Latent Dirichlet Allocation

Table of content

1. What is topic models
2. What is LDA
3. Understand LDA
4. How LDA works
5. Who’s using LDA
6. Pros & Cons
7. Implementation
    a. Python
    b. R
8. Evaluation
9. Q&A

What is Topic models

“A form of text mining in order to identifying patterns in a corpus.”
“A method for finding and tracing topics (clusters of words) in large bodies of texts”

Topic models are algorithms for discovering the main themes that pervade a large and unstructured collection of documents. Topic models can organize the collection according to the discovered themes.

Primary usage - NLP

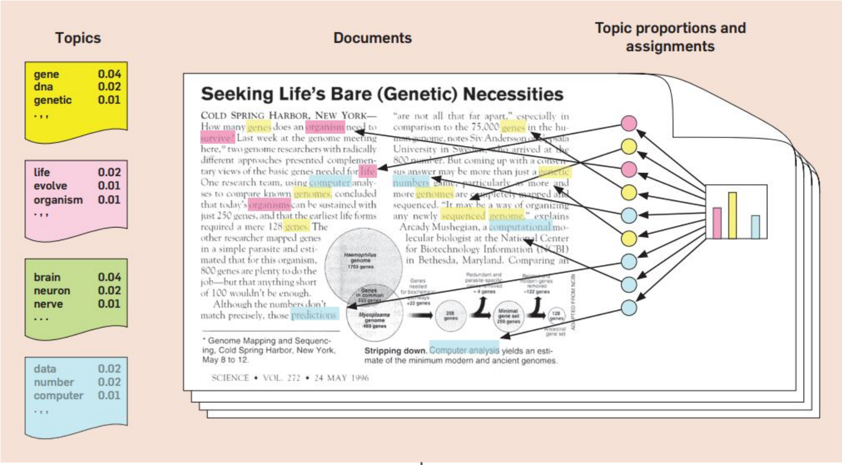
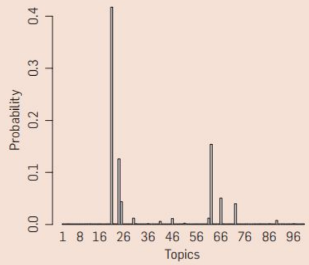
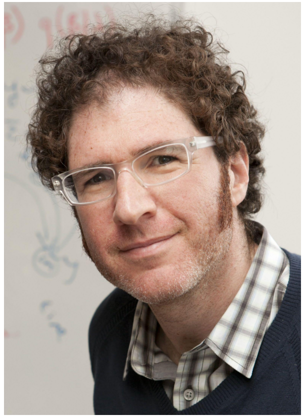

Topics
Documents
Topic proportions and assignments
gene 0.04
dna 0.02
genetic 0.01
life 0.02
evolve 0.01
organism 0.01
brain 0.04
neuron 0.02
nerve 0.01
data 0.02
number 0.02
computer 0.01
Seeking Life’s Bare (Genetic) Necessities
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.
SCIENCE • VOL. 272 • 24 MAY 1996

Figure 2. Real inference with LDA. We fit a 100-topic LDA model to 17,000 articles from the journal *Science*. At left are the inferred topic proportions for the example article in Figure 1. At right are the top 15 most frequent words from the most frequent topics found in this article.



“Genetics”	“Evolution”	“Disease”	“Computers”
human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

LDA - Latent Dirichlet Allocation



"Simplest topic model"

A three-level hierarchical Bayesian model

A generative probabilistic model for collections of discrete data such as text corpora

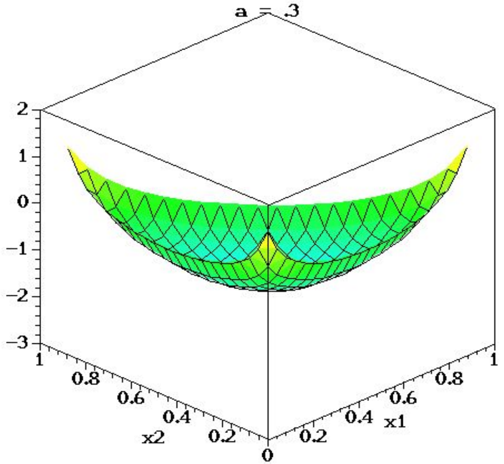
An unsupervised data mining algorithm mainly used for clustering purpose

Application:

- Used in Automatic Categorization of Software
- Bug Localization
- Face Recognition
- Video Fingerprinting
- Used in various search engines

Understand LDA

Latent topics

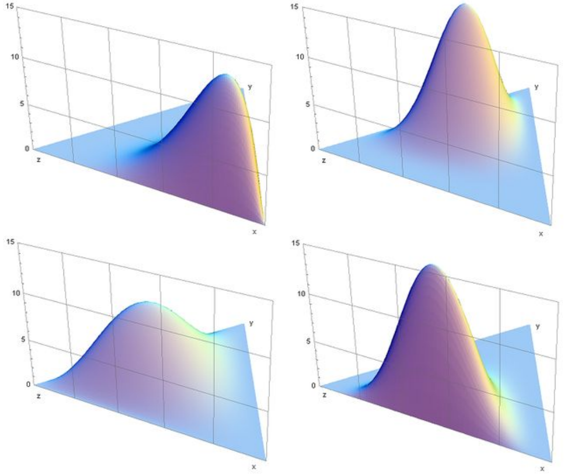


Dirichlet Distribution

$$f(x_1, x_2, \ldots, x_k; \alpha_1, \alpha_2, \ldots, \alpha_k) = \frac{1}{B(\alpha)} \prod_{i=1}^{k} x_i^{\alpha^i - 1}$$

where

$$B(\alpha) = \frac{\prod_{i=1}^{k} \Gamma(\alpha^i)}{\Gamma(\sum_{i=1}^{k} \alpha^i)}, \sum x_i = 1$$



Hypothesis: Bag of words (De Finetti Theorem)

Understand LDA

Generative model

5-steps to generate a word.

M: Number of documents.

N: Number of words.

K: Number of topics

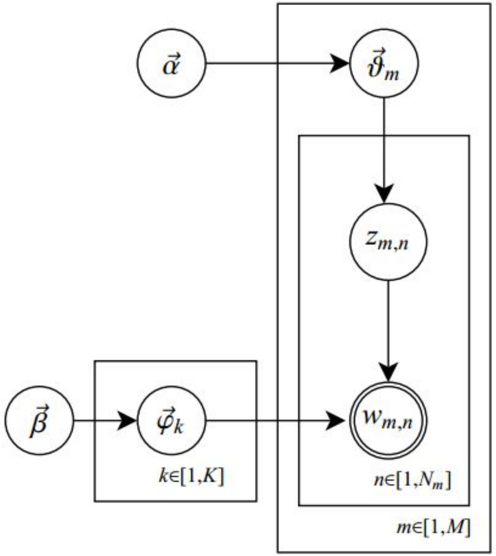
$\vec{\alpha}$ $\vec{\beta}$ Hyperparameters to generate probability distribution

$z_{m,n}$ Topic

$w_{m,n}$ $n\in[1,N_m]$ Word

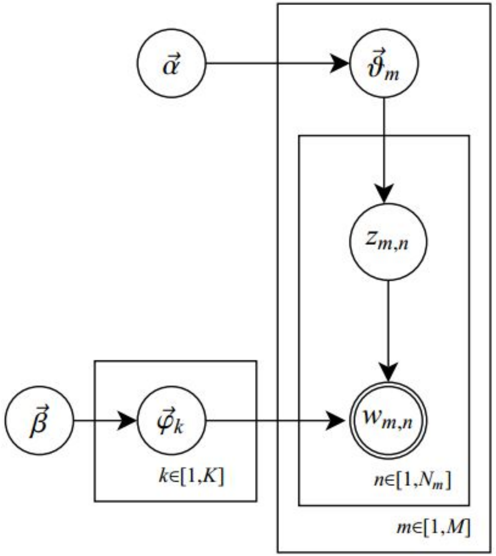
$\vec{\vartheta}_m$ Topic distribution of a document

$\vec{\varphi}_k$ $k\in[1,K]$ Word distribution of a topic



How LDA works

1. Pre-processing
2. Inference & Estimation
   a. EM v.s. Gibbs Sampling
   b. EM: the joint probability of of all latent variables
   c. Gibbs Sampling: the individual conditional probability of each latent variable
3. Experimental Results
   a. List of topics with their own group of words
   b. List of documents with each word’s topic.



Who are using LDA?

- The most common Topic Model currently in use is LDA.
- A Research Director at Google, Edward Y. Chang, is currently working on an implementation of LDA which utilizes a parallel computing architecture, allowing it to take advantage of the massive computing power Google has at its disposal
- Used in building New York Times Recommendation Engine
- By adding classified into six classes, including skin color of normal, cyan, red, yellow, black, and white and turn LDA into Supervised LDA, it helps to provides useful descriptive statistics for facial diagnosis in Traditional Chinese Medicine.

Pros & Cons

Pros:

1. LDA compared to LSI is a probabilistic model with interpretable topics that can be easily extended and embedded in other more complicated models
2. LDA is generalizable to unseen documents. The other(pLSA) assume the distribution of topics/words are following the same pattern in all documents.

Cons:

1. One of the drawback of LDA is that the algorithm fails to draw the relationship from one topic to another
2. The technique uses several tuning parameters whose impact on the resulting LDA model are not always intuitive. Values for the tuning parameters is dependent on both the problem that is being addressed and the input corpus

Evaluation

Evaluation for unsupervised algorithms like LDA is difficult.

HUMAN-IN-THE-LOOP

Word Intrusion: For each trained topic, take first ten words, substitute one of them with another randomly chosen word(intruder) and see whether a human can reliably tell which one of word is the intruder. If so then the topic is Coherent(good), if not the topic has no discernible theme(bad).

Topic Intrusion: Subjects are shown a title and snippet from a document. Along with document they are presented with four topics. Three of them are high probability topics assigned to that document. The remaining intruder topic is chosen from other low probability topics in the model.

The Mechanical Turk can be used for this purpose.