Topic Modeling and Jensen-Shannon Divergence Based Clustering on the Gutenberg Dataset using Latent Dirichlet Allocation

Eren Aldis
Department of Computer Science
Johns Hopkins University
Baltimore, MD 21218
ealdis1@jhu.edu

Abstract

This paper investigates a Latent Dirichlet Allocation (LDA) model to find topic-level, lower-dimensional representations of texts in the Gutenberg dataset. We consider a Variational Inference and a Gibbs Sampling based approach for the inference and parameter learning tasks for the model parameters. Also, different settings for the tuning parameters and hyperparameters are compared. Then, using the topic-level representations, we propose a Jensen-Shannon divergence and Mahalanobis distance-based recommendation engine which, given a sample book from the corpus, recommends other books that are similar in terms of their topic representations. We compare the results of the Gibbs Sampling and Variational Inference LDA models by their recommendations given a sample book.

1 Introduction

The immense growth of high-dimensional data over the past decade has brought an increased attention to dimensionality reduction methods and latent structure models. The analogue of this in the field of text and natural language processing has been topic modeling, aimed to find a latent description of a given document through associating the document to a set of topics. The availability of a lower-dimensional representation then can enable one to categorize and recommend documents based on similarity. One can easily imagine this being used in a bookstore or a library. Thus, we use sample texts from the Gutenberg dataset, described in section 1.1, to build such a recommendation system. For this end, we consider a model proposed by Blei, Ng, and Jordan [1] known as Latent Dirichlet Allocation (LDA). Modeling the topics as latent variables, LDA assumes each piece of text (document) has a distribution over K topics, and each topic itself has a distribution over the whole English vocabulary. This yields a probabilistic graphical model where each word in the document are considered to be observations, assumed to be generated from a mixture of these distributions. The details of the model will be further discussed in section 2.1.

In Blei, Ng, and Jordan the method described uses a Variational Bayes algorithm for the inference of the model parameters. In this paper, we will consider an MCMC approach for the inference using Gibbs sampling as well as the original Variational Bayes algorithm for the estimation of the posterior distributions of the parameters of per document topic distributions (θ) and per topic distributions over the words (φ). The details of the estimation procedure will be discussed in section 2.2.

There has been a number of studies describing ways to cluster documents together by categorizing them by their most probable topics. In this paper we consider a different approach by using the per-document topic distributions as a whole to build a clustering and a recommendation engine. Once the per-document topic distributions are available, each document can be represented as a unique

32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.

coordinate in the $K-1$ simplex where K is the number of topics. Over this space, we define distance metrics using the Jensen-Shannon divergence [6] and Mahalanobis distances, which lets us compare documents to each other, and identify which documents are most similar to one another. We will investigate this approach in section 3.

In the later sections we will present our results along with further details of our implementation and the tuning of our parameters, followed by our discussion and conclusions.

1.1 The Gutenberg Dataset

The Gutenberg Dataset [2] is a collection of 3,036 English books written by 142 authors. For the scope of this paper we have considered a sample of 200 books among the collection. This data was selected to include books of varying length, written by a variety of writers from Alduous Huxley to Abraham Lincoln on a number of different topics. The selection is aimed to replicate the nature of the entire corpus, through this variety. Nonetheless, we decided to filter some of these documents by length, to have a balanced distribution decreasing the size of the selection to 187. Still, the shortest book in the selection is 2,600 words whereas the longest is approximately 147,000 words long.

2 Model

2.1 Latent Dirichlet Allocation

Latent Dirichlet Allocation is a generative model used to infer the underlying latent structure of a given set of documents, which we call a corpus. It's most noteable for assuming that each document exhibits multiple topics. The hierarchical model, is composed of a random mixture over the set of topics, where each topic itself is a distribution over the vocabulary. The full specification with the generative model is described below. Note that a more rigorous description can be found in [1].

Given a set of documents of length D, and a vocabulary of size V containing words in the entire English vocabulary, and the number of topics K in the corpus, we assume that for $d \in \{1, \dots, D\}$, $v \in \{1, \dots, V\}$, $k \in \{1, \dots, K\}$

- The per-document topic distribution is $\theta_d \sim \text{Dir}(\alpha)$
- The per-topic distribution over the vocabulary is $\varphi_k \sim \text{Dir}(\beta)$
- For each $i \in \{1, \dots, N_d\}$, where N_d is the number of elements of the d-th document,
  - Sample a topic $z_i^{(d)} \sim \text{Multinomial}(\theta_d)$
  - Sample word from vocabulary of $z_i^{(d)}$-th topic, $w_i^{(d)} \sim \text{Multinomial}(\varphi_{z_i^{(d)}})$

  where $z_i^{(d)}$ indexes the topic and α, β are prior weights usually set constant to a number less than 1, to enforce sparser topic and word distributions. α is a K-dimensional vector, as a mixture coefficient for each topic and β is V-dimensional, components corresponding to the weights for each topic distribution over the words (φ). Hence, the

The associated graphical structure is given below,

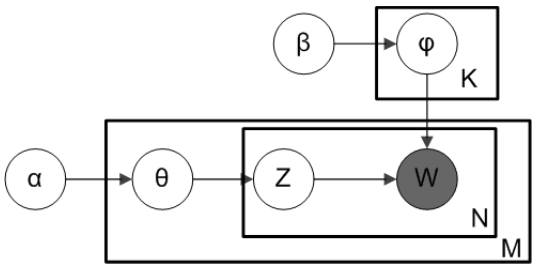

Figure 1: Graphical model for LDA

The edges indicate dependence relations and the model defines a factorization for the joint distribution of the hidden variables and observed texts. The factorization can be given as,

$$P(W, Z, \theta, \varphi, \alpha, \beta) = \prod_{k=1}^{K} P(\varphi_k; \beta) \prod_{d=1}^{D} P(\theta_j; \alpha) \prod_{i=1}^{N_d} P(Z_i^{(d)}|\theta_d) P(W_i^{(d)}|\varphi_{z_i^{(d)}}) \tag{1}$$

2.2 Inference

Given a corpus, the goal is to infer the latent topics in the documents in the corpus, using the assumptions laid out above. To do this, we must find the posterior distributions of the hidden variables θ, z, φ (ignoring the indices) (E-step or inference), and then find the optimal values that maximize these distributions (M-step). For the posterior distribution it is enough to evaluate,

$$p(\theta, z, \varphi|W, \alpha, \beta) = \frac{p(\theta, z, \varphi, W, \alpha, \beta)}{p(W|\alpha, \beta)}$$

The joint is immediately available using the factorization and assumptions laid out above. However, the marginal distribution for the observed words is intractable, and thus no closed form solution exists. A number of approaches have been proposed in literature [1],[3],[4]. We utilize two of these algorithms for the purpose of estimating the posterior distribution: variational inference described in [2] and collapsed gibbs sampling described in [3].

2.2.1 Variational Inference

According to the method proposed in the original LDA paper [1], we can use a variational bayes (VB) method, with simpler proposal distribution $q(\theta, z, \varphi|\gamma, \lambda)$, where γ, λ are free parameters for the topic distributions. Using the KL-divergence, we pose a minimization problem to minimize the KL-divergence between the proposal distribution and the true posterior p for the variational parameters γ, λ.

$$\lambda^*, \gamma^* = \arg\min_{\gamma,\lambda} D\left(q(z, \theta, \varphi|\gamma, \lambda) \,||\, p(\theta, z, \varphi|W, \alpha, \beta)\right)$$

2.2.2 Gibbs Sampling

In the Gibbs sampling approach, given the posterior distribution sought to be estimated, we sequentially sample each the variables from the known posterior distributions, conditioning on the current samples of other dependent parameters. Defining each state as a full estimation of these parameters, the MCMC method guarantees that the distribution of the variables being sampled will converge to the respective target posterior distributions.

Since the joint distribution given in (1) is only dependent on the observations through Z, we can marginalize over θ and φ to get the posterior distribution for $p(z|w, \alpha, \beta)$. Then the Gibbs sampling approach can be used to estimate this distribution by sampling from the following conditional distribution,

$$p(z_i = k|z_{-i}, w) \propto \frac{n_{-i,k}^{(v)} + \beta}{\sum_{v=1}^{V} n_{-i,k}^{(v)} + V\beta} \frac{n_{-i,k} + \alpha}{\sum_{k=1}^{K} n_{-i,k} + D\beta} \tag{2}$$

where $n_{-i,k}^{(v)}$ is the number of times word v appears in topic k and $n_{-i,k}$ is the number of times topic k is observed in a given document.

2.3 Parameter Learning (M-step)

This step is usually mentioned as a part of the inference, but for the sake of this project we will describe it as a separate process, where the parameters are optimized according to the constraints of the respective inference algorithms.

For the variational inference method, a coordinate ascent based approach is used to reach the update equations for each of the variational parameters, holding the other one constant in each update step. These update equations can be found in [1, section 5.2].

For the gibbs sampling method, the z samples are used to estimate θ and ϕ given in detail in [3, equations (6), (7)]. As both of the parameter learning methods mentioned above follow directly from the inference step, we have decided to investigate these algorithms with different dirichlet prior settings using parameters α, β. This is explained in detail in section 4.1.

3 Subsequent Inference

Once we have the topic distributions for each document in the corpus using the inference and learning methods discussed above, we can take the topic distributions to be low-dimensional representations of each document. The topic distributions are points on the $K-1$ simplex, and can be interpreted as showing the extent to which a document belongs to each topic, or exhibits that topic. Furthermore, these representations can be compared in a meaningful way on this simplex using a distance metric.

We suggest two ways of building a recommendation engine based on similarity among the topics of the document.

3.1 Dissimilarity Metric using Jensen-Shannon Divergence

The first is based on using the Jensen-Shannon divergence between two distributions (or points on the $K-1$ simplex). Using this distance metric, we can sort the documents closest to each other, and given a book from the corpus, we can recommend 10 closest books on the simplex using this measure. For topic distributions P and Q for two documents, the distance measure can be given as

$$\text{JSD}(P||Q) = \frac{1}{2}D(P||M) + \frac{1}{2}D(Q||M)$$

where $M = \frac{P+Q}{2}$.

and can be interpreted as the divergence of these distributions to the average. This, arguably provides a better way of defining closeness on the $K-1$ simplex than the traditional Euclidean distance, using the fact that topic distributions are probabilistic measures. Note that lower values mean higher similarity.

3.2 Clustering by Mahalanobis Distance

Earlier literature has suggested using the most probable topic as the class of a document for clustering. We propose an alternative approach using the distributional properties of the topic distributions like above. Mahalanobis distance uses the precision matrix for topic distributions P, Q to compute

$$M(P,Q) = \sqrt{(P-Q)^T S^{-1} (P-Q)}$$

where S^{-1} is the inverse covariance (or precision) matrix for P and Q. Using hierarchical clustering from Python's scipy library, we can define the clustering to be based on the Mahalanobis distance. Once we get the clusters, than we can then recommend books if they appear in the same clusters.

4 Preprocessing

For initial preprocessing, we remove punctuations and lemmatize each word in the document, for the documents to only contain nouns, adjectives, verbs and adverbs. This also includes removing named entities from the documents, even though this is potentially useful information. However, not removing the names caused the per-topic distributions over the vocabulary to put too much weight on these names, and thus made the topics almost impossible to distinguish in a meaningful way.

As LDA does not use the sequential nature of text data, but solely word frequencies, we then transform the data into a bag-of-words model. The bag-of-words contain only the tokens, and the frequencies/counts of each word in a document. This is then used to train both LDA models. We also

investigate the *tf-idf* model described in [5] for training the LDA model with variational inference and compare the log-likelihoods for each model.

We finally remove the words that appear in less than 20 documents or more than 25 percent of the documents. These values were selected to guarantee a good separation between the topics so that topics were distinct enough.

5 Results

5.1 Parameter Tuning

The α parameter governs the topic distribution per document. Higher values of α, imply that there is a higher probability of sampling from a higher number of topics, whereas lower values of α create sparsity in the topics, meaning each document has higher probabilities for a few topics and close to zero probabilities for the rest. We considered values of $\alpha \in \{0.01, 0.1, 0.8\}$. Setting number of topics to $K = 20$, we compare the log-likelihood results for each α.

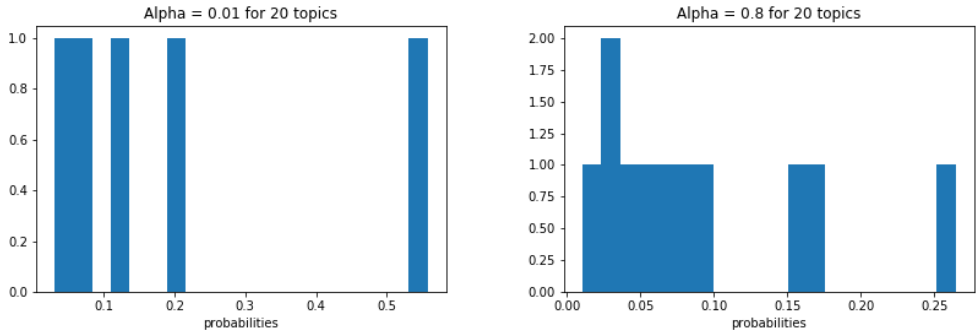

Figure 2: For smaller α the distribution is generally sparse with a few high probability peaks, whereas on the right for larger α a smoother distribution is visible. The plots show the distribution of probabilities for the topic distribution of a sample document.

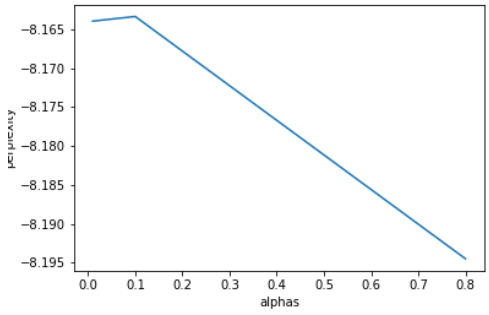

Figure 3: Plot shows perplexity ($e^{\mathscr{L}}$, $\mathscr{L}$ is log-likelihood per word), vs. α's, $\alpha = 0.1$ range seems optimal. (We want to maximize perplexity)

The β parameter governs the per-topic distributions over the vocabulary. Similarly, higher values of β result in each topic containing a mixture of a higher number of words, whereas smaller values create a sparse distribution with high probabilities for a few words. We considered values of $\beta \in \{0.01, 0.1, 0.5\}$. Setting number of topics to $K = 20$, we compare the log-likelihood results for each β in figure 4.

Even though larger β values return higher perplexity values, and thus higher log-likelihood values, we will prefer a smaller β ($\beta = 0.01$) value for the topics to be more distinct, and have less elements in common.

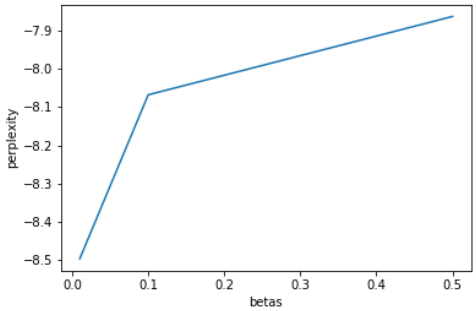

Figure 4: Plot shows perplexity ($e^{\mathscr{L}}$, $\mathscr{L}$ is log-likelihood per word), vs. β's, higher β ranges seems to return higher likelihood. (We want to maximize perplexity)

5.2 Hyperparameter Tuning (Topic Selection)

We also have to select the optimal number of topics, so we compare the bag-of-words (BOW) model to the tf-idf model for each topic. We consider number of topics $K \in \{5, 10, 20, 40, 100, 400\}$. The results are given in the figure below.

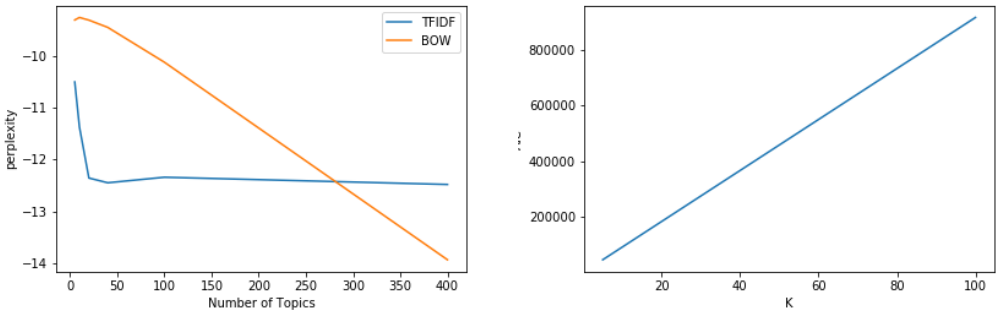

Figure 5: (Left) Perplexity vs. Number of Topics for TFIDF and Bag-of-words models, (Right) AIC vs. Number of Topics for the Gibbs Sampling Model

Based on results from figure 5, we pick number of topics $K = 10$. We also confirm this result using the Gibbs sampling LDA model, with the Akaike Information Criterion (AIC = $-\mathscr{L} + 2p$, p is the number of parameters) in the RHS of figure 6.

Furthermore, since BOW model for $K = 10$, shows better performance, we pick the BOW model to compare against the Gibbs Sampling model.

5.3 Topic Distributions for Variational Inference and Gibbs Sampling LDA Models

The mismatch in the order of magnitudes of the likelihood values from the BOW Model and the Gibbs Sampling model made a direct comparison between these two models unfeasible. But, visualizing their respective topic distributions give us a way of visually comparing the two. We project the per-topic distributions over the vocabulary, which are each >4000 dimensional, onto their two principal components using PCA. The figures are given below in Figure 7. The lower-dimensional representation of the two results are similar in that, 4 topic are visible that are distinct, and the rest seem relatively close to each other, such that their differences are not as significant. Note that, the topics from each distribution are identified by different numbers.

For example, the most dominant 10 words in topic 5 is, [formation, fig, slate, thickness, pebble, granite, layer, jean, volcanic, loop] for the Gibbs sampling model and the most dominant 10 words in

7 for the Variational inference model are [formation, roger, jane, tar, slate, thickness, pebble, layer, myth, granite]. Thus, the topics indeed seem to track each other closely.

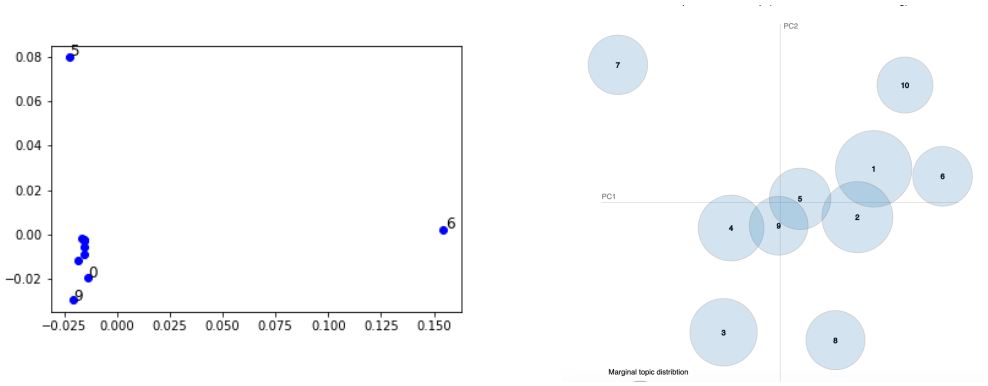

Figure 6: Distributions over the vocabulary for each topic, projected onto the principal components. (Left) Results for Gibbs sampling, (Right) Results for the Variational Inference model.

5.4 Recommendation Engine by Jensen-Shannon and Mahalanobis Distance

The goal of the recommendation engine is for it to provide books that are similar to the one that has been read. We pick Charles Darwin's Geological Observations On South America as our sample book. Below are the 5 recommendations the Variational Inference and Gibbs Sampling methods provide using the Jensen-Shannon distance based dissimilarity measure. The recommendations are almost the same, except the 5th recommendations, although they have different dissimilarity values.

Title	Dissimilarity (JS)
Sir Francis Galton, Finger Prints	0.313
John Ruskin, The Elements of Drawing	0.499
Thomas Henry Huxley, The Present Condition of Organic Nature	0.561
R M Ballantyne, Handbook to the new Gold-fields	0.584
Thomas Carlyle, History of Friedrich II of Prussia, Appendix	0.588

Table 1: Variational Inference Based LDA Recommendations for Darwin's Geological Observations On South America, using Jensen-Shannon Distance

Title	Dissimilarity (JS)
Sir Francis Galton, Finger Prints	0.256
John Ruskin, The Elements of Drawing	0.4461
Thomas Henry Huxley, The Present Condition of Organic Nature	0.512
R M Ballantyne, Handbook to the new Gold-fields	0.528
Sir Richard Francis Burton, The Land of Midian	0.547

Table 2: Gibbs Sampling Based LDA Recommendations for Darwin's Geological Observations On South America, using Jensen-Shannon Distance

The Mahalanobis distance clustering for the variational inference based LDA model, however, returns the only other element in Darwin's cluster, that is the following book: James Fenimore Cooper, The Redskins; or, Indian and Injin, Volume 1.

6 Discussion and Future Work

Using our domain knowledge about Darwin's work, it seems that the recommended books do not seem too far off, as they are almost all non-fiction, scientific books. As the books in the corpus do not have associated categories, it is always difficult to assess the goodness of the recommendations made

through clustering. A deeper survey into the selected books in the corpus can be made in future work to identify related books and to evaluate our models against it.

The lack of such a metric makes comparison of the two recommendation methods discussed above using Jensen-Shannon dissimilarity and Mahalanobis distance also not feasible. Thus, we are not able to say which clustering method one should prefer. Furthermore, we could only apply the Mahalanobis clustering to the variational inference based LDA model, as the Gibbs Sampling model yielded an extremely sparse covariance metric, which was not invertible. Thus, perhaps due to the more general usability of the Jensen-Shannon method, we could argue that we would prefer it over the Mahalanobis based method.

Moreover, we were only able to compare the recommendations of the Gibbs sampling method against the variational inference method, through specific examples, which is not an objective and universal assesment of their efficacy. A future investigation could be to combine these methods in making the recommendations by sorting them according to their dissimilarity measures.

With regards to the hyperparameter selection, a deeper analysis of the topics, and their most dominant words, should yield a better understanding of how many to select the number of topics. Furthermore, one can also imagine using different distributions than the dirichlet and multinomial in the model specification. However, due to time constraints, we leave that for future work.

Finally, our approach of recommending solely based on similarity measures gained from the LDA model, is inadequate to build a sufficiently good classifier. For this sake, other information such as the era, location, writer, along with a person's preferences should also be useful factors in building a powerful recommendation engine.

7 Conclusion

In this paper, we have used two LDA models, one using variational inference, and one using Gibbs Sampling to represent books in the Gutenberg corpus, by their topic distributions. The differences between these two weren't enough to prefer one over the other. This is not surprising as they use the same set of observations to optimize the same distributional parameters. We, then, used our lower-dimensional representation to build a similarity based recommendation system, where given a sample book, the system recommends other books that are similar in terms of their topic representations. We considered Mahalanobis distance and Jensen-Shannon distance based approaches, in building a dissimilarity score and clustering.

References

[1] Blei, David M., et al. "Latent Dirichlet Allocation." Journal of Machine Learning Research, Edited by John Lafferty, 3 Jan. 2003, pp. 993–1022.

[2] Lahiri, Shibamouli. "Complexity of Word Collocation Networks: A Preliminary Structural Analysis." Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, 2014, pp. 96–105, www.aclweb.org/anthology/E14-3011

[3] Griffiths, T. L., and M. Steyvers. "Finding Scientific Topics." Proceedings of the National Academy of Sciences, vol. 101, no. Supplement 1, 2004, pp. 5228–5235., doi:10.1073/pnas.0307752101

[4] Lin, J. "Divergence Measures Based on the Shannon Entropy." IEEE Transactions on Information Theory, vol. 37, no. 1, 1991, pp. 145–151., doi:10.1109/18.61115.

[5] G. Salton and M. McGill, editors. Introduction to Modern Information Retrieval. McGraw-Hill, 1983.

[6] Speh, Jaka, et al. "Algorithms of the LDA Model." Artificial Intelligence Laboratory Jožef Stefan Institute, 1 July 2013, pp. 1–5.