CSE 250B: Project Report
Topic Modeling using Latent Dirichlet Allocation

Syed Aziz Enam, Radheshyam Balasundaram

March 16, 2014

Abstract

A Latent Dirichlet Allocation model(LDA) is trained using collapsed Gibbs Sampling to uncover hidden topics in a corpus. The accuracy of LDA is measured on two different data sets, Classic400 and New York Times news articles. A confidence measure is also introduced to measure goodness-of-fit of the LDA model.

1 Introduction

Topic Models are algorithms used to discover the main themes that exist within a text corpus. For LDA, each theme or topic is defined by a certain distribution of words. Given the word distribution, documents are characterized as a distribution over topics. To train our LDA model a collapsed Gibbs Sampling algorithm is used. Collapsed Gibbs Sampling does not infer the topic distribution θ or the word distribution ϕ,but instead the latent topic $\vec{z}$ is learned. The following expression is used to sample z given the bag-of-words vector ω:

$$p(z_i|z_{-i},\omega) \propto \frac{q_{j\omega_i} + \beta_{\omega i}}{\sum_t q_{jt} + \beta_i}(n_{mj} + \alpha_{\omega j}) \tag{1}$$

Where n_{mj} number of times topic j occurs in the document and is q_{jt} number of times the word t appears in the topic j. The α, β values are pseudo counts added by the Dirichlet priors. Reader may refer to the lecture notes (see [Elkan(2014)]) for a derivation of the expressions and a more detailed treatment of the models.

After sufficient iterations of Gibbs sampling the distribution of z's converge to the true distribution. Given the distribution we are able to infer the parameters θ and ϕ for the respective distribution from these equations:

$$\theta_k = \frac{\sum_{i=1}^{n} I(z_i = k)}{n} \tag{2}$$

$$\phi_{k,w_i} = \frac{\text{No. of words } w_i \text{ with type } k}{\text{No. of words with type } k} \tag{3}$$

In section 2 we describe our datasets and implementation of our algorithms. In section 3 we explain the experiments that were performed to validate our model. In section 4 we discuss the results of the experiments along with our analysis.

2 Implementation Details

2.1 Data Sets

Two different datasets were used to evaluate the LDA model. The first dataset was the Classic400 dataset that consists of a vocabulary of 6205 words and a bag-of-words representation of 400 documents. Each document is also labeled with one of three topics.

The second dataset was a custom dataset which consists of news articles from the New York Times website. The articles were taken from the categories of politics, business, arts, science, and sports. A vocabulary was created by combining the top words in each category resulting in a vocabulary size of 3528 words, after eliminating stop words. The data was maintained in Compressed Sparse Column format to improve the time complexity. This helped in maintaining the running time at $O(NK)$ for one epoch, where K is the number of topics and N is the total number of word occurrances in the corpus.

3 Design of Experiments

3.1 Confidence measure for convergence

In order to measure convergence, for various values of K, we defined a *Confidence Measure* for every document.

Definition 1 (Confidence Measure)**.** For a given document m and a $\vec{z}$ for the document, let θ_k be as defined in Equation 2. We finally categorize the document to the topic k_{BEST} with the highest θ_k value. That is,

$$k_{BEST} = \arg\max_k \theta_k$$

Hence, $\max_k \theta_k$ gives a very good measure of the confidence with which which we can say that the line belongs to the topic k_{BEST}. So, we define **Confidence Measure** as $\theta_{k_{BEST}}$. Notice that the confidence measure, at any stage, lies in the range $[\frac{1}{K}, 1.0]$.

In our experiments, after every 25 epochs, we measured this confidence measure for every document. We sorted the confidence measures across the documents and looked at the $25^{th}, 50^{th}, 75^{th}$ and 100^{th} percentiles of these confidence measures. We observed that these values increase as number of epochs increased, until a certain point and then stayed the same. We used this to determine the stopping condition.

3.2 Goodness of Fit

We observed that the Confidence Measure (Definition 1) at the time of convergence is a very good indicator for measuring goodness of fit. When we experimented with different values for K and observed the top words in each topic, the correlation among the words was best when this confidence measure was high enough. We also observed that at the time of convergence, the 25^{th} percentile confidence measure was around $\frac{K-1}{K}$ (recall that the lowest can be $\frac{1}{K}$).

	Topic A	Topic B	Topic C
Topic 1	1	**96**	3
Topic 2	0	1	**99**
Topic 3	**191**	0	9

	Accuracy of Topic
Topic 1	96%
Topic 2	99%
Topic 3	95.5%

Table 1: Table on left is Categorization matrix of LDA topics trained for Classic400 Dataset. Each row corresponds to the "trueLabel" and each column corresponds to the trained label. Table on the right is the accuracy of Topic prediction.

Topic A "Aeronautics"	Topic B "Medicine"	Topic C "Scientific methods"
boundary	patients	field
layer	cases	free
wing	normal	general
mach	due	system
supersonic	ventricular	problems
effects	function	research
ratio	left	fatty
wings	nickel	scientific

Table 2: Top ten words learned for each topic in the Classic400 Dataset

4 Results

In this section, we present the results of our experiments with both the data sets.

4.1 Classic400 Dataset

We observed that the number of words in the vocabulary were 6905 and number of documents were 400. The average length of a document was close to 78. To avoid bias due to ordering of documents, we shuffled the documents in the corpus.

We first ran Gibbs Sampling with three topics topics ($K = 3$). The sampling algorithm converged after about 75 epochs. After training our model, we use θ vector to plot the documents as points in the three dimensional space based on the topics of the trained model shown in Figure 4.1.

Table 1 shows the categorization matrix, where each row corresponds to the true topic, as given in the dataset and each column corresponds to the topic predicted by our algorithm. So, "Topic 1" of input was predicted as "Topic B", "Topic 2" as "Topic C" and "Topic 3" as "Topic A".

In Table 2, we present the top 10 words that appeared in each topic. From these words, we predicted that Topic 1 was related to "Medicine", Topic 2 was "Scientific Methods" and Topic 3 was "Aeronautics".

In Table 4.1, we examine the behavoir of LDA on the dataset when we choose to train four topics instead of three i.e $K = 4$. We found that Topic 1 was split approximately between two trained topics.

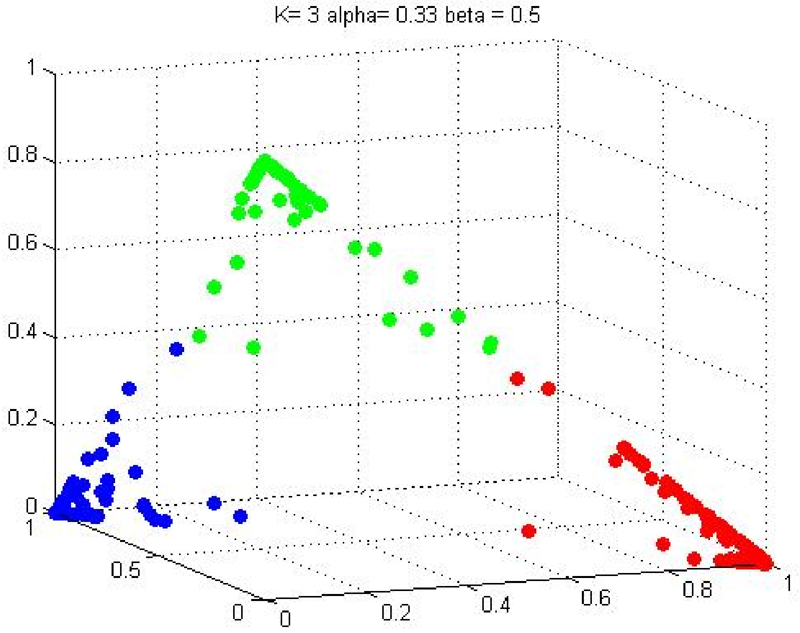

Figure 1: Plot of the documents as points in 3D space based on the the topics of the trained models.

	Topic A	Topic B	Topic C	Topic D
Topic 1	1	1	**41**	**57**
Topic 2	0	**99**	1	0
Topic 3	**192**	7	0	1

Table 3: Categorization matrix of LDA topics trained for Classic400 Dataset with $K = 4$. Each row corresponds to the "trueLabel" and each column corresponds to the trained label.

“Politics”	“Science”	“Arts”	“Sports”	“Business”
public	explain	experience	playgrounds	employers
passed	previous	manual	playoffs	company
scandal	published	producers	season	health
budget	language	theatre	stands	investment
politics	technology	digital	giants	ceremonies
said	museum	learning	athletes	talk
obama	natural	talk	baseball	foundation
president	paper	costume	madrid	arbitration

Table 4: Top ten words learned for each topic in the New York Times Dataset

4.2 New York Times Dataset

We chose the number of words in the vocabulary to be 3528 and 471 documents were extracted using New York Times web API (see [Times(2014)]). The average length of a document was close to 91. When compared the the other dataset, the length of documents in this datset were considerably high. To avoid bias due to ordering of documents, we shuffled the documents in the corpus.

Gibbs Sampling was run with five topics ($K = 5$). The sampling algorithm converged after about 100 epochs.

In Table 4, we present the top 10 words that appeared in each topic. From these words, we predicted that Topic s were related to “Politics”, “Science”, “Arts”, “Sports”,“Business”.

5 Conclusion

After evaluating LDA topic modeling on the two datasets, Classic400 and New York Times articles, we found that LDA is an effective method to cluster documents. We qualitatively assessed the effectiveness of the learned topics by examining the top words in each category and confirmed that they were coherent. We also introduced a confidence measure to quantify a goodness-of-fit.

References

[Elkan(2014)] Charles Elkan. Text mining and topic models. University Lecture, 2014. URL `http://cseweb.ucsd.edu/ elkan/250B/topicmodels.pdf`.

[Times(2014)] New York Times. Times developer network - article search api v2, February 2014. URL `http://developer.nytimes.com/docs/read/article_search_api_v2`.