```
import sys
!{sys.executable} -m pip install nltk spacy

import nltk
from nltk.tokenize import word_tokenize
from nltk.stem import PorterStemmer, WordNetLemmatizer
import spacy
from spacy.lang.en import English

# Download necessary NLTK data
nltk.download('punkt')
nltk.download('wordnet')
nltk.download('punkt_tab') # Added to resolve LookupError for punkt_tab

# Load spaCy English model
try:
    nlp = spacy.load('en_core_web_sm')
except OSError:
    print('Downloading spaCy model en_core_web_sm. This may take a while...')
    !{sys.executable} -m spacy download en_core_web_sm
    nlp = spacy.load('en_core_web_sm')

print("Libraries installed and imported successfully.")
```

```
Requirement already satisfied: nltk in /usr/local/lib/python3.12/dist-packages (3.9.1)
Requirement already satisfied: spacy in /usr/local/lib/python3.12/dist-packages (3.8.11)
Requirement already satisfied: click in /usr/local/lib/python3.12/dist-packages (from nltk) (8.3.1)
Requirement already satisfied: joblib in /usr/local/lib/python3.12/dist-packages (from nltk) (1.5.3)
Requirement already satisfied: regex>=2021.8.3 in /usr/local/lib/python3.12/dist-packages (from nltk) (2025.11.3)
Requirement already satisfied: tqdm in /usr/local/lib/python3.12/dist-packages (from nltk) (4.67.1)
Requirement already satisfied: spacy-legacy<3.1.0,>=3.0.11 in /usr/local/lib/python3.12/dist-packages (from spacy) (3.0.12)
Requirement already satisfied: spacy-loggers<2.0.0,>=1.0.0 in /usr/local/lib/python3.12/dist-packages (from spacy) (1.0.5)
Requirement already satisfied: murmurhash<1.1.0,>=0.28.0 in /usr/local/lib/python3.12/dist-packages (from spacy) (1.0.15)
Requirement already satisfied: cymem<2.1.0,>=2.0.2 in /usr/local/lib/python3.12/dist-packages (from spacy) (2.0.13)
Requirement already satisfied: preshed<3.1.0,>=3.0.2 in /usr/local/lib/python3.12/dist-packages (from spacy) (3.0.12)
Requirement already satisfied: thinc<8.4.0,>=8.3.4 in /usr/local/lib/python3.12/dist-packages (from spacy) (8.3.10)
Requirement already satisfied: wasabi<1.2.0,>=0.9.1 in /usr/local/lib/python3.12/dist-packages (from spacy) (1.1.3)
Requirement already satisfied: srsly<3.0.0,>=2.4.3 in /usr/local/lib/python3.12/dist-packages (from spacy) (2.5.2)
Requirement already satisfied: catalogue<2.1.0,>=2.0.6 in /usr/local/lib/python3.12/dist-packages (from spacy) (2.0.10)
Requirement already satisfied: weasel<0.5.0,>=0.4.2 in /usr/local/lib/python3.12/dist-packages (from spacy) (0.4.3)
Requirement already satisfied: typer-slim<1.0.0,>=0.3.0 in /usr/local/lib/python3.12/dist-packages (from spacy) (0.20.0)
Requirement already satisfied: numpy>=1.19.0 in /usr/local/lib/python3.12/dist-packages (from spacy) (2.0.2)
Requirement already satisfied: requests<3.0.0,>=2.13.0 in /usr/local/lib/python3.12/dist-packages (from spacy) (2.32.4)
Requirement already satisfied: pydantic!=1.8,!=1.8.1,<3.0.0,>=1.7.4 in /usr/local/lib/python3.12/dist-packages (from spacy)
Requirement already satisfied: jinja2 in /usr/local/lib/python3.12/dist-packages (from spacy) (3.1.6)
Requirement already satisfied: setuptools in /usr/local/lib/python3.12/dist-packages (from spacy) (75.2.0)
Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python3.12/dist-packages (from spacy) (25.0)
Requirement already satisfied: annotated-types>=0.6.0 in /usr/local/lib/python3.12/dist-packages (from pydantic!=1.8,!=1.8.1
Requirement already satisfied: pydantic-core==2.41.4 in /usr/local/lib/python3.12/dist-packages (from pydantic!=1.8,!=1.8.1,
Requirement already satisfied: typing-extensions>=4.14.1 in /usr/local/lib/python3.12/dist-packages (from pydantic!=1.8,!=1.
Requirement already satisfied: typing-inspection>=0.4.2 in /usr/local/lib/python3.12/dist-packages (from pydantic!=1.8,!=1.8
Requirement already satisfied: charset_normalizer<4,>=2 in /usr/local/lib/python3.12/dist-packages (from requests<3.0.0,>=2.
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.12/dist-packages (from requests<3.0.0,>=2.13.0->spacy)
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.12/dist-packages (from requests<3.0.0,>=2.13.0->
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.12/dist-packages (from requests<3.0.0,>=2.13.0->
Requirement already satisfied: blis<1.4.0,>=1.3.0 in /usr/local/lib/python3.12/dist-packages (from thinc<8.4.0,>=8.3.4->spac
Requirement already satisfied: confection<1.0.0,>=0.0.1 in /usr/local/lib/python3.12/dist-packages (from thinc<8.4.0,>=8.3.4
Requirement already satisfied: cloudpathlib<1.0.0,>=0.7.0 in /usr/local/lib/python3.12/dist-packages (from weasel<0.5.0,>=0.
Requirement already satisfied: smart-open<8.0.0,>=5.2.1 in /usr/local/lib/python3.12/dist-packages (from weasel<0.5.0,>=0.4.
Requirement already satisfied: MarkupSafe>=2.0 in /usr/local/lib/python3.12/dist-packages (from jinja2->spacy) (3.0.3)
Requirement already satisfied: wrapt in /usr/local/lib/python3.12/dist-packages (from smart-open<8.0.0,>=5.2.1->weasel<0.5.0
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Unzipping tokenizers/punkt.zip.
[nltk_data] Downloading package wordnet to /root/nltk_data...
[nltk_data] Downloading package punkt_tab to /root/nltk_data...
[nltk_data]   Unzipping tokenizers/punkt_tab.zip.
Libraries installed and imported successfully.
```

```
print(f"{"Original":<15} | {"Stemmed (NLTK)":<20} | {"Lemmatized (spaCy)":<20}")
print(f"{"-"*15} | {"-"*20} | {"-"*20}")

# Assuming nltk_tokens, stemmed_tokens, and spacy_lemmas are available from previous steps
# If lists are of different lengths due to spaCy's more granular tokenization, we'll align them based on NLTK tokens.

# Create a mapping for spaCy lemmas to align with NLTK tokens more easily for comparison
# This is a simplification; a direct word-to-lemma mapping is complex due to tokenization differences
# For simplicity, we'll try to match by index, acknowledging that spaCy splits some words NLTK doesn't.

# Let's create a more direct comparison for the first few tokens, as direct index comparison can be tricky
# due to spaCy splitting hyphens and adding spaces as tokens.

# For a more robust comparison, let's create a combined list, handling potential length differences.
# We'll use the shorter of the two lemma lists (NLTK or spaCy) to avoid index errors.
```

```
# Let's ensure the lists used for comparison are roughly aligned. The simplest way is to iterate over NLTK tokens
# and try to find a corresponding lemma. This is an approximation given spaCy's detailed tokenization.

# A more practical approach for comparison is to take a subset and illustrate key differences.

print(f"{"Original":<15} | {"Stemmed (NLTK)":<20} | {"Lemmatized (NLTK)":<20} | {"Lemmatized (spaCy)":<20}")
print(f"{"-"*15} | {"-"*20} | {"-"*20} | {"-"*20}")

# Iterate through the NLTK tokens and show corresponding stemmed and lemmatized forms
# Note: Alignment between NLTK tokens and spaCy lemmas can be imperfect due to tokenization differences
# We'll use a simplified alignment for demonstration purposes.
for i in range(min(len(nltk_tokens), len(stemmed_tokens), len(nltk_lemmas), len(spacy_lemmas))):
    original = nltk_tokens[i]
    stemmed = stemmed_tokens[i]
    nltk_lemma = nltk_lemmas[i]
    spacy_lemma = spacy_lemmas[i] if i < len(spacy_lemmas) else '' # Handle potential length mismatch
    print(f"{original:<15} | {stemmed:<20} | {nltk_lemma:<20} | {spacy_lemma:<20}")

print("\nKey observations from the comparison:")
print("- Stemming often results in truncated words that may not be actual words (e.g., 'presented' -> 'present', 'crackles
print("- NLTK's WordNetLemmatizer, without explicit POS tags, might not always produce the correct base form (e.g., 'preser
print("- spaCy's lemmatization is generally more accurate as it considers the word's part of speech (e.g., 'presented' ->
```

```
Original        | Stemmed (NLTK)       | Lemmatized (spaCy)
--------------- | -------------------- | --------------------
Original        | Stemmed (NLTK)       | Lemmatized (NLTK)    | Lemmatized (spaCy)
--------------- | -------------------- | -------------------- | --------------------
The             | the                  | The                  | the
patient         | patient              | patient              | patient
presented       | present              | presented            | present
with            | with                 | with                 | with
a               | a                    | a                    | a
persistent      | persist              | persistent           | persistent
cough           | cough                | cough                | cough
,               | ,                    | ,                    | ,
shortness       | short                | shortness            | shortness
of              | of                   | of                   | of
breath          | breath               | breath               | breath
,               | ,                    | ,                    | ,
and             | and                  | and                  | and
a               | a                    | a                    | a
low-grade       | low-grad             | low-grade            | low
fever           | fever                | fever                | -
for             | for                  | for                  | grade
the             | the                  | the                  | fever
past            | past                 | past                 | for
three           | three                | three                | the
days            | day                  | day                  | past
.               | .                    | .                    | three
Auscultation    | auscult              | Auscultation         | day
revealed        | reveal               | revealed             | .
bilateral       | bilater              | bilateral            |
crackles        | crackl               | crackle              | Auscultation
in              | in                   | in                   | reveal
the             | the                  | the                  | bilateral
lower           | lower                | lower                | crackle
lobes           | lobe                 | lobe                 | in
.               | .                    | .                    | the
Chest           | chest                | Chest                | low
X-ray           | x-ray                | X-ray                | lobe
showed          | show                 | showed               | .
patchy          | patchi               | patchy               | Chest
infiltrates     | infiltr              | infiltrates          | X
consistent      | consist              | consistent           | -
with            | with                 | with                 | ray
pneumonia       | pneumonia            | pneumonia            | show
.               | .                    | .                    | patchy
Antibiotic      | antibiot             | Antibiotic           | infiltrate
treatment       | treatment            | treatment            | consistent
with            | with                 | with                 | with
Azithromycin    | azithromycin         | Azithromycin         | pneumonia
was             | wa                   | wa                   | .
initiated       | initi                | initiated            |
,               | ,                    | ,                    | antibiotic
along           | along                | along                | treatment
with            | with                 | with                 | with
supportive      | support              | supportive           | Azithromycin
care            | care                 | care                 | be
including       | includ               | including            | initiate
oxygen          | oxygen               | oxygen               | ,
```

```
# NLTK Lemmatization (WordNetLemmatizer)
lemmatizer = WordNetLemmatizer()
# For simplicity, we'll try to lemmatize without explicit POS tags first
nltk_lemmas = [lemmatizer.lemmatize(word) for word in nltk_tokens]
```

```
print("\nLemmatized Tokens (NLTK WordNetLemmatizer):")
print(nltk_lemmas)

# spaCy Lemmatization
spacy_lemmas = [token.lemma_ for token in spacy_doc]
print("\nLemmatized Tokens (spaCy):")
print(spacy_lemmas)
```

```
Lemmatized Tokens (NLTK WordNetLemmatizer):
['The', 'patient', 'presented', 'with', 'a', 'persistent', 'cough', ',', 'shortness', 'of', 'breath', ',', 'and', 'a', 'low-

Lemmatized Tokens (spaCy):
['the', 'patient', 'present', 'with', 'a', 'persistent', 'cough', ',', 'shortness', 'of', 'breath', ',', 'and', 'a', 'low',
```

```
# NLTK Stemming (Porter Stemmer)
stemmer = PorterStemmer()
stemmed_tokens = [stemmer.stem(word) for word in nltk_tokens]
print("\nStemmed Tokens (NLTK Porter Stemmer):")
print(stemmed_tokens)
```

```
Stemmed Tokens (NLTK Porter Stemmer):
['the', 'patient', 'present', 'with', 'a', 'persist', 'cough', ',', 'short', 'of', 'breath', ',', 'and', 'a', 'low-grad', 'f
```

```
# NLTK Word Tokenization
nltk_tokens = word_tokenize(medical_text)
print("\nWord Tokenization (NLTK):")
print(nltk_tokens)

# spaCy Word Tokenization
spacy_doc = nlp(medical_text)
spacy_tokens = [token.text for token in spacy_doc]
print("\nWord Tokenization (spaCy):")
print(spacy_tokens)
```

```
Word Tokenization (NLTK):
['The', 'patient', 'presented', 'with', 'a', 'persistent', 'cough', ',', 'shortness', 'of', 'breath', ',', 'and', 'a', 'low-

Word Tokenization (spaCy):
['The', 'patient', 'presented', 'with', 'a', 'persistent', 'cough', ',', 'shortness', 'of', 'breath', ',', 'and', 'a', 'low'
```

```
import nltk
from nltk.tokenize import sent_tokenize

sentences = sent_tokenize(medical_text)
print("Sentence Tokenization (NLTK):")
for i, sentence in enumerate(sentences):
    print(f"Sentence {i+1}: {sentence}")
```

```
---------------------------------------------------------------------------
LookupError                               Traceback (most recent call last)
/tmp/ipython-input-4079946051.py in <cell line: 0>()
      2 from nltk.tokenize import sent_tokenize
      3
----> 4 sentences = sent_tokenize(medical_text)
      5 print("Sentence Tokenization (NLTK):")
      6 for i, sentence in enumerate(sentences):

                                 4 frames
/usr/local/lib/python3.12/dist-packages/nltk/data.py in find(resource_name, paths)
    577     sep = "*" * 70
    578     resource_not_found = f"\n{sep}\n{msg}\n{sep}\n"
--> 579     raise LookupError(resource_not_found)
    580
    581

LookupError:
**********************************************************************
  Resource punkt_tab not found.
  Please use the NLTK Downloader to obtain the resource:

  >>> import nltk
  >>> nltk.download('punkt_tab')

  For more information see: https://www.nltk.org/data.html

  Attempted to load tokenizers/punkt_tab/english/

  Searched in:
    - '/root/nltk_data'
    - '/usr/nltk_data'
    - '/usr/share/nltk_data'
    - '/usr/lib/nltk_data'
    - '/usr/share/nltk_data'
    - '/usr/local/share/nltk_data'
    - '/usr/lib/nltk_data'
    - '/usr/local/lib/nltk_data'
**********************************************************************
```