Name: Akhil
Batch: 13(AIML)
Hallticket: 2403A52344

```
# Install required libraries
!pip install nltk spacy
!python -m spacy download en_core_web_sm
```

```
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
Collecting en-core-web-sm==3.8.0
  Downloading https://github.com/explosion/spacy-models/releases/download/en_core_web_sm-3.8.0/en_core_web_sm-3.8.0-py3-none
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 12.8/12.8 MB 72.5 MB/s eta 0:00:00
✓ Download and installation successful
You can now load the package via spacy.load('en_core_web_sm')
⚠ Restart to reload dependencies
If you are in a Jupyter or Colab notebook, you may need to restart Python in
order to load all the package's dependencies. You can do this by selecting the
'Restart kernel' or 'Restart runtime' option.
```

```
import nltk
import spacy
from nltk.tokenize import sent_tokenize, word_tokenize
from nltk.stem import PorterStemmer, WordNetLemmatizer

nltk.download('punkt')
nltk.download('wordnet')
nltk.download('omw-1.4')
```

```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
[nltk_data] Downloading package wordnet to /root/nltk_data...
[nltk_data]   Package wordnet is already up-to-date!
[nltk_data] Downloading package omw-1.4 to /root/nltk_data...
[nltk_data]   Package omw-1.4 is already up-to-date!
True
```

```
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.
"""
```

```
nltk.download('punkt_tab') # Download the missing resource
nltk_sentences = sent_tokenize(medical_text)
nltk_sentences
```

```
[nltk_data] Downloading package punkt_tab to /root/nltk_data...
[nltk_data]   Package punkt_tab is already up-to-date!
['\nDiabetes 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.']
```

```
nlp = spacy.load("en_core_web_sm")
doc = nlp(medical_text)

spacy_sentences = [sent.text for sent in doc.sents]
spacy_sentences
```

```
['\nDiabetes is a chronic disease that affects how the body processes blood sugar.\n',
 'If untreated, diabetes may cause heart disease, kidney failure, nerve damage and vision problems.\n',
 'Early diagnosis and proper treatment help improve patient outcomes.\n']
```

```
nltk_words = word_tokenize(medical_text)
nltk_words
```

```
['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',
 '.']
```

```
spacy_words = [token.text for token in doc]
spacy_words
```

```
['\n',
 'Diabetes',
 'is',
 'a',
 'chronic',
 'disease',
 'that',
 'affects',
 'how',
 'the',
 'body',
 'processes',
 'blood',
 'sugar',
```

```
 '.',
 '\n',
 'If',
 'untreated',
 ',',
 'diabetes',
 'may',
 'cause',
 'heart',
 'disease',
 ',',
 'kidney',
 'failure',
 ',',
 'nerve',
 'damage',
 'and',
 'vision',
 'problems',
 '.',
 '\n',
 'Early',
 'diagnosis',
 'and',
 'proper',
 'treatment',
 'help',
 'improve',
 'patient',
 'outcomes',
 '.',
 '\n']
```

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

```
['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']
```

```
lemmatizer = WordNetLemmatizer()
nltk_lemmas = [lemmatizer.lemmatize(word) for word in nltk_words if word.isalpha()]
nltk_lemmas
```

```
['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']
```

```
spacy_lemmas = [(token.text, token.lemma_) for token in doc if token.is_alpha]
spacy_lemmas
```

```
[('Diabetes', 'Diabetes'),
 ('is', 'be'),
 ('a', 'a'),
 ('chronic', 'chronic'),
 ('disease', 'disease'),
 ('that', 'that'),
 ('affects', 'affect'),
 ('how', 'how'),
 ('the', 'the'),
 ('body', 'body'),
 ('processes', 'process'),
 ('blood', 'blood'),
 ('sugar', 'sugar'),
 ('If', 'if'),
 ('untreated', 'untreate'),
 ('diabetes', 'diabete'),
 ('may', 'may'),
 ('cause', 'cause'),
 ('heart', 'heart'),
 ('disease', 'disease'),
 ('kidney', 'kidney'),
 ('failure', 'failure'),
 ('nerve', 'nerve'),
 ('damage', 'damage'),
 ('and', 'and'),
 ('vision', 'vision'),
 ('problems', 'problem'),
 ('Early', 'early'),
 ('diagnosis', 'diagnosis'),
 ('and', 'and'),
 ('proper', 'proper'),
 ('treatment', 'treatment'),
 ('help', 'help'),
 ('improve', 'improve'),
 ('patient', 'patient'),
 ('outcomes', 'outcome')]
```

Original Word	Stemmed Form	Lemmatized Form
diabetes	diabet	diabetes
disease	diseas	disease
processes	process	process
untreated	untreated	untreated
diagnosis	diagnosi	diagnosis
outcomes	outcom	outcome

Why Lemmatization is Critical in Healthcare NLP Key Reasons

Preserves Clinical Meaning

Stemming may produce invalid or ambiguous medical roots

Lemmatization returns real dictionary terms

Medical Terminology is Highly Sensitive

Diagnosis ≠ Diagnostic ≠ Diagnose

Lemmatization understands grammatical context

Improves Clinical NLP Tasks

Named Entity Recognition (NER)

Clinical decision support

Electronic Health Record (EHR) analysis

Medical coding (ICD, SNOMED)

spaCy Advantage

Context-aware lemmatization

Better handling of biomedical text when extended with medical models (e.g., scispaCy)

Conclusion

NLTK is useful for learning and basic NLP pipelines

spaCy provides more accurate and context-aware processing

Stemming is not recommended for healthcare applications

Lemmatization is essential to preserve clinical accuracy and patient safety

```
Start coding or generate with AI.
```