NLP COMPLETE FLOW (MOST IMPORTANT ⭑)

Text Data

↓

Text Cleaning

↓

Tokenization

↓

Remove Stopwords

↓

Convert Text → Numbers

↓

ML Model

↓

Prediction

Required Libraries

- pip install nltk
- pip install scikit-learn
- pip install pandas

USES OF LIBRARY IN NLP

Library	Use
NLTK	Text processing

Library	Use
Pandas	**Dataset handle**
Scikit-learn	**ML + Vectorization**
String	**Punctuation remove**

🔥 PRACTICAL NLP START

STEP 1: TEXT INPUT

```
text = "I love Artificial Intelligence and Machine Learning!"

print(text)
```

Output:

```
I love Artificial Intelligence and Machine Learning!
```

LOWERCASE

```
text = text.lower()

print(text)
```

Output:

```
i love artificial intelligence and machine learning!
```

☞ **Why?**

Computer "Love" aur "love" ko alag samajhta hai ❌

STEP 3: TOKENIZATION

☞ CONVERT SENTENCE TO WORDS

```
import nltk
from nltk.tokenize import word_tokenize

nltk.download('punkt')

text = "i love artificial intelligence and machine learning"

tokens = word_tokenize(text)
print(tokens)
```

OUTPUT:

```
['i', 'love', 'artificial', 'intelligence', 'and', 'machine',
'learning']
```

STEP 4: REMOVE PUNCTUATION

```
import string

tokens = ['i', 'love', 'artificial', 'intelligence', '!', 'and']

clean_tokens = []

for word in tokens:
    if word not in string.punctuation:
        clean_tokens.append(word)
print(clean_tokens)
```

Output:

```
['i', 'love', 'artificial', 'intelligence',
'and']
```

STEP 5: REMOVE STOPWORDS

☞ **Stopwords = useless words**
Example:

is, am, are, and, the, in

```
from nltk.corpus import stopwords

nltk.download('stopwords')

stop_words = stopwords.words('english')

words = ['i', 'love', 'artificial', 'intelligence', 'and']

final_words = []

for word in words:
    if word not in stop_words:
        final_words.append(word)

print(final_words)
```

OUTPUT:-

```
['love', 'artificial', 'intelligence']
```

TEXT CLEANING CODE

```
import nltk
import string
from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords

nltk.download('punkt')
nltk.download('stopwords')

text = "I love Artificial Intelligence and Machine Learning!"

# Lowercase
text = text.lower()

# Tokenization
tokens = word_tokenize(text)

# Remove punctuation
tokens = [word for word in tokens if word not in string.punctuation]

# Remove stopwords
stop_words = stopwords.words('english')
clean_words = [word for word in tokens if word not in stop_words]

print("Final Clean Words:", clean_words)
```