NATURAL LANGUAGE PROCESSING

OBJECTIVE QUESTIONS (SET 01)

1. NLP stands for _____.
a. Natural Language Processing
b. Nature Language Processing
c. None Language Processing
d. None of the above
Ans: a. Natural Language Processing

2. _____, is the sub-field of AI that is focused on enabling computers to understand and process human languages.
a. Natural Language Processing
b. Data Science
c. Computer Vision
d. None of the above
Ans: a. Natural Language Processing

3. _____ is the sub-field of AI that make the interactions between computers and human (natural) languages
a. Natural Language Processing
b. Data Science
c. Computer Vision
d. None of the above
Ans: a. Natural Language Processing

4. Which of the games below is related to natural language processing?
a. Voice Assistants
b. Chatbots
c. Mystery Animal
d. Grammar Checkers
Ans: c. Mystery Animal

5. Applications of Natural Language Processing
a. Automatic Summarization
b. Sentiment Analysis
c. Text Classification
d. All of the above
Ans: d. All of the above

6. _____ Information overload is a real problem when we need to access a specific, important piece of information from a huge knowledge base.
a. Automatic Summarization
b. Sentiment Analysis
c. Text Classification
d. All of the above
Ans: a. Automatic Summarization

7. _____ is especially relevant when used to provide an overview of a news item or blog post, while avoiding redundancy from multiple sources and maximizing the diversity of content obtained.
a. Automatic Summarization
b. Sentiment Analysis
c. Text Classification

d. All of the above
Ans: a. Automatic Summarization

8. The goal of sentiment analysis is to identify sentiment among several posts or even in the same post where emotion is not always explicitly expressed.
a. Automatic Summarization
b. Sentiment Analysis
c. Text Classification
d. All of the above
Ans: b. Sentiment Analysis

9. Companies use Natural Language Processing applications, such as _____, to identify opinions and sentiment online to help them understand what customers think about their products and services
a. Automatic Summarization
b. Sentiment Analysis
c. Text Classification
d. All of the above
Ans: b. Sentiment Analysis

10. _____ 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.
a. Automatic Summarization
b. Sentiment Analysis
c. Text Classification
d. All of the above
Ans: c. Text Classification

11. _____ device helps to communicate with humans and abilities to make humans lives easier.
a. Google Assistant
b. Cortana
c. Siri
d. All of the above
Ans: d. All of the above

12. _____ is all about how machines try to understand and interpret human language and operate accordingly.
a. Natural Language Processing
b. Data Science
c. Computer Vision
d. None of the above
Ans: a. Natural Language Processing

13. By dividing up large problems into smaller ones, _____ aims to help you manage them in a more constructive manner.
a. CDP
b. CBT
c. CSP
d. CLP
Ans: b. CBT

14. CBT stands for _____.
a. Common Behavioural Therapy (CBT)
b. Cognitive Behavioural Therapy (CBT)
c. Connection Behavioural Therapy (CBT)

d. None of the above
Ans: b. Cognitive Behavioural Therapy (CBT)

15. Cognitive behavioural Therapy includes _____.
a. Your Thoughts
b. Your Behaviors
c. Your Emotions
d. All of the above
Ans: d. All of the above

16. _____ is considered to be one of the best methods to address stress as it is easy to implement on people and also gives good results.
a. Common Behavioural Therapy (CBT)
b. Cognitive Behavioural Therapy (CBT)
c. Connection Behavioural Therapy (CBT)
d. None of the above
Ans: b. Cognitive Behavioural Therapy (CBT)

17. _____ by collecting data from various reliable and authentic sources.
a. Data Acquisition
b. Database
c. Data Mining
d. None of the above
Ans: a. Data Acquisition

18. Once the textual data has been collected, it needs to be processed and cleaned so that an easier version can be sent to the machine. This is known as _____.
a. Data Acquisition
b. Data Exploration
c. Data Mining
d. None of the above
Ans: b. Data Exploration

19. Once the text has been normalized, it is then fed to an NLP based AI model. Note that in NLP, modelling requires data pre-processing only after which the data is fed to the machine.
a. Data Acquisition
b. Modelling
c. Data Mining
d. None of the above
Ans: b. Modelling

20. The model trained is then evaluated and the accuracy for the same is generated on the basis of the relevance of the answers which the machine gives to the user's responses.
a. Data Acquisition
b. Modelling
c. Evaluation
d. None of the above
Ans: c. Evaluation

21. One of the most common applications of Natural Language Processing is a chatbot, give some examples of chatbots _____.
a. Mitsuku Bot
b. CleverBot
c. Jabberwacky

d. All of the above
Ans: d. All of the above

22. There are _____ different types of chatbots.
a. 2
b. 3
c. 4
d. 5
Ans: a. 2

23. Which of the following is related to chatbots.
a. Script-bot
b. Smart-bot
c. Both a) and b)
d. None of the above
Ans: c. Both a) and b)

24. _____ bots work around a script which is programmed in them.
a. Script-bot
b. Smart-bot
c. Both a) and b)
d. None of the above
Ans: a. Script-bot

25. _____ work on bigger databases and other resources directly.
a. Script-bot
b. Smart-bot
c. Both a) and b)
d. None of the above
Ans: b. Smart-bot

26. _____ helps in cleaning up the textual data in such a way that it comes down to a level where its complexity is lower than the actual data.
a. Speech Normalization
b. Text Normalization
c. Visual Normalization
d. None of the above
Ans: b. Text Normalization

27. _____ the whole corpus is divided into sentences. Each sentence is taken as a different data so now the whole corpus gets reduced to sentences.
a. Sentence Normalization
b. Sentence Segmentation
c. Sentence Tokenization
d. All of the above
Ans: b. Sentence Segmentation

28. Under _____, every word, number and special character is considered separately and each of them is now a separate token.
a. Tokenization
b. Token normalization
c. Token segmentation
d. All of the above
Ans: a. Tokenization

29. In Tokenization each sentence is divided into _____.
a. Block
b. Tokens
c. Parts
d. None of the above
Ans: b. Tokens

30. _____ are the words which occur very frequently in the corpus but do not add any value to it.
a. Tokens
b. Words
c. Stopwords
d. None of the above
Ans: c. Stopwords

31. Stopwords are the words which occur very frequently in the corpus but do not add any value to it. for example_____.
a. Grammatical words
b. Simple words
c. Complex words
d. All of the above
Ans: a. Grammatical words

32. Applications of TFIDF are _____.
a. Document Classification
b. Topic Modelling
c. Information Retrieval System and Stop word filtering
d. All of the above
Ans: d. All of the above

33. The machine does not consider _____words as same words because of different cases.
a. Upper case
b. Lower case
c. Case sensitivity
d. None of the above
Ans: c. Case sensitivity

34. _____ is the process in which the affixes of words are removed and the words are converted to their base form.
a. Stemming
b. Stopwords
c. Case-sensitivity
d. All of the above
Ans: a. Stemming

35. Stemming and lemmatization both are _____ processes.
a. Same process
b. Alternative process
c. Other process
d. All of the above
Ans: b. Alternative process

36. _____ makes sure that lemma is a word with meaning and hence it takes a longer time to execute than stemming.
a. Stopwords
b. Stemming
c. Lemmatization
d. Token normalization
Ans: c. Lemmatization

37. _____ is a Natural Language Processing model which helps in extracting features out of the text which can be helpful in machine learning algorithms.
a. Bag of Words
b. Big Words
c. Best Words
d. All of the above
Ans: a. Bag of Words

38. Which steps we have to approach to implement the bag of words algorithm.
a. Text Normalization
b. Create Dictionary
c. Create Document Vectors
d. All of the above
Ans: d. All of the above

39. _____ identify each document in the corpus, find out how many times the word from the unique list of words has occurred.
a. Text Normalization
b. Create Dictionary
c. Document Vectors
d. All of the above
Ans: c. Document Vectors

40. TFIDF stands for _____.
a. Team Frequency and Inverse Document Frequency
b. Term Frequency and Inverse Document Frequency
c. Top Frequency and Inverse Document Frequency
d. Table Frequency and Inverse Document Frequency
Ans: b. Term Frequency and Inverse Document Frequency

OBJECTIVE QUESTIONS (SET 02)

Q1. NLP stands for _____
a. New Language Processing
b. Number Language Processing
c. Natural Language Processing
d. Neural Language Processing
Ans: c. Natural Language Processing

Q2. Which of the following is not the domain of Artificial Intelligence?
a. Data Science
b. Computer Vision
c. NLP
d. Data Vision
Ans: d. Data Vision

Q3. Which of the following domain work around numbers and tabular data?
a. Computer Vision
b. Data Science
c. NLP
d. None of the above
Ans: b. Data Science

Q4. ____ is all about visual data like images and videos.
a. Computer Vision
b. Data Science
c. NLP
d. None of the above
Ans: a. Computer Vision

Q5. What is NLP?
a. It works around numbers and tabular data.
b. It is all about visual data like images and videos.
c. It takes in the data of Natural Languages which humans use in their daily lives.
d. None of the above
Ans: c. It takes in the data of Natural Languages which humans use in their daily lives.

Q6. Which of the following is not correct about NLP?
a. It is a sub field of AI.
b. It is focused on enabling computers to understand and process human languages.
c. It takes in the data of Natural Languages which humans use in their daily lives.
d. None of the above
Ans: d. None of the above

Q7. Applications of Natural Language Processing is _____
a. Automatic Summarization
b. Sentiment Summarization
c. Text Summarization
d. All of the above
Ans: a. Automatic Summarization

Q8. Which of the following will help to access a specific, important piece of information from a huge knowledge base.
a. Sentiment Analysis
b. Text classification
c. Virtual Assistants
d. Automatic Summarization
Ans: d. Automatic Summarization

Q9. Automatic summarization is relevant _____
a. for summarizing the meaning of documents and information.
b. to understand the emotional meanings within the information, such as in collecting data from social media.
c. to provide an overview of a news item or blog post.
d. All of the above
Ans: d. All of the above

Q10. The goal of _____ is to identify sentiment among several posts.
a. Sentiment Analysis

b. Automatic Summarization
c. Text classification
d. Virtual Assistants
Ans: a. Sentiment Analysis

Q11. One of the applications of Natural Language Processing is relevant when used to provide an overview of a news item or blog post, while avoiding redundancy from multiple sources and maximizing the diversity of content obtained. Identify the application from the following
a. Sentiment Analysis
b. Virtual Assistants
c. Text classification
d. Automatic Summarization
Ans: d. Automatic Summarization

Q12. Companies use _____ application of NLP , to identify opinions and feelings/emotions online to help them understand what customers think about their products and services.
a. Sentiment Analysis
b. Automatic Summarization
c. Text classification
d. Virtual Assistants
Ans: a. Sentiment Analysis

Q13. _____ understands point of view in context to help better understand what's behind an expressed opinion.
a. Sentiment Analysis
b. Automatic Summarization
c. Text classification
d. Virtual Assistants
Ans: a. Sentiment Analysis

Q14. Which of the following 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.
a. Sentiment Analysis
b. Automatic Summarization
c. Text classification
d. Virtual Assistants
Ans: c. Text classification

Q15. Which of the following is used in spam filtering in E-mail?
a. Sentiment Analysis
b. Automatic Summarization
c. Text classification
d. Virtual Assistants
Ans: c. Text classification

Q16. Which of the following is not a Virtual Assistant?
a. Alexa
b. Cortana
c. Siri
d. Silvi
Ans: d. Silvi

Q17. _____ is a virtual assistant software application developed by Google.
a. Alexa

b. Cortana
c. Google Assistant
d. Siri
Ans: c. Google Assistant

Q18. Which of the following is a virtual assistant developed by Microsoft?
a. Siri
b. Cortana
c. Google Assistant
d. Alexa
Ans: b. Cortana

Q19. What a virtual assistants can do?
a. They can help us in keeping notes of our tasks.
b. They can make calls for us.
c. They can send messages for us.
d. All of the above.
Ans: d. All of the above.

Q20. Which out of the following is the first virtual assistant?
a. Alexa
b. Siri
c. Cortana
d. Google Assistant
Ans: b. Siri

Q21. Which of the following is an example of a voice based virtual assistant?
a. Alexa
b. Siri
c. Cortana
d. All of the above
Ans: d. All of the above

Q22. _____ is all about how machines try to understand and interpret human language and operate accordingly.
a. Natural Language Processing
b. Data Science
c. Computer Vision
d. None of the above
Ans: a. Natural Language Processing

Q23. Now a days a lot of cases are coming where people are depressed due to reasons like _____
a. Peer Pressure
b. Studies
c. Relationship
d. All of the above
Ans: d. All of the above

Q24. _____is considered to be one of the best methods to address stress as it is easy to implement on people and also gives good results.
a. CAD
b. CBT
c. CBD
d. CAM

Ans: b. CBT

Q25. CBT stands for _____
a. Cognitive Behavioural Therapy
b. Common Behavioural Therapy
c. Creative Behavioural Therapy
d. Clear Behavioural Therapy
Ans: a. Cognitive Behavioural Therapy

Q26. Cognitive Behavioural Therapy includes _____
a. understanding the behaviour of a person in their normal life.
b. understanding the emotions of a person in their normal life.
c. understanding the mindset of a person in their normal life
d. All of the above
Ans: d. All of the above

Q27. _____ is a technique used by most therapists to cure patients out of stress and depression.
a. CTB
b. CBD
c. CBT
d. BCT
Ans: c. CBT

Q28. People who are going through stress will contact _____
a. Psychiatrist
b. Physician
c. Radiologist
d. None of the above
Ans: a. Psychiatrist

Q29. To understand the sentiments of people, we need to collect their conversational data so the machine can interpret the words that they use and understand their meaning. This step is coming under _____
a. Problem Scoping
b. Data Acquisition
c. Data Exploration
d. Modelling
Ans: b. Data Acquisition

Q30. We can collect data by _____
a. Surveys
b. Databases available on the internet.
c. Interviews
d. All of the above
Ans: d. All of the above

Q31. The most common applications of Natural Language Processing is _____
a. Bat Ball
b. Chat Bot
c. Ro Bot
d. Talk Bot
Ans: b. Chat Bot

Q32. Which of the following is a chat bot?

a. Mitsuku Bot
b. CleverBot
c. Jabberwacky
d. All of the above
Ans: d. All of the above

Q33. Type of chat bot is _____
a. Script bot
b. Smart bot
c. Both of the above
d. None of the above
Ans: c. Both of the above

Q34. _____ work around a script which is programmed in them.
a. Script bots
b. Smart bots
c. Both of the above
d. None of the above
Ans: a. Script bots

Q35. _____ learn with more data.
a. Script bots
b. Smart bots
c. Both of the above
d. None of the above
Ans: b. Smart bots

Q36. Example of smart bot is _____
a. Alexa
b. Siri
c. Cortana
d. All of the above
Ans: d. All of the above

Q37. Which of the following is not an example of smart bot?
a. Siri
b. Google Assistant
c. Cortana
d. Bots which are deployed in the customer care section and answer the basic queries.
Ans: d. Bots which are deployed in the customer care section and answer the basic queries.

Q38. Our ____ keeps on processing the sounds that it hears around itself and tries to make sense out of them all the time.
a. Eyes
b. Mouth
c. Brain
d. Ear
Ans: c. Brain

Q39. Computer understands the language of _____
a. numbers
b. alphabets
c. date
d. None of the above

Ans: a. numbers

Q40. The communications made by the machines are very basic and simple. (T/F)
a. True
b. False
Ans: a. True

Q41. There are _____ types of chatbot.
a. 1
b. 2
c. 3
d. 4
Ans: b. 2

Q42. A ____ is a software or computer program that simulates human conversation.
a. Chat bot
b. Robot
c. Talk bot
d. Chatter
Ans: a. Chat bot

Q43. The possible difficulties a machine would face in processing natural language is _____
a. Arrangement of the words and meaning
b. Multiple Meanings of a word
c. Perfect Syntax, no Meaning
d. All of the above
Ans: d. All of the above

Q44. Syntax refers to the ___ of a sentence
a. Grammatical structure
b. Hindi meaning
c. Pronunciation
d. None of the above
Ans: a. Grammatical structure

Q45. _____ allows the computer to identify the different parts of a speech.
a. part-of tagging.
b. part-of-sound tagging.
c. part-of-speech tagging.
d. part-of-speak tagging.
Ans: c. part-of-speech tagging.

Q46. The following line refers to _____
5 + 6 = 6 + 5
a. Different syntax, same semantics
b. Same syntax, Different semantics
c. Different syntax, Different semantics
d. None of the above
Ans: a. Different syntax, same semantics

Q47. The following line refers to _____
2/3 (Python 2.7) ≠ 2/3 (Python 3)
a. Different syntax, same semantics
b. Same syntax, Different semantics

c. Different syntax, Different semantics
d. None of the above
Ans: b. Same syntax, Different semantics

Q48. Semantics refers to _____
a. grammar of the statement
b. meaning of the statement
c. Both of the above
d. None of the above
Ans: b. meaning of the statement

Q49. In _____ it is important to understand that a word can have multiple meanings and the meanings fit into the statement according to the context of it.
a. Natural Language
b. Computer language
c. Machine Language
d. None of the above
Ans: a. Natural Language

Q50. In Human language, a perfect balance of _____ is important for better understanding.
a. Syntax
b. Semantics
c. Both of the above
d. None of the above
Ans: c. Both of the above

Q51. _____ helps in cleaning up the textual data in such a way that it comes down to a level where its complexity is lower than the actual data.
a. Data Normalisation
b. Text Normalisation
c. Number Normalisation
d. Table Normalisation
Ans: b. Text Normalisation

Q52. The term used for the whole textual data from all the documents altogether is known as _____
a. Complete Data
b. Slab
c. Corpus
d. Cropus
Ans: c. Corpus

Q53. Which of the following is the first step for Text Normalisation?
a. Tokenisation
b. Sentence Segmentation.
c. Removing Stopwords, Special Characters and Numbers.
d. Converting text to a common case.
Ans: b. Sentence Segmentation.

Q54. In _____ the whole corpus is divided into sentences.
a. Tokenisation
b. Sentence Segmentation
c. Removing Stopwords, Special Characters and Numbers
d. Converting text to a common case
Ans: b. Sentence Segmentation.

Q55. In Tokenisation each sentence is then further divided into _____
a. Token
b. Character
c. Word
d. Numbers
Ans: a. Token

Q56. Under _____, every word, number and special character is considered separately and each of them is now a separate token.
a. Sentence Segmentation
b. Removing Stopwords, Special Characters and Numbers
c. Converting text to a common case
d. Tokenisation
Ans: d. Tokenisation

Q57. _____ are the words which occur very frequently in the corpus but do not add any value to it.
a. Special Characters
b. Stopwords
c. Roman Numbers
d. Useless Words
Ans: b. Stopwords

Q58. Which of the following is an example of stopword?
a. a
b. an
c. and
d. All of the above
Ans: d. All of the above

Q59. During Text Normalisation, which step will come after removing Stopwords, Special Characters and Numbers.
a. Converting text to a common case.
b. Stemming
c. Lemmatization
d. Tokenisation
Ans: a. Converting text to a common case.

Q60. During Text Normalisation, when we convert the whole text into a similar case, we prefer _____
a. Upper Case
b. Lower Case
c. Title Case
d. Mixed Case
Ans: b. Lower Case

Q61. _____ is the process in which the affixes of words are removed and the words are converted to their base form.
a. Lemmatization
b. Stemming
c. Both of the above
d. None of the above
Ans: c. Both of the above

Q62. After stemming, the words which we get after removing the affixes is called _____
a. Stemmed Words
b. Stemma
c. Fruit Word
d. Shoot Word
Ans: a. Stemmed Words

Q63. While stemming healed, healing and healer all were reduced to _____
a. heal
b. healed
c. heale
d. hea
Ans: a. heal

Q64. While stemming studies was reduced to _____ after the affix removal.
a. studi
b. study
c. stud
d. studys
Ans: a. studi

Q65. After Lemmatization, the words which we are get after removing the affixes is called _____
a. Lemmat
b. Lemma
c. Lemmatiz
d. Lemmatiza
Ans: b. Lemma

Q66. Which of the following statement is not correct?
a. Lemmatization makes sure that lemma is a word with meaning.
b. Lemmatization takes a longer time to execute than stemming.
c. Stemmed word is always meaningful.
d. Both Stemming and lemmatization process remove the affixes.
Ans: c. Stemmed word is always meaningful.

Q67. _____ is a Natural Language Processing model. In this we get the occurrences of each word and construct the vocabulary for the corpus.
a. Bag of Words
b. Bag of Alphabets
c. Bag of Characters
d. Bag of Numbers
Ans: a. Bag of Words

Q68. Which of the following things we are getting after 'Bag of Words' algorithm?
a. A vocabulary of words for the corpus.
b. The frequency of these words.
c. Both of the above.
d. None of the above
Ans: c. Both of the above.

Q69. Expand TFIDF
a. Term Format & Inverse Document Frequency

b. Term Frequency & Inverse Document Frequency
c. Term Frequency & Inverse Data Frequency
d. Term Frequency & Inner Document Frequency
Ans: b. Term Frequency & Inverse Document Frequency

Q70. Bag of words algorithm gives us the frequency of words in each document. It gives us an idea that if the word is occurring more in a document, _____
a. its value is more for that document
b. its value is less for that document
c. its value is not more not less for that document
d. its has no value for that document.
Ans: a. its value is more for that document

Q71. Steps to implement bag of words algorithm is given below. Choose the correct sequence.
1. Text Normalisation
2. Create document vectors
3. Create document vectors for all the documents
4. Create Dictionary
a. 1, 2, 3, 4
b. 2, 3, 1, 4
c. 1, 4, 2, 3
d. 1, 4, 3, 2
Ans: c. 1, 4, 2, 3

Q72. _____ are the words which occur the most in almost all the documents.
a. And
b. The
c. This
d. All of the above
Ans: d. All of the above

Q73. Those words which are a complete waste for machine as they do not provide any information regarding the corpus are termed as _____
a. Start words
b. End words
c. Stop words
d. Close words
Ans: c. Stop words

Q74. Which of the following type of words have more value in the document of the corpus?
a. Stop words
b. Frequent words
c. Rare words
d. All of the above
Ans: c. Rare words

Q75. Which of the following type of words have more frequency in the document of the corpus?
a. Stop words
b. Frequent words
c. Rare words
d. All of the above
Ans: a. Stop words

Q76. _____ is the frequency of a word in one document.

a. Term frequency
b. Inverse Document Frequency
c. Document Frequency
d. Inverse Frequency
Ans: a. Term frequency

Q77. _____ is the number of documents in which the word occurs irrespective of how many times it has occurred in those documents.
a. Term frequency
b. Inverse Document Frequency
c. Document Frequency
d. Inverse Frequency
Ans: c. Document Frequency

Q78. In _____, we put the document frequency in the denominator while the total number of documents in the numerator.
a. Inverse Frequency
b. Inverse Document
c. Inverse Document Frequency
d. Term Frequency
Ans: c. Inverse Document Frequency

Q79. Which of the following is an application of TFIDF?
a. Document Classification
b. Topic Modelling
c. Stop word filtering
d. All of the above
Ans: d. All of the above

Q80. _____ helps in removing the unnecessary words out of a text body.
a. Document Classification
b. Topic Modelling
c. Stop word filtering
d. Information Retrieval System
Ans: c. Stop word filtering

QUESTIONS AND ANSWERS (SET 01) - 1 mark

1. What is a Chabot?
A chatbot is a computer program that's designed to simulate human conversation through voice commands or text chats or both. Eg: Mitsuku Bot, Jabberwacky etc.

OR

A chatbot is a computer program that can learn over time how to best interact with humans. It can answer questions and troubleshoot customer problems, evaluate and qualify prospects, generate sales leads and increase sales on an ecommerce site.

OR

A chatbot is a computer program designed to simulate conversation with human users. A chatbot is also known as an artificial conversational entity (ACE), chat robot, talk bot, chatterbot or chatterbox.

OR

A chatbot is a software application used to conduct an on-line chat conversation via text or text-to-speech, in lieu of providing direct contact with a live human agent.

2. What is the full form of NLP?
Natural Language Processing

3. While working with NLP what is the meaning of?
a. Syntax
b. Semantics
Syntax: Syntax refers to the grammatical structure of a sentence.
Semantics: It refers to the meaning of the sentence.

4. What is the difference between stemming and lemmatization?
Stemming is a technique used to extract the base form of the words by removing affixes from them. It is just like cutting down the branches of a tree to its stems. For example, the stem of the words *eating, eats, eaten* is *eat*.
Lemmatization is the grouping together of different forms of the same word. In search queries, lemmatization allows end users to query any version of a base word and get relevant results.

OR

Stemming is the process in which the affixes of words are removed and the words are converted to their base form.
In lemmatization, the word we get after affix removal (also known as lemma) is a meaningful one. Lemmatization makes sure that lemma is a word with meaning and hence it takes a longer time to execute than stemming.

OR

Stemming algorithms work by cutting off the end or the beginning of the word, taking into account a list of common prefixes and suffixes that can be found in an inflected word.
Lemmatization on the other hand, takes into consideration the morphological analysis of the words. To do so, it is necessary to have detailed dictionaries which the algorithm can look through to link the form back to its lemma.

5. What is the full form of TFIDF?
Term Frequency and Inverse Document Frequency

6. What is meant by a dictionary in NLP?
Dictionary in NLP means a list of all the unique words occurring in the corpus. If some words are repeated in different documents, they are all written just once as while creating the dictionary.

7. What is term frequency?
Term frequency is the frequency of a word in one document. Term frequency can easily be found from the document vector table as in that table we mention the frequency of each word of the vocabulary in each document.

8. Which package is used for Natural Language Processing in Python programming?
Natural Language Toolkit (NLTK). NLTK is one of the leading platforms for building Python programs that can work with human language data.

9. What is a document vector table?
Document Vector Table is used while implementing Bag of Words algorithm.
In a document vector table, the header row contains the vocabulary of the corpus and other rows correspond to different documents.
If the document contains a particular word it is represented by 1 and absence of word is represented by 0 value.

OR

Document Vector Table is a table containing the frequency of each word of the vocabulary in each document.

10. What do you mean by corpus?

In Text Normalization, we undergo several steps to normalize the text to a lower level. That is, we will be working on text from multiple documents and the term used for the whole textual data from all the documents altogether is known as corpus.

OR

A corpus is a large and structured set of machine-readable texts that have been produced in a natural communicative setting.

OR

A corpus can be defined as a collection of text documents. It can be thought of as just a bunch of text files in a directory, often alongside many other directories of text files.

QUESTIONS AND ANSWERS (SET 01) - 2 marks

1. What are the types of data used for Natural Language Processing applications?

Natural Language Processing takes in the data of Natural Languages in the form of written words and spoken words which humans use in their daily lives and operates on this.

2. Differentiate between a script-bot and a smart-bot. (Any 2 differences)

Script-bot	Smart-bot
• A scripted chatbot doesn't carry even a glimpse of A.I • Script bots are easy to make • Script bot functioning is very limited as they are less powerful. • Script bots work around a script which is programmed in them • No or little language processing skills • Limited functionality	• Smart bots are built on NLP and ML. • Smart –bots are comparatively difficult to make. • Smart-bots are flexible and powerful. • Smart bots work on bigger databases and other resources directly • NLP and Machine learning skills are required. • Wide functionality

3. Give an example of the following:

- Multiple meanings of a word
- Perfect syntax, no meaning

Example of Multiple meanings of a word –

His face turns red after consuming the medicine

Meaning - Is he having an allergic reaction? Or is he not able to bear the taste of that medicine?

Example of Perfect syntax, no meaning-

Chickens feed extravagantly while the moon drinks tea.

This statement is correct grammatically but it does not make any sense. In Human language, a perfect balance of syntax and semantics is important for better understanding.

4. What is inverse document frequency?

To understand inverse document frequency, first we need to understand document frequency.

Document Frequency is the number of documents in which the word occurs irrespective of how many times it has occurred in those documents.

In case of inverse document frequency, we need to put the document frequency in the denominator while the total number of documents is the numerator.

For example, if the document frequency of a word "AMAN" is 2 in a particular document then its inverse document frequency will be 3/2. (Here no. of documents is 3)

5. Define the following:

● Stemming

● Lemmatization

Stemming: Stemming is a rudimentary rule-based process of stripping the suffixes (“ing”, “ly”, “es”, “s” etc) from a word.

Stemming is a process of reducing words to their word stem, base or root form (for example, books — book, looked — look).

Lemmatization: Lemmatization, on the other hand, is an organized & step by step procedure of obtaining the root form of the word, it makes use of vocabulary (dictionary importance of words) and morphological analysis (word structure and grammar relations).

The aim of lemmatization, like stemming, is to reduce inflectional forms to a common base form. As opposed to stemming, lemmatization does not simply chop off inflections. Instead it uses lexical knowledge bases to get the correct base forms of words.

OR

Stemming is a technique used to extract the base form of the words by removing affixes from them. It is just like cutting down the branches of a tree to its stems. For example, the stem of the words *eating, eats, eaten* is *eat*.

Lemmatization is the grouping together of different forms of the same word. In search queries, lemmatization allows end users to query any version of a base word and get relevant results.

OR

Stemming is the process in which the affixes of words are removed and the words are converted to their base form.

In lemmatization, the word we get after affix removal (also known as lemma) is a meaningful one. Lemmatization makes sure that lemma is a word with meaning and hence it takes a longer time to execute than stemming.

OR

Stemming algorithms work by cutting off the end or the beginning of the word, taking into account a list of common prefixes and suffixes that can be found in an inflected word.

Lemmatization on the other hand, takes into consideration the morphological analysis of the words. To do so, it is necessary to have detailed dictionaries which the algorithm can look through to link the form back to its lemma.

6. What do you mean by document vectors?

Document Vector contains the frequency of each word of the vocabulary in a particular document. In document vector vocabulary is written in the top row. Now, for each word in the document, if it matches with the vocabulary, put a 1 under it. If the same word appears again, increment the previous value by 1. And if the word does not occur in that document, put a 0 under it.

7. What is TFIDF? Write its formula.

Term frequency–inverse document frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus.

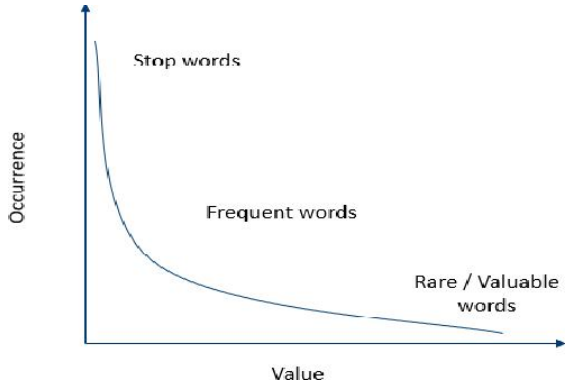
The number of times a word appears in a document divided by the total number of words in the document. Every document has its own term frequency.

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{i,j}}$$

8. Which words in a corpus have the highest values and which ones have the least?

Stop words like - and, this, is, the, etc. have highest values in a corpus. But these words do not talk about the corpus at all. Hence, these are termed as stopwords and are mostly removed at the pre-processing stage only.

Rare or valuable words occur the least but add the most importance to the corpus. Hence, when we look at the text, we take frequent and rare words into consideration.



9. Does the vocabulary of a corpus remain the same before and after text normalization? Why?
No, the vocabulary of a corpus does not remain the same before and after text normalization. Reasons are –
● In normalization the text is normalized through various steps and is lowered to minimum vocabulary since the machine does not require grammatically correct statements but the essence of it.
● In normalization Stop words, Special Characters and Numbers are removed.
● In stemming the affixes of words are removed and the words are converted to their base form.
So, after normalization, we get the reduced vocabulary.

10. What is the significance of converting the text into a common case?
In Text Normalization, we undergo several steps to normalize the text to a lower level.
After the removal of stop words, we convert the whole text into a similar case, preferably lower case. This ensures that the case-sensitivity of the machine does not consider same words as different just because of different cases.

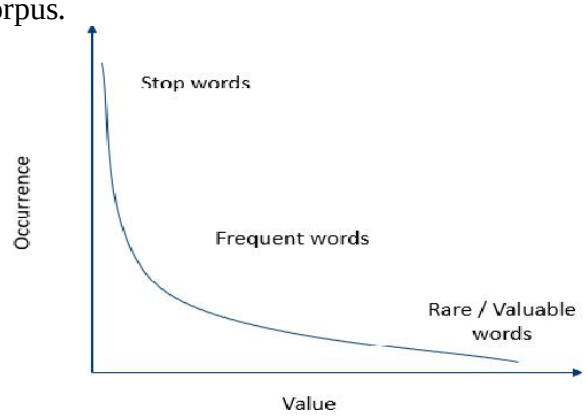
11. Mention some applications of Natural Language Processing.
Natural Language Processing Applications-
● Sentiment Analysis.
● Chatbots & Virtual Assistants.
● Text Classification.
● Text Extraction.
● Machine Translation
● Text Summarization
● Market Intelligence
● Auto-Correct

12. What is the need of text normalization in NLP?
Since we all know that the language of computers is Numerical, the very first step that comes to our mind is to convert our language to numbers.
This conversion takes a few steps to happen. The first step to it is Text Normalization.
Since human languages are complex, we need to first of all simplify them in order to make sure that the understanding becomes possible. Text Normalization helps in cleaning up the textual data in such a way that it comes down to a level where its complexity is lower than the actual data.

13. Explain the concept of Bag of Words.
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. In bag of words, we get the occurrences of each word and construct the vocabulary for the corpus.
Bag of Words just creates a set of vectors containing the count of word occurrences in the document (reviews). Bag of Words vectors are easy to interpret.

14. Explain the relation between occurrence and value of a word.

As shown in the graph, occurrence and value of a word are inversely proportional. The words which occur most (like stop words) have negligible value. As the occurrence of words drops, the value of such words rises. These words are termed as rare or valuable words. These words occur the least but add the most value to the corpus.



15. What are the applications of TFIDF?

TFIDF is commonly used in the Natural Language Processing domain. Some of its applications are:

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

16. What are stop words? Explain with the help of examples.

"Stop words" are the most common words in a language like "the", "a", "on", "is", "all". These words do not carry important meaning and are usually removed from texts. It is possible to remove stop words using Natural Language Toolkit (NLTK), a suite of libraries and programs for symbolic and statistical natural language processing.

17. Differentiate between Human Language and Computer Language.

Humans communicate through language which we process all the time. Our brain keeps on processing the sounds that it hears around itself and tries to make sense out of them all the time. On the other hand, the computer understands the language of numbers. Everything that is sent to the machine has to be converted to numbers. And while typing, if a single mistake is made, the computer throws an error and does not process that part. The communications made by the machines are very basic and simple.

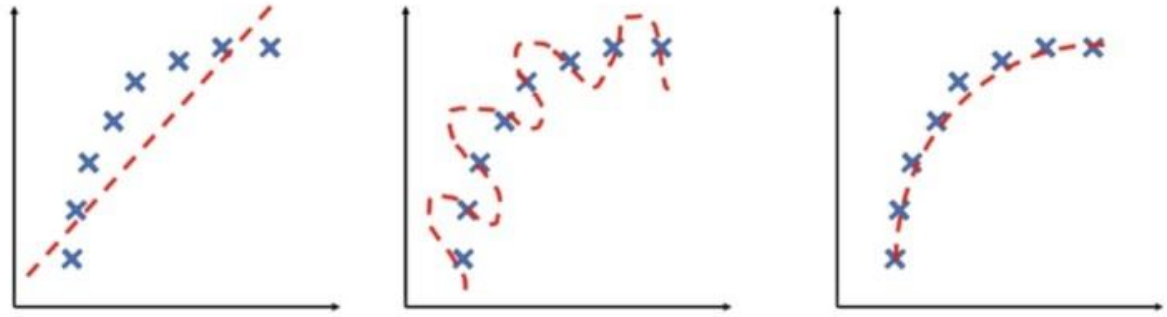
QUESTIONS AND ANSWERS (SET 01) - 3/4 marks

1. Create a document vector table for the given corpus:

Document 1: We are going to Mumbai
Document 2: Mumbai is a famous place.
Document 3: We are going to a famous place.
Document 4: I am famous in Mumbai.

We	Are	going	to	Mumbai	is	a	famous	place	I	am	in
1	1	1	1	1	0	0	0	0	0	0	0
0	0	0	0	1	1	1	1	1	0	0	0
1	1	1	1	0	0	1	1	1	0	0	0
0	0	0	0	1	0	0	1	0	1	1	1

2. Classify each of the images according to how well the model's output matches the data samples:



Here, the red dashed line is model's output while the blue crosses are actual data samples.

● The model's output does not match the true function at all. Hence the model is said to be under fitting and its accuracy is lower.

● In the second case, model performance is trying to cover all the data samples even if they are out of alignment to the true function. This model is said to be over fitting and this too has a lower accuracy

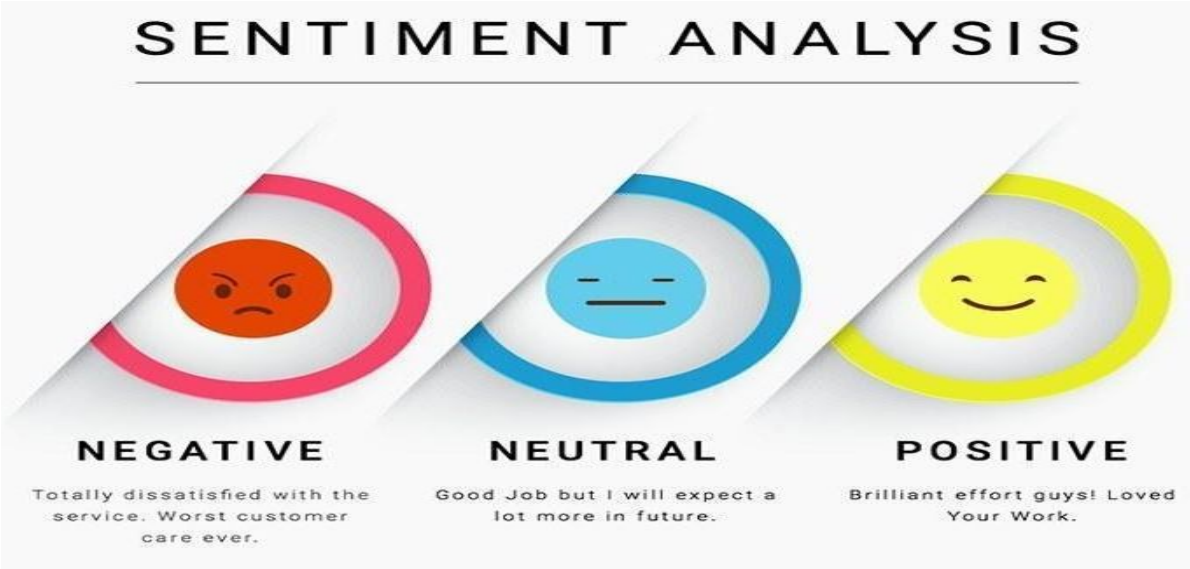
● In the third one, the model's performance matches well with the true function which states that the model has optimum accuracy and the model is called a perfect fit.

3. Explain how AI can play a role in sentiment analysis of human beings?

The goal of sentiment analysis is to identify sentiment among several posts or even in the same post where emotion is not always explicitly expressed.

Companies use Natural Language Processing applications, such as sentiment analysis, to identify opinions and sentiment online to help them understand what customers think about their products and services (i.e., "I love the new iPhone" and, a few lines later "But sometimes it doesn't work well" where the person is still talking about the iPhone) and overall *

Beyond determining simple polarity, sentiment analysis understands sentiment in context to help better understand what's behind an expressed opinion, which can be extremely relevant in understanding and driving purchasing decisions.



4. Why are human languages complicated for a computer to understand? Explain.

The communications made by the machines are very basic and simple. Human communication is complex. There are multiple characteristics of the human language that might be easy for a human to understand but extremely difficult for a computer to understand.

For machines it is difficult to understand our language. Let us take a look at some of them here:

Arrangement of the words and meaning - There are rules in human language. There are nouns, verbs, adverbs, adjectives. A word can be a noun at one time and an adjective some other time. This can create difficulty while processing by computers.
Analogy with programming language- Different syntax, same semantics: 2+3 = 3+2 Here the way these statements are written is different, but their meanings are the same that is 5. Different semantics, same syntax: 2/3 (Python 2.7) ≠ 2/3 (Python 3) Here the statements written have the same syntax but their meanings are different. In Python 2.7, this statement would result in 1 while in Python 3, it would give an output of 1.5.
Multiple Meanings of a word - In natural language, it is important to understand that a word can have multiple meanings and the meanings fit into the statement according to the context of it.
Perfect Syntax, no Meaning - Sometimes, a statement can have a perfectly correct syntax but it does not mean anything. In Human language, a perfect balance of syntax and semantics is important for better understanding.
These are some of the challenges we might have to face if we try to teach computers how to understand and interact in human language.

5. What are the steps of text Normalization? Explain them in brief.
Text Normalizationin Text Normalization, we undergo several steps to normalize the text to a lower level.
Sentence Segmentation - Under sentence segmentation, the whole corpus is divided into sentences. Each sentence is taken as a different data so now the whole corpus gets reduced to sentences.
Tokenisation- After segmenting the sentences, each sentence is then further divided into tokens. Tokens is a term used for any word or number or special character occurring in a sentence. Under tokenisation, every word, number and special character is considered separately and each of them is now a separate token.
Removing Stop words, Special Characters and Numbers - In this step, the tokens which are not necessary are removed from the token list.
Converting text to a common case -After the stop words removal, we convert the whole text into a similar case, preferably lower case. This ensures that the case-sensitivity of the machine does not consider same words as different just because of different cases.
Stemming In this step, the remaining words are reduced to their root words. In other words, stemming is the process in which the affixes of words are removed and the words are converted to their base form.
Lemmatization -in lemmatization, the word we get after affix removal (also known as lemma) is a meaningful one.
With this we have normalized our text to tokens which are the simplest form of words present in the corpus. Now it is time to convert the tokens into numbers. For this, we would use the Bag of Words algorithm

6. Through a step-by-step process, calculate TFIDF for the given corpus and mention the word(s) having highest value.
Document 1: We are going to Mumbai
Document 2: Mumbai is a famous place.
Document 3: We are going to a famous place.
Document 4: I am famous in Mumbai.
Term Frequency
Term frequency is the frequency of a word in one document. Term frequency can easily be found from the document vector table as in that table we mention the frequency of each word of the vocabulary in each document.

We	Are	Going	to	Mumbai	is	a	famous	Place	I	am	in
1	1	1	1	1	0	0	0	0	0	0	0
0	0	0	0	1	1	1	1	1	0	0	0

1	1	1	1	0	0	1	1	1	0	0	0
0	0	0	0	1	0	0	1	0	1	1	1

Inverse Document Frequency
The other half of TFIDF which is Inverse Document Frequency. For this, let us first understand what does document frequency mean. Document Frequency is the number of documents in which the word occurs irrespective of how many times it has occurred in those documents. The document frequency for the exemplar vocabulary would be:

We	Are	going	to	Mumbai	is	a	Famous	place	I	am	in
2	2	2	2	3	1	2	3	2	1	1	1

Talking about inverse document frequency, we need to put the document frequency in the denominator while the total number of documents is the numerator. Here, the total number of documents are 3, hence inverse document frequency becomes:

We	Are	going	to	Mumbai	is	a	Famous	Place	I	am	in
4/2	4/2	4/2	4/2	4/3	4/1	4/2	4/3	4/2	4/1	4/1	4/1

The formula of TFIDF for any word W becomes:
TFIDF(W) = TF(W) * log (IDF(W))
The words having highest value are – Mumbai, Famous

7. Normalize the given text and comment on the vocabulary before and after the normalization: Raj and Vijay are best friends. They play together with other friends. Raj likes to play football but Vijay prefers to play online games. Raj wants to be a footballer. Vijay wants to become an online gamer.
Normalization of the given text:
Sentence Segmentation:
1. Raj and Vijay are best friends.
2. They play together with other friends.
3. Raj likes to play football but Vijay prefers to play online games.
4. Raj wants to be a footballer.
5. Vijay wants to become an online gamer.

Tokenization:

Raj and Vijay are best friends.	*Raj*	*and*	*Vijay*	*are*	*best*	*friends*	.
They play together with other friends	*They*	*play*	*Together*	*with*	*other*	*friends*	.

Same will be done for all sentences.
Removing Stop words, Special Characters and Numbers:
In this step, the tokens which are not necessary are removed from the token list.
So, the words and, are, to, an, (Punctuation) will be removed.

Converting text to a common case:
After the stop words removal, we convert the whole text into a similar case, preferably lower case. Here we don't have words in different case so this step is not required for given text.
Stemming:
In this step, the remaining words are reduced to their root words. In other words, stemming is the process in which the affixes of words are removed and the words are converted to their base form.

Word	*Affixes*	*Stem*
Likes	*-s*	*Like*
Prefers	*-s*	*Prefer*
Wants	*-s*	*want*

In the given text Lemmatization is not required.
Given Text
Raj and Vijay are best friends. They play together with other friends. Raj likes to play football but Vijay prefers to play online games. Raj wants to be a footballer. Vijay wants to become an online gamer.
Normalized Text
Raj and Vijay best friends They play together with other friends Raj likes to play football but Vijay prefers to play online games Raj wants to be a footballer Vijay wants to become an online gamer

QUESTIONS AND ANSWERS (SET 02)

1. What do you mean by Natural Language Processing?
Answer – The area of artificial intelligence known as natural language processing, or NLP, is dedicated to making it possible for computers to comprehend and process human languages. The interaction between computers and human (natural) languages is the focus of artificial intelligence (AI), a subfield of linguistics, computer science, information engineering, and artificial intelligence. This includes learning how to programme computers to process and analyze large amounts of natural language data.

2. What are the different applications of NLP which are used in real-life scenario?
Answer – Some of the applications which is used in the real-life scenario are –
a. Automatic Summarization – Automatic summarization is useful for gathering data from social media and other online sources, as well as for summarizing the meaning of documents and other written materials. When utilized to give a summary of a news story or blog post while eliminating redundancy from different sources and enhancing the diversity of content acquired, automatic summarizing is particularly pertinent.
b. Sentiment Analysis – In posts when emotion is not always directly expressed, or even in the same post, the aim of sentiment analysis is to detect sentiment. To better comprehend what internet users are saying about a company's goods and services, businesses employ natural language processing tools like sentiment analysis.
c. Text Classification – Text classification enables you to classify a document and organize it to make it easier to find the information you need or to carry out certain tasks. Spam screening in email is one example of how text categorization is used.

d. Virtual Assistants – These days, digital assistants like Google Assistant, Cortana, Siri, and Alexa play a significant role in our lives. Not only can we communicate with them, but they can also facilitate our life. They can assist us in making notes about our responsibilities, making calls for us, sending messages, and much more by having access to our data.

3. What is Cognitive Behavioural Therapy (CBT)?
Answer – One of the most effective ways to deal with stress is cognitive behavioural therapy (CBT), which is popular since it is simple to apply to people and produces positive outcomes. Understanding a person's behaviour and mentality in daily life is part of this therapy. Therapists assist clients in overcoming stress and leading happy lives with the aid of CBT.

4. What is Problem Scoping?
Answer – Understanding a problem and identifying numerous elements that have an impact on it help define the project's purpose or objective. Who, What, Where, and Why are the 4Ws of problem scoping. These Ws make it easier and more effective to identify and understand the problem.

5. What is Data Acquisition?
Answer – We need to gather conversational data from people in order to decipher their setements and comprehend their meaning in order to grasp their feelings. This collection of information is known as Data Acquistion. Such information can be gathered in a variety of ways –
a. Surveys
b. Observing the therapist's sessions
c. Databased available on the internet

6. What is Data Exploration?
Answer – Once the textual information has been gathered using Data Acquistition, it must be cleaned up and processed before being delivered to the machine in a simpler form. As a result, the text is normalised using a number of processes, and the vocabulary is reduced to a minimum because the computer just needs the text's main ideas rather than its grammar.

7. What is Data Modelling?
Answer – After the text has been normalised, an NLP-based AI model is then fed the data. Keep in mind that in NLP, data pre-processing is only necessary after which the data is supplied to the computer. There are numerous AI models that can be used, depending on the kind of chatbot we're trying to create, to help us lay the groundwork for our project.

8. What is Data Evaluation?
Answer – The correctness of the trained model is determined based on how well the machine-generated answers match the user's input is knwon as Data Evaluation. The chatbot's proposed answers are contrasted with the correct answers to determine the model's efficacy.

9. What is Chatbot?
Answer – A chatbot is a piece of software or an agent with artificial intelligence that uses natural language processing to mimic a conversation with users or people. You can have the chat through a website, application, or messaging app. These chatbots, often known as digital assistants, can communicate with people verbally or via text.
The majority of organizations utilize AI chatbots, such the Vainubot and HDFC Eva chatbots, to give their clients virtual customer assistance around-the-clock.
Some of the example of Chatbot –
a. Mitsuku Bot
b. CleverBot
c. Jabberwacky
d. Haptik

e. Rose
f. Ochtbot

10. Explain the types of Chatbot?
Answer – There are two types of Chatbot –
a. Script Bot – An Internet bot, sometimes known as a web robot, robot, or simply bot, is a software programme that does automated operations (scripts) over the Internet, typically with the aim of simulating extensive human online activity like communicating.
b. Smart Bot – An artificial intelligence (AI) system that can learn from its surroundings and past experiences and develop new skills based on that knowledge is referred to as a smart bot. Smart bot that are intelligent enough can operate alongside people and learn from their actions.

11. Difference between human language vs computer language?
Answer – Although there is a significant difference between the languages, human language and computer language can be translated into one other very flawlessly. Human languages can be used in voice, writing, and gesture, whereas machine-based languages can only be used in written communication. A computer's textual language can communicate with vocal or visual clues depending on the situation, as in AI chatbots with procedural animation and speech synthesis. But in the end, language is still written. The languages also have different meanings. Human languages are utilized in a variety of circumstances, including this blog post, whereas machine languages are almost solely used for requests, commands, and logic.

12. What do you mean by Multiple Meanings of a word in Deep Learning?
Answer – Depending on the context, the term mouse can be used to refer to either a mammal or a computer device. Consequently, mouse is described as ambiguous. The Principle of Economical Versatility of Words states that common words have a tendency to acquire additional senses, which can create practical issues in subsequent jobs. Additionally, this meaning conflation has additional detrimental effects on correct semantic modelling, such as the pulling together in the semantic space of words that are semantically unrelated yet are comparable to distinct meanings of the same word.

13. What is Data Processing?
Answer – Making data more meaningful and informative is the effort of changing it from a given form to one that is considerably more useable and desired. This entire process can be automated using Machine Learning algorithms, mathematical modelling, and statistical expertise.

14. What is Text Normalisation?
Answer – The process of converting a text into a canonical (standard) form is known as text normalisation. For instance, the canonical form of the word "good" can be created from the words "gooood" and "gud." Another case is the reduction of terms that are nearly identical, such as "stopwords," "stop-words," and "stop words," to just "stopwords."
We must be aware that we will be working on a collection of written text in this portion before we start. As a result, we will be analysing text from a variety of papers. This collection of text from all the documents is referred to as a corpus. We would perform each stage of Text Normalization and test them on a corpus in addition to going through them all.

15. What is Sentence Segmentation in AI?
Answer – The challenge of breaking down a string of written language into its individual sentences is known as sentence segmentation. The method used in NLP to determine where sentences actually begin and end, or you can just say that this is how we divide a text into sentences. Sentence segmentation is the process in question. Using the spacy library, we implement this portion of NLP in Python.

16. What is Tokenisation in AI?
Answer – The challenge of breaking down a string of written language into its individual words is

known as word tokenization (also known as word segmentation). Space is a good approximation of a word divider in English and many other languages that use some variation of the Latin alphabet.

17. What is purpose of Stopwords?
Answer – Stopwords are words that are used frequently in a corpus but provide nothing useful. Humans utilize grammar to make their sentences clear and understandable for the other person. However, grammatical terms fall under the category of stopwords because they do not add any significance to the information that is to be communicated through the statement. Stopword examples include –
a/ an/ and/ are/ as/ for/ it/ is/ into/ in/ if/ on/ or/ such/ the/ there/ to

18. What is Stemming in AI?
Answer – The act of stripping words of their affixes and returning them to their original forms is known as stemming. The process of stemming can be carried out manually or by an algorithm that an AI system may use. Any inflected form that is encountered can be reduced to its root by using a variety of stemming techniques. A stemming algorithm can be created easily.

19. What is Lemmatization?
Answer – Stemming and lemmatization are alternate techniques to one another because they both function to remove affixes. However, lemmatization differs from both of them in that the word that results from the elimination of the affix (also known as the lemma) is meaningful.
Lemmatization takes more time to complete than stemming because it ensures that the lemma is a word with meaning.

20. What is bag of Words?
Answer – Bag of Words is a model for natural language processing that aids in removing textual elements that can be used by machine learning techniques. We obtain each word's occurrences from the bag of words and create the corpus's vocabulary.
An approach to extracting features from text for use in modelling, such as with machine learning techniques, is known as a bag-of-words model, or BoW for short. The method is really straightforward and adaptable, and it may be applied in a variety of ways to extract features from documents.

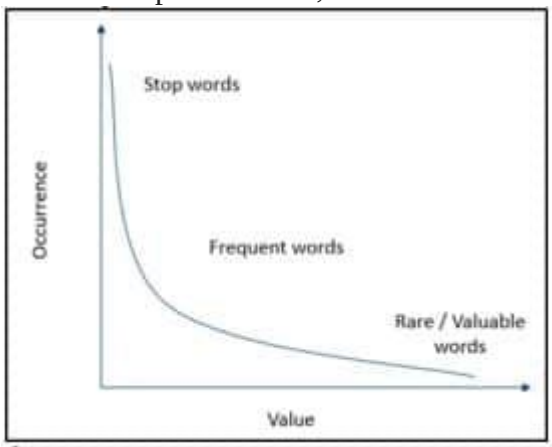
21. What is TFIDF?
Answer – TF-IDF, which stands for term frequency-inverse document frequency, is a metric that is employed in the fields of information retrieval (IR) and machine learning to quantify the significance or relevance of string representations (words, phrases, lemmas, etc.) in a document among a group of documents (also known as a corpus).

22. What are the different applications of TFIDF?
Answer – TFIDF is commonly used in the Natural Language Processing domain. Some of its applications are:
a. Document classification -Helps in classifying the type and genre of a document.
b. Topic Modelling – It helps in predicting the topic for a corpus.
c. Information Retrieval System – To extract the important information out of a corpus.
d. Steop word filtering – Helps in removing the unnecessary words out of a text body.

23. Write any two TFIDF application?
Answer – 1. Document Classification – Helps in classifying the type and genre of a document.
2. Topic Modelling – It helps in predicting the topic for a corpus.
3. Information Retrieval System – To extract the important information out of a corpus.
4. Stop word filtering – Helps in removing the unnecessary words out of a text body.

24. Write the steps necessary to implement the bag of words algorithm.
Answer – The steps to implement bag of words algorithm are as follows:
1. Text Normalisation: Collect data and pre-process it
2. Create Dictionary: Make a list of all the unique words occurring in the corpus.
3. Create document vectors: For each document in the corpus, find out how many times the word from the unique list of words has occurred.
4. Create document vectors for all the documents.

25. What is the purpose of confusion matrix? What does it serve?
Answer – The comparison between the prediction and reality's outcomes is stored in the confusion matrix. We can determine variables like recall, precision, and F1 score, which are used to assess an AI model's performance, from the confusion matrix.



26. How does the relationship between a word's value and frequency in a corpus look like in the given graph?
Answer – The graph demonstrates the inverse relationship between word frequency and word value. The most frequent terms, such as stop words, are of little significance. The value of words increases as their frequency decreases. These words are referred to as precious or uncommon words. The least frequently occurring but most valuable terms in the corpus are those.

27. In data processing, define the term "Text Normalization."
Answer – Text normalisation is the initial step in the data processing process. Text normalisation assists in reducing the complexity of the textual data to a point where it is comparable to the actual data. To lower the text's normalisation level in this, we go through numerous procedures. We work with text from several sources, and the collective textual data from all the papers is referred to as a corpus.

28. Explain the differences between lemmatization and stemming. Give an example to assist you explain.
Answer – Stemming is the process of stripping words of their affixes and returning them to their original form.
After the affix is removed during lemmatization, we are left with a meaningful word known as a lemma. Lemmatization takes more time to complete than stemming because it ensures that the lemma is a word with meaning.
The following example illustrates the distinction between stemming and lemmatization:
Caring >> Lemmatization >> Care
Caring >> Stemming >> Car

29. Imagine developing a prediction model based on AI and deploying it to monitor traffic congestion on the roadways. Now, the model's goal is to foretell whether or not there will be a traffic jam. We must now determine whether or not the predictions this model generates are

accurate in order to gauge its efficacy. Prediction and Reality are the two conditions that we need to consider.
Today, traffic jams are a regular occurrence in our life. Every time you get on the road when you live in an urban location, you have to deal with traffic. Most pupils choose to take buses to school. Due to these traffic bottlenecks, the bus frequently runs late, making it impossible for the pupils to get to school on time.
Create a Confusion Matrix for the aforementioned scenario while taking into account all potential outcomes.
Answer –
Case 1: Is there a traffic Jam?
Prediction: Yes Reality: Yes
True Positive
Case 2: Is there a traffic Jam?
Prediction: No Reality: No
True Negative
Case 3: Is there a traffic Jam?
Prediction: Yes Reality: No
False Positive
Case 4: Is there a traffic Jam?
Prediction: No Reality: Yes
False Negative

Confusion Matrix		Reality	
		Yes	No
Prediction	Yes	True Positive	False Positive
	No	False Negative	True Negative

30. Make a 4W Project Canvas.
Risks will become more concentrated in a single network as more and more innovative technologies are used. In such cases, cybersecurity becomes incredibly complex and is no longer under the authority of firewalls. It won't be able to recognise odd behaviour patterns, including data migration.
Consider how AI systems can sift through voluminous data to find user behaviour that is vulnerable. To explicitly define the scope, the method of data collection, the model, and the evaluation criteria, use an AI project cycle.
Answer –

OUR	**[stakeholders]** People who are using the new technology	WHO
HAS/ HAVE PROBLEM THAT	**[issue, problem, need]** Cyber security is the need when so much of the flow of data is not monitored or escapes the antiviruses/ firewall systems.	WHAT
WHEN/ WHILE	**[context/situation]** The problem is in the use of the latest technology where vast amounts of data is at risk.	WHERE
AN IDEAL SOLUTION WOULD	**[benefit of solution to them]** An effective AI system which is able to detect the flow of data and also report unusual activity	WHY