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

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

No. SIU-24 E 075784

Centre: SCMHRD

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

Seat No. (in words) Two Zero Four Zero Zero Four

Examination : Day & Date : Tuesday 22 Oct 2024

Programme : MBA - Business Analytics 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 :

Q1 Natural Language Processing (NLP) is a way in which a machine interacts and understand the languages of humans speak and write and produces an output which humans understands. It is a subset of Artificial Intelligence and Machine Learning. Its two real world applications are as follows :-

(1) Chatbots and Voice Assistants
NLP is being extensively used by organizations to create chatbots, improving customer interaction services and thereby reducing load on people staff.
Its benefits include :

i) Automation of customer support services upto a certain extent thereby minimizing

human interaction.

2) It has increased the efficiency of customer services by increasing capabilities to manage tasks.

Challenges:

1) Training of the model on large datasets
2) Understanding ironical and sarcastic comments / feedback

(2) Sentimental Analysis - It is the application of NLP in which the model aims to understand the sentiment / emotions behind a said sentence / phrase e.g "Good morning! How are you doing today?"
This sentence conveys a cheerful meaning suggesting that the person is wishing the other good morning and asking about the other's day.

Benefits

1) This is typically very helpful in understanding customer feedback on products
2) It is also extensively used in understanding public opinion on social media platforms like X, Instagram, etc.

Challenges

(1) Understanding lexical ambiguity in phrases / acronyms. e.g "iykyk" stands for "if you know - you know" which is a common sarcastic comment on social media platforms.
(2) Understanding ironical and sarcastic statements

P.T.O.

Commercial uses of NLP are as follows :-

Q3 POS tagging refers to Parts of Speech tagging where different parts of text are broken down into smaller units where each word is tagged to a part of speech, i.e. noun, verb, adjective, etc.

It is a part of Natural Language Understanding unit of NLP and it enhances text understanding in the following ways :

(1) It helps in sentence segmentation - i.e; breaking down the sentence into smaller parts for better understanding

(2) It helps in understanding the Syntactic part of the given text/ phrase, i.e. whether it is grammatically correct or not.

(3) It helps in stemming and lemmatization which involves breaking down of words into its simpler form while retaining context.

(4) It also helps in discourse integration, i.e; understanding how one sentence relates to the other in a text

e.g. "Good morning! When is the next meeting scheduled"

noun - meeting, meeting
as verb - scheduled

Q4 N-gram models are the models which give the probability of occuring of a word based on 'n' words or n-1 words, before it. It helps us in predicting the probability of occurence of a word based on the frequency of occurence of previous words. N-gram models consists of :-

Unigram - It describes that the probability of occurence of one word before the selected word, and hence the word 'uni'

Bigram - It describes that the probability of occurence of a word is dependent on the two words before it namely, $(n, n-1)^{th}$ words. Hence bigram model.

Trigram - Describes that the probability of occurence of a word depends upon the ~~selects~~ 3 more words before it - $(n, n-1, n-2)^{th}$ words and hence the word trigram model.

Following are the limitations of N-gram model :-

(1) Frequency of occurence of rare words cannot be modelled easily.

(2) Semantic ambiguity may arise in case if only a word of a phrase is available

(3) It requires a lot of resources to train and deploy and hence it is costly to train and implement

TO

Q7 Steps in text-preprocessing include Sentence segmentation, tokenization, Stemming and lemmatization, ~~Dynamic Analysis~~, POS Tagging, Discourse Analysis

(1) Sentence Segmentation is used to breakdown the text or paragraph into smaller units of sentences.

(2) Tokenization - Refers to the breaking down of a sentence into smaller parts or token for processing.

(3) Stemming - Refers to the breaking down of a word into its root form by removing the suffixes and the prefixes

(4) Lemmatization - Sometimes the word may not retain context when Stemming is ~~not~~ done on it so in order to retain context, lemmatization is done so that words are broken down into its smaller form while retaining the context

(5) POS Tagging - Refers to the tagging of various parts of speech such as noun, adjective, verb, etc which helps in the discourse analysis

(6) Discourse Analysis is the step in which the words are contextualized and eventually phrases and sentences are formed which ~~give~~ have meaning.

Q8

(1)

(2)

(3)

(4)

(5)

SIU SIU SIU SIU SIU SIU SIU SIU SIU SIU SIU SIU SIU SIU SIU SIU SIU SIU SIU SIU

These steps are essential to test the model trains well and is able to grasp the context of the data and produce accordingly.

Q2 / The components of Natural Language Processing are ① Lexical Analysis, Syntactic Analysis, Semantic Analysis, Discourse Analysis, Pragmatic Analysis

① Lexical Analysis - This includes the analysis of the words which are used and if they are from a recognized / trained dataset or not; i.e. if the model is able to recognize the language or not.

② Syntactic Analysis - Refers to the syntax of the language text and if words are mentioned / written properly or not, i.e. if they follow the grammar rules of the language or not.

③ Semantic Analysis - Refers to the meaning that can be inferred from the identified words (if any).

④ Discourse Analysis - This refers to the flow of text or any context in the text if it is being logically followed, i.e. text with context or references.

⑤ Pragmatic Analysis - This refers to the analysis of the statements, i.e. if they have any hidden meaning or any ironical, sarcastic or unethical statements, etc.

P.T.O.

6) Language modeling refers to the machine learning models which are created in order to understand create output based on the training set. These models can have various applications such as sentiment analysis etc. to help determining the output the different models include:

(1) Rule based model - These models consist of various rules which are programmed into them internally which help them classify and predict the output accurately.

(2) Deep Learning Models - Deep learning models deploy neural networks typically RNNs (Recurrent Neural Networks) due to their high efficiency in dealing with textual data. Deep learning models are very efficient as they apply reinforcement learning to automatically penalize wrong outputs so that the model can train itself correctly.

(3) N-gram model - N-gram model works on the principle that the probability of occurring of a word depends on the occurrence of previous 1 word (unigram), 2 words (bigram) or 3 words (trigram models).

b) Markov Model - Markov model is a model which uses probability to determine the likely occurrence of the next word. It works on the model that the occurrence of the next word only depends upon the current state and not all of the previous states / words. It assigns weights on the states to determine the outcomes.

c) Hidden Markov Models - This works on the principle of markov models with an addition that the exact value can be calculated based on the probabilities of the many 'hidden' states / layers in the markov model. It also works on the principle of adjusting weights dynamically to conclude the outcome.

Q10 / The challenges of text pre-processing when dealing with languages like Chinese or Arabic, which have complex scripts include:

① Lexical Ambiguity - Since these are scripts which includes characters which may have different phonetics when reading or writing, it is difficult to exactly understand what is written.

② Syntactic Ambiguity - Chinese and Arabic are languages which have characters instead of letters and the way of writing is also different for each of them, Chinese (left to right) and Arabic (right to left)

and hence it is difficult to understand their syntax

(3) Semantic Ambiguity - Both Chinese and Arabic are languages in which many sounds are mentioned with different characters while writing them as compared to the regular characters. Hence it becomes difficult to understand context in these languages as their pronunciation changes when they are spoken which may change the meaning altogether.

~~Remote~~ Suggestions to overcome these challenges are as follows:-

(1) Extensive training of the model on both of the languages to understand their complexities

(2) Train the model to get accustomed with the corpus or text of the languages to remove any syntactical ambiguity and unstable character understanding

(3) Train the model on the different characters which represent different sounds so that the model can identify these characters easily and make sense thereby removing semantic ambiguity

P.T.O.

SIU SIU SIU SIU SIU SIU SIU SIU SIU SIU SIU SIU SIU SIU SIU SIU SIU SIU SIU SIU SIU

Q2 Commercial use of NLP are as follows :-

(1) Chatbots - Chatbots are being used in customer service in various industries to enhance automation and reduce the resolution time of queries. Ex- Chatbots on Banking websites, IRCTC, Swiggy, Zomato, etc

(2) Text Summarization - Various tools such as ChatGPT, etc. are being used to help in summarizing the ~~text~~ data in various files and documents

(3) Virtual Assistants - Such as Alexa, Siri, Cortana, etc. are using NLP to understand the commands of the users and performing tasks accordingly

(4) Machine Translation - NLP also has its application in Machine Translation where it is converting commands given in human language to machine language for understanding and interpreting the results and then again giving results in human language. Eg. Google Assistant, Siri etc

(5) Audio to Text conversion - This is being used to give voice commands to do search where voice commands are being converted to text for doing search on search engines Ex google

Q5 Consistency Parsing refers to the processing of the text for understanding the data (NLU) and Dependency Parsing is done for generating NL data (NLG)

	Consistency Parsing	Dependency Parsing
(1)	This is done for processing an understanding of the data	(1) It helps in generating or creating the data
(2)	It is focuses on semantic part of the data for parsing	(2) It focuses on the syntactic part of the data for creating the text
(3)	It broadly performs the steps as a part of Natural Language Understanding	(3) It broadly performs the steps as a part of Natural language Generation
(4)	It aims to check for the consistency of logic and meaning of the data	(4) It aims to check for the syntax and syllables, tokens, so as to ensure correct generation of data
(5)	It aims to check for discourse analysis on the data	(5) It aims to check for pragmatic analysis on the data