IMPORTING NLTK AND SPACY

In [1]:
```
import nltk
import spacy
```

SAMPLE TEXT

In [ ]:
```
text = "The patient, a 45-year-old male, presented with a chief complaint of p
```

WORD TOKENIZATION

In [4]:
```
import spacy

# Load the English language model
# If you run into an error like "OSError: [E050] Can't find model 'en_core_web
# you might need to download it first by running: !python -m spacy download en

text = "The patient, a 45-year-old male, presented with a chief complaint of p

nlp = spacy.load("en_core_web_sm")

# The 'text' variable should already be defined from a previous cell.
# If not, uncomment the following line and paste the paragraph:
# text = "The patient, a 45-year-old male, presented with a chief complaint of

# Process the text
doc = nlp(text)

# Extract and print tokens
tokens = [token.text for token in doc]
print("Tokens (first 20):")
print(tokens[:20])
print(f"Total number of tokens: {len(tokens)}")
```

```
Tokens (first 20):
['The', 'patient', ',', 'a', '45', '-', 'year', '-', 'old', 'male', ',', 'prese
nted', 'with', 'a', 'chief', 'complaint', 'of', 'progressively', 'worsening',
'dyspnea']
Total number of tokens: 86
```

SENTENCE TOKENIZATION

In [5]:
```
# The 'doc' object from spaCy (processed text) should already be available fro
# If not, ensure the 'text' variable is defined and then re-run the previous c

# Extract and print sentences
sentences = [sent.text for sent in doc.sents]
print("Sentences:")
for i, sent in enumerate(sentences):
    print(f"Sentence {i+1}: {sent}")
print(f"Total number of sentences: {len(sentences)}")
```

```
Sentences:
Sentence 1: The patient, a 45-year-old male, presented with a chief complaint o
f progressively worsening dyspnea on exertion over the past three months.
Sentence 2: His medical history is significant for well-controlled hypertension
and a family history of coronary artery disease.
Sentence 3: Physical examination revealed bilateral fine crackles at the lung b
ases and 2+ pitting edema in both lower extremities.
Sentence 4: Initial laboratory findings showed elevated B-type natriuretic pept
ide (BNP) and a mild increase in troponin
Total number of sentences: 4
```

STEMMING

In [9]:

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

# Download necessary NLTK data if not already present
try:
    nltk.data.find('tokenizers/punkt_tab')
except LookupError:
    print("Downloading 'punkt_tab' for NLTK...")
    nltk.download('punkt_tab')

# The 'text' variable should already be defined from a previous cell.
# If not, please define it before running this cell.

# Initialize the Porter Stemmer
porter = PorterStemmer()

# Tokenize the text into words (using NLTK's word_tokenize)
words = nltk.word_tokenize(text)

# Perform stemming
stemmed_words = [porter.stem(word) for word in words]

print("Original words (first 20):")
print(words[:20])
print("\nStemmed words (first 20):")
print(stemmed_words[:20])
print(f"Total number of words after stemming: {len(stemmed_words)}")
```

```
Downloading 'punkt_tab' for NLTK...
[nltk_data] Downloading package punkt_tab to /root/nltk_data...
[nltk_data]   Unzipping tokenizers/punkt_tab.zip.
Original words (first 20):
['The', 'patient', ',', 'a', '45-year-old', 'male', ',', 'presented', 'with',
'a', 'chief', 'complaint', 'of', 'progressively', 'worsening', 'dyspnea', 'on',
'exertion', 'over', 'the']

Stemmed words (first 20):
['the', 'patient', ',', 'a', '45-year-old', 'male', ',', 'present', 'with',
'a', 'chief', 'complaint', 'of', 'progress', 'worsen', 'dyspnea', 'on', 'exer
t', 'over', 'the']
Total number of words after stemming: 77
```

LEMMATIZATION

In [10]:
```
# The 'doc' object from spaCy (processed text) should already be available fro
# If not, ensure the 'text' variable is defined and then re-run the spaCy proc

# Extract and print lemmas
lemmas = [token.lemma_ for token in doc]

print("Original words (first 20):")
print([token.text for token in doc][:20])
print("\nLemmas (first 20):")
print(lemmas[:20])
print(f"Total number of lemmas: {len(lemmas)}")
```

```
Original words (first 20):
['The', 'patient', ',', 'a', '45', '-', 'year', '-', 'old', 'male', ',', 'prese
nted', 'with', 'a', 'chief', 'complaint', 'of', 'progressively', 'worsening',
'dyspnea']

Lemmas (first 20):
['the', 'patient', ',', 'a', '45', '-', 'year', '-', 'old', 'male', ',', 'prese
nt', 'with', 'a', 'chief', 'complaint', 'of', 'progressively', 'worsen', 'dyspn
ea']
Total number of lemmas: 86
```

COMPARING OUTPUTS

In [11]:
```
# Ensure 'words', 'stemmed_words', and 'lemmas' are available from previous st
# If not, run the respective cells first.

print("{:<20} {:<20} {:<20}".format("Original Word", "Stemmed Word", "Lemma"))
print("-" * 60)

# Get a common length for comparison, using the shortest list to avoid index e
min_len = min(len(words), len(stemmed_words), len(lemmas))

# Display the first 20 comparisons or up to the minimum length if shorter
for i in range(min(20, min_len)):
    print("{:<20} {:<20} {:<20}".format(words[i], stemmed_words[i], lemmas[i]))
```

```
Original Word        Stemmed Word         Lemma
------------------------------------------------------------
The                  the                  the
patient              patient              patient
,                    ,                    ,
a                    a                    a
45-year-old          45-year-old          45
male                 male                 -
,                    ,                    year
presented            present              -
with                 with                 old
a                    a                    male
chief                chief                ,
complaint            complaint            present
of                   of                   with
progressively        progress             a
worsening            worsen               chief
dyspnea              dyspnea              complaint
on                   on                   of
exertion             exert                progressively
over                 over                 worsen
the                  the                  dyspnea
```

In [ ]: