Expert Systems With Applications 193 (2022) 116518

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

A probabilistic topic model based on short distance Co-occurrences

Marziea Rahimi *, Morteza Zahedi, Hoda Mashayekhi

Faculty of Computer Engineering, Shahrood University of Technology, Shahrood 3619995161, Iran

ARTICLE INFO

Keywords:
Probabilistic topic model
Latent Dirichlet Allocation
Document clustering
Context window
Local co-occurrence
Word order

ABSTRACT

A limitation of many probabilistic topic models such as Latent Dirichlet Allocation (LDA) is their inflexibility to use local contexts. As a result, these models cannot directly benefit from short-distance co-occurrences, which are more likely to be indicators of meaningful word relationships. Some models such as the Bigram Topic Model (BTM) consider local context by integrating language and topic models. However, due to taking the exact word order into account, such models suffer severely from sparseness. Some other models like Latent Dirichlet Co-Clustering (LDCC) try to solve the problem by adding another level of granularity assuming a document as a bag of segments, while ignoring the word order. In this paper, we introduce a new topic model which uses overlapping windows to encode local word relationships. In the proposed model, we assume a document is comprised of fixed-size overlapping windows, and formulate a new generative process accordingly. In the inference procedure, each word is sampled once in only a single window, while influencing the sampling of its other fellow co-occurring words in other windows. Word relationships are discovered in the document level, but the topic of each word is derived considering only its neighbor words in a window, to emphasize local word relationships. By using overlapping windows, without assuming an explicit dependency between adjacent words, we avoid ignoring the word order completely. The proposed model is straightforward, not severely prone to sparseness and as the experimental results show, produces more meaningful and more coherent topics compared to the three mentioned established models.

1. Introduction

The continuous growth of digitalized text data in recent decades makes it increasingly more difficult to discover the desired information. Probabilistic topic models, as unsupervised methods for modeling documents, provide a means of representing and exploring this vast amount of data. In a probabilistic topic model, a topic is a distribution over the vocabulary words, and each document is a mixture of topics. The mixture of topics is a low-dimensional representation of the otherwise high-dimensional text document, and the topics are more informative compared to single words. Searching based on topics is far more accurate and efficient than based on the words (Blei et al., 2010). Probabilistic topic models have proved to be useful in various domains such as document classification (Rubin et al., 2012; Zhang & Zhong, 2016), summarization (Harabagiu & Lacatusu, 2005; Wang et al., 2009), recommendation (Kim & Shim, 2014; Wang & Blei, 2011), text segmentation (Misra et al., 2011; Misra et al., 2009) and decision support systems (Bastani et al., 2019; Dong et al., 2018).

Topic modeling transforms documents (term-document matrix) from the specific space of words to the more general space of topics. Unlike many other transformation methods (Cerda & Varoquaux, 2020; Deerwester et al., 1990), the new dimensions generated by topic models are human-interpretable. As this is one of the key advantages of the probabilistic topic models, it is very important to design models which extract meaningful and coherent topics. Topic coherence is important in applications where the topics are directly used by experts for analysis and interpretation of data (Henrichs, 2019; Schnober & Gurevych, 2015). For example, Liu et al. (Liu et al., 2020) and Stokes et al. (Stokes et al., 2020) use the generated topics from social media to analyze people's reactions to COVID-19. It is also important for other applications where the topics or topic proportions are used as features for further processing. For example, the fundamental tasks of text clustering (Ma et al., 2014; Onan et al., 2017) and classification (Pavlinek & Podgorelec, 2017; Sinoara et al., 2019) can employ document topics as auxiliary features.

Topic models work based on word relationships and as elaborated later in this section, smaller contexts lead to extraction of more meaningful word relationships. Also, using word order information can

* Corresponding author.
E-mail addresses: marziea.rahimi@shahroodut.ac.ir (M. Rahimi), zahedi@shahroodut.ac.ir (M. Zahedi), hmashayekhi@shahroodut.ac.ir (H. Mashayekhi).

https://doi.org/10.1016/j.eswa.2022.116518
Received 31 March 2020; Received in revised form 4 September 2020; Accepted 6 January 2022
Available online 10 January 2022

further enrich the meaningfulness of the extracted relationships. The existing probabilistic topic models fail to utilize this useful information appropriately, without being trapped by sparseness. In this paper, our goal is to introduce a topic model which generates meaningful topics, benefiting from these two types of information, while avoiding their inherent issue of sparseness.

Most probabilistic topic models are inspired by LDA (Blei et al., 2003). Such models assume a generative process for the documents, and afterwards estimate the topics by reversing the process. Some simplification assumptions have to be made for this procedure to be feasible. However, it is important to define a generative process which resembles reality as much as possible. LDA assumes that words are exchangeable in the whole document, i.e. it works based on document-level word co-occurrence and neglects local word relationships. This assumption is, however, far from linguistic reality (Blei et al., 2010). According to the co-occurrence hypothesis (Schulte Im Walde & Melinger, 2008), there is a negative correlation between distance and co-occurrence strength, which means that it is more likely to find related words in a short context than a longer one. Therefore, topics which are extracted focusing on local word relationships are more coherent, accurate, and meaningful.

Some methods (Shafiei & Muios, 2006; Wallach, 2006; BTM) try to relax the word exchangeability assumption by using the concept of n-gram language models to incorporate local word relationships. N-gram language models take the exact word order into account, and thus sparseness becomes a major problem, as many possible word sequences do not occur in the training set. On the other hand, in topic modeling, exact word order in a short context, does not change the topic except for specific idioms. Topics of a text are recognizable even if the word order is twisted such that the text is meaningless (Blei et al., 2003). For example, in the following text, the words in a paragraph of two sentences are randomized. Although it is not clear what the text is exactly saying, the topic is recognizable to be on brain disorders[1].

> "an with since hard substance is antibiotics is brain and to off surgical walled it needs it cure usually gets in the drainage It deep brain abscess A infection"

However, in longer contexts, such as a complete document, exchanging words and twisting their order, would make it much difficult or even impossible to discover the topics. This is due to the fact that short-distance co-occurrences are more likely to be indicators of meaningful and robust relationships than long-distance co-occurrences (Schulte Im Walde & Melinger, 2008). Im Walde and Melinger (Schulte Im Walde and Melinger, 2008) show that strongly related words co-occur more often in smaller context windows. Increasing the window length offers little contribution to the number of co-occurrences for related words, while causing more unrelated words to co-occur in the context window. This does not strictly suggest that long-distance co-occurrences are meaningless, but rather explains that short-distance co-occurrences are more reliable.

Another approach of considering the short-distance co-occurrences is assuming a document as a set of segments where each segment is a bag of words. Topics are extracted according to words co-occurrences in the segments using LDA. Models like Latent Dirichlet Co-Clustering (LDCC) (Shafiei & Muios, 2006) have used this idea with two levels of topics, namely word-level and segment-level, to benefit from segment-level co-occurrences. However, similar to LDA, such models do not treat short-distance and long-distance co-occurrences differently. They consider which segments the words co-occur in, but do not pay attention to the co-occurrence span.

In this paper, we propose a topic model which encodes local word relationships without considering the exact word order and also without completely ignoring it. The proposed Local Latent Dirichlet Allocation (LLDA) model, considers each document as a set of overlapping windows. In each window, the topic of only a single word is extracted according to the window-level topic proportions. Contrary to LDA in which, the topic of each word directly affects topic of every other word in the document regardless of their distance, in the proposed approach, the topic of a word directly affects only topics of the words co-occurring with it in some window. Therefore, the farther two words are in the document, the fewer windows they co-occur in and thus the milder the influence they have on each other. The model is able to emphasize local word co-occurrences, which are better indicators of meaningful word relationships. This is contrary to LDA or LDCC in which, each word directly affects the topic of others in its document or segment ignoring their distance. We stress that our aim is not to perform LDA on overlapped windows violating the entirety of a document and its rich context.

LLDA is different from models which use n-grams, as it does not assume any explicit dependency between words. However, since the windows are overlapped, the order of words is not completely ignored, but the adjacent words themselves are more important than their order. Therefore, LLDA is less prone to sparseness. The highlights of this paper are as follows:

- Proposing a new topic model which focuses on short-distance co-occurrences.
- utilizing a mitigated form of word order information to avoid sparseness.
- Incorporating overlapped windows into the topic model and formulating a new generative process along with its inference process.
- Preserving the entirety of the documents by sampling each word only once but with respect to its neighbor words in a window to emphasize local word relationships.
- Performing quantitative and qualitative experiments to evaluate and compare the model intrinsically in general and also in a specific application.

Due to the model design principles, LLDA is able to extract more coherent and meaningful topics compared to the probabilistic topic models which work based on document level word co-occurrences, language models, or segment-based structures. To justify this claim, the model is compared with LDA, BTM, and LDCC as the representatives of the mentioned three categories of topic models. The results show that the proposed model is able to extract more coherent topics, and coherence does not drastically decrease by increasing the number of topics. We further asses the models in an application of text clustering, showing that the proposed model is able to produce the most similar clusters to the human-generated categories of the corpora. Text clustering is a fundamental task which can be applied with many different objectives and scenarios such as sentiment analysis (AL-Sharuee et al., 2018; Rehioui & Idrissi, 2020; Riaz et al., 2019) where the documents are categorized by their polarity or authorship attribution (Hamadache & Sayoud, 2018; Panicheva et al., 2019; Stamatatos et al., 2016) in which documents are grouped according their authorship style using clustering or classification algorithms. Text clustering is also a key step in many others like information retrieval (Djenouri et al., 2018; Reda et al., 2020) where text clustering techniques are employed to primarily group the documents based on their relevance to a query and then process the relevant groups for more accurate ranking, text summarization (Mallick et al., 2018; Rouane et al., 2019) in which text clustering is used to group similar sentences and recommendation systems (Drushku et al., 2019; Jiang et al., 2019) where text clustering is used for grouping the textual data available on user interests or their objects of interest.

In the rest of the paper, existing related models are briefly introduced in section 2. Section 3 contains a more elaborated introduction of the key base models, LDA, BTM, and LDCC, which are the representatives of

[1] The original text: "A brain abscess is an infection deep in the brain substance. It is hard to cure with antibiotics since it gets walled off and usually it needs surgical drainage."

the three main categories mentioned in section 2. In section 4, the proposed model is introduced. The process of parameter estimation for the model is also discussed in section 4, and finally, experiments and results are reported in section 5.

2. Literature review

Our main goal, in this paper, is to introduce a model which can benefit from short-distance co-occurrences as a more meaningful indicator of word association, and also a mitigated form of word-order information while avoiding the severe issue of sparseness. Accordingly, existing topic models can be grouped in three main categories. The first category consists of topic models that consider the document-level co-occurrences and ignore the word-order information completely by assuming a document as a bag of words. This category is represented by LDA. The second category is formed by the models that try to integrate language and topic models and through this integration focus more on local co-occurrences. These models suffer severely from sparseness. Most of the models in this category are inspired by BTM which is assumed as the representative of this category. The models classified in the third category assume a document is comprised of smaller segments such as paragraphs and sentences, and generally speaking, try to apply LDA on this smaller segments. However, LDA has proven to perform unsuccessfully on small texts (Qiang et al., 2020). In the following paragraphs some of the major examples of these categories are introduced and the disadvantages of each category with respect to the proposed method are discussed in more detail.

LDA is a probabilistic topic model which assumes a document is a bag of words. Following its introduction, many different probabilistic topic models have been devised. For example, dynamic topic model which is an LDA-based topic model (DTM) (Blei & Lafferty, 2006) that keeps track of topics over time, and Gaussian dynamic topic model (GDTM) (Li et al., 2019) which is an LDA-based model for community detection and forwarding prediction in social networks. Another example is the labeled LDA (Ramage et al., 2009) which takes document labels into account and models labeled documents. Hetero-Labeled LDA (Kang et al., 2014) extends this model to be able to learn from more kinds of labels than only the document labels. Supervised LDA (SLDA) (Blei & McAuliffe, 2007) is another example of LDA extensions. This model assumes an additional response variable corresponding to each document, and tries to jointly model the document and the response. The response can be interpreted differently according to dataset including the document class or category. LF_LDA (Nguyen et al., 2015) tries to improve the performance of LDA over short texts by incorporating external features into the model. There are many other examples of such models that are able to work sufficiently on short documents, mostly with the intention of using topic models for social network analysis. Many such models like LF_DMM (Nguyen et al., 2015) are inspired from Dirichlet multinomial model (DMM) (Thrun et al., 2000) which proceeds LDA and assume that the words in a sentence or short text are derived from the same topic. Many other short-text-adapted topic models are inspired by a model named bi-term topic model (Cheng et al., 2014), which is a simplification of LDA and does not directly consider the documents, and instead assumes the corpus is a collection of term pairs (bi-terms). Improving LDA using a better initialization (Belford et al., 2018) has also been considered by researchers. Incorporating prior knowledge in topic models is another direction taken by researches to improve LDA. For example, Wang et al. introduce a model (R. Wang et al., 2020) which is based on LDA but using a new version of Polya urn scheme which they call Weighted Polya Urn scheme (WPU). This model uses word similarities obtained using word embedding vectors as the prior knowledge. Hierarchical LDA (HLDA) (Griffith et al., 2004) assumes a hierarchy of topics with the topic of each word sampled through a path in this hierarchy. The path is a random variable which follows a Chinese restaurant process. Another example of such models that assume a kind of relationship between topics is the Knowledge-based Hierarchical Topic Model (KHTM) (Xu et al., 2018) which improves HLDA by incorporating prior knowledge into the model and using the generalized Polya urn model (GPU). Trying to take the document structure into account is another line of improvement over LDA. A recently introduced example is Copula-LDA (Balikas et al., 2016) which uses Copulas to model topic relationships in a coherent text segment by increasing the possibility of having similar topics for the words in a coherent text segment. Some other models try to improve the speed and performance of the inference process in LDA or other probabilistic topic models. For example, maximal latent state replication (MAX) (Rugeles et al., 2020) tries to turn LDA into a distributable algorithm by changing the sequential nature of Gibbs sampling considering a newly proposed state augmentation based inference method named State Augmentation for Marginal Estimation (SAME) (Zhao et al., 2015). Other example is a model (He et al., 2017) which tries to introduce a more efficient inference algorithm for the bi-term topic model or a model (Ha et al., 2019) which aims to reduce the time complexity of topic models by using the dropout technique from the neural network context. The common feature of these and other LDA-based models is that they treat short distance and long-distance co-occurrences equally and ignore the order of words which as explained in Section 1, introduces challenges for extracting coherent topics in longer texts.

As mentioned above, the models which take local word relationships into account can be divided into two major categories. We call them topic n-gram models and segment-based topic models. There are also Hierarchical Pitman-Yor (HPY) based models which can be considered as a subcategory of the n-gram models.

The topic n-gram models incorporate exact word order into topic models. For example, Wallach (Wallach, 2006) incorporated bigram language model into LDA and suggested the Bigram Topic Model (BTM) which assumes each word is dependent to its immediate previous word in addition to its topic. Barbieri et al. (Barbieri et al., 2013) have suggested a very similar Token-Bigram model which makes the same assumption. Grifith et al. (Griffiths et al., 2007) introduced the LDA-collocation model for incorporating collocations (Biber, 1993; Nesselhauf, 2005) into LDA. This model assumes a word is generated either from its topic or from its previous word which together form a collocation. A Bernoulli variable is employed as a switch to select one of the two options. Wang et al. proposed a generalization to the LDA-collocation model in which, a word can choose to form a collocation with its previous word or not, depending on its topic. For example, “white house” can be regarded as a collocation or as different words “white” and “house” according to the topic. Inspire by this model, Kim and Yoon (S. Kim & Yoon, 2015) propose link-topic model for disambiguation of medical abbreviations. This model assumes each word can form a pair with one of its previous words and if so, the topic of the pair is sampled from the same topic distribution. Yang et al. (Yang et al., 2015) use a similar setting in a model named contextual topic model for document summarization. In this model, each word can be generated from a unigram or a bigram language model and also from a path in a topic hierarchy. Zhu et al (Zhu et al., 2019) introduce a hierarchical topic model which models each phrase as an n-gram using an HPY process. Similar to other mentioned models in this section, this model considers bigrams and also trigrams. Jameel et al. (Jameel et al., 2015) incorporate bigram language model into a supervised topic model for document classification. Although all of the above mentioned models assume the current word is dependent only on its previous word, it is possible to incorporate higher order n-gram language models in a probabilistic topic model (Wallach, 2006) which is faced with some difficulties. N-gram language models assume that each word is generated from a sequence of n words located exactly before the word. In this generation process, sparseness is a major problem, especially when the corpus is small, because many possible sequences of the words do not appear in the corpus. Another kind of topic models (Cheng et al., 2014; He et al., 2017; Pang et al., 2019) related to this category is inspired by a

biterm topic model (Cheng et al., 2014) which assumes a corpus is a bag of word pairs (biterms). To avoid the sparseness, it ignores the documents and constructs the topics based on corpus-level word-word co-occurrences. It does not assume any explicit dependency between words and just assumes that each word pair is sampled from the same topic. The main issue with this model which has been specially introduced to extract the thematic structure of short texts is the lack of a document model.

Some topic n-gram models use the HPY process (Teh, 2006) priors (Sato & Nakagawa, 2010) and (Noji et al., 2013). These models, inspired from (Teh, 2006), are not restricted to bigrams and have been tested for different n-grams. The main goal of these models is to make domain (topic) specific language models rather than incorporating local word relationships. As a result, evaluations of these models have mostly been focused on comparing them with n-gram language models rather than other topic models. However, when word order is taken into account, sparseness becomes a major challenge. Therefore, a very large corpus is required for obtaining robust or even acceptable results, but such models are very expensive and almost impractical on large datasets (Noji et al., 2013). On the other hand, the exact word order while being very important for some applications such as statistical machine translation and speech recognition, is not very informative in topic modeling. In general topic models, the words of the context themselves are more important than their order.

Segmented topic models (Shafiei & Muios, 2006; Wallach, 2006; Shafiei & Muios, 2006; Wallach, 2006; Balikas et al., 2016; Du et al., 2010; Shafiei & Muios, 2006), in contrary to the n-gram models, ignore the word order and instead add another level of granularity to the data. These models that are mostly inspired by the LDCC model (Shafiei & Muios, 2006), assume that a document comprises several segments. Each segment, which can be a sentence or paragraph, has a topic assigned to it which is called the segment-topic. The topic of each word in the document is dependent on its segment's topic. The nature of segment-topics and word topics are different; each segment-topic is a distribution over word-topics where each word topic is a distribution over words. This means that such models consider two levels of topics. Assuming this hierarchy of topics makes it possible for the document clusters which are obtained according to segment-topics, to be expressed based on a low-dimensional representation of the text which is provided by word-topics. So, document clustering in this model is more accurate in comparison to many similar models (Banerjee et al., 2005; Chipman & Gu, 2005). Because LDCC considers a co-occurrence context shorter than a document, it mildly encodes local word relationships. But the main goal of these models is to maintain document structure in the process of word generation.

Some models have been proposed for special applications or different types of data, such as images. Usually, models which are introduced for text data are applicable to images by assuming an image as a document or a set of documents, and groups of adjacent pixels as words. But the reverse may not be applicable in a straight forward manner. For example, Jeong and Choi (Jeong & Choi, 2015) proposed a topic model for image segmentation. In this model, each image is a set of overlapped documents which is not a valid assumption for text corpora. The Spatial latent Dirichlet allocation (X. Wang & Grimson, 2008) or the spatial regularized latent topic model (Ou et al., 2015) are other examples of such models. Zuo et al. (Zuo et al., 2016) introduced a model named word network topic model for solving the sparseness and imbalance problems when employing topic models in applications involving short texts. In this model, a co-occurrence matrix is constructed, and then each word is represented as a pseudo-document of words that are co-occurred with it. LDA is applied to the pseudo documents. Park et al. (Park et al., 2019) introduced a model to deal with text data comprised of various clusters with several different topic distributions or in other words highly multi-modal topic distributions. There have also been some models to deal with topic modeling on massive data (Fuentes-pineda & Meza-ruiz, 2019; Yuan et al., 2015).

In summary, in the existing topic models, local word relationships are extracted using n-gram language models or by considering a shorter co-occurrence context. The first approach is dependent on the exact order of words and thus suffers severely from sparseness, while the latter ignores the order of words completely. In this paper, we propose a new probabilistic model which extracts local word relationships by assuming each word itself is a context window which overlaps the other words, and thus is sensitive to word orders without being very susceptible to sparseness.

3. Preliminaries

The following terminology is used to describe the models. The corpus D is a set of M exchangeable documents $D = \{d_1, d_2, \ldots, d_m, \ldots d_M\}$ containing a total of N words. Each document $d_m \in D$ is a set of N_m words $d_m = \{w_{m1}, w_{m2}, \ldots, w_{mn}, \ldots, w_{mN_m}\}$. Each word w_{mn} can accept values from vocabulary V. Each word w_{mn} is assigned a topic $z_{mn} \in \{1, 2, \ldots, K\}$. A topic is a distribution over vocabulary words and is specified by an ordered list of its I most probable words $T^k = (t_1^k, t_2^k, \ldots, t_I^k)$ which are called topic descriptors.

We will compare the proposed model with LDA, BTM and LDCC as the representatives of models based on document-level co-occurrences, models considering n-grams, and models based on co-occurrences in smaller segments, respectively. Therefore, in this section, these three models are discussed in more details.

3.1. Lda

In LDA each document d_m is a mixture over latent topics or in the other words a distribution over topics denoted by θ_m which follows a Dirichlet distribution with parameter α. Each word w_{mn} in the document is assigned a topic z_{mn}. Each topic is a distribution over words denoted by φ_k which follows a Dirichlet distribution with parameter β. LDA assumes the following generative process for documents:

- For each document d_m in the corpus
  - Choose $\theta_m \widetilde{Dirichlet}(\alpha)$
  - For each word w_{mn} in the document
    - Choose a topic $z_{mn} \widetilde{multinomial}(\theta_m)$
    - Choose the target word $w_{mn} \widetilde{multinomial}(\varphi_{z_{mn}})$

Fig. 1 shows the graphical representation of LDA. LDA works based on word co-occurrences in the documents and does not emphasize local word relationships.

3.2. BTm

In BTM, as shown in the graphical representation of the model in Fig. 2, each word is also dependent on its previous word $(w_{mn} \widetilde{multinomial}(\varphi_{z_{mn}, w_{m(n-1)}}))$. Therefore, words are not exchangeable. In BTM, in contrast to LDA, each topic is a set of $|V|$ distributions over words. In BTM, the word co-occurrence is considered in a co-occurrence window of length two.

3.3. LDCC

LDCC as shown in Fig. 3, adds an additional level of granularity to the model with the aim of maintaining document structure in generation process. In this model, a document is a bag of segments, each of which expresses a single topic y_s. For this reason, the model assumes two different types of topics, segment-topics, and word-topics. Each document d_m is a distribution ψ_m over segment-topics, each segment s is a distribution θ_s over word-topics denoted by z. Each word-topic is a distribution over words. LDCC assumes the following generative process for each document.

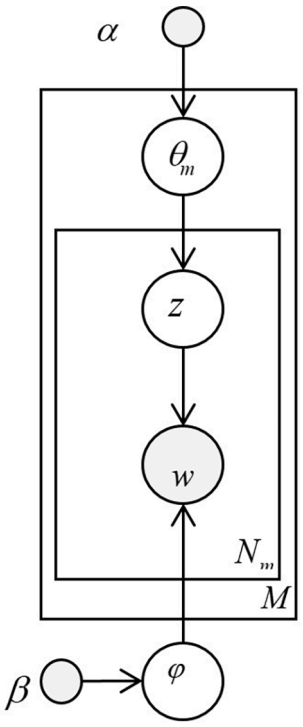

Fig. 1. Plate diagram for LDA.

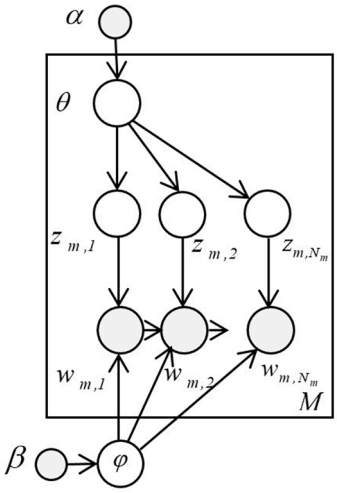

Fig. 2. Plate diagram for BTM.

- For each document d_m in the corpus D
  - o Choose $\psi_m \sim Dir(\gamma)$
  - o For each of the S segments s in d_m
    - ■ Choose a topic $y_s \sim Multinomial(\psi_m)$
    - ■ Choose $\theta_s \sim Dir(\alpha, y_s)$
    - ■ For each of the N_s words w_{sn}
      - Choose a topic $z_{sn} \sim Multinomial(\theta_s)$
      - Choose a word w_{sn} from $P(w_{sn}|z_{sn}, \varphi)$, a multinomial probability conditioned on the topic z_{sn}

Although LDCC benefits from word co-occurrences in a shorter context, the exchangeability assumption is still intact among segments and among words of each segment. In fact, LDCC is sensitive to the segment the words belong to but not to the words distances.

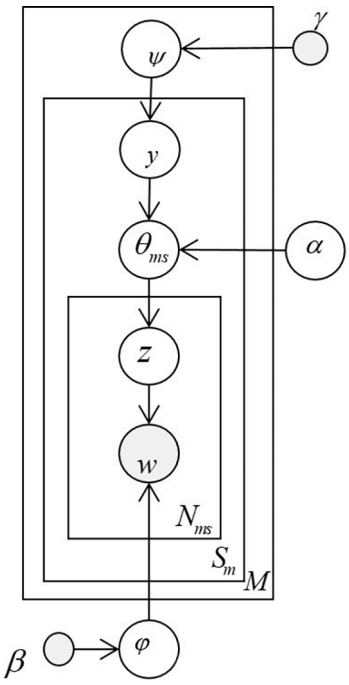

Fig. 3. Plate diagram for LDCC.

4. The local latent Dirichlet Allocation model

In this paper, we introduce a generative topic model which is able to encode local word relationships using fixed-size overlapping windows. Although the model does not use language models, it does not completely discard the word order information. This is possible by using overlapping windows. Each window has an overlap of L-1 cells with its previous window, where L denotes the window size. Using such windows, the word sequences are repeated in the document which allows the model to maintain a mild sense of word order without using language models. Changes in the order of words in a window, influences the words of overlapping windows and consequently the word co-occurrences. In this manner the effect of word order is mild and the words themselves are more important than their order.

Due to the windows overlapping, each word is covered by several windows, however it needs to be sampled only in a single window. To avoid sampling a word repeatedly in each window, we assume one single word as the target in each window. Each word may be present in different windows but is considered as target only in a single window and is sampled only in that window. In this setting, each word is assigned a topic only once, and preserves its topic in the same neighborhood. Even though each word is sampled in only a single window where it is the target, it affects the sampling of the other words which co-occurred with in other windows throughout the documents. Therefore, topics are discovered in the document level, with the topic of each word being sampled considering only its neighbor words in its own window to emphasize local word relationships.

In other words, the model assumes each word w_{mn} corresponds to a single window s_{mn} of adjacent words. We call w_{mn} the target word for s_{mn}. The location $0 \leqslant p \leqslant L-1$ of the target word, is chosen arbitrarily and is identical for all windows. In Fig. 4, a window of length L is shown. The target word of the window is w_n and the window covers words w_{n-p} to w_{n+l-p}. Each word in addition to its corresponding window, is covered by $L-1$ other windows. So, a document d_m with N_m words consists of N_m windows of length L each of which overlaps with its previous window by $L-1$ cells.

Fig. 5 shows the LLDA as a graphical model. In this model, each overlapped window is a mixture of latent topics called topic proportion

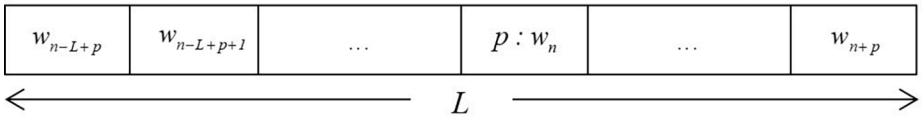

Fig. 4. Co-occurrence window. Position p in the window is where the target word is located.

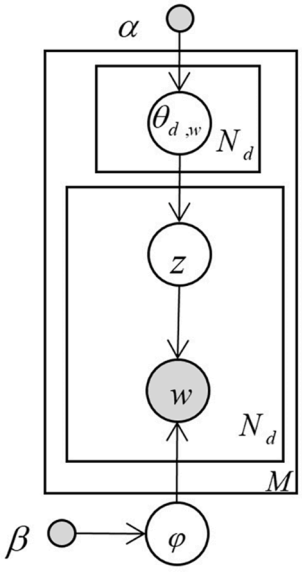

Fig. 5. Plate diagram for proposed model.

where the mixture weights follow a Dirichlet distribution with a symmetric parameter α. The mixture weights are denoted by θ where θ_{mn} is $p(z_{mn}|\theta)$. Each topic is a distribution over words denoted by φ where $\varphi_{z_{mn}w_{mn}}$ is $p(w_{mn}|z_{mn},\varphi)$ and follows a Dirichlet distribution with parameter β. The model assumes the following generative process for documents assuming φ $Dirichlet(\beta)$ is known for the corpus D:

- For each document d_m in the corpus
  - o For each overlapped window s_{mn} in the document
    - Choose $\theta_{mn}\widetilde{Dirichlet}(\alpha)$
    - For the target position of the window, choose a topic $z_{mn}\widetilde{multinomial}(\theta_{mn})$
    - Choose the target word $w_{mn}\widetilde{multinomial}(\varphi_{z_{mn}})$

Note that in each iteration of the algorithm which considers one window, only the topic of the target word for that window is sampled. However, the change to the assigned topic of a word, affects the topic distribution in all the windows containing that word. Assigning a topic to a word increases the likelihood of that topic in the windows covering the word. Therefore, the topic of each word directly affects the topic of its close neighbor words and transitively affects farther words. The farther the words are the milder the influence. This is contrary to LDA or LDCC in which, each word directly affects the topic of others in its document or segment ignoring their distance.

4.1. Inference and parameter estimation

In the proposed generative process, we assumed the topics are known. Reversing this procedure we have to find the topics that maximize the probability of words given φ, i.e., $p(w|\varphi)$ which according to the model is:

$$p(D|\varphi)=\prod_{m=1}^{M}\int\prod_{n=1}^{N_m}\left(\frac{\Gamma(\mathrm{K}\alpha)}{\Gamma(\alpha)^{\mathrm{K}}}\prod_{k=1}^{K}\theta_{m,n,k}^{\alpha-1}\right)\sum_{z_{mn}}\varphi_{z_{mn},w_{mn}}\theta_{m,n,z_{mn}}d\theta \tag{1}$$

The integral is intractable due to the coupling between φ and θ under summation, therefore φ has to be estimated using approximation algorithms. We used collapsed Gibbs sampling (Griffiths & Steyvers, 2004) because it is faster to converge compared to other common approximation methods like variational inference (Blei et al., 2003) and expectation propagation (Minka & Lafferty, 2002), despite being able to produce comparable results (Griffiths & Steyvers, 2004). For applying Gibbs sampling on the proposed model, $p(z_{xy}|z_{-xy},w)$ has to be calculated where x indicates a document, y is the position of the word in the document, and $-xy$ means every other position in the document x. According to the Bayes theorem:

$$p(z_{xy}|z_{-xy},w)=\frac{p(z_{xy},z_{-xy},w)}{p(z_{-xy},w)}\propto p(z_{xy},z_{-xy},w)=p(z,w) \tag{2}$$

With respect to the graphical representation of the model depicted in Fig. 5, we have the following equation, integrating out φ and θ.

$$p(z,w)=\int\int p(z,w,\theta,\varphi)d\theta d\varphi=\int p(\theta)p(z|\theta)d\theta\times\int p(\varphi)p(w|z,\varphi)d\varphi \tag{3}$$

Considering the conjugation between Dirichlet and Multinomial distributions gives the following equations:

$$\int p(\theta)p(z|\theta)d\theta=\left(\frac{\Gamma(\mathrm{K}\alpha)}{\Gamma(\alpha)^{\mathrm{K}}}\right)^{N}\prod_{m=1}^{M}\prod_{n=1}^{N_m}\frac{\prod_{k=1}^{K}\Gamma(\mathrm{n}_{\mathrm{k}}^{\mathrm{s_{m,n}}}+\alpha)}{\Gamma(\sum_{k=1}^{\mathrm{K}}\mathrm{n}_{\mathrm{k}}^{\mathrm{s_{m,n}}}+\alpha)} \tag{4}$$

where s_{mn} is a window of length L corresponding to the word w_{mn}, $n_k^{s_{m,n}}$ denotes the number of times a word in the window is assigned to topic k and $\Gamma(.)$ denotes the standard gamma function. Also:

$$\int(\varphi)p(w|z,\varphi)d\varphi=\left(\frac{\Gamma(|\mathrm{V}|\beta)}{\Gamma(\beta)^{|\mathrm{V}|}}\right)^{|V|}\prod_{k=1}^{K}\frac{\prod_{v=1}^{|V|}\Gamma(\mathrm{n}_{\mathrm{v}}^{\mathrm{k}}+\beta)}{\Gamma\left(\sum_{v=1}^{|\mathrm{V}|}\mathrm{n}_{\mathrm{v}}^{\mathrm{k}}+\beta\right)} \tag{5}$$

Using cancellation, we can obtain the following equation:

$$p(z_{xy}|z_{-xy},w)\propto\frac{n_{-xy,z_{x,y}}^{s_{x,y}}+\alpha}{L-1+K\alpha}\frac{n_{-xy,w_{xy}}^{z_{x,y}}+\beta}{n_{-xy,.}^{z_{x,y}}+|V|\beta} \tag{6}$$

where $n_{-xy,z_{x,y}}^{s_{x,y}}$ is the number of times a word in window $s_{x,y}$ has been assigned to topic $z_{x,y}$. Ignoring the current position, $n_{-xy,w_{xy}}^{z_{x,y}}$ is the number of times word $w_{x,y}$ has been assigned to topic $z_{x,y}$ in the whole corpus, and similarly $n_{-xy,.}^{z_{x,y}}$ is the number of times any word has been assigned to topic $z_{x,y}$ in the whole corpus. Now we can sample the new topic for each word in each document given all the other words and their topics in the document.

Fig. 6 shows the Gibbs sampling algorithm. In lines 9 and 14 in the algorithm, updating the topic of each target word, changes the counts for every overlapped window covering that word. This implies that the neighboring words of the target words are effective in counts and therefore affect the sampling of the new topic. In each state of the Markov chain φ and θ can be estimated as follows:

$$p(z_{mn}=k|\theta)=\theta_{mnk}\propto\frac{n_k^{s_{mn}}+\alpha}{L+K\alpha} \tag{7}$$

$$p(w=v|z=k,\varphi)=\varphi_{kv}\propto\frac{n_v^k+\beta}{n_\cdot^k+|V|\beta} \tag{8}$$

Randomly initialize topic assignment $z_{mn} \in \{1...K\}$ for each word w_{mn} in each document d_m ($n \in \{1...N_m\}$, $m \in \{1...M\}$)
Initialize topic counts, $n_k^{s_{mi}}$ to the number of times topic k has been assigned to any word in the sliding window s_{mn} for every window in every document
Initialize n_v^k to the number of times topic k has been assigned to word v in the corpus for every topic and every word
Initialize $n_\cdot^k$ to the number of times topic k has been assigned to any word in the corpus
Repeat up to $maxIterations$
    For each document d_m
        For each word w_{mn} in d_m
            Exclude the current word from $n_\cdot^{z_{mn}}$, $n_{w_{mn}}^{\tilde{z}_{mn}}$
            For all the overlapping windows containing w_{mn}
                Exclude current word from $n_{z_{mn}}^{s_{mi}}$
            Sample new topic from $p(z_{mn} \mid z_{-mn}, w)$ according to equation (6)
            Assign the new topic to w_{mn}
            Update $n_\cdot^{z_{mn}}$, $n_{w_{mn}}^{\tilde{z}_{mn}}$
            For all overlapping windows containing w_{mn}
                Update $n_{z_{mn}}^{s_{mi}}$

Fig. 6. Gibbs sampling algorithm for the proposed model.

5. Evaluations

In this section, we present a series of evaluations of LLDA in comparison with the three representative models LDA, BTM and LDCC. We first discuss the time complexity of the proposed model. Then focus on the experiments conducted using the mentioned model on two different datasets. We have performed three types of experiments to show the superiority of our model. First the proposed model is compared subjectively with the three mentioned models in terms of its generated topics. Then the coherences of the topics are measured which provides an intrinsic evaluation of the model and finally we investigate the ability of the model in the application of text clustering. For discussing our experiments, we first describe the dataset and the experiment settings and then the actual results are presented.

5.1. Time complexity

In this section, we compare the time complexity of Gibbs sampling in the proposed LLDA with LDA. The major time-consuming part in both algorithms is the sampling process which requires calculating the full-conditional probability $p(z_{xy}|z_{-xy}, w)$. For both algorithms, this part is executed in time $O(K)$ where k is the number of topics. This part is repeated for every word in the corpus. The total number of words in the corpus is denoted by N. Therefore, the total time required to train LDA is $O(NK)$ for a single iteration. In LLDA, every time a topic is sampled, the topics counts of every window overlapping with the corresponding window have to be updated, so the total time required for LLDA is $O(N(K + L))$. Since the order of L is usually lower than K, the execution time of LLDA is comparable to LDA.

5.2. Dataset

We use two different corpora: the 20newsgroups and the Reuters-R8. 20newsgroups consists of about 20,000 documents which are categorized in 20 categories according to their subjects. A number of 11,300 documents were used for training and the rest for testing. Ruters-R8 consists of 7674 documents from which 5485 were used for training and 2189 for testing. These documents are categorized into 8 categories. In both corpora, each document belongs only to a single category. Numbers, special characters, stopwords and words appearing in less than five documents were removed. Emails and web addresses have also been removed.

5.3. Experiment setting

Dirichlet parameters in hierarchical Dirichlet-multinomial models have smoothing effects. It is common in the topic modeling literature to consider symmetric hyperparameters as applied in this paper. Topics should be sparse, i.e., we prefer fewer words to be assigned to each topic. Sparseness of topics which is controlled by β determines how similar words have to be for the model to assign them to the same topic. Sparseness of β means the model prefers to describe each document with fewer topics. The values of β and α are related to the number of topics needed to specify the corpus. A model with sparser topics assumes

higher numbers of topics are needed to describe the corpus. So similar to what is suggested by (Griffiths, 2004) for LDA, we assume $\beta = 0.01$ and $\alpha = 1/K$. The same is used for BTM. For LDCC and LLDA, window size is another parameter influencing the results produced by the models. For bigger windows that cover more words, the model needs a smaller α to be able to describe windows by a sparse enough θ. Similar to our choices of hyperparameters values for the other two models, we assume $\beta = 0.01$ and $\alpha = 1/(K+L)$ for the LDCC and LLDA. Given these values, experiments are performed for several different numbers of topics to makes it possible to compare the four models.

In topic modeling, the model parameters are estimated in a manner that maximizes the probability of the corpus given the model. Therefore, the likelihood of the test set seems a suitable choice for evaluating a topic model. Perplexity is a widely-used metric which uses likelihood for evaluating a probabilistic model. Perplexity is the geometric mean of the likelihood of the test set given the model, and is computed by equation (9). A model with lower perplexity on the test set has a better generalization performance. Even though perplexity is not the best metric for comparing two different topic models (Chang et al., 2009; Lau et al., 2013), it has been used for model evaluation and selection (Griffiths & Steyvers, 2004; Newman et al., 2010).

$$perplexity(D_{test}) \;=\; exp - \frac{logp(w_{test}|\mathscr{M})}{N_{test}} \tag{9}$$

Obtaining the numerator in equation (10) involves summing over z (assigned topics) which is impractical. So we used importance sampling for estimating this value as suggested by Newton and Raftery (Newton & Raftery, 1994) and used widely in Monte Carlo methods (Griffiths and Steyvers 2004). In this approach S samples of z are derived from $p(z|w, \mathscr{M})$, and $P(w|\mathscr{M})$ is estimated as the harmonic mean of $p(w|z, \mathscr{M})$ over the derived samples.

$$P(w|\mathscr{M}) \simeq \mathrm{HM}\{\mathrm{P}(\mathrm{w}|\mathrm{z}^{(\mathrm{s})}, \mathscr{M})\}_{\mathrm{s}=1}^{\mathrm{S}} = \frac{\mathrm{S}}{\sum_{\mathrm{s}=1}^{\mathrm{S}} 1/\mathrm{p}(\mathrm{w}|\mathrm{z}^{(\mathrm{s})}, \mathscr{M})} \tag{10}$$

Fig. 7 shows the perplexity of the models as a function of Gibbs sampling iterations of the model according to a single Markov chain that has been run 1000 times over the two mentioned datasets. Perplexities are obtained for several iterations. Perplexities in these iterations have been obtained for 100 topics. As shown in Fig. 7, for both corpora, all four models are stabilized in 1000 iterations. However, LLDA, LDA, and BTM stabilize in less iterations than LDCC. The perplexity of our model is lower than LDA and LDCC. BTM has the lowest perplexity and converges more rapidly, but as discussed in the rest of the paper, this does not lead to more meaningful and coherent topics.

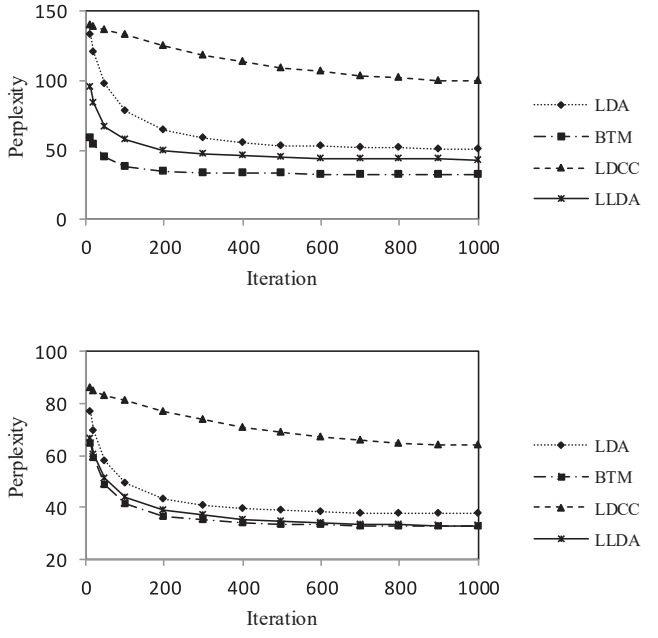

Fig. 7. Perplexity of the four models as a function of Gibbs sampling iterations on 20 newsgroups (top) and R8 (bottom).

We have also obtained the perplexity for several different window sizes in a similar way. The experiment is conducted on 10% of each corpus selected randomly. The results are shown in Fig. 8. As observed, on both datasets, initially the perplexity decreases significantly as the size of the window increases until L = 20. On R8, perplexity reaches its best value in L = 20 and increases thereafter while on 20 newsgroups, the perplexity continues to decrease in a very low pace, reaching its best value in L = 50 and increases after that. A similar pattern was observed for other number of topics where in all settings, the window size of 20 results in an acceptable model.

For the following experiments, 5 Markov chains were run 1000 times. The first 200 iterations were discarded, and then 8 samples were selected from each Markov chain. As mentioned before, we used symmetric hyperparameters. The experimental settings are summarized in Table 1.

In our experiments, the proposed model is compared with LDA, BTM, and LDCC. For LDCC to be comparable to the proposed model we assumed the segments have a fixed size identical to the window size in LLDA. According to our experiments and observations, using fixed-size windows in LDCC, generates more robust and acceptable results. Also, in LDCC we assumed that the number of word-topics is five times the number of segment-topics. We have also evaluated some other possible combinations for the number of topics, none of which resulted in significantly more coherent topics. The "number of topics" mentioned in the reports in this section is in fact the number of word-topics for LDCC.

5.4. Results

The proposed model is evaluated in three different experiments. First as we have discussed above, the convergences of the models are shown relative to perplexity changes in different iterations. Then some sample topics of the four models are compared subjectively, and the topic coherences of the models are evaluated. Finally, we also evaluated the proposed model in the application of document clustering compared to the other models. The experimental settings are summarized in Table 1.

Table 2 shows 10 samples of topics generated by the proposed model on the 20 newsgroups corpus, according to the mentioned settings, and for $K = 100$. Each topic is shown by means of its descriptors which

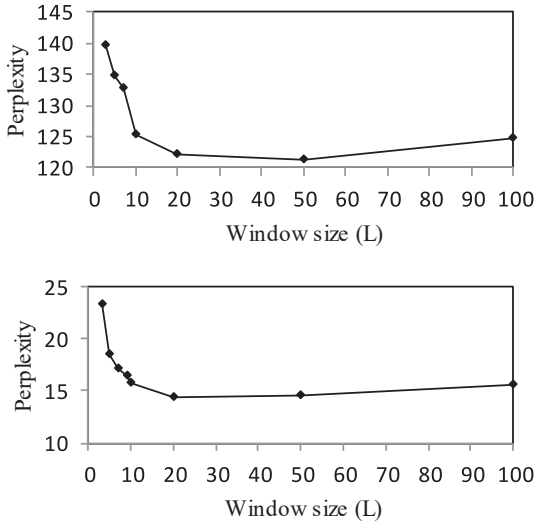

Fig. 8. Perplexity of LLDA as a function of the window size (L) on 20 newsgroups (top) and R8 (bottom).

Table 1
Experimental settings.

Parameter	Value
β	0.01
$\alpha_{LLDA,LDCC}$	$1/(K + L)$
$\alpha_{LDA,BTM}$	$1/K$
Number of iterations	1000
Burn-in	200
lag	100
L	20
P	10

appear in order of their importance in that topic. Fig. 9 and Fig. 10 show two sample documents of the corpus. In these figures, words are tagged with the topics that the proposed model has assigned to them. Words which have been removed during preprocess do not have a tag. One advantage of our model is the topic assignment. Consider the document of Fig. 9 that belongs to a category of the 20 newsgroups corpus which contains medical documents. The words assigned to topic 66 which is the main topic of this document are highlighted in the figure. Topic 66 is shown in Table 2. This topic can be named "Medicine". The other major topic, topic 31, in this document is also shown in Table 2. This topic is about scientific research and as seen, the document contains a debate about a medical claim and scientific investigation of that claim. In this example, one can see that our model has assigned sequences of adjacent words to the main topic of the document because they have appeared in the context of that topic.

Another example is shown in Fig. 10 which is a document from a category tagged as "atheism". In this example, the topics of the text are more colorful. The main topics of the first and second paragraphs are different. In the first paragraph, words tagged by topic 38, the main topic of the paragraph, are highlighted while in the second paragraph, topic 62 is highlighted as it is the main topic. Both these topics are shown in Table 2.

Table 5 each contain five samples of topics selected from 100 topics generated by LDA, LDCC, and BTM respectively, according to the mentioned settings in Table 1. These topics correspond to the first five topics of LLDA shown in Table 2. We tried to select the topics one instance of which are generated by all four models. For each such topic of LLDA, one topic from every other method is selected which has the most common words with the LLDA topic.

The proposed model generates more coherent topics than the other three as will objectively be shown in the next section, but it is also observable considering the examples here. Considering topic 27 of our model, one can conclude that this topic is about "electrical wiring". Topic 43 of LDA can be considered as the corresponding topic. However, this topic contains words like "subject" and "box", which are very common words, instead of more specific words "electrical" and "current" of the corresponding topic in LLDA. Topic 63 in LDCC corresponds to the mentioned topic. This topic contains words like "house", "hot" and "white" which not only are very common words, but also have appeared above words like "panel" and "wiring" which are more related to the topic. The topic does not have a match in BTM.

As another example, topics 34 and 95 of our model correspond to topic 24 of LDA. Topic 34 is about "Graphical User Interface (GUI) programming", topic 95 is about "coloring", and topic 24 of LDA is a combination of the two and also about GUI. It seems topic 24 in LDA has been divided in two more specific topics in our model. Topic 89 of LDCC and topic 0 of BTM correspond to this topic, however both contain general or even unrelated words to the topic and are less coherent than the corresponding topic of LLDA. There are many similar examples some of which are shown in Tables 2-5.

5.4.1. Topic coherence evaluation

Probabilistic topic models may generate many poor-quality topics, the number of which increases by increasing the number of topics (Callaghan et al., 2015). These topics are nonsensical or overgeneralized if judged by humans. Topic coherence (Mimno et al., 2011) is a metric introduced by Mimno et al. for measuring topic quality and is strongly correlated with human judgments. Using the equation below this metric evaluates topic coherence according to the document co-occurrence of the descriptor words of a topic in the subject corpus, considering topic descriptors rank in the topic.

Table 2
Examples of topics generated by the proposed model (LLDA) on 20 newsgroups corpus.

LLDA									
27	34	2	85	29	66	95	31	62	38
Ground	widget	price	engine	book	health	color	science	image	god
Wire	event	shipping	car	books	medical	visual	scientific	software	Atheists
Wiring	window	offer	gas	guide	aids	display	evidence	data	Atheism
Neutral	type	sale	battery	internet	disease	colormap	read	package	exist
Circuit	string	interest	fuel	reference	patients	image	jim	graphics	evidence
Outlets	call	ed	power	edition	national	colors	context	system	Religion
current	int	sell	problem	isbn	years	depth	work	version	belief
Connected	list	conditio	air	good	hiv	red	person	systems	people
Electrical	function	n	oil	press	children	default	fact	unix	religious
Wires	data	e-mailemail package	good	ed	research	window	article	processing	atheist

I don't doubt[31] that the placebo[66] effect[31] is alive[66] and well with medical[66] modality – estimated[66] by some to be around 20+%, but why would it be higher[66] with alternative[66] versus[66] conventional[66] medicine[31]? How do you know that it is? If you could show [31]this by careful[31] measurement[31], I suspect[31] you would have a paper[31] worthy[31] of publication[66] in a variety[66] of medical[66] journals[31]. …. Perhaps the study[66] could also include[66] how patients[66] respond[2] if they are dissatisfied[66] with a conventional[66] versus[66] an alternative[66] doctor[66], i.e. which practitioner[66] is more likely to get punched[66] in the face[44] when the success[66] of the treatment[66] doesn't meet[39] the expectations[44] of the patient[66].

Fig. 9. A sample document of the corpus and its topic assignment by the proposed model.

Science[31] is wonderful[16] at answering[31] most of our questions[14]. I'm not the type[80] to question[38] scientific[31] findings[31] very often, but... Personally[38], I find[31] the theory[38] of evolution[38] to be unfathomable. Could humans[80], a highly[31] evolved[80], complex[38] organism[80] that thinks[80], learns[14], and develops[14] truly be an organism[80] that resulted[38] from random[14] genetic[14] mutations and natural[80] selection[62]? Computers[62] are an excellent[62] example...of evolution[38] without "a" creator[38]. We did not "create[62]" computers[62]. We did not create[62] the sand[85] that goes into the silicon[62] that goes into the integrated[62] circuits[27] that go into processor[62] board[62]. We took these things[85] …

Fig. 10. Another sample document of the corpus and its topic assignment by the proposed model.

Table 3
Examples of topics generated by LDA on 20 newsgroups corpus.

LDA					
43	24	8	47	54	89
Ground	widget	sale	car	book	health
Wire	window	shipping	cars	good	medical
Wiring	visual	offer	engine	evil	center
Neutral	event	price	good	world	cancer
Outlets	application	printer	miles	time	aids
Wires	colormap	condition	drive	men	number
circuit	int	nntp-posting-host	writes	part	research
connected	set	email	driving	years	hiv
subject	display	interested	speed	church	april
box	code	cd	ford	history	newsletter

Table 4
Examples of topics generated by LDCC on 20 newsgroups corpus.

LDCC					
63	89	18	71	43	40
ground	application	shipping	car	questions	research
cable	call	sell	cars	mark	center
wire	widget	recently	price	longer	national
circuit	functions	level	front	energy	health
house	data	john	insurance	books	medical
hot	xt	disks	dealer	time	disease
connected	top	states	driving	question	cancer
panel	ax	united	miles	signature	patients
wiring	gl	ax	saturn	answers	service
white	resource	costs	told	ax	care

Table 5
Examples of topics generated by BTM on 20 newsgroups corpus.

BTM					
62	0	8	42	1	55
war	number	year	recently	good	care
military	color	distribution	cars	bible	people
secret	ago	cost	details	reading	account
time	part	works	weeks	books	average
audio	thing	megs	car	looked	medicine
attack	visual	events	drives	intended	read
provided	test	ide	government	read	state
team	system	world	tied	protection	program
ago	support	motto	hurt	fairly	ability
world	performance	tt	responses	post	regard

$$C(k, T^k) = \sum_{i=2}^{I} \sum_{j=1}^{i-1} log \frac{N(t_i^k, t_j^k) + 1}{N(t_j^k)} \tag{11}$$

where $C(k, T^k)$ is the coherence of topic k which is specified by an ordered list of its most probable words (topic descriptor) denoted by $T^k = (t_1^k, t_2^k, ..., t_I^k)$. $N(t_j^k)$ is the number of documents the word t_j^k has appeared in, for at least one time and $N(t_i^k, t_j^k)$ is the number of documents containing both t_i^k and t_j^k.

As mentioned before, short-distance co-occurrences are more likely to be indicators of meaningful and robust word relationships. As the proposed model relies more on short-distance co-occurrences, it is expected to generate more meaningful and coherent topics. We evaluated our proposed model by the coherence metric and compared it with LDCC, BTM, and LDA. Results of this experiment, for the 10 most probable words of the produced topics, are depicted in Fig. 11. As one can see in the figure, for all experiments with different numbers of topics, topic coherence of LLDA is higher than the three other methods. As mentioned before, we expect that the number of low-coherence topics to increase by increasing the number of topics, and thus the coherence is lower for higher numbers of topics in all the models. However, the degradation of coherence for LDA and LLDA are milder than the other two models on both corpora.

As mentioned above, our proposed model, LLDA, has generated the most coherent topics among all the mentioned models. The second best coherences belong to LDA. The coherence values of LLDA and LDA for the 20 newsgroups dataset, which are depicted in the upper diagram of

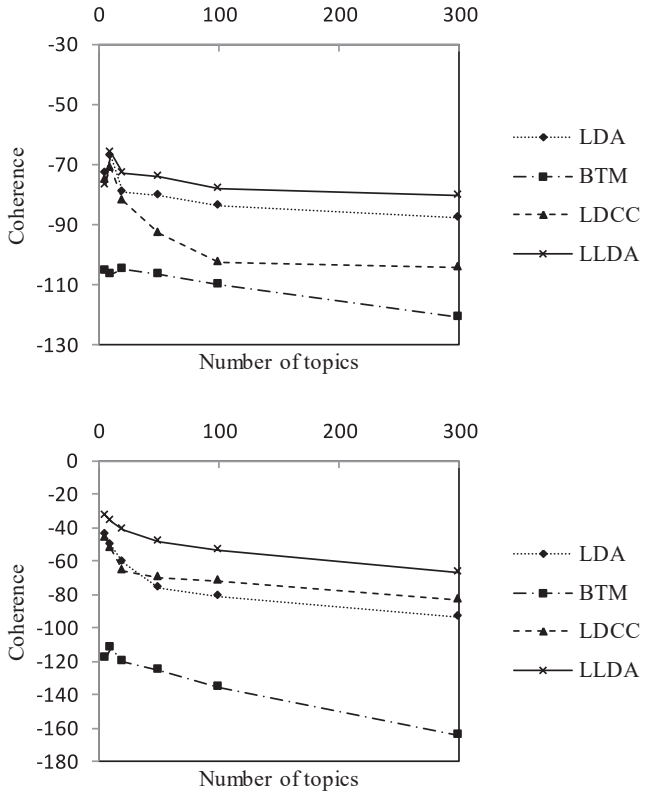

Fig. 11. Comparing mean topic coherence for several numbers of topics, on 20 newsgroups (top) and R8 (bottom).

Fig. 11, seem very close. We used two-sample *t*-test to show that these differences are neither small, nor insignificant. The values of coherences and the results of the hypothesis testing are reported in Table 6. We have also calculated the effect size using Cohen's d. The effect size proves that the differences are not only significant but also large.

The Cohen's d effect size for 20 newsgroups is 0.876 while it is 1.762 for R8. Although both fall under large category of Cohen's d, LLDA is more effective in R8. The vocabulary size of this corpus, after pre-processing, is equal to 22,923 words, and it consists of a total number of 510,106 words. These numbers for 20 newsgroups are respectively 23,710 and 2037192. According to this numbers, sparseness is more of a problem in R8 than 20 newsgroups, hence the larger effect of LLDA. This confirms the ability of LLDA in overcoming sparseness.

5.4.2. Document clustering

We expect the proposed model to be able to capture more meaningful word relationships and thus the document clusters generated by the proposed model to be more similar to the human-generated categories. We compared the clusters generated by the four models with the human-generated categories provided by the two corpora.

The topic proportions in topic models can be regarded as soft clustering of documents, while each document belongs to one single category in our corpora. Also number of topics may not be equal to number of categories. We have to use a metric which can handle the differences in number of clusters and type of clustering. Variation of information (VI) distance (Meila, 2007) is such a metric (Heinrich, 2008). VI-distance is calculated by equation (12), in which C and Z are two different cluster distributions. $H(C)$ and $H(Z)$ are entropies of C and Z respectively and $I(C,Z)$ is the mutual information between the two distributions which is calculated using Kullback-Leibler (KL) divergence, i.e., $I(C,Z) = D_{KL}\{p(c,z)||p(c)p(z)\}$. For more clarification, one can refer to (Heinrich, 2008).

$$D_{VI}(C,Z) = H(C) + H(Z) - 2I(C,Z) \tag{12}$$

In LDA and BTM, topic proportion can be considered as a soft clustering. For LDCC, segment-topics proportions can be considered as a soft clustering for documents (LDCC). Also, clusters can be formed based on word-topics by averaging over all segments in a document (LDCC_w). Both these approaches have been used in our experiments. For the proposed model, because the topic proportions are in window level and not the document level, they cannot be directly used for document clustering. Because we want to use a clustering approach for our model which is similar to the other three models, we generated a cluster (topic) distribution for each document by averaging topic proportions on all the overlapping windows located on each word in the document.

The cluster distributions generated for each one of the four models under the settings mentioned in Table 1, are compared to the corpora categories, and the obtained distances for several different numbers of topics are reported in Fig. 12. The number of topics in the figure is the number of word-topics for LDCC. As observed, the least distance according to word-topics belongs to our proposed model which means the clusters generated by LLDA are more similar to document categories formed by human judgments. This shows that LLDA is successful in generating meaningful topics by considering the local co-occurrences in a document.

Table 6
Coherence values for LDA and LLDA, with significance test and effect size results.

Coherences for 20 newsgroups					
Topic#	10	20	50	100	300
LDA	−66.983	−78.821	−80.498	−83.658	−87.798
LLDA	−65.760	−72.981	−74.238	−78.015	−80.203
H_0	$\mu_1 = \mu_2$				
T-Statistic	4.93	Two-tailed P-value		0.003936	
Cohen's d effect size		0.876			

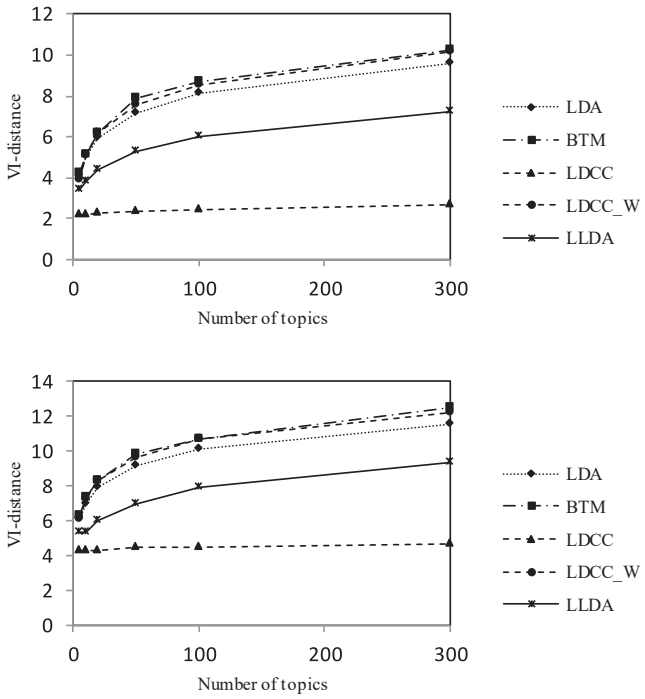

Fig. 12. Comparison of VI-distance for several values of K on 20 newsgroups (top) and R8 (bottom).

6. Conclusion

Traditional topic models, such as LDA, LDCC and BTM are used to learn the thematic structure of text corpora in expert systems, knowledge and information management systems, and many other applications. However, these topic models fail to effectively encode local word relationships and order information without suffering from sparseness. Such a drawback can limit the quality of the generated topics and the application of topic models where the word order information is a critical factor. Topic quality, or in other words coherence is important in applications where the topics are directly used for analysis, or when they are used as features in other tasks such as clustering and classification. In this paper, we proposed a new probabilistic topic model which is able to overcome this drawback.

Probabilistic topic models work based on word co-occurrences. Many of them like LDA use the whole document as the context for word co-occurrence, but according to the co-occurrence hypothesis, a shorter context indicates more robust and meaningful word relationships. Many other probabilistic topic models like BTM use shorter contexts but are restricted to use language models and word order. Such models are prone to sparseness especially for higher order language models. Some models like LDCC ignore word order but add another level of granularity by assuming a document as a bag of segments, each of which is a distribution over topics. In these kinds of models, local relationships are considered in a segment, but word exchangeability assumption is intact. We introduced a probabilistic topic model which uses overlapping windows as the co-occurrence context. The proposed model extracts word relationships in the document level, but emphasizes on the closer word relationships or in other words more local ones. Because the windows are overlapped, the word sequences are repeated in different windows. Therefore, the model can benefit mildly from word order information, and encode local word relationships without being severely prone to sparseness. Our experiments show that the proposed model extracts topics that are more coherent compared to LDA, LDCC, and BTM as representatives of the three mentioned types of topic models. The

proposed model also outperforms the other models in document clustering. This implies that it is possible to benefit from both a local context and word-order information while avoiding the issue of sparseness by using overlapping windows. And it is a successful scheme in improving the interpretability of the topics and their effectiveness in the fundamental task of document clustering. This scheme can also be used in other existing topic models to improve the coherence of the generated topics. We believe that the proposed model, in addition to text analysis applications, can be useful in modeling data like images and signals for which considering the spatial structure is vital in the modeling process.

CRediT authorship contribution statement

Marziea Rahimi: Conceptualization, Methodology, Formal analysis, Investigation, Software, Writing – original draft. **Morteza Zahedi:** Resources, Supervision, Validation. **Hoda Mashayekhi:** Supervision, Validation, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A

In this section, we will explain the inference process of the LLDA model with more details. As we have mentioned in the paper, we used collapsed Gibbs sampling to obtain the model parameters. To this end, $p(z_{xy}|z_{-xy}, w)$ has to be calculated where x indicates a document, and y is the position of the word in the document. In this section, to avoid a sophisticated notation, we use i instead of xy, where the i^{th} word of the corpus (i.e. w_i) is located in document d_i and window s_i. Also $-i$ means every other position in the corpus, and thus we need to calculate $p(z_i|z_{-i}, \boldsymbol{w})$. According to the Bayes theorem:

$$p(z_i = k|\boldsymbol{z}_{-i}, \boldsymbol{w}) = \frac{p(w_i, z_i = k|\boldsymbol{z}_{-i}, \boldsymbol{w}_{-i})}{p(w_i|\boldsymbol{z}_{-i}, \boldsymbol{w}_{-i})} \propto p(w_i, z_i = k|\boldsymbol{z}_{-i}, \boldsymbol{w}_{-i}) \tag{1}$$

According to the chain rule, the resulted term can be rewritten as follows:

$$p(w_i, z_i = k|\boldsymbol{z}_{-i}, \boldsymbol{w}_{-i}) = p(z_i = k|\boldsymbol{z}_{-i}, \boldsymbol{w}_{-i}) p(w_i|z_i = k, \boldsymbol{z}_{-i}, \boldsymbol{w}_{-i}) \tag{2}$$

According to the independencies indicated by the graphical model, the equation is rewritten as follows:

$$p(w_i, z_i = k|\boldsymbol{z}_{-i}, \boldsymbol{w}_{-i}) = p(z_i = k|\boldsymbol{z}_{-i}) p(w_i|z_i = k, \boldsymbol{z}_{-i}, \boldsymbol{w}_{-i}) \tag{3}$$

Now, each term has to be calculated. Integrating out θ which is defined over the windows, the first term can be rewritten as follows:

$$p(z_i|\boldsymbol{z}_{-i}) = \int p(z_i, \theta_{s_i}|\boldsymbol{z}_{-i}) d\theta_{s_i} = \int p(\theta_{s_i}|\boldsymbol{z}_{-i}) p(z_i|\theta_{s_i}) d\theta_{s_i} \tag{4}$$

Where,

$$p(\theta_{s_i}|z_{-i}) \propto p(\theta_{s_i}) p(z_{-i}|\theta_{s_i}) \tag{5}$$

We assumed that $p(\theta_{s_i})$ follows the Dirichlet distribution with the parameter α ($Dir(\alpha)$), and the Dirichlet distribution is the conjugate prior for the multinomial distribution. Thus the posterior distribution $p(\theta_{s_i}|z_{-i})$ above, which is the result of the product of a Dirichlet ($p(\theta_{s_i})$) and a multinomial ($p(z_{-i}|\theta_{s_i})$) component, follows the Dirichlet distribution with parameter $\alpha + n^{s_i}_{-i,k}$ i.e. $Dir\left(\alpha + n^{s_i}_{-i,k}\right)$ where $n^{s_i}_{-i,k}$ denotes the number of times any word in the window s_i of document d_i has been assigned the topic k excluding the word in position i(the current word). Thus the integral turns into the form $\int \theta_{s_i} p(\theta_{s_i}) d\theta_{s_i}$ which is $E[\theta_{s_i}]$. We have proven that $\theta_{s_i} \sim Dir\left(\alpha + n^{s_i}_{-i,k}\right)$ and thus the expectation is obtained by $\frac{n^{s_i}_{-i,k} + \alpha}{n^{s_i}_{-i,\cdot} + K\alpha}$ where $n^{s_i}_{-i,\cdot}$ is the number of words in window s_i excluding the current word and K is the number of topics. Therefore, we can obtain $p(z_i = k|z_{-i})$ as follows:

$$p(z_i = k|\boldsymbol{z}_{-i}) \propto \frac{n^{s_i}_{-i,k} + \alpha}{n^{s_i}_{-i,\cdot} + K\alpha} \tag{6}$$

For the second term of equation 3, integrating out parameter φ we have:

$$p(w_i|z_i = k,\ \boldsymbol{z}_{-i}, \boldsymbol{w}_{-i}) = \int p(w_i, \varphi|z_i = k,\ \boldsymbol{z}_{-i}, \boldsymbol{w}_{-i}) d\varphi \tag{7}$$

It can be rewritten as:

$$\int p(w_i, \varphi|z_i = k,\ \boldsymbol{z}_{-i}, \boldsymbol{w}_{-i}) d\varphi = \int p(\varphi|\ \boldsymbol{z}_{-i}, \boldsymbol{w}_{-i}) p(w_i|z_i = k, \varphi) d\varphi \tag{8}$$

The first term under the integral can be obtained by:

$$p(\varphi|\ \boldsymbol{z}_{-i}, \boldsymbol{w}_{-i}) \propto p(\varphi) p(\boldsymbol{w}_{-i}|\ \boldsymbol{z}_{-i}, \varphi) \tag{9}$$

We assumed that $p(\varphi)$ follows the Dirichlet distribution with the parameter β ($Dir(\beta)$). Thus the posterior probability $p(\varphi|\ \boldsymbol{z}_{-i}, \boldsymbol{w}_{-i})$ which is the result of the product of a Dirichlet ($p(\varphi)$) and a multinomial ($p(\boldsymbol{w}_{-i}|\ \boldsymbol{z}_{-i}, \varphi)$), follows the Dirichlet distribution with parameter $n^k_{-i,w_i} + \beta$ i.e. $Dir\left(n^k_{-i,w_i} + \beta\right)$ where n^k_{-i,w_i} denotes the number of times word w_i has been assigned the topic k excluding the current instance. Thus the integral takes on the form $\int \varphi p(\varphi) d\varphi$ which is $E[\varphi]$. We have proven that $\varphi \sim Dir\left(n^k_{-i,w_i} + \beta\right)$ and thus the expectation is obtained by $\frac{n^k_{-i,w_i} + \beta}{n^k_{-i,\cdot} + |V|\beta}$ where n^k_{-i,w_i} is the number of occurrences of word w_i excluding the current position, and V is the size of the vocabulary. Therefore, we can obtain $p(w_i|z_i = k, z_{-i}, \boldsymbol{w}_{-i})$ as follows:

$$p(w_i|z_i = k, \boldsymbol{z}_{-i}, \boldsymbol{w}_{-i}) \propto \frac{n^{k}_{-i,w_i} + \beta}{n^{k}_{-i,\cdot} + |V|\beta} \tag{10}$$

Finally, we can obtain $p(z_i|\boldsymbol{z}_{-i}, \boldsymbol{w})$ using equation 11.

$$p(z_i|z_{-i}, w) \propto \frac{n^{s_i}_{-i,z_i} + \alpha}{L - 1 + K\alpha} \times \frac{n^{z_i}_{-i,w_i} + \beta}{n^{z_i}_{-i,\cdot} + |V|\beta} \tag{11}$$

In each state of the Markov chain φ and θ can be estimated as follows:

$$p(z_i = k|\theta) = \theta_{ik} \propto \frac{n^{s_i}_{k} + \alpha}{L + K\alpha} \tag{12}$$

$$p(w_i = v|z_i = k, \varphi) = \varphi_{kv} \propto \frac{n^{k}_{v} + \beta}{n^{k}_{\cdot} + |V|\beta} \tag{13}$$

References

AL-Sharuee, M. T., Liu, F., & Pratama, M. (2018). Sentiment analysis: An automatic contextual analysis and ensemble clustering approach and comparison. *Data and Knowledge Engineering, 115*, 194–213. https://doi.org/10.1016/j.datak.2018.04.001

Balikas, G., Amoualian, H., Clausel, M., Gaussier, E., & Amini, M. R. (2016). Modeling topic dependencies in semantically coherent text spans with copulas. *COLING 2016 – 26th International Conference on Computational Linguistics, Proceedings of COLING 2016: Technical Papers*, 1767–1776.

Banerjee, A., Krumpelman, C., & Mooney, R. J. (2005). Model-based Overlapping Clustering. *Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, 532–537.

Barbieri, N., Manco, G., Ritacco, E., Carnuccio, M., & Bevacqua, A. (2013). Probabilistic topic models for sequence data. *Machine Learning, 93*(1), 5–29. https://doi.org/10.1007/s10994-013-5391-2

Bastani, K., Namavari, H., & Shaffer, J. (2019). Latent Dirichlet allocation (LDA) for topic modeling of the CFPB consumer complaints. *Expert Systems with Applications, 127*, 256–271. https://doi.org/10.1016/j.eswa.2019.03.001

Belford, M., Mac Namee, B., & Greene, D. (2018). Stability of topic modeling via matrix factorization. *Expert Systems with Applications, 91*, 159–169. https://doi.org/10.1016/j.eswa.2017.08.047

Biber, D. (1993). Co-occurrence Patterns among Collocations: A Tool for Corpus-Based Lexical Knowledge Acquisition. *Computational Linguistics, 19*(3), 531.

Blei, D. M., Carin, L., & Dunson, D. (2010). Probabilistic topic models. *IEEE Signal Processing Magazine, 27*(6), 55–65. https://doi.org/10.1109/MSP.2010.938079

Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models. *ACM International Conference Proceeding Series, 148*, 113–120. https://doi.org/10.1145/1143844.1143859

Blei, D. M., & McAuliffe, J. D. (2007). Supervised topic models. *Advances in Neural Information Processing Systems*, 121–128. http://papers.nips.cc/paper/3328-supervised-topic-models.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3*(4–5), 993–1022. https://doi.org/10.1016/b978-0-12-411519-4.00006-9

Callaghan, O., Callaghan, D. O., Greene, D., & Carthy, J. (2015). An Analysis of the Coherence of Descriptors in Topic Modeling. *Expert Systems with Applications, 42*(13), 5645–5657. https://doi.org/10.1016/j.eswa.2015.02.055

Cerda, P., & Varoquaux, G. (2020). Encoding high-cardinality string categorical variables. *IEEE Transactions on Knowledge and Data Engineering*, 1–1. https://doi.org/10.1109/tkde.2020.2992529.

Chang, J., Gerrish, S., Wang, C., & Blei, D. M. (2009). Reading tea leaves. *Advances in Neural Information Processing Systems*, 288–296.

Cheng, X., Yan, X., Lan, Y., & Guo, J. (2014). BTM: Topic modeling over short texts. *IEEE Transactions on Knowledge and Data Engineering, 26*(12), 2928–2941. https://doi.org/10.1109/TKDE.2014.2313872

Chipman, H. A., & Gu, H. (2005). *Interpretable Dimension Reduction., 32*(9), 969–987.

Deerwester, S., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). indexing by Latent semanting Analysis. *Journal of the American Society for Information Science, 41*(6), 391–407. https://doi.org/10.1017/CBO9781107415324.004

Djenouri, Y., Belhadi, A., Fournier-Viger, P., & Lin, J. C. W. (2018). Fast and effective cluster-based information retrieval using frequent closed itemsets. *Information Sciences, 453*, 154–167. https://doi.org/10.1016/j.ins.2018.04.008

Dong, L., Ji, S., Zhang, C., Zhang, Q., Qiu, L., Dong, L., … Qiu, L. (2018). An unsupervised topic-sentiment joint probabilistic model for detecting deceptive reviews. *Expert Systems with Applications, 114*, 210–223. https://doi.org/10.1016/j.eswa.2018.07.005

Drushku, K., Aligon, J., Labroche, N., Marcel, P., & Peralta, V. (2019). Interest-based recommendations for business intelligence users. *Information Systems, 86*, 79–93. https://doi.org/10.1016/j.is.2018.08.004

Du, L., Buntine, W., & Jin, H. (2010). A segmented topic model based on the two-parameter Poisson-Dirichlet process. *Machine Learning, 81*(1), 5–19. https://doi.org/10.1007/s10994-010-5197-4

Fuentes-pineda, G., & Meza-ruiz, I. V. (2019). Topic Discovery in Massive Text Corpora Based on Min-Hashing. *Expert Systems with Applications, 136*, 62–72.

Griffith, T. L., Tenenbaum, J. B., Jordan, M. I., & Blei, D. M. (2004). Hierarchical Topic Models and the Nested Chinese Restaurant Process. *Advances in Neural Information Processing Systems, 17–24*. https://doi.org/10.1016/0169-023X(89)90004-9

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America, 101*(SUPPL. 1), 5228–5235. https://doi.org/10.1073/pnas.0307752101

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review, 114*(2), 211–244. https://doi.org/10.1037/0033-295X.114.2.211

Ha, C., Tran, V. D., Ngo Van, L., & Than, K. (2019). Eliminating overfitting of probabilistic topic models on short and noisy text: The role of dropout. *International Journal of Approximate Reasoning, 112*, 85–104. https://doi.org/10.1016/j.ijar.2019.05.010

Hamadache, Z., & Sayoud, H. (2018). Authorship attribution of noisy text data with a comparative study of clustering methods. *International Journal of Knowledge and Systems Science, 9*(2), 45–69. https://doi.org/10.4018/IJKSS.2018040103

Harabagiu, S., & Lacatusu, F. (2005). Topic Themes for Multi-Document Summarization. *28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 202–209.

He, X., Xu, H., Li, J., He, L., & Yu, L. (2017). FastBTM: Reducing the sampling time for biterm topic model. *Knowledge-Based Systems, 132*, 11–20. https://doi.org/10.1016/j.knosys.2017.06.005

Heinrich, G. (2008). *Parameter estimation for text analysis*. http://www.arbylon.net/publications/text-est2.pdf.

Henrichs, A. (2019). Deforming Shakespeare's Sonnets: Topic models as poems. *Criticism, 61*(3), 387–412. https://doi.org/10.13110/criticism.61.3.0387

Jameel, S., Lam, W., & Bing, L. (2015). Supervised topic models with word order structure for document classification and retrieval learning. *Information Retrieval, 18*(4), 283–330. https://doi.org/10.1007/s10791-015-9254-2

Jeong, Y. S., & Choi, H. J. (2015). Overlapped latent Dirichlet allocation for efficient image segmentation. *Soft Computing, 19*(4), 829–838. https://doi.org/10.1007/s00500-014-1410-x

Jiang, Y., Tao, D., Liu, Y., Sun, J., & Ling, H. (2019). Cloud service recommendation based on unstructured textual information. *Future Generation Computer Systems, 97*, 387–396. https://doi.org/10.1016/j.future.2019.02.063

Kang, D., Park, Y., & Chari, S. N. (2014). Hetero-labeled LDA: A partially supervised topic model with heterogeneous labels. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 8724 LNAI*(PART 1), 640–655. https://doi.org/10.1007/978-3-662-44848-9_41.

Kim, S., & Yoon, J. (2015). Link-topic model for biomedical abbreviation disambiguation. *Journal of Biomedical Informatics, 53*, 367–380. https://doi.org/10.1016/j.jbi.2014.12.013

Kim, Y., & Shim, K. (2014). TWILITE : A recommendation system for Twitter using a probabilistic model based on latent Dirichlet allocation. *Information Systems, 42*, 59–77. https://doi.org/10.1016/j.is.2013.11.003

Lau, J. H., Baldwin, T., & Newman, D. (2013). On collocations and topic models. *ACM Transactions on Speech and Language Processing, 10*(3), 1–14. https://doi.org/10.1145/2483969.2483972

Li, Q., Liu, L., Xu, M., Wu, B., & Xiao, Y. (2019). GDTM: A Gaussian Dynamic Topic Model for Forwarding Prediction under Complex Mechanisms. *IEEE Transactions on Computational Social Systems, 6*(2), 338–349. https://doi.org/10.1109/TCSS.2019.2900299

Liu, Q., Zheng, Z., Zheng, J., Chen, Q., Liu, G., Chen, S., … Ming, W. K. (2020). Health communication through news media during the early stage of the covid-19 outbreak in China: Digital topic modeling approach. *Journal of Medical Internet Research, 22*(4). https://doi.org/10.2196/19118

Ma, Y., Wang, Y., & Jin, B. (2014). A three-phase approach to document clustering based on topic significance degree. *Expert Systems with Applications, 41*(18), 8203–8210. https://doi.org/10.1016/j.eswa.2014.07.014

Mallick, C., Das, A. K., Dutta, M., Das, A. K., & Sarkar, A. (2018). Graph-Based Text Summarization Using Modified TextRank. *Soft Computing in Data Analytics*, 137–146.

Meila, M. (2007). Comparing clusterings — an information based distance. *Journal of Multivariate Analysis, 98*, 873–895. https://doi.org/10.1016/j.jmva.2006.11.013

Mimno, D., Wallach, H. M., Talley, E., Leenders, M., & McCallum, A. (2011). Optimizing semantic coherence in topic models. *EMNLP 2011 – Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, *2*, 262–272.

Minka, T., & Lafferty, J. (2002). Expectation-Propagation for the Generative Aspect Model. *Uncertainty in Artificial Intelligence*, 352–359. https://doi.org/ISBN 1-55860-897-4.

Misra, H., Jose, J. M., & Cappé, O. (2009). Text Segmentation via Topic Modeling : An Analytical Study. *18th ACM Conference on Information and Knowledge Management.*, 1553–1556.

Misra, H., Yvon, F., Cappé, O., & Jose, J. (2011). Text segmentation: A topic modeling perspective. *Information Processing and Management, 47*(4), 528–544. https://doi.org/10.1016/j.ipm.2010.11.008

Nesselhauf, N. (2005). Structural and Functional Properties of Collocations in English. A Corpus Study of Lexical and Pragmatic Constraints on Lexical Co-occurrence. International Journal of Corpus. *Linguistics, 10*(2), 266–270. https://doi.org/10.1075/ijcl.10.2.08nes

Newman, D., Lau, J. H., Grieser, K., & Baldwin, T. (2010). *Automatic Evaluation of Topic Coherence*. *June*, 100–108.

Newton, M. A., & Raftery, A. E. (1994). Approximate Bayesian Inference with the Weighted Likelihood Bootstrap. *In Journal of the Royal Statistical Society: Series B (Methodological) (Vol., 56*(1), 3–26. https://doi.org/10.1111/j.2517-6161.1994.tb01956.x

Nguyen, D. Q., Billingsley, R., Du, L., & Johnson, M. (2015). Improving Topic Models with Latent Feature Word Representations. *Transactions of the Association for Computational Linguistics, 3*, 598–599. https://doi.org/10.1162/tacl_a_00245

Noji, H., Mochihashi, D., & Miyao, Y. (2013). Improvements to the Bayesian topic N-gram models. *EMNLP 2013 – 2013 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, *October*, 1180–1190.

Onan, A., Bulut, H., & Korukoglu, S. (2017). An improved ant algorithm with LDA-based representation for text document clustering. *Journal of Information Science, 43*(2), 275–292. https://doi.org/10.1177/0165551516638784

Ou, W., Xie, Z., & Lv, Z. (2015). Spatially Regularized Latent topic Model for Simultaneous object discovery and segmentation. *IEEE International Conference on Systems, Man, and Cybernetics, 2938–2943*. https://doi.org/10.1109/SMC.2015.511

Pang, J., Rao, Y., Xie, H., Wang, X., Wang, F. L., Wong, T.-L., & Li, Q. (2019). Fast Supervised Topic Models for Short Text Emotion Detection. *IEEE Transactions on Cybernetics, PP, 1–14*. https://doi.org/10.1109/tcyb.2019.2940520

Panicheva, P., Litvinova, O., & Litvinova, T. (2019). Author clustering with and without topical features. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Vol. 11658 LNAI*. Springer International Publishing. https://doi.org/10.1007/978-3-030-26061-3_36.

Park, H., Park, T., & Lee, Y. (2019). Partially collapsed Gibbs sampling for latent Dirichlet allocation. *Expert Systems with Applications, 131*, 208–218. https://doi.org/10.1016/j.eswa.2019.04.028

Pavlinek, M., & Podgorelec, V. (2017). Text classification method based on self-training and LDA topic models. *Expert Systems with Applications, 80*, 83–93. https://doi.org/10.1016/j.eswa.2017.03.020

Qiang, J., Qian, Z., Li, Y., Yuan, Y., & Wu, X. (2020). Short Text Topic Modeling Techniques, Applications, and Performance: A Survey. *IEEE Transactions on Knowledge and Data Engineering, 14*(8), 1–1. https://doi.org/10.1109/tkde.2020.2992485.

Ramage, D., Hall, D., Nallapati, R., & Manning, C. D. (2009). Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. *EMNLP 2009 – Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: A Meeting of SIGDAT, a Special Interest Group of ACL, Held in Conjunction with ACL-IJCNLP 2009*, *August*, 248–256.

Reda, M., Sanner, S., & Du, Y. (2020). Relevance- and interface-driven clustering for visual information retrieval. *Information Systems, 94*, Article 101592. https://doi.org/10.1016/j.is.2020.101592

Rehioui, H., & Idrissi, A. (2020). New clustering algorithms for twitter sentiment analysis. *IEEE Systems Journal, 14*(1), 530–537. https://doi.org/10.1109/JSYST.2019.2912759

Riaz, S., Fatima, M., Kamran, M., & Nisar, M. W. (2019). Opinion mining on large scale data using sentiment analysis and k-means clustering. *Cluster Computing, 22*, 7149–7164. https://doi.org/10.1007/s10586-017-1077-z

Rouane, O., Belhadef, H., & Bouakkaz, M. (2019). Combine clustering and frequent itemsets mining to enhance biomedical text summarization. *Expert Systems with Applications, 135*, 362–373. https://doi.org/10.1016/j.eswa.2019.06.002

Rubin, T. N., Chambers, A., Smyth, P., & Steyvers, M. (2012). Statistical topic models for multi-label document classification. *Machine Learning, 88*(1–2), 157–208. https://doi.org/10.1007/s10994-011-5272-5

Rugeles, D., Hai, Z., Dash, M., & Cong, G. (2020). Deterministic Inference of Topic Models via Maximal Latent State Replication. *IEEE Transactions on Knowledge and Data Engineering, XX*(XX), 1–1. https://doi.org/10.1109/tkde.2020.3000559.

Sato, I., & Nakagawa, H. (2010). Topic models with power-law using pitman-yor process. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1*, 673–681. https://doi.org/10.1145/1835804.1835890

Schnober, C., & Gurevych, I. (2015). Combining Topic Models for Corpus Exploration. *The 2015 Workshop on Topic Models: Post-Processing and Applications*, 11–20.

Schulte Im Walde, S., & Melinger, A. (2008). An in-depth look into the co-occurrence distribution of semantic associates. *Italian Journal of Linguistics*, *20*(1), 89–128.

Shafiei, M. M., & Muios, E. E. (2006). Latent dirichlet co-clustering. *Proceedings – IEEE International Conference on Data Mining, ICDM, 542–551*. https://doi.org/10.1109/ICDM.2006.94

Sinoara, R. A., Camacho-Collados, J., Rossi, R. G., Navigli, R., & Rezende, S. O. (2019). Knowledge-enhanced document embeddings for text classification. *Knowledge-Based Systems, 163*, 955–971. https://doi.org/10.1016/j.knosys.2018.10.026

Stamatatos, E., Tschuggnall, M., Verhoeven, B., Daelemans, W., Specht, G., Stein, B., & Potthast, M. (2016). Clustering by authorship within and across documents. *CEUR Workshop Proceedings, 1609*, 691–715.

Stokes, D. C., Andy, A., Guntuku, S. C., Ungar, L. H., & Merchant, R. M. (2020). Public Priorities and Concerns Regarding COVID-19 in an Online Discussion Forum: Longitudinal Topic Modeling. *Journal of General Internal Medicine, 35*(7), 2244–2247. https://doi.org/10.1007/s11606-020-05889-w

Teh, Y. W. (2006). A hierarchical Bayesian language model based on Pitman-Yor processes. *COLING/ACL 2006 – 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, *1*(July), 985–992. https://doi.org/10.3115/1220175.1220299.

Thrun, S., Mitchell, T. O. M., Nigam, K., & McCallum, A. (2000). Text classification from labeled and unlabeled documents using EM. *Machine Learning, 34*, 103–134. http://www.springerlink.com/index/p4324q3673265225.pdf.

Wallach, H. M. (2006). Topic Modeling : Beyond Bag-of-Words. *23rd International Conference on Machine Learning*, *1*, 977–984.

Wang, C., & Blei, D. M. (2011). Collaborative topic modeling for recommending scientific articles. *KDD*.

Wang, D., Zhu, S., Li, T., & Gong, Y. (2009). Multi-document summarization using sentence-based topic models. *ACL-IJCNLP 2009 – Joint Conf. of the 47th Annual Meeting of the Association for Computational Linguistics and 4th Int. Joint Conf. on Natural Language Processing of the AFNLP, Proceedings of the Conf.*, *August*, 297–300. https://doi.org/10.3115/1667583.1667675.

Wang, R., Zhou, D., & He, Y. (2020). Optimising topic coherence with Weighted Pólya Urn scheme. *Neurocomputing, 385*, 329–339. https://doi.org/10.1016/j.neucom.2019.12.013

Wang, X., & Grimson, E. (2008). *Spatial Latent Dirichlet Allocation*. 1577–1584.

Xu, Y., Yin, J., Huang, J., & Yin, Y. (2018). Hierarchical topic modeling with automatic knowledge mining. *Expert Systems with Applications, 103*, 106–117. https://doi.org/10.1016/j.eswa.2018.03.008

Yang, G., Wen, D., Kinshuk, Chen N. S., & Sutinen, E. (2015). A novel contextual topic model for multi-document summarization. *Expert Systems with Applications, 42*(3), 1340–1352. https://doi.org/10.1016/j.eswa.2014.09.015

Yuan, J., Gao, F., Ho, Q., Dai, W., Wei, J., Zheng, X., Xing, E. P., Liu, T., & Ma, W. (2015). LightLDA : Big Topic Models on Modest Computer Clusters Categories and Subject Descriptors. *24th International Conference on World Wide Web*, *1*, 1351–1361.

Zhang, H., & Zhong, G. (2016). Improving short text classification by learning vector representations of both words and hidden topics. *Knowledge-Based Systems, 102*, 76–86. https://doi.org/10.1016/j.knosys.2016.03.027

Zhao, H., Jiang, B., Canny, J., & Jaros, B. (2015). SAME but Different: Fast and high-quality gibbs parameter estimation. *The ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1495–1502*. https://doi.org/10.1145/2783258.2783416

Zhu, L., He, Y., & Zhou, D. (2019). *Hierarchical Viewpoint Discovery from Tweets Using Bayesian Modelling., 116*, 430–438.

Zuo, Y., Zhao, J., & Xu, K. (2016). Word network topic model: A simple but general solution for short and imbalanced texts. *Knowledge and Information Systems, 48*(2), 379–398. https://doi.org/10.1007/s10115-015-0882-z