3.1)

Let’s say you are given a large amount of textual data- messages, emails, books, etc. Before performing any operations on this data, it is necessary to clean and preprocess the data (removing unnecessary words or symbols, etc.). Explain how you would go about preprocessing. What different steps would be followed? Why are they necessary?

->

First we’ll import the required libararies like nltk,tensorflow and import the text.We’ll analyze the data and accordingly run the necessary steps which are:

1.Converting our text, to a text which does not have punctuations like , . ‘ ) ( etc.

2.Tokenisation of the text.

3.Removing words like is,and,any,not,yours,etc also called as stop words that do not give context of the text individually.

4.Converting all the characters to lower case

5.Lemmatization

6.Stemming

7.Converting the text to embeddings

These steps are necessary for the model to easily analyze and modify the text in its own terms and perform well on training data and test data.

3.2)

Have you ever wondered how streaming platforms like Netflix work and how they recommend movies or shows based on your current watch? How does a bank decide which customers get loans and which do not? This all is done using Unsupervised learning. Machine Learning is internally subdivided into different parts- one of them is Unsupervised learning. The technique used for these kinds of problems is known as Clustering. So, for this task, explain what clustering is and describe any two types of clustering.

-> Clustering is grouping similar data points with each other. It comes under Unsupervised machine learning where labels are not given. In clustering many algorithms like K-means clustering are used. K- means starts by randomly selecting cluster centroids. And assigning each datapoint to its nearest centroid. Then the average of each cluster is calculated through the data points. The average of datapoints of each cluster now becomes its centroid and the first step that is assigning each datapoint to it’s nearest centroid is repeated. This process is repeated till a new cluster keeps forming. The point where there are no new clusters formed is called convergence.

Two types of Clustering:

1. Hard Clustering: In this type of clustering, each datapoint is assigned to only one cluster. Datapoints are assigned based on their similarities with each other. This type is used when there is no ambiguity in data.

2.Soft Clustering: In this type of clustering, the probability of a datapoint belonging to each cluster is assigned. A single datapoint can belong to multiple clusters. This type helps us in handling the uncertainty in the data.