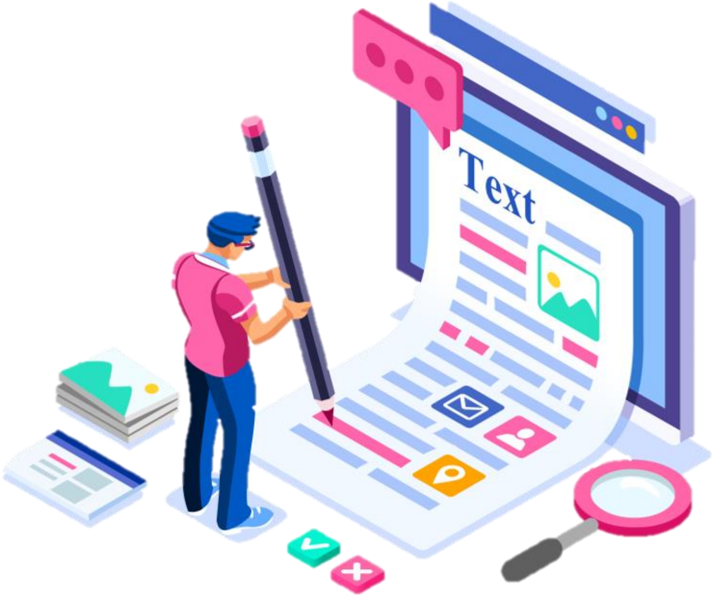
Siddhardhan

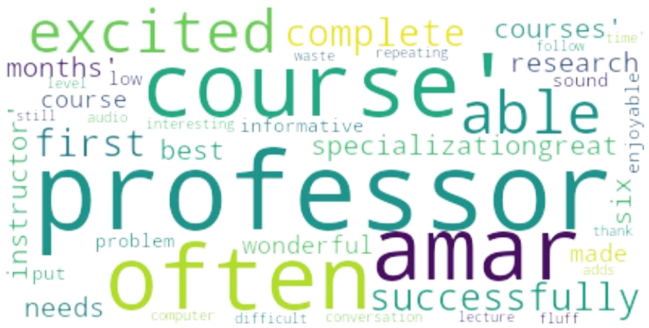
Feature Extraction Of Text Data: Tf-idf Vectorizer



Feature Extraction

The mapping from textual data to real valued vectors is called feature extraction.

Bag Of Words (BOW): list of unique words in the text corpus



Term Frequency-Inverse Document Frequency (TF-IDF):
To count the number of times each word appears in a document

Tf-idf Vectorizer

Term Frequency (TF) = (Number of times term t appears in a document)/(Number of terms in the document)

Inverse Document Frequency (IDF) **=** log(N/n), where, N is the number of documents and n is the number of documents a term t has appeared in.

The IDF value of a rare word is high, whereas the IDF of a frequent word is low.

TF-IDF value of a term = TF x IDF