# [Latent Dirichlet allocation](https://en.wikipedia.org/wiki/Latent_Dirichlet_allocation)

## Definition:

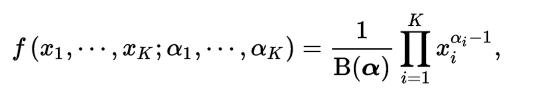
Latent Dirichlet allocation (LDA) is a  [generative statistical text model](https://en.wikipedia.org/wiki/Generative_model)[1] , that allows sets of observable variables to be explained by  Latent(unobservable ) variables , where in the texts , the words are the observables variables and topics are the latent variables. This algorithm aims to discovers the topics of the corpus or trying to find the related other corpus with this corpus.

## Algorithm Pre-processing:

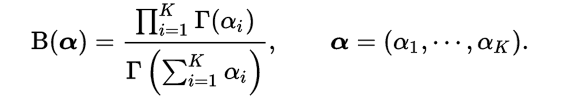
its performance depends on the vocabulary, so it is good to do parsing for all documents of the corpus and remove all words which cannot explore topic and can be found many times such as ( and , or ,.... etc) where this preprocessing will increase the performance of LDA.

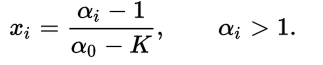
## Dirichlet distribution:

it is a [continuous](https://en.wikipedia.org/wiki/Continuous_probability_distribution) [multivariate](https://en.wikipedia.org/wiki/Multivariate_random_variable) [probability distributions](https://en.wikipedia.org/wiki/Probability_distribution) ,generalization for Beta distribution[2] but for multivariate . This distribution mainly used as prior distribution of categorical variable such as text model . Its density function follows the next density function.

****

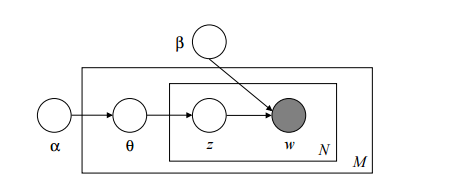
**X1,.... Xk Xi  (1,N)**  , α is called concentration parameters where αi >0 and **B(**α) is calculated using Gamma function as in the equation :

****

****

**X=(x1,x2,x3,..Xk) ~ Dir (**α)

## Graph [plate notation](https://en.wikipedia.org/wiki/Plate_notation) Model:



**Its graphical model illustrate LDA, where it has some assumption as in the following:**

M: is the number of documents , and LDA , it is estimated based on Bayes

K: is is number of the topic and LDA assumes it is given.

*α* is the parameter of the Dirichlet prior on the per-document topic distributions and *β* is the parameter of the Dirichlet prior on the per-topic word distribution, they has to be estimated to maximize the (marginal) log likelihood of the data using EM( Estimation maximization) algorithm. In a symmetric Dirichlet distribution, high alpha means that all your documents contain most topics (as opposed to documents containing a few or just a single topic). High beta means that all topics contain most of the words in the corpus. LDA works towards optimizing topic assignments based on these parameters through the iterative process.

ᶿi is the topic distribution for document i

the dimensionality k of the Dirichlet distribution is assumed to be known and fixed.

{\displaystyle \varphi \_{k}}

## Algorithm Implementation

## This algorithm consists of the following five steps:

**Step 1 Estimate (α, β)**

estimate their values as discussed in the previous section

**Step 1 define the number of the topics**

**Discover the number of the topics , it** can be given by an informed estimate (e.g. results from a previous analysis), or calculated through the trial-and-error.

In estimating or trying different values of the number of the topics , the number of the topics can be selected either when the level of interpretability is reached , or the one yielding the highest statistical certainty (i.e. log likelihood).

**Step 2:Intial assignment of the topic for each word**  
Assign in initial topic for each word to a initial topicbased on a prior Dirichlet distribution (α, β). This can lead to if a word appears more than one time , each instance of this word may be assigned to different topic.

Step 3 iterative per each word   
check and update topic assignment for each word,

* iterating through each word in every document.

For each word, its topic assignment can be changed based on

* + Finding the probability of occurrence each topic in the current document of the current word.
  + Finding the probability of occurrence each topic for this word in the current document and the others documents.
  + Weighing the values or multiplying these probabilities of each topic and the maximum multiplication is the new class for this word

Step 4 : Check on convergence of the words topics

check on if topics per words' are not converged go to step 3 , otherwise go to step 4

Step 5 : calculate the distribution matrix

It is (VxK) matrix where V is the number of the vocabularies of this corpus and K is the number of the topics for this corpus. Each cell wij  represents the number of assigning topic j for word i divided by ( V x K ) where the summation of this matrix cells has to give 1.

Summation of each row is P(occurrence word i in the corpus) . and summation of each column is P(occurrence topic j in the corpus).

Finally from this matrix , we can built this graph which represents the probability of occurrence each topic in the corpus. and it is found that for example Red topic has the highest probability and this can help to determine the main topic about the corpus.

## References:

## David M. Blei, Andrew Y. Ng, Michael I. Jordan " Latent Dirichlet Allocation" Journal of Machine Learning Research 993-1022

## *S. Kotz; N. Balakrishnan; N. L. Johnson (2000). Continuous Multivariate Distributions. Volume 1: Models and Applications. New York: Wiley.*[*ISBN*](https://en.wikipedia.org/wiki/International_Standard_Book_Number)[*0-471-18387-3*](https://en.wikipedia.org/wiki/Special:BookSources/0-471-18387-3)*.* (Chapter 49: Dirichlet and Inverted Dirichlet Distributions).