EXTENSIONS TO CROSS-COLLECTION TOPIC MODELS WITH PARALLEL INFERENCE AND DIFFERENTIAL PRIVACY USING FLEXIBLE PRIORS

Zhiwen Luo

A thesis
in
The Department
of
Concordia Institute for Information Systems Engineering

Presented in Partial Fulfillment of the Requirements
For the Degree of Master of Applied Science
(Information Systems Security)at
Concordia University
Montréal, Québec, Canada

July 2022

© Zhiwen Luo, 2022

Concordia University
School of Graduate Studies

This is to certify that the thesis prepared

By: **Zhiwen Luo**

Entitled: **Extensions to Cross-collection Topic Models with Parallel Inference and Differential Privacy using Flexible Priors**

and submitted in partial fulfillment of the requirements for the degree of

Master of Applied Science
(Information Systems Security)

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the final examining commitee:

Dr. Abdessamad Ben Hamza ________________ Chair
Dr. Abdessamad Ben Hamza ________________ Examiner
Dr. Walter Lucia ________________ Examiner
Dr. Nizar Bouguila ________________ Supervisor

Approved ________________________________
Dr. Mohammad Mannan, Graduate Program Director

____________ 2022 ______ ________________________
Dr. Mourad Debbabi, Dean
Faculty of Engineering and Computer Science

Abstract

Extensions to Cross-collection Topic Models with Parallel Inference and Differential Privacy using Flexible Priors

Zhiwen Luo

Cross-collection topic models extend previous single-collection topic models such as Latent Dirichlet Allocation (LDA) to multiple collections. The purpose of cross-collection topic modelling is to model document-topic representations and reveal similarities between topics and differences among groups. The limitation of Dirichlet prior has impeded the state-of-the-art cross-collection topic models' performance, leading to the introduction of more flexible priors.

In this thesis, we first introduce a novel topic model, GPU-based cross-collection latent generalized Dirichlet allocation (ccLGDA), exploring the similarities and differences across multiple data collections by introducing generalized Dirichlet (GD) distribution to overcome the limitations of Dirichlet prior for conventional topic models while improving computational efficiency. As a more flexible prior, the generalized Dirichlet distribution provides a more general covariance structure and valuable properties, such as capturing collection relationships between latent topics and enhancing the cross-collection topic model. Indeed, this new GD-based model utilizes the Graphics Processing Unit to perform a parallel inference on a single machine, which provides a scalable and efficient training method for massive data. Therefore, the new approach, the GPU-based ccLGDA, proposes a scheme that incorporates a thorough generative process into a robust inference process with powerful computational techniques to compare multiple data collections and find interpretable topics. Its performance in comparative text mining and document classification shows its merits.

Furthermore, the restriction of Dirichlet prior and the significant privacy risk have hampered cross-collection topic models' performance and utility. The training of those cross-collection topic models may, in particular, leak sensitive information from the training dataset. To address the two issues mentioned above, we propose another novel model, cross-collection latent Beta-Liouville allocation (ccLBLA), which operates a more powerful prior, Beta-Liouville distribution with a more general covariance structure that brings a better capability in topic correlation analysis with fewer parameters than GD distribution. To provide privacy protection for the ccLBLA model, we leverage the inherent differential privacy guarantee of the Collapsed Gibbs Sampling (CGS) inference scheme and then propose a centralized privacy-preserving algorithm for the ccLBLA model (HDP-ccLBLA) that prevents inferring data from intermediate statistics during the CGS training process without sacrificing its utility. More crucially, our technique is the first to use the cross-collection topic model in image classification applications and investigate the cross-collection topic model's capabilities. The experimental results for comparative text mining and image classification will show the merits of our proposed approach.

Acknowledgments

First of all, I would like to give my heartfelt thanks to my academic supervisor Dr. Nizar Bouguila, for his invaluable instruction and inspiration. Throughout these years, I learned a lot from his valuable tutoring. His immense knowledge and plentiful experience have encouraged me all the time in my academic research and daily life. I would like to thank him again for being such a wonderful advisor to me in so many aspects.

Many thanks to my colleagues in the lab and my friends for their support and motivation during my two years of study.

And finally, I would like to thank my family for their unconditional support throughout my studies. Their belief in me has kept my spirits and motivation high during this process.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

As social media platforms proliferate, our internet continues to collect an unprecedented quantity of information from large-scale applications, making it more important to extract knowledge and patterns from large and complex data sets. Therefore, researching efficient machine learning techniques to handle massive data collections such as text documents and images is absolutely essential. In unsupervised topic modeling, such data are generalized as documents manipulated using count vectors according to the Bag of Words (BOW) approach. The objective is to construct meaningful topics to efficiently predict unseen documents in information retrieval and document classification tasks. In further detail, topics represent the intermediate low-dimensional representations of documents [1]. A well-known topic model is Latent Dirichlet Allocation (LDA) [2] incorporating the Dirichlet distribution as conjugate prior to the multinomial distribution. In the LDA model, documents appear as a combination of topics, and topics are vocabulary distributions. Moreover, LDA is frequently used as a dimensionality reduction tool to examine documents by topic and extract useful information from a large amount of unstructured data. Recently, topic models such as LDA have been the subject of various extension techniques [3] to cluster text documents and images [4–6] through their latent topics based on words (or visual words in case of images) co-occurrence.

The origin of topic models is the latent semantic analysis (LSA) [7]. LSA model mainly utilizes Singular Value Decomposition (SVD) for data analysis, so it is not a topic model in effect. Nevertheless, its fundamental idea provides the foundation of topic models and has contributed to developing the first topic model. Based on the structure of the LSA model, Hofmann [8] proposed a probabilistic latent semantic indexing (PLSI) model. pLSI model is a probabilistic generative model by looking for a transformation from document space to topic (implicit semantic) space with different optimization goals. Therefore, the pLSI model is seen as an actual topic model. LDA model [2], as an extension of the pLSI model [8], is a complete generative probabilistic model that improves generalization capability by introducing Dirichlet prior to overcome the overfitting and the difficulty in predicting documents probability problems. In particular, the LDA model utilizes the BOW method for a variety of different applications, including text modeling and computer vision, and its generative process has been extensively documented in several articles [2,3,9]. Even though the LDA model plays a fundamental role in topic modeling and many machine learning applications, numerous studies [10, 11] have shown that the constraints of Dirichlet prior hamper the LDA's performance. Specifically, the Dirichlet has a restrictive negative covariance matrix, unable to capture the correlation between topics [12–14]. Bakhtiari and Bouguila [15] showed that using more flexible priors [16] such as Generalized Dirichlet (GD) and Beta-Liouville (BL) distributions [17] in document parameters can improve the performance of the LDA model in text modeling and computer vision applications. Moreover, Ihou and Bouguila [5, 6] proposed new models that replace the Dirichlet distribution on both the corpus and the document parameters with GD or BL priors [18], and their experiments show that those more flexible priors can perform well in topic correlated environments. The recent expansion of large-scale datasets has led to the proliferation of studies that use more efficient computational methods. Hence, Graphics Processing Units (GPUs), which have successfully accelerate various large-scale data machine learning algorithms, provides us a better platform for implementing parallel inference methods with massively built-in parallel thread processors and high-speed memory. So far, GPUs have become a high-performance parallel architecture for many applications [19]. Compared with the CPU, GPU has a much higher computational capability and memory bandwidth. However, some famous parallel implementations of LDA, such as AD-LDA [20], cannot be adapted on the GPU due to the limitation of memory consumption. Moreover, recent evidence suggests that model inversion attack [21] and membership inference attack [22], according to recent findings, can both pose a privacy issue for machine learning models in different ways. To address these privacy problems, Dwork et al. [23] proposed the differential privacy (DP) strategy for privacy preservation in machine learning models. Because differential privacy provides a mathematical framework for measuring the security of several machine learning techniques, there has been an increasing interest in applying differential privacy in topic models such as LDA.

1.2 Cross-collection Topic Model

So far, natural language processing, computer vision, pattern recognition and other disciplines are increasingly using the LDA model and its extensions, such as the LBLA. Due to different practical problems, there are more and more different new topic models inspired from LDA. For example, Zhai et al. [24] introduced a topic model, the Cross-Collection Mixture model (ccMix) based on the pLSI model [8], for handling comparative text mining problems. Due to the limitation of the ccMix model, Paul and Girju [25] presented a Cross-Cultural LDA (ccLDA) model, which is the extension of LDA and ccMix frameworks. The cross-collection topic models try to extract the common information from all collections and figure out what is unique to a specific collection in different dataset collections. As the state-of-the-art cross-collection topic model, the ccLDA model provides better generalization capabilities which is less reliant on user-defined parameters. Moreover, ccLDA model shares assumption with the LDA-Collection [26] and Topical N-Gram models [27]. Those models assume that each word can be generated from two different distributions. Based on ccLDA model, Julian and Ralf [28] offered an entropy-based ccLDA model which distinguishes collection-independent and collection-specific words according to information entropy. The BOW assumption is maintained in both ccLDA and entropy-based ccLDA models; thus, each word is dependent on different dataset collection.

The ccLDA model can both detect topics among multiple data collections and differences between those data collections. Specifically, the ccLDA model first samples a collection c (observable data), then chooses a topic z and flips a coin x to determine whether to draw from the shared topic-word distribution or the topic's collection-specific distribution. The probability of x is 1 or 0 and comes from a Beta distribution.

The generative process of the ccLDA model is based on the following steps:

- Draw a collection-independent multinomial word distribution ϕ_z from $Dirichlet(\beta)$ for each topic **z**
- Draw a collection-specific multinomial word distribution $\sigma_{z,c}$ from $Dirichlet(\delta)$ for each topic **z** and each collection **c**
- Draw a Bernoulli distribution $\psi_{z,c}$ from $Beta(\gamma_0, \gamma_1)$ for each topic **z** and each collection **c**
- For each document **d**, choose a collection **c** and draw a topic mixture θ_d from $Dirichlet(\alpha_c)$. Then for each word w_i in **d**:
  - Sample a topic z_i from $Mutl(\theta_d)$
  - Sample x_i from $Bernoulli(\psi_{z,c})$
  - If $x_i = 1$, sample a word w_i from $Mutl(\sigma_{z,c})$
    else $x_i = 0$, sample a word w_i from $Mutl(\phi_z)$

Although the ccLDA model generalizes the LDA model by adding comparative analyses of different data collections, the limitations of the Dirichlet distribution to capture the correlation between topics have impeded the performance of the ccLDA model its extensions in various text analysis or classification applications.

1.2.1 Inference schemes

Many previous inference techniques and extensions proposed to the latent Dirichlet allocation [2] considers some inference schemes, such as VB and MCMC inferences [2, 29–32]. Due to the restrictions of Dirichlet prior, those models cannot learn coherent topics and has the challenge to identify the optimal number of topics because the Dirichlet distribution has a restrictive negative correlation structure, which impedes the performance for exploring positively correlated structure datasets [33]. So, the authors in [11, 15] used a more flexible prior, generalized Dirichlet distribution to circumvent the limitation of Dirichlet prior. Those previous models have chosen variational Bayesian (VB) inference because it has a higher convergence speed than MCMC inference. However, the problem of VB is that the method suffers from a significant bias as it assumes that the latent variables and parameters are fully independent. Even though this strong assumption can bring computational advantages, it harms the overall performance of the accuracy. Therefore, the VB inference schemes could cause an inaccurate modeling result. Ihou and Bouguila proposed a CVB-LGDA model [5, 34], which takes advantage of both the VB and the MCMC as a hybrid inference scheme. Although this method is accurate, this hybrid inference scheme is too complex and inefficient for large-scale datasets because it requires second-order Taylor approximations to calculate the latent variable. Therefore, the Collapsed Gibbs sampling [29, 35] is still the first choice for many topic models because it is simple to implement and CGS can approximate a global maximum based on sampling from the actual posterior distribution rather than variational distribution in a variational Bayesian inference scheme. Nevertheless, the CGS is also inefficient for massive datasets with high computational complexity issue. This limitation mainly hinders the development and use of the topic model in real-world applications and industry, requiring highly effective performance for large-scale data. Therefore, a robust approach is to parallelize learning methods with multiple processors [20, 36].

Researchers have paid more attention to parallel algorithms for topic models inference. Newman et al. [20] propose two parallel LDA algorithms, AD-LDA and HD-LDA. The AD-LDA made an eight times speed up on a 16-processor computer. Chen et al. [37] extend the AD-LDA model with MPI and MapReduce on 32 machines and get ten times speedup. Asuncion et al. [38] provide an asynchronous distributed LDA algorithm, which makes 15-25 times speedup on 32 processors. However, those parallel algorithms require many machines. Nowadays, GPU performance has improved significantly as compared to CPU, and NVIDIA CUDA programming interface has become a powerful tool to extend topic modeling scheme, so many recent studies have shown that GPUs are a better choice for implementing the parallel algorithms for LDA model inference [19]. For instant, Yan et al. [39] have accelerated collapsed Gibbs sampling (CGS) method of LDA on the GPU. Compared with the standard LDA on the CPU, their implementation achieved a speedup of around 26 times on a single machine. Lu et al. [19] present a GLDA model, which uses GPU to accelerate the CGS-LDA training by highly reducing the memory requirement on a single GPU. Their method also can be extended to train large-scale data by involving multiple GPUs. Nevertheless, these GPU-based topic models have tended to focus on single collection rather than multiple collections data. Far too little attention has been paid to applying the cross-collection topic model on the GPU platform.

1.2.2 Topic model with Differential Privacy

Many machine learning models [40–42] have applied differential privacy to address privacy attack vulnerabilities by perturbing the model during different training parts. Specifically, there are a lot of different ways to adopt differential privacy in ML models such as output perturbation, objective perturbation [43], intermediate perturbation [44, 45] and input perturbation. In recent years, there has been an increasing interest in input perturbation and local differential privacy [46], which demonstrates that enormous randomized crowdsourced data may leak valuable statistics. By eliminating the premise of trustworthy servers, the input perturbation can give a privacy guarantee. As a classic machine learning approach, topic models also can achieve differential privacy protection by perturbing the intermediate parameters during the training process via input perturbation. For instance, by perturbing the sampling distribution in the final iteration, Zhu et al. [47] suggested a DP guarantee CGS-LDA model. While performing variational Bayesian inference scheme, Park et al. [45] used differential privacy in LDA by perturbing the adequate statistics data in each iteration. Similar to the above works, Decarolis et al. [48] altered the intermediate statistics in the spectral methodology. However, those DP guarantee methods [45, 47, 48] cannot tackle the problem of untrustworthy data curators by design. Wang et al. [49] established a locally private LDA strategy for a federated environment, but this approach is not a generic solution to standard approach for the batch-based LDA model.

Then, Zhao et al. [50] proposed a differential privacy solution for traditional batch LDA training, a hybrid privacy-preserving algorithm (HDP-LDA), which injects the noise to obfuscate the word count in each training iteration and takes advantage of the inherent randomness of Markov Chain Monte Carlo (MCMC) techniques. The inherent privacy guarantee is an essential feature of the CGS-LDA method. Recent improvements [49, 51] in intrinsic privacy have heightened that the Bayesian sampling can generate the inherent privacy guarantee without introducing further noise to sample statistics variables. Foulds et al. [52] expanded on this work, concluding that the generic MCMC mechanism may also process inherent privacy guarantees and acquire privacy protection in a way that is similar to the Laplace mechanism. Measuring the inherent privacy guarantee in a topic model such as the LDA model is still a challenge. Even though HDP-LDA [50] has been demonstrated to be effective and outperforms some methods mentioned above [45,47,48], this scheme still suffers from the restriction of Dirichlet prior and insufficient for comparative datasets analysis. In this thesis, we present a cross-collection topic model that overcomes the limitations of Dirichlet prior by adopting a more flexible prior as well as using differential privacy for privacy preservation, which can secure sensitive information from attackers who are aware of the training process.

1.3 Contributions

The main contributions of this thesis could be summarized as follows:

1. **Parallel Inference for Cross-Collection Latent Generalized Dirichlet Allocation Model and Applications**
   We present a GPU-based cross-collection latent topic model with more flexibility and scalability by providing a better prior distribution and using a parallel inference which is parallel collapsed Gibbs sampling (CGS) for handling large datasets. Our model replaces Dirichlet distribution with GD as a more flexible prior to overcome its shortcomings related to both the document and corpus parameters. It also provides an improvement to the state-of-the-art cross-collection model, CGS-ccLDA [25]. We introduce a parallel collapsed Gibbs sampling (CGS) approach for the ccLGDA model on GPUs. Our parallel approach exploits the parallel computing power of GPUs and utilizes the CGS structure of the ccLGDA learning approach, significantly reducing the computing cost and processing time. Finally, our new model is successfully applied for comparative text mining and document classification.

2. **Cross-Collection Latent Beta-Liouville Allocation Model Training with Privacy Protection and Applications**
   The generative process of LDA [2], LBLA [6,15,53], and the ccLDA [25] have all been improved by the new model. Our novel model replaces Dirichlet distribution with Beta-Liouville (BL) distribution [54,55] as more flexible prior [56] to

overcome its shortcomings related to document and corpus parameters. Compared with the state-of-the-art privacy-preserving topic model (HDP-LDA), our proposed model can discover topics' similarities and differences across multiple collections. Indeed, we deliver the first study on adopting the cross-collection topic model for image classification applications by processing each image as a separate document using the Bag of Visual Words methodology [4–6]. Our studies indicate that our proposed model (ccLBLA) can achieve a much higher generalization performance in comparative text mining, and document and image classification. Furthermore, the HDP-ccLBLA strategy can obtain a good model utility while maintaining sufficient privacy guarantees.

1.4 Thesis Overview

This thesis is structured as follows:

- Chapter 1 introduced the background knowledge regarding cross-collection topic model with different inference schemes and differential privacy.

- Chapter 2 presents a GPU-based cross-collection latent topic model with more flexibility and scalability by providing a better prior distribution and using a parallel inference which is parallel collapsed Gibbs sampling (CGS) for handling large datasets. The new approach introduces a flexible GD prior for a robust parallel inference scheme taking advantage of GPUs to show its merit in comparative text mining. Experimental results illustrate that our proposed model, GPU-based ccLGDA, outperforms ccLDA on all four quality measures on four text datasets with different domains and quantity of collections and proves the proposed method's robustness on various text datasets in other fields.

- In chapter 3, we develop a novel cross-collection topic model (ccLBLA model) that utilizes the BL distribution instead of Dirichlet for various domain text collections to improve previous cross-collection topic models. We present the first study on applying the cross-collection topic model to image classification application. What's more, our proposed model (HDP-ccLBLA) can prevent data inference from intermediate statistics during training. Indeed, our experimental studies demonstrate that the HDP-ccLBLA algorithm can achieve a good model utility under differential privacy.

- Chapter 4 demonstrates a conclusion of the thesis by summarizing the main contributions and some promising future work.

Chapter 2

Parallel Inference for Cross-Collection Latent Generalized Dirichlet Allocation Model and Applications

In this chapter, we propose a GPU-based cross-collection latent topic model with more flexibility and scalability by providing a better prior distribution and using a parallel inference which is parallel collapsed Gibbs sampling (CGS) for handling large datasets. This is a novel cross-collection topic model that combines state-of-the-art cross-collection topic model [25], the completely LGDA model [34,35] and the GLDA model [19]. Besides, This parallel inference scheme integrates the advantages of GPUs computing and Gibbs sampling with GD distributions in collapsed space [19,35]. This robust parallel inference scheme allows the ccLGDA model to analyze latent topics and discover the similarities and differences across a considerable number of collections and datasets with high computational efficiency.

2.1 The GPU-based ccLGDA Model

This section mainly describes our GPU-based Cross-Collection Latent Generalized Dirichlet Allocation (GPU-based ccLGDA) Model. Our approach integrates GLDA [19] and ccLDA [25] as a GPU-based cross-collection topic model with considering

GD distribution on both document and corpus parameters. We start with a review of the generative process of fundamental LGDA [5,11,35] and ccLDA [25] models. Then, we introduce our extension (GPU-based ccLGDA) of those two models to the parallel collapsed Gibbs sampling learning scheme applying the method on GLDA model [19], including parallel method and CGS inference schemes with GD distribution prior. Therefore, this paper will first compare CPU-based ccLGDA models (CGS-ccLGDA) to illustrate our proposed model's more comprehensive analysis and motivations by showing its merit in large-scale processing data. For helping readers to get a better understanding of our model, the variables are described in Table 2.1, and we will provide their characteristics.

C - Total number of collections

D - Total number of documents

W - Total number of words in each document

K - Total number of topics

$\mathbf{w} = w_{ij}$ - observed words

$\mathbf{z} = z_{ij}$ - latent variables

θ_j - mixing proportions

ϕ_k - corpus parameters in collection-independent distribution

$\sigma_{k,c}$ - corpus parameters in collection-specific distribution

$\psi_{k,c}$ - parameter in Bernoulli distribution

$\theta_j \sim GD(u_c, v_c)$ - generalized Dirichlet distribution

$\phi_k \sim GD(s, t)$ - generalized Dirichlet distribution

$\sigma_{k,c} \sim GD(g_c, h_c)$ - generalized Dirichlet distribution

$\psi_{k,c} \sim Beta(\gamma_0, \gamma_1)$ - Beta distribution

$x \sim Bernoulli(\psi_{ck})$ - Bernoulli distribution

$z_{jk}/\theta_{jk} \sim Mult(\theta_j)$ - multinomial distribution

$x_{jk}/z_{jk}, \phi_k, x = 0 \sim Mult(\phi_k)$ - multinomial distribution

$x_{jk}/z_{jk}, \sigma_{k,c}, x = 1 \sim Mult(\sigma_{ck})$ - multinomial distribution

Table 2.1: Model variables and definitions

2.1.1 LGDA and ccLDA models

Dirichlet distribution cannot perform well in a topic correlation analysis because of its negative covariance structure. Even though Blei et al. [57] proposed a Correlated Topic Models (CTM) to solve that problem in the topic model by introducing the normal logistic distribution. However, this distribution is not a conjugate prior to the multinomial distribution [57, 58], which makes the CTM difficult to implement.

Recent developments in the topic modeling, have focused on the need for more flexible priors. The generalized Dirichlet prior has become a popular choice. There are many extensions of the LDA model based on generalized Dirichlet prior, such as GD-LDA [10], LGDA [11], and Collapsed LGDA [5,35] models. For generalized Dirichlet distribution, in dimension $(K+1)$ space, the generalized Dirichlet distribution with hyperparameter vector $(s_1, t_1, ..., s_K, t_K)$ is defined by:

$$p(\phi_k|\mathrm{s}, \mathrm{t}) = \prod_{k=1}^{K} \frac{\Gamma(s_k + t_k)}{\Gamma(s_k)\Gamma(t_k)} \phi_k^{s_k - 1} \left(1 - \sum_{j=1}^{k} \phi_j\right)^{\gamma_k} \tag{2.1}$$

for $k = 1, \ldots, K-1,$, where $\gamma_k = t_k - s_{k+1} - t_{k+1}$, and $\gamma_K = t_K - 1$. The vector ϕ_k is the N-dimensional multinomial parameter drawn from the GD(s,t) distribution. When $t_k = s_{k+1} + t_{k+1}$, the generalized Dirichlet distribution is reduced to Dirchlet distribution [13, 59–62]. Thus, the generalized Dirichlet includes the Dirichlet distribution as a special case [63]. Compared with Dirichlet distribution, the generalized Dirichlet has more parameters and is more flexible for servel applications [12,54,64,65]. We define $\phi = (\phi_1, ..., \phi_{K+1})$ and $\phi_{K+1} = 1 - \sum_{i=1}^{k} \phi_i$. The mean and the variance of the generalized Dirichlet distribution are given by:

$$E[\phi_k] = \frac{s_k}{s_k + t_k} \prod_{i=1}^{k-1} \frac{t_i}{s_i + t_i} \tag{2.2}$$

$$Var[\phi_k] = E(\phi_k)(\frac{s_k + 1}{s_k + t_k + 1} \prod_{i=1}^{k-1} \frac{t_i + 1}{s_i + t_i + 1} - E(\phi_k)) \tag{2.3}$$

and, the covariance between ϕ_i and ϕ_j is given by

$$Cov[\phi_i, \phi_j] = E(\phi_j)(\frac{s_i + 1}{s_i + t_i + 1} \prod_{d=1}^{i-1} \frac{t_d + 1}{s_d + t_d + 1} - E(\phi_i)) \tag{2.4}$$

According to Eq.2.2 - 2.4, variables with the same mean do not need to have the same variance. Moreover, unlike the restrictive negative covariance of Dirichlet distribution [66–68], the generalized Dirichlet distribution has a more general covariance structure. Those advantages mentioned above make the generalized Dirichlet distribution more powerful and practical in topic modeling. Furthermore, both generalized Dirichlet and Dirichlet distributions are conjugate to the multinomial [69–72] and belong to the exponential family of distributions. Hence, introducing GD distribution to replace the Dirichlet prior in the LDA model provides a considerable improvement in topic correlation and is convenient practical applications. Consequently, the LGDA model can provide more practical capabilities than the original LDA model, and includes the LDA model as a particular case [11].

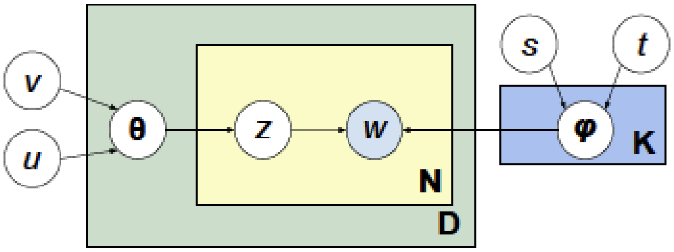


Figure 2.1: Graphical representation of LGDA

This subsection will analyze the (smoothed) LGDA model, which implements the GD distribution on both document and corpus parameters [35]. The LGDA model is a generative probabilistic model. The model generates each word of the document through the following steps:

- For each document **d**, draw a topic mixture θ_d from $GD(u, v)$.
- Draw a corpus multinomial word distribution ϕ_k from $GD(s, t)$ for each topic **z**.
- Then for each word w_i in **d**:
  - Choose a topic z_i from $Mult(\theta_d)$
  - Choose a word w_i from $Mult(\phi_k)$

Even though the LGDA model has more flexible prior to enhance the topic correlation, this model only focuses on one single collection of the dataset, which is insufficient for comparative datasets analysis. To overcome this problem in topic model scheme, Paul and Girju [25] propose the cross-collection latent Dirichlet allocation (ccLDA) model based on the ccMix [24] and LDA [2] models. The ccLDA model, can both detect topics among multiple data collections as well as differences between those data collections. Specifically, the ccLDA model first samples a collection c (observable data), then chooses a topic z and flips a coin x to determine whether to

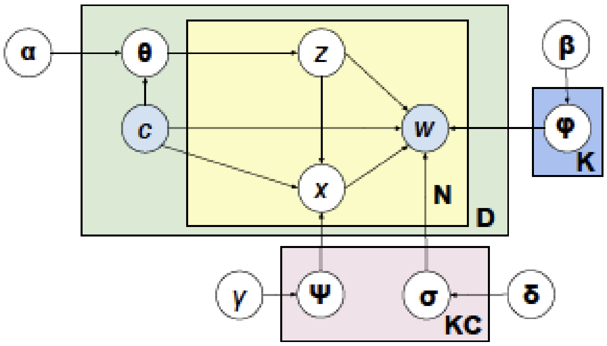


Figure 2.2: Graphical representation of ccLDA

draw from the shared topic-word distribution or from the topic's collection-specific distribution. The probability of x is 1 or 0 and comes from a Beta distribution. The generative process of the ccLDA model is based on the following steps:

- Draw a collection-independent multinomial word distribution ϕ_z from $Dirichlet(\beta)$ for each topic **z**
- Draw a collection-specific multinomial word distribution $\sigma_{z,c}$ from $Dirichlet(\delta)$ for each topic **z** and each collection **c**
- Draw a Bernoulli distribution $\psi_{z,c}$ from $Beta(\gamma_0, \gamma_1)$ for each topic **z** and each collection **c**
- For each document **d**, choose a collection **c** and draw a topic mixture θ_d from $Dirichlet(\alpha_c)$. Then for each word w_i in **d**:
  - Sample a topic z_i from $Mutl(\theta_d)$
  - Sample x_i from $Bernoulli(\psi_{z,c})$
  - If $x_i = 1$, sample a word w_i from $Mutl(\sigma_{z,c})$
    else $x_i = 0$, sample a word w_i from $Mutl(\phi_z)$

Although the ccLDA model generalizes the LDA model through adding comparative analyses of different data collections, it still suffers from an incomplete generative process due to the restricted covariance structure of Dirichlet prior. Moreover, both LGDA and ccLDA models use inefficient inference techniques to estimate the posterior of the hidden variables. For example, GD-LDA [10], LGDA [5,35], ccLDA [25] models are CPU-based. The performance of those models is still not satisfactory since they require high memory storage and extensive computational resources. Those models are inadequate for modern applications, demanding fast computation of huge datasets. To deal with the problems of LGDA and ccLDA models with CPU-based inference schemes, we will propose our GPU-based ccLGDA model in the next subsection.

2.1.2 Proposed topic model : GPU-based ccLGDA model

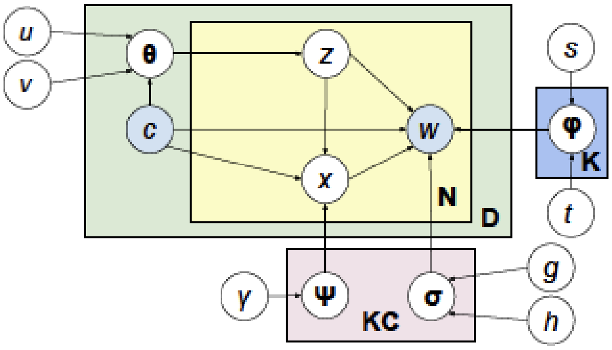


Figure 2.3: Graphical representation of ccLGDA

In this subsection, we demonstrate our extension of ccLDA [25], LGDA [5,11,34, 35], and GLDA [19] models by using a parallel inference method, in which we take advantage of collapsed Gibbs sampling (CGS) and the GPUs with GD distribution on both the document and corpus parameters in the collapsed space. For the complete analysis of the GPU-based ccLGDA model, we will first state the generative process of the ccLGDA model, and then we will analyze and compare the CPU-based ccLGDA models (CGS-ccLGDA). Finally, we will illustrate the parallel inference scheme for the ccLGDA model on single machine and show its merit.

ccLGDA model, first samples a collection **c** (observable data), then chooses a topic **z** and flips a coin **x** to determine whether to draw from the shared topic-word distribution or the topic's collection-specific distribution. The probability of **x** being 1 or 0 comes from a Beta distribution.

- Draw a collection-independent multinomial word distribution ϕ_k from $GD(s,t)$ for each topic **z**

- Draw a collection-specific multinomial word distribution $\sigma_{k,c}$ from $GD(g_c, h_c)$ for each topic **z** and each collection **c**
- Draw a Bernoulli distribution $\psi_{k,c}$ from $Beta(\gamma_0, \gamma_1)$ for each topic **z** and each collection **c**
- For each document **d**, choose a collection **c** and draw a topic mixture θ_d from $GD(u_c, v_c)$. Then for each word w_i in **d**:
  - Sample a topic z_i from $Mutl(\theta_d)$
  - Sample x_i from $Bernoulli(\psi_{k,c})$
  - If $x_i = 1$, sample a word w_i from $Mutl(\sigma_{k,c})$
    else $x_i = 0$, sample a word w_i from $Mutl(\phi_k)$

Collapsed Gibbs Sampling (MCMC-CGS) for ccLGDA

Because the estimation of the posterior distribution in the Bayesian topic model is intractable, inference methods such as VB and MCMC become the common choices to estimate the latent topics and the model parameters. For the inference of the ccLGDA model, we choose collapsed space representation because it contributes to the performance of GPU-batch models [19]. In collapsed space, the parameters are marginalized out only, leaving the latent variables that are conditionally independent [73], and the collapsed space of latent variables is a low dimensional space compared with joint space. Based on those properties of collapsed space, the computation of performing estimation is faster than in joint space. The collapsed Gibbs sampling inference algorithm computes expectations by a sampling process of the latent variables to approximate the posterior distributions using a Bayesian network. Compared with standard Gibbs sampling in the joint space, the CGS is simple to implement and computationally faster. Because the CGS inference dose not need to use digamma functions, it reduces computational consumption. Hence, the CGS inference provides an accurate estimate of the actual posterior distribution when the Markov chain reaches its stationary distribution. The ccLDA and CGS-LGDA models [25, 35] are based on CGS inference to estimate posterior distribution because of its advantages.

In the CGS-ccLGDA scheme, the conditional probabilities of latent variable z_{ij} are calculated by the current state of all variables except the particular variable

z_{ij} beging processed in the marginal joint distribution $p(\mathbf{w}, \mathbf{z}|x_{ij} = 0, u_c, v_c, s, t)$ or $p(\mathbf{w}, \mathbf{z}|x_{ij} = 1, u_c, v_c, g_c, h_c)$ between collection-common and collection-specific case. This algorithm applies the collapsed Gibbs sampler for topic assignments. The conditional probability of z_{ij} is $p(z_{ij} = k|x_i = 0, \mathbf{z}_{-ij}, \mathbf{w}, u_c, v_c, s, t)$ or $p(z_{ij} = k|x_i = 1, \mathbf{z}_{-ij}, \mathbf{w}, u_c, v_c, g_c, h_c)$. The $-ij$ represents the counts with z_{ij} excluded [73]. This conditional probability of collection-common and collection-specific are expressed as:

$$p(z_{ij} = k|x_{ij} = 0, \mathbf{z}^{-ij}, \mathbf{w}, u_c, v_c, s, t) = \frac{p(z_{ij} = k, z^{-ij}, \mathbf{w}|x_{ij} = 0, u_c, v_c, s, t)}{p(z^{-ij}, \mathbf{w}|x_{ij} = 0, u_c, v_c, s, t)} \tag{2.5}$$

$$p(z_{ij} = k|x_{ij} = 1, \mathbf{z}^{-ij}, \mathbf{w}, u_c, v_c, g_c, h_c) = \frac{p(z_{ij} = k, z^{-ij}, \mathbf{w}|x_{ij} = 1, u_c, v_c, g_c, h_c)}{p(z^{-ij}, \mathbf{w}|x_{ij} = 1, u_c, v_c, g_c, h_c)} \tag{2.6}$$

Eq.2.5 and Eq.2.6 can be simplified as following:

$$p(z_{ij} = k \mid x_{ij} = 0, \mathbf{z}^{-ij}, \mathbf{w}, u_c, v_c, s, t) \propto p(z_{ij} = k, z^{-ij}, \mathbf{w} \mid x_{ij} = 0, u_c, v_c, s, t) \tag{2.7}$$

$$p(z_{ij} = k \mid x_{ij} = 1, \mathbf{z}^{-ij}, \mathbf{w}, u_c, v_c, g_c, h_c) \propto p(z_{ij} = k, z^{-ij}, \mathbf{w} \mid x_{ij} = 1, u_c, v_c, g_c, h_c) \tag{2.8}$$

In CGS-ccLGDA model, the parameters θ, ϕ, σ are drawn from the GD distribution. To speed up the traning process, we marginalize these parameter in the collapsed space because sampling in the collapsed space is much faster than in the joint space of latent variables and parameters [5, 73]. In the collapsed space, we can integrate out θ, ϕ, σ, and ψ to get Eq.2.11 - 2.14 based on the conjugacy of the Beta/binomial and GD/multinomial distributions using the update equations from CGS-ccLDA and CGS-LGDA [25,35]. By integrating out the parameters, the Gibbs sampler's equation is obtained as an expectation expression:

$$\begin{aligned} &p(z_{ij} = k|x_{ij} = 0, \mathbf{z}^{-ij}, \mathbf{w}, u_c, v_c, s, t) = \\ &\qquad E_{p(z_{ij}=k|x_{ij}=0,\mathbf{w},u_c,v_c,s,t)}[p(z_{ij} = k|x_{ij} = 0, z^{-ij}, \mathbf{w}, u_c, v_c, s, t)] \end{aligned} \tag{2.9}$$

$$p(z_{ij} = k \mid x_{ij} = 1, \mathbf{z}^{-ij}, \mathbf{w}, u_c, v_c, g_c, h_c) = \\ E_{p(z_{ij}=k|x_{ij}=1,\mathbf{w},u_c,v_c,g_c,h_c)}[p(z_{ij} = k \mid x_{ij} = 1, z^{-ij}, \mathbf{w}, u_c, v_c, g_c, h_c)] \tag{2.10}$$

In CGS algorithm iteration, we sample new assignment of $\mathbf{z}$ and $\mathbf{x}$ alternately with the following equations:

$$p(z_i = k|x_i = 0, \mathbf{z}_{-i}, \mathbf{w}, u_c, v_c, s, t) \propto \\ \frac{(N_{jk}^{-ij} + u_{ck})(u_{ck} + \sum_{l=k+1}^{K+1} N_{jl}^{-ij})}{(u_{ck} + v_{ck} + \sum_{l=k}^{K+1} N_{jl}^{-ij})} \times \frac{(N_{kw_{ij}}^{-ij} + s_w)(t_w + \sum_{v=w+1}^{W+1} N_{kv_{ij}}^{-ij})}{(s_w + t_w + \sum_{v=w}^{W+1} N_{kv_{ij}}^{-ij})} \tag{2.11}$$

$$p(x_i = 0|x_{-i}, \mathbf{z}, \mathbf{w}, \gamma, s, t) \propto \\ \frac{N_{x=0}^{k,c} + \gamma_0}{N_{\cdot}^{k,c} + \gamma_0 + \gamma_1} \frac{(N_{kw_{ij}}^{-ij} + s_w)(t_w + \sum_{v=w+1}^{W+1} N_{kv_{ij}}^{-ij})}{(s_w + t_w + \sum_{v=w}^{W+1} N_{kv_{ij}}^{-ij})} \tag{2.12}$$

For Eq.2.11 and Eq.2.12, all counts only refer to the words for which $x_i = 0$, which are the words assigned to the topic model. Specifically, N is the total number of words for which $x_i = 0$, not the total number of words in the corpus. Same for Eq.2.13 and Eq.2.14, the count only include the words for which $x_i = 1$, which means that N is the total number of words for which $x_i = 1$.

$$p(z_i = k|x_i = 1, \mathbf{z}_{-i}, \mathbf{w}, u_c, v_c, g_c, h_c) \propto \\ \frac{(N_{jk}^{-ij} + u_{ck})(u_{ck} + \sum_{l=k+1}^{K+1} N_{jl}^{-ij})}{(u_{ck} + v_{ck} + \sum_{l=k}^{K+1} N_{jl}^{-ij})} \times \frac{(N_{ckw_{ij}}^{-ij} + g_{cw})(h_{cw} + \sum_{v=w+1}^{W+1} N_{ckv_{ij}}^{-ij})}{(g_{cw} + h_{cw} + \sum_{v=w}^{W+1} N_{ckv_{ij}}^{-ij})} \tag{2.13}$$

$$p(x_i = 1|x_{-i}, \mathbf{z}, \mathbf{w}, \gamma, g_c, h_c) \propto \\ \frac{N_{x=1}^{k,c} + \gamma_1}{N_{\cdot}^{k,c} + \gamma_0 + \gamma_1} \frac{(N_{ckw_{ij}}^{-ij} + g_{cw})(h_{cw} + \sum_{v=w+1}^{W+1} N_{ckv_{ij}}^{-ij})}{(g_{cw} + h_{cw} + \sum_{v=w}^{W+1} N_{ckv_{ij}}^{-ij})} \tag{2.14}$$

The count N_{jk}^{ij} is the number of word w_i in the document j and topic k in class c. Besides, N_{jk}^{-ij} is the total number of words in document j and topic k in class c except the word w_i being sampled. The count $N_{kw_{ij}}^{ij}$ is the number of times the word w_{ij} appears in topic k and document j. In addition, $N_{kw_{ij}}^{-ij}$ is the number of times the word w_{ij} appears in document j and topic k except beging sampled. $N^{ij}_{ckw_{ij}}$ is the number of times the word w_{ij} appears in topic k and document j in specific collection $\mathbf{c}$. In addition, $N^{-ij}_{ckw_{ij}}$ is the number of times the word w_{ij} appears in document j and topic k in specific collection $\mathbf{c}$ except beging sampled. $N^{k,c}_{x}$ is the number of $\mathbf{x}$ in topic k and collection c. $\mathbf{x}$ should be initialized as 0 for all tokens, because we initially assume that everything comes from the shared collection word distribution.

For parameters estimation, the document parameter distribution is:

$$\theta_{jk} = \frac{(N_{jk} + u_{ck})(u_{ck} + \sum_{l=k+1}^{K+1} N_{jl})}{(u_{ck} + v_{ck} + \sum_{l=k}^{K+1} N_{jl})} \tag{2.15}$$

The predictive distributions of the collection-independent and collection-specific words are:

$$\phi_{kw} = \frac{(N_{kw} + s_w)(t_w + \sum_{v=w+1}^{W+1} N_{kv})}{(s_w + t_w + \sum_{v=w}^{W+1} N_{kv})} \tag{2.16}$$

$$\sigma_{ckw} = \frac{(N_{ckw_{ij}} + g_{cw})(h_{cw} + \sum_{v=w+1}^{W+1} N_{ckv})}{(g_{cw} + h_{cw} + \sum_{v=w}^{W+1} N_{ckv})} \tag{2.17}$$

Algorithm 1 Summary of CPU-based ccLGDA Inference

```
procedure
Input: $\mathbf{w}$, $u_c, v_c, s, t, g_c, h_c$ iterMax, $K$, $V$, $N$
initialize $\mathbf{z}$, $\mathbf{x}$, $N_{jk}$, $N_{kw}$, $N_{ckw}$, $N_x$
for iter = 1 to iterMax do
    for $i = 1$ to $N$ in document $j$ in class $c$ do
        if $x_{ij} = 0$ then
            update $z_{ij}$ using Eq.2.11
        else[$x_{ij} = 1$]
            update $z_{ij}$ using Eq.2.13
        end if
        update $x_{ij}$ using Eq.2.12 and Eq.2.14
        update $N_{jk}$, $N_{kw}$, $N_{ckw}$, $N_x$
    end for
end for
Output: Parameters $\theta_j, \phi_k, \sigma_{ck}$ using Eq. 2.15 - 2.17
end procedure
```

GPU-based ccLGDA model

For showing the complete merit of ccLGDA model, we propose a method to overcome the efficiency problem of training the ccLGDA model - accleration with GPUs. Because the collapsed Gibbs sampling method is inherently sequential, each iteration depends on the previous result in the training process [19]. The study in [20] has shown that a similar accuracy could be obtained by using a parallel approximate topic model algorithm. Therefore, we integrate a related parallel topic model inference algorithm for our GPU implementation with more completely generative process. Specifically, we employ the enormous thread parallelism programming model from NVIDIA CUDA to implement our parallel algorithm.

Our parallel Gibbs sampling algorithm mainly utilizes atomic increment and decrement opreations to produce a correct result in the concurrent runing environment. Hence, we only need to maintain one copy of N_{wk} and N_{cwk} matrices in our implementation. First, we take advantage of the atomic increment and decrement operations for the correctly counter update. Then, we serialize the computation and update on the N_{wk} and N_{cwk} matrices. Algorithm 2 demonstrates our parallel ccLGDA algorithm for one interation with two modifications above. Compared with CPU-based ccLGDA model (from Algorithm 1), which requires many sequential loops to perform the result in each iteration. Our parallel algorithm makes use of the high-performance parallel architecture on GPUs to perform a concurrent running for w^p. In particular, our parallel algorithm execute global updates after each p words are sampled in p processors in parallel. This step can guarantee that updated results are correct. The inherent data parallelism of the sampling algorithm is implemented by multiple threads in each thread block because we map each thread block in CUDA to a procssor. And, the communication overhead is trivial in this implementation based on the fact that multi-processors on the GPU are tightly coupled.

2.2 Experiments

This section evaluates the cross-collection latent Generalized Dirichlet Allocation model with GPU implementation in terms of perplexity, classification accuracy, topic coherence, time efficiency, and topics examples through different datasets as compared with the ccLDA model to show our approach's merit. The experiments utilize four

Algorithm 2 Summary of GPU-based ccLGDA Inference for one iteration

```
procedure
Input: $\mathbf{w}$, $u_c, v_c, s, t, g_c, h_c$ iterMax, $K$, $V$, $N$
initialize $\mathbf{z}$, $\mathbf{x}$, $N_{jk}$, $N_{kw}$, $N_{ckw}$, $N_x$
$w^p$: word tokens assigned to the $p$th processor
for all processors in parallel do
    for each $w_{ij} \in w^p$ do
        if $x_{ij} = 0$ then
            sample $z_{ij}$ using Eq.2.11
        else[$x_{ij} = 1$]
            sample $z_{ij}$ using Eq.2.13
        end if
        sample $x_{ij}$ using Eq.2.12 and Eq.2.14
        /* Global synchronization */
        update $N_{jk}$
        Atomic update $N_{kw}$, $N_{ckw}$, $N_x$
    end for
end for
Output: Parameters $\theta_j, \phi_k, \sigma_{ck}$ using Eq. 2.15 - 2.17
end procedure
```

text datasets with different collections number, document lengths, and domains.

2.2.1 The Datasets

In the COVID-19 newspapers dataset, the first collection contains the online newspapers from the United States of America, which is from COVID-NEWS-US-NNKDATASET[1] [74]. In addition, we crawled the COVID-19 newspapers from several different British newspaper websites to model the second collection in this dataset based on the newspaper links[2]. Then, we will use this dataset for comparative content aggregation and summarization to extract common and different effects and knowledge about the virus in two various countries and demonstrate the merit of our proposed model. To display the superiority of our approach in different types of documents, the second dataset focuses on the field of computer science, including the

[1] https://github.com/nnk-dataset/usa-nnk

[2] https://www.kaggle.com/jwallib/coronavirus-newspaper-classification/data

abstracts of NeurIPS[3] and CVPR[4] papers in 2019. We take advantage of comparative text analysis to automatically discover different themes and trends in these two different conferences.

We use also a subset of the New York Times (NYT) comments dataset[5], which contains more than two million comments from 2017 to 2018. A two-month comments dataset between 2017 and 2018 is used to evaluate the accuracy of the ccLGDA model. Indeed, the larger dataset, including total month comments in 2017 and 2018, is used to assess the time efficiency of the topic model. Because the length of comments in this dataset varies greatly, some comments are discarded if the minimum number of words is less than one hundred. To make a fair comparison with the result reported in ccLDA [25], we reuse the dataset[6] crawled from an online travel platform - lonelyplanet.com. This dataset consists of three different countries' discussion forums of India, Singapore, and the UK. Each collection has thousands of threads. Therefore, our experiment utilized four domains datasets among newspapers, academic papers, customer comments, and blogs. Table 2.2 displays an overview of the datasets sizes.

Experiments Datasets			
Dataset	Collection	D	W/D
COVID-19 Newspapers	USA/UK	2731	433
Academic Papers	NIPS/CVPR	2787	91
NYT Comments	2017/2018	74895	127
Traveler Forum	India/Singapore/UK	4174	247

Table 2.2: Datasets - number of documents D and average number of words per document W/D (without stop word)

2.2.2 Experimental Setup

We preprocess the datasets by first tokenizing words with the Natural Language ToolKit(NLTK) [75], removing punctuation, stop-words, and then lemmatizing tokens to derive their common base form. Following the same setting of the asymmetric GD priors with Ihou and Bouguila [35], we implement GD priors hyperparameters as follows: $u_c = \{\frac{i}{K+1}\}_{i,\ldots,K}$; $v_c = \{\frac{i-1}{K+1}\}_{i,\ldots,K}$; s and $g_c = \{\frac{v}{V+1}\}_{i,\ldots,V}$; t and $h_c =$

[3] https://www.kaggle.com/rowhitswami/nips-papers-1987-2019-updated

[4] https://www.kaggle.com/paultimothymooney/cvpr-2019-papers

[5] https://www.kaggle.com/aashita/nyt-comments

[6] http://www.michaeljpaul.com/downloads/ccdata.php

$\{\frac{v-1}{V+1}\}_{i,\ldots,V}$ with same probability of occurrence of collection-common and collection-specific words. For Dirichlet based model, the topic distribution priors are fixed and $\alpha = 0.1$. Then, we set β and δ to 0.01; For γ_0 and γ_1, we use the same value, 1.0, for both Dirichlet and GD based models in the experiments. We experimented on an NVIDIA GeForce GTX 3070 GPU. Our proposed model, ccLGDA-GPU, is developed using NVIDIA CUDA. In our experiment, two GPU-based models, ccLGDA-GPU and GLDA (GPU-based LDA), use the same thread block setting, which is 1024 thread blocks. The CPU counterparts (CPU LDA and ccLDA[7]) are based on a widely used open-source package GibbsLDA++. In the experiment, we will compare the CPU-based ccLGDA model and GPU-based ccLGDA model to evaluate the accuracy of our parallel implementation.

For the experiment validation, we use ten-fold cross-validation, which separates each dataset with a 90% training set and 10% test set. In the Gibbs sampling, the burn-in period is five hundred, and then we collect ten samples separated by lags of ten iterations. The average of ten samples is the final result of the model. After, we calculate the document-topic parameter θ, the collection-independent word distribution parameter ϕ, the collection-specific word distribution parameter σ, and the ψ. Moreover, we can assess model perplexity, document classification accuracy, mixed topic coherence, and time efficiency based on these parameters and results.

2.2.3 Perplexity

Perplexity evaluates how well a topic model trained on training data predicts the co-occurrence of words on the unseen test data. Perplexity focuses on the topic model's ability to generate word probabilities for the unseen dataset, so a lower perplexity score indicates better generalization performance. Based on Hofmann [8], we use the "fold-in" approach for this experiment. This method evaluates the model by only learning the document-topic probabilities θ of the test dataset. All other topic model probabilities parameters keep the same from the training dataset—the validation Gibbs sampling measure only the document-topic distributions on the test documents.

In cross-collection topic model, for a test dataset of M documents, the perplexity is:

[7] http://www.michaeljpaul.com/downloads/mftm.php

$$Perplexity(D_{test}) = 2^{-\frac{1}{M} \prod_w likelihood(w|\theta_{d_{new}}, c)} \tag{2.18}$$

The M is the total number of words in all test documents. In this formula, after getting the topic probabilities θ_d and the collection c of a test document d, the likelihood of a word w in test document d is:

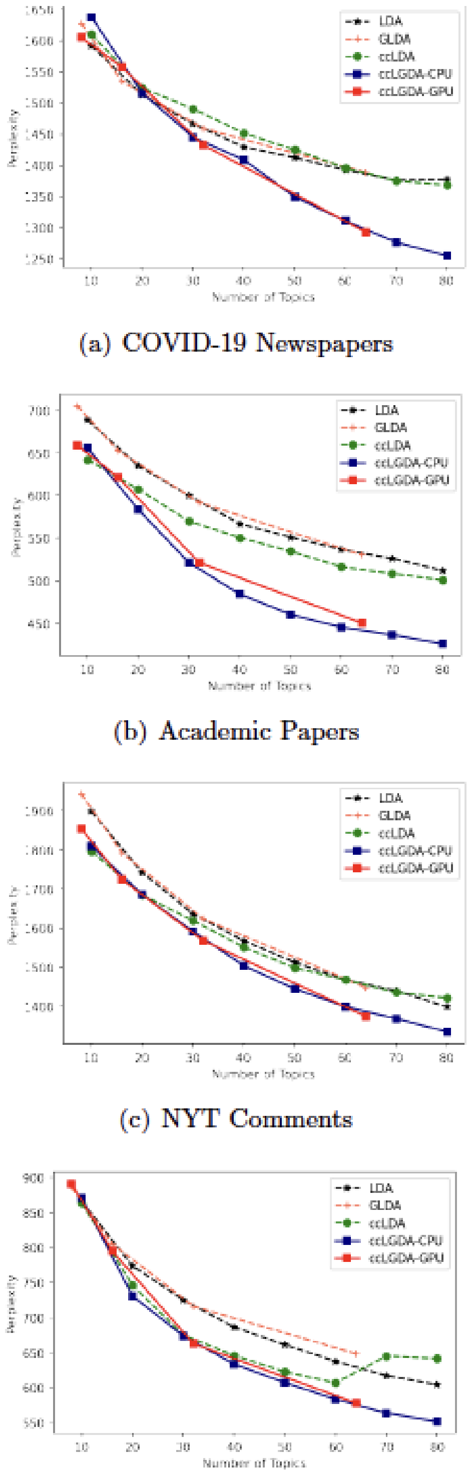
$$\begin{aligned} likelihood(w|\theta_{d_{new}}, c) = &\sum_z P(z|\theta_{d_{new}}) \\ &\times [P(w|z, x = 0)P(x = 0) + P(w|z, c, x = 1)P(x = 1)] \end{aligned} \tag{2.19}$$

$P(x = 0)$ is the probability that word w belongs to collection-independent, and $x = 1$ means the likelihood of word w being collection-specific. $P(w|z, x)$ denotes the possibility of word w sampled from collection-common or collection-specific when topic z is sampled.

Figure 2.4 presents the perplexity for each model on both corpora for different values of topics. As expected, cross-collection topic models (ccLGDA and ccLDA models) generally achieves a lower perplexity than single-collection topic model (LDA and GLDA models) because these models utilize extra information to assign a higher probability to words more likely to appear in a document. According to Figure 2.4, The ccLGDA and ccLDA models have very similar performance when the number of topics is small. With the increasing of topics, the ccLGDA models achieve lower perplexity than the ccLDA models because the GD distribution prior has better topic correlation, flexibility, generalization, and modeling capabilities [34]. This advantage can contribute to our proposed model, ccLGDA-GPU, fitting with a large dataset with a considerable number of topics. Furthermore, this experiment shows no significant difference for the perplexity results between the ccLGDA-CPU based model and the ccLGDA-GPU based model.

2.2.4 Document Classification

The cross-collection topic model can generate a document likelihood which depends on the document's collection [25], so the cross-collection models like the ccLGDA and ccLDA have the ability to make a collection prediction. In this task, each model predicts the collection of test documents based on the words. Moreover, the document



(a) COVID-19 Newspapers

(b) Academic Papers

(c) NYT Comments

(d) Traveler Forum

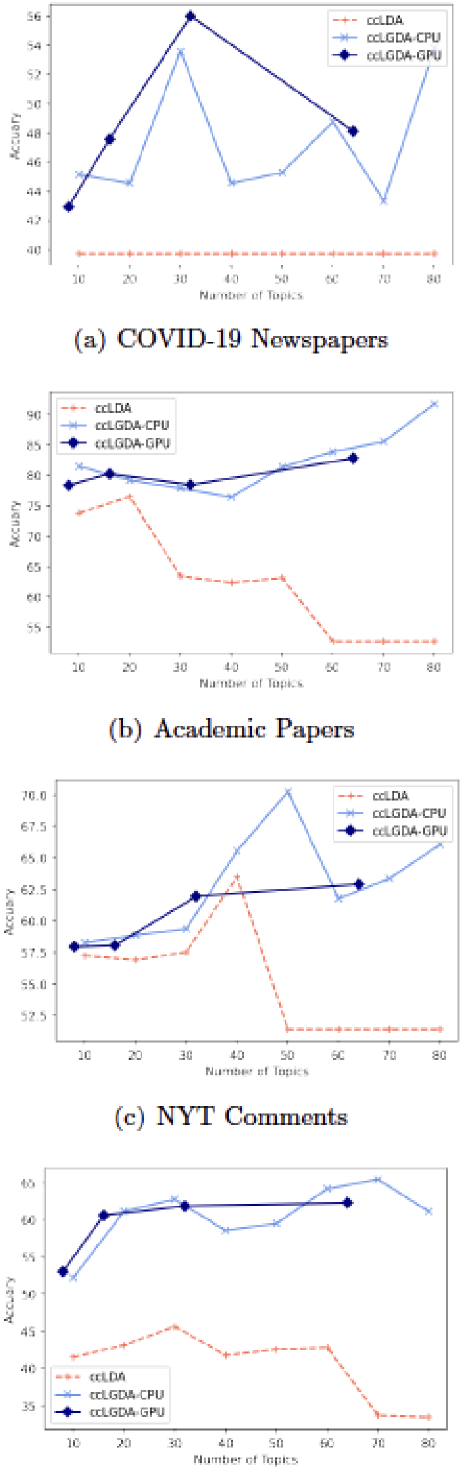
Figure 2.4: Perplexity results on four different datasets for ccLGDA-GPU based, ccLGDA-CPU based, ccLDA, LDA, and GLDA

classification accuracy can evaluate the model's separation of collection-common and collection-specific words [25, 28]. The cross-collection topic model not only assigns the most probable collection for test document, but also places a probability to each collection. This probabilistic classification allows a more detailed assessment of the degree of certainty of each topic model. Therefore, we can objectively measure the performance of these models in document classification. The cross-collection topic model calculates the category of an unlabeled document d for choosing collection c as:

$$\begin{aligned} label = arg \max_c P(c) \prod_w \sum_z P(z|\theta_{d_{new}}, c) \\ \times [P(w|z, x=0)P(x=0) + P(w|z, c, x=1)P(x=1)] \end{aligned} \tag{2.20}$$

We can get the predicted collection c by using the Eq. 2.20. Expect for $P(z|\theta_d, c)$ and $P(c)$; other probabilities are generated from the training document because $P(z|\theta_d, c)$ and $P(c)$ depend on the new test document. Following Paul's approach [25], we assign a collection c for the unlabeled document, and then we use another Gibbs sampling procedure to learn these probabilities. The classification accuracy for the new test datasets is $\frac{D_{correct}}{D_{testset}}$.

Figure 2.5 demonstrates all document classification accuracy results for four different datasets among ccLGDA and ccLDA models. The performance of the ccLGDA model is much better than the ccLDA model in the document classification task on the whole. Specifically, on academic papers and traveler forum datasets, the accuracy of our proposed GD-based model is in average 20% higher than ccLDA's accuracy. On the COVID-19 newspapers and NYT Comments datasets, the ccLGDA model also is in average 7% better than ccLDA's accuracy. Primarily, we can find the ccLDA model has a constant accuracy on the COVID-19 Newspapers dataset, and this accuracy is almost equal to the $\frac{D_{UK}}{D_{test}}$. Still, the ccLGDA model does not have the same issue and gets a reasonable accuracy. We will analyze this phenomenon in the topic analysis and discussion subsection. Furthermore, we compare the ccLGDA-CPU based model and ccLGDA-GPU based model to test the accuracy of our parallel algorithm in the cross-collection topic model. Based on Figure 2.5, we can conclude that our ccLGDA-GPU based model achieves similar performance with the CPU-based algorithm.



(a) COVID-19 Newspapers

(b) Academic Papers

(c) NYT Comments

(d) Traveler Forum

Figure 2.5: Document Classification Accuracy results on four different datasets for ccLGDA-GPU based, ccLGDA-CPU based, and ccLDA

2.2.5 Topic Coherence

The topic coherence assessment compares the ccLGDA and ccLDA models to cluster words within the collection-independent topic and between multiple collection-specific topics through semantic similarity. In particular, we evaluated the model's ability to align topics of different collections among different collection-specific topic-word distributions. However, the current topic coherence measure only considers the single word distribution per topic without handling multiple word distributions in one single topic. Therefore, we choose the mix topic coherence [28], which combines the topic representation of the collection-independent word distribution and the collection-specific word distribution. Therefore, we use the union of these representations as a single topic representation, which is independent in the different collections and distributed by specific topic terms. Then, the coherence of this union can be measured to present the current topic coherence score. Based on Risch and Krestel [28], this mixed topic coherence can also allow evaluating the topical alignment of the different collection word distributions.

For the topic coherence evaluation method, we choose the C_V method [76]. This coherence measurement is based on a sliding window, segmentation of a set of top words, indirect confirmation measures using normalized pointwise mutual information (NPMI), and cosine similarity. This coherence measure uses a sliding window and a constant window size to retrieve the co-occurrence count for a given word. These counts are used to calculate the NPMI. The segmentation of a set of top-level words results in calculating the cosine similarity between each top word vector and the sum of all full word vectors. Then C_V Coherence is the arithmetic mean of these similarities. Even though C_V coherence measurement respects correlation to human ratings, this topic coherence still has its limitations because C_V coherence assumes that words never appear together in the reference dataset are not coherent. This assumption is not suitable for some datasets with strong language contrast.

In this experiment, we use the Palmetto library[8] to evaluate the topic coherence automatically. Table 2.3 shows the C_V-based topic coherence of four datasets, which averages all topics' coherence scores. In this experiment, the number of the topic is based on the result from perplexity and document classification. From Table 2.3, we can conclude that the ccLGDA model obtains noticeably higher topic coherence values

[8] https://github.com/dice-group/Palmetto

Topic Coherence		
Dataset	ccLDA	ccLGDA
COVID-19 Newspapers	0.3832	**0.4008**
Academic Papers	0.3886	**0.4250**
NYT Comments	0.4115	**0.4174**
Traveler Forum	0.3833	**0.4015**

Table 2.3: Topic coherence comparison with ccLDA and ccLGDA models

than the ccLDA model. Especially for the COVID-19 newspaper and traveler forum dataset, our proposed model gets around 4.7% improvement. Indeed, the ccLGDA model obtains 9.6% advancement compared with the ccLDA model.

2.2.6 Time Efficiency

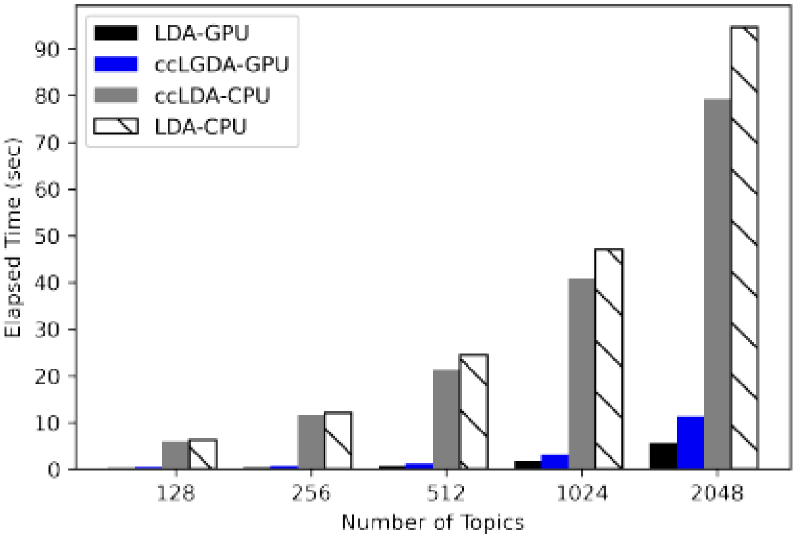


Figure 2.6: Performance comparisions for one iteration of NYT Comments dataset with number of topics varied

This section compares the time efficiency with a PC equipped with an AMD Ryzen7 5800X CPU and 16 GBytes memory. Only one CPU core is used for the CPU-based model—our time efficiency experiment on the large NYT Comments dataset for our cross-collection parallel CGS algorithm. Based on the results of perplexity and document classification, it is to be concluded that the GPU-based ccLGDA and CPU-based models have similar accuracy and effect, indicating that our parallel cross-collection topic model algorithm can produce a correct result. Figure 2.6 shows

Average Elapsed Time (sec)				
Topic	GLDA (GPU-LDA)	GPU-ccLGDA	CPU-ccLDA	CPU-LDA
128	0.23	0.41	5.9	6.3
256	0.35	0.69	11.6	12.1
512	0.68	1.2	21.3	24.5
1024	1.8	3.1	40.8	47.1
2048	5.6	11.3	79.3	94.6

Table 2.4: Average Elapsed Time comparison with GLDA (GPU-LDA), GPU-ccLGDA, CPU-ccLDA, CPU-LDA models

that our ccLGDA-GPU implementations are around 12-18X faster than CPU-ccLDA and CPU-LDA models, which are sequential programs that use only one core. Such speedup is outstanding, especially for large real-world datasets. Based on Table 2.4, we can find that our GPU-based ccLGDA model performs similarly to GLDA when the topic is below 512. Even though GLDA's performance is better than our proposed model when topic is more than 1024, GPU-ccLGDA model still keeps a reasonable elapsed time for such significant computation. This experiment demonstrates that our ccLGDA-GPU model includes high training speed and scalability for such large datasets.

2.2.7 Topics analysis and discussion

Table 2.5: ccLGDA model with three topics for COVID-19 newspapers dataset

Topic 1		**Topic 2**		**Topic 3**	
Coronaviru, health, report, viru, spread, outbread, case, public, resident, quarantine		China, world, first, chine, research, caus, report, coronaviru, year, fall		hospit, doctor, mask, nurs, die, equip, war, oxygen, surgic, healthcar	
USA	**UK**	**USA**	**UK**	**USA**	**UK**
health	case	price	anim	mask	die
office	confirm	global	Wuhan	medic	ventil
starff	contact	year	human	ventil	care
home	ship	product	sar	wear	patient
protect	infect	trade	respiratori	patient	famili
work	quarantin	market	cell	drug	nurs
depart	passeng	sale	vaccin	healthcar	intens
emerg	China	demand	bat	treat	ambul
center	flight	industri	expert	respir	away
care	patient	energi	lung	devic	age

In the COVID-19 newspapers datasets, we modeled this dataset with 30 topics according to perplexity and topic coherence results. Based on the outcome, we found the ccLDA model has a problem of word sparseness because each data source's collection-specific and collection-independent topics will be forced to align, especially for a significant gap in quantity between the collections. Because the UK collection is much more extensive than USA collections, almost all words have been assigned to the UK collection in the ccLDA model. This problem can also explain why the constant accuracy of the ccLDA model shows in the document classification experiment because all of the test documents have been assigned to one collection, which has a significant $P(c)$. Because the ccLGDA model takes advantage of GD distribution with different update equations, Eq. 2.11-2.14, the ccLGDA model overcomes the shortcoming that the collection-specific and collection-independent topic must be aligned in the ccLDA model, thereby avoiding the problem of words being scattered between the collection-specific and collection-independent topic distribution. Therefore, we mainly discuss the topics discovered by the ccLGDA model.

Table 2.5 represents the top-10 words for collection-independent and each collection local word distribution from the ccLGDA model. From the collection-independent topic words, we can quickly identify the meaning of Topic 1, which is about stopping COVID-19 from spreading. What's more, we can compare which methods are chosen between USA and UK. It is clear that the USA government let people work from home and built some emergency centers to protect public health for USA collection. The UK administration quarantines the passengers because of the confirmed cases in the ship and measures the flights from China. Topic 2 is about the China research report of Coronavirus. There is a substantial different concern between the USA and UK newspapers. We can conclude that the USA newspaper mainly focused on the virus's effect on the global economy like product price, international trade, market, and industry. The UK newspaper paid attention to the origin of the COVID-19 virus and the production of the vaccine. Besides, Topic 3 represents the treatment of COVID-19. Compared with USA newspapers, the UK collection concentrates more on the death of patients while the USA talks more about treatment equipment.

Table 2.6 compares two neural network topics learned by the ccLGDA and ccLDA model. The ccLGDA model provides better coherence in collection-specific distributions. For the collection-independent topic, both models are able to capture "neural network", "layer", "convolution", and "deep". The CVPR collection in both models manifests that many researchers pay attention to the accuracy of CNN. However,

Table 2.6: Example topics of academic papers dataset as discovered by the ccLDA and ccLGDA models

ccLDA		**ccLGDA**	
network, neural, deep, train, architectur, layer, convolut, perform. activ		network, neural, comput, layer, convolut, oper, deep, block, transform, point	
NIPS	**CVPR**	**NIPS**	**CVPR**
neural	convolut	gradient	network
weight	accuraci	method	convolute
connect	achiev	optim	neural
kernel	output	stochast	architectur
larg	result	converg	accuraci
rnn	cnn	descent	point
batch	best	nonconvex	map
recurr	separ	local	flow
care	oper	global	paramet
initi	tradit	iter	cnn

the ccLDA model shows a limitation of separating collection-specific and collection-independent words. In the ccLDA model, NIPS collection does not indicate a meaningful topic, but the ccLGDA model emphasizes optimation methods such as gradient descent and stochastic optimation in 2019 NIPS coference.

Table 2.7: Example topics of traveler forum dataset as discovered by the ccLDA and ccLGDA models

ccLDA			**ccLGDA**		
airport, flight, hour, check, arriv, luggag, time, take, termin, leav			flight, airport, book, check, arriv, leav, termin, hour, take, time		
India Collection	**Singapore Collection**	**UK Collection**	**India Collection**	**Singapore Collection**	**UK Collection**
station	airport	heathrow	book	airport	ticket
arriv	changi	gatwick	delhi	changi	airlin
mumbai	termin	allow	ticket	taxi	heathrow
airport	transit	connect	arriv	termin	london
taxi	taxi	stanst	airport	transit	global
take	hotel	checkin	mumbai	free	frill
intern	budget	think	arilin	shuttl	guid
late	citi	transfer	taxi	hotel	luggag
give	free	pari	back	train	stanst
thank	area	mean	patienc	night	connect

As shown in Table 2.7, this topic discusses transportation from the traveler forum dataset. The analysis of the ccLDA model can conclude that the ccLDA model has a good performance when the topic similarity of each data source is very high, and there is not a significant gap in quantity between the collections. For example, in the traveler forum, India, Singapore, and the United Kingdom have similar $P(c)$, and all of the collections are involved in the transportation topic. There is little difference between ccLDA and ccLGDA models in this topic, but the result of the ccLGDA model is more interpretable.

Table 2.8: ccLGDA model with three collection-independent topics for NYT Comments dataset

Topic	Topic Representation
Environment	climat, human, pollut, regul, technolog, chang, energi, scienc, natur, use, system, industri, electr, power, fossil, environment, plant, speci, much, mine
Policy	govern, constitut, right, court, nation, citizen, feder, unit, protect, free, congress, rule, secur, suprem, amend, justic, law, legal, must, first
Health	cost, health, care, price, system, compani, money, insur, expens, govern, save, afford, spend, pay, servic, free, industri, healthcar, le, tax
Crime	crime, case, call, state, would, know, crimin, deal, action, polic, claim, commit, depart, act, involv, said, victim, session, refus, person

Table 2.8 lists four different topics about public concernment from 2017 to 2018 learned by ccLGDA model on NYT Comments dataset. These topics are first 20 words from collection-independent word distribution. We can detect the public was concerning the environment, policy, health, and crime between 2017 and 2018. What's more, Table 2.9 compares world economy topics from the New York Times Comments corpus with ccLDA and ccLGDA models. Our approach, the ccLGDA model, also produces a better separation of collection-specific words and topic coherence in this dataset. ccLDA model assigns "bank" and "estate" to the 2017 collection, while the ccLGDA model assigns "worker," "job," and "wage" words for the world economy topics. Compared with estate and bank, labor cost has a much more significant effect on the world economy because real estate and banks may affect the local economy. Furthermore, both models have a similar result on the 2018 collection about China's effect on world trade. Nevertheless, ccLDA only captures the trade relationship between China and Canada. At the same time, ccLGDA assigns "China" and "global"

to 2018 collections, which is more relative to the collection-specific topic: world economy.

Table 2.9: Example topics of NYT Comments dataset as discovered by the ccLDA and ccLGDA models

ccLDA		**ccLGDA**	
busi, market, product, money, trade, compani, economi, econom, good, price		product, econom, trade, economi, manufactur, job, good, american, world, trump	
2017 Collection	**2018 Collection**	**2017 Collection**	**2018 Collection**
regul	trade	worker	trade
bank	china	job	china
estat	tariff	wage	tariff
reduc	steel	price	steel
econom	manufactur	labor	industri
growth	chine	cost	global
2008	aluminum	market	chine
doddfrank	canada	industri	countri
mortaga	industri	work	market
banker	impos	autom	deficit

Chapter 3

Cross-Collection Latent Beta-Liouville Allocation Model Training with Privacy Protection and Applications

In this chapter, to alleviate the restriction of Dirichlet prior and the significant privacy risk, we propose a cross-collection latent topic model (ccLBLA) with more flexibility and scalability by offering a better prior distribution, the Beta-Liouville distribution [77]. This is a novel enhanced cross-collection topic model that combines state-of-the-art cross-collection topic model [25], the completely LBLA model [6, 53]. To address privacy and utility issues, we present a hybrid privacy-preserving approach of the ccLBLA model (HDP-ccLBLA) based on a systematic analysis of the intrinsic differential privacy guarantee of topic model training on centralized datasets by taking advantage of HDP-LDA model [50]. The merits of our novel approach are demonstrated by experimental results in text document analysis and image recognition and categorization.

3.1 The Hybrid Privacy-preserving Approach of ccLBLA Model

This section mainly describes our Cross-Collection Latent Beta-Liouville Allocation (ccLBLA) model and the hybrid private ccLBLA framework. Our approach integrates LBLA [6,53] and ccLDA [25], and HDP-LDA [50] as a privacy preservation cross-collection topic model that takes BL distribution on both document and corpus parameters. We will start with a study of the generative process of the fundamental ccLBLA model. Then, we introduce our extension of the ccLBLA model to the hybrid privacy-preserving learning scheme applying the method on the HDP-LDA model [50], which includes cross-collection and CGS inference method with BL distribution prior. The variables in this paper are provided in Table 3.1 to allow readers to understand our models and follow easily the inference steps.

Table 3.1: Model variables and definitions

C - total number of collections
D - total number of documents
W - total number of words in each document
K - total number of topics
$\mathbf{w} = w_{ij}$ - observed words
$\mathbf{z} = z_{ij}$ - latent variables
θ_j - mixing proportions
ϕ_k - corpus parameters in collection-independent distribution
$\sigma_{k,c}$ - corpus parameters in collection-specific distribution
$\psi_{k,c}$ - parameter in Bernoulli distribution
$\theta_j \sim BL(\zeta_c)$ - generalized Dirichlet distribution
$\phi_k \sim BL(\epsilon)$ - generalized Dirichlet distribution
$\sigma_{k,c} \sim BL(\tau_c)$ - generalized Dirichlet distribution
$\psi_{k,c} \sim Beta(\gamma_0, \gamma_1)$ - Beta distribution
$x \sim Bernoulli(\psi_{ck})$ - Bernoulli distribution
$z_{jk}/\theta_{jk} \sim Mult(\theta_j)$ - multinomial distribution
$x_{jk}/z_{jk}, \phi_k, x = 0 \sim Mult(\phi_k)$ - multinomial distribution
$x_{jk}/z_{jk}, \sigma_{k,c}, x = 1 \sim Mult(\sigma_{ck})$ - multinomial distribution

3.1.1 The cross-collection LBLA model

For the complete analysis of the ccLBLA model, we will first state the generative process of the ccLBLA model, and then we will develop the inference equations when

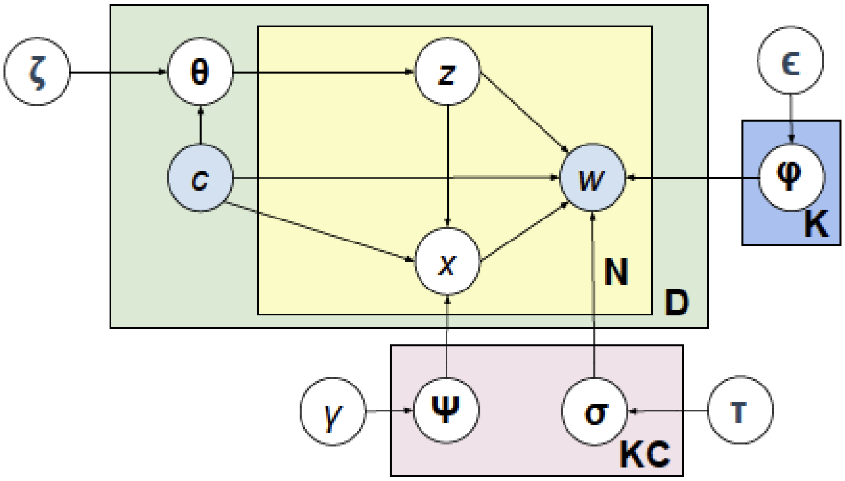


Figure 3.1: Graphical representation of ccLBLA

using the collapsed Gibbs sampling for learning (CGS-ccLBLA). The ccLBLA model first samples a collection **c** (observable data), then chooses a topic **z** and flips a coin **x** to determine whether to draw from the shared topic-word distribution or the topic's collection-specific distribution. The probability of **x** is 1 or 0 and is supported to be generated from a Bernoulli distribution.

- Draw a collection-independent multinomial word distribution ϕ_k from $BL(\epsilon)$ for each topic **z**
- Draw a collection-specific multinomial word distribution $\sigma_{k,c}$ from $BL(\tau_c)$ for each topic **z** and each collection **c**
- Draw a Bernoulli distribution $\psi_{k,c}$ from $Beta(\gamma_0, \gamma_1)$ for each topic **z** and each collection **c**
- For each document **d**, choose a collection **c** and draw a topic mixture θ_d from $BL(\zeta_c)$. Then for each word w_i in **d**:
  - Sample a topic z_i from $Mutl(\theta_d)$
  - Sample x_i from $Bernoulli(\psi_{k,c})$
  - If $x_i = 1$, sample a word w_i from $Mutl(\sigma_{k,c})$
    else $x_i = 0$, sample a word w_i from $Mutl(\phi_k)$

Because the estimation of the posterior distribution in Bayesian topic models is intractable, inference methods such as VB and MCMC have become the standard choices to estimate the latent topics and the model parameters. For the inference of the ccLBLA model, we choose collapsed space representation because it contributes to the performance of batch models [29,30]. Details about collapsed Gibbs sampling inference will be provided. Specifically, ζ_c carries the document hyperparameters α_c and β_c, ϵ includes the collection-common hyperparameters η and λ, as well as the variable τ_c holds collection-specific hyperparameters η_c and λ_c. In more detail, $(\zeta_c) = (\alpha_{c1}, ..., \alpha_{c(K-1)}, \alpha_c, \beta_c)$ means the hyperparameter set of a document with class c, and K is the number of topics. The collection-independent hyperparameter variable ϵ can be extended as $\epsilon = (\lambda_1, ..., \lambda_{V-1}, \lambda, \eta)$ while V is the size of the vocabulary or codebook. Similarly, the collection-specific hyperparameter variable ζ_c can be expressed as $\tau_c = (\lambda_{c1}, ..., \lambda_{c(V-1)}, \lambda_c, \eta_c)$ while V is also the size of the vocabulary. In our scheme, the document, topic's collection-common, and collection-specific distribution are sampled from Beta-Liouville distributions. Therefore, in our implementation, ζ_c is the $K-1$ dimensional BL hyperparameter $(\alpha_{c1}, ..., \alpha_{c(K-1)}, \alpha_c, \beta_c)$ for the document in class c in a K dimensional space. The ϵ and τ_c are the V dimensional BL hyperparameters for the vocabulary in a V dimensional space.

In collapsed space, the parameters are marginalized, leaving only the latent variables that are conditionally independent [73], and the collapsed space of latent variables is a low dimensional space as compared with joint space. Estimation in collapsed space is faster than in joint space because the parameters ϕ, σ, and θ are marginalized out. The collapsed Gibbs sampling inference approach uses a Bayesian network to estimate the posterior distributions by computing expectations through a sampling process of the latent variables. The CGS is easier to implement and computationally quicker than ordinary Gibbs sampling in the joint space. Because the CGS inference does not need the usage of digamma functions, it increases computational efficiency. As a result, when the Markov chain achieves its stationary distribution, the CGS inference delivers an accurate approximation of the actual posterior distribution. The ccLDA and its extensions [25,28] are based on CGS inference to estimate posterior distribution because of its advantages. Furthermore, in the next section, we will describe our privacy-preserving ccLBLA method by utilizing the intrinsic privacy guarantee feature of the CGS inference scheme.

In the CGS-ccLBLA scheme, the conditional probabilities of latent variable z_{ij} are calculated by the current state of all variables except the particular variable z_{ij} being processed in the marginal joint distribution $p(\mathbf{w}, \mathbf{z} \mid x_{ij} = 0, \zeta_c, \epsilon)$ or $p(\mathbf{w}, \mathbf{z} \mid x_{ij} = 1, \zeta_c, \tau_c)$ between collection-common and collection-specific case. This algorithm applies the collapsed Gibbs sampler for topic assignments. The conditional probability of z_{ij} is $p(z_{ij} = k \mid x_i = 0, \mathbf{z}_{-ij}, \mathbf{w}, \zeta_c, \epsilon)$ or $p(z_{ij} = k \mid x_i = 1, \mathbf{z}_{-ij}, \mathbf{w}, \zeta_c, \tau_c)$. The $-ij$ represents the counts with z_{ij} excluded [73]. This conditional probability of collection-common and collection-specific is expressed as:

$$p(z_{ij} = k \mid x_{ij} = 0, \mathbf{z}^{-ij}, \mathbf{w}, \zeta_c, \epsilon) = \frac{p(z_{ij} = k, z^{-ij}, \mathbf{w} \mid x_{ij} = 0, \zeta_c, \epsilon)}{p(z^{-ij}, \mathbf{w} \mid x_{ij} = 0, \zeta_c, \epsilon)} \tag{3.1}$$

$$p(z_{ij} = k \mid x_{ij} = 1, \mathbf{z}^{-ij}, \mathbf{w}, \zeta_c, \tau_c) = \frac{p(z_{ij} = k, z^{-ij}, \mathbf{w} \mid x_{ij} = 1, \zeta_c, \tau_c)}{p(z^{-ij}, \mathbf{w} \mid x_{ij} = 1, \zeta_c, \tau_c)} \tag{3.2}$$

Eq.3.1 and Eq.3.2 can be simplified as following:

$$p(z_{ij} = k \mid x_{ij} = 0, \mathbf{z}^{-ij}, \mathbf{w}, \zeta_c, \epsilon) \propto p(z_{ij} = k, z^{-ij}, \mathbf{w} \mid x_{ij} = 0, \zeta_c, \epsilon) \tag{3.3}$$

$$p(z_{ij} = k \mid x_{ij} = 1, \mathbf{z}^{-ij}, \mathbf{w}, \zeta_c, \tau_c) \propto p(z_{ij} = k, z^{-ij}, \mathbf{w} \mid x_{ij} = 1, \zeta_c, \tau_c) \tag{3.4}$$

In the CGS-ccLBLA model, the parameters θ, ϕ, and σ are drawn from the BL distribution. To speed up the training process, we marginalize these parameters in the collapsed space because sampling in the collapsed space is much faster than in the joint space of latent variables and parameters [6, 73]. By integrating out the parameters, Gibbs sampler's equations are obtained as expectation expressions:

$$\begin{aligned} &p(z_{ij} = k \mid x_{ij} = 0, \mathbf{z}^{-ij}, \mathbf{w}, \zeta_c, \epsilon) = \\ &\qquad E_{p(z_{ij}=k \mid x_{ij}=0, \mathbf{w}, \zeta_c, \epsilon)}[p(z_{ij} = k \mid x_{ij} = 0, z^{-ij}, \mathbf{w}, \zeta_c, \epsilon)] \end{aligned} \tag{3.5}$$

$$
\begin{aligned}
&p(z_{ij} = k \mid x_{ij} = 1, \mathbf{z}^{-ij}, \mathbf{w}, \zeta_c, \tau_c) = \\
&\qquad E_{p(z_{ij}=k \mid x_{ij}=1, \mathbf{w}, \zeta_c, \tau_c)}[p(z_{ij} = k \mid x_{ij} = 1, z^{-ij}, \mathbf{w}, \zeta_c, \tau_c)]
\end{aligned}
\tag{3.6}
$$

In the collapsed space, we can integrate out θ, ϕ, σ, and ψ to get Eqs.3.7 - 3.10 according to the conjugacy of the Beta/Binomial and BL/Multinomial distributions [78, 79] based on the inference equations developed for CGS-ccLDA and CGS-LBLA [6, 25]. In CGS algorithm iterations, we sample new assignment of $\mathbf{z}$ and $\mathbf{x}$ alternately with the following equations:

$$
\begin{aligned}
&p(z_i = k \mid x_i = 0, \mathbf{z}_{-i}, \mathbf{w}, \zeta_c, \epsilon) \propto \\
&\frac{(\alpha_{ck} + N_{jk}^{-ij})}{(\sum_{l=1}^{K-1} \alpha_{cl} + \sum_{l=1}^{K-1} N_{jl}^{-ij})} \times \frac{(\alpha_c + \sum_{l=1}^{K-1} N_{jl}^{-ij})}{(\alpha_c + \beta_c + \sum_{l=1}^{K} N_{jl}^{-ij})} \\
&\times \frac{(\lambda_v + N_{kv}^{-ij})}{(\sum_{l=1}^{V-1} \lambda_l + \sum_{l=1}^{V-1} N_{kl}^{-ij})} \times \frac{(\lambda + \sum_{l=1}^{V-1} N_{kl}^{-ij})}{(\lambda + \eta + \sum_{l=1}^{V} N_{kl}^{-ij})}
\end{aligned}
\tag{3.7}
$$

$$
\begin{aligned}
&p(x_i = 0 \mid x_{-i}, \mathbf{z}, \mathbf{w}, \gamma, s, t) \propto \\
&\frac{N_{x=0}^{k,c} + \gamma_0}{N_{\cdot}^{k,c} + \gamma_0 + \gamma_1} \times \frac{(\lambda_v + N_{kv}^{-ij})}{(\sum_{l=1}^{V-1} \lambda_l + \sum_{l=1}^{V-1} N_{kl}^{-ij})} \times \frac{(\lambda + \sum_{l=1}^{V-1} N_{kl}^{-ij})}{(\lambda + \eta + \sum_{l=1}^{V} N_{kl}^{-ij})}
\end{aligned}
\tag{3.8}
$$

For Eq.3.7 and Eq.3.8, all counts only refer to the words for which $x_i = 0$, which are the words assigned to the topic model. Specifically, N is the total number of words for which $x_i = 0$, not the total number of words in the corpus. Same for Eq.3.9 and Eq.3.10, the count only includes the words for which $x_i = 1$, which means that N is the total number of words for which $x_i = 1$.

$$
\begin{aligned}
&p(z_i = k \mid x_i = 1, \mathbf{z}_{-i}, \mathbf{w}, \zeta_c, \tau_c) \propto \\
&\frac{(\alpha_{ck} + N_{jk}^{-ij})}{(\sum_{l=1}^{K-1} \alpha_{cl} + \sum_{l=1}^{K-1} N_{jl}^{-ij})} \times \frac{(\alpha_c + \sum_{l=1}^{K-1} N_{jl}^{-ij})}{(\alpha_c + \beta_c + \sum_{l=1}^{K} N_{jl}^{-ij})} \\
&\times \frac{(\lambda_{cv} + N_{ckv}^{-ij})}{(\sum_{l=1}^{V-1} \lambda_{cl} + \sum_{l=1}^{V-1} N_{ckl}^{-ij})} \times \frac{(\lambda_c + \sum_{l=1}^{V-1} N_{ckl}^{-ij})}{(\lambda_c + \eta_c + \sum_{l=1}^{V} N_{ckl}^{-ij})}
\end{aligned}
\tag{3.9}
$$

$$p(x_i = 1 \mid x_{-i}, \mathbf{z}, \mathbf{w}, \gamma, \tau_c) \propto$$

$$\frac{N^{k,c}_{x=1} + \gamma_1}{N^{k,c}_{\cdot} + \gamma_0 + \gamma_1} \times \frac{(\lambda_{cv} + N^{-ij}_{ckv})}{(\sum_{l=1}^{V-1} \lambda_{cl} + \sum_{l=1}^{V-1} N^{-ij}_{ckl})} \times \frac{(\lambda_c + \sum_{l=1}^{V-1} N^{-ij}_{ckl})}{(\lambda_c + \eta_c + \sum_{l=1}^{V} N^{-ij}_{ckl})} \tag{3.10}$$

The count N^{ij}_{jk} is the number of words w_i in the document j and topic k in class c. Besides, N^{-ij}_{jk} is the total number of words in document j and topic k in class c except for the word w_i being sampled. The count $N^{ij}_{kw_{ij}}$ is the number of times the word w_{ij} appears in topic k and document j. In addition, $N^{-ij}_{kw_{ij}}$ is the number of times the word w_{ij} appears in document j and topic k except being sampled. $N^{ij}_{ckw_{ij}}$ is the number of times the word w_{ij} appears in topic k and document j in specific collection $\mathbf{c}$. In addition, $N^{-ij}_{ckw_{ij}}$ is the number of times the word w_{ij} appears in document j and topic k in specific collection $\mathbf{c}$ except being sampled. $N^{k,c}_{x}$ is the number of $\mathbf{x}$ in topic k, and collection c. $\mathbf{x}$ should be initialized as 0 for all tokens. We initially assume that everything comes from the shared collection word distribution.

For parameters estimation, the document parameter distribution is:

$$\theta_{jk} = \frac{(\alpha_{ck} + N_{jk})}{(\sum_{l=1}^{K-1} \alpha_{cl} + \sum_{l=1}^{K-1} N_{jl})} \times \frac{(\alpha_c + \sum_{l=1}^{K-1} N_{jl})}{(\alpha_c + \beta_c + \sum_{l=1}^{K} N_{jl})} \tag{3.11}$$

The predictive distributions of the collection-independent and collection-specific words are:

$$\phi_{kw} = \frac{(\lambda_v + N_{kv})}{(\sum_{l=1}^{V-1} \lambda_l + \sum_{l=1}^{V-1} N_{kl})} \times \frac{(\lambda + \sum_{l=1}^{V-1} N_{kl})}{(\lambda + \eta + \sum_{l=1}^{V} N_{kl})} \tag{3.12}$$

$$\sigma_{ckw} = \frac{(\lambda_{cv} + N_{ckv})}{(\sum_{l=1}^{V-1} \lambda_{cl} + \sum_{l=1}^{V-1} N_{ckl})} \times \frac{(\lambda_c + \sum_{l=1}^{V-1} N_{ckl})}{(\lambda_c + \eta_c + \sum_{l=1}^{V} N_{ckl})} \tag{3.13}$$

3.1.2 Hybrid Privacy-preserving ccLBLA scheme

This section will first introduce the differential privacy and exponential mechanism. Then, we demonstrate a thorough analysis of the inherent differential privacy guarantee of CGS-ccLBLA training on centralized datasets. We will present a hybrid privacy-preserving method for the cross-collection topic model (HDP-ccLBLA) based on the study above. In the HDP-ccLBLA scheme, all the intermediate statistics of

Algorithm 3 Summary of CGS-ccLBLA model

```
procedure
Input: $\mathbf{w}$, $\zeta_c$, $\tau_c$, $\epsilon$, iterMax, $K$, $V$, $N$
initialize $\mathbf{z}$, $\mathbf{x}$, $N_{jk}$, $N_{kw}$, $N_{ckw}$, $N_x$
for iter = 1 to iterMax do
    for $i = 1$ to $N$ in document $j$ in class $c$ do
        if $x_{ij} = 0$ then
            update $z_{ij}$ using Eq.3.7
        else
            update $z_{ij}$ using Eq.3.9
        end if
        update $x_{ij}$ using Eq.3.8 and Eq.3.10
        update $N_{jk}$, $N_{kw}$, $N_{ckw}$, $N_x$
    end for
end for
Output: Parameters $\theta_j, \phi_k, \sigma_{ck}$ using Eq. 3.11 - 3.13
end procedure
```

the CGS-ccLBLA model can be protected during the training process.

Differential privacy and exponential mechanism

Differential privacy [23] is a de-facto standard for privacy protection framework with a rigorous mathematical proof. So far, DP has been widely utilized in the past to assess the privacy issue of random algorithms by comparing the mathematical differences between neighboring datasets.

Theorem 1 (Differential Privacy [23]) *A randomized mechanism $f : \mathbf{D} \longrightarrow \mathbf{Y}$ offers $(\epsilon, \delta - DP)$ if for any adjacent $D, D' \in \mathbf{D}$ and $Y \in \mathbf{Y}$, there is:*

$$Pr(f(D) \in \mathbf{Y}) \leq e^{\epsilon} Pr(f(D') \in \mathbf{Y}) + \delta \tag{3.14}$$

The $Pr()$ refers to the probability and ϵ is the privacy level of f. This definition restrains an adversary's ability to infer whether the training or input dataset is D or D'.

According to Dework et al. [23], exponential machanism is a base approach to obtain $\epsilon - DP$. The main concern of exponential mechanism is to return the result sampled from a definite distribution with a fixed output set.

Theorem 2 (Exponential Machanism [23]) *Given a range R, a dataset D, a function u, and a privacy parameter ϵ, the mechanism $\mathcal{M}_E(x, u, \mathbf{R}) : D \longrightarrow R$ satisfies $\epsilon - DP$ if $\mathcal{M}_E(x, u, \mathbf{R})$ output an element $r \in \mathbf{R}$ with probability Pr satisfies that:*

$$Pr \propto exp(\frac{\epsilon}{2 \bigtriangleup u} u(x, r)) \tag{3.15}$$

where $u(x, r)$ is the utility function and $\bigtriangleup u$ is sensitivity.

Inherent Privacy of CGS inference scheme

Because Gibbs sampling has the same process with an exponential mechanism for differential privacy, Foulds et al. [52] highlighted that the Gibbs sampling method inherently generates some degree of intrinsic differential privacy. The CGS technique has the same property since it is one of the versions of Gibbs sampling. Furthermore, during each iteration of learning a topic-word distribution, the CGS inference outputs a topic from the topics set. Thus, Zhao et al. [50, 80] began to investigate the CGS process in terms of the exponential mechanism, and they sucessfully concluded the inherent privacy of the CGS algorithm in LDA model. They indeed specifically analyse the intrinsic privacy loss in each iteration before composing the privacy in total interactions of the CGS training scheme of LDA. We will employ the same concepts and then extend this idea to our proposed model so we will use the same propositions in HDP-ccLBLA model.

According to Zhao et al. [50], the intrinsic privacy of LDA's CGS inference technique has two significant drawbacks. First, because privacy loss grows linearly, the privacy loss will accumulate rapidly. Second, during the CGS inference process, there is no protection for word-count information since intrinsic privacy cannot secure the word-count data, leading to a privacy leakage issue. We will address these two potential difficulties of inherent privacy after leveraging CGS's inherent privacy feature and present a privacy-preserving solution for our model (HDP-ccLBLA).

Hybrid Privacy-preserving ccLBLA algorithm

The final hybrid privacy-preserving model (HDP-ccLBLA) described in this section integrates the inherent privacy of the CGS inference approach with external privacy provided by noise injection. We provide suitable noise in each iteration of the CGS technique to secure the word-count statistical information to overcome the possible privacy concern of intrinsic privacy. We introduce the noise to obfuscate the difference between N_{dk} or N_{cdk} in each iteration. Besides, we minimize the rapid accumulation of privacy loss by setting the upper bound of the topic-word count. We choose the same method for HDP-LDA [50], which resorts to a clipping method to restrict the inherent privacy in each iteration. Specifically, the clipping only impacts a copy of N_{dk} or N_{cdk} in the computation of sampling but not the updating of CGS inference. Algorithm 4 meets $(\epsilon_L + \epsilon_I) - DP$ in each iteration. ϵ_I is the inherent privacy loss:

$$\epsilon_I = \begin{cases} 2\log(\frac{C}{\lambda_v} + 1), & \text{if } x_{ij} = 0 \\ 2\log(\frac{C}{\lambda_{cw}} + 1), & \text{if } x_{ij} = 1 \end{cases} \tag{3.16}$$

The ϵ_L denotes the privacy loss incured by the Laplace noise, and the C is the clipping bound for N_{dk} or N_{cdk}.

In Algorithm 4, the privacy loss in the HDP-ccLBLA model includes privacy loss ϵ_L incurred by Laplace noise and the inherent privacy loss ϵ_I of CGS inference. According to Eq.3.16, we can conclude that the rapid increase of inherent privacy loss has been limited, and the word-count statistical information also gets privacy protection.

3.2 Experimental results

The cross-collection Allocation model was evaluated via perplexity, classification accuracy, and topic coherence using several applications such as comparative text mining and image classification. We also compare topic examples across multiple text datasets to demonstrate the strengths of our technique. The experiments utilize four text datasets with different collection numbers, document lengths, domains, and one well-known image dataset. In this section, we use the Scale Invariant Feature Transform (SIFT) and K-means approaches to successfully apply our cross-collection topic model (ccLBLA) to an image classification assignment using the Bag of Visual Words (BOVW) approach. Finally, we primarily validate the HDP-ccLBLA algorithm's performance in terms of model utility such as perplexity to show our

Algorithm 4 Summary of HDP-ccLBLA algorithm

```
procedure
    Input: $\mathbf{w}$, $\zeta_c$, $\tau_c$, $\epsilon$, iterMax, $K$, $V$, $N$
    initialize $\mathbf{z}$, $\mathbf{x}$, $N_{jk}$, $N_{kw}$, $N_{ckw}$, $N_x$
    for iter = 1 to iterMax do
        for $i = 1$ to $N$ in document $j$ in class $c$ do
            $\boldsymbol{\eta} \sim Lap(\frac{2}{\epsilon_L})$
            if $x_{ij} = 0$ then
                Add noise to $N_{kw}$
                $N_{kw} = N_{kw} + \boldsymbol{\eta}$
                Clip: $(N_{kw})^{temp} = min(N_{kw}, C)$
                Compute: $\epsilon_I = 2log(\frac{C}{\lambda_v} + 1)$
                update $z_{ij}$ using Eq.3.7
            else
                Add noise to $N_{ckw}$
                $N_{ckw} = N_{ckw} + \boldsymbol{\eta}$
                Clip: $(N_{ckw})^{temp} = min(N_{ckw}, C)$
                Compute: $\epsilon_I = 2log(\frac{C}{\lambda_{cv}} + 1)$
                pdate $z_{ij}$ using Eq.3.9
            end if
            update $x_{ij}$ using Eq.3.8 and Eq.3.10
            update $N_{jk}$, $N_{kw}$, $N_{ckw}$, $N_x$
        end for
    end for
    Output: Parameters $\theta_j, \phi_k, \sigma_{ck}$ using Eq. 3.11 - 3.13
    Output: Privacy loss $\epsilon = (\epsilon_L + \epsilon_I)$
end procedure
```

approach's merits.

3.2.1 The Datasets

Table 3.2 displays an overview of each dataset size for the text datasets. The COVID-19 newspapers dataset contains online newspapers from the United States of America, which is collected from COVID-NEWS-US-NNKDATASET[1] [74]. Besides, the second collection of this dataset is from several different British newspapers websites[2]. Indeed, we can use this novel dataset for comparative text mining tasks in aggregation and summarization to extract common and different effects and knowledge about the virus in two different countries and demonstrate the merits of our proposed model. Besides, the second text dataset mainly focuses on the field of computer science academic papers, including the abstracts of NeurIPS[3] and CVPR[4] papers published in 2019. We apply our model to comparative text analysis to automatically spot different topics and trends in these two different conferences. The third text dataset consists of a subset of the New York Times (NYT) comments[5], which contains more than two million comments from 2017 to 2018. Because some comments are discarded if the minimum number of words is less than one hundred, we decided to take advantage of a two-month comments dataset between 2017 and 2018 to compare the performance of the ccLBLA model with ccLDA [25], and LDA [29] models. We also reuse the dataset[6] reported in ccLDA [25] so that we can make a fair comparison. The last text dataset crawled from an online travel platform including three different countries' discussion forums of India, Singapore, and the UK, with thousands of threads in each collection [25]. Therefore, our experiment utilized four domains of datasets newspapers, academic papers, customer comments, and travel blogs, to prove that our approach can handle different types of documents.

For the image-based application, we used the famous grayscale natural scenes dataset [81]. As shown in Table 3.3 and Fig.3.2, this image dataset includes the following categories: kitchen, office, bedroom, suburb, highway, living room, street, downtown, industry, store, forest, skyscraper, coast, mountain, and rural area.

[1] https://github.com/nnk-dataset/usa-nnk

[2] https://www.kaggle.com/jwallib/coronavirus-newspaper-classification/data

[3] https://www.kaggle.com/rowhitswami/nips-papers-1987-2019-updated

[4] https://www.kaggle.com/paultimothymooney/cvpr-2019-papers

[5] https://www.kaggle.com/aashita/nyt-comments

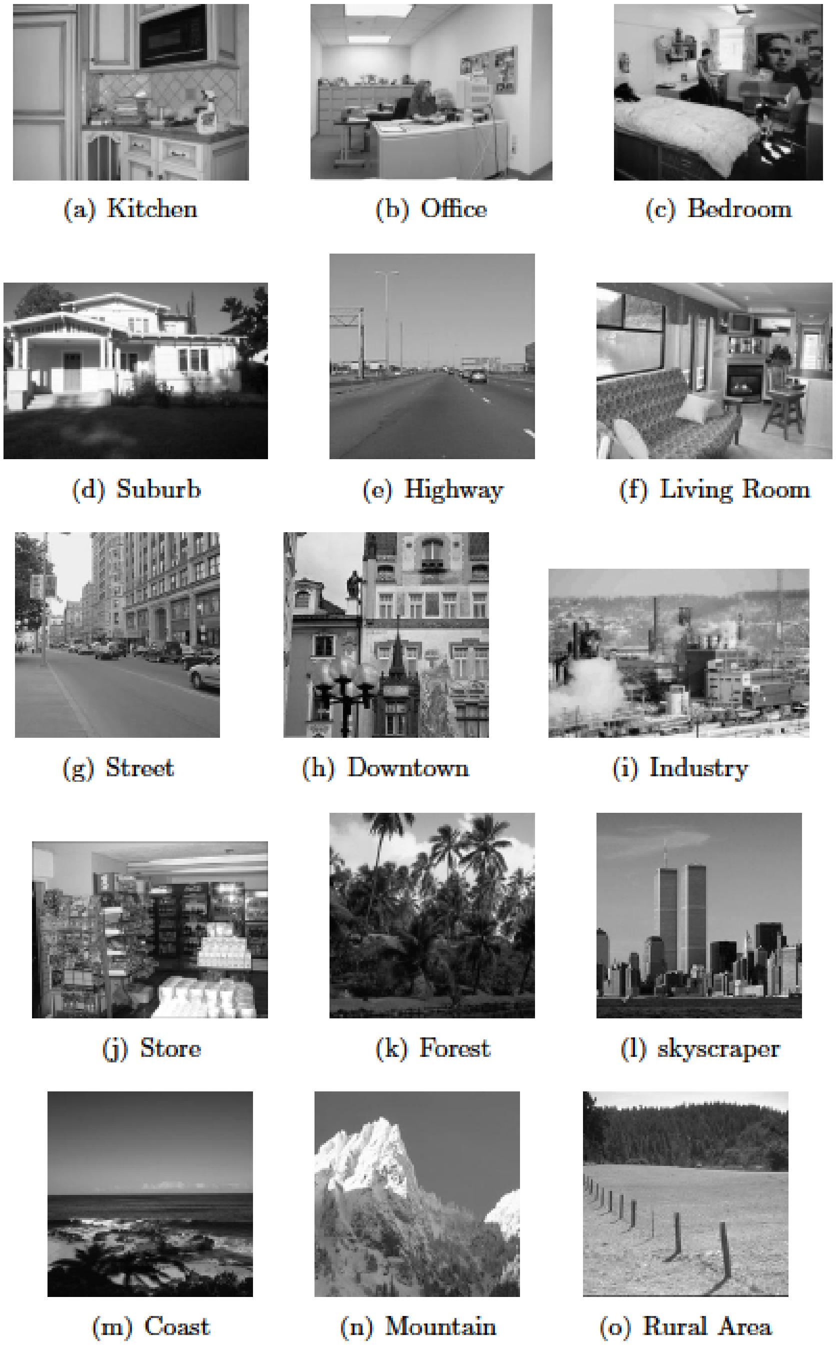
[6] http://www.michaeljpaul.com/downloads/ccdata.php

Table 3.2: Datasets - number of documents D and average number of words per document W/D (without stop words)

Text Datasets			
Dataset	Collection	D	W/D
COVID-19 Newspapers	USA/UK	2731	433
Academic Papers	NIPS/CVPR	2787	91
NYT Comments	2017/2018	74895	127
Traveler Forum	India/Singapore/UK	4174	247

Table 3.3: Size of each image category

Natural scenes images dataset	
Categories	Size
Kitchen	210
Office	215
Bedroom	216
Suburb	241
Highway	260
Living Room	289
Street	292
Downtown	308
Industry	311
Store	315
Forest	328
Skyscraper	356
Coast	360
Mountain	374
Rural Area	410



(a) Kitchen (b) Office (c) Bedroom

(d) Suburb (e) Highway (f) Living Room

(g) Street (h) Downtown (i) Industry

(j) Store (k) Forest (l) skyscraper

(m) Coast (n) Mountain (o) Rural Area

Figure 3.2: Examples from the natural scenes images dataset (Total Fifteen Categories)

3.2.2 Experiments for text mining

For comparative text mining application, we preprocess the text datasets by first tokenizing words with the Natural Language ToolKit (NLTK) [75], removing punctuation, stop-words, and then lemmatizing tokens to derive their common base form. We choose BL priors hyperparameters following the same setting of the asymmetric BL priors in [6]. For Dirichlet based model, the topic distribution priors are fixed and $\alpha = 0.1$. Then, we set β and δ to 0.01; for γ_0 and γ_1, we use the same value, 1.0. The LDA and ccLDA (LDA and ccLDA[7]) are based on a widely used open-source package GibbsLDA++. For the text experiment validation, we use ten-fold cross-validation, which separates each dataset with a 90% training set and 10% test set. In the Gibbs sampling, the burn-in period is five hundred, and then we collect ten samples separated by lags of ten iterations. The average of ten samples is the final result of the model. After, we calculate the document-topic parameter θ, the collection-independent word distribution parameter ϕ, the collection-specific word distribution parameters σ, and ψ. Moreover, we assessed model perplexity, document classification accuracy, and mixed topic coherence based on these parameters and results.

Perplexity

Perplexity evaluates how well a trained topic model predicts the co-occurrence of words on the unseen test data. Perplexity focuses on the topic model's ability to generate word probabilities for the unseen dataset, so a lower perplexity score indicates better generalization performance. Based on Hofmann [8], we use the "fold-in" approach for this experiment. This method evaluates the model by only learning the document-topic probabilities θ of the test dataset. All other topic model probabilities parameters are kept the same from the training dataset—the validation Gibbs sampling measure only the document-topic distributions on the test documents.

In cross-collection topic model, for a test dataset of M documents, the perplexity is:

$$Perplexity(D_{test}) = 2^{-\frac{1}{M}\prod_w likelihood(w|\theta_{d_{new}},c)} \tag{3.17}$$

[7] http://www.michaeljpaul.com/downloads/mftm.php

In this formula, after getting the topic probabilities θ_d and the collection c of a test document d, the likelihood of a word w in test document d is:

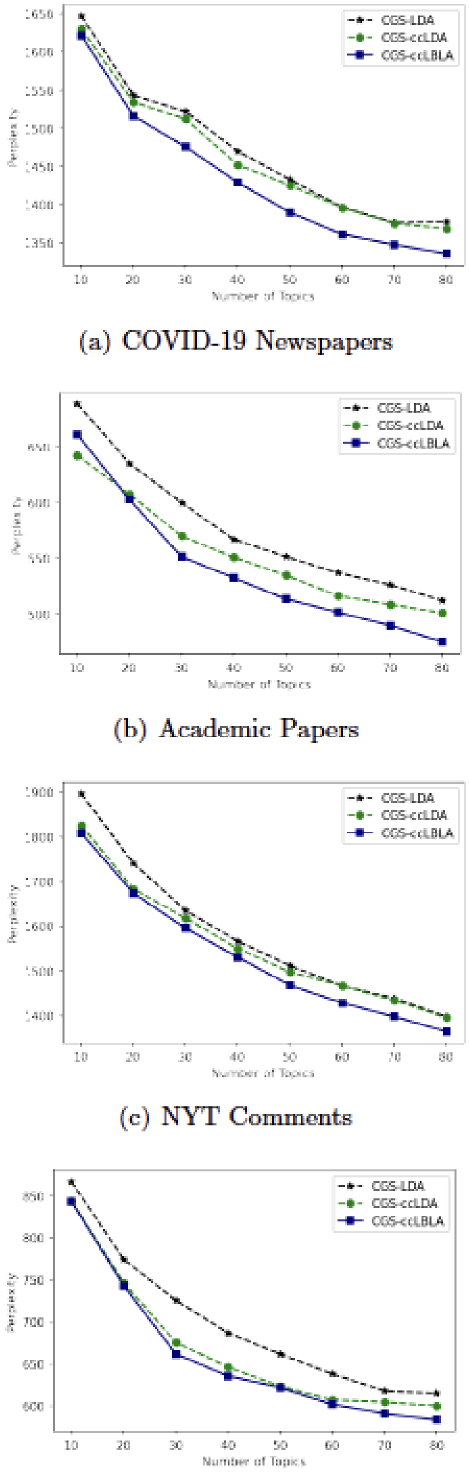
$$\begin{aligned} likelihood(w \mid \theta_{d_{new}}, c) = \sum_{z} P(z \mid \theta_{d_{new}}) \\ \times [P(w \mid z, x = 0) P(x = 0) + P(w \mid z, c, x = 1) P(x = 1)] \end{aligned} \tag{3.18}$$

$P(x = 0)$ is the probability that word w is collection-independent, and $x = 1$ means the likelihood of word w being collection-specific. $P(w \mid z, x)$ denotes the possibility of word w sampled from collection-common or collection-specific when topic z is sampled.

The perplexity for each model on both corpora for different values of topics is shown in Fig.3.3. As expected, cross-collection topic models (ccLBLA and ccLDA) generally achieve a lower perplexity than single-collection topic model (LDA) because these models utilize extra information to assign a greater probability to words more likely to exist in a document. According to Fig.3.3, The ccLBLA and ccLDA models have comparable performance when the number of topics is negligible since the topic number is not ideal for specific datasets. The ccLBLA models achieve lower perplexity than the ccLDA models as the number of topics increases. The ccLBLA and ccLDA models produce pretty similar results on the traveler forum dataset, although the difference between the two models is not significant. After examining the traveler forum dataset, we notice that each collection contains many duplicate documents, implying that this dataset cannot accurately demonstrate the capabilities of a cross-collection topic model to predict unseen documents. In the other three text datasets, ccLBLA has a lower perplexity than the ccLDA model. The main reason may be that the BL distribution prior has better topic correlation, flexibility, generalization, and modeling capabilities [6, 53].

Document Classification

Cross-collection topic models like ccLBLA and ccLDA are capable of producing collection predictions for unseen documents since they can generate a document likelihood that relies on the document's collection [25]. Each model predicts the collection



(a) COVID-19 Newspapers

(b) Academic Papers

(c) NYT Comments

(d) Traveler Forum

Figure 3.3: Perplexity results on four different datasets for LDA, ccLDA and ccLBLA

of test documents based on the words in this task. Furthermore, the document classification accuracy may be used to assess the model's separation of collection-common and collection-specific words [25, 28]. The cross-collection topic model provides a probability for each collection and assigns the most likely collection for the test document. This probabilistic classification enables a more precise assessment of each topic model's degree of certainty. Therefore, we can objectively measure the performance of these models in document classification. The cross-collection topic model calculates the category of an unlabeled document d for choosing collection c as:

$$\begin{aligned} label = arg \max_{c} P(c) \prod_{w} \sum_{z} P(z \mid \theta_{d_{new}}, c) \\ \times [P(w \mid z, x = 0) P(x = 0) + P(w \mid z, c, x = 1) P(x = 1)] \end{aligned} \tag{3.19}$$

We can get the predicted collection c by using Eq.3.19. Expect for $P(z \mid \theta_d, c)$ and $P(c)$; other probabilities are generated from the training document because $P(z \mid \theta_d, c)$ and $P(c)$ depend on the new test document. Following Paul's approach [25], we assign a collection c for the unlabeled document, and then we use another Gibbs sampling procedure to learn these probabilities. The classification accuracy for the new test datasets is $\frac{D_{correct}}{D_{testset}}$.

Table 3.4 demonstrates all document classification accuracy results for four different datasets among ccLBLA and ccLDA models. The performance of the ccLBLA model is much better than the ccLDA model in the document classification task on the whole. On the COVID-19 newspapers dataset, the document classification accuracy of the ccLBLA model is almost 45% higher than the ccLDA model. Also, the ccLBLA model achieves about 40% greater than ccLDA's accuracy. The ccLBLA model gets about 19% higher accuracy in academic papers and NYT comments datasets than the other two datasets. Based on those results, compared with the ccLDA model, we can conclude that the ccLBLA model obtains a better ability to separate collection-common and collection-specific words by introducing BL distribution.

Topic Coherence

The topic coherence evaluation compares the ccLBLA and ccLDA models for clustering words inside the collection-independent topic and between multiple collection-specific topics through semantic similarity. In particular, the model's capacity to

Table 3.4: Document classification accuracy results on four different datasets for ccLDA and ccLBLA

Document Classification Accuary		
Dataset	ccLDA	ccLBLA
COVID-19 Newspapers	0.40	**0.59**
Academic Papers	0.76	**0.91**
NYT Comments	0.63	**0.75**
Traveler Forum	0.45	**0.63**

align topics from distinct collections among different collection-specific topic-word distributions was tested. The current topic coherence metric, on the other hand, only examines a single word distribution per topic, not several word distributions inside a single topic. As a result, we use the mix topic coherence [28], which mixes the topic representation of the collection-independent word distribution with the collection-specific word distribution. As a result, we employ the union of these representations as a unified topic representation, which is distributed by particular topic terms and is independent of the individual collections. The coherence of this union can be evaluated in order to determine the current topic coherence score.

This mixed topic coherence may also be used to evaluate the topical alignment of different collection word distributions according to Risch and Krestel [28]. The C_V technique [76] is chosen as the topic coherence evaluation method. This coherence measurement is based on a sliding window, segmentation of a set of top words, indirect confirmation measures using normalized pointwise mutual information (NPMI), and cosine similarity. This coherence metric retrieves the co-occurrence count for a given word using a sliding window and a constant window size. The NPMI is calculated using these counts. When a collection of top-level words is segmented, the cosine similarity between each top word vector and the sum of all complete word vectors is calculated. The arithmetic mean of these similarities is thus C_V Coherence. Despite the fact that C_V coherence measurement takes into account human judgments, this topic coherence has limits since C_V coherence implies that words that never appear together in the reference dataset are not consistent. This assumption is not suitable for some datasets with strong language contrast.

In this experiment, we use the Palmetto library[8] to evaluate the topic coherence automatically. Table 3.5 shows the C_V-based topic coherence of four datasets, which

[8] https://github.com/dice-group/Palmetto

averages all topics' coherence scores. In the mixed topic coherence evaluation, the number of the topic is based on the result from perplexity and document classification. From Table 3.5, we can conclude that the ccLBLA model obtains slightly higher topic coherence values than the ccLDA model. Especially for the academic papers dataset, our proposed model gets around 8.3% improvement. Indeed, the ccLBLA model obtains almost 4.5% advancement compared with the ccLDA model in COVID-19 newspapers and travler forum dataset.

Table 3.5: Topic coherence comparison with ccLDA and ccLBLA models

Topic Coherence		
Dataset	ccLDA	ccLBLA
COVID-19 Newspapers	0.3832	**0.4008**
Academic Papers	0.3886	**0.4211**
NYT Comments	0.4115	**0.4182**
Traveler Forum	0.3833	**0.4013**

Topics analysis and discussion

We modeled this dataset with 30 topics based on perplexity and topic coherence findings in the COVID-19 newspapers datasets. The top-10 words for collection-independent and each collection local word distribution from the ccLBLA model are shown in Table 3.6. Topic 15, which is about maintaining public health during the Covid-19 pandemic, may be deduced from the collection-independent topic terms. Indeed, when comparing the methods used in the United Kingdom and the United States, it is evident that the United States government advises individuals to work from home and stay at a safe distance from public places to prevent the spread of Covid-19 in the USA collection. The UK government recommends that people wear masks and wash their hands to protect themselves.

Moreover, Topic 19 presents the symptom of COVID-19. The topic 23 is a Coronavirus study report. The newspapers in the United States and the United Kingdom have distinct concerns. The virus's instances and patients in China were the emphases of the US newspaper. In contrast, the COVID-19 virus's data across the world and vaccine manufacture were the focus of the UK media.

Furthermore, Table 3.7 compares ccLDA and ccLBLA models to world economic issues from the New York Times Comments dataset. Our method, the ccLBLA model,

Table 3.6: ccLBLA model with three topics for COVID-19 newspapers dataset

Topic 15		**Topic 19**		**Topic 23**	
Coronaviru, health, work, week, continu, viru, time emerg, countri, clear		Symptom, infect, viru day, ill, coronaviru, sever people, cough, fever		viru, diseas, conronaviru anim, vaccin, spread, human research, studi, scientist	
UK Collection	**USA Collection**	**UK Collection**	**USA Collection**	**UK Collection**	**USA Collection**
mask	peopl	health	hand	vaccin	infect
worker	govern	case	breath	world	China
suppli	stay	peopl	test	data	patient
protect	home	covid19	cough	use	Wuhan
wear	social	test	covid19	develop	outbreak
face	test	viru	lung	medium	ill
product	distanc	infect	suffer	research	test
hand	offic	diseas	bodi	work	pandem
equip	public	spread	throat	inform	cent
hospit	rule	death	clean	report	public

also results in superior separation of collection-specific terms and theme coherence in this dataset. The 2017 collection is assigned the terms ”bank” and ”estate” by the ccLDA model, whereas the world economy themes are assigned the words ”job,” ”work,” and ”worker” by the ccLBLA model. Labor costs have a considerably more significant impact on the global economy than real estate and banks because real estate and banks can affect the local economy. What is more, both models provide the same outcome in the 2018 collection regarding China’s impact on global commerce. The ccLDA model, on the other hand, is limited to the Sino-Canadian economic connection. ccLGDA, on the other hand, assigns ”China” and ”global” to 2018 collections, which is more relevant to the collection’s specific topic: the global economy.

3.2.3 Image classification

In this section, we successfully apply the cross-collection topic model in image classification application following the bag of visual words framework [4, 6]. Fig. 3.4 illustrates an overview of the feature extraction, clustering, and ccLBLA pipeline. Specifically, we use the Scale Invariant Feature Transform (SIFT) algorithm to extract the local features from local patches through the whole corpus collection, the vectors of counts in each image. The K-means algorithm clusters the set of training image descriptors to find the unique local feature representation. After that, we can obtain the codeword from the cluster center and the codebook or the dictionary of image vocabulary. The codebook contains a vector of counts for each image. Using this

Table 3.7: Example of topics from the NYT Comments dataset as discovered by the ccLDA and ccLBLA models

ccLDA		ccLBLA	
busi, market, product, money, trade, compani, economi, econom, good, price		econom, economi, job, polici, worker, increas, corpor, employ, product, cost	
2017 Collection	**2018 Collection**	**2017 Collection**	**2018 Collection**
regul	trade	job	trade
bank	china	work	china
estat	tariff	worker	market
reduc	steel	labor	global
econom	manufactur	class	industri
growth	chine	busi	good
2008	aluminum	incom	compani
doddfrank	canada	rate	rate
mortaga	industri	rich	cost
banker	impos	growth	product

bag of visual words approach, we can consider each image as a document and train them into our proposed ccLBLA model. Besides, in this well-known grayscale fifteen-categories natural scenes dataset, the data is separated into training and testing sets in each category: the testing set has a hundred random images while the remaining constitute the training set. In the model section, we set the range of topic numbers from 10 to 80. Then, we can use the bags of visual words representation for each image to evaluate the performance of the ccLBLA model in the image classification task based on Eq.3.19.

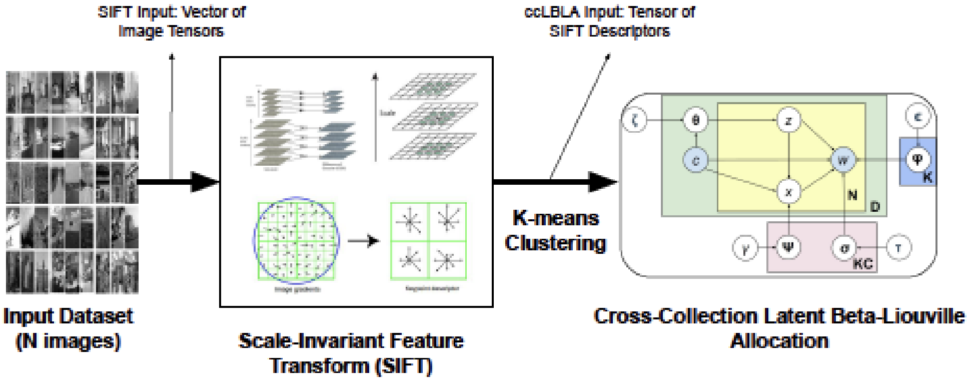

SIFT Input: Vector of Image Tensors
ccLBLA Input: Tensor of SIFT Descriptors
K-means Clustering
Input Dataset (N images)
Scale-Invariant Feature Transform (SIFT)
Cross-Collection Latent Beta-Liouville Allocation
ζ
θ
z
c
w
x
N
D
ε
φ
K
γ
ψ
σ
τ
KC

Figure 3.4: An overview of the feature extraction, clustering, and ccLBLA pipeline

$$
\begin{aligned}
class = arg\max_{c} \prod_{w} \sum_{z} P(z \mid \theta_{d_{new}}, c) \\
\times [P(w \mid z, x = 0) P(x = 0) + P(w \mid z, c, x = 1) P(x = 1)]
\end{aligned}
\tag{3.20}
$$

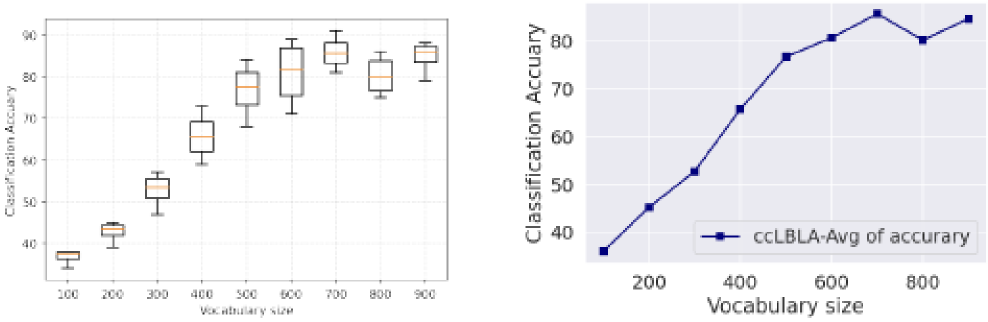
Because the cross-collection topic model can generate an image (document) likelihood which depends on the image's collection [25], cross-collection models like ccLBLA and ccLDA are capable of making collection predictions for unseen documents. Therefore, the cross-collection topic model naturally suits the classification task, and each model can predict the collection of test documents based on the visual words. Specifically, The predictive model is created by estimating the topic parameters using Eq. 3.11. The predictive topic distributions and the empirical likelihood framework lead to the estimation of the class likelihood. Based on Eq. 3.20, we can obtain the class conditionals to predict the class label of unseen images. Therefore, the collection of the unseen image is chosen with the highest class posterior distribution.

For our experiment, we use the same training and testing dataset to implement the LDA, LBLA, ccLDA, and ccLBLA models by estimating the class likelihood to predict the class label of unseen images. The highest class posterior distribution will assign the class for the unseen image.

Table 3.8: The accuracies of different tested models applied to the natural scene dataset

LDA	LBLA	ccLDA	ccLBLA
57.93%	72.67%	81.37%	90.97%

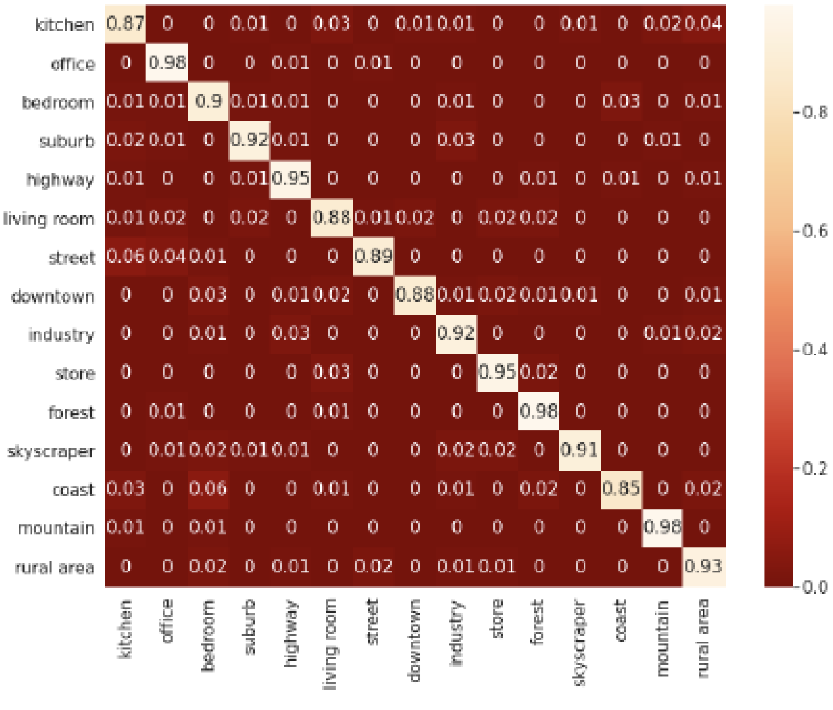
From Table 3.8, we can conclude that the ccLBLA model provides better accuracy than the other topic models. Precisely, our proposed model achieves 57% (CGS-LDA), 25% (CGS-LBLA), and 12% (CGS-ccLDA) higher accuracy. Fig.3.5 shows that the optimal vocabulary size is V=700, and we find that the optimal number of topics is K=50. The accuracy rate is 90.97, shown in the confusion matrix (Fig.3.6). These results demonstrate that the generative schemes with more flexible priors can enhance the performance of the cross-collection topic model and reinforce the concept of generalization of the ccLDA model.



(a) Accuaracy

(b) Average accuaracy

Figure 3.5: The accuracy as a function of the vocabulary size for image classification



	kitchen	office	bedroom	suburb	highway	living room	street	downtown	industry	store	forest	skyscraper	coast	mountain	rural area
kitchen	0.87	0	0	0.01	0	0.03	0	0.01	0.01	0	0	0.01	0	0.02	0.04
office	0	0.98	0	0	0.01	0	0.01	0	0	0	0	0	0	0	0
bedroom	0.01	0.01	0.9	0.01	0.01	0	0	0	0.01	0	0	0	0.03	0	0.01
suburb	0.02	0.01	0	0.92	0.01	0	0	0	0.03	0	0	0	0	0.01	0
highway	0.01	0	0	0.01	0.95	0	0	0	0	0	0.01	0	0.01	0	0.01
living room	0.01	0.02	0	0.02	0	0.88	0.01	0.02	0	0.02	0.02	0	0	0	0
street	0.06	0.04	0.01	0	0	0	0.89	0	0	0	0	0	0	0	0
downtown	0	0	0.03	0	0.01	0.02	0	0.88	0.01	0.02	0.01	0.01	0	0	0.01
industry	0	0	0.01	0	0.03	0	0	0	0.92	0	0	0	0	0.01	0.02
store	0	0	0	0	0	0.03	0	0	0	0.95	0.02	0	0	0	0
forest	0	0.01	0	0	0	0.01	0	0	0	0	0.98	0	0	0	0
skyscraper	0	0.01	0.02	0.01	0.01	0	0	0	0.02	0.02	0	0.91	0	0	0
coast	0.03	0	0.06	0	0	0.01	0	0	0.01	0	0.02	0	0.85	0	0.02
mountain	0.01	0	0.01	0	0	0	0	0	0	0	0	0	0	0.98	0
rural area	0	0	0.02	0	0.01	0	0.02	0	0.01	0.01	0	0	0	0	0.93

Figure 3.6: Confusion matrix for the natural scenes classification

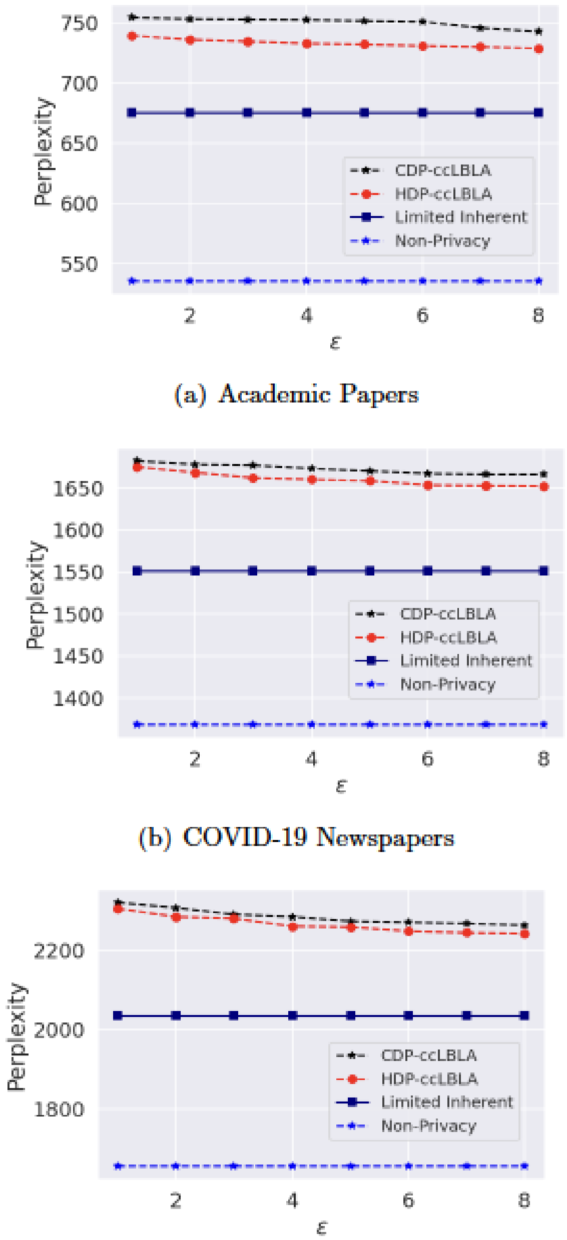
3.2.4 Performance of HDP-ccLBLA

This section details our assessment of the HDP-ccLBLA model, emphasizing its utility. We implement our method on three real-world text datasets: Covid19 newspapers, academic publications, and comments from the New York Times. The statistics for these datasets are presented in Table 3.5. Because the traveler forum dataset contains many duplicate documents which cannot accurately demonstrate the capabilities of a cross-collection topic model, we will not use this dataset in this experiment.

In this experiment, we select perplexity as the evaluation standard for topic model utility, similar to Zhao et al. [50, 80], because perplexity emphasizes the generative aspect of topic models to predict word probabilities for unseen documents in the test dataset [2, 28]. A lower perplexity indicates a higher likelihood and better model utility. To compute the perplexity and likelihood of a cross-collection topic model, we apply Eq.3.3 and Eq.3.18. To evaluate our strategy, we will compare it to CDP-ccLBLA+, which protects the training process by introducing Laplace noise into N_{dk}, N_{kw} and N_{ckw} in each iteration. In addition, we will compare the differences in topic samples between HDP-ccLBLA and Non-privacy protection ccLBLA to validate the utility of our approach.

Utility

The perplexity of HDP-ccLBLA and CDP-ccLBLA+ with different Laplace privacy ε settings is shown in Fig.3.7. In Fig.3.7, we also compare the plain CGS algorithm (Non Privacy), which lacks privacy protection. Furthermore, we employ several BL parameter configurations in ccLBLA experiments in this utility experiment. To limit the inherent privacy, we explicitly set a larger λ_w and λ_{cw}, as well as a proper clipping bound C, during the training process. Then, we set the intrinsic privacy level of HDP-ccLBLA to 10 in each iteration. Because we utilize a more significant parameter in BL distribution, the prior information can improve the model utility ability to the noise. The Imited Inherent means that the HDP-ccLBLA has the same setting for inherent privacy level but no Laplace noise for N_{kw} and N_{ckw}. From Fig.3.7, we can infer that Limited Inherent has a utility degradation compared with the plain CGS algorithm (Non Privacy) because Limited Inherent integrates a stronger inherent privacy guarantee. Even though CDP-ccLBLA+ introduces more Laplace noise and privacy loss than the HDP-ccLBLA scheme, including the intrinsic privacy loss, the utility of HDP-ccLBLA outperforms the CDP-ccLBLA+ method in that three real-world datasets based on the BL prior information.



(a) Academic Papers

(b) COVID-19 Newspapers

(c) NYT Comments

Figure 3.7: Perplexity results on three different datasets vs. Privacy level of HDP-ccLBLA

Chapter 4

Conclusion

In this thesis, to alleviate the restrictions in the traditional cross-collection topic model, we propose two cross-collection latent topic models with training acceleration and privacy protection replacing Dirichlet distribution with other more flexible prior distributions, such as generalized Dirichlet and Beta-Liouville distributions.

In Chapter 2, we proposed and implemented a novel cross-collection topic model (GPU-based ccLGDA) for multiple domain text collection to improve the original ccLDA model. Our new approach is the first GPU-based cross-collection model that utilizes the Graphics Processing Unit (GPU) to accelerate training speed significantly. The objective was to provide a powerful cross-collection topic model with more flexibility and computational efficiency to perform on various real-world, large-scale datasets. Therefore, the new approach introduces a flexible GD prior for a robust parallel inference scheme taking advantage of GPUs to show its merit in comparative text mining. The new cross-collection topic model, GPU-based ccLGDA, extends the ccLGD, GLDA, and LGDA models. These previous models suffer from the limitation of Dirichlet prior, focusing only on one individual data collection, and inefficient inference techniques, which causes a lower computational speed for large-scale applications. The GPU-based ccLGDA model provides a solution to all these shortcomings. Specifically, our new model replaced the Dirichlet distribution with the GD prior in the generative process so that our model is more flexible than the models using Dirichlet distribution. Furthermore, our new model incorporates the GPU to implement a powerful parallel inference technique that accelerates the training process on a single machine. To show the credit of our approach, we compare our result to the ccLDA model. We evaluated topic model perplexity, document classification accuracy, topic coherence, and time efficiency. Experiment results illustrate that our proposed model, GPU-based ccLGDA, outperforms ccLDA on all four quality measures on four text datasets with different domains and quantity of collections and proves the proposed method's robustness on various text datasets in other fields. In particular, the new approach overcomes the shortcoming that the collection-specific and collection-independent topic must be aligned in the ccLDA model due to the advantage of the GD in topic correlation, which produces a complete covariance structure. Indeed, our experimental studies demonstrate that the GPU-based ccLGDA model can handle such large-scale real-world datasets and provide a performance speedup of up to 18X on RTX 3070 over ccLDA and LDA on a single machine.

In chapter 3, we present and implement a novel cross-collection topic model (ccLBLA model) that utilizes the BL distribution instead of Dirichlet for various domain text collections to improve previous cross-collection topic models because the BL distribution can provide a better topic correlation representation. The ccLBLA model extends the ccLDA and LBLA models. These previous models suffer from the limitation of Dirichlet prior, or focusing only on one individual data collection. All of these issues are addressed by the ccLBLA model. In particular, our new model replaced the Dirichlet distribution with the BL prior in the generating process, making our model more flexible. We compare our experimental results to the ccLDA and LDA models to demonstrate the merits of our new technique. The perplexity of the topic model, document classification accuracy, topic coherence, and topic samples are all examined. Experimental findings show that our ccLBLA beats ccLDA and LDA models on all four quality metrics across four real-world text datasets with varying domains and number of collections. Moreover, we present the first study on applying the cross-collection topic model to image classification applications. Because of the general covariance structure in the BL distribution, the performance of the ccLBLA model in image classification demonstrates a higher classification accuracy than the ccLDA, LBLA, and LDA models. Furthermore, we investigate the privacy protection of topic models with differential privacy and propose a centralized privacy-preserving algorithm for the ccLBLA model (HDP-ccLBLA), which takes advantage of the Collapsed Gibbs Sampling inference approach's inherent differential privacy guarantee to address the privacy issue. Our HDP-ccLBLA model can prevent data inference from intermediate statistics during training. Indeed, our experimental studies demonstrate that the HDP-ccLBLA algorithm can achieve a good model utility under differential privacy.

For future work, we will continue to optimize the model parameter estimation algorithms using the variational inference. In addition, we can investigate other flexible priors to improve the performance, and propose other techniques to better separate collection-specific and collection-independent words.

Bibliography

[1] David M Blei and John D Lafferty. Topic models. In *Text mining*, pages 101–124. Chapman and Hall/CRC, 2009.

[2] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.

[3] David M. Blei. Probabilistic topic models. *Commun. ACM*, 55(4):77–84, 2012.

[4] Li Fei-Fei and Pietro Perona. A bayesian hierarchical model for learning natural scene categories. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), 20-26 June 2005, San Diego, CA, USA*, pages 524–531. IEEE Computer Society, 2005.

[5] Koffi Eddy Ihou and Nizar Bouguila. Variational-based latent generalized dirichlet allocation model in the collapsed space and applications. *Neurocomputing*, 332:372–395, 2019.

[6] Koffi Eddy Ihou and Nizar Bouguila. Stochastic topic models for large scale and nonstationary data. *Eng. Appl. Artif. Intell.*, 88, 2020.

[7] Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407, 1990.

[8] Thomas Hofmann. Probabilistic latent semantic indexing. In Fredric C. Gey, Marti A. Hearst, and Richard M. Tong, editors, *SIGIR '99: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 15-19, 1999, Berkeley, CA, USA*, pages 50–57. ACM, 1999.

[9] David M. Blei, Lawrence Carin, and David B. Dunson. Probabilistic topic models. *IEEE Signal Process. Mag.*, 27(6):55–65, 2010.

[10] Karla L. Caballero Espinosa, Joel Barajas, and Ram Akella. The generalized dirichlet distribution in enhanced topic detection. In Xue-wen Chen, Guy Lebanon, Haixun Wang, and Mohammed J. Zaki, editors, *21st ACM International Conference on Information and Knowledge Management, CIKM'12, Maui, HI, USA, October 29 - November 02, 2012*, pages 773–782. ACM, 2012.

[11] Ali Shojaee Bakhtiari and Nizar Bouguila. A variational bayes model for count data learning and classification. *Eng. Appl. Artif. Intell.*, 35:176–186, 2014.

[12] Nizar Bouguila. Clustering of count data using generalized dirichlet multinomial distributions. *IEEE Trans. Knowl. Data Eng.*, 20(4):462–474, 2008.

[13] Nizar Bouguila, Djemel Ziou, and Riad I. Hammoud. On bayesian analysis of a finite generalized dirichlet mixture via a metropolis-within-gibbs sampling. *Pattern Anal. Appl.*, 12(2):151–166, 2009.

[14] Wentao Fan and Nizar Bouguila. Variational learning of a dirichlet process of generalized dirichlet distributions for simultaneous clustering and feature selection. *Pattern Recognit.*, 46(10):2754–2769, 2013.

[15] Ali Shojaee Bakhtiari and Nizar Bouguila. Online learning for two novel latent topic models. In Linawati, Made Sudiana Mahendra, Erich J. Neuhold, A Min Tjoa, and Ilsun You, editors, *Information and Communication Technology - Second IFIP TC5/8 International Conference, ICT-EurAsia 2014, Bali, Indonesia, April 14-17, 2014. Proceedings*, volume 8407 of *Lecture Notes in Computer Science*, pages 286–295. Springer, 2014.

[16] Nuha Zamzami and Nizar Bouguila. Probabilistic modeling for frequency vectors using a flexible shifted-scaled dirichlet distribution prior. *ACM Trans. Knowl. Discov. Data*, 14(6):69:1–69:35, 2020.

[17] Wentao Fan and Nizar Bouguila. Expectation propagation learning of a dirichlet process mixture of beta-liouville distributions for proportional data clustering. *Eng. Appl. Artif. Intell.*, 43:1–14, 2015.

[18] Nizar Bouguila. On the smoothing of multinomial estimates using liouville mixture models and applications. *Pattern Anal. Appl.*, 16(3):349–363, 2013.

[19] Mian Lu, Ge Bai, Qiong Luo, Jie Tang, and Jiuxin Zhao. Accelerating topic model training on a single machine. In Yoshiharu Ishikawa, Jianzhong Li, Wei Wang, Rui Zhang, and Wenjie Zhang, editors, *Web Technologies and Applications - 15th Asia-Pacific Web Conference, APWeb 2013, Sydney, Australia, April 4-6, 2013. Proceedings*, volume 7808 of *Lecture Notes in Computer Science*, pages 184–195. Springer, 2013.

[20] David Newman, Arthur U. Asuncion, Padhraic Smyth, and Max Welling. Distributed inference for latent dirichlet allocation. In John C. Platt, Daphne Koller, Yoram Singer, and Sam T. Roweis, editors, *Advances in Neural Information Processing Systems 20, Proceedings of the Twenty-First Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 3-6, 2007*, pages 1081–1088. Curran Associates, Inc., 2007.

[21] Matthew Fredrikson, Eric Lantz, Somesh Jha, Simon M. Lin, David Page, and Thomas Ristenpart. Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing. In Kevin Fu and Jaeyeon Jung, editors, *Proceedings of the 23rd USENIX Security Symposium, San Diego, CA, USA, August 20-22, 2014*, pages 17–32, San Diego, CA, USA, 2014. USENIX Association.

[22] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017*, pages 3–18. IEEE Computer Society, 2017.

[23] Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In Serge Vaudenay, editor, *Advances in Cryptology - EUROCRYPT 2006, 25th Annual International Conference on the Theory and Applications of Cryptographic Techniques, St. Petersburg, Russia, May 28 - June 1, 2006, Proceedings*, volume 4004 of *Lecture Notes in Computer Science*, pages 486–503, St. Petersburg, Russia, 2006. Springer.

[24] ChengXiang Zhai, Atulya Velivelli, and Bei Yu. A cross-collection mixture model for comparative text mining. In Won Kim, Ron Kohavi, Johannes Gehrke, and William DuMouchel, editors, *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, Washington, USA, August 22-25, 2004*, pages 743–748. ACM, 2004.

[25] Michael J. Paul and Roxana Girju. Cross-cultural analysis of blogs and forums with mixed-collection topic models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, EMNLP 2009, 6-7 August 2009, Singapore, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1408–1417. ACL, 2009.

[26] Thomas L Griffiths, Mark Steyvers, and Joshua B Tenenbaum. Topics in semantic representation. *Psychological review*, 114(2):211, 2007.

[27] Xuerui Wang, Andrew McCallum, and Xing Wei. Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In *Proceedings of the 7th IEEE International Conference on Data Mining (ICDM 2007), October 28-31, 2007, Omaha, Nebraska, USA*, pages 697–702. IEEE Computer Society, 2007.

[28] Julian Risch and Ralf Krestel. My approach = your apparatus? In Jiangping Chen, Marcos André Gonçalves, Jeff M. Allen, Edward A. Fox, Min-Yen Kan, and Vivien Petras, editors, *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries, JCDL 2018, Fort Worth, TX, USA, June 03-07, 2018*, pages 283–292. ACM, 2018.

[29] Thomas L Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National academy of Sciences*, 101(suppl 1):5228–5235, 2004.

[30] Ian Porteous, David Newman, Alexander T. Ihler, Arthur U. Asuncion, Padhraic Smyth, and Max Welling. Fast collapsed gibbs sampling for latent dirichlet allocation. In Ying Li, Bing Liu, and Sunita Sarawagi, editors, *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, Nevada, USA, August 24-27, 2008*, pages 569–577. ACM, 2008.

[31] Limin Yao, David M. Mimno, and Andrew McCallum. Efficient methods for topic model inference on streaming document collections. In John F. Elder IV, Françoise Fogelman-Soulié, Peter A. Flach, and Mohammed Javeed Zaki, editors, *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris, France, June 28 - July 1, 2009*, pages 937–946. ACM, 2009.

[32] Thomas P. Minka. Expectation propagation for approximate bayesian inference. In Jack S. Breese and Daphne Koller, editors, *UAI '01: Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence, University of Washington, Seattle, Washington, USA, August 2-5, 2001*, pages 362–369. Morgan Kaufmann, 2001.

[33] Nizar Bouguila. Count data modeling and classification using finite mixtures of distributions. *IEEE Trans. Neural Networks*, 22(2):186–198, 2011.

[34] Koffi Eddy Ihou and Nizar Bouguila. A new latent generalized dirichlet allocation model for image classification. In *Seventh International Conference on Image Processing Theory, Tools and Applications, IPTA 2017, Montreal, QC, Canada, November 28 - December 1, 2017*, pages 1–6. IEEE, 2017.

[35] Koffi Eddy Ihou and Nizar Bouguila. A smoothed latent generalized dirichlet allocation model in the collapsed space. In *IEEE 61st International Midwest Symposium on Circuits and Systems, MWSCAS 2018, Windsor, ON, Canada, August 5-8, 2018*, pages 877–880. IEEE, 2018.

[36] Zhuolin Qiu, Bin Wu, Bai Wang, and Le Yu. Gibbs collapsed sampling for latent dirichlet allocation on spark. In Wei Fan, Albert Bifet, Qiang Yang, and Philip S. Yu, editors, *Proceedings of the 3rd International Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications, BigMine 2014, New York City, USA, August 24, 2014*, volume 36 of *JMLR Workshop and Conference Proceedings*, pages 17–28. JMLR.org, 2014.

[37] WenYen Chen, Jon-Chyuan Chu, Junyi Luan, Hongjie Bai, Yi Wang, and Edward Y. Chang. Collaborative filtering for orkut communities: discovery of user latent behavior. In Juan Quemada, Gonzalo León, Yoëlle S. Maarek, and

Wolfgang Nejdl, editors, *Proceedings of the 18th International Conference on World Wide Web, WWW 2009, Madrid, Spain, April 20-24, 2009*, pages 681–690. ACM, 2009.

[38] Arthur U. Asuncion, Padhraic Smyth, and Max Welling. Asynchronous distributed learning of topic models. In Daphne Koller, Dale Schuurmans, Yoshua Bengio, and Léon Bottou, editors, *Advances in Neural Information Processing Systems 21, Proceedings of the Twenty-Second Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 8-11, 2008*, pages 81–88. Curran Associates, Inc., 2008.

[39] Feng Yan, Ningyi Xu, and Yuan (Alan) Qi. Parallel inference for latent dirichlet allocation on graphics processing units. In Yoshua Bengio, Dale Schuurmans, John D. Lafferty, Christopher K. I. Williams, and Aron Culotta, editors, *Advances in Neural Information Processing Systems 22: 23rd Annual Conference on Neural Information Processing Systems 2009. Proceedings of a meeting held 7-10 December 2009, Vancouver, British Columbia, Canada*, pages 2134–2142. Curran Associates, Inc., 2009.

[40] Kamalika Chaudhuri, Anand D. Sarwate, and Kaushik Sinha. Near-optimal differentially private principal components. In Peter L. Bartlett, Fernando C. N. Pereira, Christopher J. C. Burges, Léon Bottou, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States*, pages 998–1006, 2012.

[41] Chugui Xu, Ju Ren, Deyu Zhang, Yaoxue Zhang, Zhan Qin, and Kui Ren. Ganobfuscator: Mitigating information leakage under GAN via differential privacy. *IEEE Trans. Inf. Forensics Secur.*, 14(9):2358–2371, 2019.

[42] Zonghao Huang, Rui Hu, Yuanxiong Guo, Eric Chan-Tin, and Yanmin Gong. DP-ADMM: admm-based distributed learning with differential privacy. *IEEE Trans. Inf. Forensics Secur.*, 15:1002–1012, 2020.

[43] Raef Bassily, Adam D. Smith, and Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *55th IEEE Annual*

Symposium on Foundations of Computer Science, FOCS 2014, Philadelphia, PA, USA, October 18-21, 2014, pages 464–473, Philadelphia, PA, USA, 2014. IEEE Computer Society.

[44] Reza Shokri and Vitaly Shmatikov. Privacy-preserving deep learning. In *53rd Annual Allerton Conference on Communication, Control, and Computing, Allerton 2015, Allerton Park & Retreat Center, Monticello, IL, USA, September 29 - October 2, 2015*, pages 909–910, Monticello, IL, USA, 2015. IEEE.

[45] Mijung Park, James R. Foulds, Kamalika Chaudhuri, and Max Welling. Variational bayes in private settings (VIPS). *J. Artif. Intell. Res.*, 68:109–157, 2020.

[46] Meng Sun and Wee Peng Tay. On the relationship between inference and data privacy in decentralized iot networks. *IEEE Trans. Inf. Forensics Secur.*, 15:852–866, 2020.

[47] Tianqing Zhu, Gang Li, Wanlei Zhou, Ping Xiong, and Cao Yuan. Privacy-preserving topic model for tagging recommender systems. *Knowl. Inf. Syst.*, 46(1):33–58, 2016.

[48] Chris Decarolis, Mukul Ram, Seyed Esmaeili, Yu-Xiang Wang, and Furong Huang. An end-to-end differentially private latent dirichlet allocation using a spectral algorithm. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 2421–2431, Online, 2020. PMLR.

[49] Yu-Xiang Wang, Stephen E. Fienberg, and Alexander J. Smola. Privacy for free: Posterior sampling and stochastic gradient monte carlo. In Francis R. Bach and David M. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 2493–2502, Lille, France, 2015. JMLR.org.

[50] Fangyuan Zhao, Xuebin Ren, Shusen Yang, Qing Han, Peng Zhao, and Xinyu Yang. Latent dirichlet allocation model training with differential privacy. *IEEE Trans. Inf. Forensics Secur.*, 16:1290–1305, 2021.

[51] Christos Dimitrakakis, Blaine Nelson, Aikaterini Mitrokotsa, and Benjamin I. P. Rubinstein. Robust and private bayesian inference. In Peter Auer, Alexander Clark, Thomas Zeugmann, and Sandra Zilles, editors, *Algorithmic Learning Theory - 25th International Conference, ALT 2014, Bled, Slovenia, October 8-10, 2014. Proceedings*, volume 8776 of *Lecture Notes in Computer Science*, pages 291–305, Bled, Slovenia, 2014. Springer.

[52] James R. Foulds, Joseph Geumlek, Max Welling, and Kamalika Chaudhuri. On the theory and practice of privacy-preserving bayesian data analysis. *CoRR*, 2016.

[53] Ali Shojaee Bakhtiari and Nizar Bouguila. A latent beta-liouville allocation model. *Expert Syst. Appl.*, 45:260–272, 2016.

[54] Elise Epaillard and Nizar Bouguila. Proportional data modeling with hidden markov models based on generalized dirichlet and beta-liouville mixtures applied to anomaly detection in public areas. *Pattern Recognit.*, 55:125–136, 2016.

[55] Nizar Bouguila. Infinite liouville mixture models with application to text and texture categorization. *Pattern Recognit. Lett.*, 33(2):103–110, 2012.

[56] Wentao Fan and Nizar Bouguila. Topic novelty detection using infinite variational inverted dirichlet mixture models. In *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*, pages 70–75, 2015.

[57] David M. Blei and John D. Lafferty. Correlated topic models. In *Advances in Neural Information Processing Systems 18 [Neural Information Processing Systems, NIPS 2005, December 5-8, 2005, Vancouver, British Columbia, Canada]*, pages 147–154, 2005.

[58] Duangmanee Putthividhya, Hagai Thomas Attias, and Srikantan S. Nagarajan. Independent factor topic models. In Andrea Pohoreckyj Danyluk, Léon Bottou, and Michael L. Littman, editors, *Proceedings of the 26th Annual International Conference on Machine Learning, ICML 2009, Montreal, Quebec, Canada, June 14-18, 2009*, volume 382 of *ACM International Conference Proceeding Series*, pages 833–840. ACM, 2009.

[59] Nizar Bouguila and Djemel Ziou. A dirichlet process mixture of generalized dirichlet distributions for proportional data modeling. *IEEE Transactions on Neural Networks*, 21(1):107–122, 2010.

[60] Nizar Bouguila and Walid ElGuebaly. Discrete data clustering using finite mixture models. *Pattern Recognit.*, 42(1):33–42, 2009.

[61] Wentao Fan, Hassen Sallay, and Nizar Bouguila. Online learning of hierarchical pitman–yor process mixture of generalized dirichlet distributions with feature selection. *IEEE Transactions on Neural Networks and Learning Systems*, 28(9):2048–2061, 2017.

[62] Fatma Najar and Nizar Bouguila. Smoothed generalized dirichlet: a novel count data model for detecting emotional states. *IEEE Transactions on Artificial Intelligence*, pages 1–1, 2021.

[63] Nizar Bouguila and Walid ElGuebaly. A generative model for spatial color image databases categorization. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 821–824, 2008.

[64] Nizar Bouguila. Hybrid generative/discriminative approaches for proportional data modeling and classification. *IEEE Transactions on Knowledge and Data Engineering*, 24(12):2184–2202, 2012.

[65] Nizar Bouguila and Walid ElGuebaly. On discrete data clustering. In Takashi Washio, Einoshin Suzuki, Kai Ming Ting, and Akihiro Inokuchi, editors, *Advances in Knowledge Discovery and Data Mining, 12th Pacific-Asia Conference, PAKDD 2008, Osaka, Japan, May 20-23, 2008 Proceedings*, volume 5012 of *Lecture Notes in Computer Science*, pages 503–510. Springer, 2008.

[66] Nuha Zamzami and Nizar Bouguila. A novel scaled dirichlet-based statistical framework for count data modeling: Unsupervised learning and exponential approximation. *Pattern Recognit.*, 95:36–47, 2019.

[67] Nuha Zamzami and Nizar Bouguila. Text modeling using multinomial scaled dirichlet distributions. In Malek Mouhoub, Samira Sadaoui, Otmane Aït Mohamed, and Moonis Ali, editors, *Recent Trends and Future Technology in Applied Intelligence - 31st International Conference on Industrial Engineering and*

Other Applications of Applied Intelligent Systems, IEA/AIE 2018, Montreal, QC, Canada, June 25-28, 2018, Proceedings, volume 10868 of *Lecture Notes in Computer Science*, pages 69–80. Springer, 2018.

[68] Nuha Zamzami and Nizar Bouguila. Model selection and application to high-dimensional count data clustering - via finite EDCM mixture models. *Appl. Intell.*, 49(4):1467–1488, 2019.

[69] N. Bouguila and D. Ziou. Improving content based image retrieval systems using finite multinomial dirichlet mixture. In *Proceedings of the 2004 14th IEEE Signal Processing Society Workshop Machine Learning for Signal Processing, 2004.*, pages 23–32, 2004.

[70] Nizar Bouguila. Spatial color image databases summarization. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, volume 1, pages I–953–I–956, 2007.

[71] N. Bouguila and D. Ziou. Mml-based approach for high-dimensional unsupervised learning using the generalized dirichlet mixture. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Workshops*, pages 53–53, 2005.

[72] Nizar Bouguila and Mukti Nath Ghimire. Discrete visual features modeling via leave-one-out likelihood estimation and applications. *J. Vis. Commun. Image Represent.*, 21(7):613–626, 2010.

[73] Yee Whye Teh, David Newman, and Max Welling. A collapsed variational bayesian inference algorithm for latent dirichlet allocation. In Bernhard Schölkopf, John C. Platt, and Thomas Hofmann, editors, *Advances in Neural Information Processing Systems 19, Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 4-7, 2006*, pages 1353–1360. MIT Press, 2006.

[74] Nafiz Sadman, Nishat Anjum, and Kishor Datta Gupta. Introduction of covid-news-us-nnk and covid-news-bd-nnk dataset. 2020.

[75] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.", 2009.

[76] Michael Röder, Andreas Both, and Alexander Hinneburg. Exploring the space of topic coherence measures. In Xueqi Cheng, Hang Li, Evgeniy Gabrilovich, and Jie Tang, editors, *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, WSDM 2015, Shanghai, China, February 2-6, 2015*, pages 399–408. ACM, 2015.

[77] Wentao Fan and Nizar Bouguila. Learning finite beta-liouville mixture models via variational bayes for proportional data clustering. In Francesca Rossi, editor, *IJCAI 2013, Proceedings of the 23rd International Joint Conference on Artificial Intelligence, Beijing, China, August 3-9, 2013*, pages 1323–1329. IJCAI/AAAI, 2013.

[78] Nuha Zamzami and Nizar Bouguila. High-dimensional count data clustering based on an exponential approximation to the multinomial beta-liouville distribution. *Inf. Sci.*, 524:116–135, 2020.

[79] Ali Shojaee Bakhtiari and Nizar Bouguila. A novel hierarchical statistical model for count data modeling and its application in image classification. In Tingwen Huang, Zhigang Zeng, Chuandong Li, and Chi-Sing Leung, editors, *Neural Information Processing - 19th International Conference, ICONIP 2012, Doha, Qatar, November 12-15, 2012, Proceedings, Part II*, volume 7664 of *Lecture Notes in Computer Science*, pages 332–340. Springer, 2012.

[80] Fangyuan Zhao, Xuebin Ren, Shusen Yang, Qing Han, Peng Zhao, and Xinyu Yang. Latent dirichlet allocation model training with differential privacy. *CoRR*, abs/2010.04391, 2020.

[81] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006), 17-22 June 2006, New York, NY, USA*, pages 2169–2178, New York, NY, USA, 2006. IEEE Computer Society.