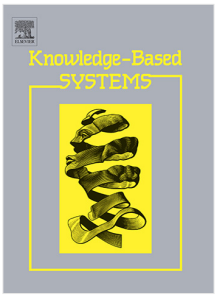
Accepted Manuscript



Experimental explorations on short text topic mining between LDA and NMF based Schemes

Yong Chen, Hui Zhang, Rui Liu, Zhiwen Ye, Jianying Lin

PII: S0950-7051(18)30407-6
DOI: https://doi.org/10.1016/j.knosys.2018.08.011
Reference: KNOSYS 4456

To appear in: *Knowledge-Based Systems*

Received date : 13 September 2017
Revised date : 8 August 2018
Accepted date : 10 August 2018

Please cite this article as: Y. Chen, et al., Experimental explorations on short text topic mining between LDA and NMF based Schemes, *Knowledge-Based Systems* (2018), https://doi.org/10.1016/j.knosys.2018.08.011

Experimental Explorations on Short Text Topic Mining between LDA and NMF based Schemes

Yong Chen[a,b,c], Hui Zhang[a,b,c], Rui Liu[a,c,*], Zhiwen Ye[a,c], Jianying Lin[a,c]

[a] *State Key Laboratory of Software Development Environment, Beihang University, Beijing 100191, P.R. China*
[b] *Beijing Advanced Innovation Center for Big Data and Brain Computing, Beihang University, Beijing 100191, P.R. China*
[c] *School of Computer Science and Engineering, Beihang University, Beijing 100191, P.R. China*

Abstract

Learning topics from short texts has become a critical and fundamental task for understanding the widely-spread streaming social messages, *e.g.*, tweets, snippets and questions/answers. Up to date, there are two distinctive topic learning schemes: generative probabilistic graphical models and geometrically linear algebra approaches, with LDA and NMF being the representative works, respectively. Since these two methods both could uncover the latent topics hidden in the unstructured short texts, some interesting doubts are coming to our minds that which one is better and why? Are there any other more effective extensions? In order to explore valuable insights between LDA and NMF based learning schemes, we comprehensively conduct a series of experiments into two parts. Specifically, the basic LDA and NMF are compared with different experimental settings on several public short text datasets in the first part which would exhibit that NMF tends to perform better than LDA; in the second part, we propose a novel model called "Knowledge-guided Non-negative Matrix Factorization for Better Short Text Topic Mining" (abbreviated as KGNMF), which leverages external knowledge as a semantic regulator with low-rank formalizations, yielding up a time-efficient algorithm. Extensive experiments are conducted on three representative corpora with currently typical short text topic models to demonstrate the effectiveness of our proposed KGNMF. Overall, learning with NMF-based schemes is another effective manner in short text topic mining in addition to the popular LDA-based paradigms.

Keywords: Short Text Mining, Topic Modeling, Latent Dirichlet Allocation (LDA), Non-negative Matrix Factorization (NMF), Knowledge-based Learning

1. Introduction

With the unprecedented development of Web 2.0, especially the real-time interactive social media, more and more short texts are flooded in our daily life, such as

*Rui Liu is the corresponding author. Email: lr@buaa.edu.cn

snippets in the search engine, social comments in Facebook/Tweet/WeChat, question-answer dialogues, just to name a few. How to drive the machine to understand the short texts and execute intelligent tasks *e.g.*, short text classification and clustering [1, 2], sentiment analysis [3], advertisement delivery [4], document summarization [5], web event mining [6, 7] and outlier detection [8], is significantly important but quite challenging in practical applications. Topic model, as a potential promising tool, holds the ability to mine the underlying structures hidden in the unstructured texts automatically and meanwhile embed each text unit into the latent semantic space, which would further benefit a battery of the aforementioned missions.

When it comes to the topic models, the most popular methods are to the probabilistic graphical models, in which topics are usually viewed as distributions over words and documents are treated to share the underlying topics with different proportions from the perspective of probability [9, 10]. Latent Dirichlet Allocation (LDA) [11] is such a typical example. Particularly, LDA models documents with fully Bayesian paradigm, and therefore is preferred for adjustable priors to closely refer to the nature of specific documents, topics and terms. LDA obtains the final outcomes of topic-word and document-topic distributions through a posterior maximization with Gibbs sampling. Notably, this kind of unsupervised statistical learning is generally considered to perform quite well only on the condition that the corpus is statistically sufficient.

In contrast, non-negative matrix factorization (NMF) [12], a distinctive topic mining method from the former probabilistic viewpoints, models the underlying components as coordinate axes and each document corresponds to a unique point in the latent linear space with a geometric perspective. More specifically, the text archive is usually firstly encoded in a term-document matrix D with TF-IDF weights and then two non-negative matrices: term-topic U and topic-document V are sought with algorithms such as Multiplicative Update (MU) [13], in which each column of U can be viewed as a topic, and each column of V can be treated as a compact embedding in the latent topic space. Remarkably, with non-negative constraints, NMF holds the special abilities to learn from data like human's cognition [12]: additive local parts constitute the whole object, which greatly enhances the interpretations for various practical scenarios.

Since the above two different models LDA and NMF could both be utilized to learn the topics behind the text collections, the questions naturally come up as below:

- Question 1: Which one (LDA or NMF) performs better on short text topic mining?
- Question 2: Why does it yield up better topics in short texts?
- Question 3: Are there any improved models that better fit short text topic mining? If not, how to propose a better one?

Motivated by such doubts, we carried out a series of experiments to comprehensively and systematically compare these two distinctive methods along with their extensions. To be more concrete, we sort out our contributions corresponding to the questions mentioned above as follows:

- Answer 1: In terms of short text topic mining, we select several publicly acceptable datasets with average length ranging from 3.46 to 14.34 terms per document

and conduct experiments with LDA and NMF, and the topic quality of the final results demonstrates that NMF is inclined to produce better topics than LDA.

- Answer 2: Short texts are classically noisy and sparse, and therefore lack sufficient information for effective statistical learning *e.g.*, LDA. However, NMF, definitely, also needs adequate information for satisfying outcomes; yet compared to LDA, NMF indeed contains much more priors such as TF-IDF (term frequency and inverse document frequency) encodings for texts (instead of the only TF in LDA), Gaussian distribution for the noises with Frobenius norm, and implicit low-rank structures w.r.t. the original term-document datasets (while LDA doesn't have this); besides, the deterministic MU [13] algorithm in NMF also contributes to more stable and better results than the stochastic Gibbs sampling without enough word co-occurrences in LDA.

- Answer 3: Due to the insufficient word co-occurrences in short texts, various priors are introduced into the "document-topic-term" hierarchical graphical models for extended LDA-based versions such as BTM [16], DMM [18], LF-LDA [20], LF-DMM [20], GPU-DMM [21] and GKLDA [29]. As will be illustrated in our experiments on short text corpus, such modified methods indeed improve a lot in regard to topic quality for the learned topics. However, for NMF, to the best of our knowledge, there has been few representative work in short text topic mining, and then we come up with a novel approach named "Knowledge-guided Non-negative Matrix Factorization for Better Short Text Topic Mining", dubbed KGNMF, in which the word-word similarities are incorporated as Graph Laplacian with low-rank forms leading to a fast efficient algorithm for topic learning. Extensive experiments showcase that our proposed KGNMF performs better than or at least comparable with the competing baselines in topics.

The remainder of this paper is organized as follows. In Sect. 2, we introduce some related work about LDA and NMF, as well as their extensions. In Sect. 3, we present our model and some analysis in details. In Sect. 4, we conduct a series of experimental explorations to proceed on the comparisons between LDA and NMF based schemes for the topic learning of short texts, and finally draw some conclusions and talk about some future explorations in Sect. 5.

2. Background

With the aim to comprehensively compare LDA and NMF based schemes for short text topic mining in this paper, we firstly introduce the basic LDA and NMF models; then two representative extended probabilistic methods DMM and BTM are shown in details; next, the well-performed paradigms for understanding short texts by exploiting external knowledge are also summarized with typical examples; As most methods are probabilistic hierarchical graphical models, we finally deliberately design an NMF-based topic modeling called KGNMF for short texts topic mining with word-word semantic regularizations for fair and overall comparisons.

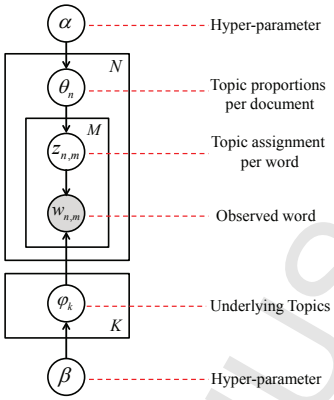


Figure 1: Hierarchical graphical model for LDA. The boxes are "plates" representing replicates. The upper outer plate represents documents, while the upper inner plate denotes the repeated choice of topics and words within a document; besides, the lower plate marks the latent topics hidden in the document collection.

2.1. Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) [9, 11] is a generative probabilistic model for a given text collection. Basically, there is an assumption that K latent topics are hidden in the given N documents corpus, and each topic is represented as a multinomial distribution over the M words in the vocabulary extracted from the above text collection. In terms of each document, it is generated by sampling a mixture of these topics and then sampling words from that mixture. More precisely, the generative process for each document in the text archive is illustrated as follows:

1. For the n-th ($n = 1, 2, \cdots, N$) document d in the whole N document-corpus, choose $\theta_n \sim Dirichlet(\alpha)$;
2. For each word $w_{n,m}$ in the document d:
   (a) Choose topic assignment $z_{n,m} \sim Multinomial(\theta_n)$;
   (b) Find the corresponding topic distribution $\varphi_{z_{n,m}} \sim Dirichlet(\beta)$;
   (c) Sample a word $w_{n,m} \sim Multinomial(\varphi_{z_{n,m}})$.

The above steps elaborate the generative process for each document and for the whole N documents in Corpus D, just repeat the procedures for N times with each corresponding to one document, which is plotted in graphical model language with "boxes" and "circles" as Fig. 1 (α and β are two Dirichlet-prior hyper-parameters). Obviously, we could easily arrive at the probability of the D corpus:

$$P(D|\alpha, \beta) = \prod_{n=1}^{N} \int P(\theta_n|\alpha) F(\theta, \varphi) d\theta_n, \tag{1}$$

where $F(\theta, \varphi) = \prod_{m=1}^{M} \sum_{z_{n,m}} P(z_{n,m}|\theta_n) P(w_{n,m}|\varphi_{z_{n,m}})$.

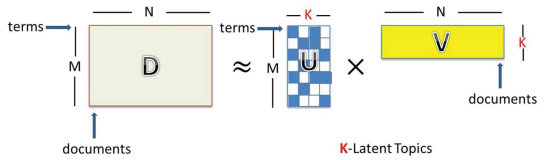


Figure 2: Non-negative matrix factorization: $D \approx UV$, with U and V elementwise non-negative. More specifically, the encoded TF-IDF term-document matrix (sized by $M \times N$) for a given text corpus is decomposed into two matrices: term-topic matrix U and topic-document matrix V, corresponding to K coordinate axes and N points (each point represents one document) in a new semantic space, respectively.

Finally, the hidden unobserved random variables: topic-word distributions φ and document-topic distributions θ could be learned through Gibbs sampling and variational EM algorithm via maximizing the probability $P(D|\alpha, \beta)$.

2.2. Non-negative Matrix Factorization

Non-negative Matrix Factorization (NMF) [12] is a linear algebra method that embeds the original high-dimensional data into a low semantic space with non-negative hidden structures, which are viewed as coordinate axes in the transformed space with geometric perspectives. Particularly, each one d_n of N documents in a given corpus is supposed to be linearly combined by K components/axes u_i with coefficients v_{kn}, where $k = 1, 2, \cdots, K$ and $n = 1, 2, \cdots, N$, *i.e.*, $d_n = \sum_{k=1}^{K} u_k v_{kn}$. The goodness of this formalization is evaluated by the square loss between the original term-document vector d_n and the linear combinations $\sum_{k=1}^{K} u_k v_{kn}$. Therefore, there exists:

$$\min \sum_{n=1}^{N} ||d_n - \sum_{k=1}^{K} u_k v_{kn}||_2^2 \tag{2}$$

$$s.t. \begin{cases} v_{kn} \geq 0, \\ u_{mk} \geq 0. \end{cases},$$

where $||\cdot||_2^2$ represents the square of a vector's L2 norm. Denote $D = [d_1, d_2, \cdots, d_n] \in R^{M \times N}$, $U = [u_1, u_2, \cdots, u_K] \in R^{M \times K}$ and $V = [v_1, v_2, \cdots, v_N] \in R^{K \times N}$, the above optimization can be transformed into a compact version as below:

$$\min \ ||D - UV||_F^2 \tag{3}$$

$$s.t. \begin{cases} V \geq 0, \\ U \geq 0. \end{cases},$$

where $||\cdot||_F^2$ marks the square of a matrix's Frobenius norm and this is illustrated as Fig. 2. Noticeably, the elementwise non-negative for matrices U and V makes the learning scheme interpretable; and more importantly, it reveals some learning philosophy as humans: parts constitute the whole. Therefore, NMF holds the ability to mine underlying topics for a given text collection.

As to the learning algorithm, Multiplicative Update [13] is designed as follows:

$$U \leftarrow U \frac{DV^T}{UVV^T}, \tag{4}$$

$$V \leftarrow V \frac{U^T D}{U^T UV}. \tag{5}$$

Iterate the above two equations until convergence and we could achieve the final term-topic matrix U and topic-document matrix V for topics mining and data representations.

2.3. Dirichlet Multinomial Mixture

Classic topic models presume that each document talks about several topics and thus treat each document as a distribution over topics; while for the short texts, it is reasonable to assume that each document is mainly on one topic. Based on such cognition, Dirichlet Mixture Model (DMM) [18, 19] is proposed to sample each word in a given specific document independently with the same topic distribution instead of the one word from one topic distribution. As to the probabilistic generations for the whole dataset, it can be summarized as follows:

1. Draw a topic distribution $\theta \sim Dirichlet(\alpha)$;
2. For each topic z:
   (a) Draw a topic-specific word distribution $\varphi_z \sim Dirichlet(\beta)$;
3. For each document d in text collection D:
   (a) Draw a topic assignment $z \sim Multinomial(\theta)$;
   (b) For each word w in document d:
      (i) Sample a word $w \sim Multinomial(\varphi_z)$.

In light of the generative procedure, we translate them into a graphical model as Fig. 3, in which α and β are two Dirichlet-priors which should be provided in advance. The likelihood for the whole collection is exhibited as:

$$P(D) = \prod_{d \in D} \left(\sum_z \left(P(z) \prod_{w \in d} P(w|\varphi_z) \right) \right), \tag{6}$$

where the latent variables in the generative process can be approximated by Gibbs sampling. Generally, this kind of learning paradigm is supposed to perform much better than traditional LDA with the assumption "one document one topic", which narrows the sampling for a word more accurately in short texts.

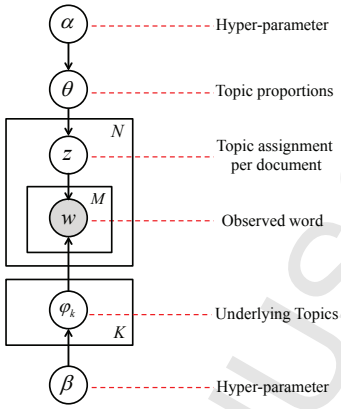


Figure 3: Hierarchical graphical model for DMM.

2.4. Biterm Topic Model

Conventional topic models explicitly describe the generate process from documents, topics to words, which implicitly capture the document-level word co-occurrences to bring the latent topics to light, and definitely would suffer severe sparsity in short documents. Biterm topic model (BTM) [16, 17] is a novel scheme to model the biterms (word-word co-occurrence patterns) in text collections instead of words for documents, which alleviate the document-level word sparsity via the biterms aggregated in the whole corpus. More specifically, for each of the $|B|$ biterms collected in the given texts, it generates as the following procedures:

1. Draw a topic distribution $\theta \sim Dirichlet(\alpha)$ for the whole document corpus;
2. For each topic z:
   (a) Draw a topic-specific word distribution $\varphi_z \sim Dirichlet(\beta)$;
3. For each biterm $b = (w_i, w_j)$ in the set B:
   (a) Draw a topic assignment $z \sim Multinomial(\theta)$;
   (b) Draw words $w_i, w_j \sim Multinomial(\varphi_z)$.

With respect to the modeling process, we illustrate it with graphical language in Fig. 4. Then the probability for each biterm b can be easily transformed as below:

$$P(b) = \sum_{z} P(z)P(w_i|z)P(w_j|z), \tag{7}$$

Therefore, the whole corpus's likelihood is:

$$P(B) = \prod_{b \in B} P(b) = \prod_{(i,j)} \sum_{z} P(z)P(w_i|z)P(w_j|z). \tag{8}$$

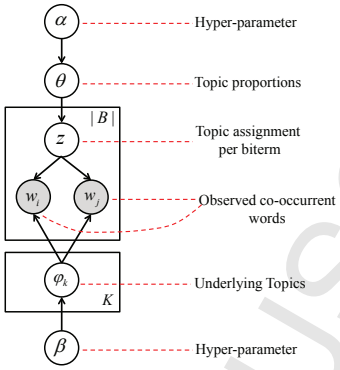


Figure 4: Hierarchical graphical model for BTM.

Similar to LDA, the parameters θ and φ are estimated with Gibbs Sampling. Although this kind of learning scheme for topics mining is reported to perform better than LDA with the reasons that biterms aggregated in the whole collection could alleviate the data sparsity, yet the texts only will bound its abilities in statistics. How to leverage knowledge from external resources as humans understand short text is quite critical for machine comprehension [14, 15].

2.5. Improving Short Text Topic Modeling with External Knowledge

Return to the fact that the short texts are suffering from severe sparsity and therefore it is significantly challenging for machine to understand the topics hidden in the unstructured data. Compared to the comprehension of human, we usually associate "apple" with "iPhone" via our brain full of knowledge when faced with insufficient information. Therefore, it is natural to let the machine learn from short texts for more accurate topics by exploiting external knowledge. Then, how to make use of external knowledge such as Wikipedia[1], BaiduBaike[2], Freebase[3], to name a few, is critical. Nguyen et al. [20] propose a method to leverage latent feature vector representations of words trained on very large corpora by Word2vec [22, 23] into the LDA and DMM models. In particular, they incorporate the latent features into the topic-to-word Dirichlet-Multinomial component with a two component mixture of a Dirichlet-Multinomial and a latent feature component. It is demonstrated that such learning models (LFLDA, LFDMM) present better topics than LDA and DMM. Besides, Li et al. [21] also focus on the external auxiliary word-word semantics for enriching short text topic modeling and propose a novel approach based on DMM. More specifically, their method is mainly to promote the semantically related words under the same topic during the Gibbs sampling process by using generalized Polya urn (GPU) model [24, 25]. This GPU-DMM method combines DMM and external knowledge simulta-

[1] https://www.wikipedia.org/
[2] https://baike.baidu.com/
[3] https://datahub.io/dataset/freebase

neously describing the short texts more accurately and enriching semantic relatedness for yielding high-quality topics. Other similar works are Refs. [26, 27, 28, 29], which all make use of external knowledge into the graphical models for better short text topic mining.

The aforementioned methods are mostly based on generative probabilistic models, and in order to make comprehensive comparisons between LDA and NMF based schemes, we propose a novel method named "Knowledge-guided Non-negative Matrix Factorization for Better Short Text Topic Mining" (KGNMF), which leverages the word-word semantics as graph laplacian regularization with low-rank formalizations and therefore is quite efficient in learning.

3. Knowledge-guided Non-negative Matrix Factorization for Better Short Text Topic Mining

3.1. Problem Statement

Given a set of documents with size N and each document is marked d_n, ($n = 1, 2, ..., N$); Suppose that we represent each document over a vocabulary with M distinctive words extracted from the document dataset in advance, then each element of d_n could be computed in TF-IDF weight denoted as d_{mn}, ($m = 1, 2, ..., M$), where the larger value d_{mn} is, the more important the m-th word is in the n-th document. Based on the above analysis, the whole text collection could be encoded in a matrix D with size $M \times N$ and each element is d_{mn}, *i.e.*, $D = [d_1, d_2, \cdots, d_N] \in R^{M \times N}$. The main problem to understand short texts is to find out the underlying topics hidden in the unstructured documents. Naturally, we could presume that each document talks about several topics and each topic denoted u_k ($k = 1, 2, ..., K$) is perceived as a liner combination of the M terms in the vocabulary, where the larger the value of u_{mk} is, the more related the word is to the topic. Then the core challenge is to seek the latent topics $U = [u_1, u_2, \cdots, u_K] \in R^{M \times K}$ and the encodings $V = [v_1, v_2, \cdots, v_N] \in R^{K \times N}$ with each v_n corresponding to a new reduced representation for the n-th document d_n.

However, in light of the short text topic mining, the document dataset solely would not be fully beneficial for traditional models (*e.g.*, NMF) due to the insufficient information. Therefore, external knowledge is supposed to be leveraged to supervise the conventional models for more accurate topics in the short-text background, which is following the human pattern of understanding short texts [31, 32]. In a nutshell, how to learn high-quality topics from short texts under the guidance of external knowledge is the core problem, which will be modeled in the following NMF-based framework.

3.2. Modeling: KGNMF

As mentioned in the former section, each topic is treated as a linear combination over terms in the vocabulary and each document is linearly combined with K different topics. The goodness between the original text corpus D and the latent topics U and new embedding V is measured via square loss, which is formalized as the optimization problem (2) with non-negative constraint for each element. This is the basic NMF for modeling the texts for mining topics, but it is not enough for short texts.

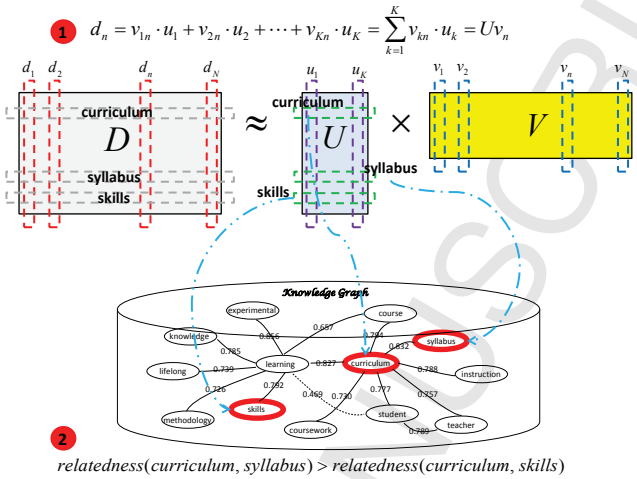


Figure 5: **KGNMF** Framework: Knowledge-guided Non-negative Matrix Factorization for Better Short Text Topic Mining.

We then further leverage external knowledge in forms of word-word pairwise semantics to supervise NMF to better comprehend short texts and propose a novel method named "Knowledge-guided NMF for Better Short Text Topic Mining", abbreviated as KGNMF, which is illustrated in Fig. 5. As can be seen clearly from the designed Framework, the upper graph is the illustration for the basic NMF with column-perspectives. In particular, each column of D represents a document, each column of U denotes a topic and each column of V is a new reduced embedding for the corresponding document in the latent semantic space. Then we conclude the basic NMF for the approximation between D and U, V.

However, on the other way around, if we turn to each row of the topic matrix U, then each row corresponds to a term over topic space; A better learning should not only preserve the underlying information in corpus, but also keep the relatedness between different word pairs; for example, the relatedness between "curriculum" and "syllabus" is much closer than the relatedness between "curriculum" and "skills". Therefore, we integrate the word-word semantic graph regularization which can be learned from external knowledge base *e.g.*, Wikipedia into the basic NMF for improvement in short text topic learning.

Formally, denote s_{ij} as the similarity between word w_i and word w_j learned from external knowledge. The similarity between each two rows of U, with each representing one word, should be consistent with the similarity from external knowledge. Specifically, we can make the following minimizations:

$$\min \frac{1}{2}\sum_{i=1}^{M}\sum_{j=1}^{M} s_{ij}||u_{i*} - u_{j*}||_2^2 = tr(U^T L U), \tag{9}$$

where u_{i*} represents the i-th row of topic matrix U, $L = diag(S \cdot \mathbb{1}) - S$, $\mathbb{1}$ denotes a column vector with each element as 1, and $S = (s_{ij})$.

Taking the corpus itself and the external knowledge into consideration, we could combine the basic NMF with word-word semantics regularization and then draw the final whole optimization as below:

$$\min \ ||D - UV||_F^2 + \lambda \cdot tr(U^T LU) \tag{10}$$
$$s.t. \begin{cases} U \geq 0, \\ V \geq 0. \end{cases},$$

where λ is a non-negative hyper-parameter to balance the importance between the information in document dataset and the external knowledge.

Note that the dimension of graph laplacian matrix L is $M \times M$, which is so large when the vocabulary is plentiful in words. This is an obstacle for designing a fast algorithm, which will be shown in the next part with a low rank formalization to represent S for efficient algorithms.

3.3. Algorithm Derivations

With respect to the optimization problem (10), it is non-convex and therefore difficult to achieve a global solution. However, we can seek for a local minimal for the practical applications. Specifically, fix one variable and optimize the other variable, then the problem is convex and we can get the unique solution for the sub-problem [38, 39]. Iterate these two sub-problems until convergence, and finally the local optimal solution could be achieved. Furthermore, let

$$J = ||D - UV||_F^2 + \lambda \cdot tr(U^T LU). \tag{11}$$

Fix U and optimize V, the objective J is equivalent to

$$J = ||D - UV||_F^2, \tag{12}$$

then the derivative of J to V can be written as:

$$\frac{\partial J}{\partial V} = 2U^T UV - 2U^T D. \tag{13}$$

In order to update V, we adopt the gradient descent method for each (ij)-th variable of V as below:

$$V_{ij} \leftarrow V_{ij} - \alpha_{V_{ij}} \frac{\partial J}{\partial V_{ij}}, \tag{14}$$

where $\alpha_{V_{ij}}$ is the iterative stepsize for updating the variable V_{ij}. Here let $\alpha_{V_{ij}} = \frac{V_{ij}}{[2U^T UV]_{ij}}$, then there holds:

$$V_{ij} \leftarrow V_{ij} - \alpha_{V_{ij}} [2U^T UV - 2U^T D]_{ij}, \tag{15}$$

$$V_{ij} \leftarrow V_{ij} - \frac{V_{ij}}{[2U^T UV]_{ij}} [2U^T UV]_{ij} + \frac{V_{ij}}{[2U^T UV]_{ij}} [2U^T D]_{ij}, \tag{16}$$

$$V_{ij} \leftarrow V_{ij} \frac{[U^T D]_{ij}}{[U^T U V]_{ij}}, \tag{17}$$

$$V \leftarrow V \odot (U^T D) \oslash (U^T U V), \tag{18}$$

where the matrix multiplication $\odot$ and division $\oslash$ are elementwise operations. Remarkably, this kind of derivations naturally preserves the non-negative constraints, which is the outstanding property for such philosophy with proper iterative stepsize.

Fix V and optimize U, the objective J is transformed to:

$$J = ||D - UV||_F^2 + \lambda \cdot tr(U^T L U), \tag{19}$$

then the derivative of J to U is

$$\frac{\partial J}{\partial U} = 2UVV^T - 2DV^T + 2\lambda \cdot diag(S \cdot \mathbb{1}) \cdot U - 2\lambda \cdot S \cdot U. \tag{20}$$

Similar to the update for V, the iterative formula for U can be drawn as below:

$$U_{ij} \leftarrow U_{ij} - \alpha_{U_{ij}} \frac{\partial J}{\partial U_{ij}}, \tag{21}$$

Set $\alpha_{U_{ij}} = \frac{U_{ij}}{[2UVV^T + 2\lambda \cdot diag(S \cdot \mathbb{1}) \cdot U]_{ij}}$, then there exists:

$$U \leftarrow U \odot (DV^T + \lambda \cdot S \cdot U) \oslash (UVV^T + \lambda \cdot diag(S \cdot \mathbb{1}) \cdot U). \tag{22}$$

Note that the term $S \cdot U$ in the iterative Eq. (22), the complexity is $O(KM^2)$, which is not tolerant when the vocabulary is too large. Luckily, we can represent S with low-rank formalization. To be specific, assume that each word is embedded as a column vector w sized by Q-dimension trained from external knowledge (Wikipedia) with word2vec tool[4]. Then for each two words marked w_i and w_j, the similarity could be defined as:

$$s_{ij} = \cos(w_i, w_j) = \frac{< w_i, w_j >}{||w_i|| \cdot ||w_j||}, \tag{23}$$

where $< w_i, w_j >= w_i^T w_j$. If each word-vector is normalize to 1, then the similarity $s_{ij} = w_i^T w_j$ and S could be transformed into the following low-rank formalizations:

$$S = WW^T, \tag{24}$$

where $W = [w_1, w_2, \cdots, w_M]^T \in R^{M \times Q}$. Replace S with Eq. (24) into Eq. (22), there holds:

$$U \leftarrow U \odot \{DV^T + \lambda \cdot W \cdot (W^T \cdot U)\} \oslash \{UVV^T + \lambda \cdot diag(W \cdot (W^T \cdot \mathbb{1})) \cdot U\}, \tag{25}$$

then the complexity of $S \cdot U$ is reduced to $O(MKQ)$ instead of $O(KM^2)$, which is much more efficient in computing because Q and K is usually much smaller than M.

Here we can summarize our algorithm for KGNMF in Algorithm 1.

[4] https://code.google.com/archive/p/word2vec/

Algorithm 1: KGNMF

Input: TF-IDF matrix D for document collection, topic number K, and word vectors W; Hyper-parameter: λ.

Output: word-topic matrix U and topic-document matrix V.

```
begin
    Randomly initialize U and V;
    for t=1,2,···,T do
        Update V according to iterative Eq. (18);
        Update U according to iterative Eq. (25);
        if convergence is satisfied then
            break;
        end
    end
end
```

3.4. Computational Complexity Analysis

As can be seen clearly from Algorithm 1, the main steps that consume much time are to update U and V iteratively. Particularly, for the iterative formula in Eq. (18), the complexity is $O(KMN + (M+N)K^2 + 2NK)$; and for the other iterative Eq. (25), the complexity is $O(KMN + 2KMQ + (M+N)K^2 + 2MK)$. Recall that K (number of topics) and Q (dimension of embedded word from external knowledge via word2vec tool, usually $Q = 100$) are much smaller than M and N, corresponding to the number of terms and the number of documents in the corpus respectively, the overall complexity is $O(TKMN)$ for the entire algorithm, where the iterative times T are set to 100 in our experiments.

4. Experiments

Generally, we use several public datasets for comprehensively comparative studies between LDA and NMF based schemes for short text topic mining. All the experiments are conducted on a workstation equipped with Intel(R) Xeon(R) CPU E5-2680 v3@2.50GHz, 24 Cores and 128G memory.

4.1. Datasets

Snippet[5] is composed by about 12K snippets drawn from Google, which has been created by Phan et al. [30]. Each snippet is on each line and each one consists of a list of words/terms plus a class label at the end in the data files. Overall, there are 8 categories marked "Business", "Computers", "Culture-Arts-Entertainment", "Education-Science", "Engineering", "Health", "Politics-Society", and "Sports".

[5] http://acube.di.unipi.it/tmn-dataset/

Table 1: Statistics of datasets. (#Docs: the total number of documents; #Words: the average number of words per document; #Vocabulary: the total number of distinctive terms in the dataset; #Label: the number of ground-truth labels or categories.)

Datasets	#Docs	#Words	#Vocabulary	#Label
Snippet	12,284	14.34	4,045	8
News	32,592	11.82	7,771	7
StackOverFlow	19,783	7.25	2,437	20
XinlangNews	8,966	4.38	2,631	6
TMNtitles	29,487	3.46	4,015	7

XinlangNews[6] is selected from the crawled original $14,664$ pieces of news from 8/23/2016 to 9/7/2016 on Sina website and then is composed by about 9K news titles after filtering too short texts with less than 2 words/title from 6 domains, *i.e.*, "entertainment" $(1,584)$, "Finance" $(2,041)$, "Sci_Tech" (949), "Society" (727), "Sports" $(3,144)$ and "Military" (521).

News[5] is a dataset of about 32K English news extracted from RSS feeds of popular newspaper websites (nyt.com, usatoday.com, reuters.com). Categories are: "Sport", "Business", "U.S.", "Health", "Sci_Tech", "World" and "Entertainment". In our experiment, we choose the title and description fields as the **News** short texts. In contrast, the titles are solely selected for a new short text collection called **TMNtitles**[5].

StackOverFlow[7] [35] is a short text dataset collected from an online question-and-answer site (StackOverflow.com) for programmers. It mainly talks about computer languages or development tools such as scala, java, matlab, spring, oracle, just to name a few (20 different classes in total).

For all the five aforementioned short-text datasets, we use NLPIR[8] and NLTK[9] tool to split the texts into terms, and filter the short texts that has less than two terms; besides, we also discard the terms, of which the IDF (Inverse Document Frequency) is less than 5, then the short texts are ready with each line corresponding to a sequence of words. Finally, we proceed on some statistical analysis as Table 1. Remarkably, the average length of each text collection ranges from 3.46 to 14.34, among which the **TMNtitles** dataset is the shortest while **Snippet** is the longest. Overall, they are all treated as short texts. By the way, for the LDA-based methods, the prepared short texts are ready for input; while for the NMF-based approaches, TF-IDF matrix is further computed for document datasets.

4.2. Experimental Setups

4.2.1. Baseline Models

Various state-of-the-art LDA-based and NMF-based topic models are adopted here for comparative studies.

[6] http://news.sina.com.cn/
[7] https://github.com/jacoxu/StackOverflow
[8] http://ictclas.nlpir.org/
[9] https://github.com/nltk/nltk

The basic topic models for short text topic mining are LDA [10] and NMF [12]. We set the number of topics to 20, 40, 60, 80 and 100, respectively, for both methods. The hyper parameters α and β in LDA are set to 1 and 0.01, respectively. With respect to the other parameters, they are set by default as in the lda-c[10]. In terms of the NMF for short text topic mining, the only parameter is the iterative times and we set it to 100 in our experiments.

For further comparisons between LDA-extended and NMF-extended methods, we select 6 other comparative baselines, *i.e.*, BTM[11] [16], LFLDA[12] [20], LFDMM[12] [20], GPUDMM[13] [21], GKLDA[14] [29] and our proposed KGNMF. For all the LDA-based baselines, the codes are publicly available and the parameters are set according to the corresponding papers. While for our NMF-based methods, we implement them with python codes. In our KGNMF, we set $\lambda = 5$ in Snippet, XinlangNews and Twitter datasets according to a series of experiments with different parameter settings.

4.2.2. Evaluation Protocols

PMI Score: Up to present, there are several automatic metrics to evaluate the topic quality such as perplexity [9], topic coherence [36, 37], and PMI scores [34]. However, as shown in [33], perplexity does not reflect the semantic coherence of a topic. It can sometimes be contrary to human judgments. In terms of the topic coherence defined in Ref. [36], it's reported to be consistent with Experts' annotations; since this kind of human evaluation is adopted in our paper, then we turn to another popular PMI score [17, 34] as the automatic metric to measure the quality of the models' topics for a more comprehensive assessment. Its specific formulation is as follows:

$$PMI(t; V^{(t)}) = \frac{2}{M(M-1)} \sum_{m=2}^{M} \sum_{l=1}^{m-1} \log \left(\frac{p(v_m^{(t)}, v_l^{(t)}) + 1}{p(v_m^{(t)}) p(v_l^{(t)})} \right), \tag{26}$$

where $V^{(t)} = (v_1^{(t)}, v_2^{(t)}, \cdots, v_M^{(t)})$ is a list of the M most probable words in topic t, and $v_m^{(t)}$, $v_l^{(t)}$ represents the m^{th} and l^{th} term of the specific topic t respectively. $p(v_m^{(t)}, v_l^{(t)})$ denotes the probability that words $v_m^{(t)}$ and $v_l^{(t)}$ appear in the same document, while $p(v_m^{(t)})$ or $p(v_l^{(t)})$ marks the probability that the m^{th} or l^{th} term occurs in the document corpus. A smoothing count of 1 is included to avoid taking the logarithm of zero. Generally speaking, the more the co-occurrence under the same topic it owns, the larger value the PMI score will be, and the better performance the mined topic it has. With respect to the K topics, the final average PMI score is then computed as the topic model's performance in topic-quality.

Expert Evaluation: Here we also evaluate the quality of learned topics based on human judgements. Two experts who are well familiar with the above short texts are invited to label the generated topics. Specifically, two tasks **"Topic Labeling"** and

[10] https://github.com/blei-lab/lda-c
[11] https://github.com/xiaohuiyan/BTM
[12] https://github.com/datquocnguyen/LFTM
[13] https://github.com/NobodyWHU/GPUDMM
[14] https://github.com/czyuan/GKLDA

Table 2: Average coherent topics evaluated by experts given topic number $K = 20$ for different short text datasets between NMF and LDA methods.

Datasets \ Methods	NMF	LDA
Snippet	**18**	**17**
XinlangNews	**20**	**19**
News	**19**	**18**
StackOverFlow	**19**	**19**
TMNtitles	**19**	**18**

"Word Labeling" are conducted by the two judges. For the first task "Topic Labeling", the experts are asked to label "coherent" if more than half of the topN (*e.g.*, topN=20) terms in a topic are semantically related; otherwise "not coherent". Then for the second task "Word Labeling", the topics that are labeled as "coherent" by both judges are used for word labeling. Particularly, each topical word is labeled as "correct" if it is coherently related to the concept represented by the topic (identified in the "Topic Labeling"); otherwise "incorrect".

After labeling, we can compute the number of the coherent topics for different methods and the average Precision for the first topN topical words (abbreviated as P@topN [40]). Finally, the overall performances are averaged by the evaluation measures from two experts.

4.3. Basic Experiments between LDA and NMF on Public Short Texts

LDA and NMF could both automatically learn topics from short texts, but they are in different manners to model texts, *i.e.*, corresponding to probabilistic generative process and geometrically linear combinations, respectively. Therefore, it is quite interesting to compare them and find out which one is better. For solidly experimental comparisons, we deliberately choose 5 public short-text datasets with different configurations as illustrated in Table 1, mainly sufficiently exhibiting the outcomes and inducing some general conclusions. As presented in Fig. 6$\sim$10, the first five subfigures are talking about the average PMI scores for the topN (5, 10, 15, 20) topical words with different topic numbers (20, 40, 60, 80, 100) corresponding to Snippet, XinlnagNews, News, StackOverFlow and TMNtitles, respectively. From these experimental results, we could easily find that the PMI values of NMF are larger than that of LDA almost on all the designed explorations, which solidly conveys a message that NMF probably performs much better than LDA under short text topic learning.

Furthermore, we also invite two experts to label the topics and their representative words for the 5 datasets; and then compute the average coherent topics as well as the precision for the topN topical terms; finally, we collected the results and visualize them in Table 2 and Fig. 6(f)$\sim$10(f). Obviously, NMF shows similar advantages over LDA on P@topN topical words, which means that NMF learned much higher-quality representative terms for the "coherent" topics than LDA. As to the number of "coherent" topics among all the mined 20 topics, NMF is still superior to LDA. Note that the av-

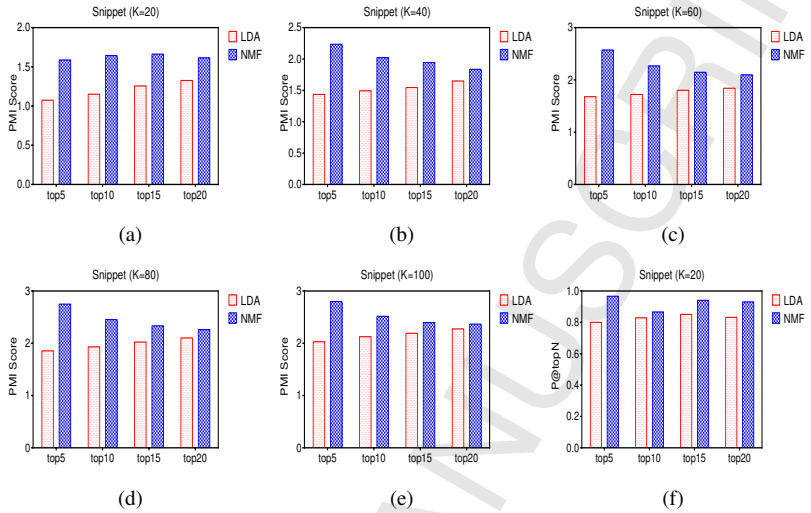


Figure 6: Topic Evaluations on **Snippet** Dataset: (1) from (a) to (e), PMI score with different numbers of topics for topN (topN=5, 10, 15 and 20) terms; (2) (f), human judgements with precision for topN (topN=5, 10, 15 and 20) terms (P@topN for short).

erage Cohen's Kappa score[15] for the topic labeling on the above experiments (topic number $K = 20$ in Table 2) is 0.9445, which indicates very good agreements between the two judges according to scale in [41].

Briefly, no matter what kind of measures for topic quality (automatic metric and expert evaluation), NMF always exhibits overwhelming advantages over LDA, which leads us to draw some analysis: Short texts are short and sparse, and thus are typically lack of information about word-word co-occurrences, which is definitely not beneficial for statistical topic models, *i.e.*, LDA. More specifically, the core method adopted in LDA is stochastic Gibbs sampling for each word in texts which would bring large variances in learning and inference especially for sparse and short texts. However, NMF firstly encoded the whole corpus with TF-IDF weights, not only considering the frequency (basic statistics, TF), but also inducing discriminative information (IDF) for each term. Moreover, NMF leverages multiplicative update algorithm, a definite learning scheme for approximating the original texts' information. In conclusion, more information encoded and definite algorithms probably make NMF tend to produce higher-quality topics than LDA (which leverages stochastic Gibbs sampling without enough word co-occurrences for learning and inference) from short texts.

4.4. Further Comparisons between LDA-based and NMF-based Methods for Short Text Topic Mining

To further explore the performance between LDA-based and NMF-based in short text topic mining, we select two representative datasets Snippet and XinlangNews, with

[15] `http://www.pmean.com/definitions/kappa.htm`

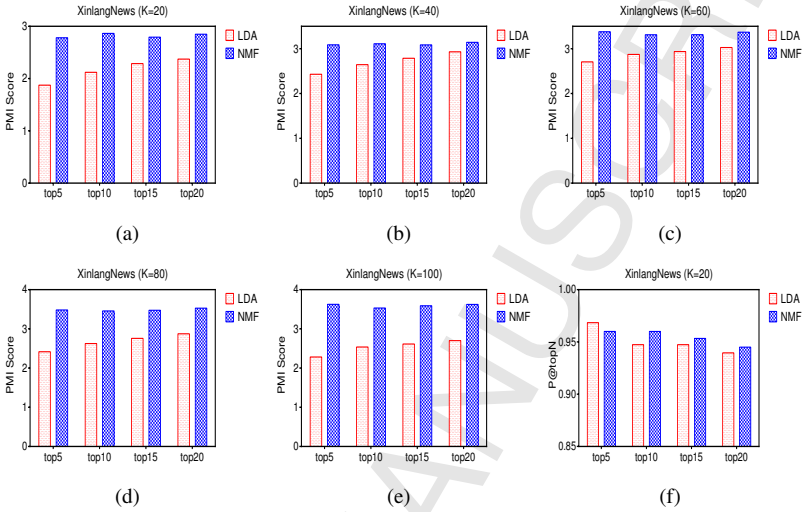


Figure 7: Topic Evaluations on **XinlangNews** Dataset: (1) from (a) to (e), PMI score with different numbers of topics for topN (topN=5, 10, 15 and 20) terms; (2) (f), human judgements with precision for topN (topN=5, 10, 15 and 20) terms (P@topN for short).

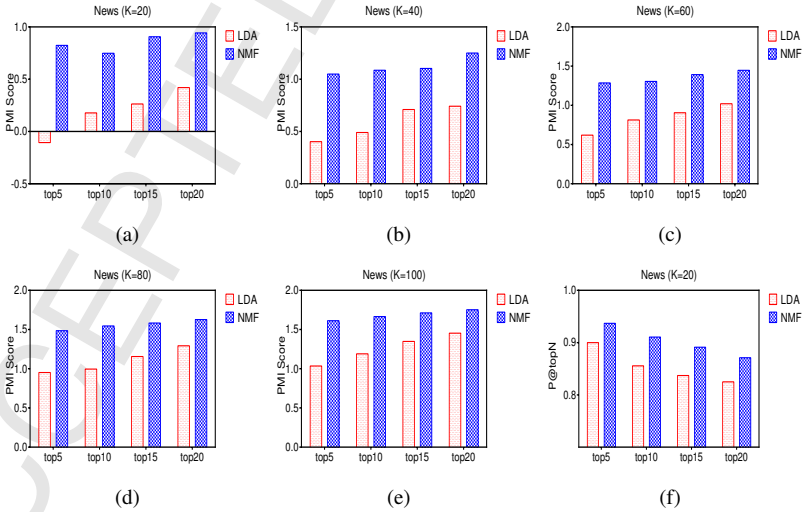


Figure 8: Topic Evaluations on **News** Dataset: (1) from (a) to (e), PMI score with different numbers of topics for topN (topN=5, 10, 15 and 20) terms; (2) (f), human judgements with precision for topN (topN=5, 10, 15 and 20) terms (P@topN for short).

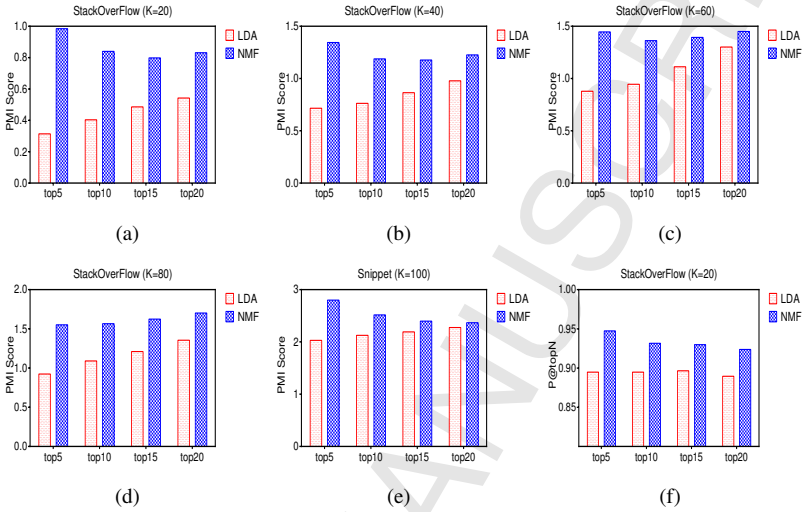


Figure 9: Topic Evaluations on **StackOverFlow** Dataset: (1) from (a) to (e), PMI score with different numbers of topics for topN (topN=5, 10, 15 and 20) terms; (2) (f), human judgements with precision for topN (topN=5, 10, 15 and 20) terms (P@topN for short).

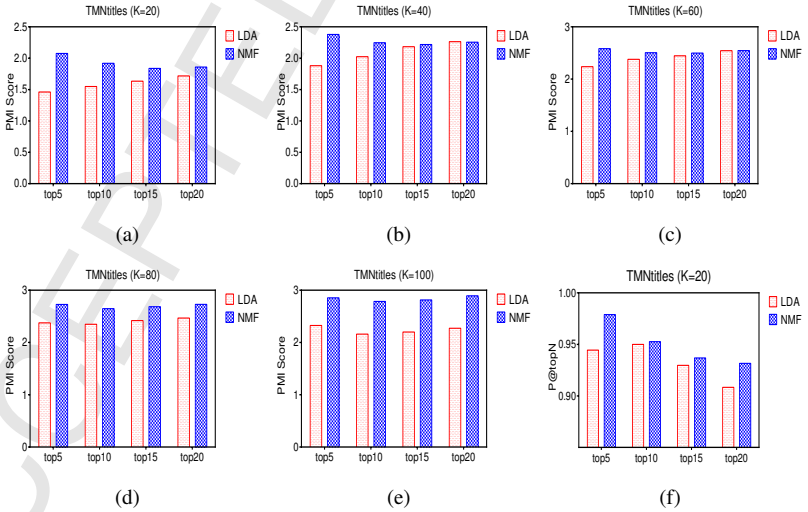
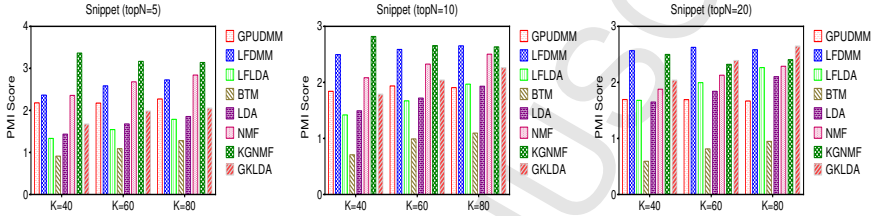


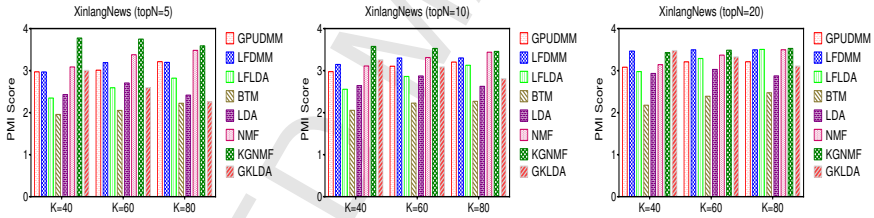
Figure 10: Topic Evaluations on **TMNtitles** Dataset: (1) from (a) to (e), PMI score with different numbers of topics for topN (topN=5, 10, 15 and 20) terms; (2) (f), human judgements with precision for topN (topN=5, 10, 15 and 20) terms (P@topN for short).

the average length of terms per document to be 14.34 and 4.38 respectively. We set the topic number to 40, 60, 80 and conduct 8 groups of experiments (LDA, BTM, LFLDA, LFDMM, GPUDMM, GKLDA, NMF and KGNMF) for each K setting. Then we collect all the experimental results and compute the PMI scores for topN (5, 10, 20) topical words; finally, we plot them in Fig. 11 and Fig. 12.



(a) (b) (c)

Figure 11: PMI scores on **Snippet** dataset among different models with different topic numbers (K=40, 60, and 80) for topN (topN=5, 10, and 20) terms.



(a) (b) (c)

Figure 12: PMI scores on **XinlangNews** dataset among different models with different topic numbers (K=40, 60, and 80) for topN (topN=5, 10, and 20) terms.

The experiments reveal a number of interesting points:

- Obviously, KGNMF showcases higher PMI scores than other LDA-based methods on different experimental configurations (especially for the top5 and top10 topical words), which demonstrates that KGNMF holds stronger ability to comprehend short texts than other competitors.
- KGNMF also performs better than NMF on both representative datasets, which indicates the effectiveness of using knowledge to guide NMF to learn better topics from short texts.
- Generally, knowledge-based GPUDMM, LFDMM, LFLDA and GKLDA exhibit better performance than BTM and LDA, and these further validate that understanding short texts with external knowledge can be enhanced.
- When comparing LFDMM to GPUDMM, the former seems to yield higher-quality topics than the latter, which implies that LFDMM adopts more effec-

tive manners to leverage knowledge than GPUDMM for better short text topic learning.

- Given the two methods LFDMM and LFLDA (which both leverage external knowledge in the same manner), LFDMM is consistently superior to LFLDA in terms of PMI scores, which validates the assumption that each document talks about one topic is quite appropriate for short texts.

Table 3: Average coherent topics and P@topN evaluated by experts given topic number $K = 60$ for different methods on **Snippet** dataset.

Snippet dataset	**Number of Coherent Topics (the total is 60)**	**P@5**	**P@10**
GPUDMM	52	0.9294	0.8686
LFDMM	55	0.9273	0.8982
LFLDA	49	0.8875	0.8417
BTM	43	0.8	0.7475
LDA	41	0.8293	0.7854
NMF	54	0.9667	0.8796
KGNMF	56	0.9607	0.9232
GKLDA	50	0.9063	0.8618

Table 4: Average coherent topics and P@topN evaluated by experts given topic number $K = 60$ for different methods on **XinlangNews** dataset.

XinlangNews dataset	**Number of Coherent Topics (the total is 60)**	**P@5**	**P@10**
GPUDMM	45	0.8279	0.7754
LFDMM	42	0.7854	0.7239
LFLDA	36	0.7264	0.6626
BTM	33	0.6916	0.6167
LDA	32	0.7014	0.6560
NMF	47	0.8846	0.8288
KGNMF	51	0.8930	0.8577
GKLDA	40	0.7822	0.7209

In this part, we also invite the two experts to label the experimental results with $K = 60$ for different models and sort out the statistics in Table 3 and Table 4, corresponding to Snippet and XinlangNews, respectively. It is worth mentioning that the average Cohen's Kappa score for the topic labeling is 0.8923, which also belongs to the high-level agreement. (Actually, the two experts' labeling are highly consistent, which exhibits the high reliability and high quality of our experiments.) From these two tables on the whole, the values in Table 3 is larger than that in Table 4, which is consistent

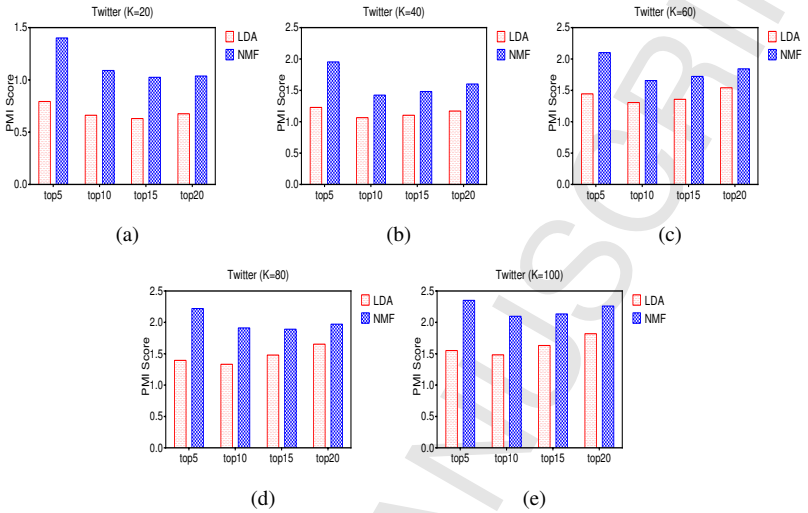


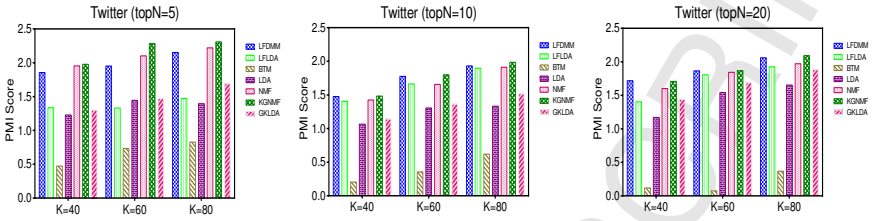
Figure 13: Topic Evaluations on **Twitter** Dataset: PMI score with different numbers of topics for topN (topN=5, 10, 15 and 20) terms.

with the average length of words per document, *i.e.*, the longer, the more semantics, and the better topics. Besides, we also discover that the top5 and top10 words that can best represent each topic in NMF-based methods are of higher quality than that in LDA-based methods, which demonstrate NMF's superiority to LDA for readability.

4.5. An Extended Case on Twitter Dataset

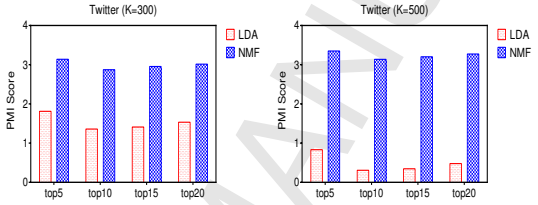
Social networks such as Twitter are abundant in short texts, which are much noisier and larger-scale than the above selected corpora. Therefore, we exploit the open APIs to extract such data from $05/07/2017$ to $05/13/2017$ about $1,886,561$ tweets in total and then conduct an extended experiment among the competing methods. With respect to the large-scale raw dataset, we further preprocess them as the former experiments and build a sub-collection with $102,384$ tweets, of which the average length for each tweet is 5.81 words. As this text collection is large enough, we not only set the topic number to 20, 40, 60, 80, and 100, but also expand them to larger values as 300 and 500. Note that the crawled Twitter corpus hasn't any annotation information which would make GPUDMM infeasible (this method needs labels to form knowledge). Besides, when the topic number is set to larger values, GKLDA would be significantly intolerable regarding time spending; then we exclude it as a baseline in the experiments with larger topic numbers. Finally, all the experimental results are collected and shown in Fig. 13$\sim$Fig. 16.

Obviously, Fig. 13 tells us that NMF yields better topics than LDA on Twitter dataset with topic numbers ranging from 20 to 100, which further strengthens the finding that "NMF tends to be superior to LDA for short text topic mining"; Besides, this advantage seems to be more obvious when the topic numbers are bigger (K=300, 500) in Fig. 15. This is probably due to the LDA's poorer performance on more sparse aver-



(a) (b) (c)

Figure 14: PMI scores on **Twitter** dataset among different models with different topic numbers (K=40, 60, and 80) for topN (topN=5, 10, and 20) terms.



(a) (b)

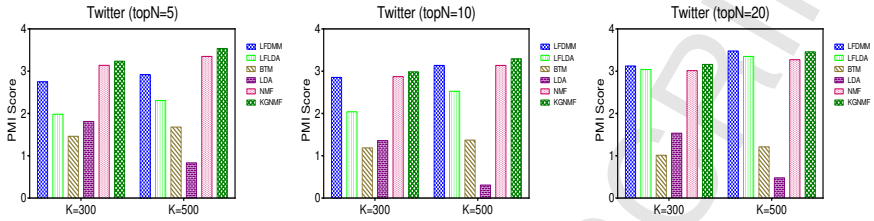
Figure 15: Topic Evaluations on **Twitter** Dataset: PMI score with larger numbers of topics (K=300, 500) for topN (topN=5, 10, 15 and 20) terms.

age texts when the topic number becomes larger; while NMF keeps stable performance for its more priors and learning algorithm. Similarly, from Fig. 14 and Fig. 16, we could also see the superiority of our proposed KGNMF to other methods both with small and large topic numbers. In general, this extended experiment further validates our previous experimental conclusions.

4.6. Summary

In order to comprehensively compare LDA-based and NMF-based topic models for short texts, we design three experimental sections: basic experiments between LDA and NMF; further comparisons between LDA-based and NMF-based approaches with representative research works, in which we propose a novel and efficient approach called KGNMF; and an extended case on Twitter dataset among all the competing methods. From the experimental outcomes, we can make a brief summarization as follows:

- Basically, NMF is likely to perform better than LDA under the same configurations in topic mining for short texts;
- Generally, external knowledge is supposed to be leveraged for better short-text topic models;
- Our proposed KGNMF, a knowledge-guided model, is superior to many state-of-the-art baselines mentioned in our experiments; and even if compared with



(a) (b) (c)

Figure 16: PMI scores on **Twitter** dataset among different models with larger topic numbers (K=300, 500) for topN (topN=5, 10, and 20) terms.

LFDMM and GPUDMM, two well-performed knowledge-based probabilistic methods, our method performs still better or at least comparable.

5. Conclusion and Future Work

In this paper, we mainly compare the LDA-based and NMF-based learning schemes for uncovering the latent topics from short texts. Extensive experiments on public short-text datasets demonstrate that (1) NMF is more inclined to produce higher-quality topics than LDA with the same experimental settings; (2) Our proposed knowledge-based model KGNMF, which is designed with word-word pairwise semantic regularization via low-rank representations (time-efficient), performs much better in learning topics than many state-of-the-art baselines; (3) generally speaking, topic modeling is supposed to better exploit external knowledge for better understanding short texts. Overall, NMF-based topic models are also significantly effective learning schemes for short text topic mining apart from the popular LDA-based methods. As to the future research, we would like to experimentally explore the differences between LDA and NMF based topic models for long/normal texts.

6. Acknowledgment

This work is supported in part by National Key R&D Project under grant No. 2017YFB1400200. We also thank to the Network Information Center (Beihang University, Beijing 100191, P.R. China) for providing high-performance servers.

References

[1] Ge Song, Yunming Ye, Xiaolin Du, Xiaohui Huang, and Shifu Bie. Short Text Classification: A Survey. Journal of Multimedia 9(5): 635-643 (2014).

[2] Prajol Shrestha, Christine Jacquin, and Beatrice Daille. Clustering Short Text and Its Evaluation. CICLing (2) 2012: 169-180.

[3] Svetlana Kiritchenko, Xiaodan Zhu, and Saif M. Mohammad. Sentiment Analysis of Short Informal Texts. J. Artif. Intell. Res. 50: 723-762 (2014).

[4] Dongxiang Zhang, Yuchen Li, Ju Fan, Lianli Gao, Fumin Shen, and Heng Tao Shen. Processing Long Queries Against Short Text: Top-k Advertisement Matching in News Stream Applications. ACM Trans. Inf. Syst. 35(3): 28:1-28:27 (2017).

[5] Cheng-Ying Liu, Ming-Syan Chen, and Chi-Yao Tseng. IncreSTS: Towards Real-Time Incremental Short Text Summarization on Comment Streams from Social Network Services. IEEE Trans. Knowl. Data Eng. 27(11): 2986-3000 (2015).

[6] Junyu Xuan, Xiangfeng Luo, Guangquan Zhang, Jie Lu, and Zheng Xu. Uncertainty Analysis for the Keyword System of Web Events. IEEE Trans. Systems, Man, and Cybernetics: Systems 46(6): 829-842 (2016).

[7] Junyu Xuan, Xiangfeng Luo, Jie Lu, and Guangquan Zhang. Explicitly and Implicitly Exploiting the Hierarchical Structure for Mining Website Interests on News Events. Inf. Sci. 420: 263-277 (2017).

[8] Zhijian Yuan, Yan Jia, and Shuqiang Yang. Online Burst Detection Over High Speed Short Text Streams. International Conference on Computational Science (3) 2007: 717-725.

[9] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet Allocation. Journal of Machine Learning Research 3: 993-1022 (2003).

[10] David M. Blei, Thomas L. Griffiths, Michael I. Jordan, and Joshua B. Tenenbaum. Hierarchical Topic Models and the Nested Chinese Restaurant Process. NIPS 2003: 17-24.

[11] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet Allocation. NIPS 2001: 601-608.

[12] Daniel D. Lee and H. Sebastian Seung. Learning the Parts of Objects by Non-negative Matrix Factorization. Nature 401: 788C791 (1999).

[13] Daniel D. Lee and H. Sebastian Seung. Algorithms for Non-negative Matrix Factorization. NIPS 2000: 556-562.

[14] Peter Norvig. Inference In Text Understanding. AAAI Spring Symposium: Machine Reading 2007: 6-10.

[15] Jun Zhang, Xiangfeng Luo and Feiyue Ye. A Machine-Oriented Text Understanding Framework Based on Human Memory and Reading Process. Web Intelligence 2012: 732-738.

[16] Xiaohui Yan, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng. A Biterm Topic Model for Short Texts. WWW 2013: 1445-1456.

[17] Xueqi Cheng, Xiaohui Yan, Yanyan Lan, and Jiafeng Guo. BTM: Topic Modeling over Short Texts. IEEE Trans. Knowl. Data Eng. 26(12): 2928-2941 (2014).

[18] Jianhua Yin, and Jianyong Wang. A Dirichlet Multinomial Mixture Model-based Approach for Short Text Clustering. KDD 2014: 233-242.

[19] Stephen G. Walker. Sampling the Dirichlet Mixture Model with Slices. Communications in Statistics - Simulation and Computation 36(1): 45-54 (2007).

[20] Dat Quoc Nguyen, Richard Billingsley, Lan Du, and Mark Johnson. Improving Topic Models with Latent Feature Word Representations. TACL 3: 299-313 (2015).

[21] Chenliang Li, Haoran Wang, Zhiqian Zhang, Aixin Sun, and Zongyang Ma. Topic Modeling for Short Texts with Auxiliary Word Embeddings. SIGIR 2016: 165-174.

[22] Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Distributed Representations of Words and Phrases and their Compositionality. NIPS 2013: 3111-3119.

[23] Quoc V. Le, and Tomas Mikolov. Distributed Representations of Sentences and Documents. ICML 2014: 1188-1196.

[24] Francois Caron, Manuel Davy, and Arnaud Doucet. Generalized Polya Urn for Time-varying Dirichlet Process Mixtures. UAI 2007: 33-40.

[25] May-Ru Chen, and Markus Kuba. On Generalized Polya Urn Models. J. Applied Probability 50(4): 1169-1186 (2013).

[26] Vivek Kumar Rangarajan Sridhar. Unsupervised Topic Modeling for Short Texts Using Distributed Representations of Words. VS@HLT-NAACL 2015: 192-200.

[27] Kang Xu, Guilin Qi, Junheng Huang, and Tianxing Wu. Incorporating Wikipedia Concepts and Categories as Prior Knowledge into Topic Models. Intell. Data Anal. 21(2): 443-461 (2017).

[28] Xiaojun Quan, Chunyu Kit, Yong Ge, and Sinno Jialin Pan. Short and Sparse Text Topic Modeling via Self-Aggregation. IJCAI 2015: 2270-2276.

[29] Zhiyuan Chen, Arjun Mukherjee, Bing Liu, Meichun Hsu, Malu Castellanos, and Riddhiman Ghosh. Discovering Coherent Topics using General Knowledge. CIKM 2013: 209-218.

[30] Xuan Hieu Phan, Minh Le Nguyen, and Susumu Horiguchi. Learning to Classify Short and Sparse Text & Web with Hidden Topics from Large-scale Data Collections. WWW 2008: 91-100.

[31] Haixun Wang. Understanding Short Texts. APWeb 2013: 1.

[32] Zheng Yu, Haixun Wang, Xuemin Lin, and Min Wang. Understanding Short Texts through Semantic Enrichment and Hashing. IEEE Trans. Knowl. Data Eng. 28(2): 566-579 (2016).

[33] Jonathan Chang, Jordan L. Boyd-Graber, Sean Gerrish, Chong Wang, and David M. Blei. Reading Tea Leaves: How Humans Interpret Topic Models. NIPS 2009: 288-296.

[34] David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. Automatic Evaluation of Topic Coherence. HLT-NAACL 2010: 100-108.

[35] Jiaming Xu, Peng Wang, Guanhua Tian, Bo Xu, Jun Zhao, Fangyuan Wang, and Hongwei Hao. Short Text Clustering via Convolutional Neural Networks. VS@HLT-NAACL 2015: 62-69.

[36] David M. Mimno, Hanna M. Wallach, Edmund M. Talley, Miriam Leenders, and Andrew McCallum. Optimizing Semantic Coherence in Topic Models. EMNLP 2011: 262-272.

[37] Jey Han Lau, David Newman, and Timothy Baldwin. Machine Reading Tea Leaves: Automatically Evaluating Topic Coherence and Topic Model Quality. EACL 2014: 530-539.

[38] Ran He, Wei-Shi Zheng, Bao-Gang Hu, and Xiangwei Kong. Nonnegative Sparse Coding for Discriminative Semi-supervised Learning. CVPR 2011: 2849-2856.

[39] Yong Chen, Hui Zhang, Junjie Wu, Xingguang Wang, Rui Liu, and Mengxiang Lin. Modeling Emerging, Evolving and Fading Topics Using Dynamic Soft Orthogonal NMF with Sparse Representation. ICDM 2015: 61-70.

[40] Zhiyuan Chen, and Bing Liu. Topic Modeling using Topics from Many Domains, Lifelong Learning and Big Data. ICML 2014: 703-711.

[41] J. Richard Landis, and Gary G. Koch. The Measurement of Observer Agreement for Categorical Data. Biometrics 33(1): 159-74 (1977).