2403A52031

AKULA MANOJ

LAB ASSIGNMENT-2.4

```
!pip install nltk
```

```
Requirement already satisfied: nltk in /usr/local/lib/python3.12/dist-
packages (3.9.1)
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)
```

```
!pip install spacy
```

```
Requirement already satisfied: spacy in
/usr/local/lib/python3.12/dist-packages (3.8.11)
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: tqdm<5.0.0,>=4.38.0 in
/usr/local/lib/python3.12/dist-packages (from spacy) (4.67.1)
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) (2.12.3)
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,<3.0.0,>=1.7.4->spacy) (0.7.0)
Requirement already satisfied: pydantic-core==2.41.4 in
/usr/local/lib/python3.12/dist-packages (from pydantic!=1.8,!
=1.8.1,<3.0.0,>=1.7.4->spacy) (2.41.4)
Requirement already satisfied: typing-extensions>=4.14.1 in
/usr/local/lib/python3.12/dist-packages (from pydantic!=1.8,!
=1.8.1,<3.0.0,>=1.7.4->spacy) (4.15.0)
Requirement already satisfied: typing-inspection>=0.4.2 in
/usr/local/lib/python3.12/dist-packages (from pydantic!=1.8,!
=1.8.1,<3.0.0,>=1.7.4->spacy) (0.4.2)
Requirement already satisfied: charset_normalizer<4,>=2 in
/usr/local/lib/python3.12/dist-packages (from requests<3.0.0,>=2.13.0-
>spacy) (3.4.4)
Requirement already satisfied: idna<4,>=2.5 in
/usr/local/lib/python3.12/dist-packages (from requests<3.0.0,>=2.13.0-
>spacy) (3.11)
Requirement already satisfied: urllib3<3,>=1.21.1 in
/usr/local/lib/python3.12/dist-packages (from requests<3.0.0,>=2.13.0-
>spacy) (2.5.0)
Requirement already satisfied: certifi>=2017.4.17 in
/usr/local/lib/python3.12/dist-packages (from requests<3.0.0,>=2.13.0-
>spacy) (2025.11.12)
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-
>spacy) (1.3.3)
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-
>spacy) (0.1.5)
Requirement already satisfied: click>=8.0.0 in
/usr/local/lib/python3.12/dist-packages (from typer-
slim<1.0.0,>=0.3.0->spacy) (8.3.1)
Requirement already satisfied: cloudpathlib<1.0.0,>=0.7.0 in
/usr/local/lib/python3.12/dist-packages (from weasel<0.5.0,>=0.4.2-
>spacy) (0.23.0)
Requirement already satisfied: smart-open<8.0.0,>=5.2.1 in
/usr/local/lib/python3.12/dist-packages (from weasel<0.5.0,>=0.4.2-
>spacy) (7.5.0)
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,>=0.4.2->spacy) (2.0.1)
```

```
import nltk

import spacy

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.
"""
print(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.
```

SENTENCE TOKENIZATION

```
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.
"""
sentences = nltk.sent_tokenize(medical_text)
print(sentences)
```

```
['\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.']
```

WORD TOKANIZATION

```
words = nltk.word_tokenize(medical_text)
print(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', '.']
```

STEMMING

```
from nltk.stem import PorterStemmer

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

print("Original words:", words)
print("Stemmed words:", stemmed_words)

Original 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', '.']
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', '.']
```

LEMMATIZATION

```
!python -m spacy download en_core_web_sm

nlp = spacy.load('en_core_web_sm')

doc = nlp(medical_text)

# Filter out newline characters explicitly for cleaner display in the
comparison
lemmatized_words = [token.lemma_ for token in doc if not
token.is_punct and token.text.strip() != '']
original_words_for_lemma = [token.text for token in doc if not
token.is_punct and token.text.strip() != '']

print("Original words:", original_words_for_lemma)
print("Lemmatized words:", lemmatized_words)

Collecting en-core-web-sm==3.8.0
  Downloading
https://github.com/explosion/spacy-models/releases/download/en_core_we
b_sm-3.8.0/en_core_web_sm-3.8.0-py3-none-any.whl (12.8 MB)
```

```
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 12.8/12.8 MB 45.9 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.
Original 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']
Lemmatized words: ['Diabetes', 'be', 'a', 'chronic', 'disease',
'that', 'affect', 'how', 'the', 'body', 'process', 'blood', 'sugar',
'if', 'untreate', 'diabete', 'may', 'cause', 'heart', 'disease',
'kidney', 'failure', 'nerve', 'damage', 'and', 'vision', 'problem',
'early', 'diagnosis', 'and', 'proper', 'treatment', 'help', 'improve',
'patient', 'outcome']
```

comparing original words, stemmed words, and lemmas

```python
import pandas as pd
import string

# Filter NLTK words and stemmed words to remove punctuation for a
cleaner comparison
# We'll consider a word as 'not punctuation' if it's not in
string.punctuation
nltk_words_filtered = [word for word in words if word not in
string.punctuation]
nltk_stemmed_words_filtered = [stemmed_words[i] for i, word in
enumerate(words) if word not in string.punctuation]

# The SpaCy lists (original_words_for_lemma and lemmatized_words) are
already filtered for punctuation and newlines
spacy_original_words_filtered = original_words_for_lemma
spacy_lemmatized_words_filtered = lemmatized_words

# Ensure all filtered lists have the same length by padding with None
if necessary
max_len_filtered = max(len(nltk_words_filtered),
len(nltk_stemmed_words_filtered),
                       len(spacy_original_words_filtered),
len(spacy_lemmatized_words_filtered))
```

```
padded_nltk_words = nltk_words_filtered + [None] * (max_len_filtered -
len(nltk_words_filtered))
padded_nltk_stemmed = nltk_stemmed_words_filtered + [None] *
(max_len_filtered - len(nltk_stemmed_words_filtered))
padded_spacy_original = spacy_original_words_filtered + [None] *
(max_len_filtered - len(spacy_original_words_filtered))
padded_spacy_lemmatized = spacy_lemmatized_words_filtered + [None] *
(max_len_filtered - len(spacy_lemmatized_words_filtered))

# Create a DataFrame for a neat comparison
data_neat = {
    'Original (NLTK Filtered)': padded_nltk_words,
    'Stemmed (NLTK Filtered)': padded_nltk_stemmed,
    'Original (SpaCy Filtered)': padded_spacy_original,
    'Lemmatized (SpaCy Filtered)': padded_spacy_lemmatized
}

df_neat = pd.DataFrame(data_neat)
print(df_neat.to_string())
```

```
   Original (NLTK Filtered) Stemmed (NLTK Filtered) Original (SpaCy
Filtered) Lemmatized (SpaCy Filtered)
0                  Diabetes                  diabet
Diabetes                    Diabetes
1                        is                      is
is                          be
2                         a                       a
a                           a
3                   chronic                 chronic
chronic                     chronic
4                   disease                  diseas
disease                     disease
5                      that                    that
that                        that
6                   affects                  affect
affects                      affect
7                       how                     how
how                         how
8                       the                     the
the                         the
9                      body                    bodi
body                        body
10                processes                 process
processes                     process
11                    blood                   blood
blood                       blood
12                    sugar                   sugar
sugar                       sugar
13                       If                      if
If                          if
```

```
14                untreated                 untreat
untreated                    untreate
15                 diabetes                  diabet
diabetes                     diabete
16                      may                     may
may                         may
17                    cause                    caus
cause                       cause
18                    heart                   heart
heart                       heart
19                  disease                  diseas
disease                     disease
20                   kidney                  kidney
kidney                      kidney
21                  failure                  failur
failure                     failure
22                    nerve                    nerv
nerve                       nerve
23                   damage                   damag
damage                      damage
24                      and                     and
and                         and
25                   vision                  vision
vision                      vision
26                 problems                 problem
problems                     problem
27                    Early                   earli
Early                       early
28                diagnosis                diagnosi
diagnosis                   diagnosis
29                      and                     and
and                         and
30                   proper                  proper
proper                      proper
31                treatment               treatment
treatment                   treatment
32                     help                    help
help                        help
33                  improve                  improv
improve                     improve
34                  patient                 patient
patient                     patient
35                 outcomes                  outcom
outcomes                     outcome
```