Day 43

DIY Solution

Q1. What is NLP?

Answer: Natural language processing (NLP) is a subfield of computer science and artificial intelligence.Where it deals with text, and voice data, and analyzes large amounts of natural language data to generate an appropriate output.

The goal is a make computers capable of "understanding" the contents of documents, including the contextual nuances of the language within them. The technology can then accurately extract information and insights contained in the documents as well as categorize and organize the documents themselves.

Q2. What is tokenization?

Answer: Tokenization is the act of breaking up a sequence of strings into pieces such as words, keywords, phrases, symbols, and other elements called tokens. Tokens can be individual words, phrases, or even whole sentences. In the process of tokenization, some characters like punctuation marks are discarded. NLTK library allows us to make tokens using word_tokenization, sentence _tokenization, etc.

Q3. State the applications of NLP.

Answer: Some of the common tasks of NLP include:

1. **Machine Translation:** This helps to translate the text of one language to another language. (language translater)
2. **Text Summarization:** Based on a large corpus, this is used to give a short summary that gives an idea of the entire text in the document.
3. **Language Modeling:** It is also helpful to generate auto-complete sentences Based on the history of previous words, this helps uncover what the further sentence will look like. A good example of this is the auto-complete sentences feature in Gmail.
4. **Topic Modelling:** This helps uncover the topical structure of a large collection of documents. This indicates what topic a piece of text is actually about.
5. **Question Answering:** This helps prepare answers automatically based on a corpus of text, and on a question that is posed.
6. **Conversational Agent:** These are basically voice assistants that we commonly see such as Alexa, Siri, Google Assistant, Cortana, etc.
7. **Information Retrieval:** This helps in fetching relevant documents based on a user's search query.
8. **Information Extraction:** This is the task of extracting relevant pieces of information from a given text, such as calendar events from emails.
9. **Text Classification:** This is used to create a bucket of categories of a given text, based on its content. This is used in a wide variety of AI-based applications such as sentiment analysis and spam detection.

Q4. What are the steps involved while preprocessing data for NLP?

Answer: Here are some common pre-processing steps used in NLP software

Preliminaries: This includes word tokenization and sentence segmentation.

Common Steps: Stop word removal, stemming and lemmatization, removing digits/punctuation, lowercasing, etc.

Processing Steps: Code mixing, normalization, language detection, transliteration, etc.

Advanced Processing: Parts of Speech (POS) tagging, coreference resolution, parsing, etc.

Q5. Explain Morphological analysis in NLP.

Answer:

Lemmatization: The task of removing inflectional endings only and to return the base dictionary form of a word which is also known as a lemma. Lemmatization is another technique for reducing words to their normalized form. But in this case, the transformation actually uses a dictionary to map words to their actual form.

Morphological segmentation: Separate words into individual morphemes and identify the class of the morphemes. The difficulty of this task depends greatly on the complexity of the morphology (i.e., the structure of words) of the language being considered. English has fairly simple morphology, especially inflectional morphology, and thus it is often possible to ignore this task entirely and simply model all possible forms of a word (e.g., "open, opens, opened, opening") as separate words. In languages such as Turkish or Meitei, a highly agglutinated Indian language, however, such an approach is not possible, as each dictionary entry has thousands of possible word forms.

Part-of-speech tagging: Given a sentence, determine the part of speech (POS) for each word. Many words, especially common ones, can serve as multiple parts of speech. For example, "book" can be a noun ("the book on the table") or verb ("to book a flight"); "set" can be a noun, verb, or adjective; and "out" can be any of at least five different parts of speech.

Stemming: The process of reducing inflected words to a base form (e.g., "close" will be the root for "closed", "closing", "close", "closer" etc.). Stemming yields similar results as lemmatization, but does so on grounds of rules, not a dictionary.