**Himanshu Kumar**

**Q1 - Explain the steps involved in Bag of Words Model?**

**Ans-** Bag of Words model signifies the binary distribution of sentences with the help of vectorisation methods.

Moreover, In this method, out of group of sentences words have been chosen which are being written in tabular form and with the help of binary distribution concurrency of each word is calculated.

There are number of steps which are being followed which are described below:-

**STEP 1 :- COLLECT DATA**

*It was the best of times,*

*it was the worst of times,*

*it was the age of wisdom,*

*it was the age of foolishness,*

**STEP 2 :- DESIGN THE VOCABULARY**

For Eg, best, times, it, worst, age, wisdom, foolishness. Was

**STEP 3 :-  CREATE DOCUMENT VECTORS**

Under this, a tabulated form distribution of words in binary form is done in order to calculate the occurrence of words.

**STEP 4 :- MANAGING VOCABULARY**

* Ignoring case
* Ignoring punctuation
* Ignoring frequent words that don’t contain much information, called stop words, like “a,” “of,” etc.
* Fixing misspelled words.
* Reducing words to their stem (e.g. “play” from “playing”) using stemming algorithms.

**SCORING WORDS**

* **Counts**. Count the number of times each word appears in a document.
* **Frequencies**. Calculate the frequency that each word appears in a document out of all the words in the document.

Moreover, bag of words further expanded to TF-IDF model, TF stands for Term Frequency whereas , IDF stands for inverse document Frequency.

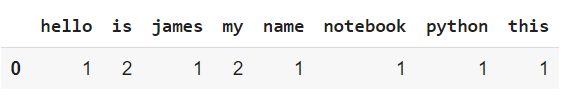
TF = number of times term “t” appear in a document divided by total no. of terms in a document

IDF = log(no. of terms in a document divided by total no. of document)

**Q2- Explain about Count vectorizer and TDF-IDF vectorizer with example?**

Ans- Count vectorizer is a method to convert text into numerical data for easy understanding by machine.

It includes breaking down a sentence or any text into words by performing preprocessing tasks like converting all words to lowercase and removing special characters. For eg:-



TF-IDF is a measure of originality of a word by comparing the number of times a word appears in a document with the number of document the word appears in.

Last date of submission of assignment is 04/09/2022.

You should prepare it early.

Here in these sentences, if we see, there are number of words which are being used , hence, with TF-IDF we can use the above formula in order to convert the concurrency in a document of two sentences along with the use of log function.

**Q3-What is semantic analysis. Explain the techniques to achieve semantic analysis 1) Name Entity Recognition 2) Co Reference Resolution?**

**Ans:-**

Semantic analysis is a method which gives machine a ability to understand natural language which is a part of NLP (natural language Processing) of which Semantic Analysis is a part.Semantic analysis means giving exact meaning to text..

Techniques to achieve semantic analysis are:-

FOPL( First order predicate Logic):-  converts sentences into logical form

Semantic Nets:- it is used to generate link nets which lets machine understand the relationship among words like Dog, Cat, pet.

Case Grammar:- defining logical relationship doesn’t simply mean that machine understands the text provided to it hence, grammar connection also plays an important role.

Name Entity Recognition is a technique which is used by several companies that uses machine learning to identify named entities in text data and classifies them into one or more predetermined categories.

Co-Reference Resolution is a task of finding all expressions that refers to same entity in a text.

**Q4 - Write a note on Word Embedding?**

**Ans-** In Word Embedding, the words which are similar in meaning shall be grouped together in vector space.Foreg:- Frog, the nearest neighbor of a frog is toad, frogs, Litoria shall be grouped nearby the parent word ‘Frog’.

Word embedding is a method of extracting features out of text so that we can input those features into a machine learning model to work with text Data.

**Q6 - What are the four components of time Series. Give 5 example of time series Data?**

**Ans-:-**  components of time Series are as follows:-

Seasonal variation

Trend variation

Cyclical variation

Random variation

Time series is a sequence taken at successively  equally spaced points of time. Moreover, it is a sequence of discrete time data wherein, we need to gather information on regular intervals , for eg, Moving average of Dow Jones, Nifty 50 stocks, movement of sensex from last 10 years, heights of ocean tides,estimation of sales for next 3 years of an insurance company.

**Q7- Explain the Exponential Smoothing Models in Time Series?**

**Ans-** Exponential Smoothing Model is a time series forecasting method for univariate data that can be extended to support data with a seasonal component. There are three types of exponential smoothing, Beta, smoothing factor for the trend., Gamma, smoothing factor for the seasonality, trend type, additive or multiplicative.

Exponential smoothing forecasting methods are similar in that a prediction is a weighted sum of past observations, but the model explicitly uses an exponentially decreasing weight for past observations

**Q8- Explain the ARIMA models in Time Series?**

**Ans-**  Autoregressive Integrated Moving Average (ARIMA) models have many uses in many industries. It is widely used in demand forecasting, such as in determining future demand in food manufacturing.

ARIMA stands for Autoregressive Integrated Moving Averages, is a form of regressive analysis that indicates the strength of a dependent variable relative to other changing variable, its objective is to predict the future time series movement.

**Q9 Using attached dataset of Nifty 50, do closing price forecasting of Nifty 50 for 9th May 2020 to 20 th May 2020?**

**Ans-**