**TOPIC MODELLING (Using SVD and NMF): TMsvdnmf.ipynb**

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

The two methods used here are SVD and NMF

**1st method –SVD (SINGULAR VALUE DECOMPOSITION)**

SVD is an unsupervised learning technique to extract key topics that

occur in a collection of documents.

This starts with Data Preprocessing of the raw data.

It involves tokenization,stopwords removal,lemmatization and stemming

We set the number of topics to 6 and number of top words to 8.

The SVD algorithm factorizes a matrix into three different matrices-

S,U and Vh,where

‘S’- represents a singular diagonal matrix

‘U’- represents matrix with orthogonal columns

‘Vh’- represnts matrix with orthogonal columns

The diagonal matrix ‘S’ contains the relative importance of each factor.

It allows us to exactly factor a matrix into orthogonal columns and orthogonal rows thereby grouping the documents into required themes/categories/topics.

For word vectorization we are using both count vectorizer and tfidf as the word vectorization methods.

“Show topics” is a customised function which goes through the vocab and returns to us the largest values of the top words for each topic.

Approaches that can be done to enhance SVD can be by the use of something called Truncated SVD.It is faster and brings with it less computational complexity.

**2nd method – NMF(Non-Negative Matrix Factorization)**

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

This starts with Data Preprocessing of the raw data.

It involves tokenization,stopwords removal,lemmatization and stemming

As it’s name suggests it is a factorization technique of a non-negative dataset into non-negative matrices .

This takes into consideration the idea ,rather than constraining factors to be othogonal we can constrain them to be non-negative.

The major positive of NMF over LDA and SVD is it’s positive factors which make the model highly interpretable.This ease of interpretability makes NMF stand out.

For word vectorization we are using both count vectorizer and tfidf as the word vectorization methods.

Here also we are giving no. of topics required by specifying value od ‘d’,in our case it being 5.

We use tf-idf to normalize term counts by taking into account how often they appear and how common or rare they are.

“Show topics” is a customised function which goes through the vocab and returns to us the largest values of the top words for each topic.

NMF is faster ,easier and most importantly it is highly interpretable due to the absence of negative factors.