What is TF-IDF and how is it used in NLP?

TF-IDF (Term Frequency-Inverse Document Frequency) is a numerical statistic used in Natural Language Processing (NLP) to evaluate the importance of a word in a document relative to a collection of documents.

Term Frequency

TF measures the frequency of a term within a document. It's calculated by dividing the number of times a term appears in a document by the total number of terms in the document.

TF helps in identifying the significance of a term within a specific document. A higher TF score indicates that the term is more relevant to that document.

TF(𝑡,𝑑)= Number of times term 𝑡 appears in document 𝑑

Total number of terms in document 𝑑TF(*t*,*d)*

Inverse Document Frequency

IDF measures the importance of a term across the entire corpus. It's calculated by taking the logarithm of the ratio of the total number of documents in the corpus to the number of documents containing the term.

IDF(t,D)=log (Total number of documents in corpus *D*

Number of documents containing the term)

TF-IDF Score

TF-IDF is computed by multiplying the TF and IDF values for each term in a document.

The TF-IDF score quantifies the importance of a term in a document relative to its importance across the corpus. Terms with higher TF-IDF scores are more significant to the document.

Significance in NLP

Information Retrieval

TF-IDF is widely used in search engines to rank documents based on their relevance to a user query. Documents with higher TF-IDF scores for the query terms are considered more relevant. It helps in retrieving documents that are most likely to satisfy the user's information needs.

Text Classification

TF-IDF is used in text classification tasks such as sentiment analysis, spam detection, or topic classification

Keyword Extraction

TF-IDF is used to extract keywords from documents by identifying terms with the highest TF-IDF scores.

Example:

Term Frequency

Consider a document containing 100 words wherein the word apple appears 5 times. The term frequency (i.e., TF) for apple is then (5 / 100) = 0.05.

Inverse Document Frequency

Assume we have 10 thousand documents, and the word apple appears in one 100 of these. Then, the inverse document frequency (i.e., IDF) is calculated as log(10,000 / 100) = 2.

TF-IDF weight is the product of these quantities: 0.05 \* 2 = 0.10