

The background of the slide is a dark gray to black gradient, overlaid with a complex, white, abstract network pattern. This pattern consists of numerous small, semi-transparent circular nodes of varying sizes, interconnected by thin, light gray lines. The nodes and lines are distributed across the entire frame, creating a sense of depth and connectivity, reminiscent of a neural network or a data visualization of relationships.

# Introduction to NLP

## Video 4: Methods of Text Preprocessing - Part 1

Instructor Video

**Text Classification Models**

**Text Generation Models**

## Text Classification Models

Instructor Video

Segregate Based on context

- Sentiment Analysis
- Spam Detection
- News Categorization, etc

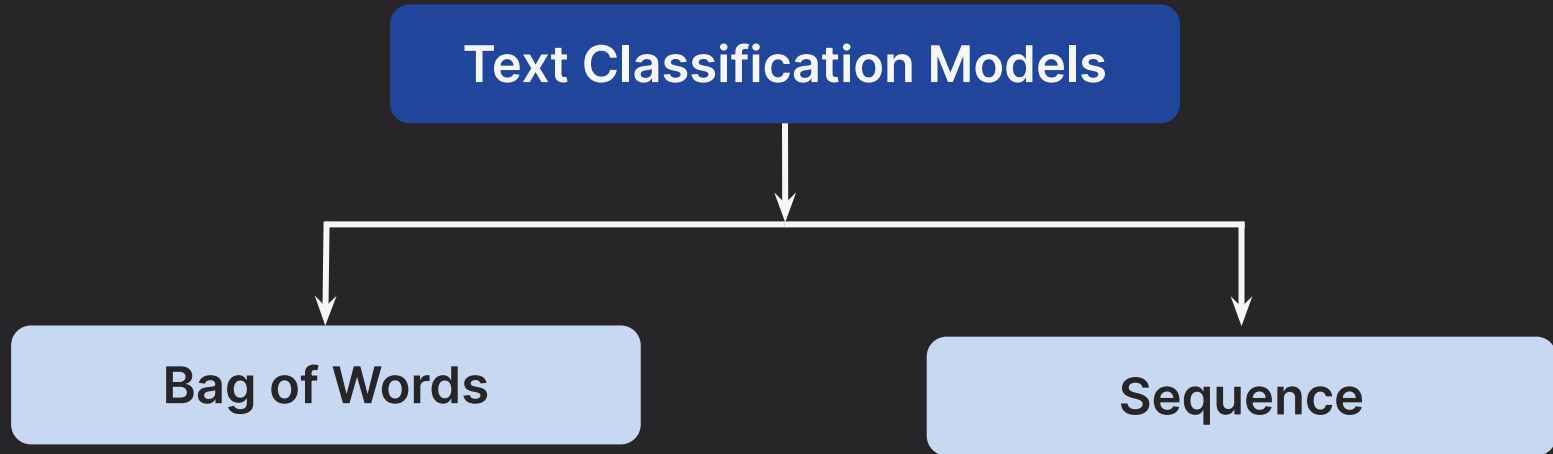
Instructor Video

## Text Generation Models

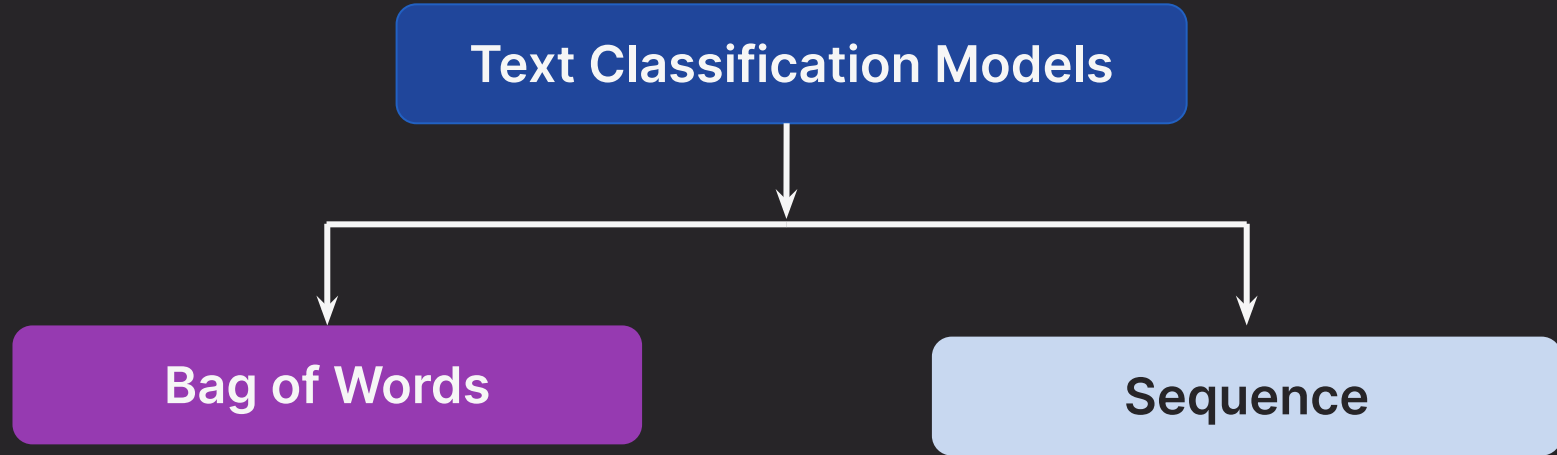
Generate text based on context

- Text summarization
- Autocomplete suggestions
- Question Answering, etc

# NLP Model



# NLP Model



# Bag of Words

The squids jumped out of the suitcases.



Sentence or document is represented as the **bag of its wordset**.

Bag of words **disregards grammar and sequence**

# Bag of Words

# Scrabble



# Bag of Words





# Tokenization

The process of dismantling the sentences, paragraphs and articles into smaller chunks is called tokenization .

# Tokenization

NLP bridges the gap between human language and machines. It allow computers to understand the meaning behind words and generate human-like text. This powerful technology has a wide range of applications. NLP is constantly evolving, playing an increasingly important role in our interactions.

## Sentence Tokenization

NLP bridges the gap between human language and machines.

It allow computers to understand the meaning behind words and generate human-like text.

This powerful technology has a wide range of applications.

NLP is constantly evolving, playing an increasingly important role in our interactions.

# Tokenization

Hello world !

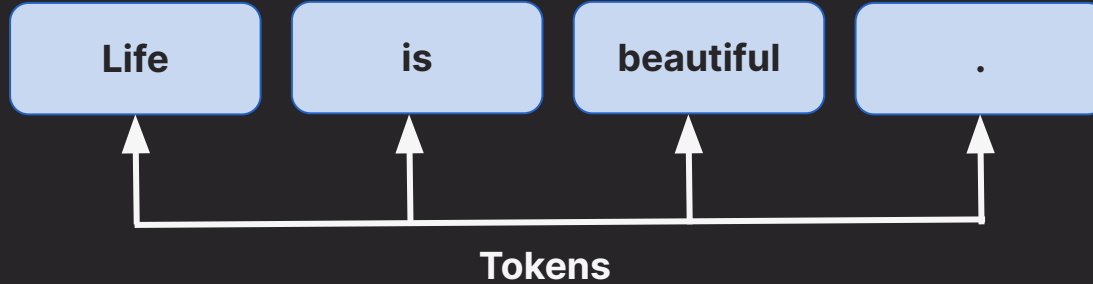
## Character Tokenization

H e l l o w o r l d !

# Tokenization

Life is beautiful.

## Word Tokenization



# Preprocessing Techniques



# Preprocessing Techniques



Lowercasing

# Preprocessing Techniques



Lowercasing



Removing Punctuation and Special Characters

# Preprocessing Techniques



Lowercasing



Removing Punctuation and Special Characters



Stop Words Removal



# Preprocessing Techniques



Lowercasing



Removing Punctuation and Special Characters



Stop Words Removal



Stemming and Lemmatization

# Preprocessing Techniques



Lowercasing



Removing Punctuation and Special Characters



Stop Words Removal



Stemming and Lemmatization



Vocabulary Building

# Preprocessing Techniques



Lowercasing



Removing Punctuation and Special Characters



Stop Words Removal



Stemming and Lemmatization



Vocabulary Building



Vectorization