VASAVI COLLEGE OF ENGINEERING (Autonomous)
IBRAHIMBAGH, HYDERABAD – 500 031
Department of Computer Science & Engineering

INNOVATION IN TEACHING

Course: Natural Language Processing **Faculty:** C. Gireesh

Topic: Tokenization and POS tagging **Semester:** VIII semester

Teaching Aid / Tool Used: NLTK

Description of the Tool:

NLTK (Natural Language Toolkit) is a suite that contains libraries and programs for statistical language processing. It is one of the most powerful NLP libraries, which contains packages to make machines understand human language and reply to it with an appropriate response.

Tools Usage in Teaching:

Tokenization is the process by which a large quantity of text is divided into smaller parts called tokens. These tokens are very useful for finding patterns and are considered as a base step for stemming and lemmatization. Tokenization also helps to substitute sensitive data elements with non-sensitive data elements.

Example python script and Output:

```
from nltk.tokenize import word_tokenize
text = "God is Great! I won a lottery."
print(word_tokenize(text))

Output: ['God', 'is', 'Great', '!', 'I', 'won', 'a', 'lottery', '.']
```

POS Tagging (Parts of Speech Tagging) is a process to mark up the words in text format for a particular part of a speech based on its definition and context. It is responsible for text reading in a language and assigning some specific token (Parts of Speech) to each word. It is also called grammatical tagging.

Some NLTK POS tagging examples are: CC, CD, EX, JJ, MD, NNP, PDT, PRP$, TO, etc.

POS tagger is used to assign grammatical information of each word of the sentence. Installing, Importing and downloading all the packages of Part of Speech tagging with NLTK is complete.

Steps Involved in the POS tagging example:

- Tokenize text (word_tokenize)
- apply pos_tag to above step that is nltk.pos_tag(tokenize_text)

Example python script and Output:

```
from nltk import pos_tag
from nltk import RegexpParser
text ="learn php from guru99 and make study easy".split()
print("After Split:",text)
tokens_tag = pos_tag(text)
print("After Token:",tokens_tag)
```

```
After Split: ['learn', 'php', 'from', 'guru99', 'and', 'make', 'study', 'easy']
After Token: [('learn', 'JJ'), ('php', 'NN'), ('from', 'IN'), ('guru99', 'NN'), ('and', 'CC'), ('make', 'VB'), ('study', 'NN'), ('easy', 'JJ')]
```