UNIT-6 **TOPIC- Natural Language Processing**

NLP (Natural Language Processing), is dedicated to making it possible for computers to comprehend and process human languages. Artificial intelligence (AI) is a subfield of linguistics, computer science, information engineering, and artificial intelligence that studies how computers interact with human (natural) languages, particularly how to train computers to handle and analyze massive volumes of natural language data.

SUB-TOPIC : Applications of Natural Language Processing

Most people utilize NLP apps on a regular basis in their daily lives. Following are a few examples of real-world uses for natural language processing:
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.
Sentiment Analysis – To better comprehend what internet users are saying about a company's goods and services, businesses use natural language processing tools like sentiment analysis to understand the customer requirement.
Indicators of their reputation – Sentiment analysis goes beyond establishing simple polarity to analyse sentiment in context to help understand what is behind an expressed view. This is very important for understanding and influencing purchasing decisions.
Text classification – Text classification enables you to classify a document and organise 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.

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.

DETAILED EXPLAINATION

Chatbots

A chatbot is one of the most widely used NLP applications. Many chatbots on the market now employ the same strategy as we did in the instance above.

There are 2 types of chatbots : Scriptbot and Smart-bot.

Scriptbot	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**

Scriptbot	Smart-bot
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

Human Language VS Computer Language

Humans need language to communicate, which we constantly process. Our brain continuously processes the sounds it hears around us and works to make sense of them. Our brain continuously processes and stores everything, even as the teacher is delivering the lesson in the classroom.

The Computer Language is understood by the computer, on the other hand. All input must be transformed to numbers before being sent to the machine. And if a single error is made while typing, the machine throws an error and skips over that area. Machines only use extremely simple and elementary forms of communication.

Data Processing

Data Processing is a method of manipulation of data. It means the conversion of raw data into meaningful and machine-readable content. It basically is a process of converting raw data into meaningful information.

Since human languages are complex, we need to first of all simplify them in order to make sure that the understanding becomes possible. Text Normalisation 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. Let us go through Text Normalisation in detail.

Text Normalisation

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 illustration is the reduction of terms that are nearly identical, such as “stopwords,” “stop-words,” and “stop words,” to just “stopwords.”

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

Sentences are first broken into segments, and then each segment is further divided into tokens. Any word, number, or special character that appears in a sentence is referred to as a token. Tokenization treats each word, integer, and special character as a separate entity and creates a token for each of them.

Removing Stopwords, Special Characters and Numbers

In this step, the tokens which are not necessary are removed from the token list. What can be the possible words which we might not require?

Stopwords are words that are used frequently in a corpus but provide nothing useful. Humans utilise 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. Stopwords include a, an, and, or, for, it, is, etc.

Converting text to a common case

After eliminating the stopwords, we change the text's case throughout, preferably to lower case. This makes sure that the machine's case-sensitivity does not treat similar terms differently solely because of varied case usage.

Stemming

The remaining words are boiled down to their root words in this step. In other words, stemming is the process of stripping words of their affixes and returning them to their original forms.

Lemmatization

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.

Bag of Words

A bag-of-words is a textual illustration that shows where words appear in a document. There are two components: a collection of well-known words. a metric for the amount of well-known words. A Natural Language Processing model called Bag of Words aids in the extraction of textual information that can be used by machine learning techniques. We gather the instances of each term from the bag of words and create the corpus's vocabulary.

Here is the step-by-step approach to implement bag of words algorithm:

1. Text Normalisation: Collect data and pre-process it
2. Create Dictionary: Make a list of all the unique words occurring in the corpus. (Vocabulary)
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.

Term Frequency

The measurement of a term's frequency inside a document is called term frequency. The simplest calculation is to count the instances of each word. However, there are ways to change that value based on the length of the document or the frequency of the term that appears the most often.

Inverse Document Frequency

A term's frequency inside a corpus of documents is determined by its inverse document frequency. It is calculated by dividing the total number of documents in the corpus by the number of documents that contain the phrase.

Applications of TFIDF

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

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 the unnecessary words out of a text body.

MCQ'S(NLP)

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

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 **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 **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 **c. Mystery Animal**

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

VERY SHORT ANSWER QUESTIONS (2 MARKS)

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 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.

3. **. What is Data Acquisition?**
   Answer – We need to gather conversational data from people in order to decipher their statements and comprehend their meaning in order to grasp their feelings. This collection of information is known as Data Acquisition. Such information can be gathered in a variety of ways –a. Surveys b. Observing the therapist's sessions c. Database available on the internet
4. **What is Data Exploration?**
   Answer – Once the textual information has been gathered using Data Acquisition, 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.
5. **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.

6. **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 known as Data Evaluation. The chatbot's proposed answers are contrasted with the correct answers to determine the model's efficacy.

SHORT ANSWER QUESTIONS (3 MARKS)

1. **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.
2. **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.
3. **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.
4. **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."
5. **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.**

LONG ANSWER QUESTIONS (4 MARKS)-

1. 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.

2. What is Chatbot? Write Types of 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
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.

3. 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.