**NEED FOR NLP**

**1.BAG OF WORDS(BoW):**

• The bag of words model is a way to represent as numerical feature vector by counting the occurrence of the words.

• Here bag means all words of the dataset irrespective of order.

• In this model only the frequency of word matters.

• The model ignores grammar, word order and case sensitivity

**• Example:**

Suppose we have two sentences:

"I love playing football"

"I love playing cricket"

Vocabulary = {I, love, playing, football, cricket}

Sentence 1 → [1, 1, 1, 1, 0]

Sentence 2 → [1, 1, 1, 0, 1]

**Advantages:**

1.easy to implement and interpret

2.useful for basic models

**Disadvantages:**

1. Ignores Context – Loses meaning because the order of words is not considered.

Example: "dog bites man" vs. "man bites dog" → treated as the same.

2. Common Words Dominate – Words like "the", "and", etc., may overshadow meaningful ones.

**When to Use Bag of Words**

• When you just need word frequency counts.

• Useful for simple text classification tasks like spam detection or sentiment analysis.

• Works well when the dataset is small to medium-sized.

**2. TF-IDF (Term Frequency – Inverse Document Frequency):**

TF-IDF improves on Bag of Words by giving importance to unique words and reducing the weight of very common words.

It has two parts:

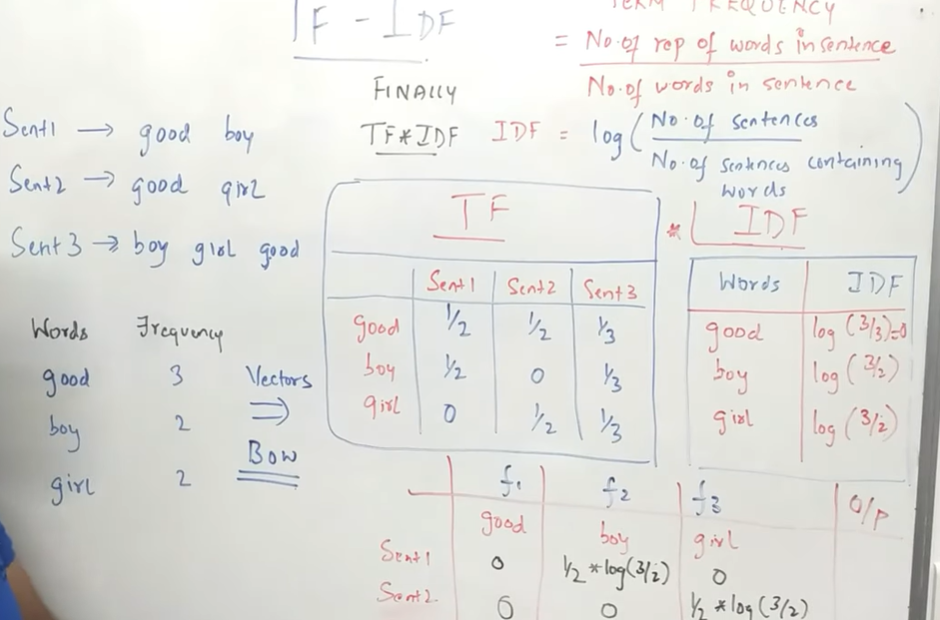
**TF (Term Frequency):** Measures how often a word appears in a document.

**IDF (Inverse Document Frequency):** Reduces the weight of words that appear too frequently across documents (like "the" or "is").

**TERM FREQUENCY =( NO.OF REPETITIONS OF WORDS IN SENTENCE / NO.OF WORDS IN SENTENCE)**

**INVERSE DOCUMENT FREQUENCY = LOG (NO.OF SENTENCES/NO.OF SENTENCES CONTAINING THE WORD**)

**FINALLY MULTIPLY IDF \* TF**

****

**When to Use TF-IDF**

* When you need to focus on important, unique words rather than just counting frequencies.
* Ideal for document classification, search engines, and information retrieval.