

# Natural Language Processing



# About the Author

Created By: Mohammad Salman

Experience: 19 Years +

Designation: Corporate Trainer



# Natural Language Processing



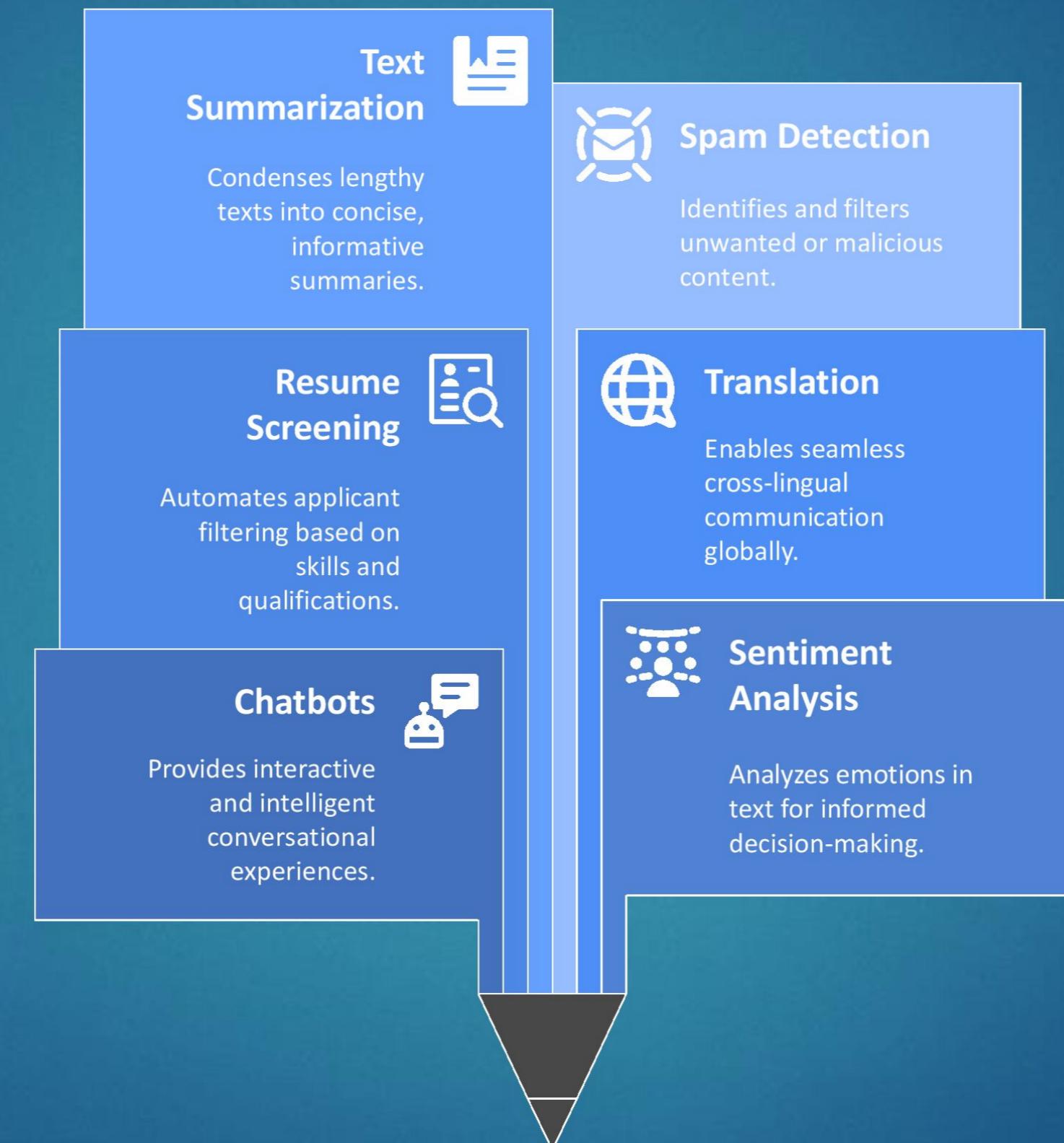
# What is Natural Language Processing

What is NLP?

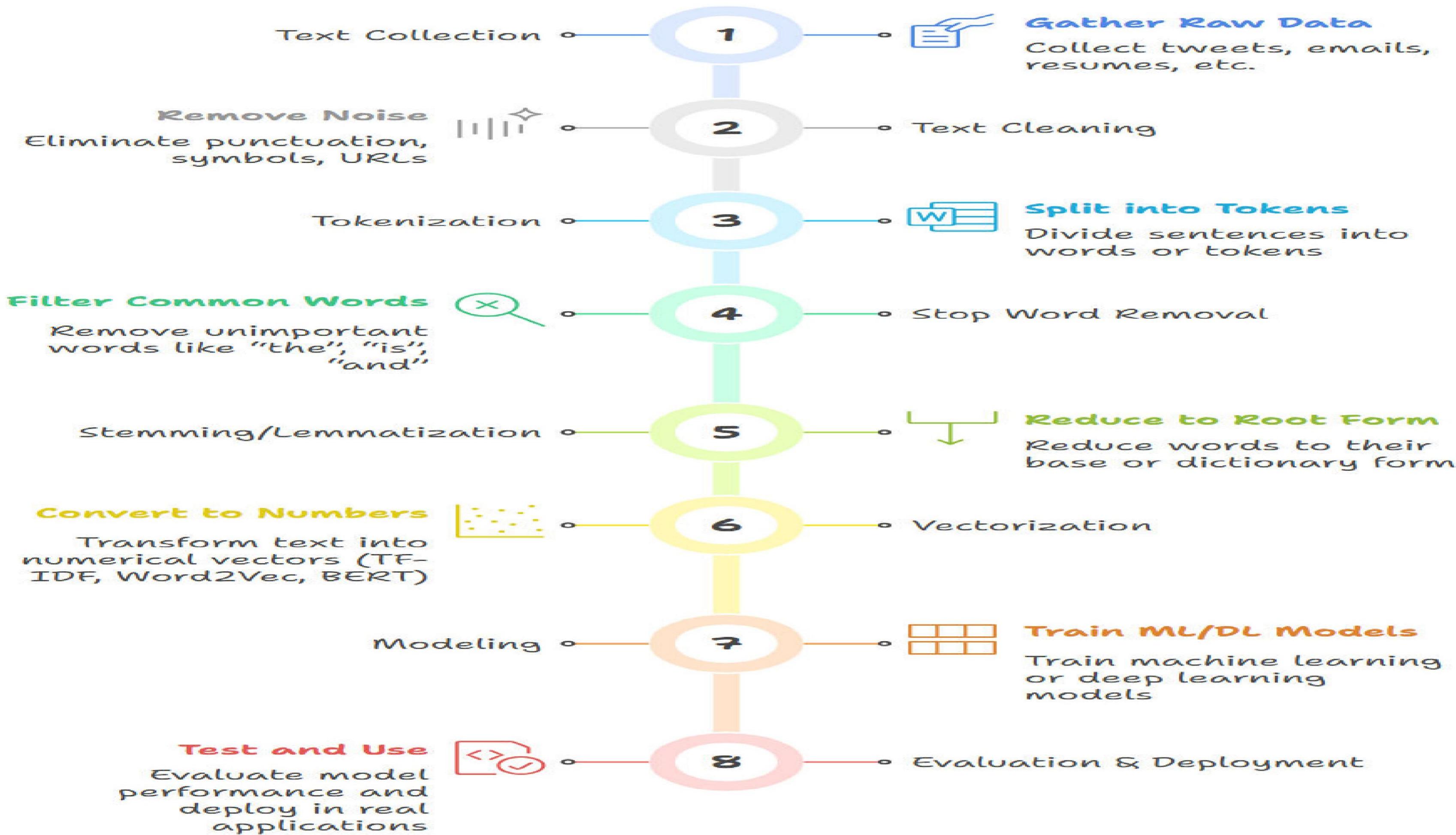
Natural Language Processing (NLP) is a branch of Artificial Intelligence (AI) that helps computers understand, interpret, and generate human language.



# NLP Applications



# NLP Network workflow



# Key NLP Preprocessing Terms

## 1. Tokenization

### **Definition:**

Tokenization is the process of splitting text into **smaller meaningful units** — usually words or sentences — called **tokens**.

### **Example:**

Sentence: "AI is changing the world."

Tokens: ["AI", "is", "changing", "the", "world"]

### **Why it's important:**

Models can't understand sentences as a whole; they need words or subwords to process patterns.

# Key NLP Preprocessing Terms

## 1. Tokenization

### **Definition:**

Tokenization is the process of splitting text into **smaller meaningful units** — usually words or sentences — called **tokens**.

### **Example:**

Sentence: "AI is changing the world."

Tokens: ["AI", "is", "changing", "the", "world"]

### **Why it's important:**

Models can't understand sentences as a whole; they need words or subwords to process patterns.

# Key NLP Preprocessing Terms

## 2. Stop Words

### **Definition:**

Stop words are **common words** in a language that add little meaning to text analysis.

### **Examples:**

“the”, “is”, “in”, “at”, “for”, “to”, “and”, “of”, “with”

### **Why remove them:**

They appear frequently but don't change the meaning of a sentence — removing them helps models focus on important words.

### **Example:**

Before: "The food is not good."

After removing stop words: "food not good"

# Key NLP Preprocessing Terms

## 3. Stemming

### **Definition:**

Stemming reduces words to their **root/base form** — by chopping off prefixes or suffixes.

The result may not be a real dictionary word.

### **Example:**

play, playing, played → play

run, running → run

# Key NLP Preprocessing Terms

## 3. Stemming

### **Definition:**

Stemming reduces words to their **root/base form** — by chopping off prefixes or suffixes.

The result may not be a real dictionary word.

### **Example:**

play, playing, played → play

run, running → run

# Key NLP Preprocessing Terms

## 4. Lemmatization

### **Definition:**

Lemmatization reduces words to their **root meaning (lemma)**, but ensures the result is a **real word**.

It uses **part-of-speech (POS)** tagging to understand context.

### **Example:**

am, are, is → be  
better → good

# Key NLP Preprocessing Terms

## 5. Vectorization

### Definition:

Converts words into **numerical form** so that ML models can process them.

#### Method

**Bag of Words (BoW)**

**TF-IDF**

**Word2Vec / GloVe**

**BERT**

#### Description

Counts word frequency

Assigns weight based on importance

Captures word meaning via neural embeddings

Deep contextual embeddings

#### Example

["AI", "world"] → [1,1]

rare words get higher weight

"king - man + woman ≈ queen"

Transformer-based

# Key NLP Preprocessing Terms

4

## Feature Extraction Techniques

Once text is clean, convert it into numerical features:

Technique	Description	Example Output
<b>Bag of Words (BoW)</b>	Word counts per document	[“data”: 2, “science’ 3]
<b>TF-IDF</b>	Weighs words by importance across docs	“data” gets lower weight if common everywhere
<b>Word Embeddings</b>	Captures meaning and context (Word2Vec, GloVe, FastText)	Similar words have nearby vectors

# Key NLP Preprocessing Terms

## 6. Part-of-Speech (POS) Tagging

### **Definition:**

Assigns each word its grammatical category (noun, verb, adjective, etc.). Helps in tasks like lemmatization and NER.

Example:

```
import nltk  
nltk.download('averaged_perceptron_tagger')
```

```
text = nltk.word_tokenize("The cat is sleeping")  
print(nltk.pos_tag(text))
```

# Key NLP Preprocessing Terms

## 6. Part-of-Speech (POS) Tagging

### **Definition:**

Assigns each word its grammatical category (noun, verb, adjective, etc.).  
Helps in tasks like lemmatization and NER.

Example:

```
import nltk  
nltk.download('averaged_perceptron_tagger')
```

```
text = nltk.word_tokenize("The cat is sleeping")  
print(nltk.pos_tag(text))
```

# Key NLP Preprocessing Terms

## 7. Named Entity Recognition (NER)

### **Definition:**

Identifies **names, dates, organizations, and locations** in text.

### **Example:**

“Apple was founded by Steve Jobs in California.”

→ Apple → ORG, Steve Jobs → PERSON, California → GPE

# Key NLP Preprocessing Terms

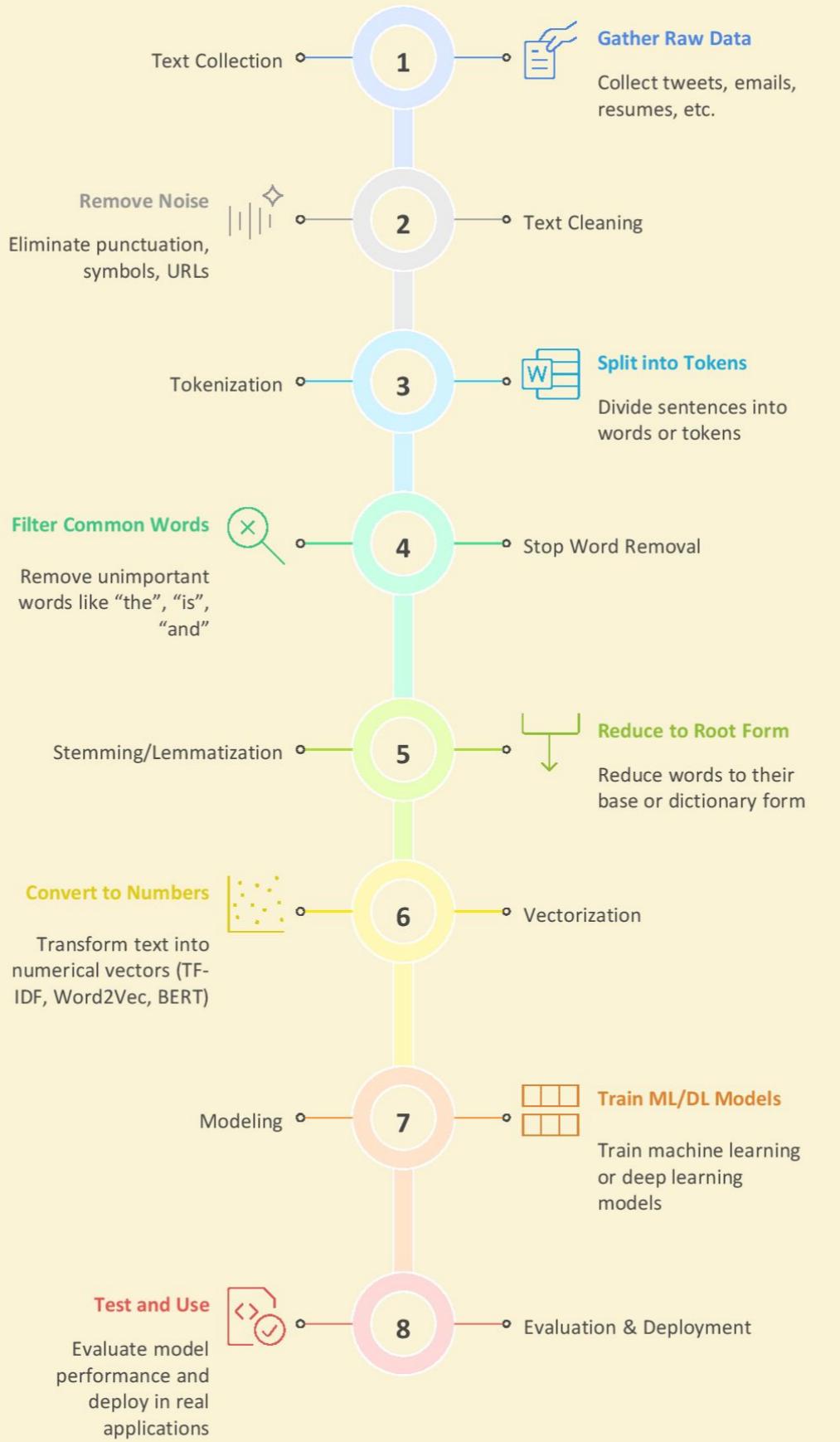
## 8. Sentence Segmentation

Splits a large paragraph into individual sentences.

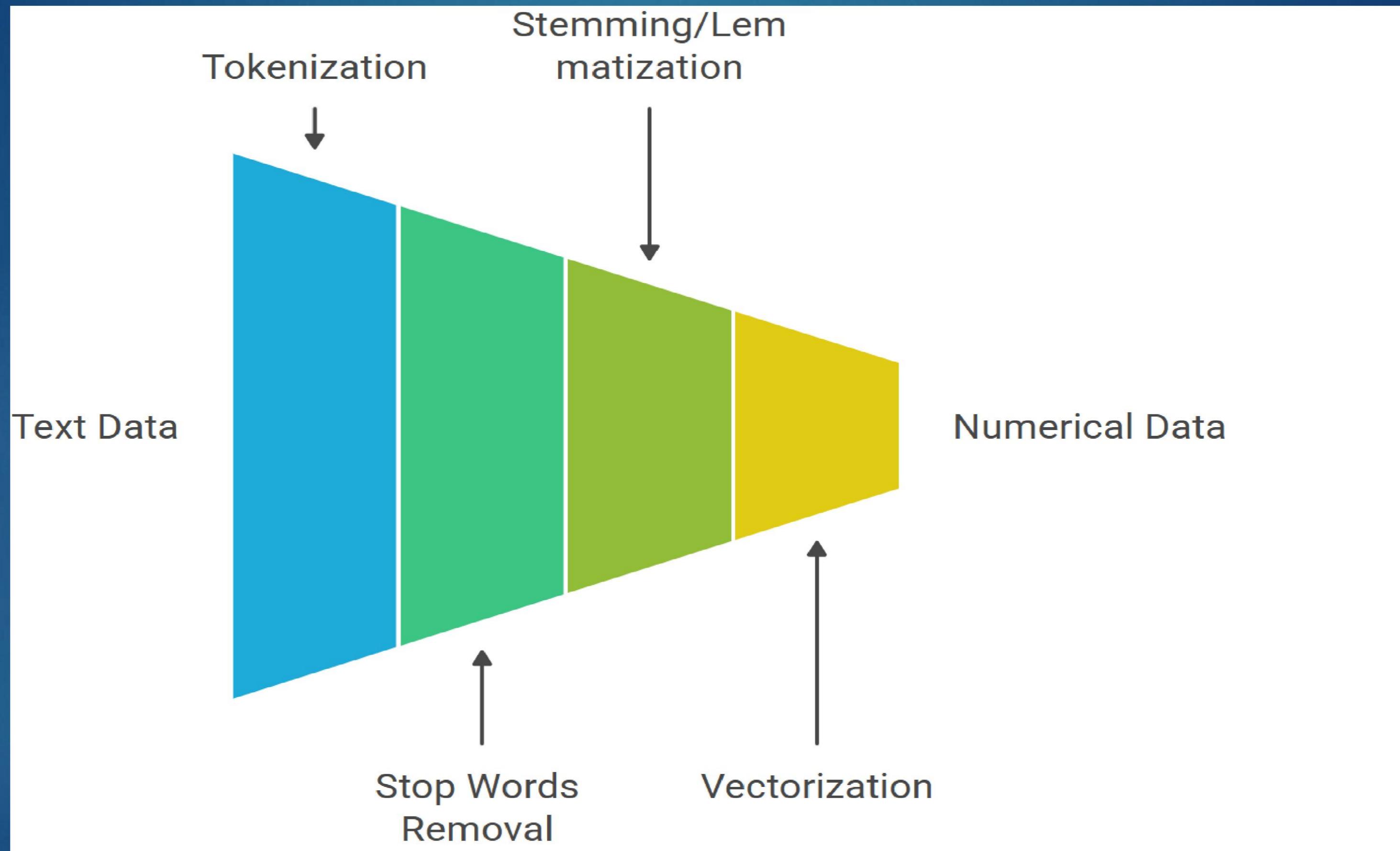
Example:

```
from nltk.tokenize import sent_tokenize  
text = "AI is powerful. It is changing the world."  
print(sent_tokenize(text))
```

# NLP Network workflow



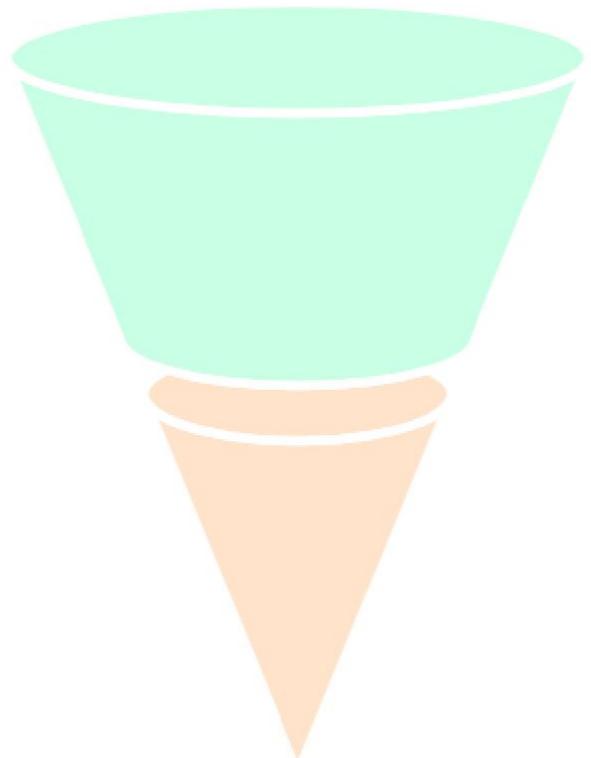
# Converting Text to Numerical Data



# Tokenization Process

## Identify Tokens

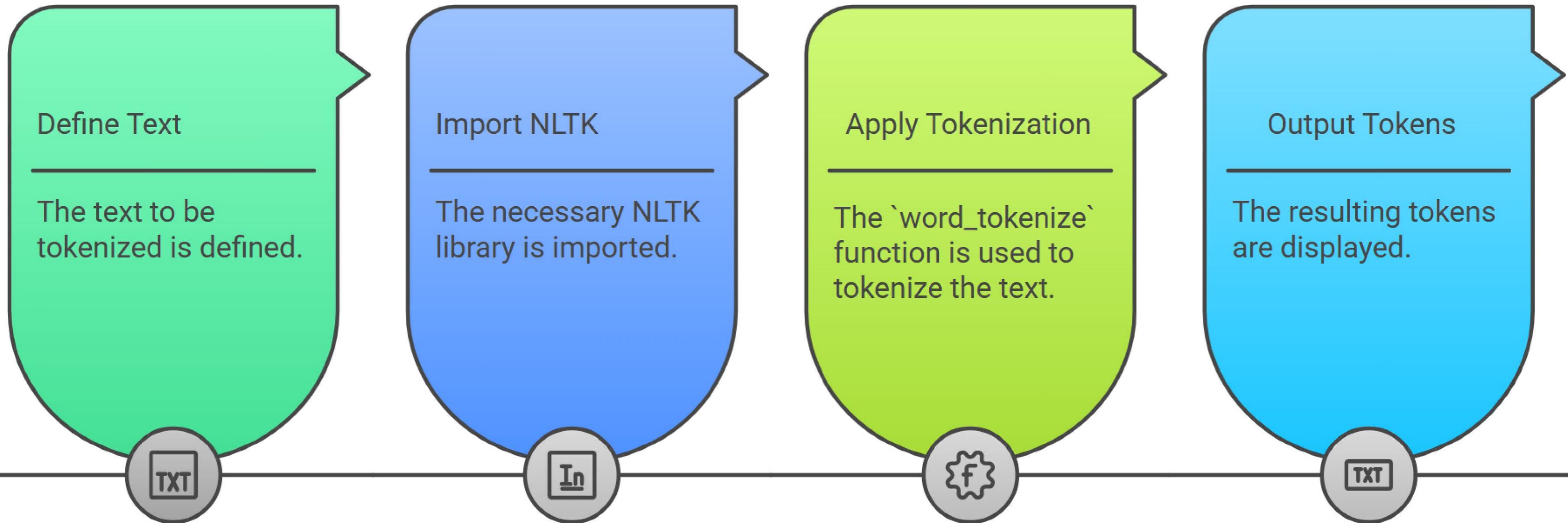
Meaningful units  
recognized as  
tokens



## Split into Units

Text divided into  
words or  
sentences

# Tokenization Process in NLP



# Stop Words Explained

What are stop words?

Common words that add little meaning to text analysis.

Why remove them?

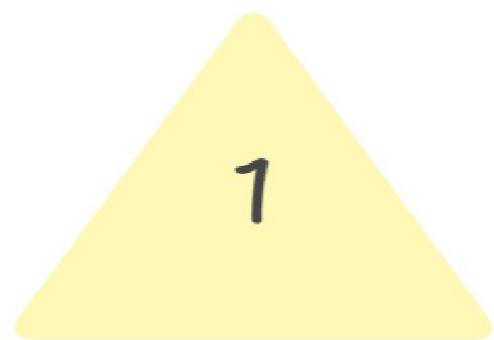
They appear frequently but don't change the meaning, helping models focus on important words.



# Stemming Process

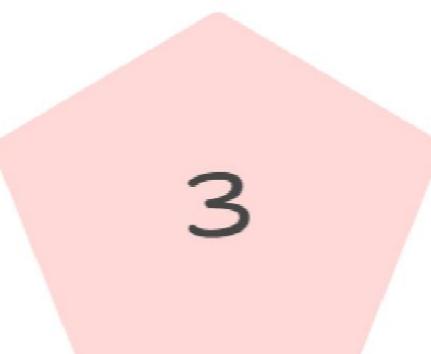
## Inflected Word

Word with  
prefixes/suffixes



## Suffix Removal

Chop off end of  
word



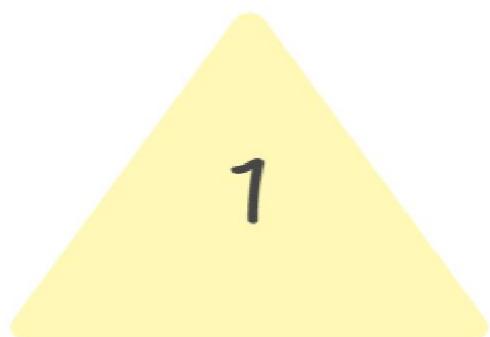
## Prefix Removal

Chop off  
beginning of  
word

# Stemming Process

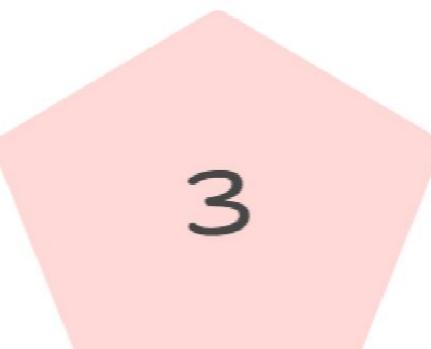
## Inflected Word

Word with  
prefixes/suffixes



## Suffix Removal

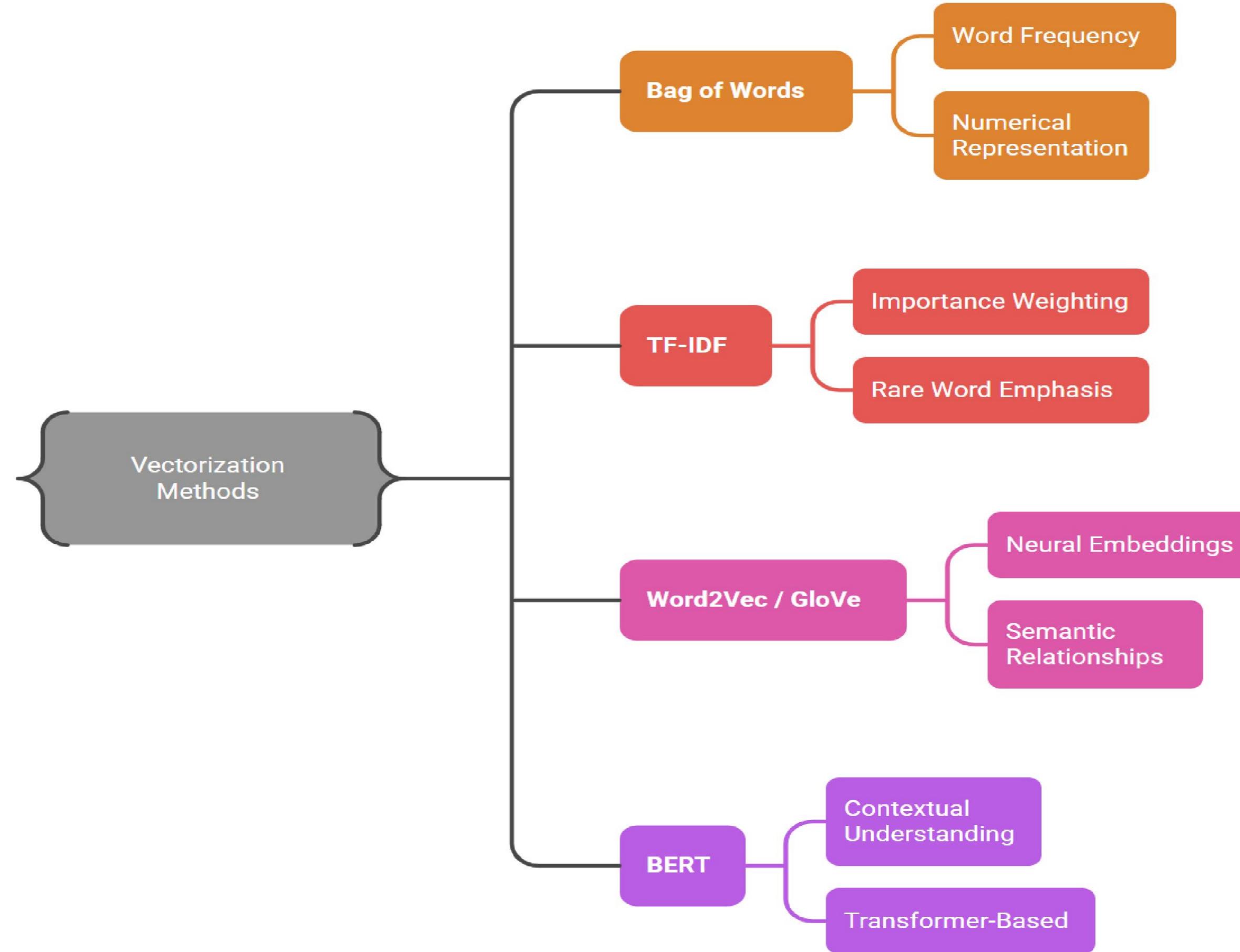
Chop off end of  
word



## Prefix Removal

Chop off  
beginning of  
word

# NLP Vectorization Methods



# Part-of-Speech (POS) Tagging

The illustration shows two stylized characters on a white background. On the left is a man with blue hair and a blue suit. On the right is a woman with blonde hair and a green dress. They are facing each other, separated by a large speech bubble. The man's speech bubble is blue and contains the question "What is POS tagging?". The woman's speech bubble is green and contains the answer "Assigning each word its grammatical category (noun, verb, adjective, etc.)". Below the woman is another blue speech bubble containing the question "Why is it important?", and a green speech bubble below it containing the answer "It helps in tasks like lemmatization and NER".

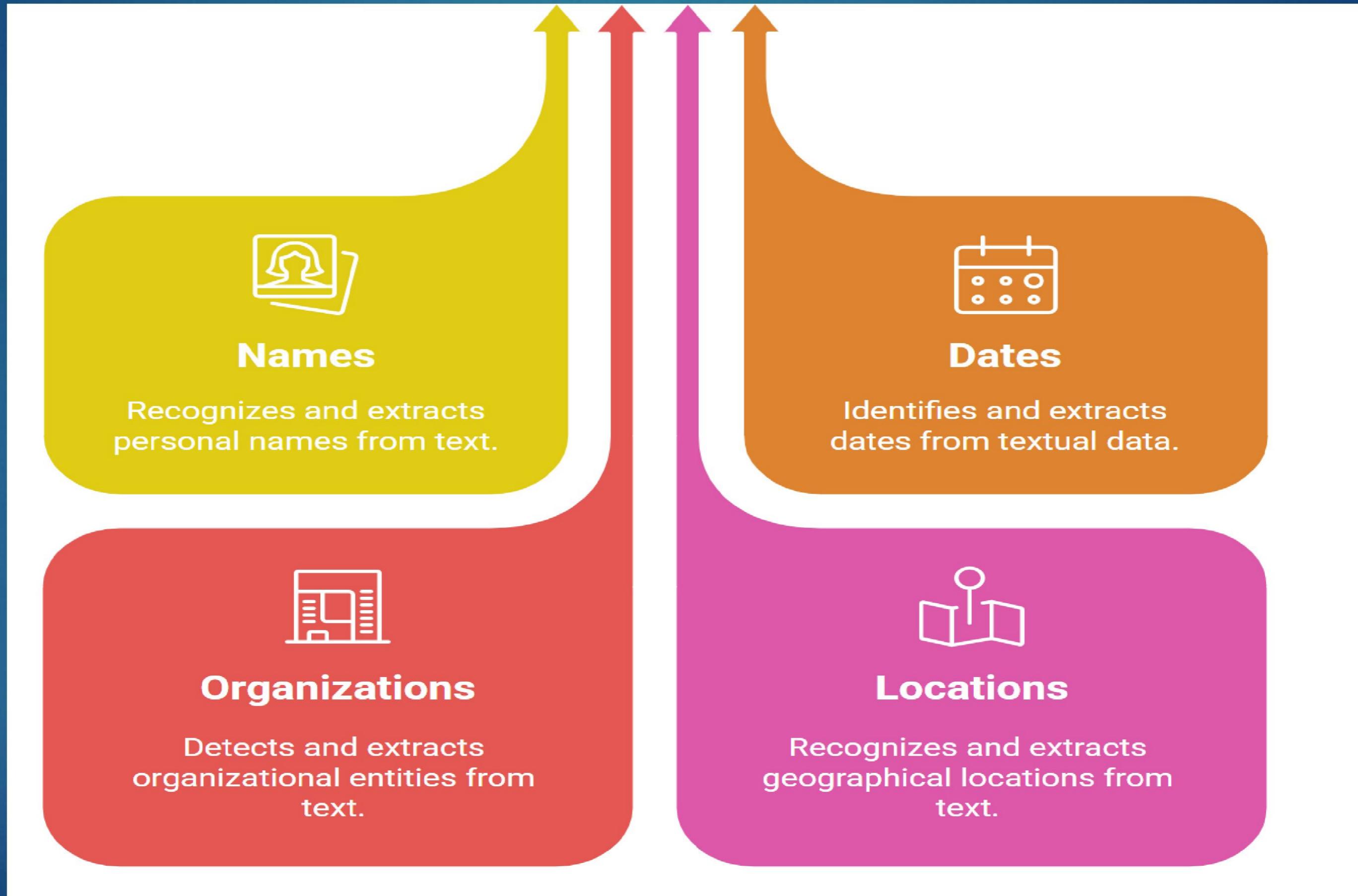
What is POS tagging?

Assigning each word its grammatical category (noun, verb, adjective, etc.).

Why is it important?

It helps in tasks like lemmatization and NER.

# Unveiling Textual Entities



# Enhancing Text Readability through Sentence Segmentation

