TASK 3

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.).

🡪

INTRODUCTION

The language we humans speak In our day to day life is something the ai or the computer systems cannot understand so for that issue to be solved and addressed Natural Language Processing (NLP) is crucial field in Artificial intelligence that makes the machines understand what the user or the human is trying to say as machines cannot directly understand raw text or sentences so it has to be converted in numerical representation which makes it ideal for the machine to process the data.so for doing the above mentioned process many methods were created like One-Hot Encoding, Bag of words ,Term Frequency – Inverse Document Frequency(TF-IDF) etc. the recent ones like BERT ,GPT which come under Contextual Embeddings .

i) One-Hot Encoding

This One Hot Encoding is the simplest and the most easiest techniques to show words as numerical values in NLP . what is does is that it basically gives each unique word in the dataset a unique vector that has a single element as 1 and the rest as 0 .That vector is called as One Hot vector .

Eg: suppose one sentence is “I am boy” and another “I am girl”

As the unique words are {I, am, boy, girl } 🡪“4”.

So every vector will have a length of 4 as each word can be represented as

I = [1,0,0,0]

am=[0,1,0,0]

boy =[0,0,1,0]

girl =[0,0,0,1]

like this each word is uniquely represented in this technique

however it is memory inefficient and also can not handle context as in how and where the same word is used.

ii) Bag of Words (BoW)

This technique is one of the easiest earlier One-Hot Encoding used represent only words but BoW represent sentences as fixed -length vetors it also considers word frequency into account. It is called a bag as words are randomly put in it ,it ignores word order .

Eg : one sentence(S1) is “The boy eat the cake”

another (S2)“The girl eat the pastry” as unique words are {the,boy,girl,eat,cake,pastry}

The vectors can be represented as :

S1 = [2,1,0,1,1,0]

S2 = [2,0,1,1,0,1]

However it lacks in grammar ,semantics and also context.

iii) Term Frequency – Inverse Document Frequency(TF-IDF)

This technique builds on the Bag of Words model as Bow only focuses on counting the word repetition and nothing else. This technique TF-IDF brings a new concept of telling how important a word Is with in a document this helps to focuse on the important words not the ones like (“the”,” am “,”is ” etc ) which are common in the entire document.

Term Frequency(TF) : it shows how frequently a term Is present in the document.

Inverse Document Frequency (IDF) : it shows how unique a word is in a document.