Natural Language Processing

About the Author

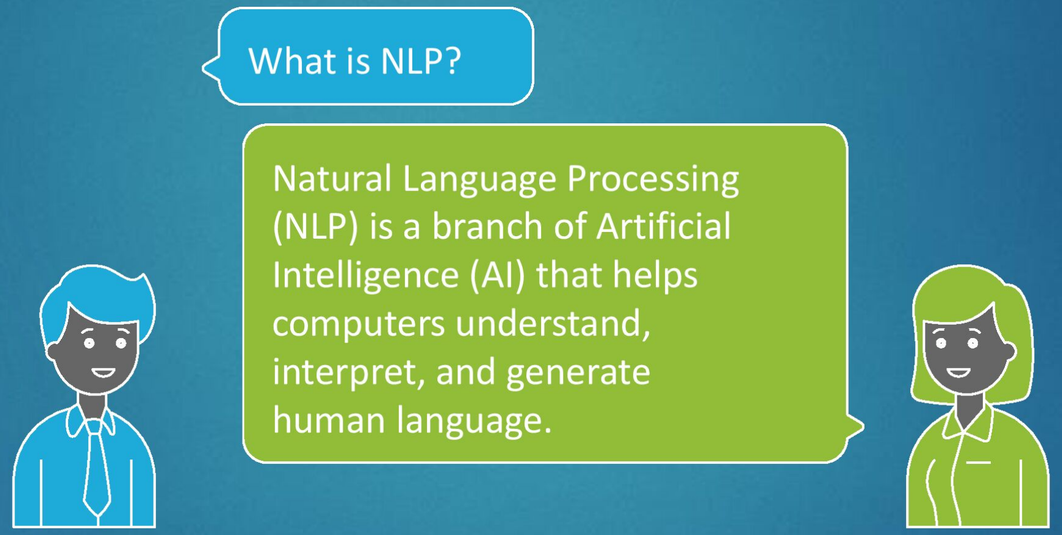
Created By: Mohammad Salman

Experience: 19 Years +

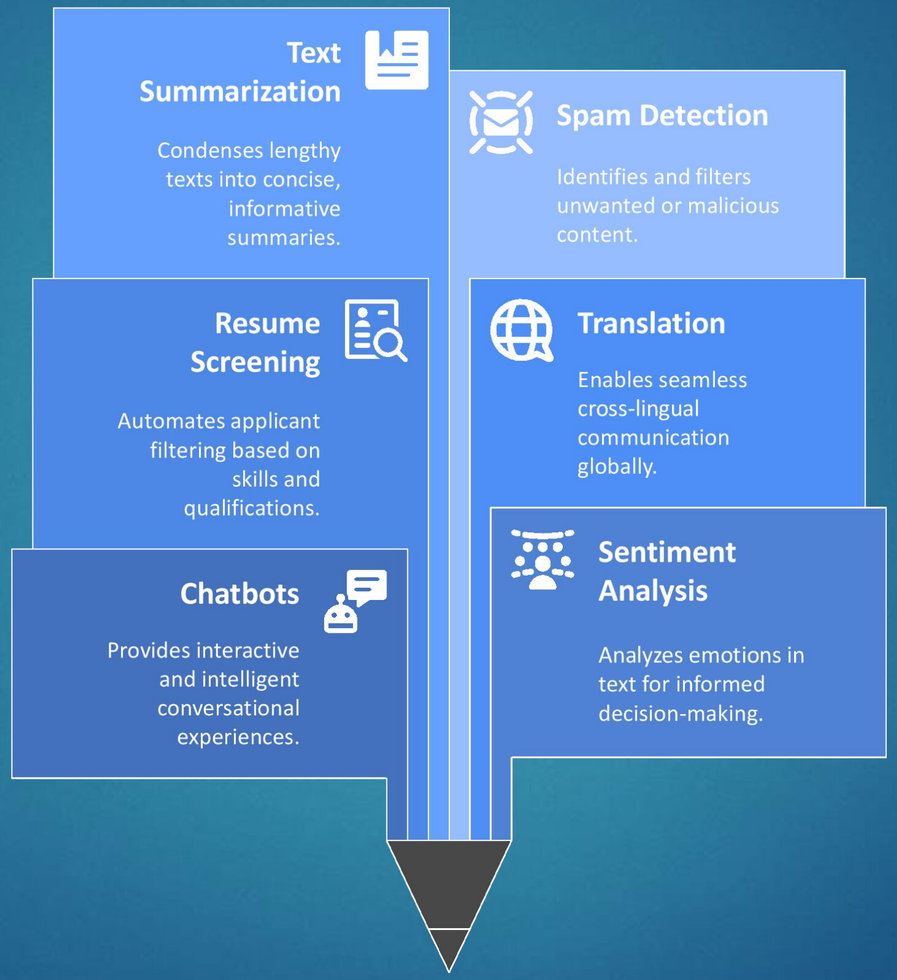
Designation: Corporate Trainer

Natural Language Processing

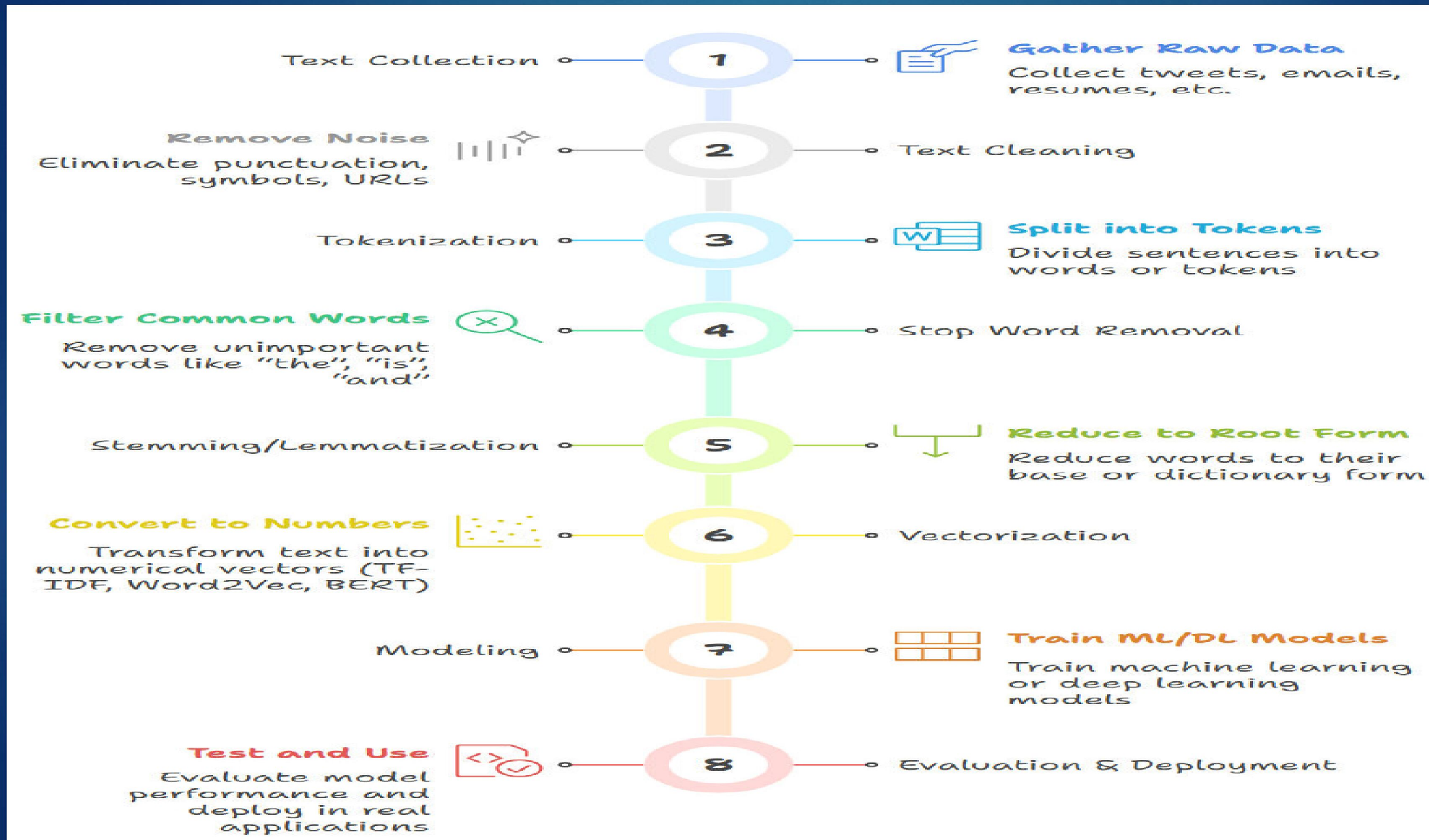
What is Natural Language Processing



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	Description	Example
Bag of Words (BoW)	Counts word frequency	["AI", "world"] → [1,1]
TF-IDF	Assigns weight based on importance	rare words get higher weight
Word2Vec / GloVe	Captures word meaning via neural embeddings	"king - man + woman ≈ queen"
BERT	Deep contextual embeddings	Transformer-based

Key NLP Preprocessing Terms

4 Feature Extraction Techniques

Once text is clean, convert it into numeriical features:

Technique	Description	Example Output
Bag of Words (BoW)	Word counts per document	[“data”: 2, “science’ 3]
TF-IDF	Weighs words by importance across docs	“data” gets lower weight if common everywhere
Word Embeddings	Captures meaning and context (Word2Vec, GloV e, 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

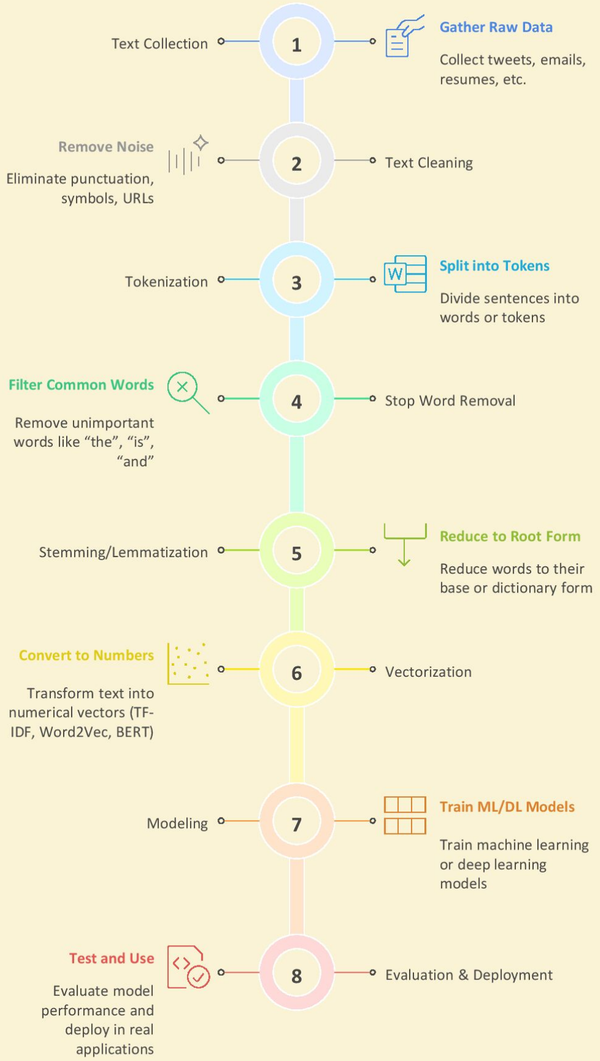
Key NLP Preprocessing Terms

8. Sentence Segmentation
Splits a large paragraph into individual sentences.

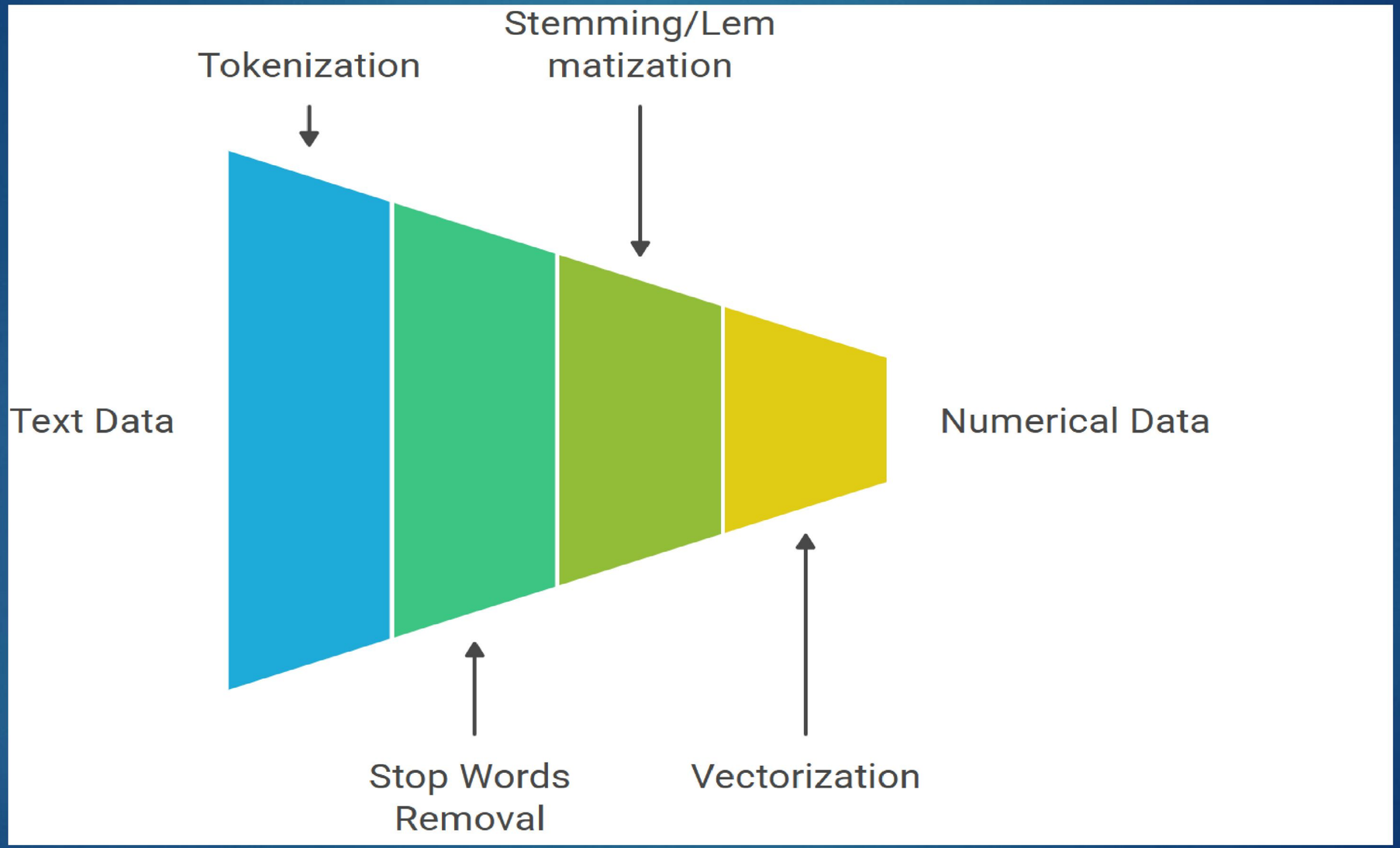
Example:

```
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

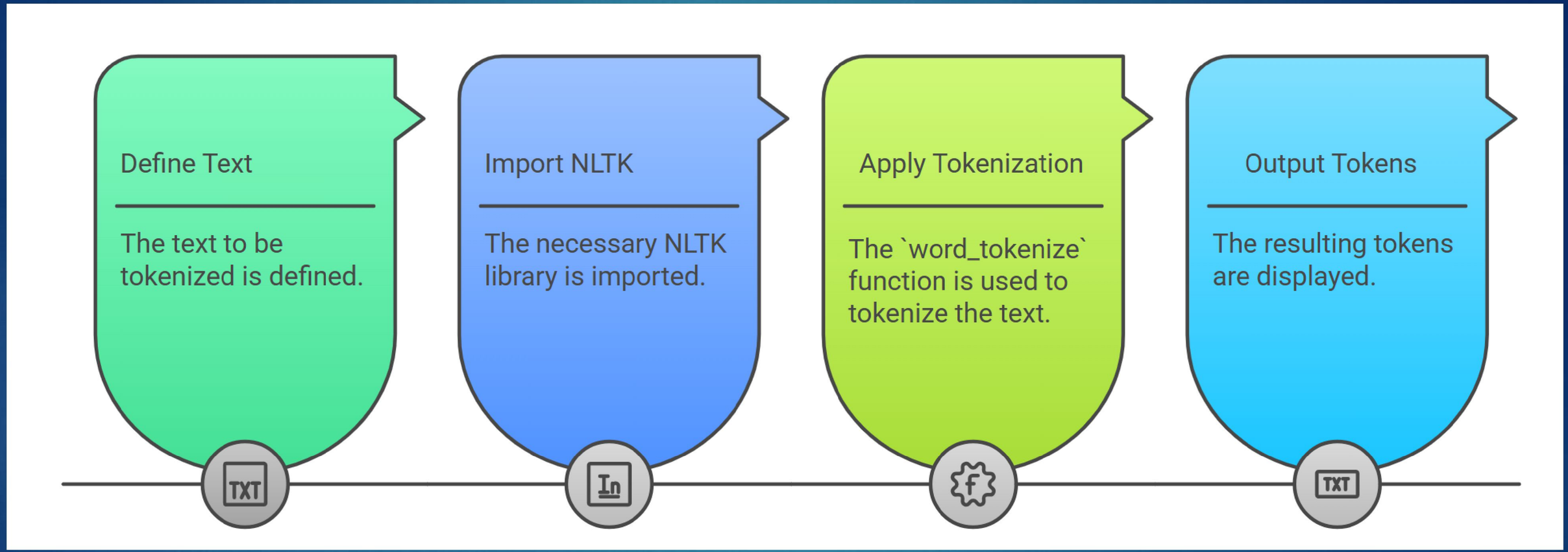
Split into Units

Text divided into words or sentences

Identify Tokens

Meaningful units recognized as tokens

Tokenization Process in NLP



Stop Words Explained

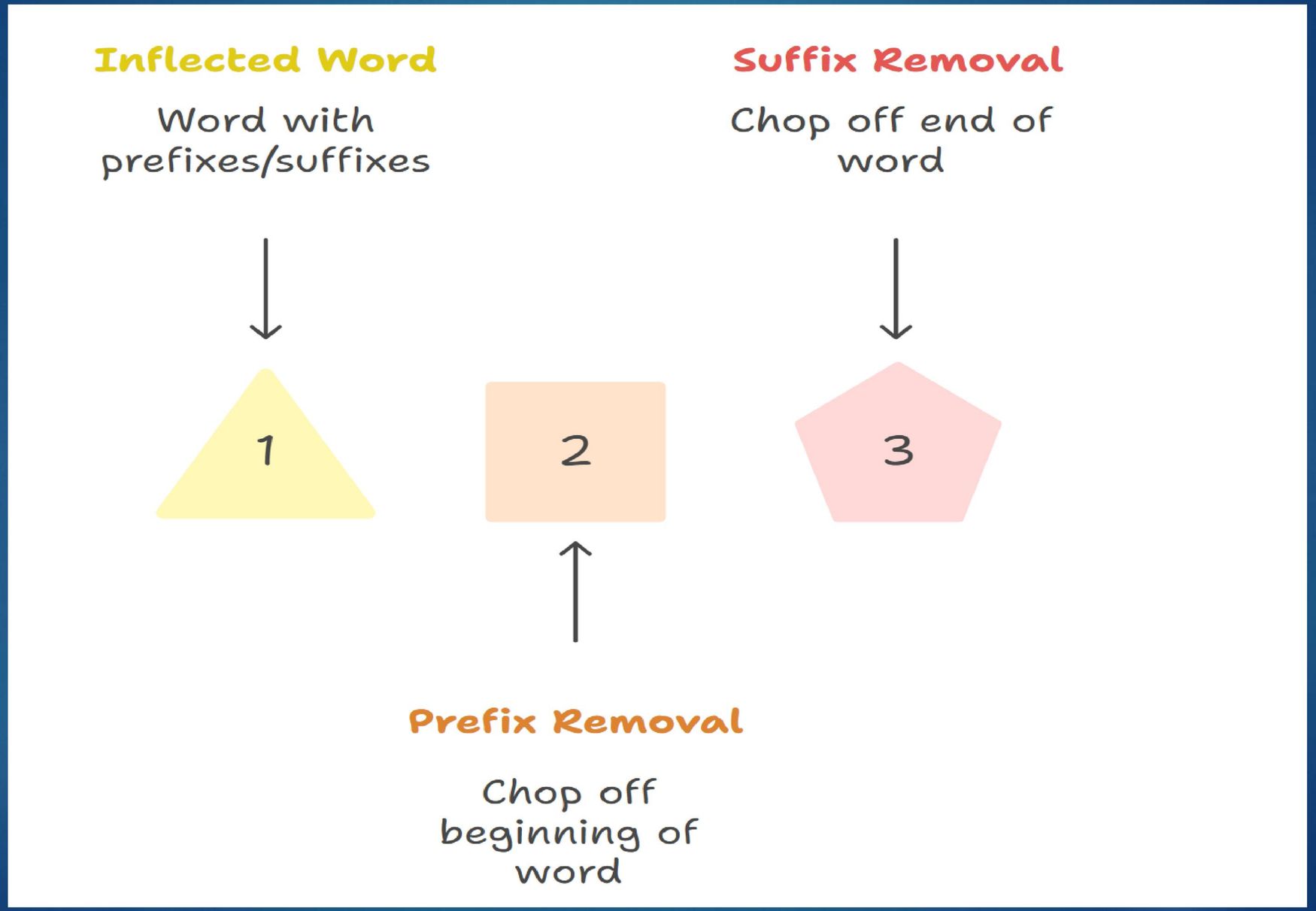
What are stop words?

Common words that add little meaning to text analysis.

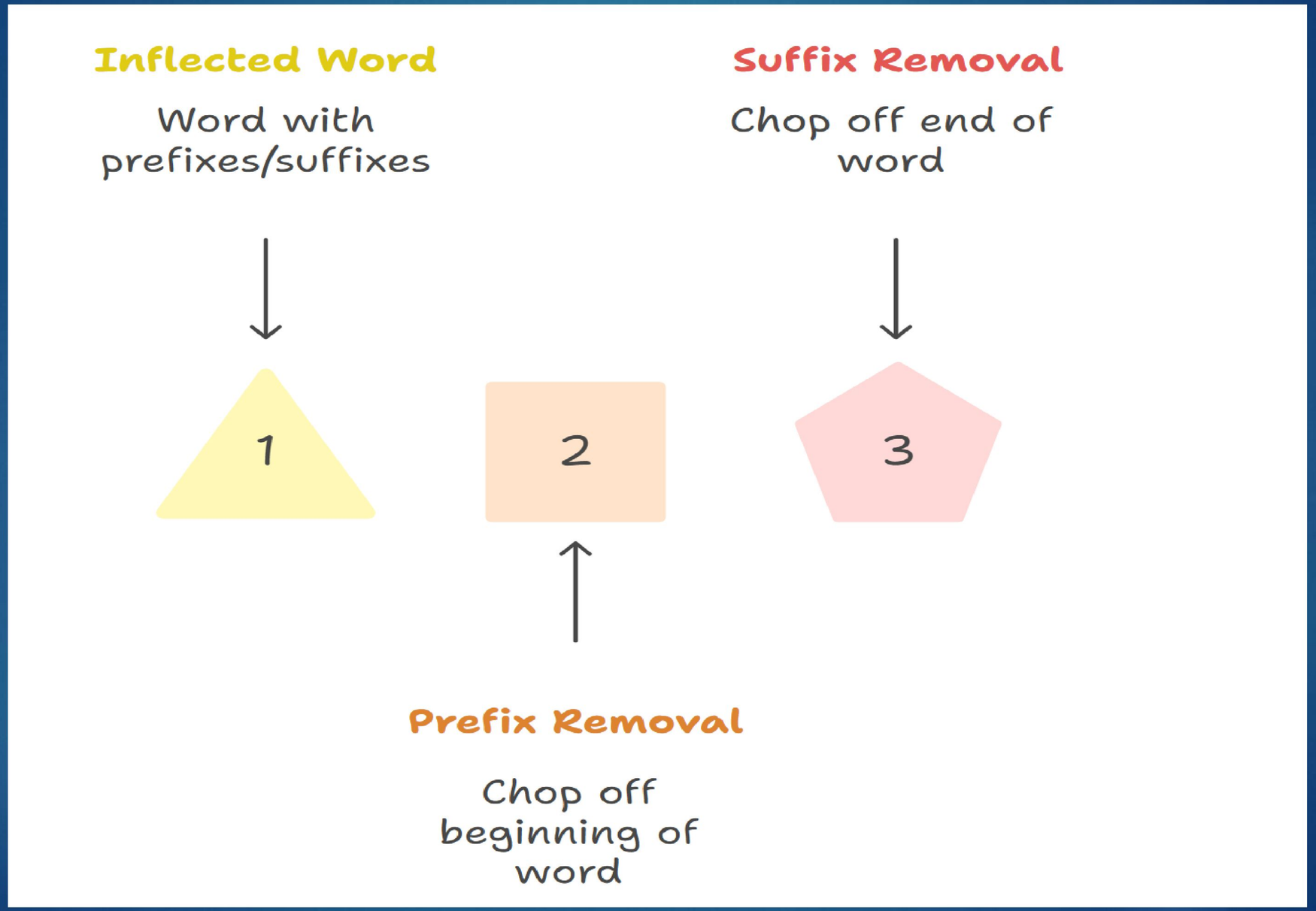
Why remove them?

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

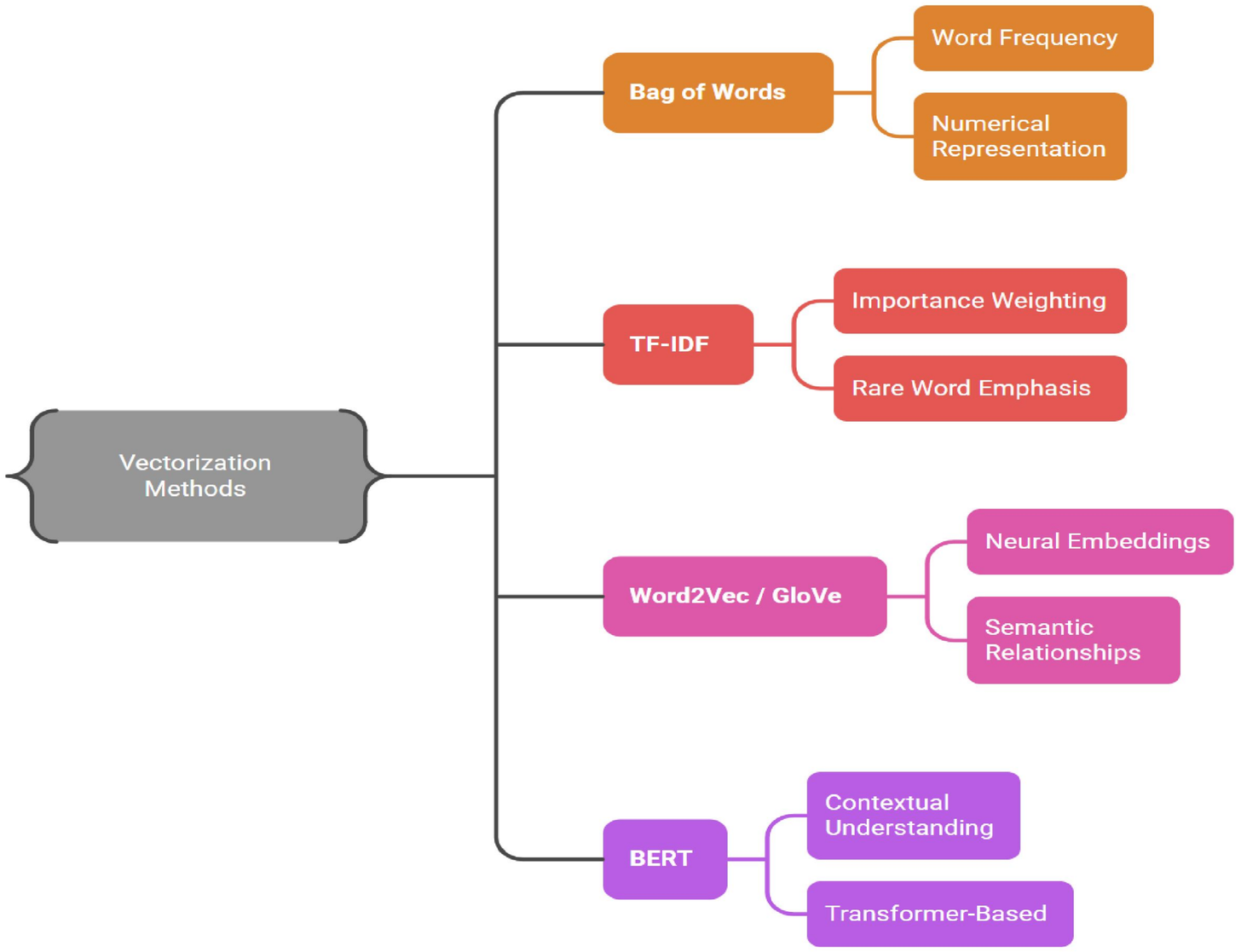
Stemming Process



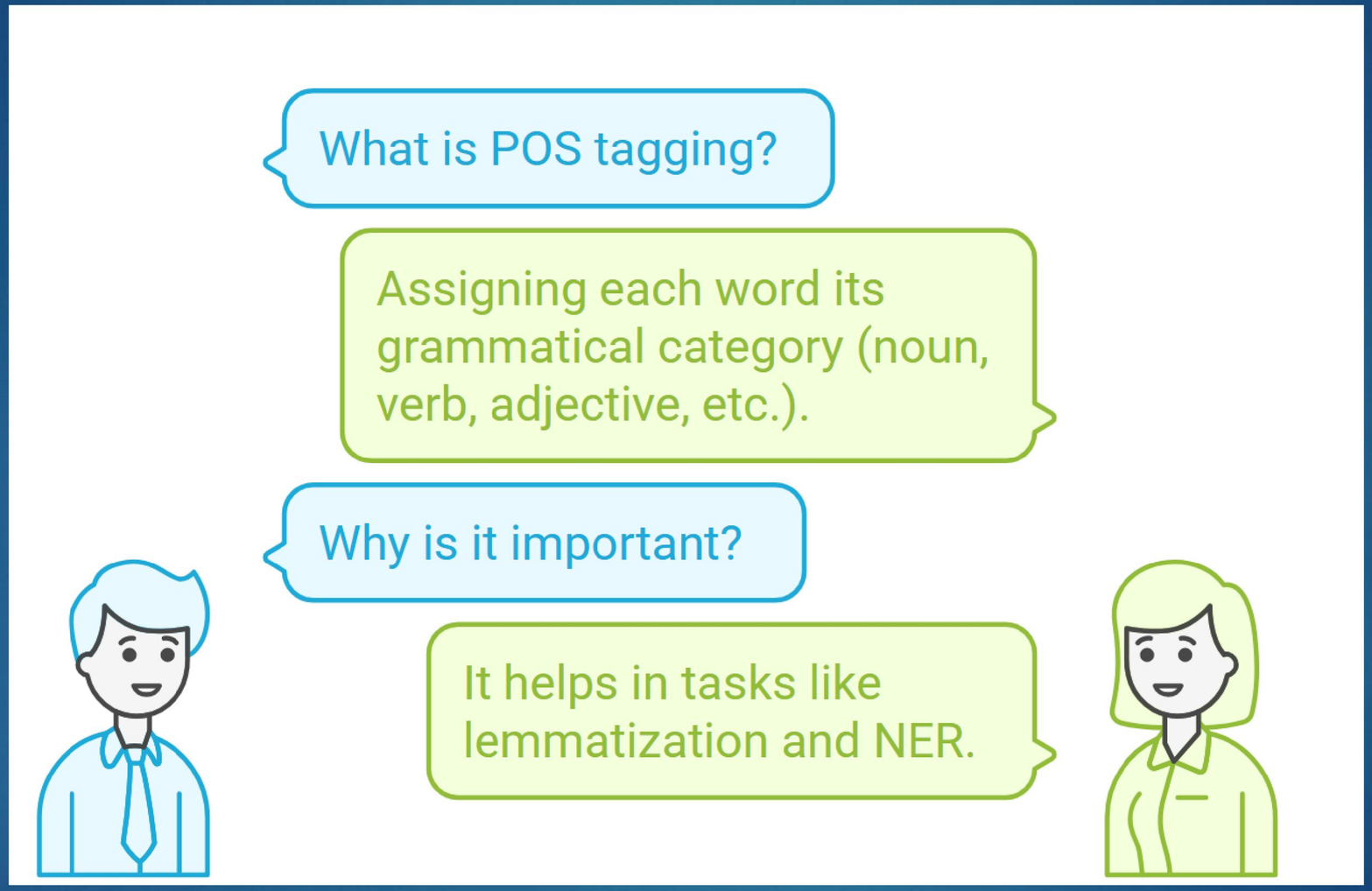
Stemming Process



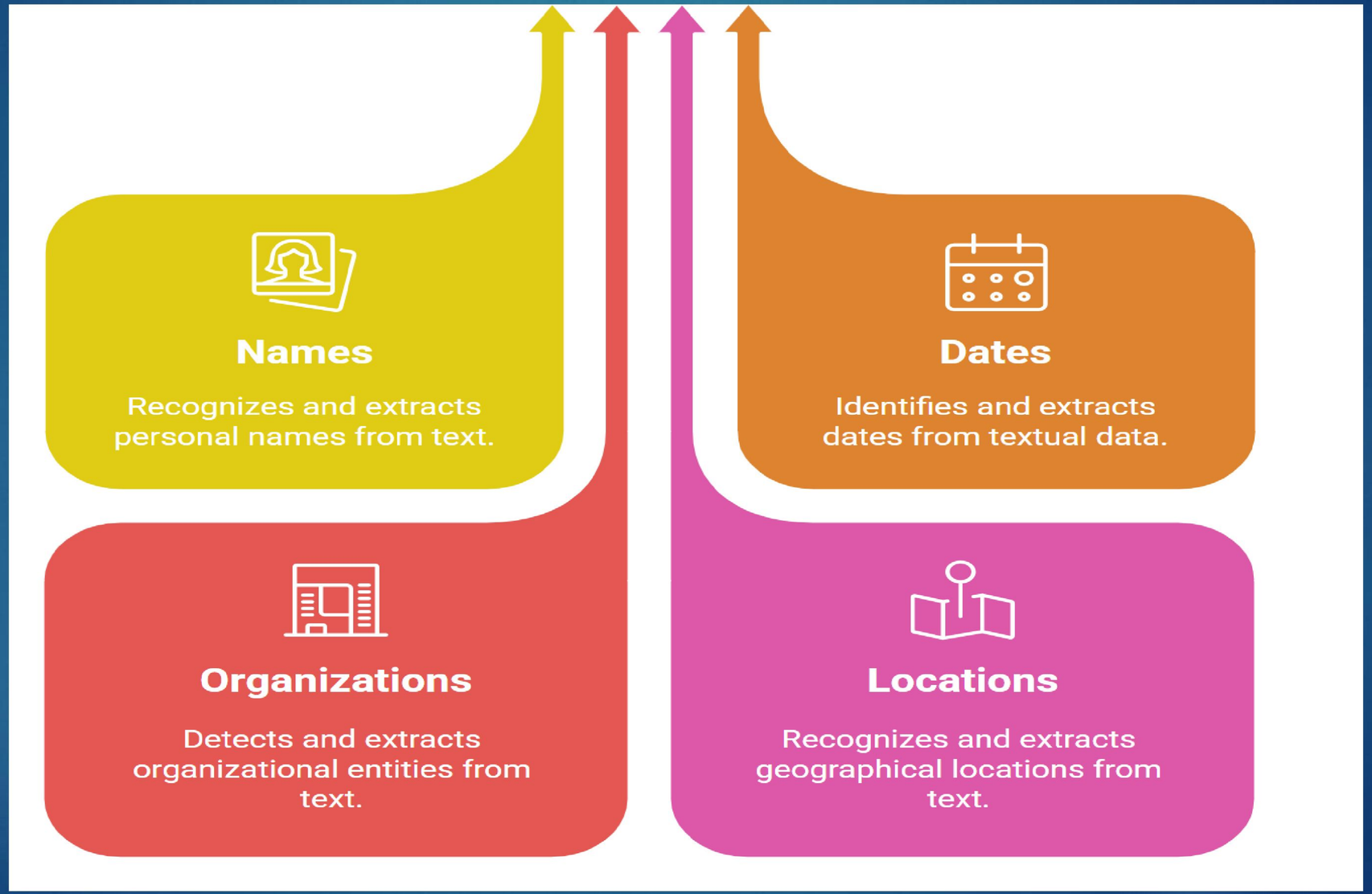
NLP Vectorization Methods



Part-of-Speech (POS) Tagging



Unveiling Textual Entities



Enhancing Text Readability through Sentence Segmentation