**TOPIC MODELLING (LDA):**

It is an unsupervised learning technique to extract key topics that occur in a collection of documents.

The data set used is the 20Newsgroup data set. It is available under sklearn data sets.

The first method used LDA (Latent Dirichlet Allocation).

It is used to classify text in a document topics/categories/themes.

It builds a topic per document model and words per topic model as dirichlet distributions.

Steps for LDA

1) Preprocessing raw data

2) It involves tokenization,stopwords removal,lemmatization and stemming

3)Filtering out Words that have fewer than 3 characters

4)Converting the preprocessed text to bag of words i.e for each document we create a dictionary reporting the frequency of occurrence of words in it.In our case we have used doc2bow for converting text to bag of words.

5)For running LDA,we can use the gensim model.We need to set some parameters for the same.

6)We need to specify how many topics are there .In our case we have set it as 8.Similarly,with no. of passes,i.e no. of training iterations over the document.In our case we have set it to 10.

7)Being an unsupervised model,LDA doesn’t give a topic name and the interpretability part needs human intervention.

8)Approaches that can be done to enhance LDA can be by the use of something called Guided LDA.