I have implemented **Latent Dirichlet Allocation** algorithm for topic modeling have converted parsed the JSON file and converted to a data frame with<id,tex> columns. Then I performed basic data preprocessing like removing stop words, tokenization, lemmatization and created a dictionary and mag of words model for the corpus and ran an LDA algorithm to get 5 topics for each text in the dataset

LDA algorithm is an unsupervised algorithm that assigns topics to given text. It converts the available **Documents->Topic->Words.** If the topics chosen is less than the length of the documents, it’s a form of dimensionality reduction where we are extracting only the important word or apply any algorithm on features obtained. also, it can be considered as a kind of k-means clustering with the clusters as the number of topics. I also hyper tuned the parameters like alpha and beta and the number of passes.

At the end the output is a data frame with the <id,topics> in the decreasing order of weights and given new document it can classify the topics.

**Sample Output(output size:(1837,2))**

