**Topic Modeling using Latent Dirichlet Allocation**

A thesis submitted in partial fulfillment of the

requirements for the award of degree of

**Bachelor of Technology**

in

**Computer Science and Engineering**

By

KARRI REVATHI (N100271)

KODUMURI TARUN KUMAR (N100426)

*Under the Supervision of*

**Mr. Udaya Kumar Ambati**

Assistant Professor Dept. of CSE



**Department of Computer Science and Engineering**

**Rajiv Gandhi University of Knowledge Technologies – Nuzvid**

**Nuzvid, Krishna – 521202**

April, 2016

**--------------------------------------------------------------------------------------------------------------------------------------**

*CERTIFICATE OF COMPLETION*

This is to certify that the work entitled,**”Topic modeling using Latent Dirichlet Allocation”** is the bonafied work of ***Karri Revathi , ID No: N100271*** , ***Kodumuri Tarun Kumar , ID No: N100426 ,c***arried out under my guidance and supervision for the partial fulfillment of the requirement for the award of the degree of **Bachelor of Technology** in the department of Computer Science and Engineering under RGUKT IIIT Nuzvid. This work is done during the academic session August 2015 – May 2016, under our guidance.

---------------------------- ---------------------------

**Mr. A.Udaya Kumar Mr. K K Singh**

Project SupervisorHead of Department

Assistant Professor Dept. of CSE Assistant Professor Dept. of CSE

RGUKT IIIT Nuzvid, Nuzvid RGUKT IIIT Nuzvid, Nuzvid

*CERTIFICATE OF EXAMINATION*

This is to certify that the work entitled, “**Topic Modeling using Latent Dirichlet Allocation”** is the bonafide work of ***Karri Revathi, ID No: N100271, Kodumuri Tarun Kumar, ID No: N100426*** and hereby accord our approval of it as a study carried out and presented in a manner required for its acceptance in the partial fulfillment of the requirement for the award of the degree of **Bachelor of Technology** for which it has been submitted. This approval does not necessarily endorse or accept every statement made, opinion expressed or conclusion drawn, as a recorded in this thesis. It only signifies the acceptance of this thesis for the purpose for which it has been submitted.

---------------------------- ---------------------------

**Mr. A.Udaya Kumar Examiner**

Project SupervisorProject Examiner

Assistant Professor Dept. of CSE Lecturer Dept. of CSE

RGUKT IIIT Nuzvid, Nuzvid RGUKT IIIT Nuzvid, Nuzvid

**--------------------------------------------------------------------------------------------------------------------------------------**

*DECLARATION*

We, ***Karri Revathi, ID No: N100271, Kodumuri Tarun Kumar, ID No: N100426*** hereby declare that the project report entitle “**Topic Modeling using Latent Dirichlet Allocation”** done by us under the guidance of **Mr. Udaya Kumar M.Tech** is submitted for the partial fulfillment of the requirement for the award of the degree of **Bachelor of Technology** in **Computer Science and Engineering** the academic session August 2015 – April 2016 at RGUKT – Nuzvid.

We also declare that this project is a result of our own effort and has not been copied or imitated from any source. Citations from any websites are mentioned in the references.

The results embodied in this project report have not been submitted to any other university or institute for the award of any degree or diploma.

**Date : \_\_\_\_\_\_\_\_\_\_\_\_ Karri Revathi [N100271]**

**Place :\_\_\_\_\_\_\_\_\_ Kodumuri Tarun Kumar [N100426]**

*ACKNOWLEDGEMENT*

We would like to express our profound gratitude and deep regards to my guide *Mr.Ambati Udaya Kumar* for his exemplary guidance, monitoring and constant encouragement throughout the course of this thesis.

We are extremely grateful for the confidence bestowed in us and entrusting our project entitled “Topic Modeling using Latent Dirichlet Allocation”.

At this juncture we feel deeply honored in expressing our sincere thanks to him for making the resources available at right time and providing valuable insights leading to the successful completion of our project.

We would like to thank RGUKT Nuzvid Director, faculty and staff for their valuable suggestions and discussions.

Last but not least I thank almighty, and I place a deep sense of gratitude to my family members and my friends who have been constant source of information during the preparation of this project work.

KARRI REVATHI

KODUMURI TARUN KUMAR

**ABSTRACT**

Topic modeling is a powerful technique for unsupervised analysis of large document collections. Topic models conceive latent topics in text using hidden random variables, and discover that structure with posterior inference. Topic models have a wide range of applications like document classification,tag recommendation, text categorization, keyword extraction and similarity search in the broad fields of text mining, Information Retrieval, statistical language modeling. In our work, a dataset with 2225 documents fall under five topics are collected from BBC news,an Email dataset of a company containing 500k mails,a dataset containing documents from different languages are used. The document models are built using LDA (Latent Dirichlet Allocation) with Gibbs sampling. Then the built model is used to extract appropriate topics from the whole corpus.

**ABBREVIATIONS**

LDA Latent Dirichlet Allocation

NLP Natural Language Processing

LSA Latent Semantic Analysis

PLSA Probabilistic Latent Semantic Analysis

MCMC Markov Chain Monte Carlo

# 

# 

# 

# 

# **1. INTRODUCTION**

As our collective knowledge continues to be digitized and stored in the form of news, blogs,web pages, scientific articles, books, images, sound, video, and social networks and it becomes more difficult to find and discover what we are looking for. We need new computational tools to help organize, search and understand these vast amounts of information.

Topic modeling is an Information Retrieval (IR) technique that discovers representative topics from a collection of documents. Thus, we expect that logically related words will co-exist in the same document more frequently than words from different topics. For example, in a document about the space, it is more possibly to find words such as: planet, satellite, universe, galaxy, and asteroid. Whereas, in a document about the wildlife, it is more likely to find words such as: ecosystem, species, animal, and plant, landscape.Topic modeling is a new trend that can offer fast and accurate results when we need to analyze large datasets.Probabilistic topic models are a suite of algorithms whose aim is to discover the hidden thematic structure in large archives of documents.

# **2. LITERATURE REVIEW**

Various techniques are there to implement the topic model.

## **2.1. Latent semantic indexing (LSI)**

LSI [[1]](https://www.colwiz.com/cite-in-google-docs/cid=f20eb6d604423aa) is an indexing and retrieval method that uses a mathematical technique called singular value decomposition (SVD) to identify patterns in the relationships between the terms and concepts contained in an unstructured collection of text. LSI is based on the principle that words that are used in the same contexts tend to have similar meanings. A key feature of LSI is its ability to extract the conceptual content of a body of text by establishing associations between those terms that occur in similar contexts.This approach can achieve significant compression in large collections. Furthermore, Deerwester et al. argue that the derived features of LSI, which are linear combinations of the original tf-idf features, can capture some aspects of basic linguistic notions such as synonymy and polysemy.To substantiate the claims regarding LSI, and to study its relative strengths and weaknesses, it is useful to develop a generative probabilistic model of text corpora and to study the ability of LSI to recover aspects of the generative model from data.A significant step forward in this regard was made by Hofmann (1999), who presented the probabilistic LSI (pLSI) model, also known as the aspect model, as an alternative to LSI.

## **2.2. Probabilistic Latent Semantic Indexing(pLSI)**

The pLSI approach [[2]](https://www.colwiz.com/cite-in-google-docs/cid=f20f281ea1193e5), models each word in a document as a sample from a mixture model, where the mixture components are multinomial random variables that can be viewed as representations of “topics.” Thus each word is generated from a single topic, and different words in a document may be generated from different topics. Each document is represented as a list of mixing proportions for these mixture components and thereby reduced to a probability distribution on a fixed set of topics. This distribution is the “reduced description” associated with the document.It is useful step toward probabilistic modeling of text, it is incomplete in that it provides no probabilistic model at the level of documents. In pLSI, each document is represented as a list of numbers (the mixing proportions for topics), and there is no generative probabilistic model for these numbers. This leads to several problems: (1) the number of parameters in the model grows linearly with the size of the corpus, which leads to serious problems such as overfitting (2) It is not clear how to assign probability to a document outside of the training set.

All theses models use bag of model approach [[3]](https://www.colwiz.com/cite-in-google-docs/cid=f20f2ce5da0776b) which means that words can be exchangeable.A classic representation theorem due to de Finetti (1990) establishes that any collection of exchangeable random variables has a representation as a mixture distribution in general an infinite mixture. Thus, if we wish to consider exchangeable representations for documents and words, we need to consider mixture models that capture the exchangeability of both words and documents.This line of thinking leads to the latent Dirichlet allocation (LDA) model.

## **2.3. Latent Dirichlet Allocation(LDA):**

It is important to emphasize that an assumption of exchangeability is not equivalent to an assumption that the random variables are independent and identically distributed. Rather, exchangeability essentially can be interpreted as meaning “conditionally independent and identically distributed,”where the conditioning is with respect to an underlying latent parameter of a probability distribution. Conditionally, the joint distribution of the random variables is simple and factored while marginally over the latent parameter, the joint distribution can be quite complex. Thus, while an assumption of exchangeability is clearly a major simplifying assumption in the domain of text modeling, and its principal justification is that it leads to methods that are computationally efficient, the exchangeability assumptions do not necessarily lead to methods that are restricted to simple frequency counts or linear operations.

# **3. PRIOR KNOWLEDGE**

## **3.1. Notations and Terminology**

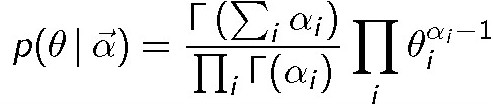
We use the language of text collections throughout the paper, referring to entities such as “words,” “documents,” and “corpora.” This is useful in that it helps to guide intuition, particularly when we introduce latent variables which aim to capture abstract notions such as topics. It is important to note, however, that the LDA model is not necessarily tied to text, and has applications to other problems involving collections of data, including data from domains such as collaborative filtering, content-based image retrieval and bioinformatics.We present experimental results in the collaborative filtering domain. Formally, we define the following terms:

* A *word* is the basic unit of discrete data, defined to be an item from a vocabulary indexed by {1,….,V}. We represent words using unit-basis vectors that have a single component equal to one and all other components equal to zero. Thus, using superscripts to denote components,the *v*th word in the vocabulary is represented by a *V*-vector *w* such that *wv* = 1 and *wu* = 0 for *u≠v*.
* A *document* is a sequence of *N* words denoted by **w** = (w1,w2,…..,wN), where *wn* is the *n*th word in the sequence.
* A *corpus* is a collection of *M* documents denoted by *D* = (w1,w2,…..,wM).

We wish to find a probabilistic model of a corpus that not only assigns high probability to members of the corpus, but also assigns high probability to other “similar” documents.Probability distributions which are used in this explained below.

## **3.2. Dirichlet Distribution**

In probability and statistics, the **Dirichlet distribution** is often denoted as Dir(𝛼), is a family of continuous multivariate probability distributions parameterized by a vector 𝛼 of positive reals.It is the multivariate generalization of the beta distribution.Dirichlet distributions are very often used as prior distributions in Bayesian statistics, and in fact the Dirichlet distribution is the conjugate prior of the multinomial distribution.



## **3.3. Multinomial Distribution:**

In probability theory, the multinomial distribution is a generalization of the binomial distribution.For example it models the probability of counts for rolling a *k* sided dice *n* times. For *n* independent trials each of which leads to a success for exactly one of *k* categories, with each category having a given fixed success probability, the multinomial distribution gives the probability of any particular combination of numbers of successes for the various categories.When *n* is 1 and *k* is 2 the **multinomial distribution** is the Bernoulli distribution. When *k* is 2 and number of trials are more than 1 it is the Binomial distribution. When *n* is 1 it is the categorical distribution. It is sometimes convenient to express the outcome of a categorical distribution as a "1-of-K" vector.

The Dirichlet distribution is a probability distribution over the space of multinomial distributions, i.e., to generate data X from a Dirichlet distribution with parameters *α*1,*α*2,*α*3 first draw a****∼*Dir*(*α*⃗ ), and then draw *X*∼*Multi*(****).

# **4. THEORETICAL BACKGROUND FOR LDA**

Latent Dirichlet allocation (LDA) [[4]](https://www.colwiz.com/cite-in-google-docs/cid=f20f5ab549d3152) is a generative probabilistic model of a corpus. The basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words.

**Model:**

Withplate notation, the dependencies among the many variables can be captured concisely. The boxes are “plates” representing replicates. The outer plate represents documents, while the inner plate represents the repeated choice of topics and words within a document. M denotes the number of documents, N the number of words in a document.

* *α* is the parameter of the Dirichlet prior on the per-document topic distributions,
* *β* is the parameter of the Dirichlet prior on the per-topic word distribution,
* 𝛳i is the topic distribution for document *i*,
* 𝜑k is the word distribution for topic *k*,
* Zij is the topic for the *j*th word in document *i*, and
* wij is the specific word.
* K denotes number of topics.

****

Figure 1: Plate Notation representing LDA

LDA assumes the following generative process for each document **w** in a corpus *D*:

1. Choose *N* ~ Poisson(𝜉).

2. Choose 𝛳i ~ Dir(𝛼).

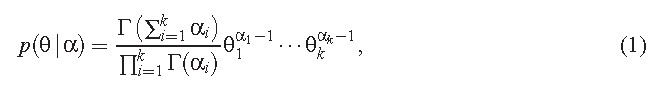
3. Choose 𝜑k ~ Dir(𝛽).

4. For each of the *N* words *wn*:

(a) Choose a topic *zij* ~Multinomial(𝛳i).

(b) Choose a word *wn* from *p*(*wn /zn;*b), a multinomial probability conditioned on the topic *zn*.

Several simplifying assumptions are made in this basic model, some of which we remove in subsequent sections. First, the dimensionality k of the Dirichlet distribution (and thus the dimensionality of the topic variable z) is assumed known and fixed. Second, the word probabilities are parameterized by a k x V matrix 𝛽 where 𝛽ij = p(wj = 1/zj = 1), which for now we treat as a fixed quantity that is to be estimated. Finally, the Poisson assumption is not critical to anything that follows and more realistic document length distributions can be used as needed. Furthermore, note that N is independent of all the other data generating variables (𝛳and z). It is thus an ancillary variable and we will generally ignore its randomness in the subsequent development.  
 A k-dimensional Dirichlet random variable 𝛳 can take values in the (k−1)-simplex (a k-vector 𝛳 lies in the (k−1)-simplex if 𝛳i⩾0, ), and has the following probability density on this simplex:

where the parameter 𝛼 is a *k*-vector with components 𝛼*i >*0, and where 𝛤(*x*) is the Gamma function.The Dirichlet is a convenient distribution on the simplex—it is in the exponential family, has finite dimensional sufficient statistics, and is conjugate to the multinomial distribution. These properties will facilitate the development of inference and parameter estimation algorithms for LDA.

Given the parameters 𝛼 and β, the joint distribution of a topic mixture 𝛳, a set of *N* topics **z**, and a set of *N* words **w** is given by:



## 4.1. Influence of Priors

If Alpha value is more a document contains mixture of most of the topics.

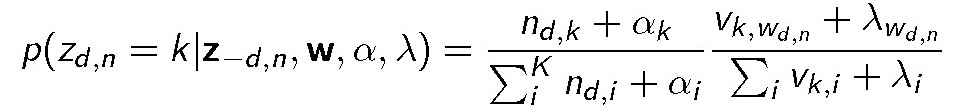
If Eta value is more,a topic is a mixture of most of the words.

default value for both is (1/num\_topics).

## **4.2. Inference**

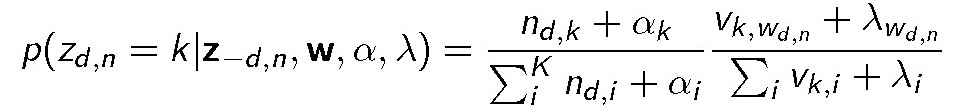
Learning the various distributions (the set of topics, their associated word probabilities, the topic of each word, and the particular topic mixture of each document) is a problem of Bayesian inference.We can use variational Bayes approximation of the posterior distribution and alternative inference techniques use Gibbs sampling and expectation maximization.We used gibbs sampling to find the posterior distribution.

## **4.3. Gibbs Sampling:**

In statistics, **Gibbs sampling**[**[5]**](https://www.colwiz.com/cite-in-google-docs/cid=f20f2ce5da26c77) or a **Gibbs sampler** is a Markov chain Monte Carlo (MCMC) algorithm for obtaining a sequence of observations which are approximated from a specified multivariate probability distribution, when direct sampling is difficult. This sequence can be used to approximate the joint distribution, to approximate the marginal distribution of one of the variables, or some subset of the variables (for example, the unknown parameters or latent variables) or to compute an integral. Typically, some of the variables correspond to observations whose values are known, and hence do not need to be sampled.

* Nd,k is number of times document d uses topic k
* Vk,Wd,n number of times topic k uses word wd,n
* 𝛼 Dirichlet parameter for document to topic distribution
* 𝛽 Dirichlet parameter for topic to word distribution
* How much the document likes topic k
* How much this topic likes word wd,n

**Algorithm:**

* For each iteration i:
  + For each document d and word n currently assigned to zold
    - Decrement nd,z\_old and vz\_oldWd,n
    - Sample znow =k with probability proportional to
    - Increment nd,z\_new and vz\_now wd,n

Numerous variations of the basic Gibbs sampler exist. The goal of these variations is to reduce the autocorrelation between samples sufficiently to overcome any added computational costs.

## **4.4. Collapsed Gibbs Sampling:**

The result of this collapsing introduces dependencies among all the categorical variables dependent on a given Dirichlet prior, and the joint distribution of these variables after collapsing is a Dirichlet-multinomial distribution. The conditional distribution of a given categorical variable in this distribution, conditioned on the others, assumes an extremely simple form that makes Gibbs sampling even easier.

# **5. OUR WORK**

## **5.1. Preprocessing**

Data preprocessing is a technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors.We applied some techniques like tokenization,stemming etc.

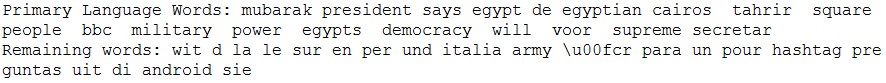
* First we split the text into individual words or sequences of words (tokenization).
* Removed all stop words.
* Stemmed the words by using porter stemmer algorithm.
* Removed the vocabulary which are present only once in the whole corpus.

This preprocessing step is common for all the below implementations before applying LDA.

## **5.2. Unsupervised Language Filtration**

It is the process of ‘purifying’ a text corpus to determine which sentences are in the primary language of the corpus and contain no foreign words or phrases.This is a requirement for building the language processing front-end of a speech synthesis system entirely automatically in a new language where linguist resources other than the text are unavailable.It is a form of text categorization,It is very useful in processing of microblogging and social media websites.Several kinds of classification approaches like markov models,hidden models etc are used to identifying the language of documents but these methods are all supervised, require clean editorially managed corpora for training and appropriate only for a limited number of languages, and require relatively large-sized documents.To be able to identify language in an unsupervised fashion we adopt and adapt a model Latent Dirichlet Allocation [[6]](https://www.colwiz.com/cite-in-google-docs/cid=f20f9d329303b4e) from the field of Topic modelling.

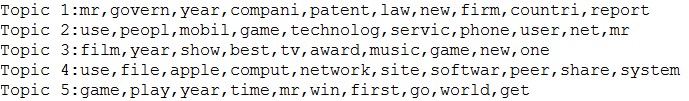
**Input:** German-10,English-40,Spanish-10,French-10,Italian-10,Dutch-10

**Output:**

## **5.3. Topic Modeling of BBC News Dataset**

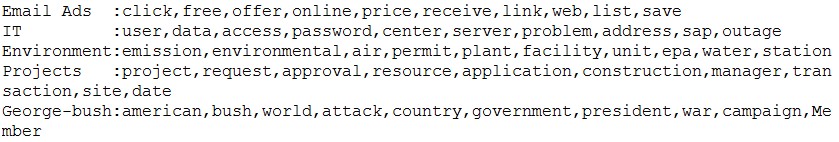
The starting point for all these approaches will be word frequency distributions. What does that mean? A news article about immigration will likely contain more geolocations, names of countries with conflict zones, and politicians who play a role in that context, while a champions league article will contain a slightly different vocabulary.We have used Latent Dirichlet Allocation with collapsed Gibbs sampling to find out the topics in the 2225 news articles which are published in BBC.We have collected the articles from five topics.But,here we are doing soft clustering means instead of assigning topic to the article,we are giving the probability that it belongs to a certain topic.

**Results:**

****

## **5.4. Topic Modeling of Enron Email Corpus**

This dataset contains about 500,000 E-mails from nearly 150 senior management at Enron and was released during the investigation into the massive fraud going on at the company in the late 90's and early 00's.By applying Latent Dirichlet Allocation with Collapsed Gibbs sampling gives the following results.We have taken number of topics K as 50.

**Results:**

# 

# 6. CONCLUSION

Topic modeling of text collections is rapidly gaining importance for a wide variety of applications.LDA aims to find structure within an unstructured collection of documents. After learning this “structure,” a topic model can answer questions such as: What a document is discussing.How similar the documents are and which documents are sufficient to know about a particular topic can be known easily.

# 

# 

# 

# 

# 7. FUTURE WORK

We have applied Latent Dirichlet Allocation on different datasets and got pretty good results.Next step is to apply efficient stemming and Named Entity recognition techniques to those results to get meaningful insights.It can be used in feature extraction.Here the labels are assigned to topics.We can improvise this to specify some labels in prior.And assign probabilities for that labels in each document.

[**References:**](https://www.colwiz.com/cite-in-google-docs/cid=colwizBiblio)

[[1] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, “Indexing by latent semantic analysis,” *J. Am. Soc. Inf. Sci.*, vol. 41, pp. 391–407, Jan. 1990.](https://www.colwiz.com/cite-in-google-docs/cid=colwizBiblio)

[[2] D. M. Blei, “Probabilistic topic models,” *Commun. ACM*, no. 4, pp. 77–84, 2012.](https://www.colwiz.com/cite-in-google-docs/cid=colwizBiblio)

[[3] H. M. Wallach, “Topic modeling: beyond bag-of-words,” in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 977–984.](https://www.colwiz.com/cite-in-google-docs/cid=colwizBiblio)

[[4] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet Allocation,” *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.](https://www.colwiz.com/cite-in-google-docs/cid=colwizBiblio)

[[5] T. Griffiths, “Gibbs sampling in the generative model of latent dirichlet allocation,” *Stanford Univ.*, vol. 518, pp. 1–3, 2002.](https://www.colwiz.com/cite-in-google-docs/cid=colwizBiblio)

[[6] W. Zhang, R. A. J. Clark, Y. Wang, and W. Li, “Unsupervised language identification based on Latent Dirichlet Allocation,” *Comput. Speech & Lang.*](https://www.colwiz.com/cite-in-google-docs/cid=colwizBiblio)

[7] "Introduction to Latent Dirichlet Allocation." *Edwin Chens Blog Atom*. Web. 02 May 2016.

[8] "Latent Dirichlet Allocation." *Wikipedia*. Wikimedia Foundation. Web. 02 May 2016.

[9] "Topic Model." *Wikipedia*. Wikimedia Foundation. Web. 02 May 2016.

[10] "Gensim: Topic Modelling for Humans." *Gensim: Experiments on the English Wikipedia*. Web. 02 May 2016.