Introduction

TF-IDF stands for Term Frequency-Inverse Document Frequency. It is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. It is calculated by multiplying two statistics: term frequency and inverse document frequency.

Term Frequency (TF): gives us the frequency of the word in each document in the corpus. It is the ratio of the number of times the word appears in a document compared to the total number of words in that document. It increases as the number of occurrences of that word within the document increases. Each document has its own tf.

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{i,j}}$$

Inverse Data

Frequency (idf): used to calculate the weight of rare words across all documents in the corpus. The words that occur rarely in the corpus have a high IDF score. It is given by the equation below.

Combining these two we come up with the TF-IDF score (w) for a word in a document in the corpus. It is the product of tf and idf:

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

$tf_{i,j}$ = number of occurrences of i in j
df_i = number of documents containing i
N = total number of documents

Real-life Example:

If We have a search engine and somebody looks for “Coke”. The search engine will return all documents containing the word “Coke”. However, some documents may contain the word “Coke” more frequently than others. In this case, TF-IDF can be used to figure out if a page titled “COKE” is about: a) Coca-Cola. b) Cocaine. c) A solid, carbon-rich residue derived from the distillation of crude oil. d) A county in Texas .

Mathematical Simulation:

There are two documents in a corpus: Text A and Text B. We will use them to create a TF-IDF matrix.

Text A: "The quick brown fox jumps over the lazy dog"
Text B: "The dog is lazy and the fox is quick"

The table below shows the values of TF for A and B, IDF, and TFIDF for A and B.

Words	TF (A)	TF (B)	IDF	TFIDF (A)	TFIDF (B)
the	2/9	2/9	log (2/2)=0	0	0
quick	1/9	1/9	log (2/2)=0	0	0
brown	1/9	0	log (2/1)=0.3	0.0333	0
fox	1/9	1/9	log (2/2)=0	0	0
jumps	1/9	0	log (2/1)=0.3	0.0333	0
over	1/9	0	log (2/1)=0.3	0.0333	0
lazy	1/9	1/9	log (2/2)=0	0	0
dog	1/9	1/9	log (2/2)=0	0	0
is	0	2/9	log (2/1)=0.3	0	0.0667
and	0	1/9	log (2/1)=0.3	0	0.0333

From the above table we can see that TFIDF of common words was zero, which shows they are not significant.On the other hand, the TFIDF of “brown”, “jumps”, “over”, “is”, “and” are non-zero.This words have more significance.