

TF*IDF is an information retrieval technique that weighs a term's frequency (TF) and its inverse document frequency (IDF). Each word or term has its respective TF and IDF score. The product of the TF and IDF scores of a term is called the TF*IDF weight of that term.

Put simply, the higher the TF*IDF score (weight), the rarer the term and vice versa.

The TF*IDF algorithm is used to weigh a keyword in any content and assign the importance to that keyword based on the number of times it appears in the document. More importantly, it checks how relevant the keyword is throughout the *corpus*.

For example, when a 100 word document contains the term "cat" 12 times, the TF for the word 'cat' is

$$TF_{cat} = 12/100 \text{ i.e. } 0.12$$

The IDF (inverse document frequency) of a word is the measure of how significant that term is in the whole corpus.

For example, say the term "cat" appears x amount of times in a 10,000,000 million document-sized corpus (i.e. web). Let's assume there are 0.3 million documents that contain the term "cat", then the IDF (i.e. $\log \{DF\}$) is given by the total number of documents (10,000,000) divided by the number of documents containing the term "cat" (300,000).

$$IDF (cat) = \log (10,000,000/300,000) = 1.52$$

$$\therefore W_{cat} = (TF*IDF)_{cat} = 0.12 * 1.52 = 0.182$$

(Note precise base of the logarithm is not important for ranking)