**Feature Extraction**: The mapping from textual data to real valued vector is called feature extraction.

**Bag of Words (BOW):** List if unique words in the text corpus.

**TF-IDF vectorization:**

**TF – Term Frequency**

TF measures how frequently a term appears in a document. It is the count of a word in the document divided by the total number of words in that document.

Formula:

TF(t)

Example:

Document 1: "I love data science and data analysis"

Let’s calculate TF for the word **"data"**:

* Count of "data" = 2
* Total words = 7

**TF("data") =**

**IDF – Inverse Document Frequency**

IDF measures how important or rare a term is across all documents. If a term appears in many documents, its IDF value is low (common words); if it appears in few documents, its IDF is high (unique words).

**IDF(t) = )**

Example:  
Suppose we have 3 documents:

* Doc 1: "I love data science"
* Doc 2: "Data science is fun"
* Doc 3: "Math and statistics are important for data"

The word "data" appears in all 3 documents, so:

**IDF("data") = ) = ) = 0**

Now, the word "statistics" appears only in Doc 3:

**IDF("statistics") = ) = ) = 1.0986**

The TF-IDF score is calculated by multiplying TF × IDF for a term.

So, if:

* TF("data") = 0.2857
* IDF("data") = 0

Then:

TF-IDF("data") **=** 0.2857×0 = 0

This means "data" is not very helpful in distinguishing documents since it appears everywhere.

But for:

* TF("statistics") = 1/7 ≈ 0.1429
* IDF("statistics") = 1.0986

Then:

TF-IDF("statistics") = .1429×1.0986 ≈ 0.157

So "statistics" has more weight and is more useful in identifying the topic of the document.