TOPIC MODELING

INTRODUCTION OF TOPIC MODELING :

Topic modeling is widely used when you have large amounts of text data and you want to label them based on the type of information they carry. This is mostly seen in customer reviews for products on the various e-commerce platforms, movie reviews, recommendations of popular places, and so on.

Similarly, imagine that you have a large corpus of scientific documents (such as research papers), and you want to build a search engine for this corpus. Suppose you can infer that a particular paper talks about “topics” such as diabetes, cardiac arrest, and obesity. With this information, conducting a topic-specific search will become easier.

TOPIC VS WORD : A topic is a group of words that define an idea.Words are observable, while topics are not. They are latent.

This exercise aimed to impart the same thinking to a machine learning algorithm. This domain of computer science is known as “information retrieval” and involves the identification of dominant themes from a sample of text.

Suppose various pieces of text are being sent from different sources. These are unstructured and unlabeled. Topic modeling can help determine the major themes or labels of this text.

An ideal example of this is customer review data. For example, reviews for a restaurant can vary according to food items, length, sentiment, etc. The job of a topic modeling algorithm is to skim through each review and figure out the major themes and keywords.

If a customer visits the site, it becomes easier for them to find relevant reviews. Suppose you are on a budget and want to know if a certain dish is affordable. “Worth the Money” would be the ideal label in this case. Or, suppose you are interested in ordering the “Fiery Paneer Tikka Wrap”; however, you are not sure if it will be any good. You can quickly figure this out by clicking on the corresponding food label (i.e., summarized based on the identified topics).

INTUTION OF TOPIC MODELING :

A topic model is an algorithm that automatically learns latent topics (themes) from a collection of documents.

It works by observing words that tend to co-appear in documents. It learns topics directly from the text and assumes that each document exhibits multiple topics.

TOPIC 1	MODEL,LANGUAGE
TOPIC 2	HAPPY

Suppose Example :

NLP is text language . I very excited to learn nlp.

The word “NLP”,”text” belongs to topic1 and the words “excited” belongs to topic2.

A document/sentence can discuss many topics represented in terms of their probabilities as shown below.

	Topic1	Topic2	Topic3
Document	0.5	0.5	0

A word can appear across many topics in terms of probability. The weights denote the importance of a word in a topic.

	model	language	text	Happiness	satisfaction	priority	Stoic	virtue	radical
Topic 1	0.33	0.33	0.33	0	0	0	0	0	0
Topic 2	0	0	0	0.33	0.33	0.33	0	0	0
Topic 3	0	0	0	0	0	0	0.33	0.33	0.33

Application of Topic Modeling in Business :

The applications of topic modeling include:

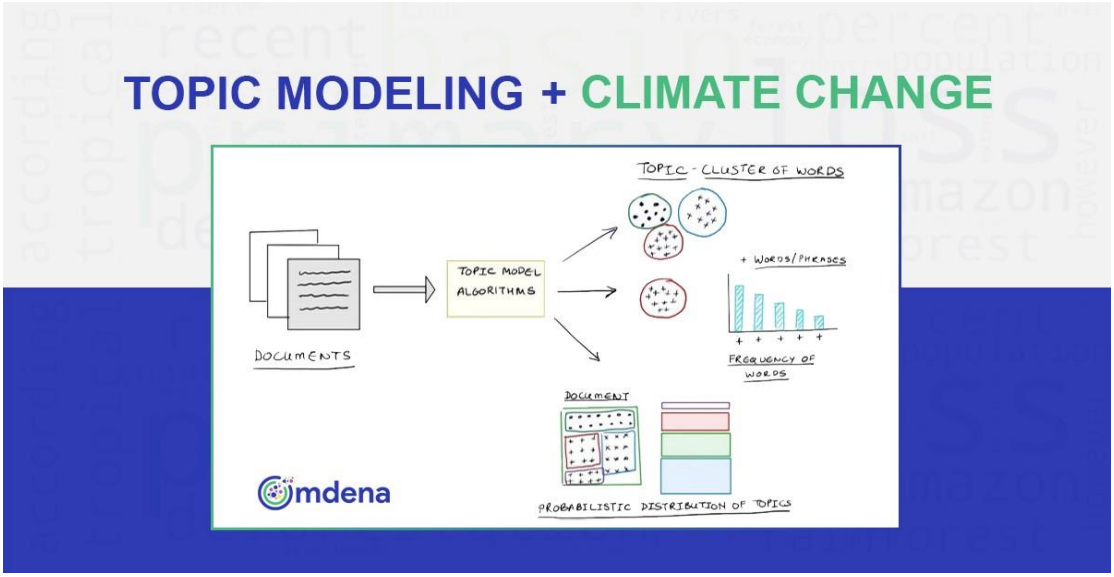
- Automatically tagging customer support tickets;
- Labeling customer reviews for a product;
- Main topics in customer survey responses;
- Twitter trending hashtags and topics
- Automatic labeling of different news articles to categories like food, travel, etc.

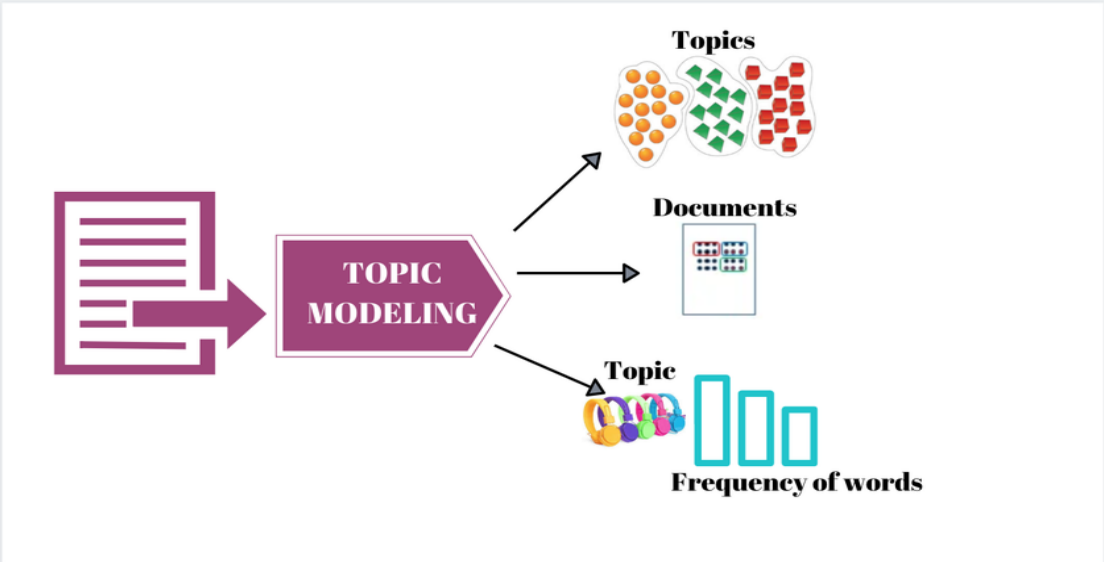
INPUT AND OUTPUT OF TOPIC MODELING :

A Set of documents are fed as input to topic models.

There are three types of output to a topic model :

- **Topic-Word-Distribution :** It is the probability associated with each unique word in every topic.
- **Document-topic-distribution** : It is the vector of topic proportions (i.e., probability) associated with each document.
- **Topic-word-assignment** : Every word in each document is assigned to a certain topic.





There are several algorithms for topic modeling as listed below:

- Latent semantic analysis (LSA)
- Non-negative matrix factorization
- Latent Dirichlet allocation (LDA)

Latent Dirichlet Allocation :

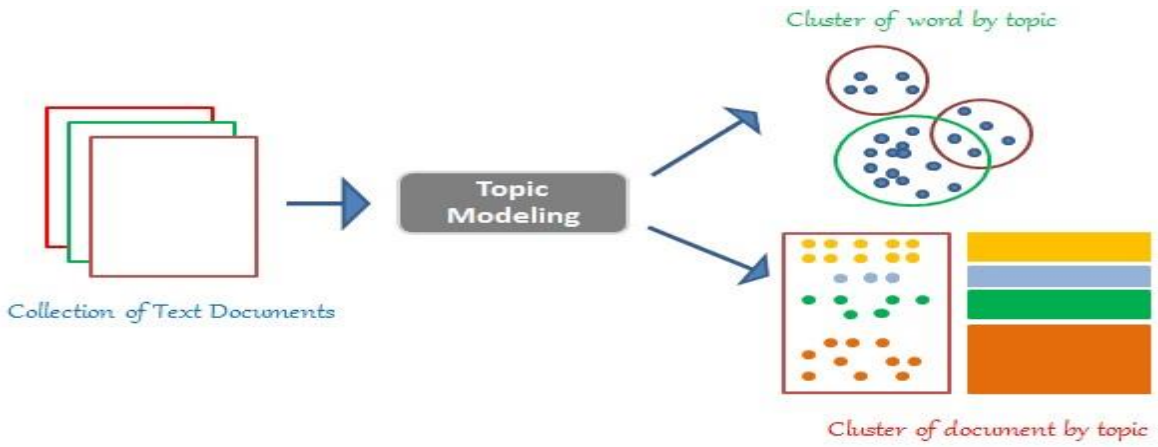
LDA is one of the algorithms used for topic modeling.

LDA is based on the Dirichlet probability distribution named after the German mathematician Johann Peter Gustav Lejeune Dirichlet.

The word “Dirichlet” means “distribution over distributions,” and in the context of topic modeling, Dirichlet is:

- The distribution of topics in documents
- The distribution of words in the topic.

The objective is to infer these two distributions using LDA. For a set of given documents, it outputs the assignment of words to different topics, topic-word probability distributions, and document-topic probability distributions.



Assumption Of LDA :

- Documents with similar topics use similar groups of words.
- Latent topics can then be found by searching for groups of words that frequently occur together in documents across the corpus.
- Documents are independent of each other.

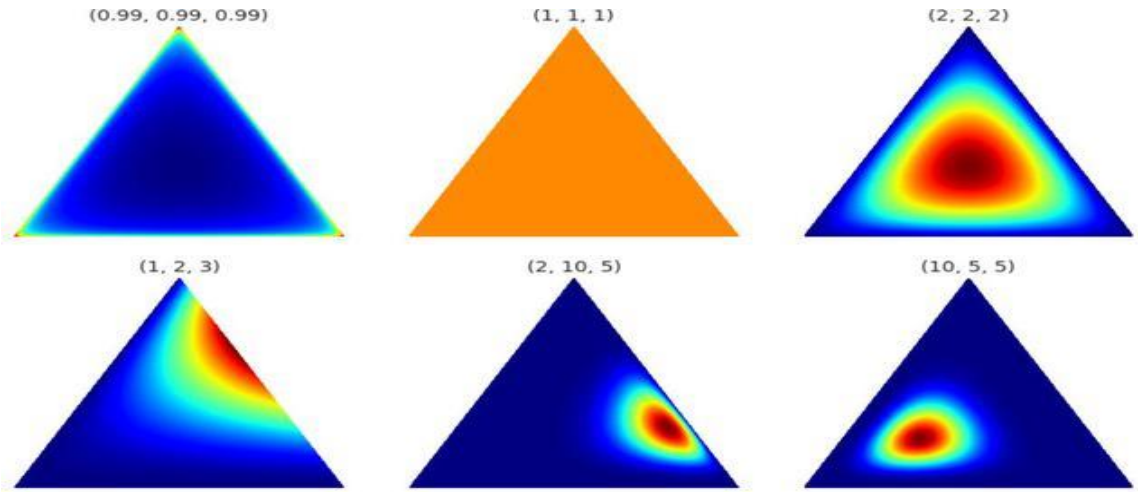
Steps for Building LDA :

- Assume that there is a set of documents.
- Choose a fixed number of K topics to discover and use LDA to learn the topic representation of each document and the words associated with each topic.
- Go through each document and randomly assign each word in the document to one of the K topics. This random assignment already gives you both topic representations of all the documents and word distributions of all the topics.
- Iterate over every word in every document to improve these topics. For every word in every document and for each topic t, you calculate:
- p(topic t | document d) = The proportion of words in document d that are currently assigned to topic t
- Next, iterate over every word in every document to improve these topics. For every word in every document and for each topic t, you calculate
- p(word w | topic t) = The proportion of assignments to topic t over all documents that come from this word w
- Reassign w a new topic, where topic t is chosen with probability p(topic t | document d) * p(words w | topic t). This is essentially the probability that topic t generated word w in document d.
- After repeating the previous step over a large number of iterations, a roughly steady state is attained where the assignments are acceptable.
- In the end, each document is assigned to one dominant topic (even other K-1 topics exist).

Hyperparameters In LDA :

Alpha is a parameter of the Dirichlet distribution that determines the document-topic distribution. Alpha is a corpus-level parameter that is chosen once. The amount of smoothing is determined by alpha

Higher alpha implies more smoothing and less distinct topics, whereas low alpha implies less smoothing and highly distinguishable topics. The recommended value of alpha is 50/T (or less if T is very small).



β is the parameter that determines the topic-word distribution. β is also a corpus-level parameter that is chosen once and determines the amount of smoothing. Higher β implies more smoothing. The recommended value of β is 0.01.

Evalution on LDA :

Perplexity measures the modeling power by calculating the inverse probability of unobserved documents. Better models have lower perplexity, suggesting fewer uncertainties about the unobserved document, as well as the better generalizability. It can be calculated as follows: