Read me file (please scroll to end to see issues being faced)

Lexical Analyzer (Lexer)

The lexical analyzer is responsible for tokenizing the input sentence. It defines tokens representing different parts of speech, such as articles, nouns, adjectives, verbs, etc.

Syntax Analyzer (Parser)

The syntax analyzer parses the tokenized input and checks whether it adheres to the specified grammar rules. It defines productions for constructing sentences with nouns, verbs, adjectives, etc.

Brown Corpus

The program utilizes NLTK's Brown Corpus to fetch a diverse set of words categorized by their parts of speech. These words are used to create patterns for matching tokens.

Token Definitions

- ARTICLE: Represents articles such as 'a', 'an', 'the'.
- NOUN: Represents nouns.
- ADJECTIVE: Represents adjectives.
- VERB: Represents verbs.
- VERBEX: Represents auxiliary verbs like 'is', 'am', 'was', etc.
- VERBC: Represents specific verb combinations like 'is sleeping', 'are crying', etc.

Sentence Structure

- A sentence can consist of a noun followed by a verb.
- A sentence can consist of a noun followed by a specific verb combination.
- A sentence can be just a noun.

Error Handling

The program handles errors such as exceeding the maximum sentence length or encountering an illegal character. It skips illegal inputs and provides informative error messages.

Additional Comments

- The program defines word lists for articles, nouns, adjectives, and verbs based on the Brown Corpus.
- Tokenization uses regular expressions and checks whether the words are valid parts of speech.
- The maximum sentence length is set to 20 words.
- Verb combinations are defined for more complex sentence structures.

How to Run

1. Ensure you have NLTK installed (`pip install nltk`).
2. Run the program.
3. Enter a sentence when prompted.

Example:-

Enter a sentence: The cat is sleeping

Input: the cat is sleeping
LexToken(ARTICLE,'the',1,0)
LexToken(NOUN,'cat',1,4)
LexToken(VERBC,'is sleeping',1,8)
Valid Statement.

Side note: The LexToken is to check if a word is correctly matching any defined set. This has been incorporated in the program to debug.

Present errors:
Presently, the program is able to identify whether a word belongs to which set. It will not place any hindi word/ or any other word in any defined category.

However, the issue is arising when defining statements to validate.
For example : is a sleep
is,a – article
sleep – verb

This will be shown as a valid statement as all words fall under some category, which should not be the case.