Fast and Modular Regularized Topic Modelling

Denis Kochedykov
J.P.Morgan
New York, USA
kochedykov@gmail.com

Murat Apishev
Moscow State University
Moscow, Russia
great-mel@yandex.ru

Lev Golitsyn
Integrated Systems
Moscow, Russia
lvgolitsyn@gmail.ru

Konstantin Vorontsov
MIPT
Moscow, Russia
vokov@forecsys.ru

***Abstract*—Topic modelling is an area of text mining that has been actively developed in the last 15 years. A probabilistic topic model extracts a set of hidden topics from a collection of text documents. It defines each topic by a probability distribution over words and describes each document with a probability distribution over topics. In applications, there are often many requirements, such as, for example, problem-specific knowledge and additional data, to be taken into account. Therefore, it is natural for topic modelling to be considered a multi-objective optimization problem. However, historically, Bayesian learning became the most popular approach for topic modelling. In the Bayesian paradigm, all requirements are formalized in terms of a probabilistic generative process. This approach is not always convenient due to some limitations and technical difficulties. In this work, we develop a non-Bayesian multi-objective approach called the Additive Regularization of Topic Models (ARTM). It is based on regularized Maximum Likelihood Estimation (MLE), and we show that many of the well-known Bayesian topic models can be re-formulated in a much simpler way using the regularization point of view. We review some of the most important types of topic models: multimodal, multilingual, temporal, hierarchical, graph-based, and short-text. The ARTM framework enables easy combination of different types of models to create new models with the desired properties for applications. This modular "lego-style" technology for topic modelling is implemented in the open-source library BigARTM.**

I. Introduction

Understanding the thematic structure of text collections is important in many applications of natural language processing and information retrieval, including searches for similar documents, navigating large text collections, and the classification, categorization and segmentation of documents. Topic modelling is an area of text mining that has been actively developed since the late 1990s. A *probabilistic topic model* extracts the hidden topic structure of a collection representing each topic by a probability distribution over words and describing each document with a probabilistic mixture of topics.

Historically, the first such model was Probabilistic Latent Semantic Analysis (PLSA), introduced by T. Hofmann in 1999 [1]. In 2003, D. Blei, A. Ng and M. Jordan proposed its Bayesian extension named the Latent Dirichlet Allocation (LDA). Since then, topic modelling has been mainly developed within the framework of graphical models and Bayesian learning. Over the past years, hundreds of extensions of LDA and PLSA have emerged which take into account the various problem-specific features of data and desired properties of the solution. A series of examples follows. When analysing topical trends in news feeds, patent databases, and academic archives, it may be quite informative to take into consideration authors, dates and sources of texts [2]. In an exploratory information search, it is important to ensure interpretability of topics and to be able to organize topics into hierarchies [3]. In multi- and cross-lingual information retrieval, parallel texts or external dictionaries are used to learn a multi-language topic model [4]. When analysing social media data, it is important to take into account the network structure, time stamps, authors, and geographical locations [5]. When topic modelling is used to classify documents, it is important that the class labels in training data be accounted for [6]. When topic model is used in recommender systems, the text description of users and items should be analysed together with user behaviour data [7]. There are some overviews of topic models and their applications in [8], [9].

There are two motivations for considering topic modelling as a multi-objective optimization problem. The first is practical — there are usually considerable requirements, problem-specific knowledge, and additional data to be taken into account [10]. For example, the topic model for an exploratory search should be simultaneously well-interpretable, hierarchical, temporal, and multimodal, while taking into account authors, categories, tags, and citations [11]. There are models that address each of these requirements, but combining them into a single model is a challenging, open problem when using the widely adopted Bayesian framework. The second motivation is theoretical. Learning hidden topics from data is an ill-posed optimization problem, which, in general, has infinitely many solutions. The standard way to address this issue is to regularize the optimization problem and make the solution more stable by adding a *regularizer* to the main optimization objective [12]. A regularizer is an additional criterion that formalizes problem-specific requirements and consequently penalizes or favours certain solutions. The above example of exploratory search shows that there may be many regularization criteria in applications.

Bayesian learning is the de facto standard in topic modelling. In this approach, one first describes the probabilistic generative model for the data, specifies prior distributions of the model parameters, and then uses Bayesian inference to obtain the posterior distributions of the parameters. The generative probabilistic process encapsulates many different types of domain-specific knowledge and requirements that come from the application. Therefore, the Bayesian inference of the posterior distributions is a difficult problem which requires unique derivations and coding for each application. There is no unified Bayesian solution that allows new functional blocks or modules to be to "plugged" into the model. Practitioners often prefer to use the simplest LDA topic model, ignoring more advanced, but impractical, solutions. At the same time, there are no strong reasons for inferring the posterior distributions

ISSN 2305-7254

in topic modelling. Researchers often go through the many technical difficulties inherent in Bayesian inference simply to obtain the point estimates of the model parameters. Apparently, Bayesian inference solves a more difficult problem than it is necessary. Hence, some limitations follow in imposing non-probabilistic constraints, specifying multiple criteria for the model, and combining models in a "lego-style" technology.

In this work, we develop a non-Bayesian multi-objective approach called Additive Regularization of Topic Models (ARTM) [13], [14]. It is based on the maximization of the log-likelihood together with a weighted sum of regularization criteria. In multi-objective optimization, this approach is known as *scalarization*. Log-likelihood describes a simple generative model usually equivalent to a matrix factorization, whereas each regularizer introduces one of the additional requirements into the model. That is, in ARTM, the difficulty of the problem is transferred from the generative process to the set of additive criteria, which can be treated separately. Many of the well-known topic models can be re-formulated in terms of regularization, and we observe that this formulation is usually much simpler. Moreover this approach allows combining topic models simply by summing up their regularizers. This gave rise to the modular technology for topic modelling implemented in the BigARTM project [15] — an open-source community-driven library for topic modelilng, available at `http://bigartm.org`.

The goal of this work is to demonstrate how a variety of topic models can be formulated using the regularization framework. The focus will be on the first and most important stage of the modelling process — formalizing the problem-specific requirements in terms of regularization criteria. The consequent steps are, in fact, almost completely automated in the ARTM framework.

II. Basics of topic modelling

Let us denote a finite set (collection) of texts by D, and a finite set (vocabulary) of terms from these texts by W. Each term can be a single word or a key phrase. Each document $d \in D$ is a sequence of n_d terms $(w_1, \ldots, w_{n_d})$ from W. Each term might appear multiple times in the same document.

Assume that each term occurrence in each document is associated with some latent topic from a finite set of topics T. Text collection is considered to be a sample of triples (w_i, d_i, t_i), $i = 1, \ldots, n$, drawn independently from a discrete distribution $p(w, d, t)$ over a finite space $W \times D \times T$. Term w and document d are observable random variables, while topic t is a *latent* (hidden) random variable. Following the "bag of words" assumption, each document is represented by a subset of terms $d \subset W$ with integers n_{dw} that count how many times the term w appears in the document d.

Conditional independence is the assumption that each topic generates terms regardless of the document: $p(w\,|\,t) = p(w\,|\,d, t)$. According to the law of total probability and the assumption of conditional independence,

$$p(w\,|\,d) = \sum_{t \in T} p(t\,|\,d)\, p(w\,|\,t). \tag{1}$$

The probabilistic model (1) describes how the collection D is generated from the known distributions $p(t\,|\,d)$ and $p(w\,|\,t)$. Learning a topic model is an inverse problem, i.e., the distributions $p(t\,|\,d)$ and $p(w\,|\,t)$ must be found, given the collection. This problem is equivalent to finding an approximate representation of the matrix of counts $F = \big(\hat{p}(w\,|\,d)\big)_{W \times D}$, $\hat{p}(w\,|\,d) = \frac{n_{dw}}{n_d}$, as a product $F \approx \Phi\Theta$ of two unknown matrices — the matrix Φ of *term probabilities for the topics* and the matrix Θ of *topic probabilities for the documents*:

$$\Phi = (\varphi_{wt})_{W \times T}, \qquad \varphi_{wt} = p(w\,|\,t);$$
$$\Theta = (\theta_{td})_{T \times D}, \qquad \theta_{td} = p(t\,|\,d).$$

Matrices F, Φ, and Θ are *probability matrices*, i.e., they have non-negative and normalized columns f_d, φ_t, and θ_d, respectively, representing discrete distributions. Usually the number of topics $|T|$ is much smaller than the collection size $|D|$ and the vocabulary size $|W|$. Thus, the problem is one of low-rank non-negative matrix factorization.

In *probabilistic latent semantic analysis*, PLSA [1], the topic model (1) is learned by log-likelihood maximization with linear constraints. The *likelihood* is the probability of the observed data as a function of model parameters Φ and Θ. Due to the independence assumption, it is equivalent to the product of the probabilities of words in the documents:

$$\prod_{i=1}^{n} p(d_i, w_i) = \prod_{d \in D} \prod_{w \in d} p(w\,|\,d)^{n_{dw}} p(d)^{n_{dw}} \;\to \max_{\Phi,\Theta}.$$

Taking the logarithm, the above becomes a sum and the terms that don't depend on the model parameter can be dropped because they don't affect optimization. We have a log-likelihood maximization subject to the linear constraints of non-negativity and normalization:

$$L(\Phi, \Theta) = \sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \varphi_{wt} \theta_{td} \;\to\; \max_{\Phi,\Theta}; \tag{2}$$

$$\sum_{w \in W} \varphi_{wt} = 1;\; \varphi_{wt} \geqslant 0; \quad \sum_{t \in T} \theta_{td} = 1;\; \theta_{td} \geqslant 0. \tag{3}$$

Note that (2) is equivalent to searching for an approximate matrix factorization $F \approx \Phi\Theta$ best in the sense of minimizing a weighted sum of the KL-divergences:

$$L(\Phi, \Theta) = \sum_{d \in D} n_d \,\mathrm{KL}\big(\hat{p}(w\,|\,d) \;\big\|\; p(w\,|\,d)\big) \;\to\; \min_{\Phi,\Theta}.$$

In Bayesian topic modelling, parameters (Φ, Θ) are assumed to be drawn from a *prior* distribution $p(\Phi, \Theta\,|\,\gamma)$ with the *hyperparameter* γ. In this case, likelihood maximization leads to the *maximum a posteriori probability* (MAP) estimate:

$$p(\Phi, \Theta\,|\,\gamma) \prod_{i=1}^{n} p(d_i, w_i\,|\,\Phi, \Theta) \;\to\; \max_{\Phi,\Theta,\gamma}.$$

Taking the logarithm, we have an extension of (2) with the log-prior playing the role of a regularizer:

$$L(\Phi, \Theta) + \ln p(\Phi, \Theta; \gamma) \;\to\; \max_{\Phi,\Theta,\gamma}. \tag{4}$$

In Bayesian topic modelling, marginal likelihood maximization is used instead of MAP. The model parameters (Φ, Θ) are first integrated out, then the log-likelihood is optimized over the hyperparameters γ. This is said to reduce both the

dimensionality of the parameter space and the risk of overfitting. Indeed, the dimensionality of γ is usually dramatically smaller than the size of matrices Φ, Θ, and it also does not depend on the size of the collection.

Bayesian inference results in the posterior $p(\Phi, \Theta \,|\, D; \gamma)$ instead of the matrices Φ, Θ themselves, although the point estimates can also be derived. Different inference techniques, give slightly different Φ, Θ estimates for the LDA model, but they are close to the straightforward MAP estimates [16]. In applications, neither posterior, interval estimates, nor median/mode estimates are used. Hence, finding posterior distributions appears to be excessive, in practice.

III. Additive regularization for topic modelling

Non-negative matrix factorization is an ill-posed problem. If $\Phi\Theta$ is a solution, then $(\Phi S)(S^{-1}\Theta)$ is another solution for an invertible matrix S such that ΦS and $S^{-1}\Theta$ are probability matrices. The standard approach for addressing ill-posed problems is to add a *regularization* criterion to the main objective [12]. Usually regularizers formalize the domain knowledge and penalize or favour certain solutions. In topic modelling, there are often many requirements which could be expressed by regularizers.

Additive regularization of topic models (ARTM) [13] is based on maximizing the log-likelihood and a weighted sum of regularizers $R_i(\Phi, \Theta)$, $i = 1, \ldots, k$:

$$\sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \varphi_{wt} \theta_{td} + \sum_{i=1}^{k} \tau_i R_i(\Phi, \Theta) \;\to\; \max_{\Phi, \Theta}; \quad (5)$$

subject to constraints (3), where the τ_i are non-negative *regularization coefficients*. The optimization problem (5), (3) is non-convex so it is only feasible to find a local maximum.

Consider a norm operator that normalizes a vector to make it a vector of probabilities:

$$p_i = \operatorname*{norm}_{i \in I}(x_i) = \frac{(x_i)_+}{\sum_{j \in I}(x_j)_+}, \text{ for all } i \in I,$$

where $(x)_+ = \max\{0, x\}$ is a truncation of negative values. If $x_i \leqslant 0$ for all $i \in I$, then $\operatorname{norm}(x)$ is the zero vector.

Theorem 1: Let regularizer $R(\Phi, \Theta)$ be differentiable. Then the local extreme (Φ, Θ) of the optimization problem (5), (3) satisfies the following system of equations with auxiliary variables $p_{tdw} = p(t \,|\, d, w)$, n_{wt}, and n_{td}:

$$p_{tdw} = \operatorname*{norm}_{t \in T}\big(\varphi_{wt}\theta_{td}\big); \quad (6)$$

$$\varphi_{wt} = \operatorname*{norm}_{w \in W}\left(n_{wt} + \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}}\right); \; n_{wt} = \sum_{d \in D} n_{dw} p_{tdw}; \quad (7)$$

$$\theta_{td} = \operatorname*{norm}_{t \in T}\left(n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right); \; n_{td} = \sum_{w \in d} n_{dw} p_{tdw}; \quad (8)$$

for all topics t and documents d that are non-degenerate in the following sense:

a) t is *degenerate*, if $n_{wt} + \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}} \leqslant 0$ for all $w \in W$;

b) d is *degenerate*, if $n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}} \leqslant 0$ for all $t \in T$.

The above degeneracy may occur when regularizer R has too strong of a sparsing effect on the model. Degenerate topics and documents can be excluded from the model. Excluding weak topics is a favourable effect of the regularization, and the degeneracy of a document may mean that it is too short or quite atypical, i.e., the model cannot describe it.

In the Expectation Maximization (EM-) algorithm, we solve equations (6)–(8) with a fixed-point iteration method, turning these equations into updates. We assign initial values to φ_{wt} and θ_{td}, and apply *E-step* (6) and *M-step* (7)–(8) as updates in a loop until convergence [17].

Note that PLSA corresponds to the ARTM case in which regularization is absent, i.e., $R(\Phi, \Theta) = 0$.

IV. Non-Bayesian generalization of LDA

The LDA model was proposed in [18] to address PLSA model overfitting. PLSA predicted word probabilities $p(w \,|\, d)$ in new documents significantly worse than in training documents. Later, it became clear that, when training with big data, both PLSA and LDA do not overfit, and the attained likelihoods don't differ by much [19], [20], [21]. Another way to achieve similar performance is to consider robust versions of the models. Some differences between PLSA and LDA manifest only for rare terms that are normally not important for inferring and interpreting the topics. In robust variants of these models such terms are ignored, and this significantly reduces both seeming overfitting and differences between PLSA and LDA models [22].

Moreover, the quality of word prediction may not be the best way to judge topic model performance. First, topic models normally are fitted not for word prediction, but for discovering semantic structures in a text collection. Second, when measuring model quality, perplexity is often used, which is known to strongly penalize underestimation of small probabilities. All of this indicates that the difference between PLSA and LDA is not as important as it was considered previously. However, LDA was widely adopted as the better alternative to PLSA.

In the LDA model, it is assumed that columns θ_d and φ_t are random vectors drawn from Dirichlet distributions with nonnegative parameters $\alpha \in \mathbb{R}^{|T|}$ and $\beta \in \mathbb{R}^{|W|}$, respectively.

According to (4), LDA corresponds to the regularizer that is equal to the logarithm of the Dirichlet prior:

$$\begin{aligned} R(\Phi, \Theta) &= \ln \prod_{t \in T} \operatorname{Dir}(\varphi_t; \beta) \prod_{d \in D} \operatorname{Dir}(\theta_d; \alpha) + \text{const} \\ &= \sum_{t,w} (\beta_w - 1) \ln \varphi_{wt} + \sum_{d,t} (\alpha_t - 1) \ln \theta_{td}. \end{aligned} \quad (9)$$

Substituting R in (7)–(8) gives us the M-step:

$$\varphi_{wt} = \operatorname*{norm}_{w \in W}\big(n_{wt} + \beta_w - 1\big); \quad (10)$$

$$\theta_{td} = \operatorname*{norm}_{t \in T}\big(n_{td} + \alpha_t - 1\big). \quad (11)$$

When $\beta_w = 1$, $\alpha_t = 1$, the Dirichlet distribution is uniform and LDA coincides with PLSA [23]. When $\beta_w > 1$, $\alpha_t > 1$, the regularizer has a smoothing effect: it makes small probabilities φ_{wt} and θ_{td} larger and brings these distributions

closer to the uniform. When $0 < \beta_w < 1$, $0 < \alpha_t < 1$, the regularizer has a sparsing effect: it makes small probabilities smaller and eventually drives some of them to zero due to negative values truncation in the norm operator.

The regularizer (9) can be equivalently represented using KL-divergences instead of log-priors:

$$R(\Phi,\Theta) = |W| \sum_{t\in T} \mathrm{KL}\big(\tfrac{1}{|W|} \,\big\|\, \varphi_{wt}\big) - \beta_0 \sum_{t\in T} \mathrm{KL}\big(\tfrac{\beta_w}{\beta_0} \,\big\|\, \varphi_{wt}\big) + |T| \sum_{d\in D} \mathrm{KL}\big(\tfrac{1}{|T|} \,\big\|\, \theta_{td}\big) - \alpha_0 \sum_{d\in D} \mathrm{KL}\big(\tfrac{\alpha_t}{\alpha_0} \,\big\|\, \theta_{td}\big).$$

This gives a non-Bayesian interpretation of LDA and the effects of regularization: columns φ_t are pushed towards distribution $\frac{\beta_w}{\beta_0}$ controlled by the coefficient β_0; columns θ_d are pushed towards $\frac{\alpha_t}{\alpha_0}$ controlled by α_0; and a weak, uncontrolled sparsing pushes all distributions away from uniform.

A. Unifying sparsing and smoothing

Dropping the restrictions that come from Bayesian inference and Dirichlet priors, we have the freedom to use negative hyperparameters in (9), as well as to mix smoothing and sparsing effects to improve topics.

Following (9), let us introduce a generalized cross-entropy regularizer for smoothing and sparsing:

$$R(\Phi,\Theta) = \sum_{t\in T}\sum_{w\in W} \beta_{wt} \ln \varphi_{wt} + \sum_{d\in D}\sum_{t\in T} \alpha_{td} \ln \theta_{td}.$$

Substituting R in (7)–(8) gives us the M-step:

$$\varphi_{wt} = \operatorname*{norm}_{w\in W}\big(n_{wt} + \beta_{wt}\big); \quad \theta_{td} = \operatorname*{norm}_{t\in T}\big(n_{td} + \alpha_{td}\big).$$

Positive α_{td} and β_{wt} correspond to smoothing distributions, negative values correspond to sparsing.

B. Semi-supervised topic learning

During the evaluation or application of a topic model, experts, assessors, or users may label some words and documents as relevant or irrelevant for some topics. This leads to *semi-supervised* topic learning with expert advice, which can be realized by the smoothing and sparsing regularizer:

$$\beta_{wt} = \beta_+[w \in W_t^+] - \beta_-[w \in W_t^-],$$
$$\alpha_{td} = \alpha_+[d \in D_t^+] - \alpha_-[d \in D_t^-],$$

where W_t^+ and D_t^+ are "white lists" of relevant terms and documents, respectively; W_t^- and D_t^- are "black lists" of irrelevant terms and documents, respectively; and $\beta_\pm$ and $\alpha_\pm$ are regularization coefficients.

Semi-supervised topic learning can be viewed as a type of topic-based information retrieval. In a query, a user provides a topic lexis in the form of a one-topic document, a set of *seed words*, or topic labels assigned to certain word positions [24]. Then the topic search engine should find and organize documents relevant to the specified topics. Semi-supervised topic learning has been used for, e.g., search and categorization of news [25], social media information on diseases and their treatments [26], crime and extremism [27], and inter-ethnic relations [28], [29], [30]. In the Ailment Topic Aspects Model (ATAM), a large corpus of medical papers was used to produce a smoothing distribution β_{wt} [26]. In semi-supervised LDA (SSLDA) and interval semi-supervised LDA (ISLDA) models, a dictionary of a few hundred ethnonyms was used to search ethnically relevant topics [28], [31], [32].

C. Separating subject topics and background topics

In order for a topic model to be more interpretable, each topic should have a *lexical kernel* consisting of words that are frequently used in the corresponding domain and rarely used in other domains. For this, the matrices Φ and Θ should have a different structure of sparseness for domain-specific subject topics $S \subset T$ and for background topics $B = T \backslash S$.

The *subject topic* $t \in S$ contains terms from a particular subject domain. The distributions $p(w\,|\,t)$ must be sparse and significantly different. Distributions $p(d\,|\,t)$ should normally also be sparse since each subject topic should be present in a relatively small fraction of documents.

The *background topic* $t \in B$ contains common words that should not be a part of subject topics. Distributions $p(w\,|\,t)$ and $p(d\,|\,t)$ for background topics are smoothed as they are typically present in most documents.

Topic model with subject and background topics can be viewed as a generalization of robust topic models [33], [22] that use a single background topic.

V. Decorrelation of topics

An interpretable topic model should not contain duplicate or very similar topics. To make topics as diverse as possible, let us minimize the sum of all topic covariances or dot products $\langle \varphi_t, \varphi_s \rangle = \sum_w \varphi_{wt}\varphi_{ws}$:

$$R(\Phi) = -\frac{\tau}{2} \sum_{t\in T} \sum_{s\in T\backslash t} \sum_{w\in W} \varphi_{wt}\varphi_{ws}.$$

Substituting R in (7) gives us the M-step:

$$\varphi_{wt} = \operatorname*{norm}_{w\in W}\Big(n_{wt} - \tau\varphi_{wt} \sum_{s\in T\backslash t} \varphi_{ws}\Big). \tag{12}$$

Decorrelation was first introduced in the Topic Weak Correlated LDA (TWC-LDA) model within the Bayesian framework [34]. In [34], a useful side-effect was observed in that decorrelation groups common words in separate topics. Later experiments with ARTM confirmed this observation [35], [17].

Combining decorrelation with smoothing background topics and sparsing subject topics improves the interpretability of topics [35], [17], [36]. A similar combination of regularizers improved the quality of exploratory search, although none of the search quality criteria were directly optimized [11].

VI. Correlated topic model

The Correlated Topic Model (CTM) formalizes the intuitive idea that documents are likely to contain certain combinations of topics more often than others [37]. For example, a document pertaining to geology is more likely to also be about archaeology than about genetics. This means that the components

of vector θ_d are correlated, whereas basic topic models like LDA assume them to be independent. Dropping this unrealistic assumption may improve the quality of topics. CTM originally employed the Bayesian framework and imposed a prior distribution on Θ with correlations between its components. One way to model correlations for probability vectors is to use the log-normal distribution:

$$p(\ln\theta_d\,|\,\mu,\Sigma) \propto \exp\bigl(-\tfrac{1}{2}(\ln\theta_d-\mu)^{\mathsf T}\Sigma^{-1}(\ln\theta_d-\mu)\bigr),$$

were μ is the mean and Σ is the covariance of the $\ln\theta$ vectors. Using this prior distribution in (4) results in the regularizer:

$$R(\Theta,\mu,\Sigma) = -\frac{\tau}{2}\sum_{d\in D}(\ln\theta_d-\mu)^{\mathsf T}\Sigma^{-1}(\ln\theta_d-\mu).$$

Substituting R in (7) gives us the M-step:

$$\theta_{td} = \operatorname*{norm}_{t\in T}\Bigl(n_{td} - \tau\sum_{s\in T}\Sigma^{-1}_{ts}\bigl(\ln\theta_{sd}-\mu_s\bigr)\Bigr), \qquad (13)$$

where Σ^{-1}_{ts} are elements of the inverse covariance matrix. Parameters Σ,μ can be found through maximum likelihood estimation (4) assuming the θ_d are known:

$$\mu = \frac{1}{|D|}\sum_{d\in D}\ln\theta_d; \quad \Sigma = \frac{1}{|D|}\sum_{d\in D}\bigl(\ln\theta_d-\mu\bigr)\bigl(\ln\theta_d-\mu\bigr)^{\mathsf T}.$$

Parameters Σ,μ can be estimated after each pass of the EM-algorithm over the collection. In [37], Lasso-type regression was used to obtain a sparse covariance matrix Σ. The covariance matrix, Σ itself, is an interesting CTM topic model output that may help to interpret the discovered topics through the relationships between them.

VII. Controlling the number of topics

A topic selection regularizer was introduced in [35] for dropping insignificant topics from the model. The regularizer is based on cross-entropy sparsing of the distribution $p(t)$, which can be easily expressed via Θ:

$$R(\Theta) = \tau n\sum_{t\in T}\frac{1}{|T|}\ln p(t), \qquad p(t) = \sum_d p(d)\theta_{td}.$$

Substituting R in (8) and replacing θ_{td} by the unbiased frequency estimate $\frac{n_{td}}{n_d}$ gives us the M-step:

$$\theta_{td} = \operatorname*{norm}_{t\in T}\Bigl(n_{td}\Bigl(1-\tau\frac{n}{n_t|T|}\Bigr)\Bigr). \qquad (14)$$

If the value of the counter n_t is small enough, all elements of the t-th row become zero, and the topic t is excluded from the model. When using this regularizer, we must start with an excessive number of topics $|T|$.

Topic selection in the ARTM framework is much simpler than in the non-parametric Bayesian Hierarchical Dirichlet Process (HDP) [38]. In both ARTM and HDP approaches, there is a hyperparameter that controls the number of topics: the regularization coefficient τ in ARTM and the hyperparameter γ in HDP.

Both HDP and ARTM can discover the true number of topics, but ARTM does it more accurately and robustly [36]. The topic selection regularizer has another useful feature: it drops duplicate, split, and correlated topics. In addition, ARTM with the topic selection regularizer is 100-times faster than the publicly available HDP implementation.

VIII. Modelling hierarchies of topics

Hierarchical models divide topics into subtopics recursively, thus simplifying information retrieval, browsing, and understanding of large multidisciplinary collections. Much work has been done on hierarchical topic modelling [39], [40], [41]. Despite this, learning a good topical hierarchy and optimizing the size and the structure of the hierarchy are still open problems. Moreover, evaluating the quality of the hierarchy remains an open problem as well [40].

There are multiple strategies for building a hierarchy: top-down vs. down-top, level-by-level vs. node-by-node, tree-based vs. multipartite graph, and document vs. term clustering. Nevertheless, there is no widely adopted best strategy.

In [42], a top-down strategy is proposed within the ARTM framework. The hierarchy is represented by a multipartite graph with a fixed number of levels and topics in each level. Each level is a flat topic model so the time for building a hierarchy is still linear in the size of the collection.

At the top level, we build an ordinary "flat" topic model. Once this parent level ℓ with topic set T is built, we build the next level $\ell+1$ with a larger number of child subtopics S. Conditional probabilities $\psi_{st} = p(s\,|\,t)$ link subtopics s with parent topics t. The requirement is that parent topic $p(w\,|\,t)$ must be accurately approximated by the probabilistic mixture of the child topics $p(w\,|\,s)$:

$$\sum_{t\in T} n_t\ \mathrm{KL}_w\Bigl(p(w\,|\,t)\ \Bigl\|\ \sum_{s\in S}p(w\,|\,s)\,p(s\,|\,t)\Bigr)$$
$$= \sum_{t\in T} n_t\ \mathrm{KL}_w\Bigl(\tfrac{n_{wt}}{n_t}\ \Bigl\|\ \sum_{s\in S}\varphi_{ws}\psi_{st}\Bigr) \ \to\ \min_{\Phi,\Psi},$$

where $\Psi = (\psi_{st})_{S\times T}$ is the *interlevel probability matrix*, which is to be estimated as an extra model parameter when learning the topic model for the $\ell+1$ level.

The above maximization problem is equivalent to the matrix factorization of the parent level matrix $\Phi^\ell = \Phi\Psi$.

Next, we add the above requirement as a regularizer to the MLE for learning the topic model of the level $\ell+1$:

$$R(\Phi,\Psi) = \tau\sum_{t\in T}\sum_{w\in W}n_{wt}\ln\sum_{s\in S}\varphi_{ws}\psi_{st}. \qquad (15)$$

This maximization problem is equivalent to the original (2) if we consider each parent topic t as a pseudo-document and insert it into the collection with assigned word frequencies $n_{wt} = \tau n_t\varphi_{wt}$. This means that it is not necessary to implement a special regularizer for building topical hierarchies. When building the model for level $\ell+1$, it is only necessary to add $|T|$ pseudo-documents to the collection. Then the connection matrix Ψ will appear in the corresponding $|T|$ columns of the estimated Θ matrix.

An additional regularizer may be used to make the interlevel connections more sparse [42]. In particular, this regularizer will force each subtopic to have a single parent; in this case, the hierarchy becomes a tree.

IX. MULTIMODAL ARTM

It is often the case that documents contain meta-data of different *modalities* apart from the text. Examples of textual modalities are: natural language words, n-grams [43], [44], tags [45], and named entities [46]. For short texts with typos, modality of character-level n-grams can be considered; this may help to improve the quality of information retrieval [47]. Non-textual modalities are: authors [48]; time stamps [2], [49]; classes, genres and categories [6]; cited or citing documents [50]; citing or cited authors [51]; users of the document, social network page, or the web-site [7]; pictures in the document; advertisements on the web-page; and so on. Clearly, meta-data can help infer topics and vise-versa, i.e., topics can help infer the semantics of the meta-data or predict the missing values of meta-data. Despite the above use-cases and data-types are quite different, they all can be easily and uniformly incorporated into the framework of multi-modal ARTM. Each document is considered as a container of tokens coming from different modalities, including natural language words.

Let M be a set of modalities. Each modality has its own vocabulary of tokens W_m, $m \in M$. These vocabularies do not overlap, and denote their union by W. Now denote the modality of a token $w \in W$ by $m(w)$. The distribution $p(t\,|\,d)$ of topics in each document is shared across modalities.

The topic model for a modality m is equivalent to (1):

$$p(w\,|\,d) = \sum_{t\in T} p(w\,|\,t)\,p(t\,|\,d) = \sum_{t\in T} \varphi_{wt}\theta_{td}. \qquad (16)$$

Stacking probability matrices $\Phi_m = \big(\varphi_{wt}\big)_{W_m\times T}$ of all the modalities vertically gives the matrix Φ of size $W\times T$.

Consider the log-likelihood for each modality m as a regularizer with coefficient τ_m:

$$\sum_{m,d}\sum_{w\in W_m} \tau_m n_{dw} \ln \sum_{t\in T} \varphi_{wt}\theta_{td} + R(\Phi,\Theta) \to \max_{\Phi,\Theta}; \qquad (17)$$

$$\sum_{w\in W_m} \varphi_{wt} = 1;\ \varphi_{wt} \geqslant 0; \quad \sum_{t\in T} \theta_{td} = 1;\ \theta_{td} \geqslant 0. \qquad (18)$$

Theorem 2: Let regularizer $R(\Phi,\Theta)$ be differentiable. Then a local extreme (Φ,Θ) of the optimization problem (17)–(18) satisfies the following conditions with auxiliary variables $p_{tdw} = p(t\,|\,d,w)$ for all non-degenerate topics and documents:

$$p_{tdw} = \operatorname*{norm}_{t\in T}\big(\varphi_{wt}\theta_{td}\big); \qquad (19)$$

$$\varphi_{wt} = \operatorname*{norm}_{w\in W_m}\bigg(\sum_{d\in D} \tau_{m(w)} n_{dw} p_{tdw} + \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}}\bigg); \qquad (20)$$

$$\theta_{td} = \operatorname*{norm}_{t\in T}\bigg(\sum_{w\in W} \tau_{m(w)} n_{dw} p_{tdw} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\bigg). \qquad (21)$$

Theorem 1 is a particular case of Theorem 2 with single modality, $|M| = 1$, and $\tau_m = 1$.

We can see that the multimodal extension of ARTM consists of two modifications: 1) breaking the matrix Φ into blocks Φ_m that are normalized separately, and 2) multiplying the data n_{dw} by weights of modalities $\tau_{m(w)}$.

A. The language modalities

One excellent idea in multi-language topic modelling is that the parallel collection of document translations is sufficient to find topics across languages and then to build the cross-language search [4]. The first multi-lingual topic models [52] considered each language as a separate modality and merged all translations of a document into one common document. Aligning parallel texts by sentences or words proved to be time-consuming and essentially did not improve the quality of the cross-lingual search.

Using a cross-language dictionary is a type of smoothing regularizer [53]. It expresses the guess that if a word u in a language k is a translation of the word w in a language ℓ, then the topic distributions of these words $p(t\,|\,u)$ and $p(t\,|\,w)$ should be close in the sense of cross-entropy:

$$R(\Phi) = \sum_{w,u}\sum_{t\in T} n_{ut} \ln \varphi_{wt}.$$

Substituting R in (7) gives us the M-step:

$$\varphi_{wt} = \operatorname*{norm}_{w\in W^\ell}\Big(n_{wt} + \tau\sum_u n_{ut}\Big).$$

We can see that probability of a word w in a topic t increases if the word has translations that also have a high probability of appearing in topic t. Experiments showed that linking parallel texts improves the search quality more effectively than using a bilingual dictionary [53].

B. The class modality

The problem of supervised document classification is one of the most important in text analysis [54]. Classifiers such as Support Vector Machine (SVM) or Regularized Logistic Regression (RLR) are generally reported to be good techniques for this task. A drawback to this approach is that performance drops rapidly as the total number of class labels and the number of labels per document increase. Topic models for multi-label document classification cope with this problem by processing class labels in the same way as words, which was done in the Dependency LDA model [6] within a Bayesian framework.

The same idea can be expressed easily in a multimodal setting by introducing the modality of class labels C. Each document $d \in D$ contains a subset of labels $C_d \subset C$.

At the training stage, we fit a topic model using both word and class modalities to obtain $\varphi_{wt} = p(w\,|\,t)$ and $\varphi_{ct} = p(c\,|\,t)$, as well as topic distributions $\theta_{td} = p(t\,|\,d)$ for each training d.

At the testing stage, we infer $\theta_{td'}$ for a new document d' with empty set $C_{d'}$ using word counts $n_{d'w}$ of the document and word distributions of topics φ_{wt}. Then the class labels for the document d' can de predicted by the probabilistic model:

$$p(c\,|\,d') = \sum_{t\in T} \varphi_{ct}\theta_{td'},$$

which is essentially a linear classifier with feature vector $\theta_{d'}$ and coefficients φ_{ct}. Next, we can choose some thresholds to convert probabilities of classes $p(c|d)$ into class labels. Alternatively, we can apply any non-linear classifier to the topic distributions $\theta_{td'}$ as feature vectors.

Experiments in [6] indicate that topic models gives superior classification quality for a multi-class problem with a large number of imbalanced, overlapping, and interdependent classes. The multimodal ARTM framework gives comparable results for the same datasets [55].

C. The time modality

Document time stamps are important for modeling the topical dynamics in newsfeeds, scientific publications, patent databases, and social media.

We introduce the time modality of time intervals as a finite set I. Assume that topics as distribution $p(w|t)$ do not change in time. In the multimodal ARTM framework, the topic dynamic over time $p(i|t) = \varphi_{it}$ appears in the t-th column of the Φ matrix according to (16):

$$p(i|d) = \sum_{t\in T} p(i|t)\, p(t|d) = \sum_{t\in T} \varphi_{it}\theta_{td}. \tag{22}$$

In Topics Over Time (TOT) [56], the dynamic of a topic is modelled by a parametric beta-distribution. This distribution family includes monotone and unimodal distributions, which are convenient for modelling event topics and a few variants of trending dynamics, but not suitable for describing more complex dynamics. Non-parametric models are more flexible and can describe arbitrary dynamics. However, some constraints must be specified to avoid overfitting. Consider two regularizers that control the topic dynamics.

First, assume that many topics correspond to short-lived events. Therefore, each interval i contains a small part of the topics from T. We essentially require the sparseness of the distributions $p(t|i)$, and we achieve this by applying a cross-entropy regularizer using Bayes rule $p(t|i) = p(i|t)\frac{p(t)}{p(i)}$:

$$R_1(\Phi) = -\tau_1 \sum_{i\in I}\sum_{t\in T} \ln \frac{\varphi_{it} n_t}{\sum_z \varphi_{iz} n_z},$$

where the n_t counter is produced in the EM-algorithm.

Second, assume that probabilities $p(i|t)$ do not change too rapidly in time i for a topic t. This requirement can be formalized by the L_1 smoothness regularizer:

$$R_2(\Phi) = -\tau_2 \sum_{i\in I}\sum_{t\in T} \big|\varphi_{it} - \varphi_{i-1,t}\big|.$$

This regularizer smooths the values $p(i|t)$ at each point of the time-series relative to the neighbouring points.

X. Author topic model

The *author topic model* (ATM) first introduced in [48] was motivated by an assumption that topics are generated by authors of documents rather than the documents themselves. Some other modality could replace authors as topic-generative, e.g., document category or any type of document source. This assumption changes the structure of the parameter space and leads to the three-matrix factorization problem.

Assume that each term w in each document d is associated, not only with a topic, but also with a category c from a given set of categories C. Assume also that for each document, we know a subset $C_d \subseteq C$ that can be associated with words in this document. For example, we may know the set of document authors. Now, observations (d_i, w_i, t_i, c_i) come from the extended space $D \times W \times T \times C$ instead of $D \times W \times T$.

Consider a topic model (1) in which probabilities of topics for a document $\theta_{td} = p(t|d)$ are calculated using mixtures of distributions $\psi_{tc} = p(t|c)$ of topics in categories (e.g., the topic profile for an author) and distributions $\pi_{cd} = p(c|d)$ of categories for documents (e.g., the contribution of each author to the document d):

$$p(w|d) = \sum_{t\in T}\sum_{c\in C_d} \varphi_{wt}\psi_{tc}\pi_{cd}. \tag{23}$$

The model is based on two conditional independence assumptions: $p(t|c,d) = p(t|c)$ and $p(w|t,c,d) = p(w|t)$.

We use the regularized log-likelihood maximization:

$$\sum_{d\in D}\sum_{w\in d} n_{dw} \ln \sum_{t\in T}\sum_{c\in C_d} \varphi_{wt}\psi_{tc}\pi_{cd} + R(\Phi,\Psi,\Pi) \to \max_{\Phi,\Psi,\Pi}$$

with the usual constraints on probability matrices Φ, Ψ, Π.

Theorem 3: Let regularizer $R(\Phi,\Psi,\Pi)$ be differentiable. Then local extreme (Φ,Ψ,Π) of the optimization problem satisfies the following conditions with auxiliary variables $p_{tcdw} = p(t,c|d,w)$ for all non-degenerate t, d, c:

$$\begin{aligned}
p_{tcdw} &= \operatorname*{norm}_{(t,c)\in T\times C_d} \varphi_{wt}\psi_{tc}\pi_{cd};\\
\varphi_{wt} &= \operatorname*{norm}_{w\in W}\Bigg(\sum_{d\in D}\sum_{c\in C_d} n_{dw}p_{tcdw} + \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}}\Bigg);\\
\psi_{tc} &= \operatorname*{norm}_{t\in T}\Bigg(\sum_{d\in D}\sum_{w\in d} n_{dw}p_{tcdw} + \psi_{tc}\frac{\partial R}{\partial \psi_{tc}}\Bigg);\\
\pi_{cd} &= \operatorname*{norm}_{c\in C_d}\Bigg(\sum_{w\in d}\sum_{t\in T} n_{dw}p_{tcdw} + \pi_{cd}\frac{\partial R}{\partial \pi_{cd}}\Bigg).
\end{aligned}$$

In the *Tag Weighted Topic Model* (TWTM) [57], tags for the document were used as a topic-generating modality. A similar model was used for video processing [58]. Documents d were defined as consecutive 1-second video clips, terms w corresponded to the visual events, topics t were interpreted as actions composed of events, and categories c were used for mining complicated behaviours. The problem as outlined was to recognize a major behaviour c in each one-second clip.

XI. Regression topic model

There are a lot of practically important problems in which it is necessary to predict some numeric value for a text document. Some application examples from e-commerce are: prediction of the user rating for a product (e.g., a consumer good, movie or book) based on the review text; prediction of the number of clicks on the ad based on its text; prediction of the salary based on the job opening description; and prediction of the number of likes on the user-review for a service.

Standard regression models use a vector document representation. Then a topic model can be used as a tool that extracts feature vectors θ_d. Another approach is to include the regression fitting criterion as a regularizer [59]. This may help to infer topics more suitable for the numerical prediction.

Assume that there is a target value $y_d \in \mathbb{R}$ that is known for each document d in the training collection that must be predicted for new documents. Often regression is fitted to minimize the squared error between predictions and target values. Consider a linear regression for simplicity:

$$R(\Theta, v) = -\tau \sum_{d \in D} \Big(y_d - \sum_{t \in T} v_t \theta_{td} \Big)^2,$$

where $v \in \mathbb{R}^T$ is a vector of regression coefficients. It can be found from likelihood maximization (4) by fixing θ_d:

$$v = (\Theta\Theta^{\mathsf{T}})^{-1}\Theta y.$$

This is the standard linear regression solution for known "data" matrix Θ. Next, we substitute the above regularizer into the equation for the M-step (8):

$$\theta_{td} = \operatorname*{norm}_t \Big(n_{td} + \tau v_t \theta_{td} \Big(y_d - \sum_{s \in T} v_s \theta_{sd} \Big) \Big).$$

Vector v can be updated after each pass of the EM-algorithm over the document collection [59].

XII. Modelling connected documents

There is often some information on links between documents that presume similarity of documents topics. This can be the fact that two documents reside in the same category, are mentioned together in another document, are hyperlinked, or one cites the other. The similarity between the topic-profiles of documents d and d' can be measured by the covariance $\sum_t \theta_{td}\theta_{td'}$. We then introduce a regularizer to maximize these covariances for linked documents:

$$R(\Theta) = \tau \sum_{d,d'} w_{dd'} \sum_{t \in T} \theta_{td}\theta_{td'},$$

where $w_{dd'}$ is the weight of the connection between documents d and d', e.g., the number of links between them.

Substituting R in (8) gives us the M-step:

$$\theta_{td} = \operatorname*{norm}_{t \in T} \Big(n_{td} + \tau \theta_{td} \sum_{d' \in D} w_{dd'} \theta_{td'} \Big).$$

This is a variant of smoothing regularizer. The θ_d distribution is pushed towards distributions $\theta_{d'}$ of documents connected with d.

A. Document network topic model

The paper [60] introduces a generic topic model NetPLSA, which accounts for a given graph structure imposed on a document collection. Consider a graph $\langle V, E \rangle$ with a set of vertices V and a set of edges E. Each vertex $v \in V$ corresponds to a subset of documents $D_v \subset D$. For example, the vertex can be a single document v, all posts of one author v, or all posts from one geographic region v.

The topic distribution of a vertex v follows from the law of total probability:

$$p(t\,|\,v) = \sum_{d \in D_v} p(t\,|\,d)\, p(d\,|\,v) = \sum_{d \in D_v} \theta_{td} p_{dv},$$

where p_{dv} can be estimated as $\operatorname*{norm}_{d \in D_v}(n_d)$ or $\frac{1}{|D_v|}$.

The NetPLSA model introduces a quadratic regularizer:

$$R(\Theta) = -\frac{\tau}{2} \sum_{(u,v) \in E} w_{uv} \sum_{t \in T} \big(p(t\,|\,v) - p(t\,|\,u) \big)^2,$$

where w_{uv} is the weight of an edge (u, v). For example, if D_v consists of all the papers of an author v, then w_{uv} can be the number of papers in which u and v are co-authors.

This regularizer requires us to have access to all documents profiles θ when processing a document d. This may be computationally inefficient for a large collection.

Alternatively, we can make the set of vertices V a modality and introduce a regularizer that depends only on the matrix Φ. Let us add into each document $d \in D_v$ a token $v \in V$ of the modality V. We express the topic distribution of a vertex using Bayes rule: $p(t\,|\,v) = \varphi_{vt}\frac{p_t}{p_v}$, and then we use frequency estimates for p_v and p_t. Substituting this into the NetPLSA regularizer makes it a function of Φ rather than Θ:

$$R(\Phi) = -\frac{\tau}{2} \sum_{(u,v) \in E} w_{uv} \sum_{t \in T} \Big(\varphi_{vt}\frac{p_t}{p_v} - \varphi_{ut}\frac{p_t}{p_u} \Big)^2. \qquad (24)$$

B. The modality of geotags and geolocations

Geographic locations associated with documents or with their authors are often used when analysing social network data. The applications may include discovering location-specific topics and the distribution of topics over locations. For example, in [61], areas of popularity of national cousins are analysed based on Flickr users posts. Another example is a reconstruction of the geographic path of the "Katrina" hurricane based on social network posts [60].

There are two common ways to specify the location of a document d. The first is the modality of *geotags*, e.g., the names of countries, regions, cities and so on. We can use directly the modality of geotags following the approach used in Section IX. The second way to specify location is to use the geographic coordinate or *geolocation* described by latitude and longitude $\ell_d = (x_d, y_d)$. We can use the regularizer (24) to account for the geographical proximity of locations. A quadratic regularizer first suggested in [61] is, in fact, the same as that used in in the NetPLSA with weights $w_{uv} = \exp(-\gamma r_{uv}^2)$ based on geographic proximity $r_{uv}^2 = (x_u - x_v)^2 + (y_u - y_v)^2$, where (u, v) can be either a document pair or a geotag pair. The NetPLSA generic approach allows both variants of modeling.

The ARTM framework allows combining both types of geographic data in the model.

XIII. Beyond bag-of-words

The bag-of-words hypothesis is probably one of most criticized assumptions in topic modelling. In response to this criticism, many advanced models of sequential text appeared. We distinguish three directions of such extensions.

The first is to consider n-*grams* or *collocations* of words rather than individual words. Topics inferred from n-grams are usually much easier to interpret than those based on

unigrams [43]. The second extension is to consider the *co-occurrence* of words according to the Harris' distributional hypothesis [62]. The development of `word2vec` [63] and other *word embeddings* techniques [64] stimulated the development of sparse interpretable topic-based word embeddings [65]. The third extension is based on the assumption that the natural language text is usually a sequence of segments, each containing only one topic or a very small number of topics. This leads to *sentence* topic models and topic-based *text segmentation* techniques.

A. Multigram topic models

The first *Bigram Topic Model* (BTM) [43] is formally equivalent to a multimodal model in which each word $v \in W$ induces a separate modality. The vocabulary $W_v \subseteq W$ of the word modality consists of all words that appear exactly after v somewhere in the collection. This multimodal representation allows us to introduce the conditional probabilities $\varphi^v_{wt} = p(w|v,t)$ over words w that go after the word v in the topic t. The log-likelihood of the bigram model can be used as a regularizer for the unigram model log-likelihood (in the original BTM it has been used as a separate bigram model objective):

$$R(\Phi,\Theta) = \sum_{d\in D}\sum_{v\in d}\sum_{w\in W_v} n_{dvw} \ln \sum_{t\in T} \varphi^v_{wt}\theta_{td},$$

where n_{dvw} counts the bigram vw in the document d. The limitation of the BTM model is that it does not consider higher order n-grams. Another problem is that the number of bigrams grows rapidly with document collection size.

In the multimodal ARTM framework, n-grams can be more naturally specified as separate modalities for each n (unigrams, bigrams, 3-grams, etc). To reduce the sizes of the vocabularies, recent fast collocation miners can be used: TopMine [66], SegPhrase [67], or AutoPhrase [68].

B. Biterm topic model for short texts

Short texts are documents that are not long enough for reliable topic inference. Examples are Twitter messages, news headers, short ads, dialog messages, and so on.

The *Biterm Topic Model* (BitermTM) is one of the most successful approaches to the problem of short texts [69]. *Biterm* is a pair of words that occur near to each other in the text. "Near" can mean in one sentence or a window of $\pm h$ words, depending on the problem at hand. The input data for the model are the counts n_{uv} of biterms (u,v) in the document collection. BitermTM describes the probability of words co-occurrence $p(u,v)$ using the conditional independence assumption $p(u,v\,|\,t) = p(u\,|\,t)\,p(v\,|\,t)$ and the law of total probability:

$$p(u,v) = \sum_{t\in T} p(u\,|\,t)\,p(v\,|\,t)\,p(t) = \sum_{t\in T}\varphi_{ut}\varphi_{vt}\pi_t,$$

where $\varphi_{wt} = p(w\,|\,t)$ and $\pi_t = p(t)$ are model parameters. This is a 3-matrix factorization $\Phi\Pi\Phi^{\mathsf{T}}$ with a diagonal matrix $\Pi = \operatorname{diag}(\pi_1,\ldots,\pi_T)$. The biterm topic model does not define the topic profiles of documents Θ and, hence, does encounter the problem of determining topic profiles for short documents.

ARTM allows us to combine the biterm and ordinary topic models and estimate the improved Θ matrix for short documents. For this, we can use the log-likelihood of the biterm topic model as a regularizer:

$$R(\Phi,\Pi) = \tau \sum_{u,v} n_{uv} \ln \sum_t \varphi_{ut}\varphi_{vt}\pi_t.$$

Substituting R in (7) gives us the M-step:

$$\varphi_{wt} = \operatorname*{norm}_{w\in W}\Big(n_{wt} + \tau \sum_{u\in W} n_{uw}p_{tuw}\Big); \tag{25}$$

$$p_{tuw} = \operatorname*{norm}_{t\in T}\big(n_t\varphi_{wt}\varphi_{ut}\big). \tag{26}$$

This can be interpreted as adding pseudo-documents to the collection. For each word $u \in W$, let us define a pseudo-document d_u containing the bag of words that appeared near the word u anywhere in the collection. The count of the word w in the pseudo-document d_u equals τn_{uw}. Then computing the auxiliary variables $p_{tuw} = p(t\,|\,u,w)$ in the EM-algorithm corresponds to the E-step processing for the pseudo-document d_u if its topic distribution is defined as $\theta_{tu} = \operatorname*{norm}_t(n_t\varphi_{ut})$. In other words, in the biterm topic model, the columns of Θ corresponding to the pseudo-documents are computed from the rows of matrix Φ using Bayes rule.

Increasing the regularization coefficient τ, we can force the matrix Φ to be estimated mainly with biterms. In the limit $\tau \to \infty$, the combined model tends to BitermTM.

C. Word network topic model

The above-mentioned idea of modelling word-context pseudo-documents instead of the original documents is at heart of the *Word Topic Model* (WTM) [70] and *Word Network Topic Model* (WNTM) [71]. Essentially, WTM and WNTM are equivalent to applying PLSA and LDA correspondingly to the collection of pseudo-documents d_u:

$$p(w\,|\,d_u) = \sum_{t\in T} p(w\,|\,t)\,p(t\,|\,d_u) = \sum_{t\in T}\varphi_{wt}\theta_{tu}.$$

Consider the log-likelihood of the model $p(w\,|\,d_u)$ as a regularizer for the original topic model:

$$R(\Phi,\Theta) = \tau \sum_{u,w\in W} n_{uw} \ln \sum_{t\in T}\varphi_{wt}\theta_{tu},$$

where n_{uw} is the count of the co-occurrence of words u,w and is defined as above for biterms. The major difference from the biterm topic model is that we explicitly infer Θ for pseudo-documents, while in the biterm topic model $\Theta = \operatorname{diag}(\pi_1,\ldots,\pi_t)\Phi^{\mathsf{T}}$. Hence, the number of estimated parameters is two times larger in WNTM. Experiments in [71] based on a collection of short texts indicate that WNTM performs slightly better than BTM and significantly better than LDA. For collections of large texts, the co-occurrence topic models do not provide significant advantage.

Both biterm and word network topic models give sparse and interpretable topic-based word embeddings [65]. Word embedding is a vector representation of a word. In the case of a topic model, the $|T|$-dimensional vector consists of conditional probabilities $p(t\,|\,w) = \varphi_{wt}\frac{p(t)}{p(w)}$. The resulting topic-based embeddings perform on par with Skip-Gram Negative Sampling (SGNS) [63] on word similarity tasks and benefit in the sparseness and interpretability of the components [65].

XIV. Experiments

We compare the BigARTM library with the latest versions of the two major public libraries for topic modelling.

Vowpal Wabbit (VW)is a library of online machine learning algorithms. For topic modelling, VW contains the VW.LDA algorithm. VW.LDA is not multi-core, but an effective single-threaded implementation in C++ made it one of the widely adopted tools for topic modelling.

Gensim [72] is a library for topic modelling and matrix factorization. It has two LDA implementations — LdaModel and LdaMulticore. Gensim is written in Python and, to speed-up calculations, it uses the NumPy library. In LdaModel, all batches are processed sequentially, and the concurrent processing is done only in NumPy. In LdaMulticore, several batches are processed concurrently, and a single aggregation thread merges the results asynchronously.

The architecture of BigARTM algorithm is based on multithreading with update delays [73].

Both Gensim and Vowpal Wabbit use the *online variational Bayes* LDA [74]. All three libraries work out-of-core, i. e., they are designed to process datasets that are too large to fit into a computers main memory at one time. This allowed us to benchmark using a fairly large collection of 3.7 million articles from the English Wikipedia.For each library, we perform a *single* pass over the collection and train a model with a fixed number of topics. The collection was split into batches with 10K documents each (`chunksize` in Gensim, `minibatch` in VW.LDA). The vocabulary consists of words that appear in at least 20 documents, but in no more than 10% of documents in the collection. The resulting dictionary was capped at the $|W| = 100\,000$ most frequent words.

Perplexity is used as the test sample quality measure:

$$\mathscr{P}(D, p) = \exp\left(-\frac{1}{n}\sum_{d\in D}\sum_{w\in d} n_{dw} \ln p(w\,|\,d)\right),$$

which is essentially an inverse of the likelihood of data, i.e., the smaller it is for the test data, the better. The size of the test sample for computing the perplexity is 100K documents.

In order to make a fair comparison, we have configured BigARTM to use only smoothing out of variety of regularizers it has, which is equivalent to the LDA model. LDA priors were fixed as $\alpha = 1/|T|,\ \beta = 1/|T|$ for all libraries.

For the experiments, we used the latest versions: VW 8.4.0, Gensim 2.3.0 (v0.10.3 under Python 2.7), and BigARTM 0.8.3. We also used a Dell Precision T5600 workstation with 2 Intel(R) Xeon(R) CPU E5-2650 0 @ 2.00Hz, 8 Cores and 16 Logical Processors.

Table I compares the performance of the libraries.

We can see that if we do not explicitly split the training between multiple processors, BigARTM is already ∼5 times faster than Gensim and ∼2 times faster than VW. The out-of-sample perplexity for all libraries is on par.

VW is not designed to explicitly split training job between processors, so its results are effectively the same in all the rows of the table. If we explicitly specify using multiple processors for training in Gensim and BigARTM, BigARTM is 5–10 times faster than VW and 10–20 times faster than Gensim. Out-of-sample perplexity is on par between BigARTM and the VW and for both it is better than the Gensim perplexity.

TABLE I. The comparison of BigARTM with VW.LDA and Gensim training time and out-of-sample quality. Cell format: "train time in minutes (test perplexity)". Rows: P — # of processors, T — # of topics.

	Gensim	VW-LDA	BigARTM	BigARTM (async)
P = 1, T= 50	142m (4945)	50m (5413)	42m (5117)	25m (5131)
P = 1, T= 100	287m (3969)	91m (4592)	52m (4093)	32m (4133)
P = 1, T= 200	637m (3241)	154m (3960)	83m (3347)	53m (3362)
P = 2, T= 50	89m (5056)		22m (5092)	13m (5160)
P = 2, T= 100	143m (4012)		29m (4107)	19m (4144)
P = 2, T= 200	325m (3297)		47m (3347)	28m (3380)
P = 4, T= 50	88m (5311)		12m (5216)	7m (5353)
P = 4, T= 100	104m (4338)		16m (4233)	10m (4357)
P = 4, T= 200	315m (3583)		26m (3520)	16m (3634)
P = 8, T= 50	88m (6344)		8m (5648)	5m (6220)
P = 8, T= 100	107m (5380)		10m (4660)	6m (5119)
P = 8, T= 200	288m (4263)		15m (3929)	10m (4309)

TABLE II. Run of BigARTM with a large number of topics. Cell format: "train time in minutes (test perplexity)".

Framework/Topics	2000	5000
BigARTM	166m (2377)	399m (1942)
BigARTM (async)	119m (2645)	281m (2216)

Finally, we run model training in BigARTM with a very large number (2000 and 5000) of topics, Table II. We can see that BigARTM is able to solve the task in a reasonable time. Asynchronous training performs better in terms of time, although it loses slightly in the out-of-sample quality comparison. Nevertheless, it was shown that the asynchronous algorithm achieves a better model in the given time-frame than the synchronous algorithm [73].

XV. Conclusion

Ater more than a decade of active development, hundreds of types of topic models have been created: hierarchical, temporal, multimodal, multilingual, supervised, semi-supervised, relational, sequential, and many others. Several major types of modern topic models were reviewed in this paper.

In applications, topic models often have to combine multiple extensions. Additive Regularization of Topic Models (ARTM) provides topic modellers with a "bag-of-regularizers" modular technology implemented in the open-source library BigARTM. A built-in set of unified regularizers enables the construction of topic models for various practical applications without tedious derivations and programming.

A lot of topic models introduced in the Bayesian framework can be reformulated much more simply in the ARTM framework, as we have tried to demonstrate in this review. We hope that ARTM will prove to be a convenient language for studying topic modelling and lowering the entry barrier for practitioners

Acknowledgements: The work was supported by the Ministry of Education and Science of the Russian Federation (project RFMEFI57915X0117).

REFERENCES

[1] T. Hofmann, "Probabilistic latent semantic indexing," in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. New York, NY, USA: ACM, 1999, pp. 50–57.

[2] W. Cui, S. Liu, L. Tan, C. Shi, Y. Song, Z. Gao, H. Qu, and X. Tong, "TextFlow: Towards better understanding of evolving topics in text." *IEEE transactions on visualization and computer graphics*, vol. 17, no. 12, pp. 2412–2421, 2011.

[3] E. E. Veas and C. di Sciascio, "Interactive topic analysis with visual analytics and recommender systems," in *2nd Workshop on Cognitive Computing and Applications for Augmented Human Intelligence, CCAAHI2015, International Joint Conference on Artificial Intelligence, IJCAI, Buenos Aires, Argentina, July 2015*. Aachen, Germany, Germany: CEUR-WS.org, 2015.

[4] I. Vulic, W. De Smet, J. Tang, and M.-F. Moens, "Probabilistic topic modeling in multilingual settings: an overview of its methodology and applications," *Information Processing & Management*, vol. 51, no. 1, pp. 111–147, 2015.

[5] J. C. L. Pinto and T. Chahed, "Modeling multi-topic information diffusion in social networks using latent Dirichlet allocation and Hawkes processes," in *Tenth International Conference on Signal-Image Technology & Internet-Based Systems*, 2014, pp. 339–346.

[6] T. N. Rubin, A. Chambers, P. Smyth, and M. Steyvers, "Statistical topic models for multi-label document classification," *Machine Learning*, vol. 88, no. 1-2, pp. 157–208, 2012.

[7] C. Wang and D. M. Blei, "Collaborative topic modeling for recommending scientific articles," in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: ACM, 2011, pp. 448–456.

[8] A. Daud, J. Li, L. Zhou, and F. Muhammad, "Knowledge discovery through directed probabilistic topic models: a survey," *Frontiers of Computer Science in China*, vol. 4, no. 2, pp. 280–301, 2010.

[9] D. M. Blei, "Probabilistic topic models," *Communications of the ACM*, vol. 55, no. 4, pp. 77–84, 2012.

[10] O. Khalifa, D. Corne, M. Chantler, and F. Halley, "Multi-objective topic modelling," in *7th International Conference Evolutionary Multi-Criterion Optimization (EMO 2013)*. Springer LNCS, 2013, pp. 51–65.

[11] A. Yanina and K. Vorontsov, "Multi-objective topic modeling for exploratory search in tech news," in *AINL-6: Artificial Intelligence and Natural Language Conference, St. Petersburg, Russia, September 20-23, 2017*, 2017 (to appear).

[12] A. N. Tikhonov and V. Y. Arsenin, *Solution of ill-posed problems*. W. H. Winston, Washington, DC, 1977.

[13] K. V. Vorontsov, "Additive regularization for topic models of text collections," *Doklady Mathematics*, vol. 89, no. 3, pp. 301–304, 2014.

[14] K. V. Vorontsov and A. A. Potapenko, "Additive regularization of topic models," *Machine Learning, Special Issue on Data Analysis and Intelligent Optimization*, 2014.

[15] K. Vorontsov, O. Frei, M. Apishev, P. Romov, and M. Suvorova, "Bigartm: Open source library for regularized multimodal topic modeling of large collections," in *AIST'2015, Analysis of Images, Social networks and Texts*. Springer International Publishing Switzerland, Communications in Computer and Information Science (CCIS), 2015, pp. 370–384.

[16] A. Asuncion, M. Welling, P. Smyth, and Y. W. Teh, "On smoothing and inference for topic models," in *Proceedings of the International Conference on Uncertainty in Artificial Intelligence*, 2009, pp. 27–34.

[17] K. V. Vorontsov and A. A. Potapenko, "Additive regularization of topic models," *Machine Learning, Special Issue on Data Analysis and Intelligent Optimization with Applications*, vol. 101, no. 1, pp. 303–323, 2015.

[18] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[19] T. Masada, S. Kiyasu, and S. Miyahara, "Comparing LDA with pLSI as a dimensionality reduction method in document clustering," in *Proceedings of the 3rd International Conference on Large-scale knowledge resources: construction and application*, ser. LKR'08. Springer-Verlag, 2008, pp. 13–26.

[20] Y. Wu, Y. Ding, X. Wang, and J. Xu, "A comparative study of topic models for topic clustering of Chinese web news," in *Computer Science and Information Technology (ICCSIT), 2010 3rd IEEE International Conference on*, vol. 5, july 2010, pp. 236–240.

[21] Y. Lu, Q. Mei, and C. Zhai, "Investigating task performance of probabilistic topic models: an empirical study of PLSA and LDA," *Information Retrieval*, vol. 14, no. 2, pp. 178–203, 2011.

[22] A. A. Potapenko and K. V. Vorontsov, "Robust PLSA performs better than LDA," in *35th European Conference on Information Retrieval, ECIR-2013, Moscow, Russia, 24-27 March 2013*. Lecture Notes in Computer Science (LNCS), Springer Verlag-Germany, 2013, pp. 784–787.

[23] M. Girolami and A. Kabán, "On an equivalence between PLSI and LDA," in *SIGIR'03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, 2003, pp. 433–434.

[24] D. Andrzejewski and X. Zhu, "Latent Dirichlet allocation with topic-in-set knowledge," in *Proceedings of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing*, ser. SemiSupLearn '09. Stroudsburg, PA, USA: Association for Computational Linguistics, 2009, pp. 43–48.

[25] J. Jagarlamudi, H. Daumé, III, and R. Udupa, "Incorporating lexical priors into topic models," in *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, ser. EACL'12. Stroudsburg, PA, USA: Association for Computational Linguistics, 2012, pp. 204–213.

[26] M. J. Paul and M. Dredze, "Discovering health topics in social media using topic models," *PLoS ONE*, vol. 9, no. 8, 2014.

[27] A. Sharma and D. M. Pawar, "Survey paper on topic modeling techniques to gain usefull forcasting information on violant extremist activities over cyber space," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 5, no. 12, pp. 429–436, 2015.

[28] S. Bodrunova, S. Koltsov, O. Koltsova, S. I. Nikolenko, and A. Shimorina, "Interval semi-supervised LDA: Classifying needles in a haystack," in *MICAI (1)*, ser. Lecture Notes in Computer Science, F. C. Espinoza, A. F. Gelbukh, and M. Gonzalez-Mendoza, Eds., vol. 8265. Springer, 2013, pp. 265–274.

[29] S. Koltcov, O. Koltsova, and S. Nikolenko, "Latent Dirichlet allocation: Stability and applications to studies of user-generated content," in *Proceedings of the 2014 ACM Conference on Web Science*, ser. WebSci'14. New York, NY, USA: ACM, 2014, pp. 161–165.

[30] S. I. Nikolenko, S. Koltcov, and O. Koltsova, "Topic modelling for qualitative studies," *Journal of Information Science*, vol. 43, no. 1, pp. 88–102, 2017.

[31] M. Apishev, S. Koltcov, O. Koltsova, S. Nikolenko, and K. Vorontsov, "Additive regularization for topic modeling in sociological studies of user-generated text content," in *MICAI 2016, 15th Mexican International Conference on Artificial Intelligence*, vol. 10061. Springer, Lecture Notes in Artificial Intelligence, 2016, pp. 166–181.

[32] ——, "Mining ethnic content online with additively regularized topic models," *Computacion y Sistemas*, vol. 20, no. 3, pp. 387–403, 2016.

[33] C. Chemudugunta, P. Smyth, and M. Steyvers, "Modeling general and specific aspects of documents with a probabilistic topic model," in *Advances in Neural Information Processing Systems*, vol. 19. MIT Press, 2007, pp. 241–248.

[34] Y. Tan and Z. Ou, "Topic-weak-correlated latent Dirichlet allocation," in *7th International Symposium Chinese Spoken Language Processing (ISCSLP)*, 2010, pp. 224–228.

[35] K. V. Vorontsov and A. A. Potapenko, "Tutorial on probabilistic topic modeling: Additive regularization for stochastic matrix factorization," in *AIST'2014, Analysis of Images, Social networks and Texts*, vol. 436. Springer International Publishing Switzerland, Communications in Computer and Information Science (CCIS), 2014, pp. 29–46.

[36] K. V. Vorontsov, A. A. Potapenko, and A. V. Plavin, "Additive regularization of topic models for topic selection and sparse factorization," in *The Third International Symposium On Learning And Data Sciences (SLDS 2015). April 20-22, 2015. Royal Holloway, University of London, UK.*, A. G. et al., Ed. Springer International Publishing Switzerland 2015, 2015, pp. 193–202.

[37] D. Blei and J. Lafferty, "A correlated topic model of Science," *Annals of Applied Statistics*, vol. 1, pp. 17–35, 2007.

[38] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Hierarchical Dirichlet processes," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1566–1581, 2006.

[39] D. M. Blei, T. Griffiths, M. Jordan, and J. Tenenbaum, "Hierarchical topic models and the nested chinese restaurant process," in *NIPS*, 2003.

[40] E. Zavitsanos, G. Paliouras, and G. A. Vouros, "Non-parametric estimation of topic hierarchies from texts with hierarchical Dirichlet processes," *Journal of Machine Learning Research*, vol. 12, pp. 2749–2775, 2011.

[41] C. Wang, X. Liu, Y. Song, and J. Han, "Scalable and robust construction of topical hierarchies," *CoRR*, vol. abs/1403.3460, 2014.

[42] N. A. Chirkova and K. V. Vorontsov, "Additive regularization for hierarchical multimodal topic modeling," *Journal Machine Learning and Data Analysis*, vol. 2, no. 2, pp. 187–200, 2016.

[43] H. M. Wallach, "Topic modeling: Beyond bag-of-words," in *Proceedings of the 23rd International Conference on Machine Learning*, ser. ICML '06. New York, NY, USA: ACM, 2006, pp. 977–984.

[44] X. Wang, A. McCallum, and X. Wei, "Topical n-grams: Phrase and topic discovery, with an application to information retrieval," in *Proceedings of the 2007 Seventh IEEE International Conference on Data Mining*. Washington, DC, USA: IEEE Computer Society, 2007, pp. 697–702.

[45] R. Krestel, P. Fankhauser, and W. Nejdl, "Latent Dirichlet allocation for tag recommendation," in *Proceedings of the third ACM conference on Recommender systems*. ACM, 2009, pp. 61–68.

[46] D. Newman, C. Chemudugunta, and P. Smyth, "Statistical entity-topic models," in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '06. New York, NY, USA: ACM, 2006, pp. 680–686.

[47] P.-S. Huang, X. He, J. Gao, L. Deng, A. Acero, and L. Heck, "Learning deep structured semantic models for web search using clickthrough data," in *Proceedings of the 22Nd ACM International Conference on Conference on Information and Knowledge Management*, ser. CIKM '13. New York, NY, USA: ACM, 2013, pp. 2333–2338.

[48] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth, "The author-topic model for authors and documents," in *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, ser. UAI '04. Arlington, Virginia, United States: AUAI Press, 2004, pp. 487–494.

[49] J. Varadarajan, R. Emonet, and J.-M. Odobez, "A sparsity constraint for topic models — application to temporal activity mining," in *NIPS-2010 Workshop on Practical Applications of Sparse Modeling: Open Issues and New Directions*, 2010.

[50] L. Dietz, S. Bickel, and T. Scheffer, "Unsupervised prediction of citation influences," in *Proceedings of the 24th international conference on Machine learning*, ser. ICML '07. New York, NY, USA: ACM, 2007, pp. 233–240.

[51] S. Kataria, P. Mitra, C. Caragea, and C. L. Giles, "Context sensitive topic models for author influence in document networks," in *Proceedings of the Twenty-Second international joint conference on Artificial Intelligence — Volume 3*, ser. IJCAI'11. AAAI Press, 2011, pp. 2274–2280.

[52] D. Mimno, H. M. Wallach, J. Naradowsky, D. A. Smith, and A. McCallum, "Polylingual topic models," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2 - Volume 2*, ser. EMNLP '09. Stroudsburg, PA, USA: Association for Computational Linguistics, 2009, pp. 880–889.

[53] M. A. Dudarenko, "Regularization of multilingual topic models," *Vychisl. Metody Programm. (Numerical methods and programming)*, vol. 16, pp. 26–38, 2015.

[54] F. Sebastiani, "Machine learning in automated text categorization," *ACM Computing Surveys*, vol. 34, no. 1, pp. 1–47, 2002.

[55] K. Vorontsov, O. Frei, M. Apishev, P. Romov, M. Suvorova, and A. Yanina, "Non-bayesian additive regularization for multimodal topic modeling of large collections," in *Proceedings of the 2015 Workshop on Topic Models: Post-Processing and Applications*. New York, NY, USA: ACM, 2015, pp. 29–37.

[56] X. Wang and A. McCallum, "Topics over time: A non-markov *continuous-time model of topical trends," in Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery andData Mining*, ser. KDD '06. New York, NY, USA: ACM, 2006, pp. 424–433.

[57] S. Li, J. Li, and R. Pan, "Tag-weighted topic model for mining semi-structured documents," in *IJCAI'13 Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*. AAAI Press, 2013, pp. 2855–2861.

[58] T. Hospedales, S. Gong, and T. Xiang, "Video behaviour mining using a dynamic topic model," *International Journal of Computer Vision*, vol. 98, no. 3, pp. 303–323, 2012.

[59] E. Sokolov and L. Bogolubsky, "Topic models regularization and initialization for regression problems," in *Proceedings of the 2015 Workshop on Topic Models: Post-Processing and Applications*. New York, NY, USA: ACM, 2015, pp. 21–27.

[60] Q. Mei, D. Cai, D. Zhang, and C. Zhai, "Topic modeling with network regularization," in *Proceedings of the 17th International Conference on World Wide Web*, ser. WWW'08. New York, NY, USA: ACM, 2008, pp. 101–110.

[61] Z. Yin, L. Cao, J. Han, C. Zhai, and T. Huang, "Geographical topic discovery and comparison," in *Proceedings of the 20th international conference on World wide web*. ACM, 2011, pp. 247–256.

[62] Z. Harris, "Distributional structure," *Word*, vol. 10, no. 23, pp. 146–162, 1954.

[63] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *CoRR*, vol. abs/1310.4546, 2013.

[64] Y. Liu, Z. Liu, T.-S. Chua, and M. Sun, "Topical word embeddings," in *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, ser. AAAI'15. AAAI Press, 2015, pp. 2418–2424.

[65] A. Potapenko, A. Popov, and K. Vorontsov, "Interpretable probabilistic embeddings: bridging the gap between topic models and neural networks," in *AINL-6: Artificial Intelligence and Natural Language Conference, St. Petersburg, Russia, September 20-23, 2017*, 2017 (to appear).

[66] A. El-Kishky, Y. Song, C. Wang, C. R. Voss, and J. Han, "Scalable topical phrase mining from text corpora," *Proc. VLDB Endowment*, vol. 8, no. 3, pp. 305–316, 2014.

[67] J. Liu, J. Shang, C. Wang, X. Ren, and J. Han, "Mining quality phrases from massive text corpora," in *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, ser. SIGMOD '15. New York, NY, USA: ACM, 2015, pp. 1729–1744.

[68] J. Shang, J. Liu, M. Jiang, X. Ren, C. R. Voss, and J. Han, "Automated phrase mining from massive text corpora," *CoRR*, vol. abs/1702.04457, 2017.

[69] X. Yan, J. Guo, Y. Lan, and X. Cheng, "A biterm topic model for short texts," in *Proceedings of the 22Nd International Conference on World Wide Web*, ser. WWW '13. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 2013, pp. 1445–1456.

[70] B. Chen, "Word topic models for spoken document retrieval and transcription," vol. 8, no. 1, pp. 2:1–2:27, 2009.

[71] Y. Zuo, J. Zhao, and K. Xu, "Word network topic model: A simple but general solution for short and imbalanced texts," *Knowledge and Information Systems*, vol. 48, no. 2, pp. 379–398, 2016.

[72] R. Řehůřek and P. Sojka, "Software framework for topic modelling with large corpora," in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Valletta, Malta: ELRA, 2010, pp. 45–50.

[73] O. Frei and M. Apishev, "Parallel non-blocking deterministic algorithm for online topic modeling," in *AIST'2016, Analysis of Images, Social networks and Texts*, vol. 661. Springer International Publishing Switzerland, Communications in Computer and Information Science (CCIS), 2016, pp. 132–144.

[74] M. D. Hoffman, D. M. Blei, and F. R. Bach, "Online learning for latent Dirichlet allocation," in *NIPS*. Curran Associates, Inc., 2010, pp. 856–864.