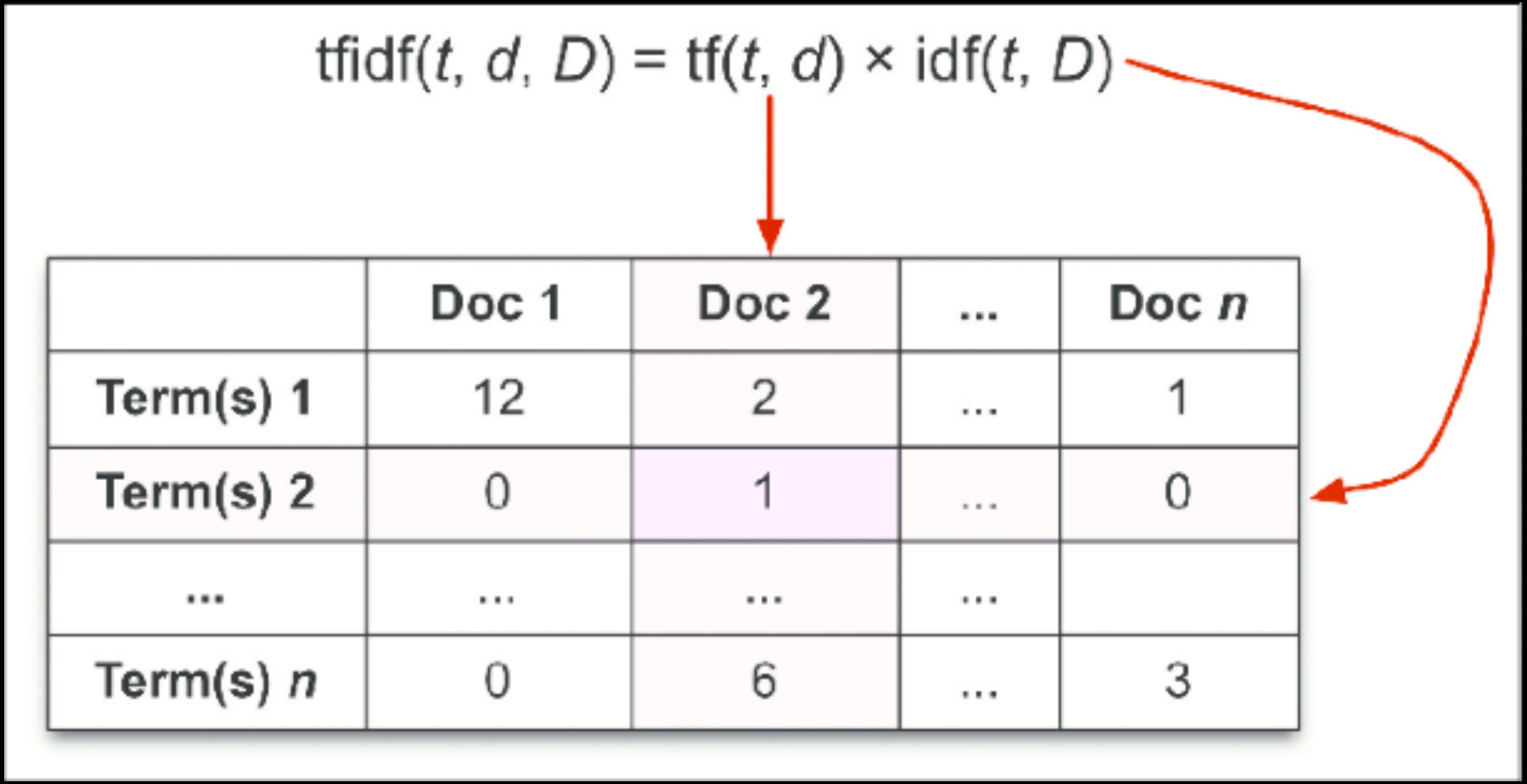
- We often want to count terms in the **documents** which they are contained.

- However, just counting how many times a word appears in a corpus often **does not always** capture whether this term is important.

- So, how do we resolve this?



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

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

- A **TF-IDF** measure (**term-frequency inverse [TF] document frequency [IDF]**) is a commonly used counting technique intended to reflect how relevant a term is in a given document.

- In other words, it is a **document specific measure** that tells you how important an **n-gram** is in a document **relative** to a **corpus**.

- **A TF-IDF** measure (**term-frequency inverse [TF] document frequency [IDF]**) is a commonly used counting technique intended to reflect how relevant a term is in a given document.

- In other words, it is a **document specific measure** that tells you how important an **n-gram** is in a document **relative** to a **corpus**.