Topic Embedding for Documents

July 8, 2015

1 Introduction

In the previous chapter, a generative word embedding model is presented, along with a learning algorithm to find a set of word embeddings. In this chapter, we extend this model by incorporating topics of a document, into this generative model, and develop a continuous counterpart of Latent Dirichlet Allocation (LDA). Through learning the latent topics, the semantics of a document will be summarized as a few topic vectors, which could be used in different applications.

2 Notations

We assume each word in a document is semantically similar to a *topic embedding* in the embedding space. We often refer to topic embeddings simply as *topic*s. Specifically, each document has K candidate topics, arranged in the matrix form $\boldsymbol{T}_i = (\boldsymbol{t}_{i1} \cdots \boldsymbol{t}_{iK})$, referred to as the *topic matrix*. Particularly, we fix $\boldsymbol{t}_{i1} = \boldsymbol{0}$, referred to as the *null topic*. As there are many words which have no obvious semantics, these words can be assigned to this null topic. Similar to words, each topic $\boldsymbol{t}_{ik}$ accompanies a residual $r_{i,k}$. In addition, there is a topic weight β, a hyperparameter controling their degree of impact to the distribution of words.

The above assumption that each word is semantically similar to a topic, is formulated as follows. In a document d_i, each word w_{ij} is assigned to a topic indexed by $z_{ij} \in \{1, \cdots, K\}$. Geometrically this means the embedding $\boldsymbol{v}_{w_{ij}}$ tends to align with the direction of $\boldsymbol{t}_{i,z_{ij}}$. Each topic $\boldsymbol{t}_{ik}$ has a document-specific prior probability to be assigned to a word, denoted as $\phi_{ik} = P(k|d_i)$. The vector $\boldsymbol{\phi}_i = (\phi_{i1}, \cdots, \phi_{iK})$ is referred to as the *mixing proportions* of these topics in document d_i. As in LDA, $\boldsymbol{\phi}_i$ is governed by a Dirichlet prior $\mathrm{Dir}(\boldsymbol{\alpha})$.

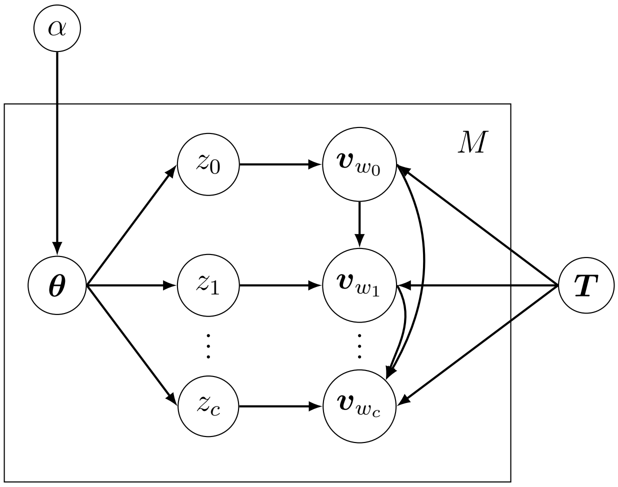

Figure 1: The Graphical Model of Topic Embedding

3 Distribution of a Text Window Parameterized by Word and Topic Embeddings

3.1 Conditional Distribution of a Word Given Context and Topic

Using the similar idea, we extend eq.(7) in [1] to incorporate the impact of the topic:

$$P(w_c \mid w_0{:}w_{c-1}, z_c, d_i) = P(w_c) \exp\left\{ \boldsymbol{v}_{w_c}^\top \left(\sum_{i=0}^{c-1} \boldsymbol{v}_{w_i} + \beta \boldsymbol{t}_{i,z_c} \right) + \sum_{i=0}^{c-1} a_{w_i w_c} + r_{i,z_c} \right\}, \tag{1}$$

where d_i is the current document, and $\beta > 0$ is a hyperparameter, named the *topic weight*, controling their degree of impact to the distribution of w_c. The topic residual r_{i,z_c} only depends on the topic assignment z_c, but not on the value of w_c.

The topic weight β determines the “polarity” of the topics: a bigger β means that if a word is assigned to topic k, then its embedding is more strongly driven towards the direction of $\boldsymbol{t}_{ik}$. In particular, when $\beta = 0$, our model reduces to a model without topics.

This equation is equivalent to

$$\log \frac{P(w_c \mid w_0{:}w_{c-1}, z_c, d_i)}{P(w_c)} = \boldsymbol{v}_{w_c}^\top \left(\sum_{i=0}^{c-1} \boldsymbol{v}_{w_i} + \beta \boldsymbol{t}_{i,z_c} \right) + \sum_{i=0}^{c-1} a_{w_i w_c} + r_{i,z_c}. \tag{2}$$

In order to estimate r_{ik}, we let the context size $c = 0$ and $z_c = k$, and then (1) becomes:

$$P(s_j \mid k, d_i) = P(s_j) \exp\left\{\beta \boldsymbol{v}_{s_j}^{\top} \boldsymbol{t}_{ik} + r_{ik}\right\}. \tag{3}$$

It is required that $\sum_{s_j \in \boldsymbol{S}} P(s_j \mid k, d_i) = 1$ to make (3) a distribution. It follows that

$$r_{ik} = -\log\Big(\sum_{s_j \in \boldsymbol{S}} P(s_j) \exp\{\beta \boldsymbol{v}_{s_j}^{\top} \boldsymbol{t}_{ik}\}\Big). \tag{4}$$

That is, r_{ik} is uniquely determined by β and $\boldsymbol{t}_{ik}$. Specifically, when $\beta = 0$, $r_{ik} = 0$. Remind that when $\forall i, \boldsymbol{t}_{i1} = 0$, and thus $r_{i1} = 0$.

Our decision of making r_{ik} invariant to different values of w_c is a trade-off between computational efficiency and modeling accuracy. Intuitively, the distribution of w_c is primarily determined by its context $w_0{:}w_{c-1}$, and less influenced by the topic $\boldsymbol{t}_{ik}$. Then the magnitude of $\beta \boldsymbol{v}_{w_c}^{\top} \boldsymbol{t}_{ik} + r_{ik}$ should usually be smaller than the that of the context vectors. Within this expression, the magnitude of r_{ik} should also be smaller than the residuals between two words. As such, approximating it by a constant value will not result in big errors of the distribution of w_c.

4 The Generative Process

Now we have proposed the basic distributions of the words. Before the generative process begins, a few hyperparameters need to be specified:

1. The parameter $\boldsymbol{\alpha}$ of the Dirichlet prior of the mixing proportions $\boldsymbol{\phi}_i$, $\mathrm{Dir}(\boldsymbol{\alpha})$;

2. The topic weight β;

The generative process is as follows:

1. Draw the residual matrix $\boldsymbol{A}$ from the Truncated Gaussian prior $\mathcal{N}_{\mathrm{Fea}(\boldsymbol{G},N)}(\boldsymbol{A}; 0, \boldsymbol{H})$;

2. Draw the embeddings $\boldsymbol{V}$ uniformly from the solution set $\mathrm{Sol}(\boldsymbol{V}; \boldsymbol{G}, \boldsymbol{A})$, of $\boldsymbol{V}^{\top}\boldsymbol{V} = \boldsymbol{G} - \boldsymbol{A}$;

3. For each document d_i:

   (a) Draw the mixing proportions $\boldsymbol{\phi}_i$ from the Dirichlet prior $\mathrm{Dir}(\boldsymbol{\alpha})$;

   (b) For the j-th word, do the following:

      i. Draw topic assignment z_{ij} from the categorical distribution $\mathrm{Cat}(\boldsymbol{\phi}_i)$;

      ii. Draw word w_{ij} with probability $P(w_{ij} \mid w_{i,j-c}{:}w_{i,j-1}, z_{ij}, d_i)$.

5 Likelihood Function

Given the embeddings $\boldsymbol{V}$ and the bigram residuals $\boldsymbol{A}$, the topics $\boldsymbol{T}$ and the hyperparamters $\boldsymbol{\alpha}, \beta$, the complete-data likelihood of a document d_i is:

$$
\begin{aligned}
&p(d_i, \boldsymbol{Z}_i, \boldsymbol{\phi}_i | \boldsymbol{\alpha}, \beta, \boldsymbol{V}, \boldsymbol{A}, \boldsymbol{T}_i) \\
=&p(\boldsymbol{\phi}_i|\boldsymbol{\alpha}) p(\boldsymbol{Z}_i|\boldsymbol{\phi}_i) p(d_i|\beta, \boldsymbol{V}, \boldsymbol{A}, \boldsymbol{T}_i, \boldsymbol{Z}_i) \\
=&\frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{j=1}^K \phi_{ij}^{\alpha_j - 1} \cdot \prod_{j=1}^{L_i} \Bigg(\phi_{i,z_{ij}} P(w_{ij}) \\
&\cdot \exp\Bigg\{ \boldsymbol{v}_{w_{ij}}^\top \Big(\sum_{k=j-c}^{j-1} \boldsymbol{v}_{w_{ik}} + \beta \boldsymbol{t}_{z_{ij}} \Big) + \sum_{k=j-c}^{j-1} a_{w_{ik} w_{ij}} + r_{i,z_{ij}} \Bigg\} \Bigg),
\end{aligned}
\tag{5}
$$

where $\boldsymbol{Z}_i = (z_{i1}, \cdots, z_{iL_i})$, and $\Gamma(\cdot)$ is the Gamma function. The topic residuals $\boldsymbol{r}_i = \{r_{ik}\}_k$ are uniquely determined by $\boldsymbol{T}_i$ and β, and thus are implicit in the likelihood functions.

We denote the latent variables of all documents $\{\boldsymbol{Z}_i\}_{i=1}^M$ collectively by $\boldsymbol{Z}$, and all the document-specific $\{\boldsymbol{\phi}_i\}_{i=1}^M$ by $\boldsymbol{\phi}$. Then the complete-data likelihood of the whole corpus is:

$$
\begin{aligned}
&p(\boldsymbol{D}, \boldsymbol{B}, \boldsymbol{A}, \boldsymbol{V}, \boldsymbol{Z}, \boldsymbol{\phi} | \boldsymbol{\alpha}, \beta, \boldsymbol{T}) \\
=&\mathcal{N}_{\mathrm{Fea}(\boldsymbol{G},N)}(\boldsymbol{A}; 0, \boldsymbol{H}) \cdot U(\mathrm{Sol}(\boldsymbol{V}; \boldsymbol{G}, \boldsymbol{A})) \\
&\cdot \prod_{i=1}^M \{ p(\boldsymbol{\phi}_i|\boldsymbol{\alpha}) p(\boldsymbol{Z}_i|\boldsymbol{\phi}_i) p(d_i|\beta, \boldsymbol{V}, \boldsymbol{A}, \boldsymbol{T}_i, \boldsymbol{Z}_i) \} \\
=&\frac{1}{\mathcal{Z}(\boldsymbol{A}, \boldsymbol{V}; \boldsymbol{B})} \exp\{ - \sum_{i,j=1}^{W,W} f(h_{i,j}) a_{s_i s_j}^2 \} \prod_{i=1}^M \Bigg\{ \frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{j=1}^K \phi_{ij}^{\alpha_j - 1} \\
&\cdot \prod_{j=1}^{L_i} \Bigg(\phi_{i,z_{ij}} P(w_{ij}) \cdot \exp\Bigg\{ \boldsymbol{v}_{w_{ij}}^\top \Big(\sum_{k=j-c}^{j-1} \boldsymbol{v}_{w_{ik}} + \beta \boldsymbol{t}_{z_{ij}} \Big) + \sum_{k=j-c}^{j-1} a_{w_{ik} w_{ij}} + r_{i,z_{ij}} \Bigg\} \Bigg) \Bigg\},
\end{aligned}
\tag{6}
$$

where $U(\mathrm{Sol}(\boldsymbol{V}; \boldsymbol{G}, \boldsymbol{A}))$ is a uniform distribution over $\mathrm{Sol}(\boldsymbol{V}; \boldsymbol{G}, \boldsymbol{A})$, and $\mathcal{Z}(\boldsymbol{A}, \boldsymbol{V}; \boldsymbol{B})$ is the normalizing function of $\mathcal{N}_{\mathrm{Fea}(\boldsymbol{G},N)}(\boldsymbol{A}; 0, \boldsymbol{H}) \cdot U(\mathrm{Sol}(\boldsymbol{V}; \boldsymbol{G}, \boldsymbol{A}))$:

$$
\mathcal{Z}(\boldsymbol{B}, \boldsymbol{A}, \boldsymbol{V}) = \int_{\mathrm{Fea}(\boldsymbol{G},N)} \exp\{ -||\boldsymbol{A}||_{f(\boldsymbol{H})}^2 \} \cdot \lambda(\mathrm{Sol}(\boldsymbol{V}; \boldsymbol{G}, \boldsymbol{A})) d\boldsymbol{A}, \tag{7}
$$

where $\lambda(\mathrm{Sol}(\boldsymbol{V}; \boldsymbol{G}, \boldsymbol{A}))$ is the Lebesgue measure of $\mathrm{Sol}(\boldsymbol{V}; \boldsymbol{G}, \boldsymbol{A})$.

Taking the logarithm of both sides, we obtain

$$
\begin{aligned}
&\log p(\boldsymbol{D}, \boldsymbol{B}, \boldsymbol{A}, \boldsymbol{V}, \boldsymbol{Z}, \boldsymbol{\phi}|\boldsymbol{\alpha}, \beta, \boldsymbol{T}) \\
=&C_0 - \log \mathcal{Z}(\boldsymbol{B}, \boldsymbol{A}, \boldsymbol{V}) - ||\boldsymbol{A}||^2_{f(\boldsymbol{H})} + \sum_{i=1}^{M} \Bigg\{ \log \phi_{ik} \cdot \sum_{k=1}^{K} (m_{ik} + \alpha_{0k} - 1) \\
&+ \sum_{j=1}^{L_i} \Bigg(\boldsymbol{v}_{w_{ij}}^{\top} \Big(\sum_{k=j-c}^{j-1} \boldsymbol{v}_{w_{ik}} + \beta \boldsymbol{t}_{z_{ij}} \Big) + \sum_{k=j-c}^{j-1} a_{w_{ik} w_{ij}} + r_{i, z_{ij}} \Bigg) \Bigg\},
\end{aligned} \tag{8}
$$

where $m_{ik} = \sum_{j=1}^{L_i} \delta(z_{ij} = k)$ counts the number of words assigned with the k-th topic in d_i, $C_0 = M \log \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} + \sum_{i,j=1}^{M,L_i} \log P(w_{ij})$ is constant given $\boldsymbol{\alpha}$.

6 Two Stage Learning Algorithm

6.1 Learning Objective and Process

Given the hyperparameters $\boldsymbol{\alpha}, \beta$, the learning objective is to find the estimates of the bigram probabilities $\boldsymbol{B}$, the embeddings and residuals $\boldsymbol{V}, \boldsymbol{A}$, the topics $\boldsymbol{T}$, and the word-topic and document-topic distributions $p(\boldsymbol{Z}_i, \boldsymbol{\phi}_i | d_i, \boldsymbol{B}, \boldsymbol{A}, \boldsymbol{V}, \boldsymbol{T})$. Here the hyperparameters $\boldsymbol{\alpha}, \beta$ are fixed after specified manually and effectively constants, and hence we hide them in the distribution notations.

We denote $\{\boldsymbol{Z}_i, \boldsymbol{\phi}_i\}_{i=1}^{M}$ collectively as $\boldsymbol{Z}, \boldsymbol{\phi}$. Then the above objective is to find the optimal $\boldsymbol{B}^*, \boldsymbol{A}^*, \boldsymbol{V}^*, \boldsymbol{T}^*$ and the posterior $p(\boldsymbol{Z}, \boldsymbol{\phi} | \boldsymbol{D}, \boldsymbol{B}^*, \boldsymbol{A}^*, \boldsymbol{V}^*, \boldsymbol{T}^*)$. This posterior is analytically intractable, and we use a simpler variational distribution $q(\boldsymbol{Z}, \boldsymbol{\phi})$ to approximate it.

The coupling between $\boldsymbol{A}, \boldsymbol{V}$ and $\boldsymbol{T}, \boldsymbol{Z}, \boldsymbol{\phi}$ in (8) makes it very difficult to find the optimal $\boldsymbol{A}^*, \boldsymbol{V}^*, \boldsymbol{T}^*$ and the corresponding posterior of $\boldsymbol{Z}, \boldsymbol{\phi}$. To get around this difficulty, we divide the learning into *two stages*.

1. In the first stage, considering that the topics have relatively small impact to word distributions, we simplify the model by disabling topics temporarily, and obtain the optimal solution $\boldsymbol{B}^*, \boldsymbol{A}^*, \boldsymbol{V}^*$ of this reduced model. The optimal solution could be calculated in closed-form;

2. In the second stage, we use $\boldsymbol{B}^*, \boldsymbol{A}^*, \boldsymbol{V}^*$ as an approximate solution, and then enable the topics, and find the corresponding optimal $\boldsymbol{T}^*$, $p(\boldsymbol{Z}, \boldsymbol{\phi} | \boldsymbol{D}, \boldsymbol{B}^*, \boldsymbol{A}^*, \boldsymbol{V}^*, \boldsymbol{T}^*)$ of the full model. In the presence of a lot of hidden variables, a variational EM algorithm is pertinent. During the VEM iterations, we fix $\boldsymbol{B} = \boldsymbol{B}^*, \boldsymbol{A} = \boldsymbol{A}^*, \boldsymbol{V} = \boldsymbol{V}^*$.

6.2 Estimating $\boldsymbol{B}, \boldsymbol{A}, \boldsymbol{V}$ on the Reduced Model with Topics Disabled

As the first step, we disable topics by setting the topic weight β temporarily to 0. In this reduced model, different choices of the topic embeddings $\boldsymbol{T}$, document-topic distributions $\boldsymbol{\phi}$ and topic assignments $\boldsymbol{Z}$ only bring a constant offset to the log-likelihood of the corpus, so they are chosen arbitrarily as $\boldsymbol{T}_0$,$\boldsymbol{\phi}_0, \boldsymbol{Z}_0$.

The matrix $\boldsymbol{B}$ is estimated using the Maximum Likelihood Estimation, and $\boldsymbol{A}, \boldsymbol{V}$ are estimated using the Low Rank Positive Semidefinite Approximation algorithm in Section 5, [1].

6.3 Estimating $\boldsymbol{T}, \boldsymbol{Z}, \boldsymbol{\phi}$ using Variational EM Algorithm on the Full Model

In this stage, we use $\boldsymbol{B}^*, \boldsymbol{A}^*, \boldsymbol{V}^*$ obtained in the previous subsection as their approximate solutions, and then enable the topics by setting β to the prespecified value. Then we proceed to find the corresponding optimal $\boldsymbol{T}^*, p(\boldsymbol{Z}, \boldsymbol{\phi}|\boldsymbol{D}, \boldsymbol{B}^*, \boldsymbol{A}^*, \boldsymbol{V}^*, \boldsymbol{T}^*)$ of this full model. In the presence of a lot of hidden variables, a variational EM algorithm is pertinent. During the VEM iterations, we fix $\boldsymbol{B} = \boldsymbol{B}^*, \boldsymbol{A} = \boldsymbol{A}^*, \boldsymbol{V} = \boldsymbol{V}^*$.

To simplify notation, in the following, we make the hyperparameters $\boldsymbol{\alpha}, \beta$, and the fixed parameters $\boldsymbol{B}^*, \boldsymbol{A}^*, \boldsymbol{V}^*$ implicit in the probabilistic functions. As the topic residuals $\boldsymbol{r} = \{r_{ik}\}_{i,k}$ are uniquely determined by $\boldsymbol{T}$ and β, they are also kept implicit whenever they are irrelevant to the discussion.

We use p to denote the posterior $p(\boldsymbol{Z}, \boldsymbol{\phi}|\boldsymbol{D}, \boldsymbol{T})$ when it is clear from context. Then for an arbitrary variational distribution $q(\boldsymbol{Z}, \boldsymbol{\phi})$, the following equalities hold

$$\begin{aligned} & E_q \log \left[\frac{p(\boldsymbol{D}, \boldsymbol{Z}, \boldsymbol{\phi}|\boldsymbol{T})}{q(\boldsymbol{Z}, \boldsymbol{\phi})} \right] \\ =& E_q \left[\log p(\boldsymbol{D}, \boldsymbol{Z}, \boldsymbol{\phi}|\boldsymbol{T}) \right] + \mathcal{H}(q) \\ =& \log p(\boldsymbol{D}|\boldsymbol{T}) - \mathrm{KL}(q||p), \end{aligned} \tag{9}$$

which implies

$$\mathrm{KL}(q||p) = \log p(\boldsymbol{D}|\boldsymbol{T}) - \Big(E_q \left[\log p(\boldsymbol{D}, \boldsymbol{Z}, \boldsymbol{\phi}|\boldsymbol{T}) \right] + \mathcal{H}(q) \Big). \tag{10}$$

In (10), $E_q \left[\log p(\boldsymbol{D}, \boldsymbol{Z}, \boldsymbol{\phi}|\boldsymbol{T}) \right] + \mathcal{H}(q)$ is usually referred to as the *variational free energy* $\mathcal{L}(q, \boldsymbol{T})$, which is a lower bound of $\log p(\boldsymbol{D}|\boldsymbol{T})$. Directly maximizing $\log p(\boldsymbol{D}|\boldsymbol{T})$ w.r.t. $\boldsymbol{T}$ is intractable due to the hidden variables $\boldsymbol{Z}, \boldsymbol{\phi}$, so we maximize its lower bound $\mathcal{L}(q, \boldsymbol{T})$ instead. We adopt a mean-field approximation of the true posterior as the variational distribution, and use a Variational Expectation Maximization (VEM) algorithm to find $q^*, \boldsymbol{T}^*$ maximizing $\mathcal{L}(q, \boldsymbol{T})$.

6.3.1 Mean-Field Approximation and VEM Algorithm

We assume that the mean-field approximation of the true posterior factorizes as follows:

$$q(\boldsymbol{Z}, \boldsymbol{\phi}; \boldsymbol{\pi}, \boldsymbol{\theta}) = q(\boldsymbol{\phi}; \boldsymbol{\theta}) q(\boldsymbol{Z}; \boldsymbol{\pi}) = \prod_{i=1}^{M} \left\{ \mathrm{Dir}(\boldsymbol{\phi}_i; \boldsymbol{\theta}_i) \prod_{j=1}^{L_i} \mathrm{Cat}(z_{ij}; \boldsymbol{\pi}_{ij}) \right\}.$$

Taking the logarithm of both sides, we obtain

$$\begin{aligned} \log q(\boldsymbol{Z}, \boldsymbol{\phi}; \boldsymbol{\pi}, \boldsymbol{\theta}) = \sum_{i=1}^{M} \Bigg\{ & \log \Gamma(\theta_{i0}) - \sum_{k=1}^{K} \log \Gamma(\theta_{ik}) \\ & + \sum_{k=1}^{K} (\theta_{ik} - 1) \log \phi_{ik} + \sum_{j,k=1}^{L_i, K} \delta(z_{ij} = k) \log \pi_{ij}^k \Bigg\}, \end{aligned} \tag{11}$$

where $\theta_{i0} = \sum_{k=1}^{K} \theta_{ik}$, π_{ij}^k is the k-th component of $\boldsymbol{\pi}_{ij}$.

It follows that

$$\begin{aligned} & \mathcal{H}(q) \\ = & -E_q[\log q(\boldsymbol{Z}, \boldsymbol{\phi}; \boldsymbol{\pi}, \boldsymbol{\theta})] \\ = & \sum_{i=1}^{M} \left\{ \sum_{k=1}^{K} \log \Gamma(\theta_{ik}) - \log \Gamma(\theta_{i0}) - \sum_{k=1}^{K} (\theta_{ik} - 1) \psi(\theta_{ik}) + (\theta_{i0} - K) \psi(\theta_{i0}) - \sum_{j,k=1}^{L_i, K} \pi_{ij}^k \log \pi_{ij}^k \right\}. \end{aligned} \tag{12}$$

Plugging q into $\mathcal{L}(q, \boldsymbol{T})$, we have

$$
\begin{aligned}
&\mathcal{L}(q, \boldsymbol{T}) \\
=&\mathcal{H}(q) + E_q \left[\log p(\boldsymbol{Z}, \boldsymbol{\phi}|\boldsymbol{T})\right] \\
=&\mathcal{H}(q) + C_0 - \log \mathcal{Z}(\boldsymbol{B}^*, \boldsymbol{A}^*, \boldsymbol{V}^*) - ||\boldsymbol{A}||^2_{f(\boldsymbol{H})} \\
&+ \sum_{i=1}^{M} \left\{ \sum_{k=1}^{K} \left(E_{q(\boldsymbol{Z}_i|\boldsymbol{\pi}_i)}[m_{ik}] + \alpha_{0k} - 1 \right) \cdot E_{q(\phi_{ik}|\boldsymbol{\theta}_i)}[\log \phi_{ik}] \right. \\
&\left. + \sum_{j=1}^{L_i} \left(\boldsymbol{v}_{w_{ij}}^\top \Big(\sum_{k=j-c}^{j-1} \boldsymbol{v}_{w_{ik}} + \beta E_{q(z_{ij}|\boldsymbol{\pi}_{ij})}[\boldsymbol{t}_{z_{ij}}] \Big) + \sum_{k=j-c}^{j-1} a_{w_{ik}w_{ij}} + E_{q(z_{ij}|\boldsymbol{\pi}_{ij})}[r_{i,z_{ij}}] \right) \right\} \\
=&C_1 + \mathcal{H}(q) + \sum_{i=1}^{M} \left\{ \sum_{k=1}^{K} \Big(\sum_{j=1}^{L_i} \pi_{ij}^k + \alpha_{0k} - 1 \Big) \Big(\psi(\theta_{ik}) - \psi(\theta_{i0}) \Big) + \sum_{j=1}^{L_i} \Big(\beta \boldsymbol{v}_{w_{ij}}^\top \boldsymbol{T}_i \boldsymbol{\pi}_{ij} + \boldsymbol{r}_i^\top \boldsymbol{\pi}_{ij} \Big) \right\},
\end{aligned}
\tag{13}
$$

where $\boldsymbol{T}_i$ is the topic matrix of the i-th document, and $\boldsymbol{r}_i$ is the vector constructed by concatenating all the topic residuals r_{ik}. $C_1 = C_0 - \log \mathcal{Z}(\boldsymbol{B}^*, \boldsymbol{A}^*, \boldsymbol{V}^*) - ||\boldsymbol{A}||^2_{f(\boldsymbol{H})} + \sum_{i,j=1}^{M,L_i} \Big(\boldsymbol{v}_{w_{ij}}^\top \sum_{k=j-c}^{j-1} \boldsymbol{v}_{w_{ik}} + \sum_{k=j-c}^{j-1} a_{w_{ik}w_{ij}} \Big)$ is constant. $\psi(\cdot)$ is the digamma function.

Then the Variational EM algorithm alternately optimize w.r.t. q and $\boldsymbol{T}, \boldsymbol{r}$ as follows:

1. Initialize all the topics $\boldsymbol{T}_i = \boldsymbol{0}$, and correspondingly their residuals $\boldsymbol{r}_i = \boldsymbol{0}$;

2. Iterate over the following two steps until convergence. In the l-th step:

   (a) Let the topics and residuals be $\boldsymbol{T} = \boldsymbol{T}^{(l-1)}, \boldsymbol{r} = \boldsymbol{r}^{(l-1)}$, find $q^{(l)}(\boldsymbol{Z}, \boldsymbol{\phi})$ that maximizes $\mathcal{L}(q, \boldsymbol{T}^{(l-1)})$. This is the Expectation step (E-step). In this step, $\log p(\boldsymbol{D}|\boldsymbol{T})$ is constant. Then the q that maximizes $\mathcal{L}(q, \boldsymbol{T}^{(l)})$ will minimize $\mathrm{KL}(q||p)$, i.e. such a q is the closest variational distribution to p measured by KL-divergence;

   (b) Given the variational distribution $q^{(l)}(\boldsymbol{Z}, \boldsymbol{\phi})$, find $\boldsymbol{T}^{(l)}, \boldsymbol{r}^{(l)}$ that maximizes $\mathcal{L}(q^{(l)}, \boldsymbol{T})$. This is the Maximization step (M-step). In this step, $\boldsymbol{\pi}, \boldsymbol{\theta}, \mathcal{H}(q)$ are constant;

6.3.2 Update Equations of $\boldsymbol{\pi}, \boldsymbol{\theta}$ in E-Step

In the E-step, $\boldsymbol{T} = \boldsymbol{T}^{(l-1)}, \boldsymbol{r} = \boldsymbol{r}^{(l-1)}$ are constant. For notational simplicity, we drop their superscripts (l) and denote them as $\boldsymbol{T}, \boldsymbol{r}$.

Plugging (12) into (13), we obtain

$$
\begin{aligned}
&\mathcal{L}(q, \boldsymbol{T}^{(l-1)}) \\
=&\sum_{i=1}^{M}\Bigg\{\sum_{k=1}^{K}\log\Gamma(\theta_{ik}) - \log\Gamma(\theta_{i0}) - \sum_{k=1}^{K}(\theta_{ik}-1)\psi(\theta_{ik}) + (\theta_{i0}-K)\psi(\theta_{i0}) - \sum_{j,k=1}^{L_i,K}\pi_{ij}^{k}\log\pi_{ij}^{k} \\
&+\sum_{k=1}^{K}\Big(\sum_{j=1}^{L_i}\pi_{ij}^{k} + \alpha_{0k} - 1\Big)\Big(\psi(\theta_{ik}) - \psi(\theta_{i0})\Big) + \sum_{j=1}^{L_i}\Big(\beta\boldsymbol{v}_{w_{ij}}^{\top}\boldsymbol{T}_i\boldsymbol{\pi}_{ij} + \boldsymbol{r}_i^{\top}\boldsymbol{\pi}_{ij}\Big)\Bigg\} + C_5.
\end{aligned}
\tag{14}
$$

We first maximize (14) w.r.t. π_{ij}^{k}, the probability that the j-th word in the i-th document takes the k-th latent topic. Note that this optimization is subject to the normalization constraint that $\sum_{k=1}^{K}\pi_{ij}^{k} = 1$.

We isolate terms containing $\boldsymbol{\pi}_{ij}$, and form a Lagrange function by incorporating the normalization constraint:

$$
\Lambda(\boldsymbol{\pi}_{ij}) = -\sum_{k=1}^{K}\pi_{ij}^{k}\log\pi_{ij}^{k} + \sum_{k=1}^{K}\Big(\psi(\theta_{ik}) - \psi(\theta_{i0})\Big)\pi_{ij}^{k} + \beta\boldsymbol{v}_{w_{ij}}^{\top}\boldsymbol{T}_i\boldsymbol{\pi}_{ij} + \boldsymbol{r}_i^{\top}\boldsymbol{\pi}_{ij} + \lambda_{ij}(\sum_{k=1}^{K}\pi_{ij}^{k} - 1). \tag{15}
$$

Taking the derivative w.r.t. π_{ij}^{k}, we obtain

$$
\frac{\partial\Lambda(\boldsymbol{\pi}_{ij})}{\partial\pi_{ij}^{k}} = -1 - \log\pi_{ij}^{k} + \psi(\theta_{ik}) - \psi(\theta_{i0}) + \beta\boldsymbol{v}_{w_{ij}}^{\top}\boldsymbol{t}_{ik} + r_{ik} + \lambda_{ij}. \tag{16}
$$

Setting this derivative to 0 yields the maximizing value of π_{ij}^{k}:

$$
\pi_{ij}^{k} \propto \exp\{\psi(\theta_{ik}) + \beta\boldsymbol{v}_{w_{ij}}^{\top}\boldsymbol{t}_{ik} + r_{ik}\}. \tag{17}
$$

Next, we maximize (14) w.r.t. θ_{ik}, the k-th component of the posterior Dirichlet parameter:

$$
\begin{aligned}
&\frac{\partial\mathcal{L}(q, \boldsymbol{T}^{(l-1)})}{\partial\theta_{ik}} \\
=&\frac{\partial}{\partial\theta_{ik}}\Bigg\{\log\Gamma(\theta_{ik}) - \log\Gamma(\theta_{i0}) + \Big(\sum_{j=1}^{L_i}\pi_{ij}^{k} + \alpha_{0k} - \theta_{ik}\Big)\psi(\theta_{ik}) - \Big(L_i + \sum_{k}\alpha_{0k} - \theta_{i0}\Big)\psi(\theta_{i0})\Bigg\} \\
=&\Big(\sum_{j=1}^{L_i}\pi_{ij}^{k} + \alpha_{0k} - \theta_{ik}\Big)\psi'(\theta_{ik}) - \Big(L_i + \sum_{k}\alpha_{0k} - \theta_{i0}\Big)\psi'(\theta_{i0}),
\end{aligned}
\tag{18}
$$

where $\psi'(\cdot)$ is the derivative of the digamma function $\psi(\cdot)$, commonly referred to as the *trigamma function*.

Setting (18) to 0 yields a maximum at

$$\theta_{ik} = \sum_{j=1}^{L_i} \pi_{ij}^k + \alpha_{0k}. \tag{19}$$

Note this solution depends on the values of π_{ij}^k, which in turn depends on θ_{ik} in (17). Then we have to alternate between (17) and (19) until convergence.

6.3.3 Update Equations of $\boldsymbol{T}_i, \boldsymbol{r}_i$ in M-Step

In the M-step, $\boldsymbol{\pi} = \boldsymbol{\pi}^{(l)}, \boldsymbol{\theta} = \boldsymbol{\theta}^{(l)}$ are constant. For notational simplicity, we drop their superscripts (l) and denote them as $\boldsymbol{\pi}, \boldsymbol{\theta}$.

Given these parameter values, (13) is a constant plus the sum of many $\beta \boldsymbol{v}_{w_{ij}}^\top \boldsymbol{T}_i \boldsymbol{\pi}_{ij} + \boldsymbol{r}_i^\top \boldsymbol{\pi}_{ij}$, each of which in turn is a linear transformation of the vector $\beta \boldsymbol{v}_{w_{ij}}^\top \boldsymbol{T}_i + \boldsymbol{r}_i^\top$. The k-th component of this vector is $\log \frac{\exp\{\beta \boldsymbol{v}_s^\top \boldsymbol{t}_{ik}\}}{E_{P(s)}[\exp\{\beta \boldsymbol{v}_s^\top \boldsymbol{t}_{ik}\}]}$, the logarithm of a softmax function of $\boldsymbol{t}_{ik}$. As a softmax function is concave w.r.t. the weight $\boldsymbol{t}_{ik}$, this component is concave, and so is $\beta \boldsymbol{v}_{w_{ij}}^\top \boldsymbol{T}_i + \boldsymbol{r}_i^\top$. Therefore $\mathcal{L}(q^{(l)}, \boldsymbol{T})$ is a concave function of $\boldsymbol{T}$, and its maximum is achieved when its derivative w.r.t. $\boldsymbol{T}$ is 0.

The topic residuals $\boldsymbol{r}_i$ are uniquely determined by $\boldsymbol{T}_i$ and β. Thus we first solve $\boldsymbol{T}_i$, and then $\boldsymbol{r}_i$ is readily determined.

As the first column of $\boldsymbol{T}_i$ is fixed to 0, we only need to find the maximum w.r.t. other columns. We denote the submatrix of all columns of $\boldsymbol{T}_i$ except the first column as $\boldsymbol{T}_{\text{-}1,i}$. To find this maximum, we take the derivative of (13) w.r.t. $\boldsymbol{T}_{\text{-}1,i}$:

$$\begin{aligned}
&\frac{\partial \mathcal{L}(q^{(l)}, \boldsymbol{T})}{\partial \boldsymbol{T}_{\text{-}1,i}} \\
=&\frac{\partial \sum_{j=1}^{L_i} \left(\beta \boldsymbol{v}_{w_{ij}}^\top \boldsymbol{T}_i \boldsymbol{\pi}_{ij} + \boldsymbol{\pi}_{ij}^\top \boldsymbol{r}_i \right)}{\partial \boldsymbol{T}_{\text{-}1,i}} \\
=&\beta \frac{\partial}{\partial \boldsymbol{T}_{\text{-}1,i}} \mathrm{Tr}(\boldsymbol{T}_i \sum_{j=1}^{L_i} \boldsymbol{\pi}_{ij} \boldsymbol{v}_{w_{ij}}^\top) + (\sum_{j=1}^{L_i} \boldsymbol{\pi}_{ij})^\top \frac{\partial \boldsymbol{r}_i}{\partial \boldsymbol{T}_{\text{-}1,i}} \\
=&\beta \sum_{j=1}^{L_i} \boldsymbol{v}_{w_{ij}} \boldsymbol{\pi}_{\text{-}1,ij}^\top + (\sum_{j=1}^{L_i} \boldsymbol{\pi}_{ij})^\top \frac{\partial \boldsymbol{r}_i}{\partial \boldsymbol{T}_{\text{-}1,i}} \\
=&\beta \sum_{j=1}^{L_i} \boldsymbol{v}_{w_{ij}} \boldsymbol{\pi}_{\text{-}1,ij}^\top + \sum_{k=2}^{K} \bar{\pi}_i^k \frac{\partial r_{ik}}{\partial \boldsymbol{T}_{\text{-}1,i}},
\end{aligned} \tag{20}$$

where $\bar{\pi}_i^k = \sum_{j=1}^{L_i} \pi_{ij}^k$, the sum of the variational probabilities of each word being assigned to the k-th topic in the i-th document. $\boldsymbol{\pi}_{\text{-}1,ij}^\top$ is the subvector of all elements of $\boldsymbol{\pi}_{ij}$ except the first: $(\pi_{ij}^2, \cdots, \pi_{ij}^K)^\top$. The index of k in the second term in (20) starts from 2 because r_{i1} is fixed to be 0.

Solving the critical point $\boldsymbol{T}_{\text{-}1,i}$ of (20) requires the computation of $\frac{\partial r_{ik}}{\partial \boldsymbol{T}_i}$. (4) states that $r_{ik} = -\log\Big(E_{P(s)}[\exp\{\beta \boldsymbol{v}_s^\top \boldsymbol{t}_{ik}\}]\Big)$. Then the derivative of r_{ik} w.r.t. $\boldsymbol{T}_i$ is difficult to compute. Alternatively we use a second-order approximation to ease the computation.

As discussed above, $||\beta \boldsymbol{t}_{ik}||$ is small, and thus $||\beta \boldsymbol{v}_s^\top \boldsymbol{t}_{ik}||$ is usually small too (the Gaussian prior over $\boldsymbol{v}_s$ strongly discourage big $||\boldsymbol{v}_s||$). Then a second-order approximation to $\exp\{\beta \boldsymbol{v}_s^\top \boldsymbol{t}_{ik}\}$ is appropriate: $\exp\{\beta \boldsymbol{v}_s^\top \boldsymbol{t}_{ik}\} \approx 1 + \beta \boldsymbol{v}_s^\top \boldsymbol{t}_{ik} + \frac{1}{2}\beta^2 (\boldsymbol{v}_s^\top \boldsymbol{t}_{ik})^2$. It follows that

$$\begin{aligned} & E_{P(s)}[\exp\{\beta \boldsymbol{v}_s^\top \boldsymbol{t}_{ik}\}] \\ \approx & 1 + \beta \boldsymbol{t}_{ik}^\top E_{P(s)}[\boldsymbol{v}_s] + \frac{1}{2}\beta^2 \boldsymbol{t}_{ik}^\top E_{P(s)}[\boldsymbol{v}_s \boldsymbol{v}_s^\top] \boldsymbol{t}_{ik}. \\ = & 1 + \beta \boldsymbol{t}_{ik}^\top \bar{\boldsymbol{v}} + \frac{1}{2}\beta^2 \boldsymbol{t}_{ik}^\top \boldsymbol{X} \boldsymbol{t}_{ik}, \end{aligned} \tag{21}$$

where $\bar{\boldsymbol{v}} = E_{P(s)}[\boldsymbol{v}_s]$ and $\boldsymbol{X} = E_{P(s)}[\boldsymbol{v}_s \boldsymbol{v}_s^\top]$. As $\boldsymbol{V}$ is fixed, $\bar{\boldsymbol{v}}$ and $\boldsymbol{X}$ can be precomputed. The dimensionality of $\boldsymbol{X}$ is $N \times N$, and N is usually chosen as hundreds. Thus $\boldsymbol{X}$ can easily fit into the memory.

It follows that

$$\begin{aligned} \frac{\partial r_{ik}}{\partial \boldsymbol{t}_{ik}} &= -\frac{1}{E_{P(s)}[\exp\{\beta \boldsymbol{v}_s^\top \boldsymbol{t}_{ik}\}]} \frac{\partial}{\partial \boldsymbol{t}_{ik}} E_{P(s)}[\exp\{\beta \boldsymbol{v}_s^\top \boldsymbol{t}_{ik}\}] \\ &\approx -e^{r_{ik}} \cdot \beta(\bar{\boldsymbol{v}} + \beta \boldsymbol{X} \boldsymbol{t}_{ik}). \end{aligned} \tag{22}$$

To summarize, $\dfrac{\partial r_{ik}}{\partial \boldsymbol{t}_{ij}}$ are divided into two cases:

$$\begin{cases} \dfrac{\partial r_{ik}}{\partial \boldsymbol{t}_{ik}} \approx -e^{r_{ik}} \cdot \beta(\bar{\boldsymbol{v}} + \beta \boldsymbol{X} \boldsymbol{t}_{ik}), & k \neq 1 \\ \dfrac{\partial r_{ik}}{\partial \boldsymbol{t}_{ij}} = 0, & k = 1 \text{ or } j \neq k. \end{cases} \tag{23}$$

Plugging (23) into (20), we obtain

$$\frac{\partial \mathcal{L}(q^{(l)}, \boldsymbol{T})}{\partial \boldsymbol{T}_{\text{-}1,i}} \approx \beta \sum_{j=1}^{L_i} \boldsymbol{v}_{w_{ij}} \boldsymbol{\pi}_{\text{-}1,ij}^\top - \beta(\bar{\boldsymbol{V}} + \beta \boldsymbol{X} \boldsymbol{T}_{\text{-}1,i})\Pi_i, \tag{24}$$

where $\bar{\boldsymbol{V}} = (\bar{\boldsymbol{v}} \cdots \bar{\boldsymbol{v}})_{N\times(K-1)}$, whose first column is 0 and other columns are all $\bar{\boldsymbol{v}}$, and $\Pi_i = \begin{pmatrix} \bar{\pi}_i^2 e^{r_{i2}} & & 0 \\ & \ddots & \\ 0 & & \bar{\pi}_i^K e^{r_{iK}} \end{pmatrix} = \mathrm{diag}(\bar{\boldsymbol{\pi}}_{\text{-}1,i})\mathrm{diag}(\exp\{\boldsymbol{r}_{\text{-}1,i}\})$. Here $\boldsymbol{r}_{\text{-}1,i}$ is the subvector of all elements of $\boldsymbol{r}_i$ except the first.

Setting the RHS of (24) to 0 leads to an equation whose solution is near $\arg\max_{\boldsymbol{T}_{\text{-}1,i}} \mathcal{L}(q^{(l)}, \boldsymbol{T})$:

$$(\bar{\boldsymbol{V}} + \beta \boldsymbol{X}\boldsymbol{T}_{\text{-}1,i})\Pi_i = \sum_{j=1}^{L_i} \boldsymbol{v}_{w_{ij}} \boldsymbol{\pi}_{\text{-}1,ij}^{\top}. \tag{25}$$

However, (25) cannot be solved directly, because the terms $e^{r_{ik}}$ in Π_i are complicated functions of $\boldsymbol{t}_{ik}$. To circumvent this complexity, we adopt an iterative algorithm. In the m-th iteration, $\boldsymbol{r}_{\text{-}1,i}$ take the values $\boldsymbol{r}_{\text{-}1,i}^{(m-1)}$ found in the (m-1)-th iteration (if $m = 1$, then $\boldsymbol{r}_{\text{-}1,i}$ take the values computed in the last E-step), yielding a solution

$$\boldsymbol{T}_{\text{-}1,i}^{(m)} = \frac{1}{\beta}\boldsymbol{X}^{-1}\left\{ \left(\sum_{j=1}^{L_i} \boldsymbol{v}_{w_{ij}} \boldsymbol{\pi}_{\text{-}1,ij}^{\top} \right) \mathrm{diag}(\bar{\boldsymbol{\pi}}_{\text{-}1,i})^{-1} \mathrm{diag}(\exp\{-\boldsymbol{r}_{\text{-}1,i}^{(m-1)}\}) - \bar{\boldsymbol{V}} \right\}. \tag{26}$$

In the next iteration, $\boldsymbol{r}_{\text{-}1,i}^{(m)}$ is computed using (4). This iterative process continues until convergence.

References

[1] Anonymous. A generative word embedding model and its low rank positive semidefinite solution. Submitted to EMNLP'2015.