Practical No:4

Title:

Perform Lemmatization and Stemming. Identify parts-of Speech using Penn Treebank tag set.

Aim:

1. To perform **lemmatization** and **stemming** on a given text.
2. To identify and tag parts of speech (POS) using the **Penn Treebank Tag Set**.

Pre-requisites:

1. Understanding of Natural Language Processing (NLP) techniques, specifically **lemmatization**, **stemming**, and **POS tagging**.
2. Familiarity with libraries such as nltk for NLP tasks.
3. Knowledge of the **Penn Treebank Tag Set**, which is a standard POS tagging system used in English corpora.

Theory:

Lemmatization:

- **Lemmatization** is the process of reducing a word to its base or dictionary form, known as the **lemma**. Lemmatization considers the **context** and **part of speech** of the word, which results in more meaningful reductions compared to stemming.
  - Example: The word *running* becomes *run*, and *better* becomes *good*.

Stemming:

- **Stemming** is a more rudimentary process of chopping off the end of words to reduce them to their root forms (known as **stems**). Stemming does not consider the context of the word and often results in non-real words.
  - Example: The word *running* becomes *run*, and *studies* becomes *studi*.

Part-of-Speech (POS) Tagging:

- POS tagging is the process of assigning a tag to each word in a sentence that corresponds to its grammatical category. These categories include nouns, verbs, adjectives, etc.
- **Penn Treebank Tag Set** is a commonly used POS tag set that includes 36 POS tags, such as:
  - **NN**: Noun, singular
  - **VB**: Verb, base form
  - **JJ**: Adjective
  - **RB**: Adverb
  - **PRP**: Personal pronoun

Steps for Lemmatization and Stemming:

1. **Text Preprocessing:**
   - Input text is tokenized into individual words.
   - Convert the text into lowercase for uniformity.
2. **Perform Lemmatization:**
   - Use lemmatization to convert each word into its base form using a lemmatizer like `WordNetLemmatizer`.
3. **Perform Stemming:**

    - Apply stemming to each word using a stemmer like `PorterStemmer`.

Steps for POS Tagging Using Penn Treebank Tag Set:

1. **Text Preprocessing:**
    - Tokenize the sentence into words.
2. **POS Tagging:**
    - Use a POS tagger from `nltk` to assign parts-of-speech tags to each word according to the Penn Treebank Tag Set.
3. **Output Tags:**
    - Display the tagged sentence with the POS categories of each word.

Conclusion:

Lemmatization and stemming are essential techniques in text normalization for many NLP applications. Lemmatization produces linguistically meaningful base forms, while stemming reduces words more crudely to their stems. POS tagging, using the Penn Treebank Tag Set, provides grammatical information about each word in a sentence, making it crucial for tasks like syntactic parsing, named entity recognition, and machine translation. These processes are foundational steps in building sophisticated language models.