NLP assignment-2.4

Roll Number: 2403A52011
Batch Number: 02

Python Implementation

```python
import nltk
from nltk.stem import PorterStemmer
from nltk.stem import WordNetLemmatizer

# Ensure required resources are downloaded
nltk.download('punkt')
nltk.download('punkt_tab')
nltk.download('wordnet')

medical_text = """
Diabetes is a chronic disease that affects how the body processes blood sugar.
If untreated, diabetes may cause heart disease, kidney failure, nerve damage and vision problems.
Early diagnosis and proper treatment help improve patient outcomes.
"""

# Tokenization
tokens = nltk.word_tokenize(medical_text)

# Apply stemming
stemmer = PorterStemmer()
stemmed_tokens = [stemmer.stem(token) for token in tokens]

# Apply lemmatization
lemmatizer = WordNetLemmatizer()
lemmatized_tokens = [lemmatizer.lemmatize(token) for token in tokens]

# Output results
print("Stemmed Tokens:", stemmed_tokens)
print("Lemmatized Tokens:", lemmatized_tokens)
```

Output

```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
[nltk_data] Downloading package punkt_tab to /root/nltk_data...
[nltk_data]   Package punkt_tab is already up-to-date!
[nltk_data] Downloading package wordnet to /root/nltk_data...

Stemmed Tokens: ['diabet', 'is', 'a', 'chronic', 'diseas', 'that', 'affect', 'how', 'the'
    , 'bodi', 'process', 'blood', 'sugar', '.', 'if', 'untreat', ',', 'diabet', 'may', '
    caus', 'heart', 'diseas', ',', 'kidney', 'failur', ',', 'nerv', 'damag', 'and', '
    vision', 'problem', '.', 'earli', 'diagnosi', 'and', 'proper', 'treatment', 'help', '
    improv', 'patient', 'outcom', '.']
```

```
Lemmatized Tokens: ['Diabetes', 'is', 'a', 'chronic', 'disease', 'that', 'affect', 'how',
     'the', 'body', 'process', 'blood', 'sugar', '.', 'If', 'untreated', ',', 'diabetes',
     'may', 'cause', 'heart', 'disease', ',', 'kidney', 'failure', ',', 'nerve', 'damage'
    , 'and', 'vision', 'problem', '.', 'Early', 'diagnosis', 'and', 'proper', 'treatment'
    , 'help', 'improve', 'patient', 'outcome', '.']
```

Difference between Stemming and Lemmatization

Feature	Stemming	Lemmatization
Process	Cuts off the ends of words (suffixes) using heuristic rules.	Analyzes word context (POS tagging) to find the actual dictionary form (lemma).
Output	Often results in a non-dictionary, meaningless root (e.g., ”studies” $\rightarrow$ ”study”).	Always a valid, meaningful word (e.g., ”better” $\rightarrow$ ”good”, ”caring” $\rightarrow$ ”care”).
Speed	Very fast, as it’s a simple rule-based process.	Slower, as it requires linguistic knowledge and dictionary lookups.
Use Case	Information retrieval, search engines where speed is crucial.	Applications needing high accuracy, like chatbots, sentiment analysis.