SOC: Navigating the Waters of AI
TF-IDF Score

Ashutosh Gandhe

July 31, 2023

1 TF-IDF

It stands for Term Frequency- Inverse Document Frequency. This term is used in natural language processing to asses the importance of a certain word in a particular document.

2 Term Frequency

This quantity measures how frequently a term or word appears in a document. Formula for TF is as follows:
$TF(term, document) = \frac{Number\ of\ occurences\ of\ the\ term\ in\ the\ document}{Total\ number\ of\ terms\ in\ the\ document}$
TF value is higher for more frequently occurring terms in the document.

3 Inverse Document Frequency

This quantity is a measure of the importance of a term in the entire collection of documents. It gives more weight to rare terms and less weight to common terms. Formula for IDF is as follows:
$IDF(term) = \log \frac{Total\ number\ of\ documents}{Number\ of\ documents\ containing\ the\ term}$

4 TF-IDF Score

$TF - IDF(term, document) = TF(term, document) \cdot IDF(term)$
This TF-IDF score indicates the importance of a specific word related to the entire corpus. If a word has a high TF-IDF score for a particular document, it means the word is both frequently occurring in that document and relatively rare across other documents in the collection.

5 Example

Consider the following 3 documents.
Document 1: She sells sea shells on the sea shore.
Document 2: The sea is very calm.
Document 3:She sells shells.
Total number of documents $N = 3$. Now we will calculate the TF-IDF scores of the words sea and shells.

$TF(sea, Document1) = \frac{2}{8} = 0.25$
$TF(sea, Document2) = \frac{1}{5} = 0.2$
$TF(sea, Document3) = 0$
$IDF(sea) = \log(\frac{3}{2}) = 0.405$

$TF - IDF(sea, Document1) = \frac{0.25}{0.405} = 0.617$
$TF - IDF(sea, Document2) = \frac{0.2}{0.405} = 0.493$
$TF - IDF(sea, Document1) = 0$

Inferences: The word has a greater score in document1, so it is of greater importance in it, whereas it is not considered important in document 3, where it does not appear.

6 Working of the model

The given final model gives answer to the queries in following rough steps:

- Process the documents and make a list of all the words present in it.

- Remove stopwords (like is, if, but, and etc)and use tokenization to standardize the words.
- Calculate the TF-IDF scores for each word.
- Processing the user queries by removing stopwords and finding TF-IDF score of remaining words.
- Comparing TF-IDF scores of each particular word in the prompt and those in the document, and sort the documents according to relevance with the given query.
- The top ranked document has the key words for the answer, sentences are generated to give the final answer to the user.