1. What is topic models

* Topic models are algorithms for discovering the main themes that creates a large and unstructured collection of documents.
* By applying topic models into a collection of documents, the hidden value beneath every word can be discovered.
* The primary usage of topic models is natural language processing.

Example:

* Science magazine uses topic model (LDA) to extract 100 topics among 17,000 papers. We can see the example of four topics here.

2. What is LDA

* A generative probabilistic model for collections of discrete data such as text
* An unsupervised data mining algorithm mainly usecorporad for clustering purpose

* LDA is a three-level hierarchical Bayesian model, in which

Each document of a text corpora is modeled as a finite mixture over some topics

Each topic is, modeled as an infinite mixture over an set of topic probabilities.

* The topic probabilities provide an explicit representation of a document.

3. Understand LDA

* What is Latent topic - Hidden value of a single unit of measurement. In most case, of a word.
* What is De Finetti Theorem/ Bag of words hypothesis: words in a document are infinitely exchangeable, then the joint probability of a specific topic is represented by a mixture of words.
* What is Dirichlet Distribution - the collection/distribution of probability distributions.

In probability and statistics, the Dirichlet distribution is a family of continuous multivariate probability distributions.

A specific distribution is decided by a vector alfa.

4. Understand LDA - Generative model

based on Latent topic, bags of words and Dirichlet Distribution

Documentations can be created using the **generative model.**

5 steps. 1. Alpha -- topic distribution theta in a document

2. Sample theta get topic Zm,n

3. Beta -- word distribution in topic Zm,n fie

4. Sample fie to get the word

5. Repeat for each word of a doc, also for all documents.

5. How LDA Works

* With known words. The whole process is to infer and estimate every word's hidden topic & the distribution of topics in each document.
* Modeling process

1. preprocessing

2. EM & Gibbs Sampling.

3. Present result

2 formats: list of topics with their own group of words

list of documents with each word's hidden topic