What is TF-IDF 10 marks

1. TFIDF or Term Frequency-Inverse Document Frequency indicates the importance of a word in a set. It helps in information retrieval with numerical statistics. For a specific document, TF-IDF shows a frequency that helps identify the keywords in a document. The major use of TF-IDF in NLP is the extraction of useful information from crucial documents by statistical data. It is ideally used to classify and summarize the text in documents and filter out stop words.

TF helps calculate the ratio of the frequency of a term in a document and the total number of terms. Whereas, IDF denotes the importance of the term in a document.

The formula for calculating TF-IDF:

TF(W) = (Frequency of W in a document)/(The total number of terms in the document)

IDF(W) = log\_e(The total number of documents/The number of documents having the term W)

When TF\*IDF is high, the frequency of the term is less and vice versa.

Google uses TF-IDF to decide the index of search results according to the relevancy of pages. The design of the TF-IDF algorithm helps optimize the search results in Google. It helps quality content rank up in search results.

What are unigrams, bigrams, trigrams, and n-grams in NLP 10 marks

1. When we parse a sentence one word at a time, then it is called a unigram. The sentence parsed two words at a time is a bigram.

When the sentence is parsed three words at a time, then it is a trigram. Similarly, n-gram refers to the parsing of n words at a time.



Code

1. Load the date from !kaggle datasets download -d kazanova/sentiment140
   1. LOAD
   2. CLEANING
   3. Prepare the model

Soln : .<https://colab.research.google.com/github/mborkhat/NLP/blob/main/MTech_AI_NLP_Lec_3_18_08_2024_Sentiment.ipynb#scrollTo=AiTawj9XEKOq>

2.

| Load the data from <https://www.kaggle.com/datasets/selener/consumer-complaint-database>   * 1. LOAD   2. CLEANING   3. Prepare the model |
| --- |

Son:<https://colab.research.google.com/drive/1JkqgW4AJShePsjtVmj283CQxMbcLJW-O>