Roi Pillar, 308292135

**A summarization of the original LDA paper by Blei et al. 2003**

The original paper describes Latent Dirichlet Allocation (LDA). LDA is a topic model, which is a statistical model for identifying the different "topics" that occur in a given document (LDA is also applied in fields other than text modeling – this will be discussed). LDA is a probabilistic model, which in this context means a document is modeled as a probability distribution over all topics (as opposed to a one-hot representation), and it is a generative model – this will be discussed further on.

The main goals of topic modeling in text corpora are:

1. Finding structure in large collections of text that can't be realistically categorized manually (by humans).
2. Extracting compact representations of documents that can be used by downstream applications (dimensionality reduction).
3. Being able to predict the topics of previously unseen documents.

TDF-IF offers some assistance in achieving these goals. Namely, it reduces documents of arbitrary length to a fixed size list (of size |V|), while capturing the notion of words that are highly representative of the ducoment they are featured in (this is done by taking into account not only the amount of times a given words has been seen in a document, but also the number of times it has been seen across the entire corpus). However, TDF-IF does not provide information about the statistical structure of documents – and this is where topic models come in.

The first technique for topic modeling discussed in the paper is Latent Semantic Indexing (LSI). While employing the distributional hypothesis, LSI uses SVD to reduce the number of rows in the word-count matrix of a document. The major difference between the next 2 techniques that will be presented is, that in LSI topics are orthogonal to each other (like words in the BoW scheme are orthogonal to each other), and there is no notion of distance" between 2 topics, like there is in the following 2 techniques.

The second modeling technique is pLSI (Probabilistic Latent Semantic Indexing), which evolved from LSI, satisfying the need to minimize the information loss stemming from the dimensionality reduction, by defining topics as probability distributions over words in the vocabulary. This technique is the "main competitor" or predecessor (at least according to the paper) of LDA in the field of topic modeling. In pLSI, topics are distributions over words, and documents are distributions over topics (in all 3 techniques, there is a fixed amount of topics). There are 2 major practical problems with pLSI, the paper states:

1. The model does not contain a mechanism for assigning a probability to an unseen document – i.e. all documents in the world have an equal probability of existing - they are uniformly distributed.
2. The number of parameters of the model grows linearly with the number of documents – which creates problems of overfitting on the training data.

LDA seeks to solve both of these problems, using the Dirichlet distribution.

One big difference between LDA and the other 2 models, is that LDA is a generative model. (Note: pLSI *is* generative of the documents in a given corpus, but it is not generative on new documents). This means, that it assumes the entire corpus has been generated by some process (see below). As with other generative models we discussed in class, it uses the Bayesian method of utilizing the "prior" probability – the probability of seeing some observation as input. This also means that both inputs (=documents) and outputs (=topic distribution) can be generated by the model using some hidden parameters, as opposed to discriminative models which can only infer outputs given inputs (generative models can do this as well as we will see).

Before we get to the generative story, let's present a brief explanation of the Dirichlet distribution.

Informally, the Dirichlet distribution is a distribution over distributions, i.e. it assigns probabilities to distributions. In this context, the Dirichlet is used to obtain a sparse prior probability for both the topic distribution (Dir(α)), and for the word distribution in topics (Dir(β)). The importance of the prior here is that documents are assumed to belong to a small number of topics with a high probability (rather than belong to a large number of topics with a low probability), and every topic is assumed to be composed of a small number of discriminative words. When αi<1 (α is a vector), the model tends to generalize better than pLSI (which is just LDA with α=1). Herein lies the difference between pLSI and LDA – pLSI assumes documents have been generated from topic distributions that have been seen in training data, whereas LDA does not. It instead assumes the documents are generated from arbitrary topic distributions obtained by this Dirichlet prior.

(source: <http://dirichlet.net/pdf/wallach09rethinking.pdf>)

LDA's generative story then for each document di in a corpus D consisting of M documents of length Ni each, with K total latent topics is as follows:

1. Choose {\displaystyle \theta \_{i}\sim \operatorname {Dir} (\alpha )} θi∼Dir(α), where i is an index into the list of documents, and Dir(α) is the Dirichlet distribution with the hyperparameter α.
2. For each word position in the document:
   1. Choose a topic ϕk∼Multinomial(θi) , where k is an index into the list of topics, and ϕk is selected from a multinomial distribution that was sampled in step 1.
   2. Choose a word wj from p(wj | ϕk,β).

Legend:

α - A vector of size K whose elements are all strictly positive. This is a corpus-level parameter, it is sampled once per corpus. It is the parameter of the Dirichlet prior for topic distributions in documents.

β - A matrix of size kxV, in which the rows represent the topics, and the columns represent the vocabulary, and β(i,j) = p(wj = 1| ϕi = 1), meaning the i,j cell of the matrix is the probability of the word j being sampled from topic i.

θi – The topic distribution for document i sampled from the Dirichlet

ϕk – The topic (distribution over words) sampled from θi.

Wj – The word (BoW representation) sampled from ϕk and the β prior.

We can see that the LDA model's parameters do not grow linearly with the number of documents in the training data – the model's parameters are α which is of size k and β which is of size kxV, both of which are learned on the training data (as opposed to the kM+kV parameters of the pLSI model).

The paper describes the main issue with LDA – which is computing the posterior – that is p(θ, z|w,α,β) – the distributions of topics across documents and words across topics, given the learned parameters α, β and the training corpus. This posterior cannot be computed tractably – and so the paper describes a variational inference algorithm to overcome this.

In order to compute the best alpha and beta values, an EM algorithm is presented, which finds α,β s.t ∑log p(Wj |α,β) is maximized on the training data.

In order for LDA to not assign probability 0 to documents with unseen words (=words that are not in the training vocabulary), it has to be used along with some smoothing technique to distribute the probability mass of documents in a more equal manner. The paper states the best results are achieved through Dirichlet smoothing (Laplace smoothing with a Dirichlet prior).

The empirical results shown in the paper demonstrate LDA's superiority over the other methods (pLSI, a unigram model and a mixture of unigrams) in Perplexity as a function of number of topics (when number of topics is <200), which is indicative of the fact that LDA generalizes better due to its use of the Dirichlet prior.