```
import nltk
import spacy
from nltk.tokenize import sent_tokenize, word_tokenize
from nltk.stem import PorterStemmer
from nltk.corpus import stopwords
from nltk.stem import WordNetLemmatizer
nltk.download('wordnet')
nltk.download("stopwords")
nltk.download('punkt_tab')
```

```
[nltk_data] Downloading package wordnet to /root/nltk_data...
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Package stopwords 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!
True
```

```
medical_text = """
Diabetes mellitus is a chronic condition characterized by elevated blood glucose levels.
Patients with diabetes often require insulin therapy and regular monitoring.
Early diagnosis and lifestyle intervention can reduce complications.
"""
```

```
nltk_sentences = sent_tokenize(medical_text)
print("NLTK Sentences:")
nltk_sentences
```

```
NLTK Sentences:
['\nDiabetes mellitus is a chronic condition characterized by elevated blood glucose levels.',
 'Patients with diabetes often require insulin therapy and regular monitoring.',
 'Early diagnosis and lifestyle intervention can reduce complications.']
```

```
nltk_words = word_tokenize(medical_text)
print("NLTK Words:")
nltk_words
```

```
NLTK Words:
['Diabetes',
 'mellitus',
 'is',
 'a',
 'chronic',
 'condition',
 'characterized',
 'by',
 'elevated',
 'blood',
 'glucose',
 'levels',
 '.',
 'Patients',
 'with',
 'diabetes',
 'often',
 'require',
 'insulin',
 'therapy',
 'and',
 'regular',
 'monitoring',
 '.',
 'Early',
 'diagnosis',
 'and',
 'lifestyle',
 'intervention',
 'can',
 'reduce',
 'complications',
 '.']
```

```
stemmer = PorterStemmer()
words = nltk_words
stemmed_words = [stemmer.stem(word) for word in words]
stemmed_words
```

```
['diabet',
 'mellitu',
 'is',
 'a',
 'chronic',
```

```
 'condit',
 'character',
 'by',
 'elev',
 'blood',
 'glucos',
 'level',
 '.',
 'patient',
 'with',
 'diabet',
 'often',
 'requir',
 'insulin',
 'therapi',
 'and',
 'regular',
 'monitor',
 '.',
 'earli',
 'diagnosi',
 'and',
 'lifestyl',
 'intervent',
 'can',
 'reduc',
 'complic',
 '.']
```

```
lemmatizer = WordNetLemmatizer()
nltk_lemmas = [lemmatizer.lemmatize(token) for token in nltk_words]
print("NLTK Lemmatization Output:")
nltk_lemmas
```

```
NLTK Lemmatization Output:
['Diabetes',
 'mellitus',
 'is',
 'a',
 'chronic',
 'condition',
 'characterized',
 'by',
 'elevated',
 'blood',
 'glucose',
 'level',
 '.',
 'Patients',
 'with',
 'diabetes',
 'often',
 'require',
 'insulin',
 'therapy',
 'and',
 'regular',
 'monitoring',
 '.',
 'Early',
 'diagnosis',
 'and',
 'lifestyle',
 'intervention',
 'can',
 'reduce',
 'complication',
 '.']
```

Why Lemmatization is Critical in Healthcare NLP

Key Observations:

Stemming:

- Rule-based and aggressive
- Can truncate medical terms incorrectly
- Produces non-dictionary words (e.g., “diagnos”) **Lemmatization:**
- Context-aware
- Preserves medically valid root forms
- Produces dictionary-standard terminology

Importance in Healthcare NLP

```
Lemmatization is critical in healthcare NLP because:
```

Clinical Accuracy: Medical terms must retain semantic correctness. Incorrect stems can distort diagnoses, symptoms, or treatments.

Interoperability: Healthcare NLP systems integrate with ontologies (e.g., ICD, SNOMED). Lemmas align better with standardized vocabularies.

Patient Safety: Misinterpretation of clinical language can lead to incorrect insights or automated decisions.

Improved Model Performance: Lemmatized tokens improve downstream tasks such as:

- Clinical entity recognition
- Medical document classification
- Clinical decision support systems