



2ª série

Inteligência Artificial

Processamento de Linguagem Natural (PLN)

Rildo Oliveira



02/05/2024

OBJETO DO CONHECIMENTO: Processamento de Linguagem Natural (PLN)

HABILIDADE:

PCRPO3 - Identificar, entender e explicar em que situações o computador pode ou não ser utilizado para solucionar um problema.

EF05HI06 - Comparar o uso de diferentes linguagens e tecnologias no processo de comunicação e avaliar os significados sociais, políticos e culturais atribuídos a elas.

OBJETIVOS:

- Entender os fundamentos do processamento de linguagem natural;
- Explorar técnicas comuns de pré-processamento de texto e sua influência na qualidade das análises de PLN.

DA TEORIA À PRÁTICA: Uso de imagens, texto e conceitos para um melhor entendimento do tema abordado.

Introdução ao PLN

O Processamento de Linguagem Natural (PLN) é um campo da Inteligência Artificial (IA) que se concentra na interação entre computadores e linguagem humana.

E nessa loucura de dizer que não te quero, vou negando as aparências, disfarçando as evidências. Mas pra que viver fingindo, se eu não posso enganar meu coração? Eu sei que te amo.

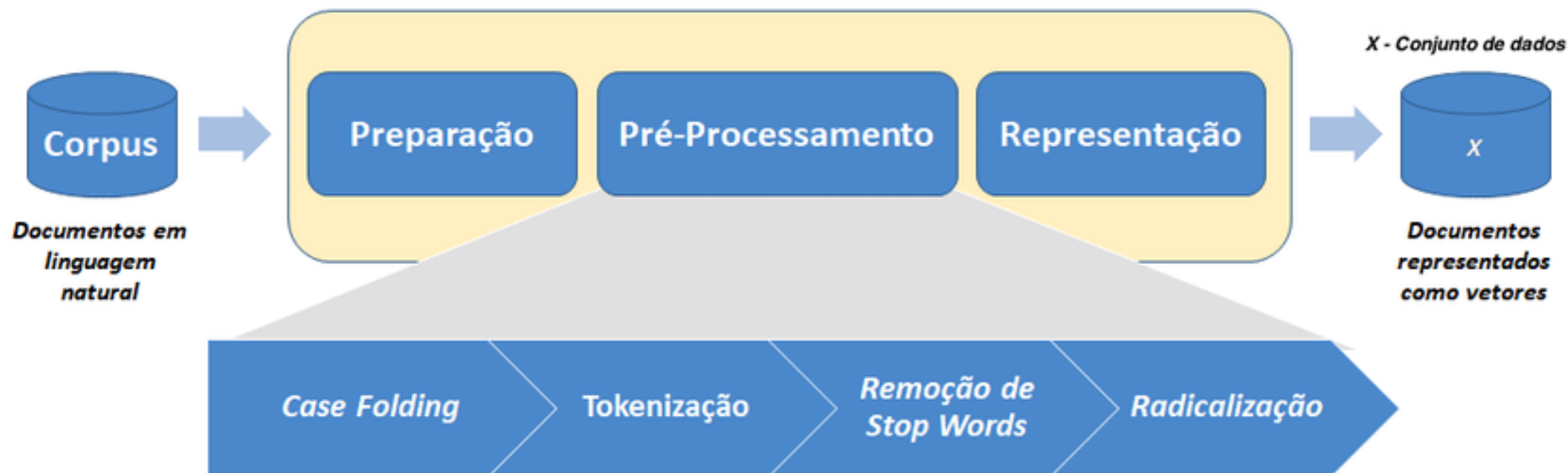
Definição de PLN

O PLN refere-se à capacidade dos computadores de entender, interpretar e gerar texto em linguagem natural. Ele permite que os sistemas computacionais compreendam comandos, respondam a perguntas e até mesmo gerem texto de maneira semelhante à humana. A importância do PLN na área de IA reside na sua capacidade de permitir a comunicação eficaz entre humanos e máquinas, abrindo portas para uma série de aplicações práticas.



Pré-processamento de Texto

Antes que um computador possa entender o texto em linguagem natural, é necessário realizar uma série de etapas de pré-processamento para limpar e preparar os dados de entrada.



Tokenização

A tokenização é o processo de dividir um texto em unidades significativas, como palavras, frases ou símbolos. Isso facilita a manipulação e a análise do texto pelos computadores.

M E U N O M E É H U G O C A L I X T O

" M E U "

" N O M E "

" É "

" H U G O "

" C A L I X T O "

**CADA PALAVRA
É UM TOKEN**

Limpeza de texto

Durante o pré-processamento, é comum remover pontuações e stopwords (palavras comuns que geralmente não contribuem para o significado do texto) para simplificar a análise posterior. Além disso, outras tarefas de limpeza, como correção ortográfica e normalização de texto, podem ser realizadas.

Limpeza de texto

Frase original:

O processamento de linguagem natural é uma área fascinante da inteligência artificial.

Frase após tokenização e remoção de stopwords:

"processamento linguagem natural área fascinante inteligência artificial."

Neste exemplo, as palavras "O", "de", "é", "uma", "da", foram removidas por serem stopwords, e a frase foi reduzida às palavras restantes após o processo de tokenização e limpeza.

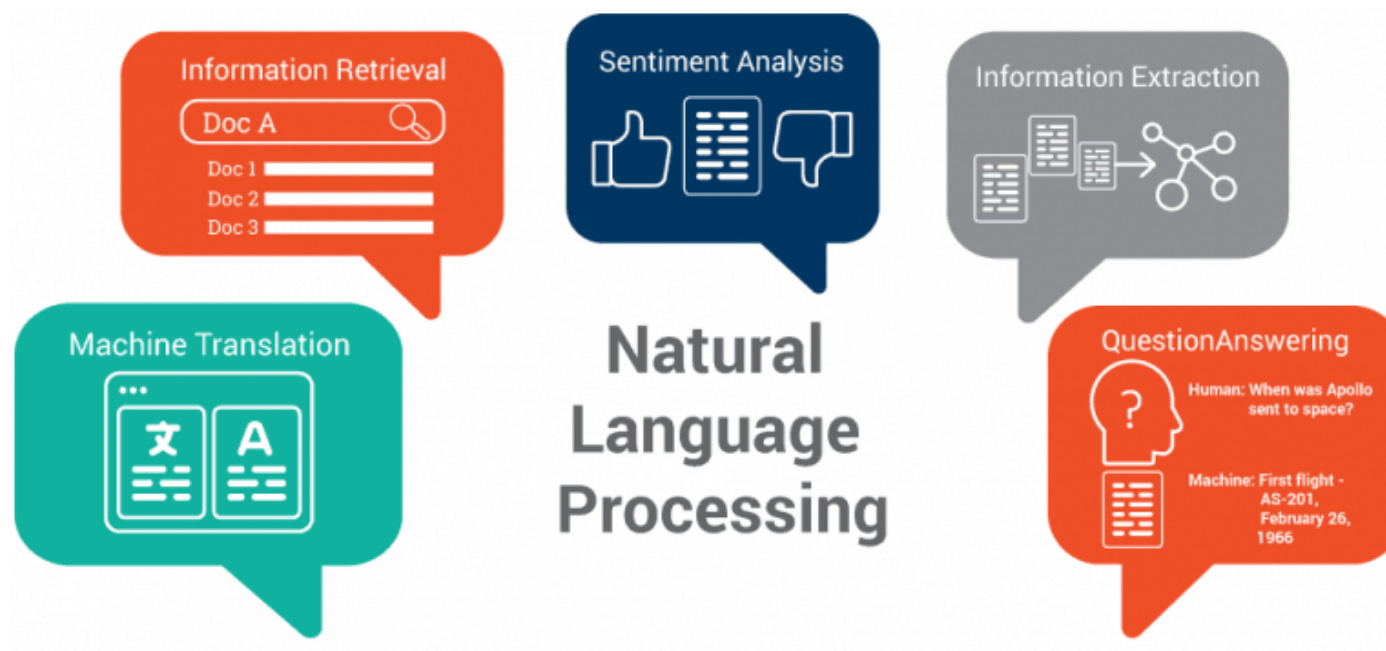
Normalização de Texto: Lematização e Stemming

A normalização de texto envolve a redução de palavras em suas formas base ou raízes. A lematização e o stemming são técnicas comuns usadas para alcançar esse objetivo, ajudando a reduzir a dimensionalidade do espaço de recursos e melhorar a precisão das análises de texto.

Palavra Original	Lematização	Stemming
correndo	correr	corr
andando	andar	and
inteligentemente	inteligente	intelig
programadora	programador	program
processando	processar	process

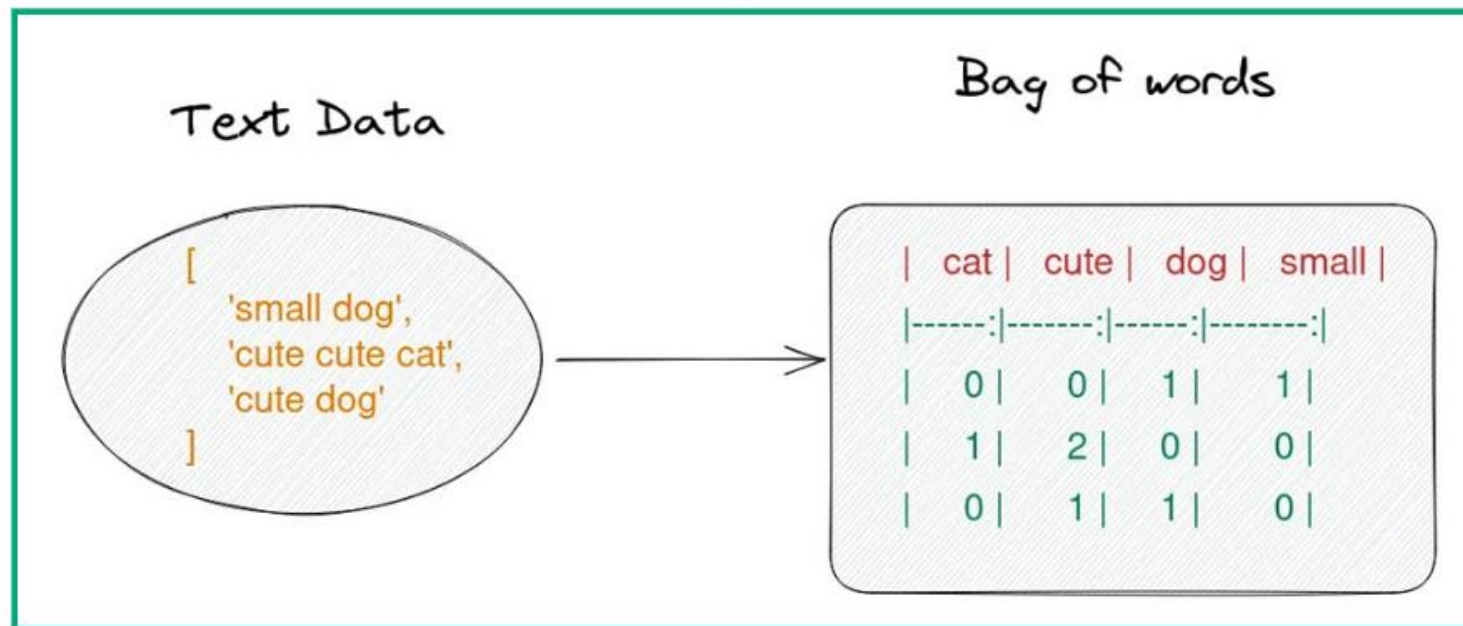
Representação de Texto

A representação de texto é fundamental no PLN para transformar o texto em formatos compreensíveis para os algoritmos de aprendizado de máquina.



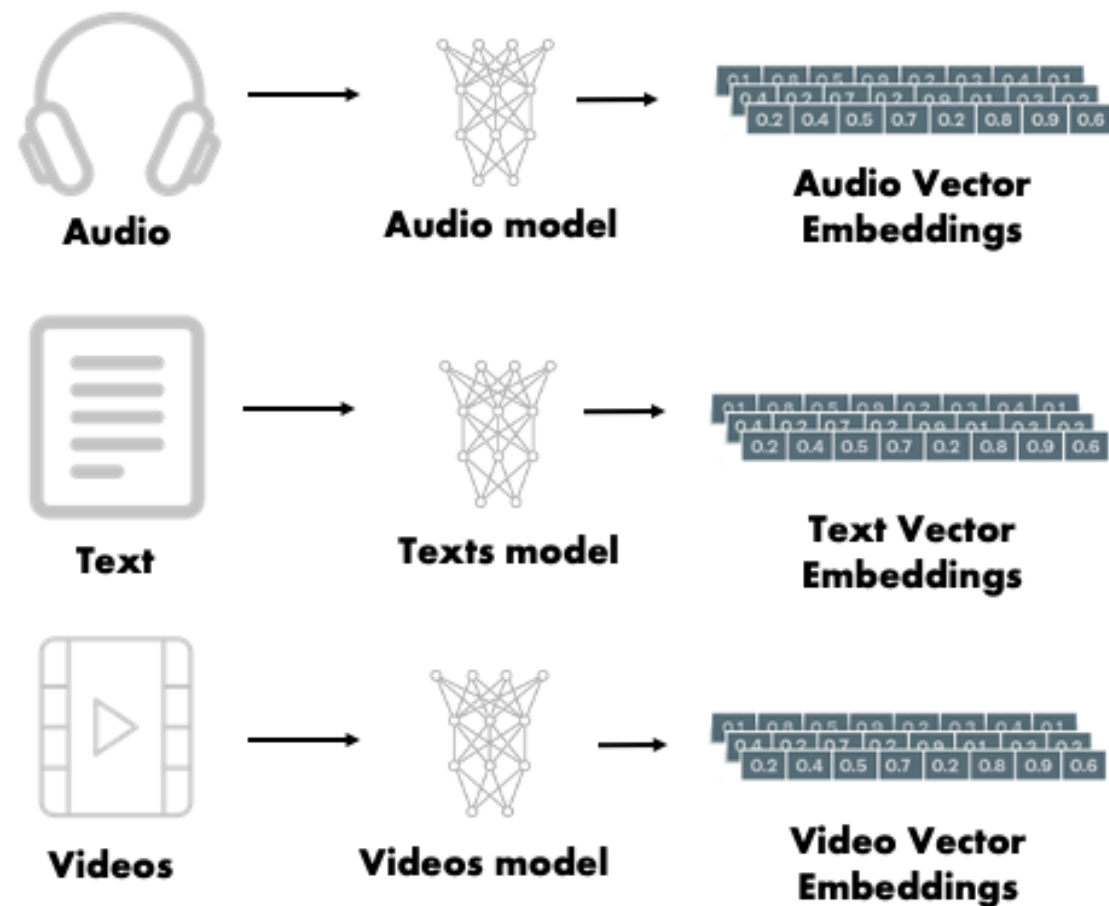
Saco de Palavras (Bag of Words)

Os modelos de saco de palavras representam um documento como um vetor contendo a contagem de ocorrências de cada palavra. O TF-IDF é uma variação que leva em consideração a frequência das palavras atribuindo pesos diferentes às palavras com base em sua importância.



Representação Vetorial de Palavras

Word embeddings são técnicas que representam palavras como vetores densos em um espaço vetorial de alta dimensão. Eles capturam o significado semântico e relações entre palavras com base em seu contexto de uso.



Modelos de Linguagem Neural, como BERT e GPT

Os modelos de linguagem neural, como BERT (Bidirectional Encoder Representations from Transformers) e GPT (Generative Pre-trained Transformer), são redes neurais pré-treinadas que alcançaram resultados impressionantes em uma variedade de tarefas de PLN. Eles capturam o contexto bidirecional das palavras em uma frase e geram texto coerente e relevante.



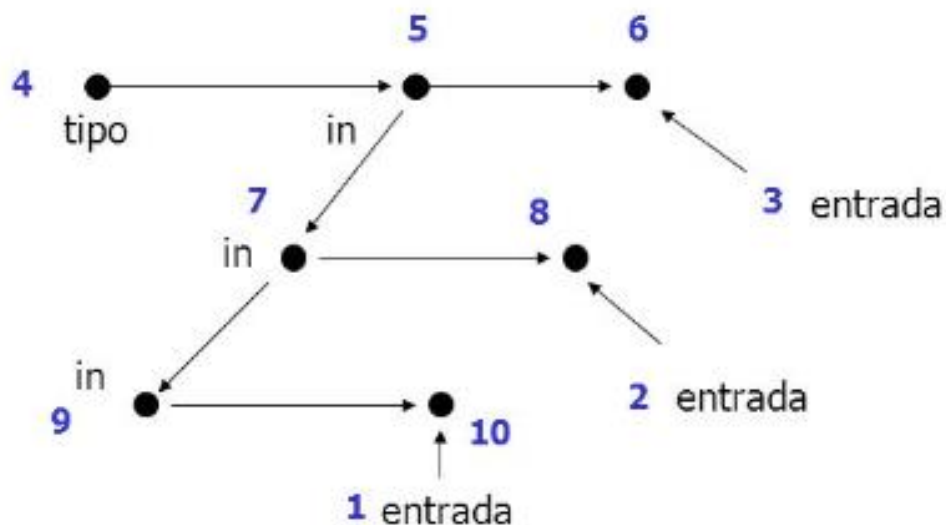
Análise de Partes do Discurso (POS Tagging)

A análise de partes do discurso é o processo de atribuir a cada palavra em um texto uma categoria gramatical, como substantivo, verbo, adjetivo, etc. Isso ajuda a entender a estrutura gramatical das frases.

Palavra	POS Tagging (NLTK)
Eu	Pronome (PRP)
estou	Verbo (VB)
correndo	Verbo (VB)
e	Conjunção (CC)
ela	Pronome (PRP)
está	Verbo (VB)
andando	Verbo (VB)
inteligentemente	Advérbio (RB)

Análise de Dependência

A análise de dependência identifica as relações sintáticas entre as palavras em uma frase, representando-as como um grafo de dependência. Isso ajuda a entender como as palavras estão conectadas umas às outras na estrutura da frase.



Reconhecimento de Entidades Nomeadas (NER)

O reconhecimento de entidades nomeadas é o processo de identificar e classificar entidades importantes em um texto, como nomes de pessoas, locais, organizações, datas, etc. Isso é útil para extrair informações importantes de textos não estruturados.

EXPERIENCE 1

LANGUAGE 2

HARD SKILL 3

SOFT SKILL 4

Pessoa Engenheira de Dados EXPERIENCE ↵

Hard skills: Linguagem Python HARD SKILL ; SQL HARD SKILL ; Modelagem de Dados HARD SKILL (Lakehouse HARD SKILL , Data Warehouse HARD SKILL). ↵

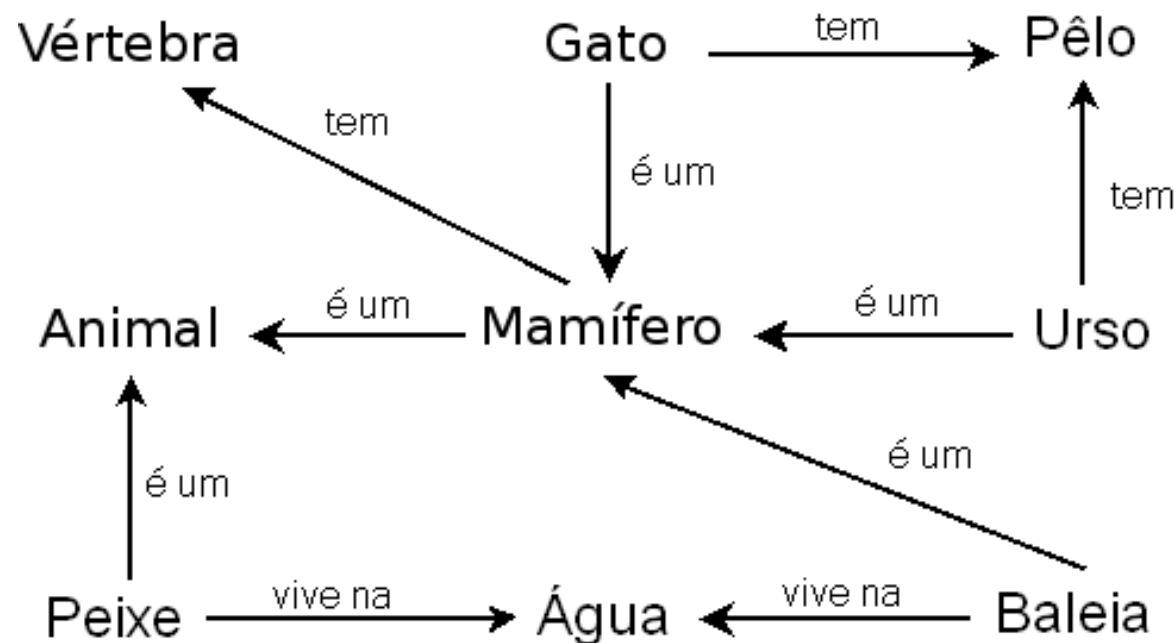
Soft skills: Trabalho em equipe SOFT SKILL ; Pensamento crítico SOFT SKILL ; Comunicação SOFT SKILL ; Ética profissional SOFT SKILL ; ↵

Idioma: Inglês LANGUAGE OU espanhol LANGUAGE .

SOURCE: manual_annotation JOB_ID: 1

Análise Semântica

A análise semântica aborda o significado das palavras e das frases em um contexto mais amplo. Isso inclui tarefas como compreensão de sentenças, análise de sentimentos e resolução de ambiguidades semânticas.



Tradução Automática e Alinhamento de Idiomas

A tradução automática é uma das aplicações mais conhecidas do PLN, permitindo a tradução instantânea de texto entre diferentes idiomas. O alinhamento de idiomas é o processo de encontrar correspondências entre textos em diferentes idiomas para facilitar a tradução.



Geração Automática de Texto

O PLN também é usado para gerar automaticamente texto novo e relevante em várias áreas, como resumos de texto, redação de artigos e criação de conteúdo para chatbots.



You

crie um texto sobre moralismo



ChatGPT

Moralismo é um conceito que permeia diversas esferas da vida humana, desde a ética individual até as estruturas sociais e políticas. Refere-se à aplicação rigorosa de padrões morais ou éticos, muitas vezes de forma inflexível ou dogmática. Esses padrões podem derivar de fontes diversas, como religião, filosofia, tradição cultural ou valores pessoais.

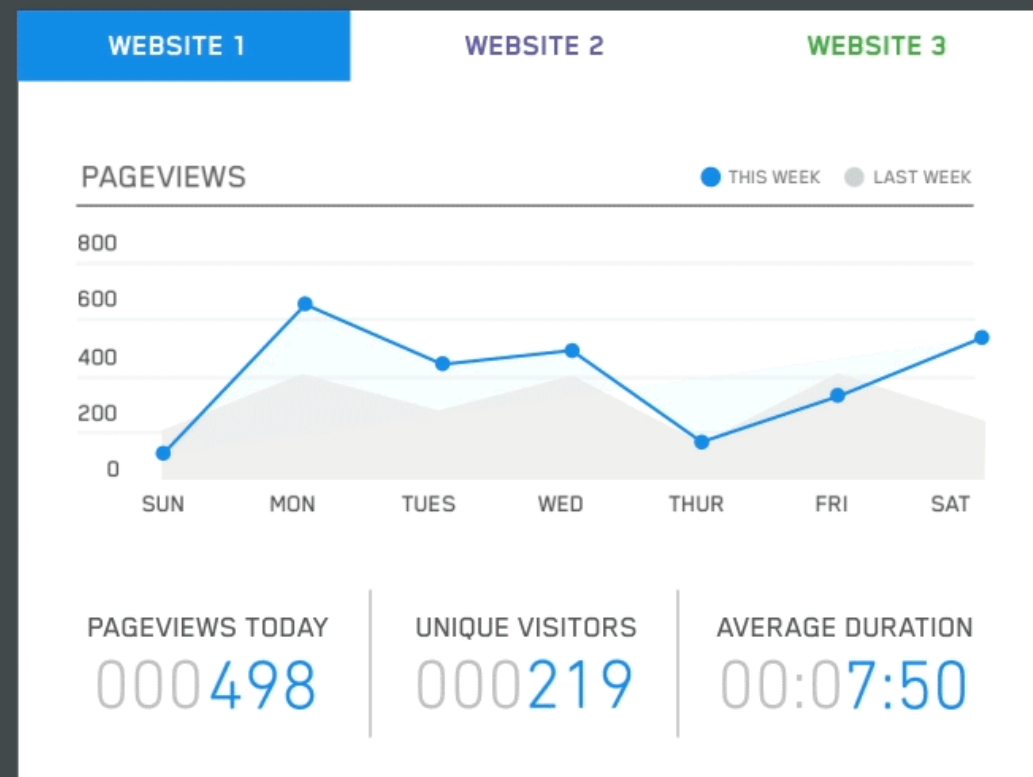
Chatbots e Assistentes Virtuais

Chatbots e assistentes virtuais são sistemas de software baseados em PLN que interagem com usuários em linguagem natural para responder a perguntas, fornecer informações ou executar tarefas específicas.



PLN em Big Data

O PLN também é usado para extrair informações úteis e insights de grandes volumes de texto não estruturado, como análise de sentimentos em mídias sociais, mineração de opiniões em avaliações de produtos e muito mais.



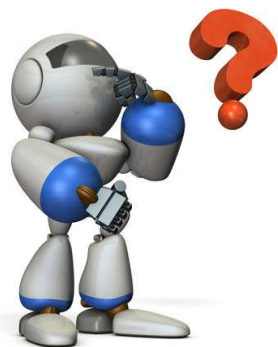
Polissemia e Ambiguidade na Linguagem Natural

A polissemia refere-se à ocorrência de uma mesma palavra com diferentes significados, enquanto a ambiguidade ocorre quando uma frase pode ter interpretações múltiplas. Lidar com essas nuances da linguagem natural é um desafio para os sistemas de PLN.



Variações Linguísticas, como Gírias e Jargões

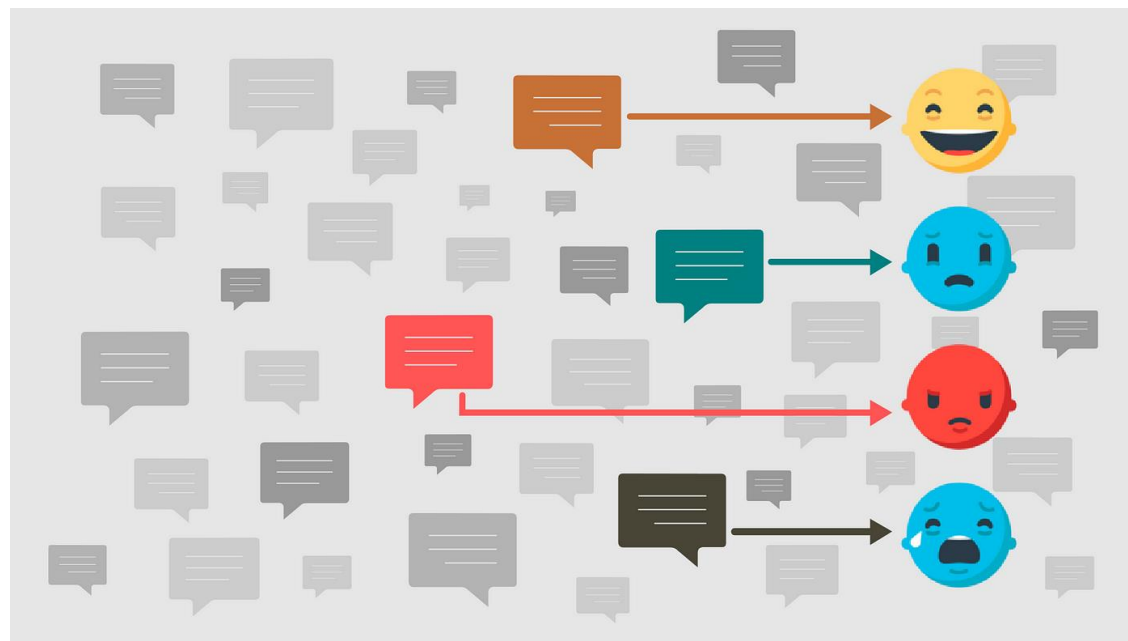
A linguagem natural é diversificada e pode variar de acordo com fatores como região geográfica, cultura e contexto social. Gírias, jargões e dialetos representam desafios adicionais para os sistemas de PLN.



**DÁ UMA
MORAL
(AÊ!)**

Textos Não Estruturados e Contextos Complexos

O PLN muitas vezes lida com textos não estruturados, como postagens em redes sociais, e precisa compreender o contexto e a intenção por trás das palavras para uma análise precisa. Isso pode ser especialmente desafiador em contextos complexos ou ambíguos.



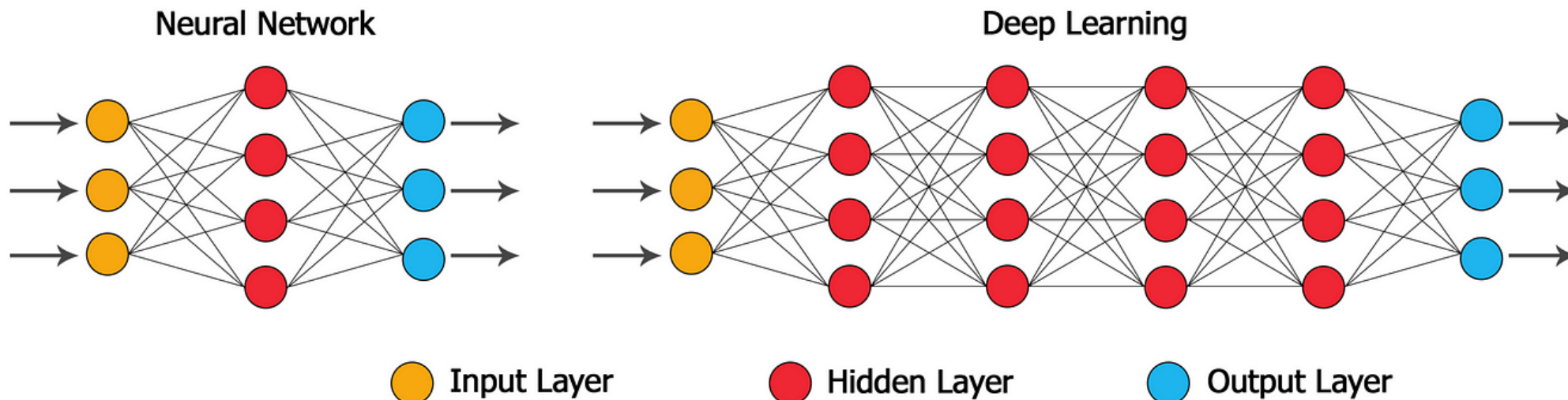
Privacidade e Segurança na Análise de Dados Pessoais

A análise de dados pessoais por meio do PLN levanta preocupações sobre privacidade e segurança. É essencial garantir que os dados dos usuários sejam protegidos e que a análise seja realizada de forma ética e transparente.



Avanços em Modelos de Linguagem Neural

Os modelos de linguagem neural estão em constante desenvolvimento, com novas arquiteturas e técnicas emergindo para melhorar a capacidade de compreensão e geração de texto em linguagem natural.



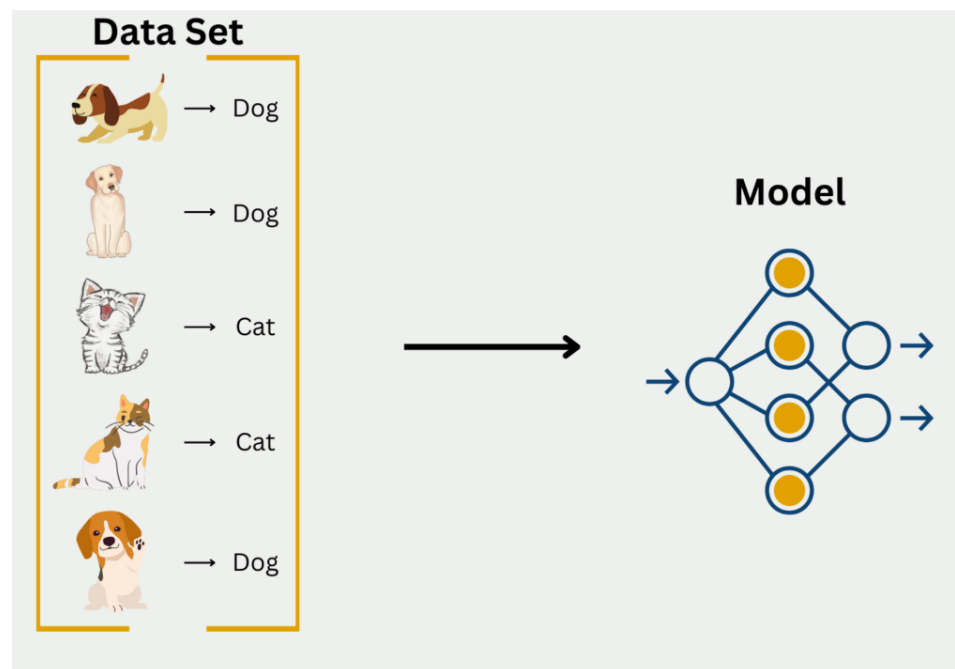
Integração de PLN com outras áreas da IA

O PLN está sendo cada vez mais integrado com outras áreas da Inteligência Artificial, como visão computacional e robótica, para criar sistemas mais inteligentes e abrangentes.

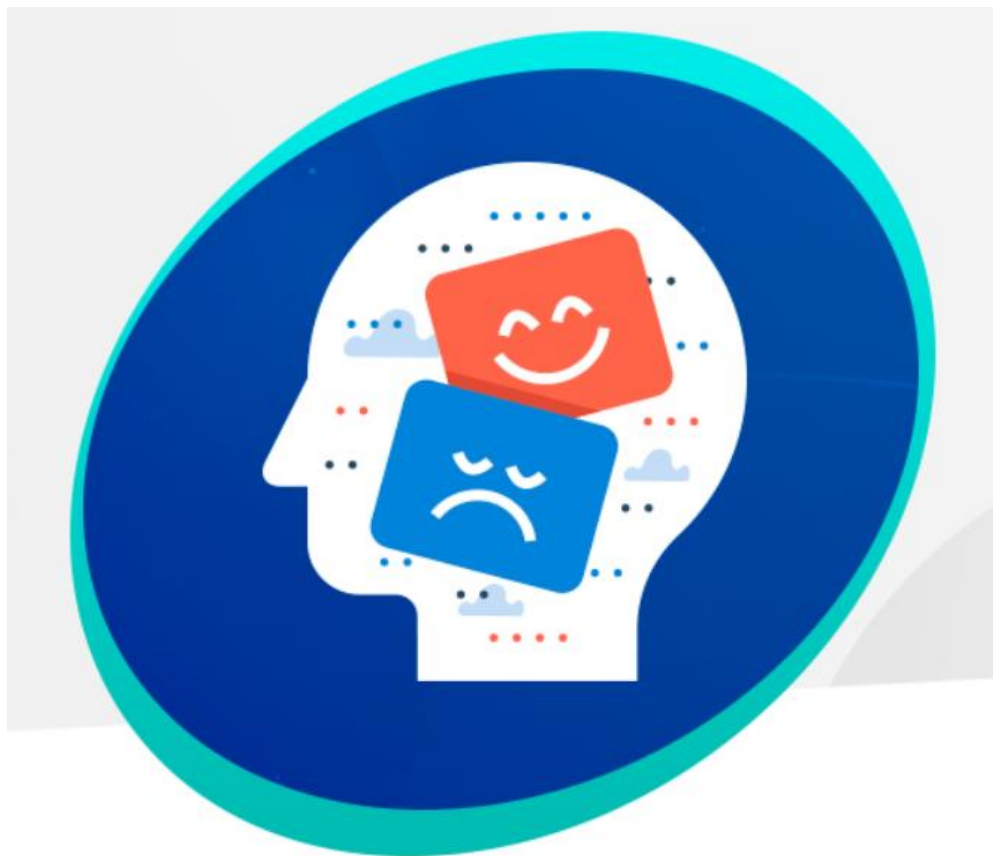


Técnicas de Transfer Learning

O Transfer Learning serve para aproveitar o conhecimento adquirido durante o treinamento de um modelo em uma tarefa específica e aplicá-lo a uma tarefa relacionada, geralmente com conjuntos de dados menores ou diferentes.



Gotit AI



<https://gotit.ai/pt-br/Home/Sentiment>

Referências Bibliográficas

1. Jurafsky, D., & Martin, J. H. (2020). Speech and Language Processing (3rd ed.). Pearson.
2. Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to Information Retrieval. Cambridge University Press.
3. Goldberg, Y. (2017). Neural Network Methods for Natural Language Processing. Synthesis Lectures on Human Language Technologies, 10(1), 1–309.
4. Bird, S., Klein, E., & Loper, E. (2009). Natural Language Processing with Python. O'Reilly Media, Inc.
5. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is All You Need. In Advances in Neural Information Processing Systems (pp. 5998-6008).
6. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners. OpenAI.



ATÉ A PRÓXIMA AULA!