Research Task SYNAPSE – Anish Chandurkar

Option 1 : Machines cannot directly understand raw text the way humans do. To process language, we use the field of Natural Language Processing (NLP), where words are transformed into numerical representations that algorithms can work with. Over time, several techniques have been developed to represent words as vectors. Explain various techniques used to convert words into numbers.(e.g Bag of Words, TF-IDF etc.).

We communicate with words, and so we can understand each other but computers cannot do that naturally. They operate with numbers, we need to convert text into something numerical and so we have different techniques for this.

**1)Bag Of Words: -**  
In this technique the frequency of each word is noted down. Let’s say 2 sentences, “His name is Anish”, “Anish loves programming”. In this Anish is repeated twice and while the other words are repeated once. Thus, the computer takes a note of it and converts it into vectors, which aids them in understanding.

**2)Term Frequency – Inverse Document Frequency:-**(also referred to as TF-IDF), you calculate the TF by the formula ,   
and you can calculate IDF by the formula, . Then for each word in each of the sentences, you calculate (TF \* IDF) value, which gives a decimal and feed it to the model for learning.

**3)One – Hot Encoding: -**   
In this technique, each word is given a unique vector. Let’s say the vocabulary is [His, name, Anish]. Then the word “His” is represented as [1,0,0], “name” as [0,1,0], and “Anish” as [0,0,1]. Thus, the computer converts every word into such binary vectors, which helps in processing the text, though it does not capture meaning or similarity between words.