations ($W_1 + W_2$). For example, in the first topic on the economic crisis, the conservative side is more likely to reminisce about the Reagan presidency.

In our experiments, we find that running our model without observed covariates results in coherent topics which tend to be highly polarized: they are used primarily by one side or the other. Introducing deviations for ideology ($W_1$) captures some common trends across all topics: we find both sides are more likely to refer to figures from the opposing ideological perspectives, i.e., “bush”, “cheney”, and “rove”, vs. “clintons”, “hillary”, and “ayers”. Adding deviations for interactions between topics and perspective ($W_2$) results in topics that are more evenly used across perspectives, and provides an indications of how these topics vary in usage between the left and the right as illustrated in Table 3.

NIPS data: Table 4 shows some examples from our model with a temporal regularizer applied to the NIPS data. The top line shows two topics, and each following line shows variations for the corresponding three-year period ($W_1 + W_2$). We find we can detect the clear emergence of certain trends, like $svm$ and $cnn$. We can also effectively control the smoothness of variation over time, although this sometimes results in somewhat mixed topics, as in the right column of Table 4: “nonlinear” has switched from the analog setting to the optimization setting.

4 Additional Related Work

In addition to SAGE, other variations on topic models have been proposed for multi-facted models, including hierarchical models (Blei et al., 2010; Nguyen et al., 2013, 2015). The correlated topic model...
Table 3: Examples of topics found in the CMU political blogs corpus along with variations for each ideological perspective.

| Years          | primal lagrangian margin nesterov hinge | nonlinear sigmoidal analog vlsi multiplication |
|----------------|----------------------------------------|-----------------------------------------------|
| 1987-89        | pixel kernel convex optimization primal | nonlinear analog nonlinearity vlsi hardware   |
| 1990-92        | margin regularization optimization handwritten | chip vlsi circuits sigmoidal neuron node     |
| 1993-95        | convolutional speaker convex manifold tangent | nonlinear sigmoidal analog vlsi chip digital |
| 1996-98        | svm convex riemannian relaxation margin svms | analog vlsi chip sigmoidal multilayer circuit |
| 1999-01        | convex primal margin svm regularized tangent | circuit nonlinear sigmoidal kernel smola   |
| 2002-04        | margin convex laplacian affinity eigenvectors | circuit nonlinear kernel circuits vlsi analog |
| 2005-07        | convex primal margin hinge dual laplacian | vlsi ica analog chip nonlinear subspace pca |
| 2008-10        | eigenvectors primal hinge nesterov eigenvector | cnn nonlinear kernels lipschitz deep vlsi |
| 2011-13        | lagrangian primal nesterov margin laplacian | cnn hilbert tangent pca ica nonlinear kernels |
| 2014-16        | primal bregman lagrangian nesterov duality | backpropagation kernel compression hilbert |

Table 4: Two examples of gradually changing variations on base topic over ten 3-year periods.

More recently, Srivastava and Sutton (2017) also applied VAE-style inference to topic models, demonstrating they could approximate the original LDA model. They also introduced a new model, Product of Experts LDA. Although their perplexity numbers are inferior to vanilla LDA, they argue for their model primarily in terms of topic coherence. However, they do not investigate sparsity or the type of multi-facted modeling that we present here. Others have also applied VAE-style inference to a variety of other NLP tasks (Bowman et al., 2016; Marcheggiani and Titov, 2016).

Finally, LDA can also be cast as a matrix factorization problem, for example using the Poisson-gamma framework (Cemgil, 2009; Paisley et al., 2014; Ranganath et al., 2015). Unfortunately, the reparameterization trick cannot be applied directly to the gamma or Dirichlet distributions. A few recent papers have sought to get around this restriction, either through transformations (Kucukelbir et al., 2016) or a generalization of reparameterization (Ruiz et al., 2016). Black-box and VAE-style inference have been implemented in at least two general purpose tools designed to allow rapid exploration and evaluation of models (Kucukelbir et al., 2015; Tran et al., 2016).

5 Conclusion

We have presented a neural framework for generalized topic models to enable rapid exploration of models with covariates, interactions, and customized regularizers. We take advantage of sampling-based variational inference to develop a general algorithm for our framework such that variations do not require any model-specific algorithm derivations. Overall, our framework demonstrates strong performance in modeling texts. Our work can further support experimenting with diverse modeling options to incorporate multiple facets, and the use of sparsity allows us to trade off perplexity against topic coherence.
Acknowledgements

This research was made possible by a Natural Sciences and Engineering Research Council of Canada Postgraduate Scholarship (to D.C.) and a University of Washington Innovation Award.

Appendix

Preprocessing for 20 newsgroups. For the 20 newsgroups dataset, we download the articles using the scikit-learn interface, without removing headers, footers or quotes. We parse the text using spaCy and convert all characters to lower case. Optionally, we then exclude stopwords using the list of standard stopwords in Mallet. We then keep the 2000 words which appear in the largest number of documents.

Preprocessing for NIPS. For this dataset, we use the same processing as above, except that we use a vocabulary size of 10,000 words, and we exclude all tokens which involve any symbols other than alphabetic characters, and drop all tokens of length less than 3, in order to avoid ambiguous tokens like section numbers and mathematical symbols.

Residual network style short cuts. In both the generator network and the encoding network, we employ a shortcut variation that is inspired by residual networks. Figure 2 uses a four-layer generator network as an example. The key idea is to add a linear shortcut between $z$ and the output so that the deep network only learns the residual and the feedback can be efficiently backpropagated to the parameters.

Figure 2: A four-layer generator network with shortcuts.

KL cost annealing and variance initialization. We follow Bowman et al. (2016) and use the KL cost annealing trick in training. Specifically, we add a weight variable to the KL divergence term in Equation (5) of the main paper. We set the weight to zero at the start of training, and increases the weight linearly to 1 in the first 20 epochs. This allows the model to encode information in $z$ as if it were an auto-encoder without being constrained by the prior initially, and gradually optimize the variational lower bound in subsequent epochs.

In a similar spirit, for the final layer of the inference network that produces the variance of the variational distribution ($\log \sigma^2$), we initialize the bias term to be $-4$ and the weight matrix to be 0 so that the sampler approximates a deterministic scenario and effectively learn the parameters for $\mu$ values.

References

Amr Ahmed and Eric P. Xing. Staying informed: Supervised and semi-supervised multi-view topical analysis of ideological perspective. In Proceedings of EMNLP, 2010.

David M. Blei. Build, compute, critique, repeat: Data analysis with latent variable models. Annual Review of Statistics and its Application, 1:203–232, January 2014.

David M. Blei and John D. Lafferty. Dynamic topic models. In Proceedings of ICML, 2006.

6 https://spacy.io/

7 A line of research has reported perplexity numbers based on preprocessing proposed in Hinton and Salakhutdinov (2009). In that paper they describe their preprocessing as removing stop-words, stemming, filtering the vocabulary to the 2000 most common words, and then transforming the counts by apply the transformation $c_{ij} = \log(1 + c_{ij})$. They do not provide a list of stopwords, however, and stemming may produce different results depending on the software used. We do not apply this transformation, and find that stop word removal has a major effect on perplexity, as has past work, thus it is essential to standardize for a fair comparison.
David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.

David M. Blei, Thomas L. Griffiths, and Michael I. Jordan. The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. *J. ACM*, 57(2), February 2010.

Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew M. Dai, Rafal Józefowicz, and Samy Bengio. Generating sentences from a continuous space. In *Proceedings of CONLL*, 2016.

Tamara Broderick, Nicholas Boyd, Andre Wibisono, Ashia C Wilson, and Michael I Jordan. Streaming variational bayes. In *Proceedings of NIPS*. 2013.

Ali Taylan Cemgil. Bayesian inference for nonnegative matrix factorisati models. *Computational Intelligence and Neuroscience*, pages 4:1–4:17, January 2009.

Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L Boyd-graber, and David M Blei. Reading Tea Leaves: How Humans Interpret Topic Models. In *Proceedings of NIPS*, 2009.

Michael Collins, Sanjoy Dasgupta, and Robert E. Schapire. A generalization of principal component analysis to the exponential family. In *Proceedings of NIPS*, 2001.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. Technical report, EECS Department, University of California, Berkeley, 2010.

Jacob Eisenstein and Eric Xing. The CMU 2008 political blog corpus. Technical report, Carnegie Mellon University, 2010.

Jacob Eisenstein, Amr Ahmed, and Eric P. Xing. Sparse additive generative models of text. In *Proceedings of ICML*, 2011.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

Geoffrey E Hinton and Ruslan R Salakhutdinov. Replicated Softmax: An Undirected Topic Model. In *Proceedings of NIPS*, 2009.

Matthew Hoffman, Francis R Bach, and David M Blei. Online Learning for Latent Dirichlet Allocation. In *Proceedings of NIPS*, 2010.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *Proceedings of ICLR*, 2014.

Alp Kucukelbir, Rajesh Ranganath, Andrew Gelman, and David Blei. Automatic variational inference in stan. In *Proceedings of NIPS*, 2015.

Alp Kucukelbir, Dustin Tran, Rajesh Ranganath, Andrew Gelman, and David M. Blei. Automatic differentiation variational inference. *CoRR*, abs/1603.00788, 2016. URL [http://arxiv.org/abs/1603.00788](http://arxiv.org/abs/1603.00788).

John D. Lafferty and David M. Blei. Correlated topic models. In *Proceedings of NIPS*, 2006.

Jey Han Lau, David Newman, and Timothy Baldwin. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of EACL*, 2014.

Diego Marcheggiani and Ivan Titov. Discrete-state variational autoencoders for joint discovery and factorization of relations. *TACL*, 4:231–244, 2016.

Jon D. McAuliffe and David M. Blei. Supervised topic models. In *Proceedings of NIPS*, 2008.

Yishu Miao, Lei Yu, and Phil Blunsom. Neural variational inference for text processing. In *Proceedings of ICML*, 2016.

Andriy Mnih and Karol Gregor. Neural variational inference and learning in belief networks. In *Proceedings of ICML*, 2014.

David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. Automatic evaluation of topic coherence. In *Proceedings of ACL*, 2010.

Viet-An Nguyen, Jordan L. Boyd-Graber, and Philip Resnik. Lexical and hierarchical topic regression. In *Proceedings of NIPS*, 2013.
Viet-An Nguyen, Jordan Boyd-Graber, Philip Resnik, and Kristina Miler. Tea party in the house: A hierarchical ideal point topic model and its application to Republican legislators in the 112th congress. In Proceedings of ACL, 2015.

John William Paisley, David M. Blei, and Michael I. Jordan. Bayesian nonnegative matrix factorization with stochastic variational inference. In Handbook of Mixed Membership Models and Their Applications., pages 205--224. 2014.

Rajesh Ranganath, Sean Gerrish, and David Blei. Black Box Variational Inference. In Proceedings of AISTATS, 2014.

Rajesh Ranganath, Linpeng Tang, Laurent Charlin, and David M Blei. Deep exponential families. In Proceedings of AISTATS, 2015.

Danilo J. Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In Proceedings of ICML, 2014.

Molly Roberts, Brandon Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Gadrarian, Bethany Albertson, and David Rand. Structural topic models for open ended survey responses. American Journal of Political Science, 58:1064--1082, 2014.

Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. The author-topic model for authors and documents. In Proceedings of UAI, 2004.

Francisco J R Ruiz, Michalis K Titsias, and David M Blei. The generalized reparameterization gradient. In Proceedings of NIPS, 2016.

Alexandra Schofield, Måns Magnusson, and David Mimno. Pulling out the stops: Rethinking stopword removal for topic models. In Proceedings of EACL, 2017.

Akash Srivastava and Charles Sutton. Autoencoding variational inference for topic models. In Proceedings of ICLR, 2017.

Chenhao Tan, Dallas Card, and Noah A. Smith. Friendships, rivalries, and trysts: Characterizing relations between ideas in texts. In Proceedings of ACL, 2017.

Dustin Tran, Alp Kucukelbir, Adji B. Dieng, Maja Rudolph, Dawen Liang, and David M. Blei. Edward: A library for probabilistic modeling, inference, and criticism. arXiv preprint arXiv:1610.09787, 2016.

Hanna M Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. Evaluation Methods for Topic Models. In Proceedings of ICML, 2009.

Dani Yogatama, Bryan R. Routledge, , and Noah A. Smith. A sparse and adaptive prior for time-dependent model parameters. In NIPS Time Series Workshop, 2015.