Introduction to NLP

Video 4: Methods of Text Preprocessing - Part 1

Analytics Vidhya

Analytics Vidhya

Instructor Video

Text Classification Models

Text Generation Models

Analytics Vidhya

Instructor Video

Text Classification Models

Segregate Based on context

- Sentiment Analysis
- Spam Detection
- News Categorization, etc

Analytics Vidhya

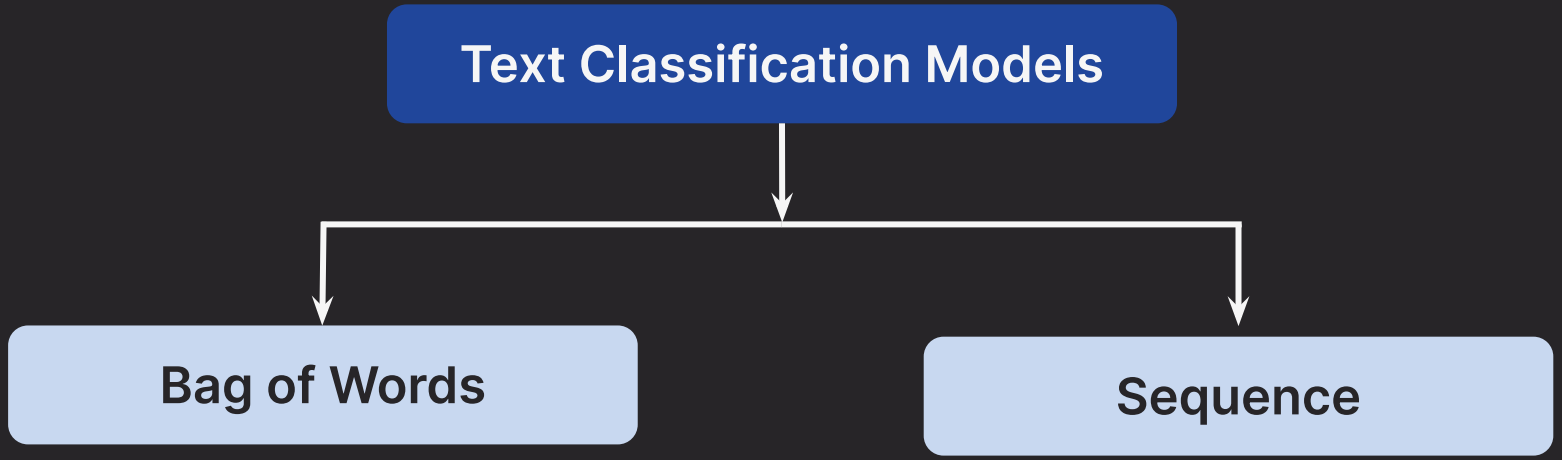
Instructor Video

Text Generation Models

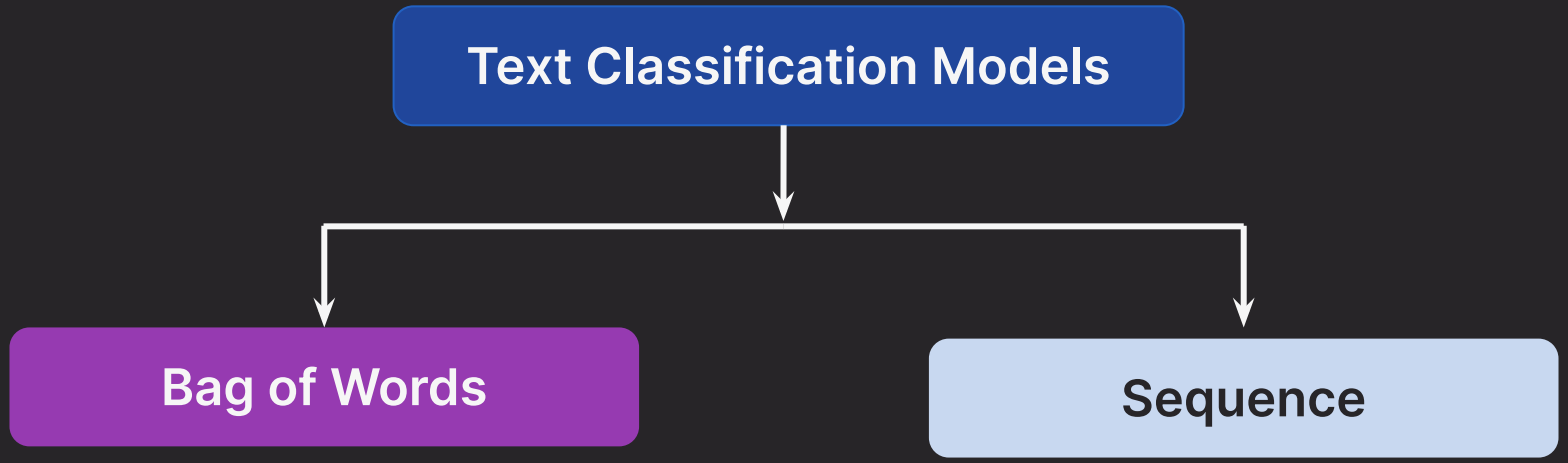
Generate text based on context

- Text summarization
- Autocomplete suggestions
- Question Answering, etc

NLP Model



NLP Model



Bag of Words

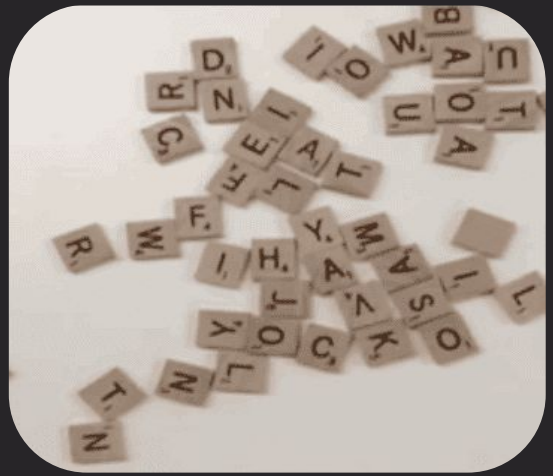
The squids jumped out of the suitcases.

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 .

Analytics Vidhya

Tokenization

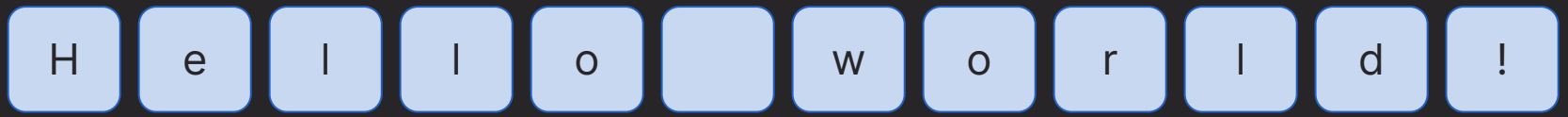
NLP bridges the gap between human language and machines. It allow computers to understand the meaning behind words and generate human-like text. This powerful technology has a wide range of applications. NLP is constantly evolving, playing an increasingly important role in our interactions.

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 !

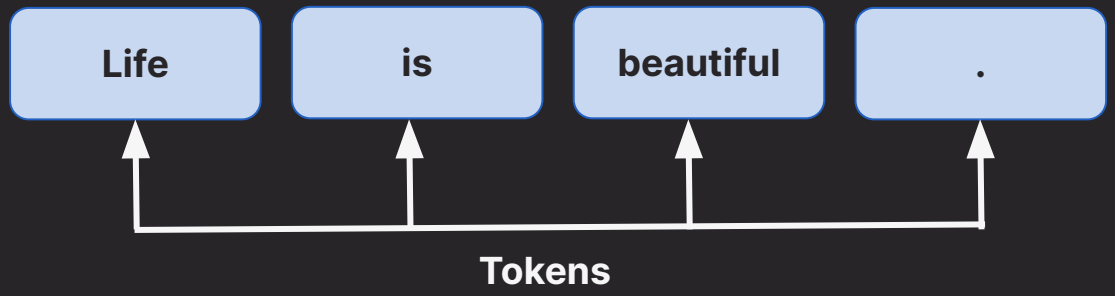
Character Tokenization



Tokenization

Life is beautiful.

Word Tokenization



Analytics Vidhya

Preprocessing Techniques

Analytics Vidhya

Preprocessing Techniques

- ✓ Lowercasing

Analytics Vidhya

Preprocessing Techniques

- ✓ Lowercasing
- ✓ Removing Punctuation and Special Characters

Analytics Vidhya

Preprocessing Techniques

- ✓ Lowercasing
- ✓ Removing Punctuation and Special Characters
- ✓ Stop Words Removal

Analytics Vidhya

Preprocessing Techniques

- ✓ Lowercasing
- ✓ Removing Punctuation and Special Characters
- ✓ Stop Words Removal
- ✓ Stemming and Lemmatization

Analytics Vidhya

Preprocessing Techniques

- ✓ Lowercasing
- ✓ Removing Punctuation and Special Characters
- ✓ Stop Words Removal
- ✓ Stemming and Lemmatization
- ✓ Vocabulary Building

Analytics Vidhya

Preprocessing Techniques

- ✓ Lowercasing
- ✓ Removing Punctuation and Special Characters
- ✓ Stop Words Removal
- ✓ Stemming and Lemmatization
- ✓ Vocabulary Building
- ✓ Vectorization