Total Pages 12 (Twelve)

SYMBIOSIS INTERNATIONAL (DEEMED UNIVERSITY)

(Established under Section 3 of the UGC Act 1956)

Re-accredited by NAAC with 'A++' Grade | Awarded Category - I by UGC

वसुधैव कुटुम्बकम्

Verified all entries & found correct
Jr. Supervisor's Signature & Date
22/10/24

No. 075663
SIU-24 E

Centre: SCMHRD, Pune

Seat No. (in figures) 2 0 4 0 0 8

Seat No. (in words) Four zero zero eight

Examination : Day & Date : Tuesday - 22/10/2024

Programme : MBA BA Semester : III

Course : Natural Language Processing

Main Ans. Script 1 + No. of Supplementary Answer Scripts ______ =Total

Q.No.																Total	Signature of Examiner
Max. Marks																	
Marks Obtained																	

INSTRUCTIONS

1. Mention your details only in the space provided for in the main answer script & the supplement. If any other details (including seat number, name, contact details, etc.) are written anywhere else in the answer script and or supplement it will be treated as adoption of unfair means and the performance will be treated as null and void for the entire examination.

2. Write answer in legible hand. Answers written in an illegible and undecipherable hand are liable to be marked as zero.

3. An act of copying or of impersonation at an examination is punishable under the Maharashtra Prevention of Malpractices at University, Board and other specified examinations Act 1982.

4. Candidates should write answers in BLUE/BLACK ink only. Use of Pencil and other colors are permitted only in case of diagrams, graphs etc. Answer Scripts written with pencil or ink of other colors will not be evaluated.

5. Write on both sides of paper.

BEGIN WRITING HERE :

③Ans- Pos Tagging - Pos tagging is a method by which the words of a text are tagged with their appropriate parts of speech such noun, pronoun, verb, adjective, etc. By adding an additional layer, it helps us understand the use of word in a sentence.

Example - "Students are studying NLP". This text is tagged with Parts of Speech. Students → Noun, studying - Verb, NLP → Noun

PoS Tagging in NLP removes ambiguities and enhances text understanding in NLP. For example, consider this sentence -

"John is scared of flying planes"

Here, flying can mean an action word - verb or an adjective - flying planes (planes that fly). With PoS tagging, it gives a clearer picture and helps the model understand better to give accurate results

PoS tagging is a linguistic method of analysis It is a step in text preprocessing and an essential step as it also helps in overcoming the problem of lexical ambiguity.

④ Ans- N-gram Model -

N-gram model is a type of statiscal language Modeling. In this type, the continuous sequence of 'n' items are taken from the text and the probability of each word occuring is calculated. N-gram model is also best used to predict the next words when few set of words are given. For example - "I like to study NLP" is input text. N-gram model will break it down to - "I", "like", "to", "study", "NLP" ($n=1$). In case of binary model where $n=2$, the result will be "I like", "to study", "study NLP".

In this, the Markov assumption states that the current word is affected by the previous words. Ex - If "San" is given to the model, there is high probability that it will predict "Francisco" and less probability of predicting "Delhi". There are various measures like Entropy, Cross Entropy and perplexity to determine the accuracy of this model.

Limitation - The limitation of this model is that it does not deal with missing values. Also, if the ~~value doesnot~~ model doesnot identify any word correctly, Then it gives it a probability of zero. This reduces the accuracy. For example, in the sentence "I ate Icecream" The probability is calculated as

$$P(I|<s>) = \frac{1}{3}$$

$$= \frac{1}{3} \times \frac{1}{2} \times 0 = 0$$

$$P(ate|I) = \frac{1}{2}$$

$$P(icecream|<E>) = 0$$

5 ans

Constituency Parsing	Dependency Parsing
1. It focuses on the relation-ship between the constituents or phrases - Nounphrase, verb Phrase - of a sentence	1. It focuses on the relationship between the words of a sentence by identifying subject, object, verb
2. It uses phrase based grammar structure like Context-free grammar.	2. It uses dependency grammar structure
3. It follows top-down approach for Constiuency Parse tree - root to leaves	3. It follows bottom-up approach for it is Parse tree - leaves to roots
4. The graph is represented by non-overlapping constituents	4. The graph is represented by nodes (words) and edges (connection between nodes)
5. It is suitable for Natural Language Understanding (NLU)	5. It is suitable for Natural Language Generation (NLG)
6. It is more expressive and complex to interpret	6. It is simple and efficient but less expressive
7. It is used for languages with high morphology like Korean, Finnish	7 It is used for languages with less morphology like English, Chinese
8. It is used for traditional NLP tasks - NER, Pos tagging, etc	8 It is used for advanced NLP tasks - Speech recognition, prediction modelling, etc
9 Follows a simple syntactical structure	9 Follows a complex syntactical structure

(6) Ans- Naive Bayes Classifier - or a set of multiple models

Naive Bayes Classifier is a classification model which is used for text classification. The model determines the probability of a given instance from the set of features. It is used in high dimension datasets as one word is one feature. It is easy to calculate and speed. It is used for text classification, Spam filteration, text processing etc. For example, the task is to identify whether or not an email is a spam. It classifies the emails as "Spam" and "not Spam" by calculating the probabilities.

The assumptions underlying Naive Bayes Classifier -

1. All the features are independent of each other.
2. All the features are important for calculating the probabilities.
3. Continuous data follows a normal distribution.
4. Discrete data follows a multinomial distribution.
5. There are no missing values.

The accuracy of this measured using confusion matrix by determining - True Positive, True Negative, False Positive, False Negative.

(1) ans- Text Pre-Processing is the first step in NLP. The raw text that is provided as an input has to be pre-processed in order for the machine to understand better. Removing spaces, unnecessary words, missing values, tokenization, etc are few of the steps in text pre-processing. Broadly the steps in text pre-processing include –

1. Document ~~Triage~~ Processing 2. Sentence Processing

1. Document ~~Triage~~ Processing – In this step, the input text is converted into structured document. This is done to retrieve information and ignore the parts that are not important. It is important for identifying structure, retrieve information and process tasks. This step involves –

1. Data Encoding Identification – The structurture encoding is identified that is the byte size of the characters
2. Language Identification – The language of the text is identified
3. Removing unnecessary data like images, GIFs, etc

2. Sentence Processing – In this step, the text is broken into sentences – sentence tokenization, and then further into words – word tokenization. Here, the focus is on syntactic structure of the words. It involves, grammar and spelling corrector, sentence formation, POS tagging, etc. These are done on the basis of grammar rules. The words and sentences are arranged in a structured way. For example, "I am church go" – "I am going to church".

As a part of pre-processing, the model also eliminated filler words and also keeps important words that makes it easier to understand –

Ex. "I going Church" – 'am' and 'to' is removed

The importance of text-pre-processing is that it clears out anamolies by using various methods like POS tagging, Named Entity Recognition, Stemming, Lemmatization, etc.

The noise from the data is also removed during this step. This improves the accuracy of model.

(8) Ans- Components of NLP -

The main two components of NLP are -

(1) Natural Language Understanding - In this, the machine tries to understand the human langugage that is given as an input in order to generate results. It converts the human language to a language that is understable to the machine (for example, in binary digits) This is the first step in NLP.

It handles various anamolies -

1. Lexical anamoly - when the same words contained in a text have a different meaning.
Ex - "Rupa is going to bank" Bank can mean a financial institution or Bank of a river.

2. Syntactic anamoly - when the sentence has different meanings.
Ex - "John is scared of flying planes" flying plane can mean planes that fly or an activity of flying planes.

3. Referential anamoly - when the pronouns are used to refer to something and it becomes unclear.
Ex - "Pooja and Sashi are in the park. She fell down!" She can refer to either Pooja or Shashi.

(2) Natural Language Generation - In this, the machine generates the input that is understable to humans. It gives out the output after processing it. NLG converts the machine understable language to human language.

this is done by breaking the text into sentences (sentence sequencing) Then further the sentences are broken down into words (words sequencing) Then these are arranged in a meaningful structure

$$\boxed{NLP = NLU + NLG}$$

NLU is harder than NLG

Example of NLG - "Ingredients essential dish tasty → Ingredients essence dish tasty → Ingredients are the essence of a dish that makes it tasty

(9) ans Language Modelling -

Language modelling is the probabilistic description of language phenomena. It finds the probability of words from given alternatives. Language modelling is used in many cases like speech recognition, text classification, retrieve information, sentiment analysis, etc

Example -

Probability of choosing right text

1. Spelling Correction - I lvoe NLP < I love NLP

2. Sentence correction - I have two ice and one nose < I have two eyes and one nose

3. Word Prediction - Please turn right the light < Please turn off the light

4. Sentence Formation - I sweets eat < I eat sweets

Different methods of language modeling –

1. Statistical Language Modeling - In this type, the probability of each word occuring in a given text from the alternatives is calculated. N-gram is one the most common type of this modeling. The ~~word~~ sentence is broken down into 'n' items to predict the words. This model also predicts the next words based on previous words.

$P(w|w_1, w_2, w_3 \ldots w_n)$ – This means calculating the probability of 'w' based on $w_1, w_2, w_3 \ldots$ which are given. This is a traditional method

2. Neural Language Modeling – In this type, the probability is calculated of each word occuring. This is more advanced method of language modeling. This is used in advanced applications of NLP like likelihood maximization, probability analysis, etc. These used word embedders to calculate the probability. The words are assigned vectors here.

①Ans- Natural Language Processing (NLP) is branch of AI. It is way of interaction between humans and computers. NLP encompasses methods or techniques that help the machine to read, understand and interpret human language. It uses statistical methods and grammar rules to process the text/speech. Today, NLP is being used in various domains.

Real world application –

1. Speech Recognition – NLP is used in speech recognition and further in speech-to-text conversion. For example, Siri, Cortana, Alexa are used for speech recognition. It obeys

commands from the user to simplify their tasks Example, "Siri, remind me to pick groceries at 4 pm" - The user is simplifying his task by giving a task to the machine

Limitation - Sometimes, the machine may not understand the commands that were not used to train or misinterpret. Language barrier can be another limitation.

2. Spelling and Grammar Correction - NLP is used to check for spellings and grammar for a text. Grammarly is a famous tool. This is used by users to draft important work emails ensuring there are no spelling mistakes. Limitation - sometimes, the model may show an error to words that are not in its dictionary. Example. abbrievations, social media language, etc.

②ans Commercial Uses of NLP -

1. Chatbots - NLP is being used in Chatbots to interact with humans. They are used by many businesses to solve queries, give prompt replies, give basic information by reducing manual intervention.
   Ex- Zomato, Swiggy uses chatbots to assist their customers with complains related to food delivery
2. Text prediction - NLP is used by various companies like Google, samsung, etc. on their own keyboards that are embedded with this feature. They predict the sentences as the user types
3. ChatGPT - NLP is used in ChatGPT that answers the questions of the user

It gives only specific and relevant information. It also carries out features like text summarization etc.

4. Transcripts – Many applications like youtube, Ms Teams, Zoom converts speech to texts in the form of transcripts. This helps the user to understand the speech effectively with text generation.

10 Ans The challenge of text pre-processing when dealing with complex languages —

1. These complex languages may not have spaces between them like in English. This makes it difficult to perform word tokenization.
2. The letters in complex languages are in symbolic (logographic) format. The Encoding structure may vary from that of english letters. Example, english letter may take 3 byte size and the chinese letter may take 1 byte size.
3. A single symbol might mean an entire word in Chinese. This becomes a challenge to identify.

To overcome these challenges, the machine should be trained on these languages by providing corpus. The corpus dependent learning should be specific to chinese or arabic scripts which are complex.