**NLTK Detailed Step-by-Step Guide**

### 1. Installation Commands

Before using NLTK, we need to install the library and download necessary datasets.

pip install nltk

After installation, open Python and run:

import nltk  
nltk.download('punkt') # For tokenization  
nltk.download('wordnet') # For lemmatization  
nltk.download('stopwords') # For removing common words

### 2. Tokenization

Tokenization is the process of breaking text into smaller units.

from nltk.tokenize import sent\_tokenize, word\_tokenize, TreebankWordTokenizer  
  
text = "NLTK is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces."  
  
# Sentence Tokenization  
sentences = sent\_tokenize(text)  
print(sentences)  
  
# Word Tokenization  
words = word\_tokenize(text)  
print(words)  
  
# Treebank Word Tokenization  
tokenizer = TreebankWordTokenizer()  
treebank\_tokens = tokenizer.tokenize(text)  
print(treebank\_tokens)

### 3. Stemming

Stemming reduces words to their base/root form.

from nltk.stem import PorterStemmer, RegexpStemmer, SnowballStemmer  
  
ps = PorterStemmer()  
print(ps.stem("running"))  
  
regex\_stemmer = RegexpStemmer('ing$')  
print(regex\_stemmer.stem("running"))  
  
snowball = SnowballStemmer('english')  
print(snowball.stem("running"))

### 4. Lemmatization

Lemmatization reduces words to their dictionary form using vocabulary and morphological analysis.

from nltk.stem import WordNetLemmatizer  
  
lemmatizer = WordNetLemmatizer()  
print(lemmatizer.lemmatize("running", pos="v"))  
print(lemmatizer.lemmatize("better", pos="a"))

### 5. Stopwords Removal

Stopwords are common words (like “is”, “the”, “and”) usually removed in NLP.

from nltk.corpus import stopwords  
  
stop\_words = set(stopwords.words('english'))  
filtered\_words = [word for word in words if word.lower() not in stop\_words]  
print(filtered\_words)

### 6. Text Preprocessing Example

A complete example of preprocessing.

text = "NLTK provides easy-to-use interfaces for over 50 corpora and lexical resources!"  
  
# Tokenize  
words = word\_tokenize(text)  
  
# Lowercase  
words = [w.lower() for w in words]  
  
# Remove stopwords  
words = [w for w in words if w not in stop\_words]  
  
# Lemmatize  
words = [lemmatizer.lemmatize(w) for w in words]  
  
print(words)