

# Conceitos

## Conceitos Fundamentais

Nesta seção, o objetivo é compreender o que são modelos de linguagem e como eles representam o texto de forma numérica e probabilística.

## Modelos de Linguagem

- **Definição:** Um modelo de linguagem é uma função que atribui uma probabilidade a uma sequência de palavras.
- **Exemplo:**

$$(P(w_1, w_2, \dots, w_n) = P(w_1) \times P(w_2|w_1) \times \dots \times P(w_n|w_1, \dots, w_{n-1}))$$

- $P$  → probabilidade de ocorrência de uma palavra ou sequência de palavras.
- $w_i$  → palavra na posição  $i$  da sequência (exemplo:  $w_1$  é a primeira palavra,  $w_2$  a segunda, etc.).
- $P(w_n|w_1, \dots, w_{n-1})$  → probabilidade da  $n$ -ésima palavra dado o contexto anterior.

Essa decomposição usa a **regra da cadeia de probabilidade**, permitindo calcular a chance de uma frase ocorrer com base nas dependências entre palavras.

- **Tipos principais:**
  - **Modelos N-grama:** baseiam-se em janelas de contexto fixas (ex: bigramas, trigramas).
  - **Modelos probabilísticos:** usam estimativas de frequência para prever a próxima palavra.
  - **Modelos neurais:** usam embeddings e redes neurais (ex: RNNs, Transformers).

## Modelos Estatísticos vs. Modelos Baseados em Regras


Tipo	Descrição	Exemplo
Baseados em regras(HMM)	Usam gramáticas e dicionários definidos manualmente.	Análise sintática, POS tagging com expressões regulares

Tipo	Descrição	Exemplo
Estatísticos(Naive Bayes)	Aprendem padrões a partir de dados.	Modelos de linguagem, classificação de texto

Hoje, os modelos neurais superam os baseados em regras em tarefas complexas, mas os dois podem se complementar.

## Dicionários e Vocabulário

- **Vocabulário:** conjunto de todas as palavras conhecidas pelo modelo.
- **Out-of-Vocabulary (OOV):** palavras não vistas durante o treinamento.
- **Técnicas de redução do vocabulário:**
  - Remoção de stopwords
  - Subword tokenization (WordPiece, BPE)

 *Dica prática:* use o `nltk.FreqDist` ou `Counter` do Python para analisar as palavras mais comuns de um corpus.