Before moving on to NLP techniques, lets start with understanding what it is and what it helps us with?

What is NLP?

NLP stands for Natural Language Processing, it is a branch of Artificial Intelligence (AI) that focuses on enabling machines to understand, interpret, and generate human language. In simple words, we humans talk in normal languages like english, hindi, mandarin etc, but the computer cannot understand these they only understand binary (0s and 1s). So how do we communicate with it? Thats where NLP jumps in, it acts as a **bridge** between the two, converting human language into something machines can process. It uses computational linguistics and machine learning to these human communications.
Common examples are Google Translate, AI chatbots (siri, gemini, alexa) etc.

The NLP Techniques

1) **Bag Of Words (BoW)**

   It turns sentences into a collection of words and counts each word's frequency irrespective of its order or grammar
   Vocabulary: It is the list of all unique words from the entire dataset
   Document Representation: A vector where each element shows a frequency of words from the vocabulary used as a feature.
   Working: After selecting the most frequent words, you choose the top n words with the highest frequency and build a BoW model from it by a binary matrix where each row corresponds to a sentence and each column represents one of the top N frequent words. A one in the matrix shows that the word is present in the sentence and a zero shows its absence and then we finally create a word cloud to identify most common words.

   Advantages
   ○ It is computationally efficient, versatile
   ○ Easy to implement, interpret
   Limitations
   ○ It ignores context, does not capture meanings so might miss important relations. With large datasets this method will be too confusing which would make it inefficient

2) One-Hot Encoding

- Converts categorical data into binary vectors to feed as input to neural networks
- Working: First stage is to determine the categorical variables, then count the frequency of distinct words to identify potential groups, the third stage is to make a binary vector for each of the categories called one hot encoded vector
- Drawbacks:

  1)High dimensionality: If vocabulary is very large, the vectors become very big and **sparse** (mostly zeros).
  2)No semantic meaning: It cannot capture relationships between words. For example, "king" and "queen" will be as unrelated as "king" and "apple".