Different ways to convert words into numbers :

1. One hot encoding : In this method each word is represented as binary vector.

For eg : [“Car”,”Toy”,”Bat”]

Here Car = [1,0,0]

Here Toy = [0,1,0]

Here Bat = [0,0,1]

But the disadvantage is that it still cannot differentiate between car and toy.

1. Bag of words :

In this it counts the occurrence of the word in the sentence. For eg

S1 : I love eating pizza

S2 : I love the food.

Vocabulary : (I, love, eating, pizza, the, food)

Vector : S1 [1,1,1,1,0,0]

S2 [1,1,0,0,1,1]

But the disadvantage is still order and word meaning is ignored.

1. TF-IDF (Term Frequency – Inverse Document Frequency)

In tf-idf it adds weight to each word. The word which gets repeated many times in the document gets less weightage, whereas rare word gets more weightage.

TFIDF=TF× (logN)/DF

Where tf = term frequency (how many times the word repeat in the document)

N = No of document

DF = no of document in which the word exist

Advantage – better than bag of words

Disadvantage – still ignores the meaning and order

1. Subword Models :

In this the words are broken down into many parts and then interpreted

Eg unhappiness is broken into “un” + “happy” + “ness”

So even if it does not know the meaning of unhappiness it interprets by breaking the word.

Advantage – ordered and meaningful