**BACKGROUND OF LDA**

* LDA is an unsupervised statistical method of extracting topics, concepts and other types of meaning from unstructured data. It doesn’t understand syntax or any other aspect of human language. It’s looking for patterns, and it does that equally well no matter what language text is in, or even if it consists of just symbols rather than characters.
* Initially LDA was developed for text analysis and population genetics, LDA has since been extended and used in applications from time series to image analysis

**History**:

* LDA, the most common topic model currently in use developed by David Blei, Andrew Ng and Michael I. Jordan. LDA is a generalization of pLSI (Probabilistic Latent Semantic Indexing) which has been evolved from LSA (Latent Semantic Analysis) which is also known as distributional semantics.
* LSA :- Applications for LSA are data clustering, document classification, cross language retrieval, synonymy and polysemy, information retrieval, etc.

**Data type for LDA**:

* The document used for LDA algorithm should be of document term matrix for understanding how frequently each term occurs within each document
* It is kind of lookup grid which shows which term appears in which documents and how many times
* More sophisticated weights can be used, for example tf-idf scheme which calculates the weight of the term in the documents
* The document-term matrix helps in improving search results and finding topics
* LDA can process structured, semi-structured and unstructured data from any number of sources
* Various methods used for Topic Modelling are LDA, Expectation-Maximization, Gibbs Sampling, Variational approaches, document matrix and method of moments
* LDA model is Bayesian version of pLSA model
* LDA model can perform better
* The goal/primary usage of LDA is to find a way to characterize each document using short descriptions that can preserve essential statistical relationships of the document
* 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.
* **LDA**:- In the original Latent Dirichlet Allocation (LDA) model, an unsupervised, statistical approach is proposed for modeling text corpora by discovering latent semantic topics in large collections of text documents. The key insight into LDA is the premise that words contain strong semantic information about the document. Therefore, it is reasonable to assume that documents on roughly similar topics will use the same group of words. Latent topics are thus discovered by identifying groups of words in the corpus that frequently occur together within documents.
* **Q. Why LDA is unsupervised model?** Learning in this model is unsupervised because the input data is incomplete: the corpus provides only the words within documents; there is no training set with topic or subject annotations.

* **Background**:- **a**. Before LDA, a number of methods, both probabilistic and non-probabilistic, have been proposed for the task of text document modeling. The goal of these methods remain the same across both categories: to find a way to characterize each document using short descriptions that can preserve essential statistical relationships. This is necessary for basic tasks such as document classification, summarization, and similarity and relevance judgment.

**b**. Non-probabilistic methods mainly use word counts as features. In the popular td-idf scheme, each document in the corpus is reduced to a fixed-length list of numbers, roughly corresponding to the frequency of appearance for a basic set of words within that document. Later, the LSI model uses the singular value decomposition (SVD) of the word-by-document matrix from td-idf to identify a subset of the feature space that captures the most variance.

**c**. Probabilistic models improved on these previous models. They mainly fall under the category of latent variable models, and introduce the idea of topics. The mixture of unigrams model assumes that all 2 words from a document are drawn independently from a single topic, where each topic is a random variable that has its own particular distribution over words. Probabilist[[1]](#footnote-0)ic latent semantic indexing (pLSI), which allows each document to exhibit multiple topics, followed naturally. In pLSI, each word of a training document is generated from a different, randomly chosen topic, where topics are drawn from a document specific distribution over latent topics shared by the whole corpus. This document-specific distribution, however, had many drawbacks as it allowed no natural way to assign probabilities to previously unseen documents.

**Notation & Terminology**:

1. A word w ∈ {1, . . . , V } is the most basic unit of discrete data. For cleaner notation, w is a V - dimensional unit-based vector. If w takes on the ith element in the vocabulary, then w i = 1 and w j = 0 for all j 6= i.
2. A document is a sequence of N words denoted by w = (w1, w2, . . . , wN ), where wn is the nth word in the sequence.
3. A corpus is a collection of M documents denoted by D = {w1, w2, . . . , wM}.
4. A topic z ∈ {1, . . . , K} is a probability distribution over the vocabulary of V words. Topics model particular groups of words that frequently occur together in documents, and thus can be interpreted as “subjects.”. The generative process imagines that each word within a document is generated by its own topic, and so z = (z1, z2, . . . , zN ) denotes the sequence of topics across all words in a document.

**Advantages and Disadvantages:**

* One disadvantage of LDA is that it does not derive any interrelationship between the extracted topics nor identify topics at various level of granularity

**Visualizations that can be done in Topic Modeling**

* Stacked Bar Chart
* Heat Map

1. **Quoting and Paraphrasing:** University of California, San Diego:- http://cseweb.ucsd.edu/~dhu/docs/research\_exam09.pdf [↑](#footnote-ref-0)