**TOPIC MODEL WORKING**

It involves counting words and grouping similar word patterns to infer topics within unstructured data. Topic modelling refers to the process of diving a corpus of documents in two:

1. A list of the topics covered by the documents in the corpus

2. Several sets of documents from the corpus grouped by the topics they cover.

It try to figure out which topics are present in the documents of the corpus and how strong that presence is.

**TWO TOPIC MODELLING METHODS**

**1. LATENT SEMANTIC ANALYSIS (LSA)**

**2. LATENT DIRICHLET ALLOCATION(LDA)**

LSA: It is based on distributional hypothesis which means the semantic of words can be grasped by looking at the contexts the words appear in . OR in other words , semantics of two words will be similar if they tend to occur in similar contexts.

It computes how frequently words occur in the documents and the whole corpus and assumes that similar documents will contain approximately the same distribution of word frequencies for certain words. The syntactic information (example: word order) and semantic information (example: the multiplicity of meanings of a given word) are ignored and each document is treated as a bag of words. The standard method for computing word frequency is what is known as tf-idf which computes frequencies by taking into consideration not only how frequent words are in a given document, but also how frequent words are in all the corpus documents.

Words with higher frequency in the full corpus will be better candidates for document representations than less frequent words , regardless of how many times they are appear in individual documents. Tf-idf value are much better than those that only take into consideration word frequencies at document level.

Once it has been computed , we can create a Document- term matrix which shows the tf-idf value for each term in a given document . This matrix will have rows for every document in the corpus and columns for every term considered.

LDA: It is based on distribution hypothesis and the statistical mixture hypothesis for which a statistical distribution can be determined. The purpose of LDA is mapping each document in our corpus to set of topics which cover a good deal of the words in the document.It assigns topics to arrangement of words. It also assumes that all words in the document can be assigned a probability of belonging to a topic. The goal of LDA is to determine the mixture of topics that a document contains.