Total Pages 12 (Twelve)

वसुधैव कुटुम्बकम्

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
27/10/24

No. 075662
SIU-24 E

Centre: SCMHRD

Seat No. (in figures) | 2 | 0 | 4 | 0 | 0 | 7 |

Seat No. (in words) two zero four zero zero seven

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

Programme : Business Analytics Semester : III

Course : Natural Language processing

Main Ans. Script 1 + No. of Supplementary Answer Scripts 1 =Total 2.

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 branch of computer science that make use of computational linguistics and statistics to understand the human language by the machines.

It has taught the Machines to understand, process and generate human language.

Computer science, Advanced modelling & Computational Linguistics make NLP.

Advanced Modelling can be statistical way or by using Artificial Intelligence like Machine Learning.

It is consist of Natural Language Understanding & Natural Language Generation.

In Natural
Machine
input text
meaning

Some of

Text class
Classifie
Eg: SPA
Sentimen
of the
ne
extract
words
the
on Sc

Benefits:
th
analyz
meanin
& un
This f
cre
This d
langu

Challen
were
valu
ov
other

In Natural Language understanding: Machine tries to break down the input text to understand the structure, meaning & context of the language.

Some of Real world Applications are:

Text classification: the input text is classified based on common features. Eg: SPAM Detection

Sentiment Analysis: The sentiments of the text such as positive, negative or neutral can be extracted by finding the sentiment words (positive, negative) & scoring the sentence, Document based on scores.

Benefits:

Helps to understand human Language, analyze & extract the visible & hidden meanings. Automates & helps in processing & understanding volumes of text. This process will help in Model creation. This also helps Machine understand Human Language.

Challenges:

* Ambiguity in Language, same words having different meanings.

* Complex languages, & huge volumes of data required for accurate model.

* Cannot recognize sarcasm and other subjective aspects of communication

Natural language Generation.
Machine generates the human language to Interact with the humans.

Some Real world Applications

Machine translation: The translation of text from one language to other.

Content Generation: The content is generated based on given prompt

Benefits:

- Reduces the Manual work of humans, used in text summarization.
- GPTs make content generation useful to interact with humans.
- Chatbots to generate & attend to human queries.

challenges:

- Depends on the training dataset, it may contain baisness in answers/replies
- Complex task requires computational power.

Q.2. Commercial uses Of NLP:

- Text-speech conversion } Speech
- Speech-to-text conversion } recognition
- text classification
- Sentiment Analysis
- Machine translation
- Word or sentence prediction/generation
- Text Summarization.
- ~~Speech recognition~~

~~language mo~~
- Document classification

Text-speech or Speech-text conversion:

NLP is used in conversion of 'text' of a language into speech using the inbuilt models of computer science.

In python, the following inbuit language models are used for text-speech conversion.

NLTK → Natural Language toolkit

spacy, Gensim are the other python packages that contains inbuilt models for text Analysis.

The broken down text is converted to digitized form into numbers then to audio signals, then passed to speaker for audio output.

Application in voice recognition, voice enabled automation.

here the human speech is converted to digital signal then to text then processed. The commands of voice is then executed.

Eg: siri, Alexa.

Python Package 'Sphinx' can do speech recognition.

Documents classification:

Documents which are collection of text into separate files can also be classified & categorized using nlp

Example, content of the document is rated, plagirism check, etc.

In plagiarism checker the words & sentence are checked with Documents found in Internet, if it matches then it flags & classifies as "plagirism detected".

Text summarization:

Text corpus is summarized to form the idea behind the text. It will help reduce the long text into short & precise format help us to label the group of text.

Q.3 Significance of POS tagging:

Parts of speech tagging used to tag the words to the grammatical categories, as Nouns, pronouns verbs, adverbs, adjectives, interjections and conjuction.

It is a disambiguation method used to address one of the major chalenges of NLP – Ambiguity.

The words and its context & role in the sentence is recognized then ambiguity is cleared.

collection of
s can also
ed using NLP.

document is
, etc.

words & sentences
s found in
n if flags &
deleted.

mmarized
d the text.
long
e format help
roup of text.

gging:

g used to
e grammatical
nouns, pronouns,
es, interjections

method used
major
mbiguity.

xt & role
gnized then

Example:
"I want to fly, like a fly!"

Here without POS tagging the word 'fly' is considered as a same entity though it has 2 meanings an insect fly and action fly.

Now when we do POS tagging using ~~NEP~~ ~~or~~ NLTK or spacey, we get.

I –	noun.	Here the 'fly' is seperated to verb & noun.
want –	verb	
to –	preposition.	
fly –	verb	
like –	verb.	
a –	Articles	
fly –	noun	

Thus ambigiety is rectified

POS also helps the data tobe applicable for NER (Named entity recognition)

~~Since the Noun - fly and verb-fly are separated, the NER will be more accurate.~~

Example:

Mr. plant's office is near the ~~B~~biogas plant, where coffee was planted

Here ~~Mr~~ Mr.plant → noun

biogas plant → noun

planted → verb.

Here NER since tagged from pos recognizer Mr. plant as a name

Pos also categorizes the word → (word category tagging) used to further syntactic Analysis

Q4. N-grams are sequence of n items such as words, letters and base pairs from sentences.

when N=1, unigrams

Eg: 'NLP', 'is', 'an', 'interesting', 'subject' → unigm

when N=2, Bigrams

Twin towers is a large building. Twin towers was constructed in 19th Century.

Here Twin towers → Bigrams as they always appear in a sentence together

Also, when we divide the words in a sontance in a pair of two or three or more, n-grams concept is used.

for same example:
NLP is an interesting subject

(NLP, is, an, interesting, subject) unigram

(NLP, is),
(is, an),
(interesting, subject),
etc.
} Bigrams.

N-grams come under statistical modelling of language modelling where the next word or sequence of words are predicted using their probability of occurance that comes from frequency of words.

Limitations:
N-grams cannot be applied for huge dataset as it makes the computation more complex.

Approach:
The next word prediction using n-gram here unigram,

(NLP, is, a, good, Subject) unigram.

P('NLP' | 'is', 'a', good', subject')

probablity of occurance of 'NLP' given the sequence of sentence is (is, a good Subject).

Here we saw, the model then trained for ~~new~~ data it also creates baisness in data as the training happens due to proximity of words

Cannot work on real life test data, unless trained on huge volumes of datasets

Q5.

Q6

Consistency parsing	Dependency parsing
☆ It identifies the constituents of the sentences from words. Into noun phrases verb phrases	☆ It helps to identify the relation (or) grammitical relation among the words in a sentence. example subject-verb relation
☆ Used in Natural Language understanding	☆ Used in Natural Language generation
☆ The tree (phrase tree) follows top down approach flows from root node to leaves	☆ The phrase tree follows bottom up approach The flows from leaves to root node.
☆ It follows tree structure with nodes → root node & leaves.	☆ It follows graph structure where nodes - words edges - relations.
☆ Extracts complex information & characteristics of words, of a language. Thus Rich Morphological language of rich Syntactic Structure can be used.	☆ extracts simple information & characteristics thus, poor morphological & poor syntacticaly structure langages like 'English' 'Chinese' are used.
☆ Old traditional model approach	☆ New Advanced model approach (machine Learning)

Q6. Application of Naïve Bayes classifier:

It is supervised Machine learning algorithm of probabilistic classification modelling make use of Bayes theorem to predict the classes of target variables. Used in classification. mare suitable in text classification.

Usage:

- Naive bayes used in high ~~[illegible]~~ dimentional features of data.
- Speed of prediction is high
- process & predict complex relations
- features should not be conditionally dependant & data must not contain missing values
- The features should be equaly weighted or given importance.
- ~~f~~ Continous features should be normally distributed & discrete features should be multinomial distributed.

Text classification with an Example:

multiple applications include:

- ~~Text classification~~
- Document Classification
- content rating
- SPAM Detection.
- Sentiments classification.

SPAM Detection:

Here the Naive bayes classifies the new email or SMS as a 'spam' or 'not spam', using training dataset

Total Pages 8 (Eight)

SYMBIOSIS INTERNATIONAL (DEEMED UNIVERSITY)

(Established under Section 3 of the UGC Act 1956)

Re-accredited by NAAC with 'A' grade (3.58/4) | Awarded Category - I by UGC

Verified all entries & found correct
Jr. Supervisor's Signature & Date
20/10/24

Supplement No. SIU-19 A
04432

Centre: ScMHRD

Seat No. (in figures) | 2 | 0 | 4 | 0 | 0 | 7 |

Seat No. (in words) twozero fourzero zero seven

Semester : III

Programme : BA

Course : NLP

Supplementary Answer Script No. I / ~~II / III~~ ✓

BEGIN WRITING HERE :

assuming the features of SPAM class,

i) unknown resource → yes, no

ii) suspicious email → yes, no

iii) Occurance of $\overset{\text{spam}}{\text{words}}$ → 'Discount', offer,

↳ 'Discount, offer, lottery'

So, using Bayes theorem, the new email has the following

~~P(Spam~~ features (unknown, suspicious, occurance)

test data ⇒ (yes, yes, offer) ⇒ New email

~~then~~

$$P\left(\frac{\text{spam}}{\text{new email}}\right) = \frac{P\left(\frac{\text{unknown}}{\text{spam}}\right) \times P\left(\frac{\text{suspicious}}{\text{spam}}\right) \times P\left(\frac{\text{occ}}{\text{spam}}\right) \times P(\text{spam})}{P(\text{new email})}$$

Similarly

we also calculate for $P\left(\frac{\text{Not spam}}{\text{new email}}\right)$

finally in this case $P\left(\frac{\text{spam}}{\text{new email}}\right) > P\left(\frac{\text{not spam}}{\text{new email}}\right)$

thus it Identifies as 'SPAM'

Similarly based on occorance of positive & negative words, sentiment of document is classified as positive or negative.

Q.7 Steps in text-preprocessing:

1. Tokenization:

The text preprocessing first starts with,

i) Docement triage

ii) text extraction/tokenization

Documents triage:

The doments, are converted to digitized format that can be processed by the computers software for Analysis, ~~and~~, stores them into corpus

Text extraction:

The text data in decoments are extracted & stored in machine or software accessable formats into object & variables.

Sentence tokenization:

The sentences are extracted & Stored in a list by breaking down a document or passage using delimiters

Importance:

Organizes the text into suitable format for text preprocessing. Breaking down also helps to analyze sentence wise or wordwise. Helps for further text Analysis.

word tokenization:

Breakdown of sentence into words and letters,

2. Lemmatization & Stemming.

Reduces the words into its base form that are meaningful words for Lemmatization.

Example : fly → fly
flew → fly
flown → fly
flying – fly.

~~Importan~~ If it reduces the words removing suffixes to meaningless words known as stemming.

Studying → studi

Importance: Reduces the complexity in words & different forms of words into their base forms reduces variance & errors.

3. Stopwords removal:
Removes filler words & insignificant words from the dataset

Importance: Removes words that does not add any meaning, reducing volume of dataset.

punctuations can also be removed in similar way.

4. POS & NER tagging:
the words then ~~to~~ categorized based on parts of speech and entities like name, place, & location can be extracted using Named entity recognition

Significance of text preprocessing:
Removes Bais & variance improves performance of model

Q.8. The main components of NLP
(NLU) Natural Language understanding:
It understands the human Language by breaking it down to machine readable & recognizable structure.

This stage consists of meaning extraction & understanding role & structure of words in sentences.

POS tagging} Parts of speech tagging helps to identify words as parts of speech removing ~~its~~ ambiguity

Semantic Analysis used to extract meaning & context.

Syntactic Analysis used to extract the grammatical structure.

Natural language Generation (NLG)
: While the model generates the language or text from the understanding of input data.

Text planning:
Plans the text → words that should be present in language generation.

Sentence planning: The sentence generation sequence & order of words planned

Realization: The structure of sentence words put into language that conveys meaningful message & generated

Q.9. Language Modelling:
It is the probabilistic Method of assigning probabilities to words in a sentence based on count or frequencies used to predict the ~~next~~ probability of next word or sequence of words generated.

Two Methods
Statistical Method
Neural Method.

Statistical Method Make use of probability to predict next word from sequence of words or input string.

More the concept of n-grams are used. n-grams are sequence of n items.

assuming are

n grams can be bigrams, trigrams, etc.

Now, Probablity of sentence = P(W)
Probabilty of words = $P(w_1, w_2, \ldots$

$P(W) = P(w_1, w_2, \ldots w_n)$

$P(W)$ & $P(w_1, w_2, \ldots w_n)$ is Language Modelling.

Using Markov, we

Neural Model: Neural Method

It overcomes the limitations of statistical Method, by using advanced & complex modelling done using Machine Learning or deep learning. But it also needs huge computation power & huge dataset to train.

Q.10 Languages like Chinese and Arabic does not have structure (or) order of words in a sentence.

The order of words in a sentence does not change the meaning of sentence.

Out of them these are orthographic Language.

[illegible] input sence of string next [illegible] → P(w) = P([illegible])

Natural → Aty, complex, ML, [illegible] translation of [illegible]

Lang. model: Prob. [illegible] of [illegible]

joint Prob. [illegible] word prob.

12

EXAMPLES

Rough. ALL

(15 [illegible])

Text class

content generation

Speech suggest

Sentiment

machine translations.

① Computational linguistics.

comp science advanced (ML AI) / machine

② Machine translate

→ Google translate

→ chatbots.

→ SPAM Detection

Naive Bayes

→ Speech recognition

⑧

NLU	NLG
→ POS → tokenization → Semantic mng.	text planning sentence plan Realization

③ disambiguation

word categorising [illegible]

Syntactic analysis

linguistics

noun adj adverb.

(Radius NER)

7) Document triage & [illegible] technique

· Tokenization

– Sentence

– word – tree

Loading lang model NLTK,

→ [illegible]

→ S/o

statistical → Ngram prob. of [illegible]

Lang → Neural

④ angram

unigram. eg.

sequence of n items letters [illegible] base pairs

find so predict next.

string of words.

process of conversion of raw text into digitized

Documents

Importance

– understanding [illegible]

→ pos → remove Ambiguity

→ Readable data for ML & data Anlys

→ [illegible] ↑ Performance of model

Statistical & limitations

cannot applied huge data & complex relations.

⑩ selection words & algorithm sub-word

NLU basic

Graph struc. [illegible] edge

G = (V,E) [illegible] info & [illegible]

[illegible] simple

advanced → ML → [illegible] M/c translation.

⑤ constituents

1) NP VP

NLU

3) TOP down

mountain tree

2) [illegible] root [illegible]

4) tree struc [illegible]

5) extract complex info

6) old traditional [illegible]

7)

Supervised ML algorithm based on probab.

prob. class → text classification

Bayes theorem.

⑥ App. of Naive Bayes. →

Assumptions

ideal

text class

high dimension features complex.

speed

[illegible]

confusion matrix.

Y/N

→ Spam detection.

→ Document classification.

sentiment classification

→ content rating.

10) logographic

alphabetic

Syllabic

Ideographic

→ Sees not have order [illegible] to words