INDIAN SCHOOL AL WADI AL KABIR

Woksheet No:1	ARTIFICIAL INTELLIGENCE (417) CLASS X CHAPTER 4: NATURAL LANGUAGE PROCESSING

1 Mark questions

1. Which technique is used to assess the meaningfulness of the input text?
   a. Pragmatic analysis b. Lexical Analysis c. **Semantic analysis** d. Discourse Integration

2. Natural Language Processing majorly deals with_____ processing.
   a. Numeric data b. **Textual data** c. Image data d. Visual data

3. What is the first stage of Natural Language Processing (NLP)?
   a. Semantic Analysis b. Pragmatic Analysis c. **Lexical Analysis** d. Syntactic Analysis

4. Words that we want to filter out before doing any analysis of the text are called_____.
   a. Rare words b. **Stop words** c. Frequent words d. Filter words

5. What does discourse integration involve in the context of sentence formation?
   a. Identifying individual words in a sentence
   b. Forming a coherent story within a sentence
   c. **Establishing relationships between preceding and succeeding sentences**
   d. Applying punctuation and grammar rules to a sentence

6. The _____ domain of Artificial Intelligence, that is focused on enabling computers to understand and process human languages.
   a. Data Science b. Computer Vision c. **Natural Language Processing** d. None of these

7. Which of the following is an application of Natural Language Processing?
   a. Automatic Summarization b. Sentiment Analysis c. Text classification **d. All of these**

8. _____are software applications that mimic written or spoken human speech for the purposes of simulating a conversation or interaction with a real person
   a. Face filter b. Google lens c. **chatbots** d. Facial recognition

9. _____ work around a script which is programmed in them
   a. smart bots b. chat bots c. **script bots** d. robots

10. _____refers to the grammatical structure of a sentence
   a. Syntax b. Semantics c. Object detection d. Normalization

11. Pick the odd one out.
   a. Stemming **b. Localization** c. Lemmatization d. removal of stop words.

12. _____ is a term used for any word or number or special character occurring in a sentence.
   a. corpus b. statement c. **token** d. stop word

13. Which algorithms result in two things, a vocabulary of words and frequency of the words in the corpus?
    a. Sentence segmentation b. Tokenisation **c. Bag of words** d. Text normalisation

14. Which feature of NLP helps in understanding the emotions of the people mentioned with the feedback?
(a) Virtual Assistants (b) **Sentiment Analysis** (c) Text classification (d) Automatic Summarization

15. What will be the output after stemming the word studies?
    a. study b. studie c. **studi** d. none of these

16. What is the primary challenge faced by computers in understanding human languages?

    a. Complexity of human languages

    b. Lack of computational power

    c. Incompatibility with numerical data

    d. Limited vocabulary

17. How do voice assistants utilize NLP?

    a. To analyze visual data b. To process numerical data

    c. **To understand natural language** d. To execute tasks based on computer code

18. Which of the following is NOT a step in Text Normalisation?

    a.Tokenization b. Lemmatization c. Punctuation removal d. **Document summarization**

19. In the context of text processing, what is the purpose of tokenisation?

    a. To convert text into numerical data
    b. To segment sentences into smaller units
    c. To translate text into multiple languages
    d. To summarize documents for analysis

20. What distinguishes lemmatization from stemming?
    a. **Lemmatization produces meaningful words after affix removal, while stemming does not**.
    b. Lemmatization is faster than stemming.
    c. Stemming ensures the accuracy of the final word.
    d. Stemming generates shorter words compared to lemmatization.

21. What is the primary purpose of the Bag of Words model in Natural Language Processing?
    a. To translate text into multiple languages
    b. To extract features from text for machine learning algorithms
    c. To summarize documents for analysis
    d. To remove punctuation marks from text

22. In the context of text processing, what are stop words?
    a. Words with the frequent occurrence in the corpus
    b. Words with negligible value that are often removed during preprocessing
    c. Words with the lowest occurrence in the corpus
    d. Words with the most value added to the corpus

23. What is the characteristic of rare or valuable words in the described plot?
    a. They have the highest occurrence in the corpus
    b. They are often considered stop words
    c. They occur the least but add the most value to the corpus
    d. They are typically removed during preprocessing

24. What information does the document vector table provide?
    a. The frequency of each word across all documents
    b. The frequency of each word in a single document
    c. The total number of words in the entire corpus
    d. The average word length in the entire corpus

25. What is the primary purpose of TFIDF in text processing?
    a. To identify the presence of stop words in documents
    b. To remove punctuation marks from text
    c. To identify the value of each word in a document
    d. To translate text into multiple languages

2 Mark questions

26. Write note on Natural language processing

Natural Language Processing (commonly called NLP) takes in the data of Natural Languages which humans use in their daily lives and operates on this. Natural Language Processing, or NLP, is the sub-field of AI that is focused on enabling computers to understand and process human languages.

27.Write notes on text classification and virtual assistants.

Text classification makes it possible to assign predefined categories to a document and organize it to help you find the information you need or simplify some activities. For example, an application of text categorization is spam filtering in email.

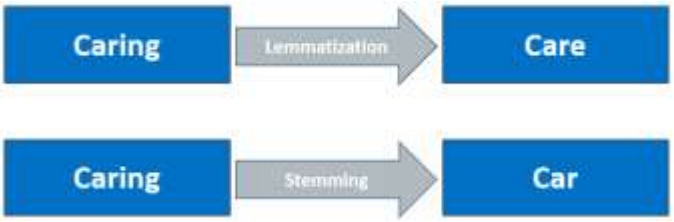
Virtual Assistants: Accessing our data, helps us in keeping notes of our tasks, make calls for us, send messages and a lot more. With the help of speech recognition, these assistants can not only detect our speech but can also make sense out of it. Ex: Google Assistant, Cortana, Siri, Alexa, etc

28. Differentiate Stemming and Lemmatization

Stemming: is the process in which the affixes of words are removed and the words are converted to their base form which is known as **Stem**. Stemming does not consider if the stemmed word is meaningful or not It just removes the affixes hence it is faster.

Lemmatization: is the process in which the affixes of words are removed and the words are converted to their base form which is known as **lemma**. Lemmatization makes sure that lemma is a word with meaning and hence it takes a longer time to execute than stemming.

Eg:



29. Identify any two stop words which should not be removed from the given sentence and why?

Stopwords in the given sentence which should not be removed are:
@, . (fullstop),_(underscore) , 123(numbers) These tokens are generally considered as stopwords, but in the above sentence, these tokens are part of email id. removing these tokens may lead to invalid website address and email ID. So these words should not be removed from the above sentence.
(1 mark for identifying any two stop words from the above, and 1 mark for the valid justification.)

30. Write note on Bag of Words Algorithm.

Bag of Words is a Natural Language Processing model which helps in extracting features out of the text which can be helpful in machine learning algorithms.

As we put The normalized corpus which we have got after going through all the steps of text processing, into the bag of words algorithm, the algorithm returns to us the unique words out of the corpus and their occurrences in it. Thus, we can say that the bag of words gives us two things:

1. A vocabulary of words for the corpus

2. The frequency of these words (number of times it has occurred in the whole corpus).

31. Describe any four applications of TFIDF.

Document Classification	Topic Modelling	Information Retrieval System	Stop word filtering
Helps in classifying the type and genre of a document.	It helps in predicting the topic for a corpus.	To extract the important information out of a corpus.	Helps in removing unnecessary words from a text body.

4 Marks Questions

32. Samiksha, a student of class X was exploring the Natural Language Processing domain. She got stuck while performing the text normalization. Help her to normalize the text on the segmented sentences given below:

Document 1: Akash and Ajay are best friends

Document 2: Akash likes to play football but Ajay prefers to play online games

1. Tokenisation

Akash, and, Ajay, are, best, friends, Akash, likes, to, play, football, but, Ajay, prefers, to, play, online, games

2. Removal of stopwords

Akash, Ajay, best, friends, Akash, likes, play, football, Ajay, prefers, play, online, games

3. Converting text to a common case

akash, ajay, best, friends akash, likes, play, football, ajay, prefers, play, online, games

4. Stemming/Lemmatisation

akash, ajay, best, friend, akash, like, play, football, ajay, prefer, play, online, game

(1 mark for each step; 1*4=4)

33. Differentiate script bot and smart bot.

Script-bot	Smart- bot
Script bots are easy to make	Smart-bots are flexible and powerful
Script bots work around a script which is programmed in them	Smart bots work on bigger databases and other resources directly
Mostly they are free and are easy to integrate to a messaging platform	Smart bots learn with more data
No or little language processing skills	Coding is required to take this up on board
Limited functionality	Wide functionality

34. **Through a step-by-step process, calculate TFIDF for the given corpus Document 1: Johny Johny Yes Papa, Document 2: Eating sugar? No Papa Document 3: Telling lies? No Papa**
Document 4: Open your mouth, Ha! Ha! Ha!

1) Tokenized documents

1. Doc1: `johny johny yes papa` → tokens = [johny, johny, yes, papa] (4 tokens)
2. Doc2: `eating sugar no papa` → tokens = [eating, sugar, no, papa] (4 tokens)
3. Doc3: `telling lies no papa` → tokens = [telling, lies, no, papa] (4 tokens)
4. Doc4: `open your mouth ha ha ha` → tokens = [open, your, mouth, ha, ha, ha] (6 tokens)

Vocabulary (unique terms):
`johny, yes, papa, eating, sugar, no, telling, lies, open, your, mouth, ha`

---

2) Document frequencies (df) and IDF

N = 4 documents.

- df(johny) = 1 → IDF = ln(4/1) = ln(4) = **1.38629436112**
- df(yes) = 1 → IDF = ln(4) = **1.38629436112**
- df(papa) = 3 → IDF = ln(4/3) = **0.28768207245**
- df(eating) = 1 → IDF = ln(4) = **1.38629436112**
- df(sugar) = 1 → IDF = ln(4) = **1.38629436112**
- df(no) = 2 → IDF = ln(4/2) = ln(2) = **0.69314718056**
- df(telling) = 1 → IDF = ln(4) = **1.38629436112**
- df(lies) = 1 → IDF = ln(4) = **1.38629436112**
- df(open) = 1 → IDF = ln(4) = **1.38629436112**
- df(your) = 1 → IDF = ln(4) = **1.38629436112**
- df(mouth) = 1 → IDF = ln(4) = **1.38629436112**
- df(ha) = 1 → IDF = ln(4) = **1.38629436112**

(kept high precision for intermediate steps)

---

3) TF (normalized) for terms present in each doc

Doc1 (4 tokens):

- TF(johny, D1) = 2 / 4 = **0.5**
- TF(yes, D1) = 1 / 4 = **0.25**
- TF(papa, D1) = 1 / 4 = **0.25**

Doc2 (4 tokens):

- TF(eating, D2) = 1 / 4 = **0.25**
- TF(sugar, D2) = 1 / 4 = **0.25**
- TF(no, D2) = 1 / 4 = **0.25**
- TF(papa, D2) = 1 / 4 = **0.25**

Doc3 (4 tokens):

- TF(telling, D3) = 1 / 4 = **0.25**
- TF(lies, D3) = 1 / 4 = **0.25**
- TF(no, D3) = 1 / 4 = **0.25**
- TF(papa, D3) = 1 / 4 = **0.25**

Doc4 (6 tokens):

- TF(open, D4) = 1 / 6 ≈ **0.16666666667**
- TF(your, D4) = 1 / 6 ≈ **0.16666666667**
- TF(mouth, D4) = 1 / 6 ≈ **0.16666666667**
- TF(ha, D4) = 3 / 6 = **0.5**

---

4) TF–IDF = TF × IDF (selected nonzero entries)

Doc 1

- TF–IDF(johny, D1) = 0.5 × 1.38629436112 = **0.69314718056**
- TF–IDF(yes, D1) = 0.25 × 1.38629436112 = **0.34657359028**
- TF–IDF(papa, D1) = 0.25 × 0.28768207245 = **0.07192051811**

Doc 2

- TF–IDF(eating, D2) = 0.25 × 1.38629436112 = **0.34657359028**
- TF–IDF(sugar, D2) = 0.25 × 1.38629436112 = **0.34657359028**
- TF–IDF(no, D2) = 0.25 × 0.69314718056 = **0.17328679514**
- TF–IDF(papa, D2) = 0.25 × 0.28768207245 = **0.07192051811**

Doc 3

- TF–IDF(telling, D3) = 0.25 × 1.38629436112 = **0.34657359028**
- TF–IDF(lies, D3) = 0.25 × 1.38629436112 = **0.34657359028**
- TF–IDF(no, D3) = 0.25 × 0.69314718056 = **0.17328679514**

- TF–IDF(papa, D3) = 0.25 × 0.28768207245 = **0.07192051811**

Doc 4

- TF–IDF(open, D4) = (1/6) × 1.38629436112 ≈ **0.23104906019**
- TF–IDF(your, D4) = (1/6) × 1.38629436112 ≈ **0.23104906019**
- TF–IDF(mouth, D4) = (1/6) × 1.38629436112 ≈ **0.23104906019**
- TF–IDF(ha, D4) = 0.5 × 1.38629436112 = **0.69314718056**

---

5) Compact TF–IDF matrix (terms × docs, zeros omitted)

- D1: {johny: 0.693147181, yes: 0.346573590, papa: 0.071920518}
- D2: {eating: 0.346573590, sugar: 0.346573590, no: 0.173286795, papa: 0.071920518}
- D3: {telling: 0.346573590, lies: 0.346573590, no: 0.173286795, papa: 0.071920518}
- D4: {open: 0.231049060, your: 0.231049060, mouth: 0.231049060, ha: 0.693147181}

****