VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

Topic Modelling using LDA, NMF and BERT and Article Recommendation using TF-IDF

Natural Language Processing
CSE4022

Siddharth Mandal: 20BDS0157
Rishabh Ajay Agarwal – 20BCE0372
Aditi Parashar – 20BCE0426
Devavrat Kaustubh Dubale – 20BCE0660

Slot: E1+TE1
Faculty: Dr. Rajesh Kannan R

Abstract

Analysing the contents and the topics covered in given articles and documents is a very important part of our daily lives. But going through tonnes of documents manually to understand the topic they cover and labelling them can be a tiresome job. Topic Modelling can help solve this problem as it is used for recognizing the words from the topics present in the document or the corpus of data. This is useful because extracting the words from a document takes more time and is much more complex than extracting them from topics present in the document. Topic Modelling is an unsupervised technique of recognizing or extracting topics by identifying the patterns like clustering algorithms which divide the data into different parts. This is done by finding out the patterns of word clusters and frequencies of words in the documents.

In our project we will be using LDA, NMF and BERT which will be using input from TF-IDF model which will contain the TF-IDF scores which will indicate the importance of the terms in the document with respect to the corpus or in other words which are three different topic models which can be used to classify text in a model to a particular topic. Finally we will create an article recommendation engine where after giving a keyword, the engine would suggest the best documents from the pool of documents.

LDA(Latent Dirichlet Allocation) is an example of topic model and is used to classify text in a document to a particular topic. It builds a topic per model and words per document model, which are modelled as Dirichlet distributions. Non-Negative Matrix Factorization (NMF) is an unsupervised technique so there are no labeling of topics that the model will be trained on. The way it works is that, NMF decomposes (or factorizes) high-dimensional vectors into a lower-dimensional representation. These lower-dimensional vectors are non-negative which also means their coefficients are non-negative. BERTopic is the last topic modelling technique we will use that uses BERT embeddings and c-TF-IDF to create easily interpretable topics whilst keeping the important words in topic description. TF-IDF is one of the best metrics to determine how significant a term is to a text in a series or a corpus. TF-IDF is a weighting system that assigns a weight to each word in a document based on its term frequency (tf) and the reciprocal document frequency (idf).

We will be using the three topic modelling techniques because of their various advantages and disadvantages, for example, the number of topics should be known beforehand in LDA and NMF, whereas every document gets assigned to a variety of topics in LDA and NMF and BERT assigns every document only one topic. We can visualise our topics in LDA and BERT whereas not in NMF.