TF-IDF

Computers are **good with numbers**, but **not that much with textual data**. In natural language processing (NLP), **TF–IDF**, short for **term frequency–inverse document frequency**, is a **numerical statistic** that is intended to reflect **how important** a word is to a document in a **corpus**.

From our intuition, we think that the words which **appear more often should have a greater weight** in textual data analysis, **but that's not always the case**. Words such as "the", "will", and "you" — called **stopwords** — appear the most in a corpus of text, but are of **very little significance**. Instead, the **words which are rare** are the ones that actually help in **distinguishing between the data**, and **carry more weight**.

Term Frequency (TF) gives us the **frequency of the word in each document** in the **corpus**. It is the **ratio of 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. The TF of **each word** in a document is defined as:

$$\textbf{TF} = \frac{\text{Number of times the word appears in a document}}{\text{Total number of words in the document}}$$

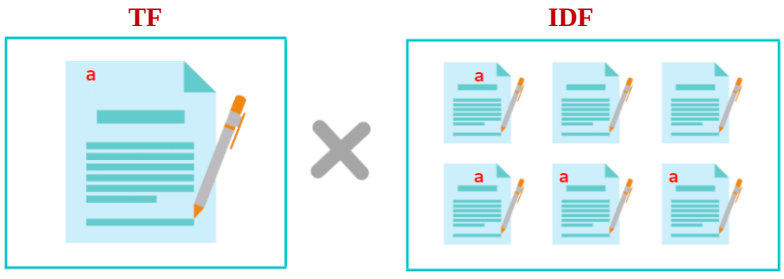
Each document has **its own tf**.

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**.
The IDF of **each word** is defined as:

$$\textbf{IDF} = \frac{\text{Total number of documents}}{\text{Number of documents containing that word}}$$

Combining these two we come up with the **TF-IDF score** for a word in a document in the corpus.

$$\textbf{TF–IDF} = \text{TF} \times \text{IDF}$$



Frequency of a word within the document

Frequency of a word across the documents

Let’s take an example to get a clearer understanding.

Sentence 1: The car is driven on the road.

Sentence 2: The truck is driven on the highway.

We will now calculate the **TF-IDF** for the above two documents, which represent our **corpus**.

Word	TF		IDF	TF × IDF	
	Sentence 1	Sentence 2		Sentence 1	Sentence 2
The	1/7	1/7	Log(2/2) = 0	0	0
car	1/7	0	Log(2/1) = 0.3	0.043	0
truck	0	1/7	Log(2/1) = 0.3	0	0.043
is	1/7	1/7	Log(2/2) = 0	0	0
driven	1/7	1/7	Log(2/2) = 0	0	0
on	1/7	1/7	Log(2/2) = 0	0	0
the	1/7	1/7	Log(2/2) = 0	0	0
road	1/7	0	Log(2/1) = 0.3	0.043	0
highway	0	1/7	Log(2/1) = 0.3	0	0.043

From the above table, we can see that TF-IDF of **common words is zero**, which shows they are **not significant**. On the other hand, the TF-IDF of “**car**”, “**truck**”, “**road**”, and “**highway**” are **non-zero**. These words have **more significance**.

Write a program that calculates the TF-IDF of collection of documents.

Link: freecodecamp

Link: Linguist

Link: Github