***TF-IDF Vectorization Example with Sklearn***

**What is TF-IDF?**

**TF-IDF** stands for:

* **Term Frequency - Inverse Document Frequency**

It is a numerical statistic used in **Natural Language Processing (NLP)** to evaluate how important a word is to a document **relative to a collection of documents (called a corpus)**.

In simple terms:

* Words that occur **frequently in a document but rarely in other documents** are considered **important** and given **higher scores**.
* Words that are common across many documents (like "is", "the", "and") get **lower scores**, because they don't carry much useful information.

**How does TF-IDF work?**

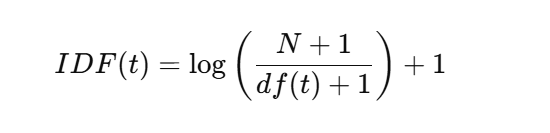
**1. Term Frequency (TF)**

It measures how many times a term appears in a document.

### 

### 2. ****Inverse Document Frequency (IDF)****

It measures how **unique or rare** a term is across all documents.



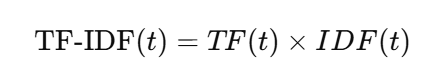
Where:

* NNN = Total number of documents
* df(t)df(t)df(t) = Number of documents that contain the term ttt

➡️ If a term appears in **many documents**, the IDF score becomes **low**.  
➡️ If a term appears in **few documents**, the IDF score becomes **high**.

### 3. ****TF-IDF Calculation****

Finally, the **TF-IDF score** for a term is:



**Why do we use TF-IDF?**

* It helps to **remove common words** (stopwords) and focuses on the **unique, meaningful words**.
* Converts **text data into numerical format**, so we can apply machine learning models (like classification, clustering, etc.).
* Helps in **feature extraction** from text.