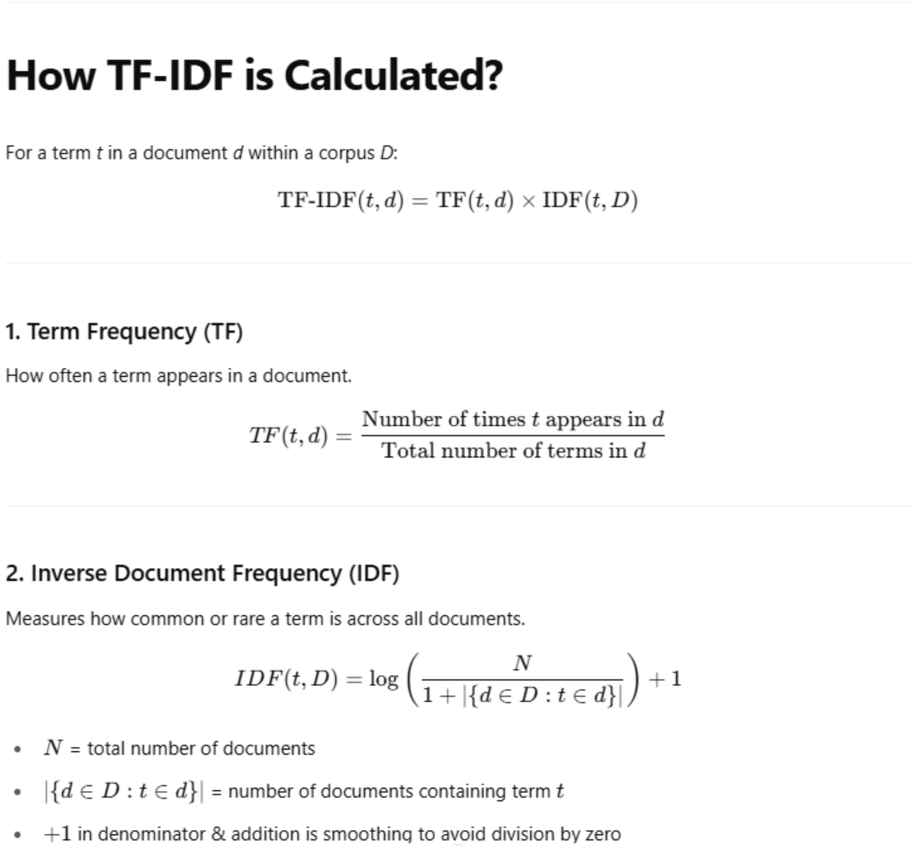
What is TF-IDF?

TF-IDF stands for **Term Frequency-Inverse Document Frequency**. It's a statistic that reflects how important a word is to a document relative to a collection (corpus) of documents.

It's widely used in text processing to convert text into numbers (vectors) to feed into machine learning models.



How TF-IDF is Calculated?

For a term *t* in a document *d* within a corpus *D*:

$$\text{TF-IDF}(t, d) = \text{TF}(t, d) \times \text{IDF}(t, D)$$

1. Term Frequency (TF)

How often a term appears in a document.

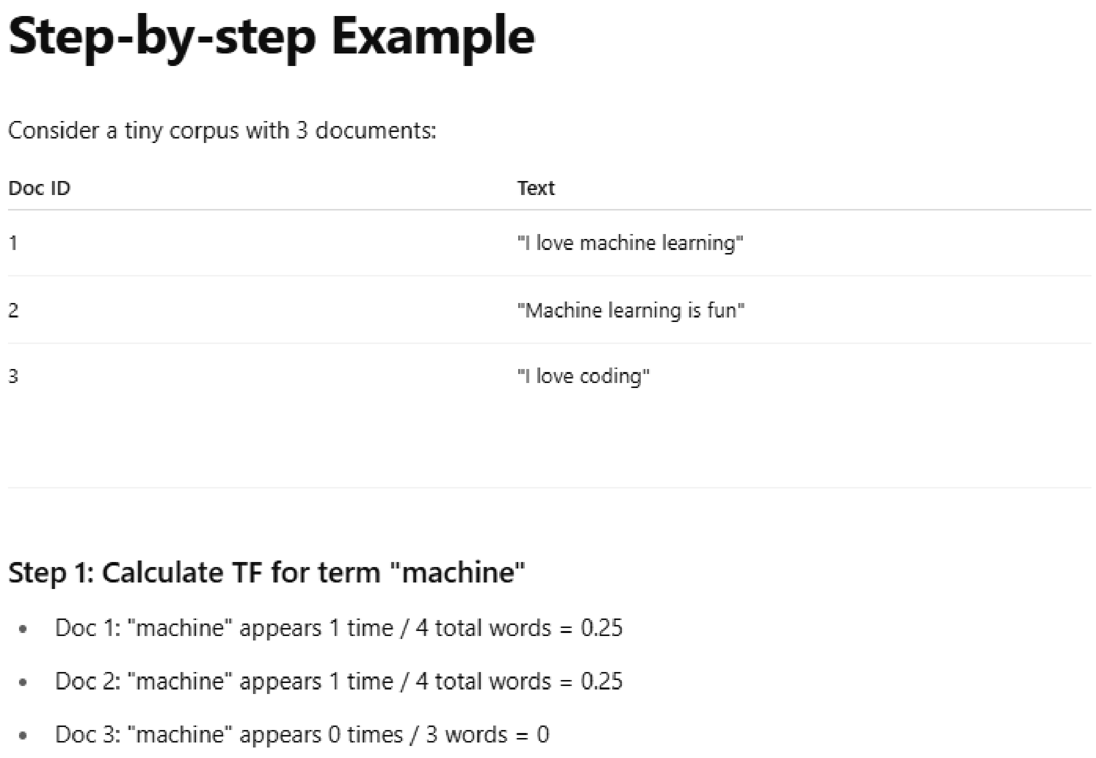
$$TF(t, d) = \frac{\text{Number of times } t \text{ appears in } d}{\text{Total number of terms in } d}$$

2. Inverse Document Frequency (IDF)

Measures how common or rare a term is across all documents.

$$IDF(t, D) = \log\left(\frac{N}{1 + |\{d \in D : t \in d\}|}\right) + 1$$

- N = total number of documents
- $|\{d \in D : t \in d\}|$ = number of documents containing term t
- $+1$ in denominator & addition is smoothing to avoid division by zero



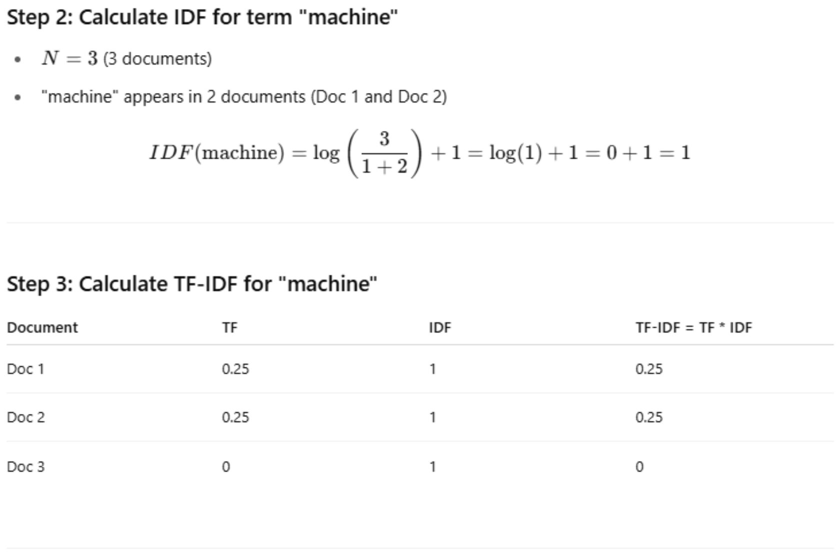
Step-by-step Example

Consider a tiny corpus with 3 documents:

Doc ID	Text
1	"I love machine learning"
2	"Machine learning is fun"
3	"I love coding"

Step 1: Calculate TF for term "machine"

- Doc 1: "machine" appears 1 time / 4 total words = 0.25
- Doc 2: "machine" appears 1 time / 4 total words = 0.25
- Doc 3: "machine" appears 0 times / 3 words = 0



Step 2: Calculate IDF for term "machine"

- $N = 3$ (3 documents)
- "machine" appears in 2 documents (Doc 1 and Doc 2)

$$IDF(\text{machine}) = \log\left(\frac{3}{1+2}\right) + 1 = \log(1) + 1 = 0 + 1 = 1$$

Step 3: Calculate TF-IDF for "machine"

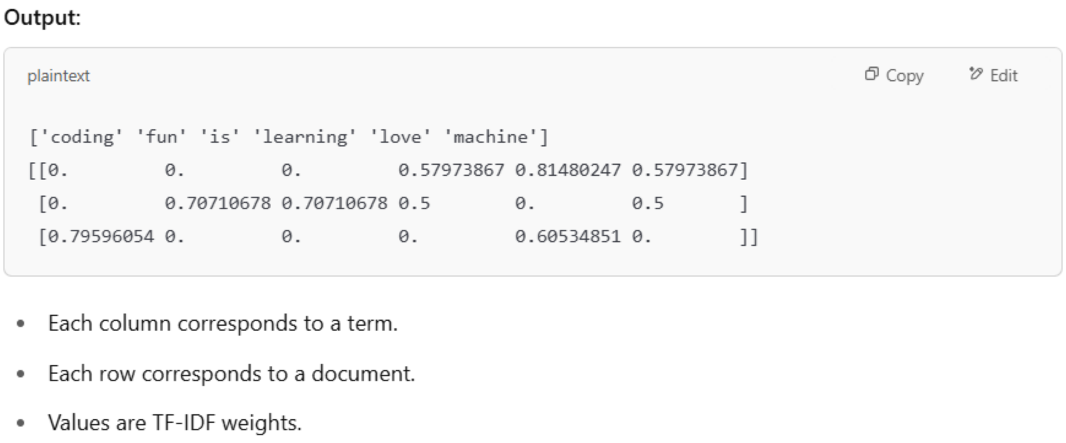
Document	TF	IDF	TF-IDF = TF * IDF
Doc 1	0.25	1	0.25
Doc 2	0.25	1	0.25
Doc 3	0	1	0

```
from sklearn.feature_extraction.text import TfidfVectorizer


corpus = [
    "I love machine learning",
    "Machine learning is fun",
    "I love coding"
]

vectorizer = TfidfVectorizer()
X = vectorizer.fit_transform(corpus)

print(vectorizer.get_feature_names_out())
print(X.toarray())
```



Output:

```
['coding' 'fun' 'is' 'learning' 'love' 'machine']
[[0.         0.         0.         0.57973867 0.81480247 0.57973867]
 [0.         0.70710678 0.70710678 0.5        0.         0.5       ]
 [0.79596054 0.         0.         0.         0.60534851 0.        ]]
```

- Each column corresponds to a term.
- Each row corresponds to a document.
- Values are TF-IDF weights.

Summary

- **TF** tells how frequent a word is in a document.
- **IDF** downweights common words across many documents.
- **TF-IDF** highlights important words unique to each document.