LDA with stochastic optimization*

CS510 Final Project Report†

Meiruo Xiang
Department of Statistics,
University of Illinois
at Urbana-Champaign
Urbana, Illinois
mxiang3@illinois.edu

Yiqin Wu
Department of Statistics,
University of Illinois
at Urbana-Champaign
Urbana, Illinois
yiqinwu2@illinois.edu

Meng Du
Department of Statistics,
University of Illinois
at Urbana-Champaign
Urbana, Illinois
mengd2@illinois.edu

ABSTRACT

Based on online variational Bayes LDA, we modify the way of handling incoming unseen words, that is, put the new unseen words into a seperate identity instead of ignoring out of bag words and inspect the evolving topics by perplexity of this modified structure using in online VB LDA. Then we compare this modified method's performance with original online VB LDA (Hoffman et al. 2010) and online LDA with infinite vocabulary (Zhai and Boyd-Graber 2013) based on pointwise mutual information . [1]

CCS CONCEPTS

• **Information systems** → **Document topic models**; *Language models*; • **Theory of computation** → *Probabilistic computation*; • **Computing methodologies** → *Topic modeling*;

KEYWORDS

Latent Dirichlet Allocation, topic modeling, stochastic

ACM Reference Format:
Meiruo Xiang, Yiqin Wu, and Meng Du. 2018. LDA with stochastic optimization: CS510 Final Project Report. In *Proceedings of UIUC*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.475/123_4

1 INTRODUCTION

Clustering the dynamic text data can be very helpful in text mining and information retrieval area. It is a fundamental step for text crawling, document organization and topic detection. An efficient algorithm for clustering text streams can be widely used in education, entertainment, business and other fields. The aim of our research is to construct a topic model for dynamic text streams, which takes time effect into account and try to improve clustering stability based on existing models.

There are some researches that relates to this research question. Blei and Lafferty[1] proposed a dynamic topic model, using logistic normal distribution to model the change of hyperparameters of word distribution and topic distribution instead of Dirichlet distribution to show the topic evolution over time. Hoffman, Blei and Bach[2] developed an online variational Bayes algorithm for LDA, which is based on online stochastic optimization with a natural gradient step. This online LDA model can analyze massive document collections, including those arrivs in a stream. Based on this model, K.Zhai and J.Graber[6] extend LDA by drawing topics from a Dirichlet process whose base distribution is a distribution over all strings rather than from a finite Dirichlet distribution. There model can successfully incorporate new words and has a better performance than topic models with finite vocabularies. Also, Shangsong Liang, Emine Yilmaz and Evangelos Kanoulas[4] proposed a new dynamic clustering topic model that enables tracking the time-varying distributions of topics over documents and words over topics by using a short-term or long-term dependency model over sequential data.Their model overcomes the difficulty of handling short text.

In addition to these methods, we have developed a new method which is also based on Hofman's online LDA model[2]. Since this online LDA method will skip the words when new words in document comes in, inspired by idea of add-one smoothing, we developed a new way to include new words into vocabulary. Our methods has three steps: variational inference, stochastic optimization, and dynamic method for LDA. Among these three steps, our proposed idea is embedded for including new words.

After introducing our improved method, we will compare this method along with the original online LDA method proposed by Blei [3] and the upgraded online LDA with infinity vocabulary method proposed by Zhai [6]. Using perplexity which is introduce in [2] and PMI with co-occurrence statistics which is introduced by Newman [5] to evaluate the goodness-of-fit and topic coherence respectively. Finally, we will use visualization to show our results.

2 METHODOLOGY

2.1 Topic modeling

A topic model is a statistical model used in text mining to find the hidden semantic structures in a collection of documents. The "topic" produced by topic modeling techniques is a cluster of similar words, here the topic is the main idea discussed in text data. Many application requires discovery of topics in text, for example, find out the major topics in presidential election debating. The thematic model captures this intuition in a mathematical framework that allows checking a set of documents based on the statistics of words in each word and finding out what those topics might be and what

*Produces the permission block, and copyright information
†The full version of the author's guide is available as `acmart.pdf` document
[1]This is an abstract footnote

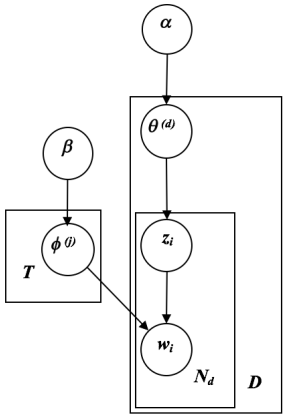
Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
UIUC, Dec 2017, CS510 Project
© 2017 Copyright held by the owner/author(s).
ACM ISBN 123-4567-24-567/08/06.
https://doi.org/10.475/123_4

the thematic balance for each document is. There are many techniques to do topic modeling, the most widely used and one of the simplest method is Latent Dirichlet Allocation (LDA), which we are going to introduce in next section.

2.2 LDA

Latent Dirichlet Allocation is generative topic modeling method, using Bayesian inference to infer the topic coverage and topic word distributions. We can generate the words for document by sampling a topic assignment z from the topic distribution θ, and sampling a word from word distribution ϕ with respect to the corresponding topic, where topic distribution and word distribution has a prior distribution which is Dirichlet distribution.

It can be interpret as the following graph model, where α and β are Dirichlet priors for distribution over topics for each document and distribution over words for each topic respectively. $\phi^{(j)} \sim Dirichlet(\beta)$ is the distribution over words for topic j, and $\theta^{(d)} \sim Dirichlet(\alpha)$ is the distribution over topics for document d. $z_i \sim Discrete(\theta^{(d)})$ is the topic assignment for each word. $w_i \sim Discrete(\phi^{(zi)})$ is the word generated from assigned topic.



2.3 Online LDA model

2.3.1 Variational inference. To implement the LDA model, we choose to use the variantional inference method. One of the basic methods of variantional inference is naÃŕve mean-field methods. The naive mean-field method will use the product of several independent distributions to approximate the target distribution. Using the tractable distribution to approximate the complex distribution will help the computation in the model learning become feasible. Here is the basic concept of naive mean-field methods:

$$p(x) \quad x \in x^n \tag{1}$$

is our target distribution, which will be approximated by

$$q(x) = \prod_{i=1}^{n} q_i(x_i) \tag{2}$$

And the method for inference is to find the best $\{q_i\}_{i=1}^n$ that will minimize the KL divergence of q and p that is $D(q||p)$. This is an optimization problem whose target function $D(q||p)$ is convex. However, the tractable set of q is not convex. Fortunately for each term in the product, minimizing KL divergence over q_i alone (with the other coordinates fixed) is a convex program. Then, we can derive the following equations:

$$\begin{aligned} D(q||p) =& E_q[\log \frac{\prod_{i=1}^n q_i(x_i)}{p(X)}] \\ =& \sum_{i=1}^{n} E_q[logq_i(X_i)] - E_q[logq_i(X_i)] \\ =& \sum_{i=1}^{n} \sum_{x_i \in X} q_i(x_i) logq_i(x_i) - \sum_{x \in X^n} (\prod_{i=1}^{n} q_i(x_i)) logp(x) \end{aligned}$$

Define the Lagrangian

$$L(q, \lambda) = D(q||p) + \sum_{i=1}^{n} \lambda_i (\sum_{x_i} q_i(x_i) - 1) \tag{3}$$

where the Lagrange multipliers $\lambda_i, 1 \le i \le n$ are associated with the equality constraints $\sum_{x_i} qi(xi) = 1$ for each i. The non-negativity constraints on $\{q_i\}$ are momentarily ignored. In our problem the true posterior is approximated by $q(z, \theta, \beta)$. To minimize the KL-divergence between the target distribution, the posterior distribution $p(z, \theta, \beta|w, \alpha, \eta)$ in this problem and $q(z, \theta, \beta)$ can be considered as a optimization problem over maximizing the Evidence Lower BOund(ELBO):

$$\log p(w|\alpha, \eta) \ge L(w, \phi, \gamma, \lambda) \tag{4}$$

$$\stackrel{\Delta}{=} E_q[\log p(w, z, \theta, \beta|\alpha, \eta)] - E_q[\log q(z, \theta, \beta)] \tag{5}$$

And for the tractable sets of q, we will use the following:

$$q(z_{di} = k) = \phi_{dw_{wi}k} \tag{6}$$

$$q(\theta_d) = \text{Dirichlet}(\theta_d; \gamma_d) \tag{7}$$

$$q(\beta_k) = \text{Dirichlet}(\beta_k; \lambda_k) \tag{8}$$

The posterior over the per-word topic assignments z follows the distribution with parameter ϕ, the posterior over the per document topic weights θ follows the distribution with parameter θ, and the posterior over the topics β follows the distribution with parameter λ. The the ELBO can be written as:

$$\begin{aligned} \mathcal{L}(w, \phi, \gamma, \lambda) \quad = \quad & \sum_d \{E_q[\log p(w_d|\theta_d, z_d, \beta)] + E_q[\log p(z_d|\theta_d)] \\ - \quad & E_q[\log q(z_d)] + E_q[\log p(\theta_d|\alpha)] \\ - \quad & E_q[\log q(\theta_d)] + (E_q[\log p(\beta|\eta)] \\ - \quad & E_q[\log q(\beta)])/D\} \end{aligned}$$

To further apply the variational inference methods into the dynamic data streams, the above equation is based on data in one batch or in one incoming stream. D in the equation stands for the number of documents.

Then we can put the assumptions in (6)-(8) into this equation:

$$\begin{aligned}\mathcal{L} &= \sum_d \sum_w n_d w \sum_k \phi_{dwk}(E_q[\log\theta_{dk}] + E_q[\log\beta_{kw}] \\ &\quad - \log\phi_{dwk}) - \log\Gamma(\sum_k \gamma_{dk}) + \sum_k (\alpha - \gamma_{dk})E_q[\log\theta_{dk}] \\ &\quad + \log\Gamma(\gamma_{dk}) + (\sum_k - \log\Gamma(\sum_w \lambda_{kw}) \\ &\quad + \sum_w (\eta - \lambda_{kw})E_q[\log\beta_{kw}] + \log\Gamma(\lambda_{kw}))/D \\ &\quad + \log\Gamma(K\alpha) - K\log\Gamma(\alpha) + (\log\Gamma(W\eta) - W\log\Gamma(\eta))/D \\ &= \sum_d l(n_d, \phi_d, \gamma_d, \lambda)\end{aligned}$$

where W is the size of the vocabulary and D is the number of the documents. And $l(n_d, \phi_d, \gamma_d, \lambda)$ stands for the ELBO of each documents.

In variational inference, the optimization problem for naive mean field is a convex problem, based on the Lagrangian equation, the closed formed optimization answer can be found:

$$\begin{aligned}0 = \frac{\partial L(q,\lambda)}{\partial q_i(x_i)} &= 1 + \log q_i(x_i) \\ &\quad - \sum_{x':x'_i=x_i} (\prod_{i\neq j} (q_j(x'_j))\log p(x') + \lambda_i),\end{aligned}$$

$$1 \le i \le n, x_i \in X$$

so, the result will be:

$$q_i(x_i) = \frac{1}{Z_i} exp\{E_{\prod_{j\neq i} q_j[\log p(X_{1:i-1}, x_i, X_{i+1:n})]}\}$$

$$1 \le i \le n, x_i \in X$$

So, in this case,we will have the following result:

$$\phi_{dwk} \propto exp\{E_q[\log\theta_{dk}] + E_q[\log\beta_{kw}]\} \tag{9}$$

$$\gamma_{dk} = \alpha + \sum_w n_{dw}\phi_{dwk} \tag{10}$$

$$\lambda_{kw} = \eta + \sum_d n_{dw}\phi_{dwk} \tag{11}$$

And the expectation result for $\log\theta$ and $\log\beta$ are:

$$E_q[\log\theta_{dk}] = \Psi(\gamma_{dk}) - \Psi(\sum_{i=1}^{K}\gamma_{di}) \tag{12}$$

$$E_q[\log\beta_{kw}] = \Psi(\lambda_{kw}) - \Psi(\sum_{i=1}^{W}\lambda_{ki}) \tag{13}$$

So, to do the update based on the streaming data, we can use EM algorithm based on the equations above. This is a iterative method and will converge due to the features of EM algorithm.

2.3.2 Stochastic optimization. The optimization algorithm above is still for the fixed documents. To further extend the algorithm, instead of training the whole data set repetitively, we will only train the incoming new data and use some stochastic method to modify the result. Consider the problem of minimizing a target function $f(x)$. By the increasing of the data, it will be more expensive to compute the optimization function based on the whole data. So, we can compute it based on the data coming at time t. This can be considered as noisy version of the original optimization target function $g(x, Z)$, Here Z is a random variable drawn from some distribution P such that $f(x) = E[g(x; Z)]$. In the dynamic streaming situation, the $g(x; Z)$ can be considered as $f(x_t)$. And this will be guaranteed to converge by the Law of Large Numbers. So, instead of computing the $\nabla f(x)$, we will calculate the $\nabla f(x)$. This means, we will do the variational inference for the incoming data at time t instead of all the data. After optimize the noisy version of target function, we will modify the result based on the stochastic gradient decent(SGD) method. Here is the SGD algorithm: where a_n is the decent step. And here are two limitation for the a_n:

ALGORITHM 1: Stochastic Gradient Decent

```
Initialize X_1;
for n = 2, 3... do
    draw Z_n P;
    X_{n+1} = X_n − a_n ∇g(X_n, Z_n)
end
```

$$\sum_{i=1}^{\infty} a_n = \infty$$

$$\sum_{i=1}^{\infty} a_n^2 \le \infty$$

These two remarks are very important for the convergence of the algorithm.

2.3.3 Dynamic method for LDA . In section 2.3.2, we have discussed use the learning period for time t, which is similar to the method in section 2.3.1. The following work will be combing the single learning with the dynamic updating. First, we try to update λ by using the SGD, here we define the decent step to be $\rho_t = (\tau_0 + t)^{-\kappa}$, to satisfy the convergence condition, we set $\kappa \in (0.5, 1)$. Then the update for λ will be:

$$\widetilde{\lambda}_{kw} = \eta + \frac{D}{S}\sum_s n_{tsk}\phi_{tskw} \tag{14}$$

Then the algorithm will be in the table.

ALGORITHM 2: Online Variational Inference for LDA

```
Define ρ_t = (τ_0 + t)^{−κ}
Initialize lambda randomly;
for t = 0 to ∞ do
    E step: Initialize γ_tk = 1 repeat
        Set φ_twk ∝ exp{E_q[log θ_tk] + E_q[log β_kw]};
        Set γ_tk = α + Σ_w φ_twk n_tw;
    until (1/K) Σ_k |change in γ_tk| < 0.00001;
    M step:
    Compute λ̃_kw = η + D Σ_s n_tsk φ_tskw;
    Set λ = (1 − ρ_t) + ρ_t λ̃
end
```

To update the hyperparameters α and η, we can use the linear-time Newton-Raphson method if γ and λ are given in [2].

Then the update for α and η will be:

$$\alpha \leftarrow \alpha - \rho_t \widetilde{\alpha}(\gamma_t) \tag{15}$$

$$\eta \leftarrow \eta - \rho_t \widetilde{\eta}(\lambda) \tag{16}$$

where $\widetilde{\alpha}(\gamma_t)$ is the inverse of the Hessian times the gradient $\nabla_\alpha l(n_t, \gamma_t, \phi_t, \lambda)$, $\widetilde{\eta}(\lambda)$ is the inverse of the Hessian times the gradient$\nabla_\eta \mathcal{L}$

2.4 Online LDA with infinite vocabulary

For online LDA with infinite vocabulary (Zhai and Boyd-Graber, 2013), the only difference of its model from traditional one is the way to define the word distribution of each topic, β_k, $k = 1, ..., K$ since the goal is to make the distribution be able to extend to any number of words in vocabulary. Here author used Dirichlet process DP(α^β, G_0) (Ferguson 1973), α^β is the scale parameter and G_0 is the base distribution. Now β is the parameter vector of an infinite multinomial,

$$\beta_i \equiv b_i \prod_{j=1}^{i-1}(1 - b_j),\ G \equiv \sum_i \beta_i \delta_{\rho_i} \tag{17}$$

where $b_1, ..., b_i, ... \sim Beta(1, \alpha^\beta)$, $\rho_1, ..., \rho_i, ... \sim G_0$. β_i give the probability of selecting word atom ρ_i from the base distribution. To define G_0, Zhai uses a modified character language model,

$$G_0(\rho) \equiv p(l = |\rho||\boldsymbol{\lambda}) \prod_{i=1}^{|\rho|} p(c_i|\mathbf{c_{i-n}, ..., i-1})$$

where $p(l = |\rho||\boldsymbol{\lambda})$ is a multinomial distribution, then $\sum_l \lambda_l = 1$ and $|\rho|$ is the word's length.

$$\boldsymbol{\lambda} = argmin_{\boldsymbol{\lambda}} \sum_\rho |p_c(\rho) - p(\rho|\boldsymbol{\lambda})|^2$$

For variational inference of infinite vocalulary, it used a truncation ordered set (TOS) τ_k for each topic's vocabulary that map every unique word w to an integer t, which is the index of the atom ρ_{kt} that corresponds to w. Suppose there are D documents in the corpus with N_d words each, the joint likelihood is:

$$\begin{aligned} & p(W, \rho, \beta, \theta, z) \\ = & \prod_{k=1}^{K}[\prod_{t=1}^{\infty} p(\rho_{kt}|G_0) \cdot p(\beta_{kt}|\alpha^\beta)] \\ & \cdot[\prod_{d=1}^{D} p(\boldsymbol{\theta}_d|\alpha^\theta) \\ & \cdot \prod_{n=1}^{N_d} p(z_{dn}|\boldsymbol{\theta}_d) p(w_{dn}|z_{dn}, \beta_{z_{dn}})] \end{aligned}$$

Denote the latent variables $\beta, \theta, \mathbf{z}$ as $\mathbf{Z}$, and its distribution as q($\mathbf{Z}$). Then to minimize the Kullback-Leibler(KL) divergence between p($\mathbf{Z}$) and q($\mathbf{Z}$), it is same to maximize $\mathcal{L}$, which is the evidence lower bound (ELBO),

$$\mathcal{L} = E_{q(\mathbf{Z})}[log\, p(\mathbf{W}, \mathbf{Z})] - E_{q(\mathbf{Z})}[q] \tag{18}$$

and let

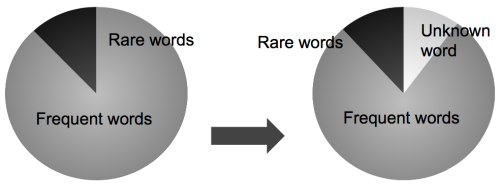
$$q(\mathbf{Z}) \equiv \mathbf{q}(\beta, \mathbf{z}) = \prod_{\mathbf{D}} \mathbf{q}(\mathbf{z_d}|\eta) \prod_{\mathbf{K}} \mathbf{q}(\mathbf{b_k}|v_{\mathbf{k}}^1, v_{\mathbf{k}}^2). \tag{19}$$

where $q(b_k|v_k^1, v_k^2)$ is a beta distribution and $q(z_d|\eta)$ is a signle distribution over K^{N_d} possible topic components. With TOS, use MCMC sampling to calculate $q(z_d|\eta)$ and stochastic variation inference for $q(\mathbf{b}|v)$ (Zhai and Boyd-Graber, 2013).

2.5 Processing of new incoming words

Since the online LDA method proposed by Blei and Hofman will skip the words when new words in document comes in, we developed a new way to include new words into vocabulary. Our method is inspired by idea of add-one smoothing. The major process is adding a new word 'unknowwordsss' which represents the set of unknown words we would meet when reading new documents into our vocabulary.

In order to let this created new word 'unknowwordsss' perform as any other normal words in corpus, we need to assign this word a frequency. We define the frequency of this word at time t=0 as the number of rare words in the stream of documents at time t=0. The following picture can directly illustrate this idea:



Rare words can be defined as words that has frequency less than a very small positive integer, here we define them as words that appeared only once in the stream of documents. The rare words will not be replaced. Then, when new stream of documents comes in, we can replace the word which is not in vocabulary with this created word 'unknowwordsss'. By this means, we can gather the information of new words.

3 IMPLEMENTATION

3.1 Task description

We will run the dynamic LDA method and the online LDA with infinity vocabulary method on the same data. The target for the simulation is:

- Compare the performance of different methods
- In the dynamic methods, we want to see the change for the LDA model when more data coming
- For the same method, we want to see the different performance amount different parameters

3.2 Dataset

We will use the DBLP data set for our simulation. This data set is available on http://dblp.uni-trier.de/

3.3 Implementation process

We will monitor streaming situation for the data simulation. In the streaming, 100 documents will be loaded in each time. Our method will train the incoming data every and have some outputs both for the model itself and the evaluation metrics.

For all models, we train the models into K=10 topics and use same topic choice distribution with Dirichlet prior with α^{θ}=1/K, learning rate κ=0.75.

3.4 Evaluation

3.4.1 Perplexity . Perplexity is measurement of goodness-of-fit for a probability model. It is calculated by the inverse of the probability of the test set, and normalized by the number of tokens in the test set. Here is how it calculates:

$$Perplexity(w_1 ... w_N) = exp\left\{-\frac{\sum_{i=1}^{N} logp(w_i|\alpha, \beta)}{N}\right\}$$

A language model with lower perplexity is better as the test set is assigned with higher probability. The perplexity of two language models can be compared only when they have same vocabulary.

3.4.2 Point-wise mutual information. Point-wise mutual information(PMI) has always been used to measure the association between two words. It is calculated by the log value of ratio of the joint probability of two words and the product of marginal probabilities of this two words. It can be write as the following formula:

$$PMI(x, y) = log\frac{p(x, y)}{p(x)p(y)}$$

Higher PMI value indicates the two words shares higher similar information. In our research, we use PMI to measure the coherence of each topic that LDA gives. As the Newman[1] proposed in his paper, for each topic generated by LDA model, we will choose the top ten words with the highest probability, and calculate the PMI value of every combination of two words chosen from the topic's top ten words. And then average the PMI value of each combination in order to get the PMI of one topic. In order to calculate PMI for each pair of words, we count the co-occurrence of this two words in a 5-word window in our dataset.

3.4.3 Topic based similarity. Using Hellinger distance, we can calculate two document's similarity based on their topic coverage respectively. Denote that there are K topics, then for any two documents in the corpus d_i and d_j, Hellinger distance is as below:

$$dist(d_i, d_j) = \frac{1}{2}\sum_{t=1}^{K}(\sqrt{p(t|d_i} - \sqrt{p(t|d_j)})^2$$

For documents which are more likely classified into the same topic, we expect that the Hellinger distance between them should be small and for those two which are more likely to be in difference topic, the corresponding Hellinger distance should be much larger. An obvious difference between the distances of those in same topics and different topics implies that the topic coverage of the document is mainly focused on the most likely topic (that is, less fuzzy the topic allocation is), implying a strong confidence of the most likely topic it belongs to.

3.5 Result

3.5.1 Perplexity . Following is the evolution of perplexity when new data coming. We can find that by the coming of new data, perplexing is keeping decreasing. This result may be even lower, if we used more data. Also, the performance of the method treating unseen words as a single new word is much better than the original online LDA. And we should notice that, when computing the perplexity of the LDA with unknown words, we did not use the output of "unknown words". Because only with the same vocabulary can the method be compared.

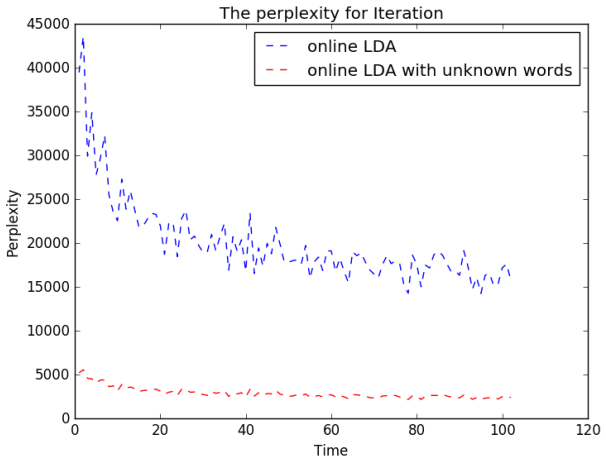


Figure 1: original online LDA and online LDA with unknown words

3.5.2 Point-wise mutual information. This is the result of PMI. In the evaluation procedure, we only use the training data set to be compute the concurrence statistics. Because our training set is not large enough, some of the output for PMI are zero. So, the result may not be reasonable enough to do the comparison.

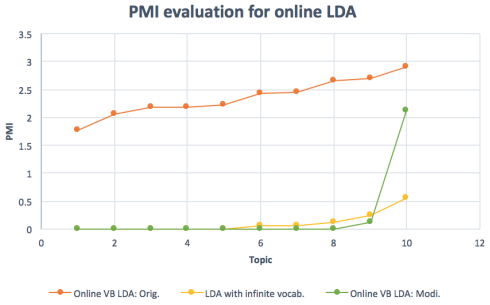


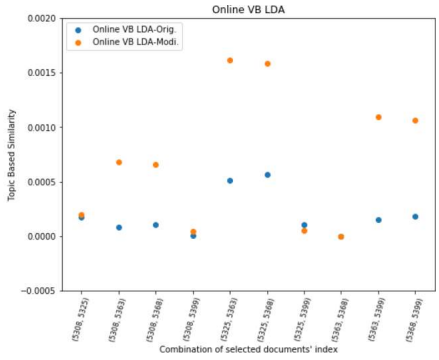
Figure 2: Point-wise mutual information of 3 methods: Original online VB LDA, Modified online VB LDA, online LDA with infinite vocabulary

3.5.3 Topic based similarity. For topic based similarity, we randomly select 5 documents in the last batch and inspect their similarities with each other. Here we selected documents 5308, 5325, 5363, 5368 and 5399 (index begins from 0). From the expectation of γ_d, all docs are in same topic in both online VB LDA model, and we can find that their similarities with each other are all very low. In Figure 3a, blue points represent original online LDA gather closer

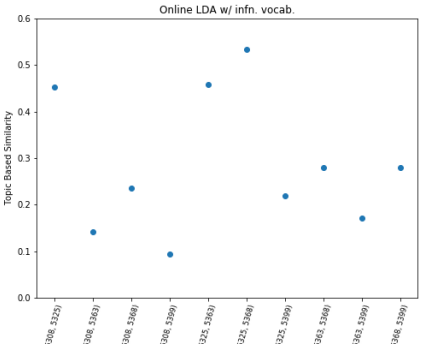
around 5e-4 while our modified method has orange points assemble more dispersively. In this case the original method is better than our modified method based on topic based similarity because closer points means more similar topic coverages for documents in the same topic.

In online LDA with infinite vocabulary model, doc 5325 is in topic 4 , different from other docs which are in topic 5. In Figure 3b, the plots which represent the similarities of 5325 with other documents 5308, 5363 and 5368 are obviously higher than other other plots. This makes sense since the topic coverage of 5325 if different from others. Although 3b shows explicit difference of similarities for those in same topic and those not, the limit of y axis simply shows it has more variance in topic coverage.

Figure 3 is the result for topic based similarity. The result is not very good. So, we used two images to illustrate the result. The range for similarity of LDA with infinity words is quite different from original LDA models. This is because of the vocabulary in this method will keep increasing which may have big impact on the topic based similarity.



(a) S=100, K=10, η=1/K, κ=0.75



(b) S=100, K=10, α^{β}=1k, κ=0.75

Figure 3: Topic based similarity among 5 selected documents of 3 methods: Original online VB LDA, Modified online VB LDA, online LDA with infinite vocabulary

4 CONCLUSION AND DISCUSSION

4.1 Conclusion

The online LDA model is very useful in modern days. Because of the data quantity has become larger and larger. Besides the streaming data situation, the online LDA model can also be applied to a fixed but big data set in which compute the variational inference result of the whole data set is not computational tractable. Then, we can use the dynamic method to do the inference for a set of data which is not too big repetitively.

The online LDA method can be improved a lot even by adding a new label to all the unknown words. Our simulation proved this.

LDA using a Dirichlet process to fit the infinity vocabulary model is a very good approach to adapt to the real world situation. However, due to the limitation of time, our project did not find the most suitable evaluation metric to show that.

4.2 Future work

This project used a small data set, which may not show the advantage of the online LDA method. The future work will be doing the simulation with a bigger data set and also choosing several different method to do some comparison.

Online LDA method with unknown words performs better in the simulation. However, we do not know whether this is a coincidence. In the future, we will do more simulation to test the result. And if the future simulation give us a positive respond. Then we will try to prove this in a theoretical way.

The final thing may need to be improved is the evaluation metric in our project. Perplexity has a limitation that only the models with same vocabulary can be compared. So, this is not tractable for the infinity vocabulary method. PMI is a good method if we can collect a large corpus set which may be computational costing. For topic based similarity, since some topic may have common words as top words which means it represents a noise when compare the similarity between topics, a possible way to solve the large range of distance is to use useful topics when calculating Hellinger distance instead of all topics we set, which is mentioned in [5].

REFERENCES

[1] David M Blei and John D Lafferty. 2006. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*. ACM, 113–120.
[2] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3, Jan (2003), 993–1022.
[3] Matthew Hoffman, Francis R Bach, and David M Blei. 2010. Online learning for latent dirichlet allocation. In *advances in neural information processing systems*. 856–864.
[4] Shangsong Liang, Emine Yilmaz, and Evangelos Kanoulas. 2016. Dynamic clustering of streaming short documents. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 995–1004.
[5] David Newman, Sarvnaz Karimi, and Lawrence Cavedon. 2009. External evaluation of topic models. In *in Australasian Doc. Comp. Symp., 2009*. Citeseer.
[6] Ke Zhai and Jordan Boyd-Graber. 2013. Online latent Dirichlet allocation with infinite vocabulary. In *International Conference on Machine Learning*. 561–569.