# UPDATE: Variational Autoencoders for Topic Modeling

*Kelly Geyer*
klgeyer@bu.edu

## 1. Task

In this project, we want to explore potential improvements for topic modeling, using Variational Autoencoders (VAEs) [1]. Topic models are widely used for unsupervised learning of groupings, or *topics*, for a given corpus of documents. They are used in a variety of applications, which include computational social science and bioinformatics. While there are several approaches to topic modeling, in this project we focus on improving the probabilistic topic models that assume the generative process of Latent Dirichlet Analysis (LDA) [2]. In particular, this class of topic models uses Baysian modeling to estimate both the word-topic and document-topic probabilities. In turn, these topic probabilities may be either visualized or clustered to summarize a large corpus of text. In recent works, probabilistic topic modeling has been improved in both performance and computational efficiency using VAEs [4,5,6].

**Motivation:** VAEs are a viable alternative to traditional LDA topic models. They improve upon several of the challenges of LDA, as explained below:

1. *Computational efficiency* - Probabilistic topic models such as LDA are estimated using Gibbs sampling. The draw-back to this approach is that it is very slow in practice. VAEs use techniques that frame estimation as an optimization problem, and can be solved much more quickly by means of Variational Inference (VI).
2. *Interpretability* - The document-topic and word-topic probabilities are intuitive variables for summarizing a corpus of documents. However, in practice the resulting topic distributions may appear as nonsensical, and the output may be prone to subjectivity in interpretation [3]. Enforcing sparsity upon the topic probability matrices, or having topics overlap less, may make the output more interpretable. To counter this, there is an emerging body on extending VAEs to sparse data and regularization: [7,8,9]. Another issue of interpretability is selecting the number of topics. However, we will control the number topics, as explained later in Section 4.

## 2. Related Work

There are efforts of topic modeling that have stood out in my literature search. This project draws significant inspiration from LDA topic modeling [1]. As discussed in Section 1, this is a generative model for estimating the document-topic and document-word probabilities. In the motivation, discussed how this model can fall short in expectations of interpretable results. Additionally, using an MCMC sampler for this model is slow to converge.

In the work [10], the authors improve the scalability and speed of LDA by using stochastic variational inference to estimate posterior distributions.

The approach from [5], called ProdLDA improves training and topic interpretation. It is the first VAE inference method specifically for topic models (in cases where K is fixed). ProdLDA explicitly approximates the Dirichlet prior (i.e., topic generating distribution), which largely provides more interpretation information. Additionally, ProdLDA produces better topics than traditional LDA (by their metrics), as well as improves the speed of model fitting.

## 3. Approach

At the moment, my approach is the following:

1. Prepare the 20 Newsgroup data set in an easy-to-use format.
2. Perform topic modeling using traditional LDA, and stochastic VI LDA. Python's Gensim implements LDA.
3. Implement ProdLDA from [5], and fit the model to the 20 Newsgroup data set.
4. Interpret and compare results. Understanding the results of these unsupervised models will require thoughtful and clear graphics. I will likely create these plots myself rather than relying on existing packages.
5. If time permits, repeat steps 1-4 using word embeddings instead of bag-of-word features (see Section 4 for details about data)

## 4. Dataset and Metric

The data set used for evaluation is a collection of news articles, covering various topics. In particular, this project uses the 20 Newsgroups data set [11]. This is a collection of nearly 20,000 articles, which cover 20 different topics such as sports, politics, technology, and several others. Mostly likely, I will use a subset of this data, say 3-7 topics, to evaluate the methods. A label, or *topic,* is provided for each article.

**Data-preprocessing:** The pre-processing of this data consists of creating a document-term frequency matrix of vocabulary count per document will be constructed from the articles. Stop-words (e.g. *and, or, there*, etc.) will be omitted from this matrix. Ultimately, this processing of data leads to a bag-of-words assumption. I expect that this will result in a very sparse matrix of counts. While there are more sophisticated and representative features that may be derived from text data (e.g., word embeddings), I will use simple features due to the short term nature of this project.

**Metrics:** Recall that our analysis is unsupervised, and we cannot use traditional error metrics. Instead, we want to evaluate the interpretability of the model output. We will consider there metrics of evaluation:

1. *Perplexity* - For this metric, the articles are divided into training and testing sets. A topic model is fitted to the training data, and perplexity is calculated using the log-likelihood. Ideally, a high value for log-likelihood would imply that the model is representative. Perplexity is inversely related to the log-likelihood so lower scores are desired,
$$perplexity(test\ docs) = exp(-L(test\ docs)/\#\ of\ words)$$
2. *Topic Coherence* - Topic coherence is a measure of how similar words are within a topic. Ideally, we want words within a topic to be highly associated in meaning.
3. *Qualitative analysis* - We will use two visual evaluations of the latent topics. First, we will create histograms of topic distributions over the vocabulary. Secondly, we can plot latent variables using t-SNE to observe clustering behavior against the assigned topic labels [12].
4. *Computational Time* - In addition to model fit, we are interested in computational efficiency. We will measure and compare computational time across methods.

**5. Preliminary Results**

So far, I have been able to complete steps 1-3 listed in the section Approach. I'm still working on interpretation of results (step 4).

**6. Detailed Timeline and Roles**

| Task | Deadline |
|---|---|
| Obtain results for LDA-type models | 03/09/20 |
| Obtain results for ProdLDA | 3/16/20 |
| Create and interpret results of topic models<br>● Code perplexity<br>● Code topic coherence | 04/25/20 |
| Prepare presentation | 4/27/20 |
| Prepare report | 05/01/20 |

**References**


1) Kingma, Diederik P., and Max Welling. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013).
2) Blei, David M. Probabilistic topic models, Communications of the ACM 55.4: 77-84, 2012.
3) Chang, Jonathan, Sean Gerrish, Chong Wang, Jordan L. Boyd-Graber, and David M. Blei. "Reading tea leaves: How humans interpret topic models." In *Advances in neural information processing systems*, pp. 288-296. 2009.
4) Miao, Yishu, Edward Grefenstette, and Phil Blunsom. Discovering discrete latent topics with neural variational inference. In Proceedings of the 34th International Conference on Machine Learning-Volume 70, pp. 2410-2419. JMLR. org, 2017.
5) Srivastava, Akash, and Charles Sutton. Autoencoding variational inference for topic models. arXiv preprint arXiv:1703.01488 (2017).
6) Burkhardt, Sophie, and Stefan Kramer. "Decoupling sparsity and smoothness in the dirichlet variational autoencoder topic model." *Journal of Machine Learning Research* 20, no. 131 (2019): 1-27.
7) Antelmi, Luigi, Nicholas Ayache, Philippe Robert, and Marco Lorenzi. "Sparse Multi-Channel Variational Autoencoder for the Joint Analysis of Heterogeneous Data." (2019).
8) Dai, Bin, Yu Wang, John Aston, Gang Hua, and David Wipf. "Connections with robust PCA and the role of emergent sparsity in variational autoencoder models." *The Journal of Machine Learning Research* 19, no. 1 (2018): 1573-1614.

9) Zhao, He, Piyush Rai, Lan Du, Wray Buntine, and Mingyuan Zhou. "Variational Autoencoders for Sparse and Overdispersed Discrete Data." *arXiv preprint arXiv:1905.00616* (2019).
10) Hoffman, Matthew D., David M. Blei, Chong Wang, and John Paisley. "Stochastic variational inference." *The Journal of Machine Learning Research* 14, no. 1 (2013): 1303-1347
11) Lang, Ken. Newsweeder: Learning to filter netnews. Machine Learning Proceedings. Morgan Kaufmann, 1995. 331-339, 1995.
12) Maaten, Laurens van der, and Geoffrey Hinton. Visualizing data using t-SNE. Journal of machine learning research 9, no.: 2579-2605, 2008.