TF-IDF

TF-IDF is a numerical statistics that reflects the importance of a word is document.
It is a powerful technique to identify the most important words in a document.
By assigning weights to word based on their frequency and rarity , we can extract meaningful information from unstructured text data.

TF : Frequency of word in a document.
TF(t,d) = number of occurrences of word(t) in document(d)/ total number of terms in document(d)

IDF : Frequency of word across all documents.
IDF(t,D) = ln(total number of documents in corpus(D)/number of documents in corpus(D) containing the term(t))

TFIDF(t,D) = TF(t,d)*IDF(t,D)

Advantages:

1. Measures Relevance: TF and IDF together helps to identify which terms are relevant in the document.
2. Handle large text corpus: TF IDF is scalable and can handle large text corpus.
3. Handle stop-words: TF-IDF automatically down-weights the common word that occurs frequently in the document.
4. Can be used for various applications: Text classification, Information retrieval , document clustering.
5. Interpretable: Scores generated are easy to interpret.

Disadvantages:

1. Ignores the context: Does not take into the context in which the term appears.
2. Assumes Independence: Assumes that the terms in the document are independent of each other.
3. High Dimensionality: The vocabulary size can become very large when working with large datasets which can lead to high dimensional features.
4. No concept of word order:
5. Limited to frequency.

Applications OF TF-IDF

1. Search Engines: Used to rank the documents.
2. Text Classification: To identify most important feature in the document.
3. Information/Keyword extraction : To extract most important information/keyword in document.