import nltk

from nltk.tokenize import WhitespaceTokenizer, WordPunctTokenizer, TreebankWordTokenizer, TweetTokenizer

from nltk.tokenize import MWETokenizer

from nltk.stem import PorterStemmer, SnowballStemmer

from nltk.stem.wordnet import WordNetLemmatizer

# Sample text

text = "Tokenization is a key task in NLP. It breaks text into tokens, which can be words, phrases, or symbols."

# Tokenization

# Whitespace Tokenizer

whitespace\_tokenizer = WhitespaceTokenizer()

whitespace\_tokens = whitespace\_tokenizer.tokenize(text)

print("Whitespace Tokenizer:", whitespace\_tokens)

# Punctuation-based Tokenizer

punct\_tokenizer = WordPunctTokenizer()

punct\_tokens = punct\_tokenizer.tokenize(text)

print("Punctuation-based Tokenizer:", punct\_tokens)

# Treebank Tokenizer

treebank\_tokenizer = TreebankWordTokenizer()

treebank\_tokens = treebank\_tokenizer.tokenize(text)

print("Treebank Tokenizer:", treebank\_tokens)

# Tweet Tokenizer

tweet\_tokenizer = TweetTokenizer()

tweet\_tokens = tweet\_tokenizer.tokenize(text)

print("Tweet Tokenizer:", tweet\_tokens)

# Multi-Word Expression Tokenizer

mwe\_tokenizer = MWETokenizer()

mwe\_tokenizer.add\_mwe(("key", "task"))

mwe\_tokens = mwe\_tokenizer.tokenize(text.split())

print("MWE Tokenizer:", mwe\_tokens)

# Stemming

porter\_stemmer = PorterStemmer()

snowball\_stemmer = SnowballStemmer("english")

porter\_stems = [porter\_stemmer.stem(token) for token in treebank\_tokens]

print("Porter Stemmer:", porter\_stems)

snowball\_stems = [snowball\_stemmer.stem(token) for token in treebank\_tokens]

print("Snowball Stemmer:", snowball\_stems)

# Lemmatization

lemmatizer = WordNetLemmatizer()

lemmatized\_tokens = [lemmatizer.lemmatize(token) for token in treebank\_tokens]

print("Lemmatization:", lemmatized\_tokens)