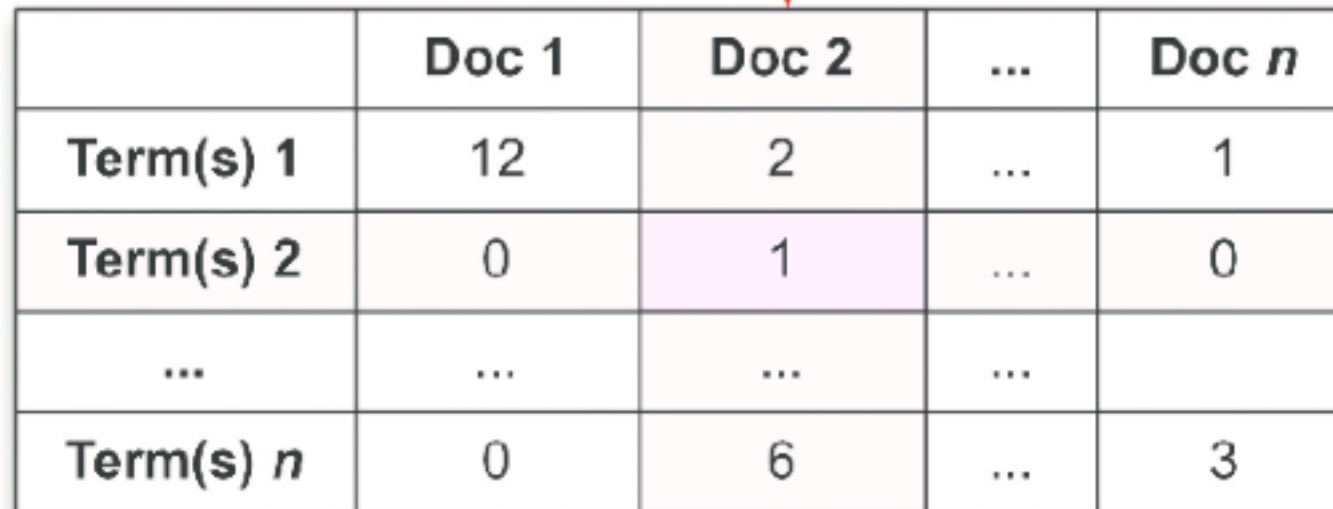


$$\text{tfidf}(t, d, D) = \text{tf}(t, d) \times \text{idf}(t, D)$$

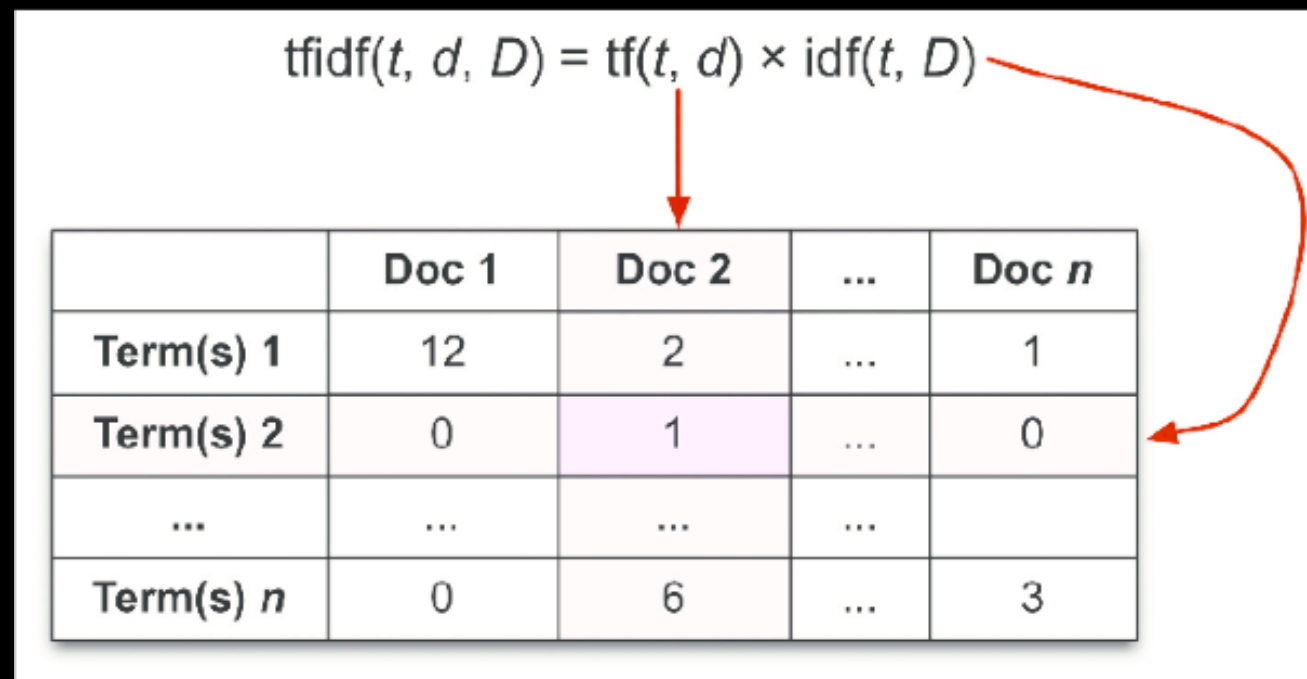


	Doc 1	Doc 2	...	Doc n
Term(s) 1	12	2	...	1
Term(s) 2	0	1	...	0
...	
Term(s) n	0	6	...	3

- **However**, if an **n-gram** occurs many times in not just one document **but many others** in the corpus, it could be that this **n-gram** is merely a frequent term and **not necessarily meaningful**. This is what the **IDF** balances for.
- In other words, with the most frequent n-grams (**TF**) we get a first approximation, but the **IDF** should help us to refine and get better results.

Calculating a **TF-IDF**

$$\text{tfidf}(t, d, D) = \text{tf}(t, d) \times \text{idf}(t, D)$$



	Doc 1	Doc 2	...	Doc n
Term(s) 1	12	2	...	1
Term(s) 2	0	1	...	0
...
Term(s) n	0	6	...	3

- **TF**(n-gram \mathbf{w}) = (Number of times n-gram \mathbf{w} appears in a document) / (Total number of n-grams in the document).
- **IDF**(n-gram \mathbf{w}) = \log_e [(Total number of documents in the corpus) / (Number of documents with n-gram \mathbf{w} in it)]

