AN EXTENSION OF STANDARD LATENT DIRICHLET ALLOCATION TO MULTIPLE CORPORA

ADAM FOSTER[1], HANGJIAN LI[2], GEORG MAIERHOFER[3], AND MEGAN SHEARER[4]

ACADEMIC MENTOR: RUSSEL E. CAFLISCH[5]

This work is dedicated to our academic mentor, Russel E. Caflisch, and to our families and friends who made this all possible through their constant support.

Abstract. Latent Dirichlet Allocation (LDA) is a highly successful topic modeling framework. We describe a new extension to LDA which supports multiple subcorpora, each containing a different type of document. As in LDA, this multiple-corpora LDA (mLDA) model assumes document topic proportions follow a symmetric Dirichlet distribution. However, in mLDA, the Dirichlet parameter is subcorpus dependent. An online algorithm for training mLDA models is derived. The algorithm is applied to data from the USC Shoah Foundation's Visual History Archive. Results show mLDA produced a better language model than standard LDA for this data. Using the same data, the mLDA topic model is used to construct an information retrieval system. Search results from this system outperform those obtained from traditional string-based search systems. A novel approach to the visualization of topics is outlined and visualizations are presented. As a novel development in natural language processing, mLDA will allow the power of topic modeling to be applied to a huge range of fields with diverse data by incorporating more information into a single topic model. It also enhances the applicability of topic modeling to information retrieval.

Key words. Latent Dirichlet Allocation, generative models, natural language processing, information retrieval

1. Introduction. Topic models are generative models in natural language processing which posit the existence of latent topics to explain an observed corpus of documents. Latent Dirichlet Allocation (LDA) is a topic model in which topics and topic proportions are assumed to follow Dirichlet distributions [8]. It has become the standard topic modeling framework [4]. Nevertheless, many authors have found it beneficial to modify or extend the basic LDA model. For example, changing the Dirichlet distribution of the topic proportions within documents to a log-normal allows modelers to uncover correlations between topics [6]. Authorship information can be included in another extension of LDA due to Rosen-Zvi et al. [16]. In hierarchical LDA [9], the topics are arranged in a tree structure. Paths through the tree are random samples from a nested Chinese restaurant process. A given document contains a mixture of the topics on one particular path through the tree. Such models may be better suited to modeling syntactic features of language. The related Hierarchical Dirichlet Process circumvents the need to specify in advance the number of topics in the model [18].

Variational inference can be applied to the estimation of parameters in the LDA model [8]. An important feature of this method is the treatment of hyperparameters: these may be fixed constants, or updated using a Newton-Raphson method [8]. A more modern variant of the original variational algorithm allows training documents to be streamed, meaning extremely large corpora can be used to train topic models [11].

In common with other generative models, LDA may be applied to information retrieval [13]. One approach is to use the Hellinger distance between topic distributions as a measure of similarity between a document and a query [7].

This paper outlines a new extension of LDA, multiple-corpora LDA (mLDA), in which the corpus is comprised of several subcorpora of different document types. The topic distributions for a subcorpus follow a symmetric Dirichlet distribution, but with a subcorpus-dependent parameter. Topics are common to the entire corpus. This allows the modeler to build a topic model using multiple document collections, incorporating the distinct nature of each collection into the model.

Frequently, information concerning a domain of interest does not reside in a single homogeneous corpus. Instead, a model which uses all of the available information must consider multiple

[1]University of Cambridge, Queens' College, Cambridge CB3 9ET, U.K. (aef39@cam.ac.uk)
[2]University of California, Los Angeles, 520 Portola Plaza, CA 90095 (lihangjian123@ucla.edu)
[3]University of Cambridge, Trinity College, Cambridge CB2 1TQ, U.K. (gam37@cam.ac.uk)
[4]University of Arizona, 617 N. Santa Rita Ave. Tucson, AZ 85721 (megshearer@email.arizona.edu)
[5]Department of Mathematics, University of California, Los Angeles, 460 Portola Plaza 1158A, CA 90095-7121 (rcaflisch@ipam.ucla.edu)

subcorpora. For example, a model for the language of science writing could incorporate *Science* articles, scientific textbooks and Wikipedia articles on scientific subjects. Unlike LDA, mLDA allows the construction of a topic model from all of these sources of information without the unrealistic assumption that the topic distributions are identically distributed across subcorpora.

This work also offers greater flexibility in the application of topic modeling to information retrieval. Whilst a query may be assumed to contain the same latent topics which are present in the corpus to be searched, it is not necessarily the case that its topic distribution follows the same distribution as the documents in the corpus. Without abandoning the generative hypothesis, that documents and queries arise from the same generative model, mLDA may more accurately reflect the different natures of queries and documents.

This paper lays out the theory of mLDA. Results obtained from a dataset from the USC Shoah Foundation are presented. A model is trained using two distinct subcorpora and is compared with a model trained by treating the entire corpus as homogeneous. The model is applied to information retrieval. The topic proportions of search queries and potential search results are inferred by assuming that these collections form two further subcorpora which were not used to train the model.

The paper is organized as follows. Section 2 provides detailed mathematical background including a description of the LDA model. Section 3 describes mLDA and the adaptation of the variational algorithm of [11] required to train parameters under the new model. In Section 4, theoretical evaluation of mLDA as a language model is conducted, whilst Section 5 details the implementation of an information retrieval system based on mLDA and evaluation of mLDA from this perspective. Visualizations of mLDA topics are presented and discussed. Concluding remarks are contained in Section 6 and Section 7 is reserved for acknowledgements.

2. Background.

2.1. Standard LDA. Latent Dirichlet Allocation (LDA) is a topic modelling technique that was first described by Blei, Ng and Jordan in 2003 [8]. It is a hidden random variable model for natural language processing. The goal of LDA is to automatically identify topics within a corpus of documents. In the LDA model topics are considered to be probability distributions over the finite vocabulary. We denote by $\mathbb{V}$, of cardinality V, the vocabulary of the corpus and $\mathbf{W}_1, \ldots, \mathbf{W}_D \in \mathbb{V}^N$ the documents in the corpus (each assumed to contain N words). We denote by $W_{d,n}$ (for $1 \leq d \leq D, 1 \leq n \leq N$) the n^{th} entry in $\boldsymbol{W}_d$, i.e. the n^{th} word in the d^{th} document of the corpus. Furthermore, we denote the $(n-1)$-simplex by $\boldsymbol{\Delta}^{n-1}$ (an n-vector θ lies in the $(n-1)$-simplex if $\theta_i \geq 0$, $\sum_{i=1}^{n} \theta_i = 1$), and let $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K \in \boldsymbol{\Delta}^{V-1}$ be the topics (which are distributions over words), $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_D \in \boldsymbol{\Delta}^{K-1}$ the topic proportions (which are distributions over topics), and $Z_{d,n} \in \{1, \ldots, K\}$ for $1 \leq d \leq D, 1 \leq n \leq N$, the topic of word n in document d.

Standard LDA assumes that producing a document is a random process described by the following generative model:

1. Choose topics $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K \sim \text{Dir}(\eta)$, where $\eta \in \mathbb{R}^+$ is a parameter.
2. For $d = 1, \ldots, D$: Choose the topic distribution of document d as $\boldsymbol{\theta}_d \sim Dir(\alpha)$, where $\alpha \in \mathbb{R}^+$ is a parameter that does not depend on d.
   (a) Choose the topic of the n^{th} word, $Z_{d,n} \sim \text{Multinomial}(\boldsymbol{\theta}_d)$.
   (b) Choose the n^{th} word, $W_{d,n} \sim \text{Multinomial}(\boldsymbol{\beta}_{Z_{d,n}})$.

The probabilistic graphical model of this process is shown in Figure 2.1 below, where the plates symbolize the "level" at which the random variable (r.v.) is chosen (K is topic level, D is corpus level and N is document level). The blank and filled dots represent latent and observed r.v.s respectively.

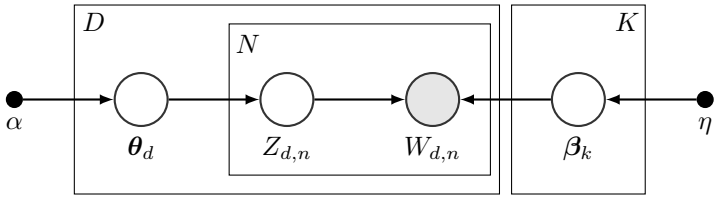


Fig. 2.1: Generative model of LDA

The goal of LDA is to gain information about the topics $\boldsymbol{\beta}_k$ as well as the assignment of topics to each document $\boldsymbol{\theta}_d$ within this model. The standard way to retrieve this information is by assuming given values for the hyperparameters α, η and then making estimates of the random variables under these assumptions.[6] Note that, intuitively speaking, these hyperparameters change the shape of the Dirichlet distributions, which reflect how many topics we expect to be in each document, α, and how many highly relevant words we expect in each topic, η (cf. Section 2.2). We then want to find good estimates for these random variables and the most suitable method for this is variational Bayesian inference (VB), as described in [8, p. 1003-1005].

Optimization algorithms for VB in the case of standard LDA return vectors $\hat{\boldsymbol{\beta}}_k$, and $\hat{\boldsymbol{\theta}}_d$ which are its best estimates for topic k and for the topic distribution of document d respectively, cf. [12]. This means that $\hat{\boldsymbol{\beta}}_k$ tells us which words the topic k is likely to produce and $\hat{\boldsymbol{\theta}}_d$ tells us which topics are likely to be present in document d.

2.2. Dirichlet parameters. The Dirichlet distribution is incremental in the model of LDA. It has unique properties, one of which will be of special importance when constructing our mLDA model in Section 3.2.

DEFINITION 2.1 (Dirichlet distribution). *We say a random variable $\boldsymbol{X}$ follows a Dirichlet distribution, $\boldsymbol{X} \sim Dir(\boldsymbol{\alpha})$, for $\boldsymbol{\alpha} \in \mathbb{R}^K_{>0}$, if its probability density function is non-zero when $\sum_{i=1}^K x_i = 1$ and given by*

$$p(\boldsymbol{x}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{i=1}^{K} x_i^{\alpha_i - 1},$$

where $B(\boldsymbol{\alpha}) = \frac{\prod_{i=1}^K \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^K \alpha_i)}$ is the beta function.

Note that this means $\boldsymbol{X}$ takes values in the $(K-1)$-simplex $\boldsymbol{\Delta}^{K-1}$. The distribution is called symmetric if $\boldsymbol{\alpha}$ is of the form $\boldsymbol{\alpha} = (\alpha, \ldots, \alpha) \in \mathbb{R}^N_{>0}$. In Figures 2.2, 2.3 and 2.4 below we can see several samples of a symmetric Dirichlet distributions with fixed dimension $K = 12$ and varying Dirichlet parameter α.

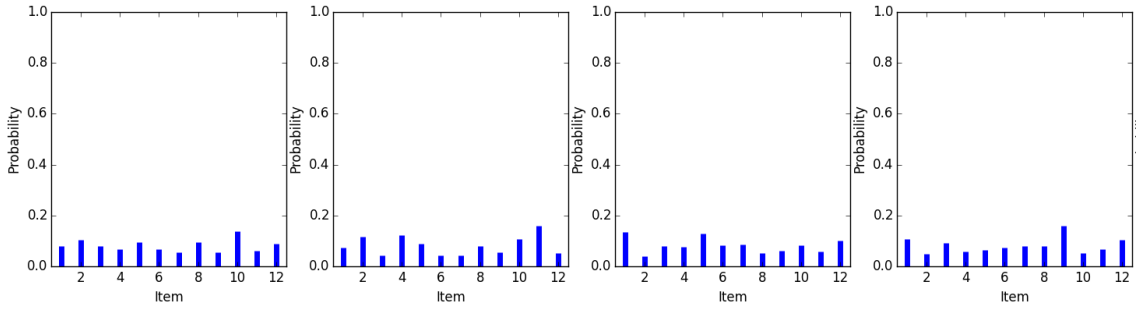


Fig. 2.2: Symmetric Dirichlet Distribution with $K = 12, \alpha = 10$

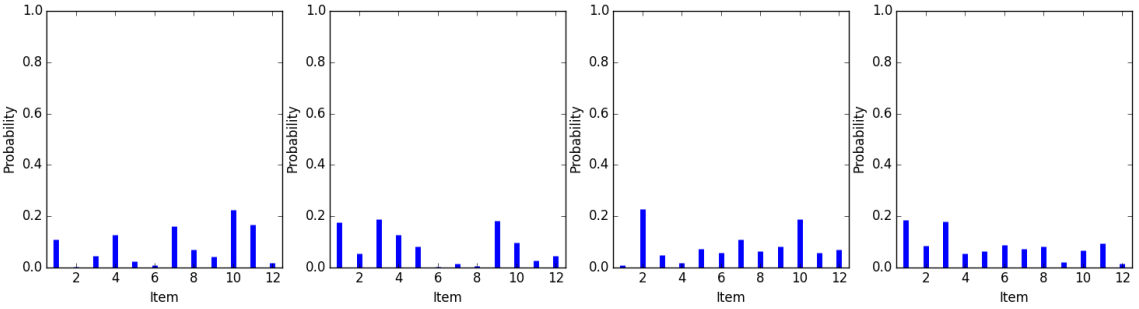


Fig. 2.3: Symmetric Dirichlet Distribution with $K = 12, \alpha = 1$

[6] An alternative method would be to find maximum likelihood estimators for the hyperparameters α, η and then to proceed with statistical inference, cf. [8, p. 1005-1006].

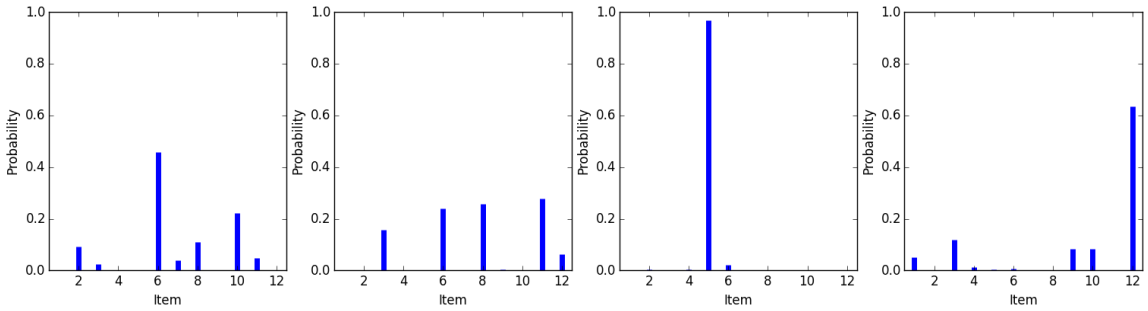


Fig. 2.4: Symmetric Dirichlet Distribution with $K = 12, \alpha = 0.1$

These graphs reflect an important property of the parameter α: it influences a feature of the Dirichlet distribution often referred to as "peakiness", cf. [4]. The samples for large α, such as $\alpha = 10$ (shown in Figure 2.2) are quite uniform. For smaller values of α as in Figure 2.3 and Figure 2.4 we observe more sparsity – only a few items have non-negligible probability.

3. Multiple-corpora LDA.

3.1. Motivation. The standard LDA algorithm is good for many different applications, however there are cases when the corpus of documents can be separated into various subcorpora of different document types. In this section, we first introduce a mixture LDA model to help demonstrate a different approach for analyzing corpora containing different types of subcorpra, and then propose the novel multiple-corpora LDA model which generalizes this approach. To begin with consider the following two examples:

EXAMPLE 3.1. *Suppose we wish to model language arising in scientific literature. We may find popular science books and research papers to be useful training data. Topics such as cosmology, anthropology, geophysics and many others may arise in a popular science book (3.1a). Dedicated research papers (3.1b & 3.1c), on the other hand, are likely to be about only one of these areas. Table 3.1 gives an example. To construct a realistic model, we wish to enforce that popular science books can contain a mixture of topics, but research papers contain only one topic. The topics are common across the entire corpus.*

civilization	hominid
sun	*Homo sapiens*
Big Bang	rock
galaxy	volcano
tectonic	Neanderthal

(a) A popular science book

sun
supernova
nebulae
galaxy
atom

(b) Paper on cosmology

human
culture
ethology
prehistoric
kinship

(c) Paper on anthropology

Table 3.1: Two distinct classes of documents featuring common topics (cosmology, anthropology and geophysics)

EXAMPLE 3.2. *Suppose we need to analyze an archive of documents, which comes with a categorisation of terms in the vocabulary, which are topically related. The documents in the archive arise as before in the standard LDA way as mixtures of topics, however each category - grouping words of similar topic - arises from just one topic.* A model for data of these features in above examples is given as follows, where the documents in the second subcorpus are given as $\boldsymbol{W}_d^{(2)}$ (which again are viewed as N_2-sets of words from the vocabulary):

1. For each document in the second subcorpus ($d = 1, \ldots, D_2$) choose its topic $T_d \sim$ Multinomial($\boldsymbol{\phi}$), where $\boldsymbol{\phi} \in \boldsymbol{\Delta}^{K-1}$ is a category independent parameter.
2. For each place in the document ($n = 1, \ldots, N_2$) choose the actual word $W_{d,n}^{(2)} \sim$ Multinomial($\boldsymbol{\beta}_{T_d}$).

where the topics $\boldsymbol{\beta}_k$ are chosen as in Section 2.1.

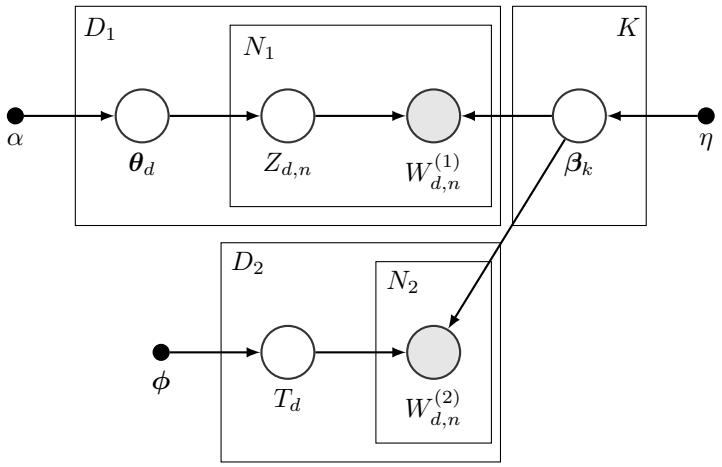


Fig. 3.1: Mixture LDA model

This process is visualized in Figure 3.1 and we call this the **mixture** LDA model. It is reasonable to assume that $\boldsymbol{\phi} = \frac{1}{K}$, i.e. that the T_d have symmetric distribution. In this case above model turns out to be limit case of a more general model which is described in the next section.

3.2. Generative model of multiple-corpora LDA. The previous section motivates the development of a more general model featuring multiple subcorpora. Unlike in mixture LDA, we do not need to enforce the hard constraint that documents in one subcorpus contain *exactly one* topic. Table 3.2 gives an illustration of the new idea (in comparison to Table 3.1):

civilization	hominid
sun	*Homo sapiens*
Big Bang	rock
galaxy	volcano
tectonic	Neanderthal

(a) A popular science book

Homo sapiens
ethnic
sociocultural
holistic
continental drift

(b) Paper on anthropology, mentioning geophysics

Table 3.2: Two distinct classes of documents featuring common topics (cosmology, anthropology and geophysics)

Let us now recall from Section 2.2 the hyperparameter α in the standard LDA model specifies the properties of samples from the Dirichlet distribution. In particular it – roughly speaking – specifies how many topics we expect to find in documents coming from the according model – smaller α means we expect fewer topics. This motivates the following model - which we call **multiple-corpora LDA (mLDA)**. Given a corpus of $L \in \mathbb{N}$ subcorpora with distinct features, we describe the generative process of creating this corpus as follows:

1. Choose topics $\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_K \sim \mathrm{Dir}(\eta)$, where $\eta \in \mathbb{R}^+$, and the distribution is over $\boldsymbol{\Delta}^{V-1}$.
2. For Subcorpus l, where $l = 1, \dots, L$: Choose the topic distribution of document d as $\boldsymbol{\theta}_d^{(l)} \sim \mathrm{Dir}(\alpha_l)$, where $\alpha_l \in \mathbb{R}^+$ is a parameter that does not depend on d, $d = 1, \dots, D_l$.
   (a) Choose the topic of the n^{th} word, $Z_{d,n}^{(l)} \sim \mathrm{Multinomial}(\boldsymbol{\theta}_d^{(l)})$, where $n = 1, \dots, N_l$.
   (b) Choose the n^{th} word, $W_{d,n}^{(l)} \sim \mathrm{Multinomial}(\boldsymbol{\beta}_{Z_{d,n}^{(l)}})$.

Splitting the data into multiple corpora in this way allows us to specify the hyperparameters $\alpha_1, \dots, \alpha_L$ individually (before estimating the random variables), in order to incorporate our prior knowledge or belief of the topic proportions in each subcorpus. This model is visualised in Figure 3.2 below.

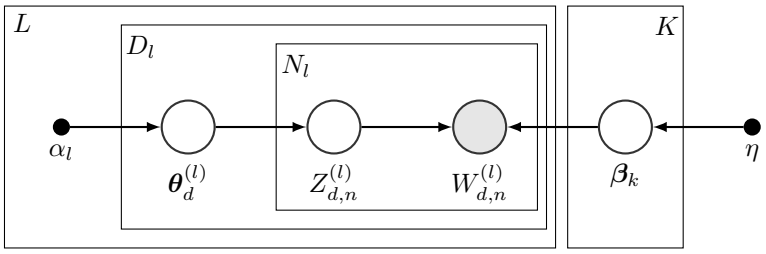


Fig. 3.2: mLDA model

As mentioned before in the case when $L = 2, \alpha_1 = \alpha, \alpha_2 \to 0$, the above model becomes the mixture LDA model for $\boldsymbol{\phi} = \frac{1}{K}$, as described in Section 3.1. A proof of this statement is given in Appendix A.

3.3. Variational Inference in mLDA. In order to make use of this new model we need to develop an algorithm that allows us to estimate the random variables $\boldsymbol{\theta}_d^{(l)}, \boldsymbol{\beta}_k$ in mLDA. For this we use Variational Bayes inference as used by Blei et al. [8, p. 1003-1005]. Here the true posterior is approximated by a variational distribution with free parameters $\boldsymbol{\phi}^{(l)}, \boldsymbol{\gamma}^{(l)}, \boldsymbol{\lambda}$, where $l = 1, \dots, L$. We then optimize those to maximize the Evidence Lower Bound (ELBO, cf. [12, p. 3]):[7]

$$\begin{aligned} \log p(\boldsymbol{w}|\boldsymbol{\alpha}, \eta) &\geq \mathcal{L}(\boldsymbol{w}, \boldsymbol{\phi}, \boldsymbol{\gamma}, \boldsymbol{\lambda}) \\ &\equiv \mathbb{E}_q[\log p(\boldsymbol{w}, \boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\beta}|\boldsymbol{\alpha}, \eta)] - \mathbb{E}_q[\log q(\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\beta})] = -\mathrm{KL}(q\|p) \end{aligned} \tag{3.1}$$

Here $p(\boldsymbol{w}, \boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\beta}|\boldsymbol{\alpha}, \eta)$ is the true posterior distribution of the model[8], and $q(\boldsymbol{w}, \boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\beta})$ is an arbitrary variational distribution with parameters $\boldsymbol{\phi}, \boldsymbol{\gamma}, \boldsymbol{\lambda}$. $\mathrm{KL}(q\|p)$ is the Kullback-Leibler divergence - a measure for the distance of two continuous probability distributions [3, p. 55]. Hence maximizing the ELBO is equivalent to minimizing the Kullback-Leibler divergence between the variational distribution and the true posterior.

Following [8, p. 1007] we choose a separable distribution for q, in the form (for $l = 1, \dots, L$):

$$q(z_{dn}^{(l)} = k) = \phi^{(l)}_{dw_{d,n}^{(l)}k}; \quad q(\theta_d^{(l)}) = \mathrm{Dirichlet}(\theta_d^{(l)}; \gamma_d^{(l)}); \quad q(\beta_k) = \mathrm{Dirichlet}(\beta_k; \lambda_k) \tag{3.2}$$

where $\phi_{dn}^{(l)} \in \boldsymbol{\Delta}^{K-1}, \gamma_d^{(l)} \in \mathbb{R}^K_{>0}, \lambda_k \in \mathbb{R}^V_{>0}$ and $d = 1, \dots, D_l, n = 1, \dots N_l, k = 1, \dots, K$.
This means the posterior of the word-topic assignments is parametrised by $\boldsymbol{\phi}$, the posterior of the topic proportions is parametrised by $\boldsymbol{\gamma}$ and the posterior of the topics is parametrised by $\boldsymbol{\lambda}$. Moreover conditioned upon those parameters the random variables are independent w.r.t. q. With this choice of variational distribution the ELBO can be written as:

$$\begin{aligned} \mathcal{L}(\boldsymbol{w}, \boldsymbol{\phi}, \boldsymbol{\gamma}, \boldsymbol{\lambda}) = \sum_{l=1}^{L} \sum_{d=1}^{D_l} \{ &\mathbb{E}_q[\log p(w_d^{(l)}|\theta_d^{(l)}, z_d^{(l)}, \boldsymbol{\beta})] + \mathbb{E}_q[\log p(z_d^{(l)}|\theta_d^{(l)})] - \mathbb{E}_q[\log q(z_d^{(l)})] \\ &+ \mathbb{E}_q[\log p(\theta_d^{(l)}|\alpha_l)] - \mathbb{E}_q[\log q(\theta_d^{(l)})] + (\mathbb{E}_q[\log p(\boldsymbol{\beta}|\eta)] - \mathbb{E}_q[\log q(\boldsymbol{\beta})]) / (\textstyle\sum_{l=1}^{L} D_l) \}. \end{aligned}$$

Here the dependence on the variational parameters is given implicitly and we have taken the last term into the summation (dividing by the number of summations). We can now expand above expectations in terms of the variational parameters to find - writing $n_{dv}^{(l)}$ for the number of times

[7] Note that if we use boldface for variables in more than one dimension, we refer to the whole object, whereas if we use lightface and neglect indices, we refer to the multi-dimensional object consisting of all elements of indices that have been neglected.

[8] By neglecting one of the arguments of p we mean the marginal p.d.f. of the remaining arguments, e.g. $p(\boldsymbol{w}|\boldsymbol{\alpha}, \eta)$ is the marginal p.d.f. of $\boldsymbol{w}$.

word $v \in \mathbb{V}$ appears in document d of subcorpus l:

$$\begin{aligned}\mathcal{L} = \sum_l \sum_d \Bigg[& \sum_v n_{dv}^{(l)} \sum_k \phi_{dvk}^{(l)} (\mathbb{E}_q[\log \theta_{dk}^{(l)}] + \mathbb{E}_q[\log \beta_{kv}] - \log \phi_{dvk}^{(l)}) \\ & - \log \Gamma \left(\sum_k \gamma_{dk}^{(l)} \right) + \sum_k (\alpha_l - \gamma_{dk}^{(l)}) \mathbb{E}[\log \theta_{dk}^{(l)}] + \log \Gamma(\gamma_{dk}^{(l)}) \\ & + \left(\sum_k - \log \Gamma(\sum_v \lambda_{kv}) + \sum_v (\eta - \lambda_{kv}) \mathbb{E}[\log \beta_{kv}] + \log \Gamma(\lambda_{kv}) \right) / (\textstyle\sum_{l=1}^{L} D_l) \\ & + \log \Gamma(K\alpha_l) - K \log \Gamma(\alpha_l) + (\log \Gamma(V\eta) - V \log \Gamma(\eta)) / (\textstyle\sum_{l=1}^{L} D_l) \Bigg] \\ =: & \sum_l \sum_d l(n_d^{(l)}, \phi_d^{(l)}, \gamma_d^{(l)}, \boldsymbol{\lambda}).\end{aligned}$$

Where V is, as before, the number of words in the vocabulary. Here $l(n_d^{(l)}, \phi_d^{(l)}, \gamma_d^{(l)}, \boldsymbol{\lambda})$ denotes the contribution of document d in subcorpus l to $\mathcal{L}$. We can hence, analogously to [8, p. 1004] optimize $\mathcal{L}$ using coordinate ascent over the variational parameters $\boldsymbol{\phi}, \boldsymbol{\gamma}, \boldsymbol{\lambda}$ in the following way:

$$\begin{aligned}\phi_{dvk}^{(l)} &\propto \exp\{\mathbb{E}_q[\log \theta_{dk}^{(l)}] + \mathbb{E}_q[\log \beta_{kv}]\} \\ \gamma_{dk}^{(l)} &= \alpha_l + \sum_v n_{dv}^{(l)} \phi_{dvk}^{(l)} \\ \lambda_{kv} &= \eta + \sum_l \sum_d n_{dv}^{(l)} \phi_{dvk}^{(l)}\end{aligned}$$

where

$$\begin{aligned}\mathbb{E}_q[\log \theta_{dk}^{(l)}] &= \Psi(\gamma_{dk}^{(l)}) - \Psi(\textstyle\sum_{i=1}^{K} \gamma_{di}^{(l)}) \\ \mathbb{E}_q[\log \beta_{kv}] &= \Psi(\lambda_{kv}) - \Psi(\textstyle\sum_{i=1}^{V} \lambda_{ki})\end{aligned}$$

and Ψ denotes the digamma function - the first derivative of the gamma function.

We can (similarly to the onlineLDA algorithm in [12, p. 5]) use an algorithm of the following form to numerically find optimizing solutions for this problem: We can partition our algorithm into an "E" step - iteratively update γ and ϕ until convergence, holding $\boldsymbol{\lambda}$ fixed - and an "M" step - update $\boldsymbol{\lambda}$ given ϕ, by analogy with an Expectation-Maximization algorithm. As in [12, p. 5] we propose an "online" algorithm for mLDA to optimize these parameters. This means that the algorithm is given information for each document in each subcorpus just once, then updates the document specific parameters until convergence (E step) and afterwards updates $\boldsymbol{\lambda}$ as a weighted average of the new value and the value obtained from the documents previously seen (M step).

This, our proposed, algorithm for the optimisation step of mLDA is described in Algorithm 1 below. Here the parameter κ specifies the rate at which information from previous documents is forgotten by the algorithm. Note that we need to choose $\kappa \in (0.5, 1]$ to ensure convergence of the optimisation algorithm to a local maximum (cf. [12, p. 4]).

Once we have optimised with respect to the variational parameters we can calculate our estimates for the original random variables as expectations of the variational distribution:

$$\begin{aligned}\hat{\theta}_d^{(l)} &= \mathbb{E}_q[\theta_d^{(l)} | \gamma_d^{(l)}] = \frac{\gamma_d^{(l)}}{\|\gamma_d^{(l)}\|_1} \\ \hat{\beta}_k^{(l)} &= \mathbb{E}_q[\beta_k^{(l)} | \lambda_k^{(l)}] = \frac{\lambda_k^{(l)}}{\|\lambda_k^{(l)}\|_1},\end{aligned}$$

since the variational distributions of both are Dirichlet (3.2).[9]

[9] Note that all of $\hat{\theta}_d^{(l)}, \gamma_d^{(l)}, \hat{\beta}_k^{(l)}, \lambda_k^{(l)}$ represent vectors in the appropriate dimensions (for notational clarity the bold font has been neglected).

Algorithm 1: Online mLDA

Define $\rho(d) \equiv (\tau_0 + d)^{-\kappa}$
Initialize $\boldsymbol{\lambda}$ randomly
for $l = 1 : L$ **do**
  for $d = 1 : D_l$ **do**
    E step:
    Initialize $\gamma_{dk}^{(l)} = 1$
    while $\frac{1}{K}\sum_k |change\ in\ \gamma_{dk}^{(l)}| > \epsilon$ **do**
      Set $\phi_{dvk}^{(l)} \propto \exp\{\mathbb{E}_q[\log \theta_{dk}^{(l)}] + \mathbb{E}_q[\log \beta_{kv}]\}$
      Set $\gamma_{dk}^{(l)} = \alpha_l + \sum_v n_{dv}^{(l)} \phi_{dvk}^{(l)}$
    end
    M step:
    Compute $\tilde{\lambda}_{kv} = \eta + D n_{tv}^{(l)} \phi_{tvk}^{(l)}$
    Set $\boldsymbol{\lambda} = (1 - \rho(d + \sum_{i=1}^{l-1} D_i))\boldsymbol{\lambda} + \rho(d + \sum_{i=1}^{l-1} D_i)\tilde{\boldsymbol{\lambda}}$
  end
end

3.4. Implementation. The algorithm outlined in Section 3.3 was implemented in Python as an extension of the popular package `gensim`. The code is publicly available [14].

4. Theoretical evaluation. To determine the validity and usefulness of mLDA, we used data from the USC Shoah Foundation (USC SF) to train an mLDA model. This data demonstrates a similar structure to the one described in Example 3.2. We go on to estimate the perplexity of the model (with various parameter choices). In Section 5, we discuss practical applications of our work.

4.1. The USC SF data set. The USC SF has collected over 52,000 video testimonies of survivors and witnesses of the Holocaust and other genocides. These testimonies are accessible to the public through the USC SF's Visual History Archive (VHA). A testimony is an interview with a survivor, and testimonies are broken down into one minute segments. Segments can also be tagged with keywords; we use these keywords as our data set (rather than working directly with video).

In order to manage the thousands of available keywords that can be used to tag video segments, the keywords are organized into a *keyword hierarchy*. This is a tree-like structure. A keyword has one parent (occasionally several), which may itself have a parent, and so on. For example, "visas" has parent "documents and artifacts" which has parent "objects". The keyword hierarchy is the existing method used to classify keywords and make them accessible to archivists.

Accurate models trained from the USC SF data should use information from both sources: testimonies and hierarchy.

The precise collections of documents used were:

1. **Testimony documents** consisting of multiple one minute video segments. After careful consideration, it was decided to break a testimony into documents by beginning a new document whenever a time or place related keyword was encountered. This was a compromise between excessively long documents (testimonies) and excessively short ones (one minute segments).
2. **Hierarchy documents** consisting of keywords with the same *direct* parent in the keyword hierarchy. Notice that keywords on different levels fall into different documents.

In the language of Section 3.2, the vocabulary consists of keywords. There are $L = 2$ subcorpora, with subcorpus 1 consisting of testimony documents and subcorpus 2 of hierarchy documents. This data is suitable for the testing of mLDA as it exactly exhibits two different subcorpora of distinct document types, which is what the mLDA model has been developed for.

(Note that in this data set not every document does necessarily contain the same number of words. However the previous theory generalises straightforwardly to this case and our implementation is written in sufficient generality to accommodate this feature.)

4.2. Evaluation of mLDA language model. The evaluation of probabilistic models often involves calculation, or estimation, of the likelihood of some held-out test sample. The closely related **perplexity** is often used in language models [12, p. 7].

DEFINITION 4.1 (Perplexity). *Let n_i^{test} be the vector of word-counts for the i^{th} document in the test set, and α_i the value of the Dirichlet parameter in the subcorpus of the i^{th} document in the test set, then the perplexity is given as:*

$$\textit{perplexity}(\boldsymbol{w}^{test}, \boldsymbol{\lambda}, \boldsymbol{\alpha}) \equiv \exp\left[-\frac{(\sum_i \log p(w_i^{test}|\alpha_i, \boldsymbol{\beta}))}{\left(\sum_{i,v} n_{iv}^{test}\right)}\right].$$

Intuitively speaking, a smaller perplexity corresponds to a test sample that is more probable under the given model. Models with smaller perplexity fit the data better.

It is computationally expensive to evaluate this measure directly, hence we use the lower bound on the log-likelihood (3.1) to give an upper bound for the perplexity:

$$\text{perplexity}(\boldsymbol{w}^{test}, \boldsymbol{\lambda}, \boldsymbol{\alpha}) \leq \exp\left[-\frac{(\sum_i \mathbb{E}_q[\log p(w_i^{test}, \theta_i, z_i|\alpha_i, \boldsymbol{\beta})] - \mathbb{E}_q[\log q(\theta_i, z_i)])}{\left(\sum_{i,v} n_{iv}^{test}\right)}\right] \tag{4.1}$$

Here the topics $\boldsymbol{\lambda}$ are found on the training corpus and then held fixed to infer γ_i, ϕ_i for the test corpus using the E-Step (and appropriate α_i) in Algorithm 1. In our implementation we choose the learning parameter $\kappa = 0.5$.

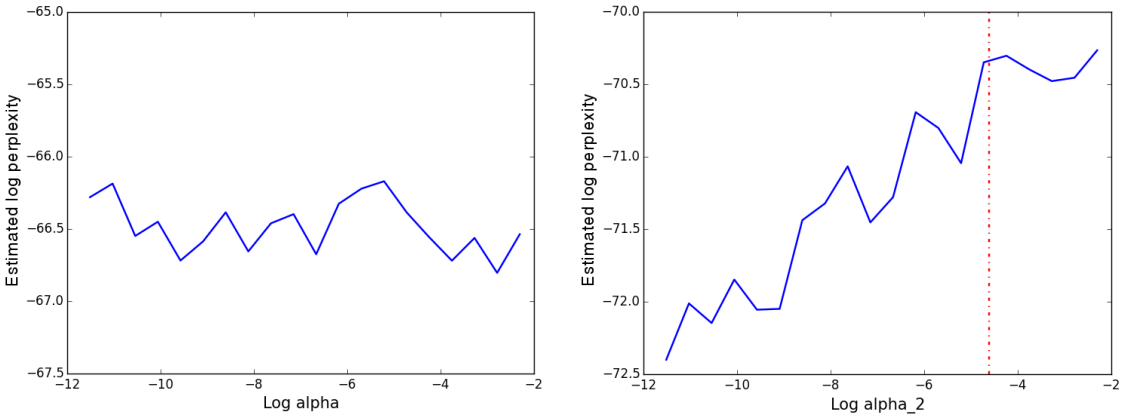
The quantity on the right hand side of (4.1) is referred to as the **variational Bayesian bound** on the perplexity.

In the large data limit, as the number of training samples tends to infinity, the variational Bayesian bound on the negative log-likelihood evaluated on the training data converges to the Bayesian information criterion (BIC) for the model [1, p. 23]. An intuitive explanation can be found in [2, p. 75-76]. The BIC is widely used as a criterion for model selection, and arises naturally from Bayesian considerations [17].

Under the assumption that the training and test data arise from the same distribution, the large data limit of the variational Bayesian bound on the negative log-likelihood of the *test data* also converges to the BIC. Therefore the variational Bayesian bound on the log perplexity converges to the BIC. The variational Bayesian bound is an estimator for a well-known model selection criterion and is the criterion which will be used in this paper, because the true likelihood and BIC are intractable. We refer to the variational Bayesian bound on the perplexity as the **estimated perplexity**.

4.3. Perplexity estimation for the USC SF data set. Figure 4.1 below allows us to compare the performance of mLDA and LDA with our data from the USC SF archive. In both graphs we apply the methods to the whole data set (total number of documents $> 60,000$), holding out a total of 400 documents in each case for testing. We evaluate the model using $K = 100$ topics.

In Figure 4.1a we can see the estimated perplexity of LDA which is trained on all hierarchy documents and all but 400 testimony documents (multiple segments), which are used as a test corpus. The documents from both classes are in this case treated equally and the perplexity is estimated for various values of the Dirichlet parameter α. We can see that there is no clear dependence of the estimated perplexity on the value of α. In Figure 4.1b we see the estimated perplexity of mLDA, which is trained on all but 200 hierarchy documents and all but 200 testimony documents, these 400 are used as test corpus. The value of the Dirichlet parameter for the testimony documents is held fixed, $\alpha_1 = 0.01$, and the Dirichlet parameter for the hierarchy documents is allowed to vary, $10^{-5} \leq \alpha_2 \leq 0.1$. The red line marks where the mLDA model is equivalent to the regular LDA model. We can see a clear dependence of the estimated perplexity on the value of α_2 and as we had expected earlier, the model fits better if α_2 takes smaller values. This reflects the intuitive idea that each hierarchy document is built from a single topic (not a mixture as a testimony segment might be) and confirms that mLDA performs better on this example.



(a) LDA: 400 testimony doc. (b) mLDA: 200 testimony, 200 hierarchy doc.

Fig. 4.1: Variational Bayesian bounds on log perplexity of LDA and mLDA for various α values

Furthermore the variational Bayesian bound on the perplexity allows us to evaluate how well the algorithm finds optimizing solutions (i.e. estimates for the latent r.v.s) for the given problem. In Figure 4.2 we see the estimated perplexity of the mLDA model as a function of the training corpus size. Both training and test corpus consist to 4% from hierarchy and 96% from testimony documents. This reflects the real life sizes of the corpora. Moreover we hold the Dirichlet parameters fixed at $\alpha_1 = 0.01$, $\alpha_2 = 0.0001$ and run the model for $K = 100$ topics. We can observe that indeed the estimated perplexity decreases for larger training corpora, which verifies our model and choice of parameters.

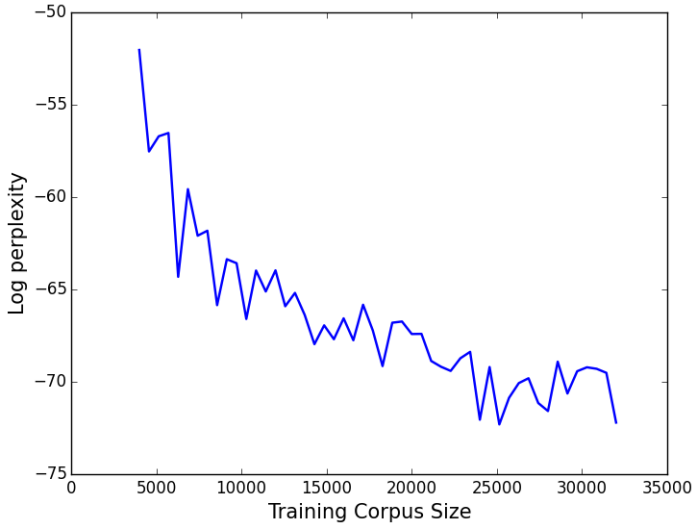


Fig. 4.2: Held-out perplexity as a function of training corpus size

5. Applications.

5.1. Application to search. Generative models such as mLDA are typically stepping stones towards information retrieval systems. The mLDA model was applied to information retrieval on the USC SF data set. Given a search query consisting of keywords, the set of video segments should be ordered by relevance.

Figure 5.1 gives an overview of the method used to construct a search system from mLDA. First, an mLDA model is trained. Given a topic model, the search system computes and stores the topic distributions of every searchable document (these are typically distinct from the training documents). When a user enters a query, the topic distribution of the query is also calculated.

The Hellinger distance between a document's topic distribution and the query's is used to order the searchable corpus. This distance measure is standard in this context [7].

DEFINITION 5.1 (Hellinger distance). *Let $\boldsymbol{\theta}, \boldsymbol{\psi} \in \boldsymbol{\Delta}^{K-1}$ be topic proportions. The Hellinger distance between $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ is given by:*

$$\textit{distance}(\boldsymbol{\theta}, \boldsymbol{\psi}) = \sum_{j=1}^{K} \left(\sqrt{\theta_j} - \sqrt{\psi_j} \right)^2 .$$

In contrast to the LDA model, the mLDA model treats the following subcorpora as distinct:

- L training subcorpora,
- the searchable subcorpus, and
- queries.

For the USC SF data, there were two training subcorpora: the testimony and hierarchy corpora with hyperparameters α_1 and α_2 respectively. It has already been noted in Section 4 that mLDA generates a more accurate model than LDA. The new model also offers greater flexibility in information retrieval. For example, for the USC SF data, it is acceptable to assume that search queries typically refer to only one topic. By using $\alpha_{\text{query}} \ll 1$, this observation may be incorporated into the search system. For the USC SF data set, the searchable subcorpus was different from both training corpora described in Section 4.1, as it consisted of individual one minute video segments.

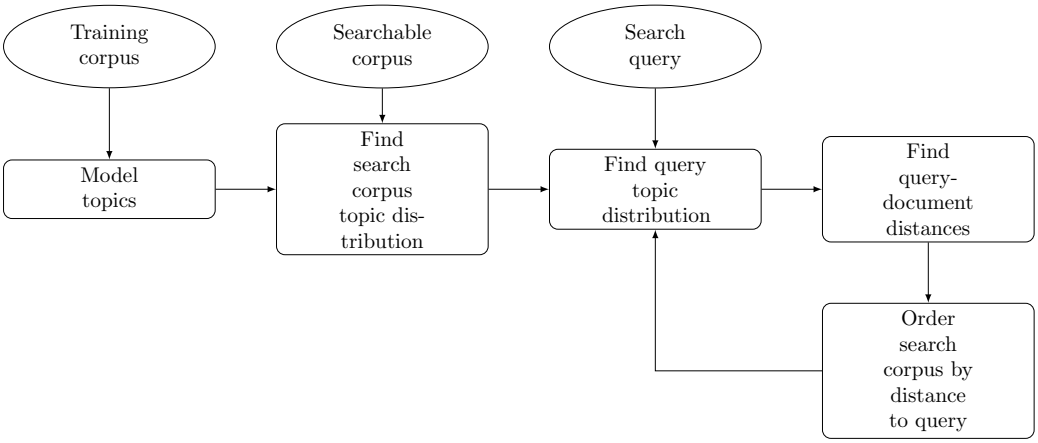


Fig. 5.1: A high level description of the search algorithm

Search systems using both mLDA and LDA were constructed and compared to a simple direct keyword matching algorithm. Results are displayed in Table 5.1. It can be seen that, in response to the example query "war criminals", LDA and mLDA models behave very differently to keyword level algorithms. This is because string level algorithms require the presence of the literal term "war criminals". Topic models return results which are related, but not by exact language. Note that the choice of parameters for LDA/mLDA means that exact string matches will rank *worse* than more broadly related results. This choice was made to emphasise the novelty and possible advantages of topic based search systems.

Whilst LDA and mLDA bear many resemblances in search performance, it can be seen that mLDA results are more well focused to the search query. In every case, the name of at least one war criminals is returned in the segment when mLDA was used. The standard LDA search system returns a noticeably irrelevant result "trucks, Soviet resistance fighters, ...". The additional flexibility given by mLDA is considered the main reason for this improved performance.

Direct keyword matching
war criminals, war crimes trials, ...
war criminals, Holocaust history
war criminals
Standard LDA
Dachau Trial, post-liberation trial reflections, ...
war crimes investigations, interviewee occupations, ...
trial participants, Ilse Koch, ...
trucks, Soviet resistance fighters, Pustkow, ...
war crimes trial-related psychological reactions, war crimes investigations, Dachau, ...
mLDA
Albert Speer, trial defendants, Nuremberg, ...
Walther Funk, trial procedures, Hjalmar Schacht, ...
Julius Streicher, Wilhelm Keitel, Franz von Papen, ...
Rudolf Hess, Holocaust history, Otto Ohlendorf, ...
Flick Trial, Doctors Trial, IG Farben Trial, trial verdicts, ...

Table 5.1: The keywords of the highest ranked segments of three search systems, in response to the query "war criminals". mLDA used $\alpha_1 = 1, \alpha_2 = 0.05$, the search corpus used $\alpha = 0.5$ and search queries $\alpha = 0.01$. LDA used the same setting with the absence of hierarchy data.

5.2. Visualizing topics. Based on the mLDA model, a dynamic, graph-based visualization was built to interact with the keywords from the USC SF data. This visualization is intended to help users navigate VHA's keywords and mLDA topics, and find connections between testimonies and subjects of interest.

The visualization displays a topic as in Figure 5.2. Each purple node represents a keyword from

Fig. 5.2: Visualization of a single topic

the USC SF data, indicated by its text label. The six keywords that have the highest probability of occurring within this particular topic i are displayed. One keyword is chosen as the center node. In Figure 5.2 this is "camp forced labor". (The centre node for topic i is chosen to be $\text{argmax}_{v \in V}(\boldsymbol{\beta}_{i,v} \cdot m_v)$ where $\boldsymbol{\beta}_i$ is topic i, a distribution over the vocabulary as described in Section 2, and m_v is the total frequency of keyword v in the corpus.)

Initially, the visualization displays only the center nodes, as shown in Figure 5.3. The visualization in this example contains 10 topics modeled using mLDA. Different colors represent different topics. This form of visualization demonstrates connections between keywords because keywords in the same topic relate closely to one another. It can also help to identify common themes present in the Archive. For instance, if a topic or set of topics contain practices by leaders, events, or other actions that occurred before both the Holocaust and the Armenian Genocide, this can provide insight into the mechanics of such atrocities.

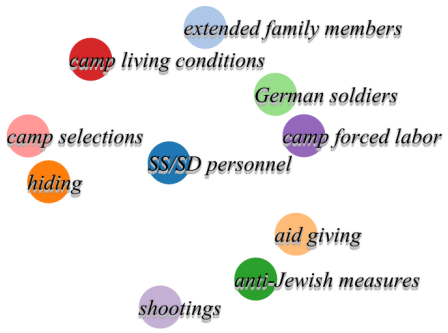


Fig. 5.3: Keywords initially displayed on visualization

6. Conclusion and future work. The multiple-corpora LDA model is an extension of the LDA model which allows the Dirichlet parameters α to be set independently for each subcorpus within a corpus of documents. The freedom to choose different α values enables the model to account for the variation of topic distributions among subcorpora. Following [12, p. 7], we compared the perplexities of a standard LDA model trained from the USC SF data with an mLDA model trained from the same data. The training data naturally consisted of two distinct subcorpora, and mLDA gave significantly better results when appropriate Dirichlet parameters were chosen for each subcorpus. Because of the nature of the second subcorpus, we manually chose a much smaller α value than for the first subcorpus, reflecting our intuition that documents in the second subcorpus contained fewer topics. The mLDA model yielded considerably better results than standard LDA when α_2 was chosen over 100 times smaller than α_1. Based on the mLDA model, we developed a search engine and a visualization of the USC SF keyword structure, exemplifying two possible applications of mLDA. The search engine treated search queries as a third subcorpus with α_{query} distinct from the α values used to train the model.

In our experiments, the mLDA model was tested on a data set containing two subcorpora. To examine its robustness on more complex data, we need a mechanism to automatically set α values for each subcorpus. Future work will focus on modeling the Dirichlet parameter α as a random variable for each subcorpus so as to reduce the problem of selecting parameter values. In general, mLDA presents a new direction in topic modeling which illustrates the potential of incorporating multiple sources of information into a single topic model. It makes topic models more flexible in their treatment of training data and data used purely for inference (such as search queries). mLDA will unlock topic modeling as a language processing tool for those applying machine learning to diverse data.

7. Acknowledgements. This work is based upon the RIPS 2015 project conducted by the authors at IPAM, UCLA (California), supervised by Prof. Russel E. Caflisch, and jointly supported by USC Shoah Foundation and NSF Grant DMS-0931852. We would like to thank Dr. Michael R. Raugh, for the enthusiasm with which he directed the RIPS program and for his experienced advice throughout our project. We would also like to thank the USC Shoah Foundation for providing access to their data and for the helpful advice we received from our industrial mentors Sam Gustman and Mills Shih-Chung Chang. Furthermore we would like to express our gratitude to Krishna Bhogaonker for his guidance and to our colleagues at RIPS 2015 together with the staff of IPAM for creating a highly stimulating research environment during this program.

REFERENCES

[1] Hagai Attias, *Inferring parameters and structure of latent variable models by variational bayes*, in Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence, Morgan Kaufmann Publishers Inc., 1999, pp. 21–30.

[2] Matthew J Beal, *Variational algorithms for approximate Bayesian inference*, University of London, 2003.

[3] Christopher M Bishop, *Pattern recognition and machine learning*, Springer, 2006.

[4] David M. Blei, *Topic models*, 2009. ⟨`http://videolectures.net/mlss09uk_blei_tm`⟩. Accessed August 4, 2015. Lecture at the Machine Learning Summer School, Cambridge.

[5] ———, *Probabilistic topic models*, Commun. ACM, 55 (2012), pp. 77–84.

[6] David M Blei and John D Lafferty, *Correlated topic models*, Advances in neural information processing systems, 18 (2006), p. 147.

[7] ———, *Topic models*, Text mining: classification, clustering, and applications, 10 (2009), p. 34.

[8] David M. Blei, Andrew Y. Ng, and Michael I. Jordan, *Latent dirichlet allocation*, The Journal of Machine Learning Research, 3 (2003), pp. 993–1022.

[9] Griffiths Thomas L Jordan Michael I Blei, David M and Joshua B Tenenbaum, *Hierarchical topic models and the nested chinese restaurant process*, Advances in neural information processing systems, 16 (2004), p. 17.

[10] Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L Boyd-Graber, and David M Blei, *Reading tea leaves: How humans interpret topic models*, in Advances in neural information processing systems, 2009, pp. 288–296.

[11] Matthew Hoffman, Francis R Bach, and David M Blei, *Online learning for latent dirichlet allocation*, in advances in neural information processing systems, 2010, pp. 856–864.

[12] ———, *Online learning for latent dirichlet allocation*, in advances in neural information processing systems, 2010, pp. 856–864.

[13] Victor Lavrenko, *A generative theory of relevance*, vol. 26, Springer Science & Business Media, 2008.

[14] Radim Rehurek and Adam Foster, *Gensim*, 2015. (`https://github.com/octavius-1993/gensim`). Public repository, accessed September 9, 2015.

[15] Willi Richert, *Building Machine Learning Systems with Python*, Packt Publishing Ltd, 2013.

[16] Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth, *The author-topic model for authors and documents*, in Proceedings of the 20th conference on Uncertainty in artificial intelligence, AUAI Press, 2004, pp. 487–494.

[17] Gideon Schwarz et al., *Estimating the dimension of a model*, The annals of statistics, 6 (1978), pp. 461–464.

[18] Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei, *Hierarchical dirichlet processes*, Journal of the american statistical association, (2012).

Appendix A. Mixture LDA as a limit of mLDA. We want to show that in the limit $\alpha_2 \to 0$ the mLDA model is equivalent to the mixture LDA model with symmetric $\boldsymbol{\phi} = \frac{1}{K}$, where the models are as described in Section 3. This is equivalent to showing that the limiting distribution of the actual words in the second subcorpus $W_{d,n}^{(2)}$ in mLDA, as $\alpha_2 \to 0$, is the same as in mixture LDA with symmetric $\boldsymbol{\phi}$. As individual documents are independent in our model it is sufficient to show this for just one document (here $W_1, \ldots, W_N$ correspond to the words of a document in subcorpus 2 in mLDA, with control distribution $W'_1, \ldots, W'_N$ corresponding to words of a document in subcorpus 2 of mixture LDA). The following proposition provides precisely this result:[10]

Propostion A.1. *Let $K, N, V \in \mathbb{N}$, $\alpha \in \mathbb{R}^+$ and $\boldsymbol{\beta}^{(1)}, \ldots, \boldsymbol{\beta}^{(K)} \in \boldsymbol{\Delta}^{V-1}$. Suppose $T \sim Multinomial(1/K)$ is a r.v. and $W'_n \sim Multinomial(\boldsymbol{\beta}^{(T)})$, for $1 \leq n \leq N$, are i.i.d. r.v.s. Moreover suppose that $\boldsymbol{\theta} \sim Dir(\alpha)$, with $\boldsymbol{\theta} \in \boldsymbol{\Delta}^{K-1}$ and $Z_n \sim Multinomial(\boldsymbol{\theta})$, for $1 \leq n \leq N$, are i.i.d. r.v.s and that $W_n \sim Multinomial(\boldsymbol{\beta}^{(Z_n)})$, for $n = 1, \ldots, N$. Then*

$$\boldsymbol{W} \xrightarrow{\mathcal{D}} \boldsymbol{W'} \text{ as } \alpha \to 0,$$

where $\boldsymbol{W} := (W_1, \ldots, W_N)$ and $\boldsymbol{W'} := (W'_1, \ldots, W'_N)$.

[10] Note that we have adopted a notation that is slightly more suitable for this purpose.

Proof. Let us firstly construct the joint p.d.f.s of $\boldsymbol{W}$ and $\boldsymbol{W'}$ respectively. For $\boldsymbol{W'}$ we have:

$$\begin{aligned}\mathbb{P}(\boldsymbol{W'}=\boldsymbol{w}) &= \sum_{j=1}^{K}\mathbb{P}(\boldsymbol{W'}=\boldsymbol{w}|T=j)\mathbb{P}(T=j)\\ &= \sum_{j=1}^{K}\mathbb{P}(\boldsymbol{W'}=\boldsymbol{w}|T=j)\frac{1}{K}\\ \mathbb{P}(\boldsymbol{W'}=\boldsymbol{w}) &= \sum_{j=1}^{K}\frac{1}{K}\prod_{i=1}^{N}\beta_{w_i}^{(j)}. \qquad \text{(A.1)}\end{aligned}$$

And for $\boldsymbol{W}$ we have:

$$\begin{aligned}\mathbb{P}(\boldsymbol{W}=\boldsymbol{w}) &= \sum_{\boldsymbol{z}}\mathbb{P}(\boldsymbol{W}=\boldsymbol{w}|\boldsymbol{Z}=\boldsymbol{z})\mathbb{P}(\boldsymbol{Z}=\boldsymbol{z})\\ &= \sum_{\boldsymbol{z}}\mathbb{P}(\boldsymbol{W}=\boldsymbol{w}|\boldsymbol{Z}=\boldsymbol{z})\int_{\boldsymbol{\Delta}^{K-1}}\mathbb{P}(\boldsymbol{Z}=\boldsymbol{z}|\boldsymbol{\theta})p(\boldsymbol{\theta})\,d\boldsymbol{\theta}\\ &= \sum_{\boldsymbol{z}}\prod_{j=1}^{N}\beta_{w_j}^{(z_j)}\int_{\boldsymbol{\Delta}^{K-1}}p(\boldsymbol{\theta})\prod_{k=1}^{N}\theta_{z_k}\,d\boldsymbol{\theta} \qquad \text{(A.2)}\end{aligned}$$

where $\boldsymbol{Z}=(Z_1,\dots,Z_N)$ and p is the p.d.f. of $\boldsymbol{\theta}$. Now, letting $N_i=\sum_{k=1}^{N}\mathbb{1}_{\{i\}}(z_k)$ for $i=1,\dots,K$ (i.e. the number of occurrences of index i)[11] we find:

$$\int_{\boldsymbol{\Delta}^{K-1}}p(\boldsymbol{\theta})\prod_{k=1}^{N}\theta_{z_k}\,d\boldsymbol{\theta}=\mathbb{E}\left[\prod_{j=1}^{K}\theta_j^{N_j}\right]=\frac{\Gamma(K\alpha)}{\Gamma(K\alpha+N)}\prod_{j=1}^{K}\frac{\Gamma(\alpha+N_j)}{\Gamma(\alpha)}$$

where the second equality reflects the well-known formula for moments of the Dirichlet distribution. Using the "recursive" property of the Gamma function we can write this as:

$$\begin{aligned}\int_{\boldsymbol{\Delta}^{K-1}}p(\boldsymbol{\theta})\prod_{k=1}^{N}\theta_{z_k}\,d\boldsymbol{\theta} &= \frac{\Gamma(K\alpha)}{(K\alpha+N-1)\cdots(K\alpha)\Gamma(K\alpha)}\prod_{N_j\neq 0}\frac{(\alpha+N_j-1)\cdots(\alpha)\Gamma(\alpha)}{\Gamma(\alpha)}\\ &= \frac{\prod_{N_j\neq 0}(\alpha+N_j-1)\cdots(\alpha)}{(K\alpha+N-1)\cdots(K\alpha)}\end{aligned}$$

Since $\sum_{i=1}^{K}N_i=N$, there is at least one $N_i\neq 0$. From this it is easy to see that as $\alpha\to 0$

$$\int_{\boldsymbol{\Delta}^{K-1}}p(\boldsymbol{\theta})\prod_{k=1}^{N}\theta_{z_k}\,d\boldsymbol{\theta}\to\begin{cases}\frac{1}{K}\text{ if }\forall j\neq i,\ N_j=0\\ 0\text{ o.w.}\end{cases}$$

Hence by (A.2) we immediately can conclude:

$$\mathbb{P}(\boldsymbol{W}=\boldsymbol{w})\to\sum_{j=1}^{K}\frac{1}{K}\prod_{i=1}^{N}\beta_{w_i}^{(j)}\overset{\text{(A.1)}}{=}\mathbb{P}(\boldsymbol{W'}=\boldsymbol{w})\text{ as }\alpha\to 0$$

so

$$\boldsymbol{W}\xrightarrow{\mathcal{D}}\boldsymbol{W'}\text{ as }\alpha\to 0.$$

□

[11] Note that consequently $\sum_{i=1}^{K}N_i=N$.