Practical No:5

Title:

Implement HMM for POS tagging. Build a Chunker

Aim:

1. To implement a **Hidden Markov Model (HMM)** for Part-of-Speech (POS) tagging.
2. To build a **Chunker** for identifying syntactic structures such as noun phrases and verb phrases.

Pre-requisites:

1. Understanding of **Hidden Markov Models (HMM)** and their applications in sequence modeling.
2. Familiarity with **Part-of-Speech tagging** and how POS tags help in identifying syntactic structures.
3. Understanding of **Chunking**, the process of extracting short phrases from a sentence.
4. Knowledge of NLP libraries like nltk to implement POS tagging and chunking.

Theory:

1. Hidden Markov Model (HMM) for POS Tagging

A **Hidden Markov Model (HMM)** is a probabilistic model used to represent a sequence of observed events, such as words, based on hidden states, such as part-of-speech tags. HMM assumes that:

- The sequence of words depends on a hidden sequence of POS tags.

- Each word in the sentence is emitted from a corresponding POS tag.

2. Chunking (Shallow Parsing)

Chunking is the process of identifying short phrases in a sentence, such as noun phrases (NP), verb phrases (VP), etc., by grouping POS tags. The process is often referred to as shallow parsing because it does not provide the deep hierarchical structure of a full syntactic parse.

- Noun Phrase (NP): A chunk consisting of a noun and any preceding adjectives or determiners (e.g., "the big dog").
- Verb Phrase (VP): A chunk consisting of a verb and any associated objects or complements (e.g., "is running quickly").

In chunking, regular expressions can be defined based on POS tag sequences to extract these chunks from the sentence.

Steps for HMM POS Tagging:

1. Text Preprocessing:
   - Tokenize the sentence into words.
2. HMM Training:
   - Train the HMM on a tagged corpus to estimate transition probabilities between POS tags and emission probabilities from POS tags to words.
3. POS Tagging:
   - Use the Viterbi algorithm to predict the most likely sequence of POS tags for a given sentence.

Steps for Chunker:

1. Text Preprocessing:
    - Tokenize the sentence into words and apply POS tagging.
2. Chunking:
    - Define regular expression patterns to extract noun and verb phrases based on the POS tag sequences.
3. Chunk Extraction:
    - Apply the chunking patterns to extract chunks such as noun phrases and verb phrases.

Conclusion:

Implementing **HMM for POS tagging** is useful for predicting the most likely sequence of tags for a sentence, especially when the sequence of words is known, but the tags are hidden. **Chunking** helps in extracting useful syntactic structures like noun and verb phrases, which can be used in further tasks such as named entity recognition (NER) and shallow parsing. Together, these techniques provide foundational tools for syntactic analysis in NLP tasks.