PYTHON FOR COMPUTATIONAL PROBLEM SOLVING

Introduction to Natural Language Toolkit(NLTK)

UE25CS151A

Department of Computer Science and Engineering
Prof. Kundhavai K R, CSE Department

Introduction to NLTK

What is NLTK?

NLTK (Natural Language Toolkit) is a Python library for working with text data.
Useful for:

- Splitting text into words/sentences
- Removing common words
- Counting frequent terms

Installation and Setup

Install NLTK

```
pip install nltk
```

Download Required Datasets

```
import nltk
nltk.download('punkt')       # Tokenizers
nltk.download('stopwords')   # Stop word lists
nltk.download('punkt_tab')   # Extra tokenizer data
print("NLTK is ready!")
```

Notes : Run pip in terminal

- nltk.download() fetches data needed for NLP tasks
- No errors → installation successful

Tokenization

What is Tokenization?

Breaking text into smaller units called **tokens**.

Types

- **Word Tokenization** → splits into words/punctuation
- **Sentence Tokenization** → splits into full sentences

Word Tokenization Example

```
from nltk.tokenize import word_tokenize

text = "Python is fun to learn!"

words = word_tokenize(text)

print("Words:", words)
```

Output

```
['Python', 'is', 'fun', 'to', 'learn', '!']
```

Sentence Tokenization Example

Sentence Tokenization Example

```
from nltk.tokenize import sent_tokenize
text = "I love Python. NLTK is awesome!"
sentences = sent_tokenize(text)
print("Sentences:", sentences)
```

Output

```
['I love Python.', 'NLTK is awesome!']
```

Stop Word Removal

Stop Words

Common words like:

- the, is, and, to, a

These words don’t add much meaning → removed to focus on important terms.

Stop Word Removal

Stop Word Removal Example

```
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
text = "The dog runs fast and jumps high."
stop_words = set(stopwords.words('english'))
words = word_tokenize(text)
filtered_words = []
for word in words:
    if word.lower() not in stop_words:
        filtered_words.append(word)
print("Filtered Words:", filtered_words)
```

Output

```
['dog', 'runs', 'fast', 'jumps', 'high', '.']
```

Word Frequency Analysis

Concept : Counts how often each word appears — helps find key themes.

Word Frequency Example

```
from nltk.tokenize import word_tokenize

from nltk.probability import FreqDist

text = "Python is fun. Python is easy. I love Python."

words = word_tokenize(text.lower())  # Case-insensitive

freq = FreqDist(words)

print("Top 3 Words:", freq.most_common(3))
```

Output

```
[('python', 3), ('.', 3), ('is', 2)]
```

THANK YOU

Department of Computer Science and Engineering

Prof. Kundhavai K R, CSE Department

Ack: Teaching Assistant:
Adithya Jeyaramsankar– PES2UG22CS029