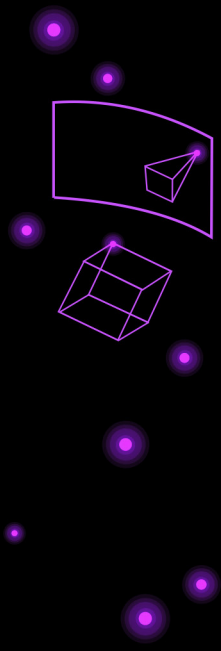
AKCIT
CENTRO DE COMPETÊNCIA EMBRAPII EM TECNOLOGIAS IMERSIVAS

Resolução de Problemas com Processamento de Linguagem Natural

Autoria:

Manoel Verissimo dos Santos Neto



Organizadores:

Deborah Silva Alves Fernandes
Taciana Novo Kudo
Renata Dutra Braga
Cristiane Bastos Rocha Ferreira
Arlindo Rodrigues Galvão Filho

Cegraf UFG

Resolução de problemas com Processamento de Linguagem Natural

Autoria:
Manoel Verissimo dos Santos Neto

Organizadores:
Deborah Silva Alves Fernandes
Taciana Novo Kudo
Renata Dutra Braga
Cristiane Bastos Rocha Ferreira
Arlindo Rodrigues Galvão Filho

Cegraf UFG

2024

Revisão Técnica
Cristiane Bastos Rocha Ferreira
Deborah Silva Alves Fernandes

Revisão Editorial
Ana Laura de Sene Amâncio Zara Brisolla

Capa
Iuri Vaz Miranda

Editoração Eletrônica
Luma Wanderley de Oliveira
Layane Grazielle Souza Dias

https://doi.org/10.5216/SAN.res.ebook.978-85-495-0960-4/2024

Dados Internacionais de Catalogação na Publicação (CIP)
(Câmara Brasileira do Livro, SP, Brasil)

Santos Neto, Manoel Verissimo dos
Resolução de problemas com processamento de linguagem natural [livro eletrônico] / Manoel Verissimo dos Santos Neto ; organizadores Deborah Silva Alves Fernandes...[et al.]. -- Goiânia, GO : Cegraf UFG, 2024.
PDF

Outros organizadores: Taciana Novo Kudo, Renata Dutra Braga, Cristiane Bastos Rocha Ferreira, Arlindo Rodrigues Galvão Filho.
Bibliografia.
ISBN 978-85-495-0960-4

1. Linguagem e línguas 2. Linguagem de programação para computadores 3. Linguística I. Fernandes, Deborah Silva Alves. II. Kudo, Taciana Novo. III. Braga, Renata Dutra. IV. Ferreira, Cristiane Bastos Rocha. V. Galvão Filho, Arlindo Rodrigues.

24-219347 CDD-005.133

Índices para catálogo sistemático:

1. Linguagem de programação : Computadores : Processamento de dados 005.133

Eliane de Freitas Leite - Bibliotecária - CRB 8/8415

Resolução de Problemas com Processamento de Linguagem Natural

Instituições responsáveis

- Universidade Federal de Goiás (UFG)
  - Centro de Competência Embrapii em Tecnologias Imersivas, denominado AKCIT (Advanced Knowledge Center for Immersive Technologies)
  - Centro de Excelência em Inteligência Artificial (CEIA)

Instituições financiadoras

- Empresa Brasileira de Pesquisa e Inovação Industrial (Embrapii)
- Governo do Estado de Goiás
- Empresas parceiras do AKCIT

Apoio

- Universidade Federal de Goiás (UFG)
  - Pró-Reitoria de Pesquisa e Inovação (PRPI-UFG)
  - Instituto de Informática (INF-UFG)

Lista de Abreviaturas

BERT	*Bidirectional Encoder Representations from Transformers* - Representações de Codificadores Bidirecionais de Transformadores
Bi-LSTM	*Bi-directional Long Short-Term Memory* - Memória de Longo Prazo Bidirecional
BPE	*Byte Pair Encoding* - Codificação de Pares de Bytes
GPT	*Generative Pre-trained Transformer* - Transformador Pré-treinado Generativo
GRU	*Gated Recurrent Units* - Unidades Recorrentes Fechadas
IA	Inteligência Artificial
IEEE	*Institute of Electrical and Electronics Engineers* - Instituto de Engenheiros Eletricistas e Eletrônicos
IR	*Informations Retrieval* - Retorno de Informações
LLaMA	*Large Language Model Meta AI* - Grande Modelo de Linguagem da META AI
LLM	*Large Language Models* - Grandes Modelos de Linguagem
LSTM	*Long Short-Term Memory* - Memória de Longo Prazo

MPT	*MosaicML Pretrained Transformer*
NER	*Named Entity Recognition* - Reconhecimento de Entidade Nomeada
NLP	*Natural Language Processing* - Processamento da Linguagem Natural
RAG	*Retrieval-Augmented Generation* - Recuperação Aumentada de Geração
RNN	*Recurrent Neural Networks* - Redes Neurais Recorrentes
TF-IDF	Term Frequency - *Inverse Document Frequency*

Lista de Figuras

Sumário

Apresentação

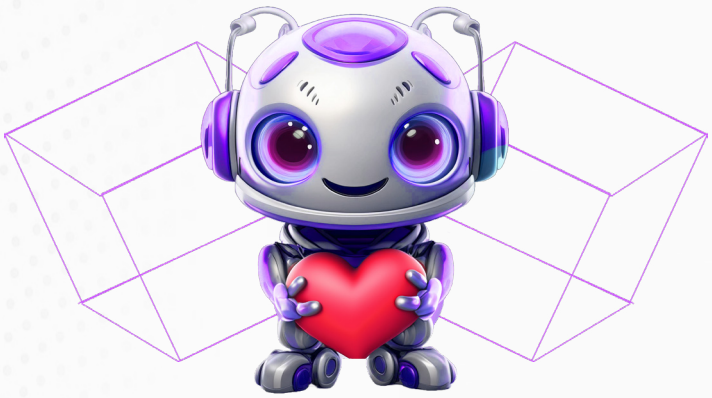
Prezado(a) Participante,

Seja bem-vindo(a) ao Microcurso **Resolução de Problemas com Processamento de Linguagem Natural**!

Este Microcurso faz parte da Coleção Formação e Capacitação do Centro de Competências Imersivas, uma parceria entre a Embrapii e a Universidade Federal de Goiás (UFG).

A sua oferta foi motivada com o objetivo de explicitar a aplicação dos princípios teóricos e práticos do Processamento de Linguagem Natural (NLP) na resolução de problemas complexos.

Esperamos que ao final deste Microcurso, você consiga entender a definição de NLP e os diferentes contextos e problemas para os quais NLP pode ser utilizado.



Desejamos um excelente estudo!!!

AKCIT
CENTRO DE COMPETÊNCIA EMBRAPII EM TECNOLOGIAS IMERSIVAS

Resolução de Problemas com Processamento de Linguagem Natural

Unidade I

Introdução à Linguagem Natural e ao Processamento de Linguagem Natural

Unidade I - Introdução à Linguagem Natural e ao Processamento de Linguagem Natural

Nesta Unidade, embarcamos em uma jornada fascinante pelo mundo da Linguagem Natural e do Processamento de Linguagem Natural (NLP). A linguagem natural, a forma como nos comunicamos cotidianamente, é complexa e rica em nuances. O NLP, por sua vez, é o campo de estudo que permite aos computadores compreenderem, interpretarem e responderem ao nosso idioma de maneira significativa. Vamos explorar como essas tecnologias estão revolucionando diversas áreas, desde assistentes virtuais até traduções automáticas, e compreender os princípios fundamentais que permitem essas interações avançadas entre humanos e máquinas. Prepare-se para descobrir como a união de linguística e inteligência artificial (IA) está moldando o futuro da comunicação.

Neste *ebook*, optamos por utilizar a sigla NLP, do inglês *Natural Language Processing*, em vez de PLN (Processamento de Linguagem Natural). Essa escolha se deve à ampla adoção e reconhecimento da sigla NLP na literatura acadêmica e na indústria global de tecnologia, especialmente em artigos, tutoriais e bibliotecas de códigos populares. A utilização de NLP facilita a busca por recursos e materiais adicionais, além de promover uma maior integração com comunidades internacionais e conteúdos em inglês, que dominam a maioria das publicações e inovações nessa área.

1.1 Visão geral da Linguagem Natural

A linguagem natural pode ser definida como todo e qualquer idioma que surja de forma espontânea entre os seres humanos, sem planejamento prévio, como produto da capacidade inata para a comunicação presente na mente humana.

A linguagem natural está intimamente relacionada com a ciência da linguística, que investiga os sistemas de comunicação inerentes aos seres humanos. Esse campo de estudo se aprofunda na compreensão de como as línguas são estruturadas, utilizadas e evoluem, enfatizando a intrincada relação entre linguagem e pensamento humano. Dentro da linguística, a análise da linguagem natural abrange diversos níveis estruturais, tais como: fonética e fonologia que estuda os sons; morfologia que examina a construção das palavras a partir de unidades mínimas de significado; sintaxe que explora a organização dessas palavras em frases; e semântica que se ocupa dos significados que essas combinações comunicam.

Um outro ponto importante da linguagem natural é a pragmática, que explora como o contexto afeta o entendimento das palavras. Isso significa que uma mesma frase pode ser interpretada de maneiras diferentes dependendo da situação em que é usada. Esse

campo mostra como a linguagem se ajusta para atender às diferentes necessidades de comunicação das pessoas em variados contextos.

Vale a pena ressaltar que, no contexto da linguística, existem variações e mudanças, de forma que a linguística revela a incrível adaptabilidade e a dinâmica da linguagem natural. Essas mudanças e variações não apenas demonstram como a linguagem varia entre diferentes grupos sociais, regiões ou períodos históricos, mas, também, analisam as transformações que ocorrem nas línguas ao longo do tempo. Essa compreensão mais profunda da linguagem humana ajuda a iluminar sua riqueza e complexidade, bem como a maneira como ela evolui e se adapta continuamente.

1.2 Visão Geral do Processamento de Linguagem Natural (NLP), Termos-chave e Conceitos Fundamentais

O NLP é uma subárea da inteligência artificial (IA) e da linguística computacional que se concentra na interação entre computadores e linguagem humana, sendo uma tecnologia que dá a capacidade aos computadores de "ler, entender e processar" a linguagem natural humana.

O NLP tem suas origens nas décadas de 1950 e 1960, com os primeiros sistemas de tradução automática e a proposta do "Teste de Turing" como critério de inteligência. Na década de 1980, surgiram abordagens estatísticas, como os modelos ocultos de Markov. O crescimento da *web*, na década de 2000, impulsionou métodos baseados em dados, levando ao desenvolvimento de sistemas de busca e análise de sentimentos. Com a ascensão do *deep learning* e o surgimento da arquitetura *Transformers* em 2017, iniciou-se uma revolução em NLP. O *Bidirectional Encoder Representations from Transformers* (BERT), que é um modelo baseado na arquitetura *Transformers*, alcançou avanços significativos em tarefas como tradução automática e sumarização de texto, dentre outras.

Para se chegar nessa arquitetura que revolucionou a resolução de problemas de NLP, foram utilizadas outras arquiteturas e técnicas nesses problemas, exemplificadas na Figura 1.

Figura 1 - Arquiteturas e técnicas de Processamento de Linguagem Natural

Fonte: autoria própria.

Bag of Words - Técnica que representa documentos como vetores, ignorando a estrutura e a ordem das palavras, e focando apenas na frequência de ocorrência das palavras individuais. Essa abordagem cria um "saco" (*bag*) de palavras únicas presentes no texto e, em seguida, gera um vetor para cada documento, onde cada componente do vetor representa a frequência ou a contagem das palavras do "saco" no documento. Em termos simples, é uma coleção de palavras para representar uma frase, com contagem de palavras e, na maioria das vezes, desconsiderando a ordem em que aparecem. Essa técnica foi amplamente utilizada em tarefas de classificação de texto, recuperação de informação e análise de sentimento.

Word Embeddings - Representações distribuídas de palavras em um espaço vetorial de alta dimensão, no qual palavras semanticamente semelhantes são mapeadas para vetores próximos. Essas representações capturam o significado e a semântica das palavras com base em seu contexto em um corpus de texto. Uma técnica comum para criar *word embeddings* é o Word2Vec, que utiliza redes neurais para aprender representações vetoriais de palavras a partir de grandes volumes de texto não rotulados.

Por exemplo, no espaço de *word embeddings*, palavras semanticamente relacionadas, como "rei" e "rainha", tendem a estar próximas umas das outras, enquanto palavras com significados diferentes, como "cachorro" e "computador", estarão distantes.

Essas representações densas e semânticas de palavras têm várias aplicações em NLP, como, por exemplo, recuperação de informação, tradução automática e análise de sentimentos.

RNN - Modelos baseados em Redes Neurais Recorrentes (RNN) são uma classe de modelos de aprendizado profundo que processam sequências de dados, como texto ou áudio, levando em consideração a dependência temporal entre os elementos da sequência. As RNNs são capazes de lidar com entradas de comprimento variável e capturar informações contextuais de longo prazo por meio de mecanismos de realimentação.

A arquitetura das RNNs permite que informações de etapas anteriores sejam levadas em consideração ao processar cada elemento da sequência, tornando-as eficazes em tarefas que exigem compreensão de contexto, como tradução automática, geração de texto e análise de sentimentos.

No entanto, as RNNs tradicionais enfrentam dificuldades em lidar com dependências de longo prazo devido a problemas de treinamento do modelo. Esse problema está relacionado a cálculos matemáticos complexos que resultam em erros durante o processo de treinamento. Esses erros impedem que o modelo realize efetivamente o "aprendizado", comprometendo sua capacidade de processar informações que dependem de sequências longas de dados. Para contornar essa limitação, surgiram variantes de RNNs, como as redes *Long Short-Term Memory* (LSTM) e *Gated Recurrent Units* (GRU), que foram projetadas para capturar dependências de longo prazo de forma mais eficaz.

LSTM - É uma arquitetura especial de RNN que foi projetada para resolver os problemas matemáticos complexos das RNNs tradicionais, permitindo que as redes capturem dependências de longo prazo em sequências de dados. Elas foram introduzidas por Hochreiter e Schmidhuber em 1997.

Ao contrário das RNNs tradicionais, que podem ter dificuldades em manter informações relevantes em longas sequências de dados, as LSTMs possuem unidades de memória especializadas que podem lembrar informações importantes por longos períodos de tempo. Essas unidades de memória possuem três portas principais: a porta de entrada (*input gate*), a porta de esquecimento (*forget gate*) e a porta de saída (*output gate*), que controlam o fluxo de informações dentro da célula LSTM.

As LSTMs têm sido amplamente utilizadas em uma variedade de tarefas de NLP, como tradução automática, geração de texto e análise de sentimentos, devido à sua capacidade de capturar dependências de longo prazo e lidar com sequências de dados de comprimento variável.

Bi-directional LSTM - As RNN Bidirecionais (Bi-LSTM) são uma extensão das redes LSTM que permitem a fluência da informação tanto para frente quanto para trás na sequência de entrada. Isso significa que as Bi-LSTMs podem capturar informações contextuais não apenas do passado, como as LSTMs tradicionais, mas também do futuro, o que pode ser útil em muitas tarefas de NLP, como Reconhecimento de Entidades Nomeadas (NER), análise de sentimentos e tradução automática.

Na arquitetura das Bi-LSTMs, duas camadas LSTM são empregadas: uma que processa a sequência na direção original (*forward*) e outra que processa a sequência na direção reversa (*backward*). As saídas dessas duas camadas são então combinadas, geralmente concatenadas, para fornecer uma representação contextual abrangente de cada elemento na sequência de entrada.

As Bi-LSTMs são amplamente utilizadas em tarefas onde a compreensão do contexto é crucial, pois permitem que o modelo capture informações de contexto tanto antes quanto depois de cada palavra em uma frase ou sequência de texto.

Attention Based - O conceito de "*attention mechanism*" ou mecanismo de atenção em redes neurais é um componente importante no campo da aprendizagem profunda, especialmente em tarefas que envolvem sequências de dados, em NLP e também em análise de séries temporais. Ele foi projetado para melhorar a capacidade dos modelos de focar em diferentes partes de uma entrada para uma representação mais eficiente.

Em sua essência, o mecanismo de atenção permite que um modelo de rede neural se concentre em aspectos específicos de sua entrada, o que é particularmente útil em contextos onde a entrada é uma sequência de dados. Por exemplo, ao traduzir uma sentença de um idioma para outro, pode ser útil focar em uma palavra específica ou em uma parte da sentença enquanto se ignora o resto temporariamente. Isso é semelhante à maneira como os humanos se concentram em informações pertinentes enquanto ignoram o irrelevante.

O mecanismo de atenção geralmente calcula pesos para diferentes partes da entrada, indicando quão importantes elas são para uma tarefa específica. Esses pesos são então usados para criar uma combinação ponderada dos vetores de entrada, que é passada para as próximas camadas do modelo. Essencialmente, isso permite que o modelo adapte dinamicamente o foco da atenção aos dados relevantes.

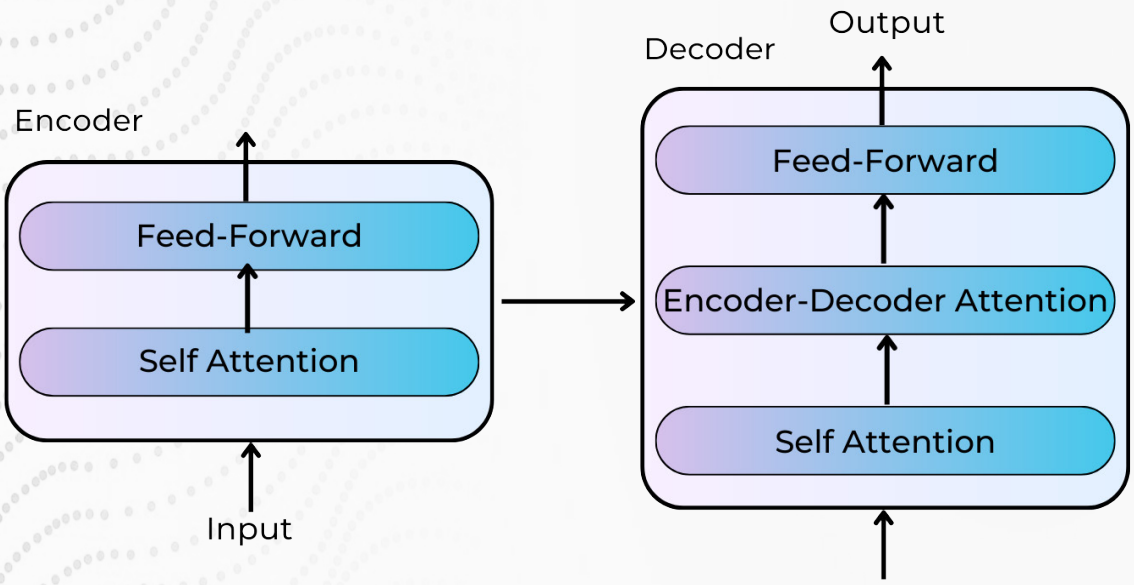
Transformers - Essa arquitetura é uma grande inovação no campo de NLP, que se destaca por algumas características principais:

- **Mecanismo de atenção:** em vez de processar sequencialmente como RNNs, *os transformers* usam um mecanismo de atenção para identificar quais partes de uma sequência de entrada são mais importantes para entender a sequência. Isso permite que o modelo identifique toda a sequência de uma vez e se concentre nas partes mais relevantes;
- **Paralelismo:** diferente das RNNs, que processam dados sequencialmente, os *Transformers* permitem processamento paralelo, o que acelera significativamente o treinamento e a execução;
- **Codificadores e decodificadores:** a arquitetura é composta por duas partes principais - codificadores (que leem e processam a entrada) e decodificadores (que geram a saída). Essa estrutura é especialmente útil para tarefas como tradução de idiomas;
- **Representações posição-dependentes:** como os *transformers* não possuem um mecanismo interno para entender a ordem dos dados, usam *embeddings* de posição para incluir informações sobre a posição de cada palavra na sequência.

Essas características tornam os *transformers* extremamente eficazes e flexíveis para várias tarefas de linguagem natural, incluindo tradução, resumo de texto, e resposta a perguntas.

Na Figura 2, observa-se um desenho simplificado da arquitetura *transformers*, com toda a complexidade em torno desta arquitetura que revolucionou NLP nos últimos anos. Essa arquitetura é usada para criar Grandes Modelos de Linguagem (LLMs), como o ChatGPT®, que é um dos mais conhecidos. O ChatGPT® utiliza essa estrutura nas diferentes versões do modelo GPT.

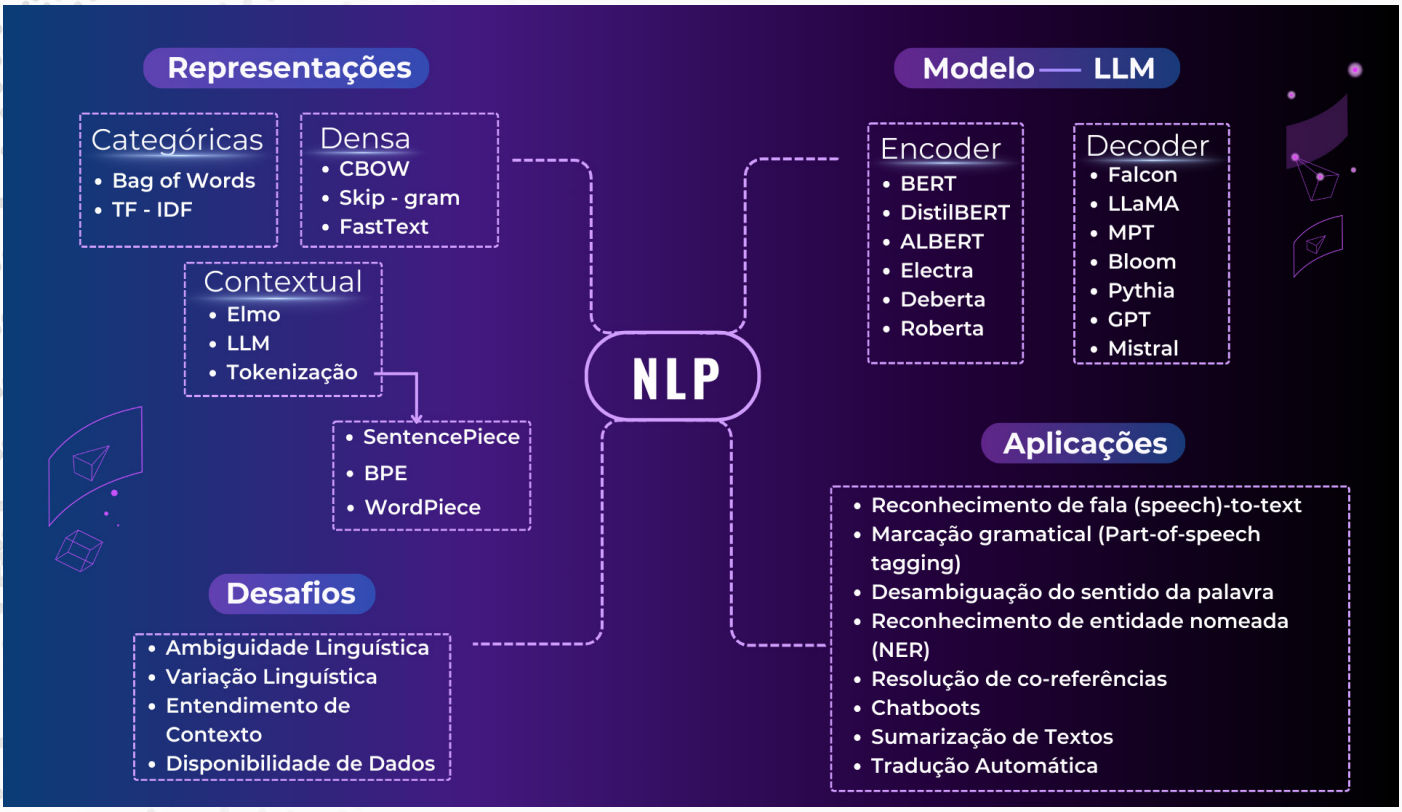
Figura 2 - Arquitetura *transformers*



Fonte: autoria própria.

Na Figura 3, encontra-se um resumo sobre NLP, onde são mostradas algumas técnicas e diversas aplicações.

Figura 3 - Resumo de Processamento de Linguagem Natural



Fonte: autoria própria.

Saiba mais...

- Uma introdução ao Bag of Words e como programá-lo em Python para PLN (Dubey, 2018; Trad. Sousa, 2023).
- Word Embedding: fazendo o computador entender o significado das palavras (Fonseca, 2021).
- Redes Neurais Recorrentes — LSTM (Olah, 2015; Trad. Junior, 2019)
- Um breve estudo dos Transformers (Alves, 2022).

Resolução de Problemas com Processamento de Linguagem Natural

Unidade II

Principais Técnicas em Processamento de Linguagem Natural

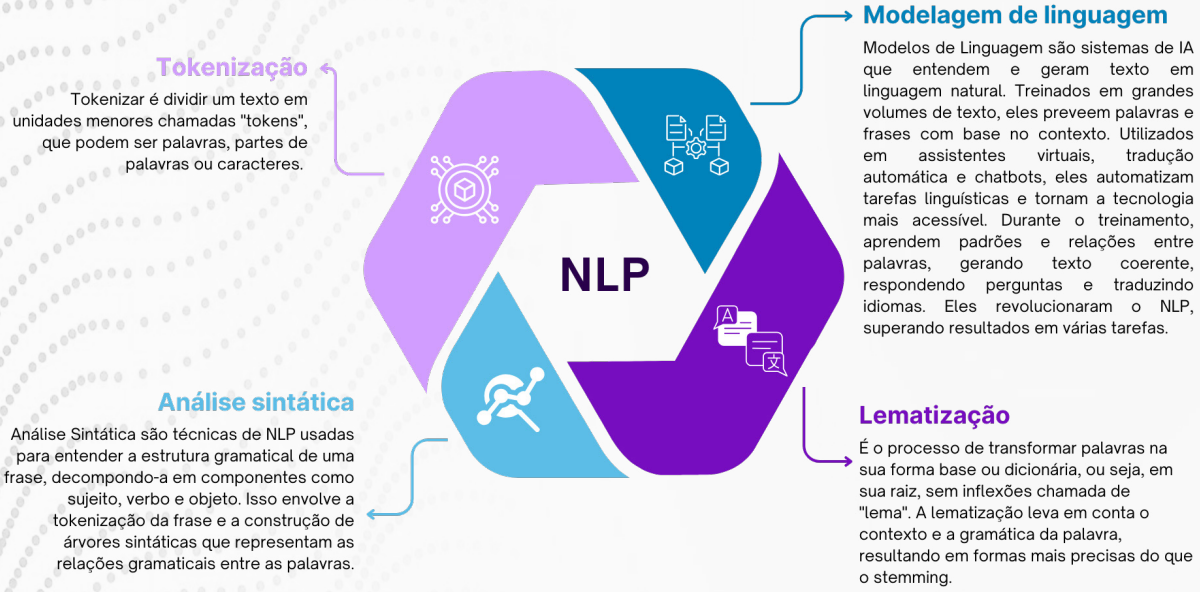
Unidade II - Principais Técnicas em Processamento de Linguagem Natural

Nesta Unidade, vamos explorar as principais técnicas utilizadas em NLP, oferecendo uma visão abrangente das suas definições, acompanhadas de exemplos ilustrativos. Nas seções seguintes, serão apresentados estudos de caso e aplicações práticas que demonstram como essas técnicas são implementadas em contextos reais, destacando sua relevância e impacto em diversas áreas. Essa abordagem visa proporcionar uma compreensão clara e prática das ferramentas e metodologias que formam a base do NLP, facilitando a aplicação desses conhecimentos em projetos futuros.

2.1 Principais Técnicas Utilizadas em NLP

As principais técnicas utilizadas em NLP estão apresentadas no infográfico a seguir (Figura 4). Começando pela **tokenização**, essa divide o texto em unidades menores, como palavras ou frases, facilitando a análise subsequente. A **remoção de *stop words*** elimina palavras comuns e frequentemente usadas que não agregam muito significado, como "e", "de" e "o", ajudando a focar nos termos mais importantes. Técnicas de ***stemming* e lematização** são usadas para reduzir palavras às suas formas básicas, melhorando a consistência dos dados. ***Parsing* e análise sintática** examinam a estrutura gramatical das frases, permitindo uma compreensão mais profunda do texto. ***Word embeddings*** transformam palavras em vetores numéricos, capturando seus significados e relações contextuais. Finalmente, os **modelos de linguagem** são sistemas avançados que entendem e geram texto coerente em linguagem natural, baseando-se em grandes volumes de dados para realizar uma ampla gama de tarefas de NLP. Juntas, essas técnicas formam a base das aplicações modernas em NLP, proporcionando ferramentas poderosas para a análise e geração da linguagem natural.

Figura 4 - Principais técnicas utilizadas em Processamento de Linguagem Natural



Fonte: autoria própria.

2.1.1 Tokenização

"Tokenizar" é o processo de dividir um texto em unidades menores, chamadas de "tokens". Esses tokens podem ser palavras, partes de palavras ou até caracteres individuais, dependendo do nível de granularidade desejado. A tokenização é um passo fundamental em muitas tarefas de NLP porque transforma o texto bruto em um formato que pode ser analisado pelos algoritmos de aprendizado de máquina. Na prática e de forma simples, essa técnica é uma maneira de transformar um texto em "número" para que algoritmos de aprendizado de máquina possam realizar o aprendizado.

2.1.2 Remoção de Stop Words

Remover *stop words* é o processo de eliminar palavras comuns e frequentemente usadas em um texto que, geralmente, não agregam muito significado ao conteúdo. Exemplos de *stop words* em português incluem "o", "a", "de", "que", "e", entre outros. A remoção dessas palavras ajuda a focar nos termos mais importantes e significativos do texto. Além disso, essa remoção reduz a quantidade de dados a ser processada, tornando os modelos mais rápidos e eficientes.

2.1.3 Stemming e Lematização

Stemming e lematização são técnicas usadas em NLP para reduzir palavras às suas formas básicas ou raízes. Ambas ajudam a normalizar o texto, tornando mais fácil para os algoritmos entenderem o significado das palavras. ***Stemming*** é o processo de cortar o final ou os afixos de palavras para reduzir variantes de uma palavra à sua raiz ou "*stem*".

Por exemplo, as palavras “correndo”, “corre” e “corrida” podem ser reduzidas à raiz “corr”. **Lematização** é o processo de transformar palavras na sua forma base ou dicionária, chamada de “lema”. A lematização leva em conta o contexto e a gramática da palavra, resultando em formas mais precisas do que o *stemming*.

2.1.4 *Parsing* e Análise Sintática

Parsing e análise sintática são técnicas de NLP usadas para entender a estrutura gramatical de uma frase, decompondo-a em componentes como sujeito, verbo e objeto. Isso envolve a tokenização da frase e a construção de árvores sintáticas que representam as relações gramaticais entre as palavras. *Parsing* é essencial para a compreensão de texto, resolução de ambiguidades gramaticais e extração de informações, sendo utilizado em tarefas como tradução automática e resposta a perguntas. As abordagens incluem *parsers* baseados em regras, *parsers* estatísticos que utilizam aprendizado de máquina, e *parsers* de *deep learning* que capturam relações complexas entre palavras.

2.1.5 Word Embeddings

Word embeddings são representações numéricas de palavras em um espaço vetorial que capturam o significado e as relações entre elas. Essas representações transformam palavras em vetores de números, onde palavras com significados semelhantes têm vetores próximos no espaço vetorial. Treinados em grandes corpora de texto, os *word embeddings* permitem algoritmos de NLP entenderem o contexto das palavras de maneira mais precisa, melhorando tarefas como tradução automática, análise de sentimentos e resposta a perguntas. Modelos populares, como Word2Vec, GloVe e FastText, utilizam diferentes técnicas para gerar esses vetores, permitindo capturar nuances semânticas e contextuais das palavras de forma eficiente e versátil.

2.1.6 Modelagem de Linguagem

Modelos de linguagem são sistemas de IA que entendem e geram texto em linguagem natural, treinados em grandes volumes de texto para prever palavras e frases com base no contexto. Eles são usados em assistentes virtuais, tradução automática e *chatbots*, permitindo a automação de tarefas linguísticas e tornando a tecnologia mais acessível. Durante o treinamento, os modelos aprendem padrões e relações entre palavras, permitindo-lhes gerar texto coerente, responder perguntas e traduzir idiomas. Os modelos de linguagem fizeram uma revolução em NLP, no quais vários resultados de várias tarefas foram superados com o uso desses modelos.

2.2 Demonstração da aplicação prática das técnicas em NLP

Nesta seção, as principais técnicas utilizadas em NLP serão demonstradas de forma prática para tangibilizar o conhecimento e facilitar o entendimento. A tokenização será abordada, mostrando como dividir textos em unidades menores, como palavras e frases.

A remoção de *stop words* será exemplificada, explicando como eliminar palavras comuns que pouco agregam ao significado do texto. Técnicas de *stemming* e lematização serão detalhadas, destacando como reduzir palavras às suas formas básicas. A análise de sentimentos será realizada, mostrando como identificar emoções e opiniões em textos. Aplicaremos o NER para demonstrar como identificar e classificar nomes de pessoas, locais e organizações. A análise sintática será abordada, detalhando como examinar a estrutura gramatical das frases. *Word embeddings* serão demonstrados, explicando como transformar palavras em vetores numéricos para capturar significados contextuais. Também serão apresentados exemplos de tradução automática e sumarização de textos, mostrando como converter textos entre idiomas e criar versões condensadas. Por fim, serão discutidos modelos de linguagem avançados, que entendem e geram texto natural, aplicáveis em diversas tarefas de NLP.

2.2.1 Tokenização

Ao escrever um documento, o editor pode usar a tokenização para dividir o texto em palavras individuais. Isso permite que o *software* identifique palavras incorretas e sugira correções. Por exemplo, se você digita “teh” em vez de “the”, o sistema tokeniza essa palavra e a compara com um dicionário, sugerindo a correção apropriada. Esse processo melhora a precisão da escrita e auxilia na correção de erros ortográficos de forma automática e eficiente.

2.2.2 Remoção de Stop Words

A remoção de *stop words* é crucial em tarefas de mineração de texto e análise de sentimentos, na qual a eficiência e a precisão são essenciais. Por exemplo, em um sistema de análise de opiniões de produtos, comentários de usuários são processados para determinar a satisfação geral. *Stop words*, como “e”, “mas”, “o”, “de”, são removidas porque não carregam informações significativas e podem prejudicar a análise. Ao eliminá-las, o algoritmo se concentra em palavras relevantes, como “excelente”, “ruim”, “qualidade”, permitindo uma análise mais rápida e precisa dos sentimentos expressos, ajudando empresas a entender melhor o *feedback* dos clientes e a tomar decisões informadas.

2.2.3 *Stemming* e Lematização

Stemming e lematização são técnicas essenciais em sistemas de busca para melhorar a correspondência de consultas e resultados. Em um motor de busca de uma biblioteca digital, por exemplo, essas técnicas são usadas para normalizar palavras de forma que variações morfológicas sejam tratadas como equivalentes. Se um usuário pesquisa por “correndo”, o sistema aplica *stemming* ou lematização para reduzir a palavra à sua raiz ou forma base, “correr”. Isso permite que o motor de busca retorne resultados que contenham

"correr", "correu", "corrida", entre outras variações, garantindo uma cobertura mais ampla e precisa dos documentos relevantes, melhorando significativamente a experiência do usuário ao encontrar informações pertinentes.

2.2.4 *Parsing* e Análise Sintática

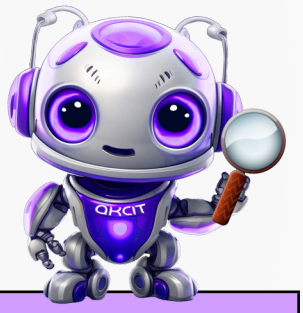
Parsing e análise sintática são fundamentais na construção de sistemas de tradução automática, na qual a compreensão precisa da estrutura gramatical das frases é crucial. Por exemplo, ao traduzir textos de um idioma para outro, um sistema de tradução automática utiliza *parsing* para decompor as frases de entrada em suas estruturas gramaticais, identificando sujeitos, verbos, objetos e outros componentes sintáticos. Essa análise permite que o sistema compreenda o significado e a relação entre as palavras, produzindo traduções mais precisas e gramaticalmente corretas. Dessa forma, *parsing* e análise sintática melhoram significativamente a qualidade das traduções, facilitando a comunicação e o entendimento entre falantes de diferentes idiomas.

2.2.5 *Word Embeddings*

Word embeddings são amplamente utilizadas em sistemas de recomendação de conteúdo, como em plataformas de *streaming* de música ou vídeo. Por exemplo, em um serviço de música, *word embeddings* podem ser aplicadas para analisar e entender o contexto e a similaridade entre palavras nas descrições de músicas, gêneros e artistas. Com essas representações vetoriais, o sistema pode identificar relações semânticas, como artistas semelhantes ou músicas com temas líricos parecidos. Isso permite que a plataforma recomende novas músicas ou artistas que estejam alinhados com as preferências e o histórico de escuta do usuário, proporcionando uma experiência de descoberta de conteúdo mais personalizada e envolvente.

2.2.6 Modelagem de Linguagem

A modelagem de linguagem é fundamental na criação de assistentes virtuais inteligentes, como o Google Assistant® ou a Alexa®, da Amazon®. Esses assistentes utilizam modelos de linguagem avançados para compreender e gerar texto ou fala em linguagem natural. Por exemplo, quando um usuário faz uma pergunta ou dá um comando de voz, o modelo de linguagem processa a entrada, compreendendo o contexto e a intenção por trás das palavras. Em seguida, o assistente gera uma resposta relevante ou executa uma ação específica, como definir um alarme, fornecer informações sobre o clima ou tocar uma música. A modelagem de linguagem permite que esses assistentes entendam nuances, sinônimos e a estrutura da linguagem humana, oferecendo interações mais naturais e eficientes.



Saiba mais...

- A tokenização no processo de linguagem natural e análise de texto (Santos, 2022).
- Otimização do processamento e a filtragem de *Stop Words*: meus estudos em spaCy e NLP — Parte 3 (Surreaux, 2024).
- *Stemming* e lematização em Python (Pykes, 2024).
- *Word Embeddings* — Representação vetorial de textos para *Machine Learning* (*Technology and Artificial Intelligence League* - TAIL, 2020)

Unidade III

Problemas Comuns Resolvidos com Processamento de Linguagem Natural

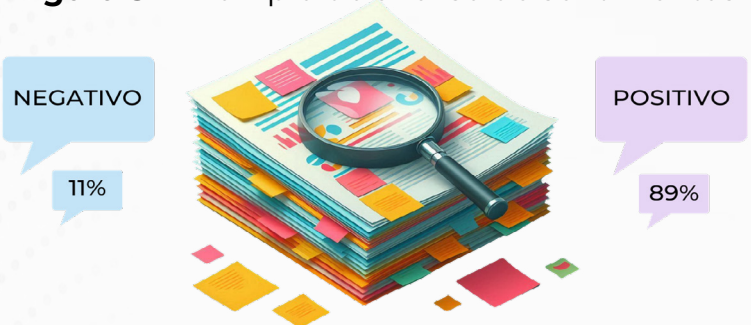
Unidade III - Problemas Comuns Resolvidos com Processamento de Linguagem Natural

No campo de NLP, vários desafios comuns emergem ao lidar com a complexidade e sutilezas da linguagem humana. A **análise de sentimentos**, por exemplo, enfrenta dificuldades em detectar ironia, sarcasmo e nuances emocionais, muitas vezes, resultando em interpretações imprecisas. O **NER** lida com a ambiguidade, em que palavras ou nomes podem ter múltiplos significados ou pertencer a diferentes categorias, complicando a tarefa de classificação correta. O **retorno de informações (IR)** luta com a relevância e a precisão, especialmente em consultas vagamente definidas ou em contextos onde a desambiguação é necessária. A **sumarização de documentos** enfrenta o desafio de manter a coerência e a integridade do texto original, muitas vezes, omitindo informações críticas ou gerando resumos que não capturam totalmente o contexto. Os **assistentes virtuais e *chatbots*** lidam com a compreensão do contexto, adaptabilidade e a capacidade de fornecer respostas precisas e relevantes em conversas dinâmicas e multifacetadas. Esses problemas ilustram a complexidade do NLP e a necessidade contínua de aprimorar técnicas e modelos para lidar com as diversas nuances da linguagem humana.

3.1 Análise de Sentimentos

Análise de Sentimentos, conforme exemplificado na Figura 5, é uma técnica avançada de NLP que tem como objetivo identificar e extrair sentimentos, opiniões e emoções expressas em um texto. O objetivo principal é avaliar a atitude humana em relação a um tema específico, classificando o sentimento como positivo, negativo ou neutro. Este tipo de análise é amplamente utilizado em diversas áreas, como *marketing*, para entender a satisfação dos clientes com produtos e campanhas; atendimento ao cliente, para melhorar a qualidade do suporte oferecido; política, para monitorar a opinião pública sobre políticas e candidatos; e saúde, para rastrear o bem-estar emocional das pessoas e identificar necessidades de apoio psicológico. Por meio da análise de sentimentos, é possível obter *insights* valiosos sobre as percepções e emoções das pessoas, auxiliando na tomada de decisões estratégicas.

Figura 5 - Exemplo de análise de sentimentos



Fonte: autoria própria.

Uma excelente aplicação da análise de sentimentos é em uma plataforma de comércio eletrônico que vende uma ampla variedade de produtos, desde eletrônicos até vestuário. Para melhorar a experiência do cliente e aumentar as vendas, a empresa pode utilizar os *insights* obtidos da análise de sentimentos aplicada aos comentários e avaliações dos clientes. Com essas informações, é possível:

- **Melhorar processos:** se muitos comentários negativos mencionam problemas de entrega, a empresa pode trabalhar para otimizar a logística e acelerar o tempo de entrega;
- **Aprimorar produtos:** comentários negativos sobre a qualidade de um produto específico podem levar a uma revisão desse produto pelo departamento de qualidade;
- **Direcionar *marketing*:** comentários positivos sobre determinados produtos podem ser usados em campanhas de *marketing* para destacar esses produtos.

Além disso, a análise de sentimentos permite obter *feedback* em tempo real, implementando um sistema de monitoramento contínuo para analisar novos comentários assim que eles são postados. Isso permite respostas rápidas a problemas emergentes, ajudando a manter a satisfação do cliente em alta. Assim, é possível obter benefícios da análise de sentimentos para as empresas tais como:

- **Melhoria da satisfação do cliente:** ao entender melhor os sentimentos dos clientes, a empresa pode tomar ações para melhorar a experiência de compra;
- **Aumento das vendas:** produtos com *feedback* positivo podem ser promovidos mais agressivamente, enquanto produtos com *feedback* negativo podem ser melhorados ou retirados do mercado;
- **Redução de custos:** identificar e resolver rapidamente problemas recorrentes podem reduzir custos associados a devoluções e reclamações;
- **Decisões estratégicas informadas:** *insights* detalhados sobre as percepções dos clientes ajudam a empresa a tomar decisões mais informadas sobre *marketing*, desenvolvimento de produtos e atendimento ao cliente.

3.2 Reconhecimento de Entidades Nomeadas (NER)

O NER é uma técnica de NLP usada para identificar e classificar entidades específicas mencionadas em um texto. Essas entidades incluem nomes de pessoas, organizações, locais, datas, valores monetários e outras categorias relevantes. O objetivo do NER é transformar dados textuais não estruturados em informações estruturadas, facilitando a análise e a extração de conhecimento.

Uma aplicação dessa técnica é na análise de notícias, em que é possível extrair nomes de pessoas, organizações e locais de artigos, ajudando a entender eventos e suas relações. No atendimento ao cliente, o NER pode identificar produtos e datas em reclamações, automatizando a categorização e priorização de problemas. Na pesquisa acadêmica, a técnica é utilizada para extrair nomes de autores, instituições e publicações de textos

científicos, facilitando a indexação e busca de informações. Por fim, em aplicações financeiras, o NER permite identificar valores monetários e datas em relatórios, permitindo uma análise automatizada e eficiente.

Um exemplo de aplicação de NER é em uma companhia de seguros que oferece diversos produtos, como seguros de vida, saúde, automóveis e propriedades. O NER pode ser usado com o objetivo de melhorar a eficiência no processamento de sinistros e aprimorar a qualidade do atendimento ao cliente utilizando percepções obtidas em documentos de sinistros e interações com clientes. Nesse sentido, é possível realizar análise e processamento de documentos, conforme a seguir:

- **Automatização de sinistros:** extrair e categorizar automaticamente as informações relevantes dos documentos de sinistros, como valores, datas, locais e nomes, para acelerar o processamento e aprovação dos sinistros;
- **Priorização de atendimento:** identificar e priorizar automaticamente os casos de atendimento ao cliente com base na identificação de entidades críticas, como valores de apólices e datas de vencimento.

Também é possível realizar tomada de ações de acordo com os *insights* gerados por meio do uso de NER e também processos de melhoria contínua. Isso pode, por exemplo, ser realizado conforme a seguir:

- **Melhoria de processos:** utilizar as informações estruturadas para otimizar os processos internos, reduzindo o tempo de resposta e aumentando a eficiência operacional;
- **Aprimoramento do atendimento ao cliente:** fornecer aos agentes de atendimento informações detalhadas e precisas sobre as interações anteriores com os clientes, melhorando a personalização e a qualidade do atendimento;
- **Detecção de fraudes:** analisar padrões e inconsistências nos dados de sinistros para identificar possíveis fraudes, aumentando a segurança e a confiabilidade dos serviços oferecidos;
- ***Feedback* em tempo real:** implementar um sistema de monitoramento contínuo para analisar novos sinistros e interações de clientes em tempo real, permitindo respostas rápidas e eficientes a novas demandas;
- **Ajustes e melhorias:** ajustar continuamente os modelos de NER com base nos *feedbacks* recebidos para melhorar a precisão e a relevância dos *insights*.

Utilizando NER é possível obter alguns benefícios da aplicação da técnica em uma empresa de seguros. A seguir, alguns exemplos desses benefícios:

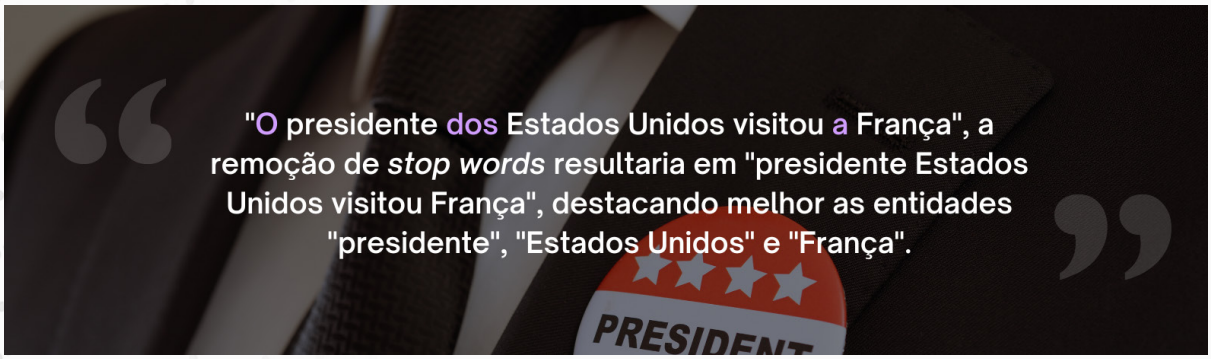
- **Melhoria da eficiência operacional:** redução significativa do tempo necessário para processar sinistros, permitindo uma resposta mais rápida aos clientes;
- **Aprimoramento da qualidade do atendimento:** acesso rápido e preciso às informações relevantes melhora a satisfação do cliente e a experiência geral;

- **Redução de fraudes:** identificação de padrões suspeitos e de inconsistências nos sinistros ajuda a prevenir fraudes, economizando recursos;
- **Decisões estratégicas informadas:** *insights* detalhados sobre as interações com clientes e sinistros ajudam a empresa a tomar decisões mais informadas sobre estratégias de negócios e melhorias de processos.

Conforme mencionado na seção 2.2, a remoção de *stop words* é uma técnica de pré-processamento de texto que pode ser bastante útil no NER. *Stop words* são palavras comuns e frequentemente usadas em um idioma, como “o” , “a”, “de”, “que” em português, que não carregam muita informação significativa por si só. Essas palavras aparecem com alta frequência em textos, mas não contribuem significativamente para o entendimento semântico específico.

Remover essas *stop words* antes de aplicar o NER ajuda a reduzir o ruído no texto. *Stop words* podem adicionar complexidade desnecessária ao processo de identificação de entidades nomeadas. Ao removê-las, o modelo de NER pode focar mais facilmente nas palavras que são mais prováveis de serem entidades nomeadas, como nomes de pessoas, organizações, locais e datas. Por exemplo, em uma frase como “O presidente dos Estados Unidos visitou a França”, a remoção de *stop words* resultaria em “presidente Estados Unidos visitou França”, destacando melhor as entidades “presidente”, “Estados Unidos” e “França” (Figura 6).

Figura 6 - Exemplo de identificação de *stop words* em uma frase



Fonte: autoria própria.

Além disso, a remoção de *stop words* melhora a eficiência de processamento. Ao eliminar palavras desnecessárias, o volume de dados a ser processado é reduzido, o que pode acelerar os algoritmos de NER. Isso resulta em menos computação e memória necessárias para analisar o texto. Menos *tokens* para processar significa que o modelo pode executar mais rapidamente e com menos recursos computacionais.

Outro benefício é o foco em palavras relevantes. As palavras que permanecem após a remoção das *stop words* são, geralmente, mais relevantes e carregam mais significado semântico. Isso facilita a identificação de padrões e contextos que ajudam na correta classificação das entidades. Por exemplo, em «João foi à escola com Maria», removendo «foi à com» deixa «João escola Maria», tornando mais claro que «João» e «Maria» são nomes de pessoas e «escola» é um possível local.

Em resumo, a remoção de *stop words* é uma técnica simples, mas poderosa no pré-processamento de texto que pode melhorar significativamente o desempenho do reconhecimento de entidades nomeadas (Figura 7). Ao focar o modelo em palavras mais significativas, reduzindo o ruído e aumentando a eficiência de processamento, é possível obter melhores resultados na identificação de entidades em textos.

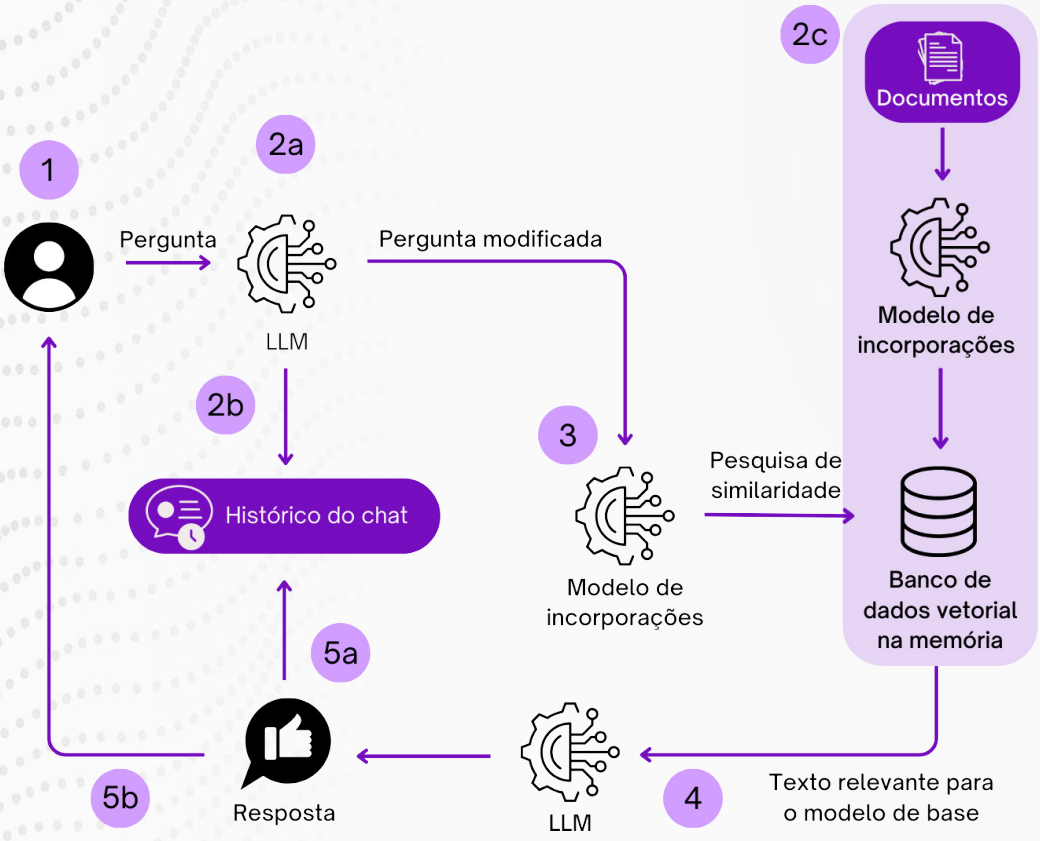
Figura 7 - Reconhecimento de entidades nomeadas (NER)

Fonte: autoria própria.

3.3 Retorno de Informações

Retorno de informações (*Information Retrieval* [IR]), conforme exemplificado na Figura 8, é o processo de obter recursos (geralmente documentos) que são relevantes para uma necessidade de informação a partir de um grande repositório de dados, como uma biblioteca digital ou a internet. A tarefa de IR é fundamental em diversos contextos, incluindo motores de busca na *web*, sistemas de recomendação e bibliotecas digitais.

Figura 8 - Retorno de informações (IR)



Fonte: autoria própria.

O objetivo principal é encontrar os documentos mais relevantes para uma consulta específica feita pelo usuário. Essa técnica é popular entre buscadores na internet, como por exemplo, Google® e Bing®, que encontram páginas da *web* relevantes com base nas consultas dos usuários, utilizando algoritmos sofisticados que consideram fatores como a frequência de palavras-chave e a autoridade dos *sites*.

Além disso, os sistemas de recomendação, utilizados na Netflix® e Amazon®, utilizam IR para recomendar filmes, livros ou produtos com base nos interesses e histórico de compras do usuário. As bibliotecas digitais, como PubMed e IEEE Xplore, permitem que pesquisadores encontrem artigos científicos relevantes por meio de consultas baseadas em palavras-chave e tópicos de interesse. Os assistentes virtuais, como Siri® e Alexa®, também utilizam IR para buscar informações e responder perguntas dos usuários.

Uma plataforma de comércio eletrônico que vende uma ampla gama de produtos, desde eletrônicos a vestuário, de livros a utensílios domésticos, pode usar IR com o objetivo de melhorar a experiência do usuário e aumentar as vendas, pois permite aos clientes encontrar rapidamente os produtos que desejam comprar. A busca realiza o ranqueamento dos produtos, pois são ordenados de acordo com sua relevância para a consulta do usuário. Produtos com alta relevância aparecem no topo da lista. A exibição dos resultados são apresentados ao usuário, frequentemente com trechos das descrições que destacam as palavras-chave encontradas e avaliações dos clientes.

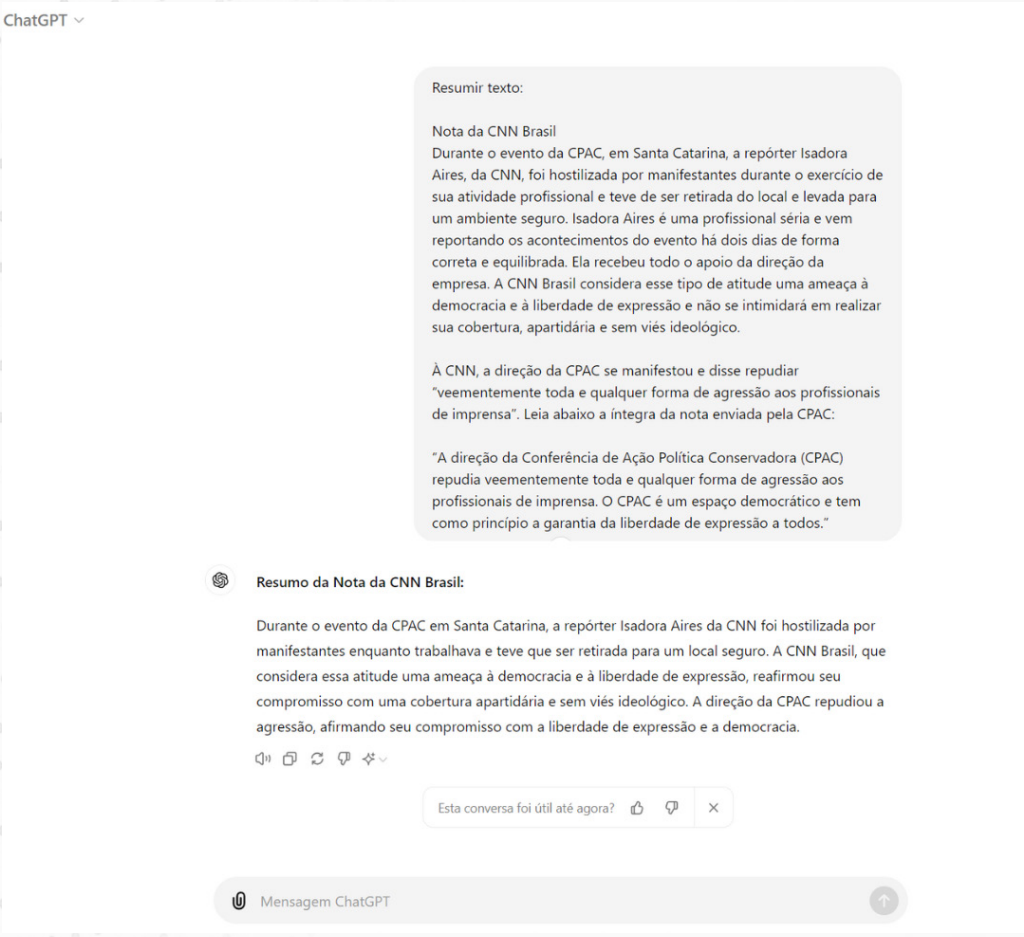
Quando um usuário realiza uma busca, o algoritmo de *stemming*, citado na seção 2.3, reduz todas as palavras na consulta e no índice de documentos às suas raízes. Isso ajuda a garantir que diferentes formas de uma palavra sejam tratadas como equivalentes. No caso da lematização, as palavras são transformadas na sua forma base na consulta e nos documentos para melhorar a precisão da busca, pois considera o contexto gramatical, fornecendo resultados mais relevantes. Ambas as técnicas ajudam a reduzir a redundância no índice de documentos, uma vez que palavras com diferentes formas são consolidadas em uma única entrada. Isso torna o índice mais compacto e eficiente.

Atualmente, com o crescimento do uso dos grandes modelos de linguagem (LLMs), IR aumentou sua importância com a técnica de Recuperação Aumentada de Geração (RAG), que é crucial para maximizar a eficácia, precisão e relevância das respostas geradas. Os LLMs são capazes de gerar texto coerente e contextualizado a partir de *prompts* de entrada. No entanto, esses modelos não têm acesso a informações em tempo real ou a dados fora do seu treinamento, que podem estar desatualizados. Nesse contexto, IR, por outro lado, é a técnica que permite recuperar dados relevantes a partir de grandes repositórios, como bancos de dados, documentos ou a *web*. IR busca, classifica e recupera informações específicas que atendem às consultas dos usuários e podem ser enviadas aos modelos para retornarem respostas mais assertivas.

3.4 Sumarização de Documentos

Sumarização de documentos, conforme exemplificado na Figura 9, é uma técnica de NLP que visa condensar um texto longo em uma versão mais curta, mantendo as informações mais importantes e essenciais. O objetivo é criar um resumo que capture as ideias principais do texto original sem perder seu significado. A sumarização é amplamente utilizada em diversas áreas, como notícias, pesquisa acadêmica e atendimento ao cliente, para facilitar a rápida compreensão de grandes volumes de texto.

Figura 9 - Exemplo de sumarização de documentos utilizando o ChatGPT®



Fonte: reportagem disponível em Rádio Itatiaia (2024).

Existem dois tipos de sumarização, descritos a seguir:

- **Sumarização extrativa:** seleciona e combina frases ou parágrafos diretamente do texto original para criar o resumo. Não gera novo texto, apenas seleciona partes existentes. Exemplo: se um artigo de notícias contém várias frases sobre um evento, a sumarização extrativa escolherá as frases mais relevantes para compor o resumo;
- **Sumarização abstrativa:** gera novo texto que captura as ideias principais do original, criando uma versão condensada com frases reformuladas e, possivelmente, novas expressões. Exemplo: em vez de, simplesmente, selecionar frases, a sumarização abstrativa reescreveria o texto de forma mais concisa, mantendo o significado principal.

Existem diversas aplicações práticas para a sumarização. Por exemplo, na área do jornalismo, pode ser aplicado para resumir notícias, para facilitar a leitura rápida e a disseminação de informações. Já na pesquisa acadêmica, pode ser usada na geração de resumos de artigos científicos para ajudar pesquisadores a identificarem rapidamente trabalhos relevantes. No atendimento ao cliente, é possível gerar resumos de longos e-mails ou transcrições de chamadas para facilitar a rápida resposta a consultas dos

clientes. E por fim, na gestão de documentos, é possível realizar análise das informações com a criação de resumos de relatórios extensos para que executivos e gerentes possam tomar decisões informadas mais rapidamente.

Uma empresa com uma plataforma de notícias online que publica uma vasta gama de artigos sobre política, economia, esportes, tecnologia e entretenimento pode usar essa técnica de NLP com o objetivo de melhorar a experiência do usuário e aumentar o engajamento no *site*, realizando resumos automáticos de artigos de notícias. Isso ajuda leitores a obterem rapidamente uma visão geral das notícias mais recentes, incentivando-os a lerem os artigos completos.

Os benefícios da sumarização de documentos para empresa podem ser:

- **Melhoria da experiência do usuário:** leitores podem obter rapidamente uma visão geral das notícias mais recentes, economizando tempo e facilitando a leitura;
- **Aumento do engajamento:** resumos atraentes incentivam leitores a clicarem nos artigos completos para obterem mais detalhes, aumentando o tempo de permanência no *site*;
- **Eficiência operacional:** automatização do processo de criação de resumos permite que editores se concentrem em tarefas mais estratégicas, como a curadoria de conteúdo e a análise de tendências;
- **Personalização:** oferecer resumos personalizados com base nos interesses dos leitores aumenta a relevância e a satisfação do usuário, potencialmente aumentando a fidelidade e a retenção de leitores.

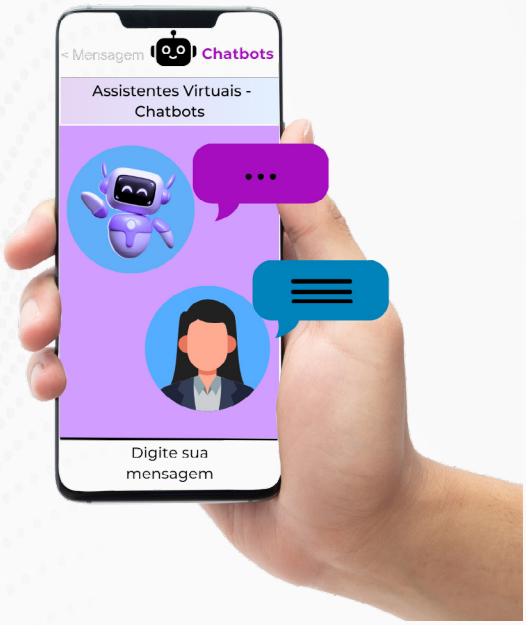
A implementação da técnica de sumarização de documentos em uma plataforma de notícias online pode transformar a forma como os leitores interagem com o conteúdo, proporcionando uma experiência mais eficiente e personalizada. Ao oferecer resumos automáticos de alta qualidade, é possível melhorar significativamente o engajamento do usuário, aumentar o tempo de permanência no *site* e, consequentemente, impulsionar suas métricas de desempenho.

3.5 Assistentes Virtuais - *Chatbots*

Assistentes virtuais - *chatbots*, conforme exemplificado na Figura 10, são sistemas de *software* projetados para interagir com seres humanos por meio de conversas naturais, geralmente, por meio de texto ou voz. Esses sistemas utilizam técnicas de NLP para entender e responder às consultas dos usuários, oferecendo uma ampla gama de funcionalidades que vão desde responder perguntas simples até realizar tarefas complexas, como agendamento de compromissos ou suporte ao cliente. Esses sistemas, podem ser usados no suporte ao cliente, para responder a perguntas frequentes, solucionar problemas técnicos e guiar os clientes por meio de processos de atendimento. Isso permite uma redução de tempo de espera e melhoria na eficiência do atendimento. Outro exemplo, são os assistentes pessoais, que ajudam os usuários a realizar tarefas cotidianas, como definir

lembretes, tocar música e controlar dispositivos domésticos inteligentes, fazendo com que haja um aumento da conveniência e automação de tarefas diárias. As plataformas educacionais podem, também, utilizar os *chatbots* para fornecer assistência a estudantes, responder perguntas sobre materiais de estudo e oferecer *feedback* em tempo real, oferecendo acesso imediato a suporte educacional e personalização da aprendizagem.

Figura 10 - Assistentes virtuais - *chatbots*



Fonte: autoria própria.

Os *chatbots* utilizam tecnologias com a capacidade de transformar a maneira como os seres humanos interagem com sistemas de *software* e serviços. Utilizando técnicas avançadas de NLP, esses sistemas são capazes de entender e responder às consultas dos usuários, de maneira natural e eficaz. Com aplicações que vão desde suporte ao cliente até assistência pessoal e educação, os *chatbots* têm o potencial de melhorar significativamente a eficiência e a qualidade do atendimento, proporcionando experiências mais personalizadas e convenientes para os usuários.

Uma empresa de telecomunicações que oferece serviços de telefonia móvel, internet e televisão pode usar este tipo de tecnologia para melhorar a eficiência do atendimento ao cliente e aumentar a satisfação dos usuários. Utilizando um *chatbot* como assistente virtual, é possível resolver problemas comuns, responder perguntas frequentes e processar solicitações de serviço.

Considerando esse cenário de uso, seguem alguns benefícios da implementação de um *chatbot*:

- **Redução de custo:** diminuição do volume de chamadas e e-mails para o suporte ao cliente, reduzindo a necessidade de uma grande equipe de atendimento;
- **Atendimento 24/7:** disponibilidade constante para responder a consultas e resolver problemas, independentemente do horário;

- **Respostas imediatas:** o *chatbot* oferece respostas rápidas e precisas, reduzindo o tempo de espera e melhorando a experiência do cliente;
- **Interação personalizada:** utilização de dados históricos e preferências do cliente para personalizar as interações e fornecer recomendações relevantes;
- **Facilidade de uso:** interface amigável e intuitiva que facilita a navegação e interação dos clientes com os serviços da empresa;
- **Promoções e ofertas:** o *chatbot* pode informar os clientes sobre promoções e ofertas especiais, incentivando *upgrades* e novas assinaturas;
- **Análise de interações:** coleta de dados sobre as interações dos clientes para identificar tendências e áreas de melhoria;
- **Feedback em tempo real:** capacidade de coletar *feedback* imediato dos clientes após cada interação, permitindo ajustes rápidos e contínuos.

A técnica de tokenização, citada na seção 2.1, é um processo fundamental no funcionamento de assistentes virtuais e *chatbots*, especialmente no contexto de modelos de linguagem. Quando um usuário interage com um *chatbot*, ele insere frases ou perguntas que precisam ser compreendidas pelo sistema para fornecer respostas adequadas. A tokenização desempenha um papel crucial nesse processo ao dividir o texto de entrada em unidades menores e manejáveis, chamadas *tokens*, que geralmente são palavras, frases ou até caracteres.

Ao dividir o texto em *tokens*, o *chatbot* consegue analisar cada parte individualmente, facilitando a identificação de palavras-chave e a compreensão da estrutura da frase. Por exemplo, se um usuário pergunta “Qual é a previsão do tempo para amanhã?”, a tokenização pode separar essa frase em: [“Qual”, “é”, “a”, “previsão”, “do”, “tempo”, “para”, “amanhã”]. Esse detalhamento permite que o modelo de linguagem entenda melhor o contexto e a intenção por trás da consulta realizada pelo usuário.

Os modelos de linguagem, como GPT-3® e BERT®, dependem fortemente da tokenização para processar e gerar texto. Esses modelos são treinados em grandes volumes de dados textuais, onde a tokenização é utilizada para transformar o texto em sequências de *tokens* que podem ser analisadas e compreendidas pelo algoritmo. Ao aplicar tokenização ao texto de entrada, os modelos de linguagem podem identificar padrões, relações entre palavras e contextos, permitindo que eles compreendam e respondam de maneira mais precisa e natural.

No caso dos *chatbots*, após a tokenização do texto de entrada, o modelo de linguagem analisa os *tokens* e gera uma resposta apropriada com base em seu treinamento. A resposta gerada também passa pelo processo inverso, no qual os *tokens* são combinados para formar uma frase coerente e compreensível para o usuário. Esse ciclo de tokenização e processamento é contínuo, permitindo que o *chatbot* mantenha conversas fluidas e relevantes.

Em suma, a tokenização facilita a compreensão e a geração de texto pelos modelos de linguagem em assistentes virtuais e *chatbots*, dividindo o texto em partes menores que são mais fáceis de processar e analisar. Isso permite que o sistema entenda melhor as consultas dos usuários e forneça respostas mais precisas e contextualmente adequadas.

Saiba mais...

- Introdução à análise de sentimentos (Rebouças, 2023).
- Reconhecimento de Entidades Nomeadas: entidades, subentidades, relacionamentos e ambiguidade (Albuquerque, 2022).
- Introdução à Recuperação de Informações (Zerbinati, 2019)
- Utilizando processamento de linguagem natural para criar um sumarização automática de textos (Lima, 2017)
- Chatbots e Assistentes Virtuais Inteligentes: por onde começar? (Rocha, 2017)

AKCIT
CENTRO DE COMPETÊNCIA EMBRAPII EM TECNOLOGIAS IMERSIVAS

Resolução de Problemas com Processamento de Linguagem Natural

Unidade IV

Desafios Éticos em Processamento de Linguagem Natural

Unidade IV - Desafios Éticos em Processamento de Linguagem Natural

Nesta Unidade, diversos desafios éticos associados à NLP serão abordados. Discutiremos as dificuldades relacionadas à privacidade dos dados, viés algorítmico e a potencial disseminação de desinformação. Também exploraremos as implicações do uso inadequado da tecnologia, incluindo a manipulação de informações, violação de direitos de privacidade e a perpetuação de preconceitos sociais. A compreensão dessas questões é fundamental para garantir que as aplicações de NLP sejam desenvolvidas e implementadas de maneira responsável e ética, minimizando os riscos e maximizando os benefícios para a sociedade.

4.1 Principais Desafios em NLP

NLP é uma área que traz inúmeros benefícios em diversas aplicações, mas também enfrenta desafios éticos significativos. Esses desafios abrangem questões de privacidade, viés, manipulação de informações, segurança e responsabilidade. A seguir, são descritos, em detalhes, os principais desafios éticos no NLP.

4.1.1 Privacidade e Segurança de Dados

NLP frequentemente envolve a coleta e a análise de grandes volumes de dados textuais que podem conter informações pessoais e sensíveis. As aplicações de assistentes virtuais que coletam dados de conversas podem inadvertidamente armazenar informações sensíveis como, por exemplo, números de cartões de crédito ou detalhes sobre a saúde das pessoas, assim como as plataformas de mídias sociais analisam mensagens e postagens para detectar tendências ou opiniões.

Sendo assim, é importante manter a privacidade e a segurança dos dados e, para isso, existem desafios que precisam ser observados conforme a seguir:

- **Anonimização:** garantir que os dados sejam anonimizados corretamente para proteger a privacidade dos indivíduos;
- **Consentimento:** obter consentimento explícito dos usuários para a coleta e utilização de seus dados;
- **Armazenamento seguro:** implementar práticas de armazenamento seguro para prevenir vazamentos de dados.

4.1.2 Viés e Discriminação

Os modelos de NLP podem incorporar e amplificar vieses presentes nos dados de treinamento, resultando em discriminação contra certos grupos de pessoas. Um modelo de NLP treinado predominantemente em textos de uma determinada região ou grupo social pode não funcionar bem para outros grupos, perpetuando estereótipos. Assim como, ferramentas de recrutamento automatizadas que usam NLP podem discriminar candidatos com base em gênero ou raça, se os dados históricos de contratação forem enviesados. Nesse sentido, seguem dois desafios que precisam ser observados:

- **Detecção e mitigação de viés:** desenvolver técnicas para identificar e corrigir vieses nos dados e nos modelos;
- **Diversidade nos dados de treinamento:** garantir que os dados de treinamento sejam representativos de diferentes grupos demográficos e contextos culturais.

4.1.3 Manipulação e Desinformação

Tecnologias de NLP podem ser usadas para criar e disseminar informações falsas ou enganosas. Modelos de linguagem como GPT-3® podem ser usados para gerar notícias falsas ou manipular opiniões públicas. Os *bots* em redes sociais podem usar NLP para espalhar propaganda ou manipular discussões online. Por isso, é importante identificar e tratar os desafios, conforme listados a seguir:

- **Detecção de *fake news*:** desenvolver sistemas robustos para detectar e mitigar a disseminação de *fake news* e desinformação;
- **Responsabilidade:** as empresas e desenvolvedores devem ser responsabilizados pelo uso de suas tecnologias para a disseminação de desinformação.

4.1.4 Transparência e Explicabilidade

Muitos modelos de NLP, especialmente aqueles baseados em *deep learning*, são complexos e difíceis de interpretar. Nesse sentido, os sistemas de recomendação que sugerem conteúdo ou produtos sem explicar claramente como a recomendação foi feita, representa um desafio ético do uso da tecnologia. Assim como os assistentes virtuais que fornecem respostas ou conselhos sem que os usuários entendam a base dessas respostas. Dessa maneira, é de suma importância lidar com os itens a seguir:

- **Explicabilidade:** desenvolver métodos para tornar os modelos de NLP mais transparentes e suas decisões mais compreensíveis para os usuários;
- **Justificativas para decisões:** fornecer justificativas claras e compreensíveis para as decisões tomadas pelos modelos de NLP.

4.1.5 Uso Ético e Responsável

Em aplicações de NLP, é importante garantir que elas sejam usadas de maneira ética e responsável, considerando o impacto social e moral de suas implementações. A implementação de *chatbots* em serviços de saúde precisam garantir que as informações fornecidas sejam precisas e não causem danos aos usuários. As ferramentas de monitoramento de funcionários que usam NLP não podem invadir a privacidade ou criar um ambiente de trabalho opressivo. Por isso, é importante garantir que os itens a seguir sejam de fato observados, pois são desafios que não devem ser negligenciados:

- **Governança e regulamentação:** estabelecer diretrizes e regulamentações claras para o uso ético de tecnologias de NLP;
- **Educação e conscientização:** promover a conscientização sobre os potenciais impactos éticos e sociais do NLP entre desenvolvedores e usuários.

4.2 Análise dos Desafios Éticos e Sociais Associados ao Desenvolvimento e à Aplicação de Tecnologias de NLP

O desenvolvimento e a aplicação de tecnologias de NLP têm avançado significativamente, no entanto, esses avanços também trazem à tona uma série de desafios éticos e sociais que precisam ser cuidadosamente considerados e gerenciados. Portanto, é importante tratar des mitigar todos os riscos citados no item 4.1, conforme a seguir:

4.2.1 Privacidade e Segurança de Dados

Um dos principais desafios éticos em NLP é a privacidade e a segurança dos dados. Modelos de NLP frequentemente requerem grandes quantidades de dados textuais para treinamento, que, muitas vezes, contêm informações sensíveis ou pessoais. A coleta, armazenamento e uso desses dados levantam preocupações significativas sobre a privacidade dos indivíduos. Além disso, a segurança dos dados deve ser garantida para evitar vazamentos ou acessos não autorizados que possam comprometer a privacidade dos usuários.

Para mitigar esses riscos, é essencial implementar práticas robustas de anonimização de dados, garantir o consentimento informado dos usuários e adotar políticas de gerenciamento de dados que priorizem a proteção da privacidade. Além disso, a transparência sobre como os dados são coletados, armazenados e utilizados é crucial para construir e manter a confiança dos usuários.

4.2.2 Viés e Discriminação

Os modelos de NLP são treinados em grandes corporações de texto que refletem o uso real da linguagem, incluindo seus preconceitos e estereótipos. Como resultado, esses modelos podem aprender e perpetuar vieses existentes, levando a decisões

discriminatórias. Por exemplo, um modelo de NLP pode gerar respostas enviesadas contra determinados grupos demográficos se os dados de treinamento contiverem tais preconceitos.

O viés e a discriminação em NLP não são apenas questões técnicas, mas também éticas e sociais, pois podem reforçar e amplificar desigualdades existentes na sociedade. Para enfrentar esses desafios, é importante adotar práticas de desenvolvimento ético, como a auditoria contínua dos modelos para vieses, a diversificação dos dados de treinamento e a implementação de técnicas para mitigar vieses, como o balanceamento de dados e a inclusão de critérios de equidade nos algoritmos.

4.2.3 Manipulação e Desinformação

As tecnologias de NLP podem ser usadas para criar e disseminar desinformação de maneira rápida e eficiente. Exemplos incluem a geração automática de notícias falsas ou a manipulação de informações em redes sociais para influenciar opiniões públicas ou eleições. A capacidade de criar textos que parecem autênticos, mas que contêm informações enganosas, representa um desafio significativo para a integridade da informação na era digital.

Combater a manipulação e a desinformação requer uma abordagem multifacetada, incluindo a educação dos usuários para reconhecer conteúdos suspeitos, a implementação de sistemas de verificação de fatos automatizados e a colaboração entre plataformas de tecnologia e reguladores para desenvolver e aplicar políticas contra a disseminação de desinformação.

4.2.4 Transparência e Explicabilidade

Outro desafio crítico é a transparência e a explicabilidade dos modelos de NLP. Muitos desses modelos, especialmente aqueles baseados em *deep learning*, são frequentemente descritos como “caixas-pretas” devido à sua complexidade e à dificuldade de interpretar como eles chegam a determinadas conclusões. A falta de transparência pode dificultar a identificação de vieses, erros e outros problemas éticos.

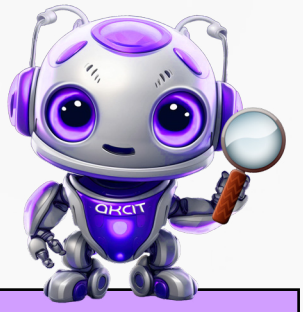
Para abordar esse problema, é essencial desenvolver técnicas que melhorem a interpretabilidade dos modelos de NLP, como métodos de explicação de decisões e visualizações que ajudem a entender o comportamento dos modelos. Além disso, os desenvolvedores devem fornecer documentação clara e detalhada sobre os dados e métodos utilizados no treinamento dos modelos.

4.2.5 Uso Ético e Responsável

Finalmente, o uso ético e responsável das tecnologias de NLP envolve garantir que essas ferramentas sejam desenvolvidas e aplicadas de maneira a beneficiar a sociedade como

um todo, minimizando danos potenciais. Isso inclui considerar os impactos sociais e éticos durante todas as fases do desenvolvimento, desde a concepção até a implementação e uso final.

Os desenvolvedores e pesquisadores têm a responsabilidade de considerar os potenciais efeitos negativos de suas tecnologias e de tomar medidas proativas para evitar abusos. Isso pode incluir a implementação de princípios éticos em códigos de conduta, a realização de avaliações de impacto ético e a promoção de uma cultura de responsabilidade e sensibilidade ética em toda a comunidade de NLP.



Saiba mais...

- A importância da ética no desenvolvimento do processamento de linguagem natural: construindo um futuro responsável (Santos, 2024).
- Explorando a interseção entre processamento de linguagem natural e ética: o caminho para uma IA responsável (Santos, 2023).

AKCIT
CENTRO DE COMPETÊNCIA EMBRAPII EM TECNOLOGIAS IMERSIVAS

Resolução de Problemas com Processamento de Linguagem Natural

Unidade V
Problemas em Aberto e Resumo

Unidade V - Problemas em Aberto e Resumo

Nesta Unidade, serão citados os desafios e os problemas em aberto para os principais temas que foram abordados ao longo do *ebook*. Contudo, é importante apenas recapitular algumas técnicas e os problemas abordados. Exploramos as ***word embeddings***, que transformam palavras em vetores numéricos para capturar seus significados e relações contextuais. Em seguida, exploramos as **RNNs** e suas variantes avançadas, como **LSTM**, que são projetadas para processar sequências de dados e capturar dependências de longo prazo. A introdução dos **transformers** revolucionou o NLP, permitindo o processamento paralelo de sequências e melhorando significativamente a eficiência e a precisão dos modelos de linguagem. Discutimos também as técnicas de ***stemming*** **e lematização**, que simplificam palavras às suas formas básicas, facilitando a análise textual. A **análise de sentimentos** é uma aplicação-chave que identifica e classifica as emoções expressas em textos, enquanto a **sumarização de texto** visa condensar informações extensas em resumos concisos e informativos. Os **modelos de linguagem**, incluindo avanços recentes como GPT-3®, desempenham um papel crucial na geração e compreensão de texto. O **NER** permite a identificação de entidades específicas, como nomes e locais, dentro dos textos. Finalmente, os **assistentes virtuais e *chatbots*** combinam várias dessas técnicas para interagir de maneira eficaz e natural com os usuários, automatizando e melhorando a qualidade do atendimento e da assistência digital.

A seguir, serão elencados alguns problemas e desafios para as tarefas: **modelos de linguagem, análise de sentimentos, NER**, **assistentes virtuais e *chatbots***.

5.1 Modelos de Linguagem

- **Compreensão do contexto e coerência** - Modelos de linguagem, como GPT-3®, têm demonstrado uma capacidade impressionante de gerar texto coerente e fluente. No entanto, eles ainda enfrentam desafios em manter a coerência ao longo de textos longos e em entender contextos complexos. Um problema em aberto é a dificuldade dos modelos em lembrar e utilizar informações mencionadas anteriormente em uma conversa ou texto extenso, o que pode resultar em respostas incoerentes ou repetitivas.
- **Viés e discriminação** - Os modelos de linguagem são treinados em grandes corpora de texto que refletem os vieses presentes na sociedade. Como resultado, esses modelos podem perpetuar ou até amplificar esses vieses, levando a respostas discriminatórias ou preconceituosas. A identificação e mitigação de vieses em modelos de linguagem é um problema crítico em aberto. É necessário desenvolver técnicas que não apenas detectem vieses, mas também ajustem os modelos para redução desses efeitos.

- **Explicabilidade e transparência** - A maioria dos modelos de linguagem modernos, especialmente aqueles baseados em *deep learning*, é considerada "caixas-pretas" devido à sua complexidade. Isso levanta questões sobre a transparência e a confiabilidade das respostas geradas. Há uma necessidade crescente de desenvolver métodos que tornem esses modelos mais explicáveis, permitindo que os usuários entendam como e por quê certas respostas são geradas.
- **Generalização para novos domínios** - Modelos de linguagem geralmente são treinados em dados genéricos da *web* e podem não se adaptar bem a domínios específicos, como medicina ou finanças. Adaptar esses modelos para funcionar eficientemente em novos domínios sem a necessidade de grandes quantidades de dados anotados é um desafio contínuo. Técnicas de aprendizado transferível e *fine-tuning*, que é uma categoria de treino para ajustar o modelo para uma tarefa específica, são áreas de pesquisa ativas para abordar esse problema.

5.2 Análise de Sentimentos

- **Detecção de ironia e sarcasmo** - A análise de sentimentos enfrenta grandes desafios ao tentar detectar ironia e sarcasmo. Essas nuances linguísticas podem inverter o significado aparente de uma frase, tornando a tarefa de classificação de sentimentos particularmente difícil. Modelos de NLP precisam de mais avanços para entenderem melhor o contexto e identificarem essas figuras de linguagem.
- **Ambiguidade linguística** - Palavras e frases, muitas vezes, têm múltiplos significados dependendo do contexto em que são usadas. Essa ambiguidade pode levar a erros na análise de sentimentos. Desenvolver modelos que possam desambiguar eficazmente o significado das palavras em diferentes contextos é um problema em aberto que requer soluções inovadoras.
- **Contexto cultural e regional** - Sentimentos podem variar significativamente com base em contextos culturais e regionais. Um modelo treinado em dados de uma região pode não funcionar bem em outra devido a diferenças na expressão de emoções e opiniões. Abordar essa variabilidade e criar modelos que sejam culturalmente adaptáveis é um desafio significativo na análise de sentimentos.
- **Análise de sentimentos em múltiplos idiomas** - A maioria das pesquisas em análise de sentimentos tem se concentrado em idiomas específicos, principalmente em inglês. Há uma necessidade crescente de modelos que funcionem eficientemente em múltiplos idiomas e que possam lidar com textos multilíngues de maneira eficaz. Isso inclui a necessidade de grandes corpora, anotados em diferentes idiomas para treinamento.

5.3 Reconhecimento de Entidades Nomeadas (NER)

- **Ambiguidade de entidades** - Uma das principais dificuldades do NER é a ambiguidade das entidades. O mesmo termo pode se referir a diferentes entidades dependendo do contexto. Por exemplo, "Apple" pode se referir a uma fruta ou à empresa de tecnologia. Resolver essa ambiguidade requer modelos que possam entender e desambiguar entidades com base no contexto.
- **Entidades raras ou novas** - Muitos modelos de NER têm dificuldade em reconhecer entidades raras ou novas que não estavam presentes nos dados de treinamento. Desenvolver modelos que possam identificar e aprender sobre novas entidades de maneira dinâmica é um problema em aberto. Isso é especialmente importante em domínios como notícias, onde novas entidades aparecem frequentemente.
- **Multilíngue e *cross-domain* NER** - A maioria dos sistemas de NER é treinada para funcionar em um único idioma e domínio. No entanto, há uma necessidade crescente de sistemas que possam funcionar eficientemente em múltiplos idiomas e em diferentes domínios, como medicina, direito e finanças. Criar modelos robustos que possam ser facilmente adaptados a novos idiomas e domínios continua sendo um desafio significativo.
- **Robustez contra erros de ortografia e sintaxe** - Textos do mundo real frequentemente contêm erros de ortografia e gramática, o que pode afetar a precisão dos modelos de NER. Desenvolver sistemas que sejam robustos contra esses erros e que possam identificar corretamente as entidades nomeadas mesmo em presença de texto ruidoso é um problema importante.

5.4 Assistentes Virtuais e *Chatbots*

- **Manutenção do contexto em conversas prolongadas** - Assistentes virtuais e *chatbots* ainda enfrentam dificuldades em manter o contexto ao longo de conversas prolongadas. Muitas vezes, eles não conseguem "lembrar" as informações mencionadas anteriormente ou perdem o fio da conversa, o que leva a respostas incoerentes ou irrelevantes. Melhorar a capacidade de manutenção de contexto é crucial para a eficácia desses sistemas.
- **Personalização e adaptabilidade** - Para proporcionar uma experiência mais satisfatória, assistentes virtuais precisam ser capazes de personalizar as respostas com base no histórico e nas preferências do usuário. Desenvolver algoritmos que possam aprender e se adaptar continuamente ao comportamento e às necessidades dos usuários é um desafio em aberto significativo.
- **Compreensão e resolução de ambiguidade** - *Chatbots*, muitas vezes, lutam para compreender e resolver ambiguidades nas consultas dos usuários. Isso inclui entender intenções vagas ou múltiplas, interpretar perguntas mal formuladas e lidar com respostas ambíguas. Criar sistemas que possam lidar

eficazmente com essas ambiguidades é essencial para melhorar a interação e a satisfação do usuário.

- **Garantia de privacidade e segurança** - Como os assistentes virtuais e *chatbots* frequentemente lidam com informações pessoais sensíveis, garantir a privacidade e a segurança dos dados dos usuários é de extrema importância. Desenvolver métodos robustos para proteger esses dados e garantir que as interações sejam seguras é um problema contínuo.
- **Explicabilidade e transparência** - Para ganhar a confiança dos usuários, os assistentes virtuais precisam ser transparentes e explicáveis em suas respostas e ações. Os usuários devem ser capazes de entender por qual razão o assistente virtual está fornecendo uma determinada resposta ou recomendação. Melhorar a explicabilidade dos sistemas de IA é um desafio significativo que precisa ser abordado.

Os problemas em aberto no NLP são variados e complexos, refletindo a natureza multifacetada da linguagem humana e a diversidade de aplicações práticas. Resolver esses desafios requer uma abordagem multidisciplinar, combinando avanços em algoritmos, aprendizado de máquina, ética e regulamentação. À medida que a pesquisa e o desenvolvimento continuam, espera-se que esses problemas sejam gradualmente superados, levando a sistemas de NLP mais robustos, precisos e éticos.

Saiba mais...

- *Large language models: a survey* (Minaee *et al.*, 2024).
- *A survey of sentiment analysis: approaches, datasets, and future research* (Tan *et al.*, 2023).
- *Comprehensive overview of named entity recognition: models, domain-specific applications and challenges* (Pakhale, 2023).
- *Conversational AI: a survey* (Bhagia *et al.*, 2024).

AKCIT
CENTRO DE COMPETÊNCIA EMBRAPII EM TECNOLOGIAS IMERSIVAS

Resolução de Problemas com Processamento de Linguagem Natural

Unidade VI
Encerramento

Unidade VI - Encerramento

Nesta Unidade de Encerramento, revisamos os conceitos, técnicas e desafios sobre **Resolução de problemas com NLP,** abordados ao longo desse *ebook*. O objetivo é consolidar o conhecimento adquirido e refletir sobre a importância e o impacto dessas tecnologias em diversos contextos.

6.1 Revisão dos Conceitos Fundamentais

O *ebook* começou com uma introdução abrangente ao NLP, destacando sua origem, evolução e os principais marcos históricos. Abordamos a definição de linguagem natural como qualquer idioma que surge espontaneamente entre os seres humanos e exploramos como o NLP se baseia em fundamentos da linguística e da IA para permitir que computadores compreendam e processem a linguagem humana.

6.2 Principais Técnicas de NLP

Discutimos diversas técnicas essenciais no NLP, incluindo:

- **Tokenização:** divisão de texto em unidades menores, facilitando a análise subsequente;
- **Remoção de *stop words*:** eliminação de palavras comuns que não agregam significado ao conteúdo;
- ***Stemming* e lematização:** redução de palavras às suas formas básicas ou raízes;
- ***Parsing* e análise sintática:** compreensão da estrutura gramatical das frases;
- ***Word embeddings*:** transformação de palavras em vetores numéricos que capturam significados contextuais;
- **Modelagem de linguagem:** criação de modelos que entendem e geram texto coerente em linguagem natural.

6.3 Aplicações Práticas

Exploramos como essas técnicas são aplicadas em contextos reais, tais como:

- **Análise de sentimentos:** identificação e extração de emoções e opiniões em textos, auxiliando no *marketing*, atendimento ao cliente e análise política;
- **NER:** identificação e classificação de nomes de pessoas, organizações e locais em textos, utilizado em análise de notícias, atendimento ao cliente e pesquisa acadêmica;
- **IR:** busca e recuperação de dados relevantes a partir de repositórios, essencial em motores de busca e sistemas de recomendação;
- **Sumarização de documentos:** criação de versões condensadas de textos longos, informações essenciais, aplicada em jornalismo, pesquisa acadêmica e atendimento ao cliente;
- **Assistentes virtuais - *chatbots*:** interação com usuários por meio de conversas naturais oferecendo suporte ao cliente, assistência pessoal e educação.

6.4 Desafios Éticos

Abordamos, também, os desafios éticos no desenvolvimento e na aplicação de tecnologias de NLP, incluindo:

- **Privacidade e segurança de dados:** proteção de informações pessoais e sensíveis dos usuários;
- **Viés e discriminação:** evitar a perpetuação de preconceitos sociais nos modelos de linguagem;
- **Manipulação e desinformação:** prevenção da criação e da disseminação de informações falsas;
- **Transparência e explicabilidade:** garantir que os modelos de NLP sejam claros e compreensíveis;
- **Uso ético e responsável:** promover a aplicação das tecnologias de NLP de maneira que não causem danos sociais ou individuais.

6.5 Conclusão

O NLP é uma área dinâmica e crucial da IA, com aplicações que transformam a maneira como interagimos com a tecnologia e a informação. As técnicas e desafios abordados nesse *ebook* ilustram tanto o potencial quanto a responsabilidade envolvidos no desenvolvimento de sistemas de NLP.

Esperamos que este *ebook* tenha proporcionado uma compreensão profunda e prática das principais técnicas de NLP e dos desafios associados. O conhecimento adquirido poderá ser aplicado em diversos projetos e contextos, contribuindo para avanços significativos na análise e na geração de linguagem natural.

6.6 Reflexão Final

Encerramos com uma reflexão sobre o futuro do NLP. À medida que as tecnologias continuam a evoluir, novas oportunidades e desafios surgirão. É essencial continuar a explorar, inovar e aplicar os princípios éticos discutidos para garantir que o NLP beneficie a sociedade de maneira justa e equitativa. A integração contínua de linguística, IA e ética será fundamental para o sucesso e a sustentabilidade das tecnologias de NLP.

Referências

ALBUQUERQUE, B.. Reconhecimento de Entidades Nomeadas: entidades, subentidades, relacionamentos e ambiguidade [Internet]. **Medium.com.** 2022. Disponível em: https://medium.com/data-hackers/reconhecimento-de-entidades-nomeadas-entidades-subentidades-relacionamentos-e-ambiguidade-bc4f302d0f9b. Acesso em: 16 jul. 2024.

ALVES, E. Um breve estudo dos Transformers [Internet]. **Medium.com.** 2022. Disponível em: https://erika-gl-alves.medium.com/um-breve-estudo-dos-transformers-6abbf1b77512. Acesso em: 16 jul. 2024.

BHAGIA, Y.; ABBAS, S. M.; KUMAR, S.; MAHESHWARI, S.. Conversational AI: a survey. **International Research Journal of Engineering and Technology (IRJET)**. 2024, v. 11, n. 4, p. 833-842. Disponível em: https://www.irjet.net/archives/V11/i4/IRJET-V11I4139.pdf. Acesso em: 16 jul. 2024.

BENGIO, Y.; DUCHARME, R.; VINCENT, P.; JAUVIN, C.. A neural probabilistic language model. **Journal of Machine Learning Research**. 2003, v. 3, n. 2003, p. 1137-1155. Disponível em: https://www.jmlr.org/papers/volume3/bengio03a/bengio03a.pdf. Acesso em: 16 jul. 2024.

DUBEY, P.. An introduction to Bag of Words and how to code it in Python for NLP [Internet]. **freeCodeCamp.org.**: Machine Learning. 2018. Disponível em: https://www.freecodecamp.org/news/an-introduction-to-bag-of-words-and-how-to-code-it-in-python-for-nlp-282e87a9da04/. Acesso em: 16 jul. 2024.

FONSECA, C.. Word Embedding: fazendo o computador entender o significado das palavras - Uma introdução a conceitos muito importantes em NLP: embeddings e word2vec [Internet]. **Turing Talks**. 2021. Disponível em: https://medium.com/turing-talks/word-embedding-fazendo-o-computador-entender-o-significado-das-palavras-92fe22745057. Acesso em: 16 jul. 2024.

GRAVES, A.. SCHMIDHUBER, J.. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. **Neural Networks**. 2005, v. 18, n. 5-6, p. 602-610. doi: 10.1016/j.neunet.2005.06.042. Acesso em: 16 jul. 2024.

HOCHREITER, S.; SCHMIDHUBER, J.. Long short-term memory. **Neural Computation**. 1997, v. 9, n. 8, p. 1735-1780. doi: 10.1162/neco.1997.9.8.1735. Acesso em: 16 jul. 2024.

HOUAISS, A.. **Dicionário Houaiss da Língua Portuguesa**. Rio de Janeiro: Editora Objetiva, 2001.

JOACHIMS, T.. Text categorization with support vector machines: learning with many relevant features. *In*: NÉDELLEC, C.; ROUVEIROL, C. (eds.). **Machine Learning: ECML-98**. Springer, Berlin, Heidelberg. Lecture Notes in Computer Science. 2005, v. 1398, p. 137-142. doi: 10.1007/BFb0026683. Acesso em: 16 jul. 2024.

JONES, K. S.. A statistical interpretation of term specificity and its application in retrieval. **Journal of Documentation**. 1972, v. 28, n. 1, p. 11-21. Disponível em: https://www.staff.city.ac.uk/~sbrp622/idfpapers/ksj_orig.pdf. Acesso em: 16 jul. 2024.

KLEIN, D.; MANNING, C. D.. Accurate unlexicalized parsing. *In*: **Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics**. Saporo, Japão. 2003, p. 423-430. Disponível em: https://aclanthology.org/P03-1054/. Acesso em: 16 jul. 2024.

KOEHN, P.. **Statistical machine translation**. Cambridge: Cambridge University Press, 2009.

LIMA, V. R.. Utilizando processamento de linguagem natural para criar uma sumarização automática de textos [Internet]. **Medium.com**. 2017. Disponível em: https://medium.com/@empowerpython/utilizando-processamento-de-linguagem-natural-para-criar-um-sumariza%C3%A7%C3%A3o-autom%C3%A1tica-de-textos-775cb428c84e. Acesso em: 16 jul. 2024.

MIKOLOV, T.; CHEN, K.; CORRADO, G.; DEAN, J.. Efficient estimation of word representations in vector space. **Proceedings of Workshop at ICLR**. 2013. doi: 10.48550/arXiv.1301.3781. Acesso em: 16 jul. 2024.

MIKOLOV, T.; SUTSKEVER, I.; CHEN, K.; CORRADO, G. S.; DEAN, J.. Distributed representations of words and phrases and their compositionality. *In*: **Advances in Neural Information Processing Systems 26 (NIPS 2013)**. 2013, p. 3111-3119. Disponível em: https://papers.nips.cc/paper_files/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf. Acesso em: 16 jul. 2024.

MINAEE, S.; MIKOLOV, T.; NIKZAD, N.; CHENAGHLU, M.; SOCHER, R.; AMATRIAIN, X. *et al.*. Large language models: a survey. **arXiv preprint**. arXiv:2402.06196v2. 2024. Disponível em: https://arxiv.org/pdf/2402.06196. Acesso em: 16 jul. 2024.

NADEAU, D.; SEKINE, S.. A survey of named entity recognition and classification. **Lingvisticae Investigationes**. 2007, v. 30, p. 3-26. Disponível em: https://nlp.cs.nyu.edu/sekine/papers/li07.pdf. Acesso em: 16 jul. 2024.

OLAH, C.. Understanding LSTM Networks. **Colah's Blog** [Internet]. 2015. Disponível em: https://colah.github.io/posts/2015-08-Understanding-LSTMs/. Acesso em: 16 jul. 2024.

PAKHALE, K.. Comprehensive overview of named entity recognition: models, domain-specific applications and challenges. **arXiv preprint**. arXiv:2309.14084v1. 2023. Disponível em: https://arxiv.org/pdf/2309.14084. Acesso em: 16 jul. 2024.

PANG, B.; LEE, L.. Opinion mining and sentiment analysis. **Foundations and Trends in Information Retrieval.** 2008, v. 2, n. 1–2, p. 1–135. doi: 10.1561/1500000011. Acesso em: 16 jul. 2024.

PENNINGTON, J.; SOCHER, R.; MANNING, C.. GloVe: global vectors for word representation. *In*: **Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)**. Doha, Qatar. 2014, p. 1532-1543. doi: 10.3115/v1/D14-1162. Acesso em: 16 jul. 2024.

PYKES, K.. Stemming e lematização em Python [Internet]. **DataCamp Inc.**. 2024. Disponível em: https://www.datacamp.com/pt/tutorial/stemming-lemmatization-python. Acesso em: 16 jul. 2024.

REBOUÇAS, G.. Introdução à análise de sentimentos [Internet]. **Medium.com**. 2023. Disponível em: https://medium.com/@gabriellareboucas6/introdu%C3%A7%C3%A3o-%C3%A0-an%C3%A1lise-de-sentimentos-f968bb9624c3. Acesso em: 16 jul. 2024.

ROCHA, P.. Chatbots e Assistentes Virtuais Inteligentes: por onde começar? [Internet]. **Medium.com**. 2017. Disponível em: https://medium.com/@pedrogomesrocha/chatbots-e-assistentes-virtuais-inteligentes-por-onde-comecar-be5c07368672. Acesso em: 16 jul. 2024.

RUSH, A. M.; CHOPRA, S.; WESTON, J.. A neural attention model for abstractive sentence summarization. *In:* **Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)**. Lisboa, Portugal. 2015, p. 379–389. doi: 10.18653/v1/D15-1044. Acesso em: 16 jul. 2024.

SALTON, G.; MCGILL, M. J.. **Introduction to modern information retrieval.** McGraw-Hill, 1983. 448 p.

SANTOS, D.. A tokenização no processo de linguagem natural e análise de texto [Internet]. **Medium.com**. 2022. Disponível em: https://medium.com/@dheiver.santos_10420/a-tokeniza%C3%A7%C3%A3o-no-processo-de-linguagem-natural-e-an%C3%A1lise-de-texto-41ca71e18501. Acesso em: 16 jul. 2024.

SANTOS, D.. Explorando a interseção entre processamento de linguagem natural e ética: o caminho para uma IA responsável [Internet]. **Medium.com**. 2023. Disponível em: https://medium.com/@dheiver.santos_10420/t%C3%ADtulo-explorando-a-interse%C3%A7%C3%A3o-entre-processamento-de-linguagem-natural-e-%C3%A9tica-o-caminho-para-1790cc3a162f. Acesso em: 16 jul. 2024.

SANTOS, D.. A importância da ética no desenvolvimento do processamento de linguagem natural: construindo um futuro responsável. [Internet]. **Medium.com**. 2022. Disponível em: https://medium.com/@dheiver.santos_10420/t%C3%ADtulo-a-import%C3%A2ncia-

da-%C3%A9tica-no-desenvolvimento-do-processamento-de-linguagem-natural-929f9fa2c80d. Acesso em: 16 jul. 2024.

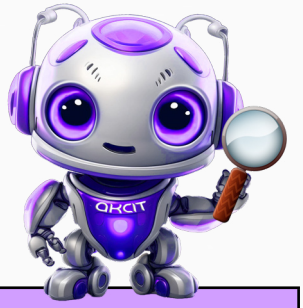
SURREAUX, P.. Otimização do processamento e a filtragem de Stop Words: meus estudos em spaCy e NLP — Parte 3 [Internet]. **Medium.com**. 2024. Disponível em: https://medium.com/@surreauxpp/otimiza%C3%A7%C3%A3o-do-processamento-e-a-filtragem-de-stop-words-meus-estudos-em-spacy-e-nlp-parte-3-872b29ad67d0. Acesso em: 16 jul. 2024.

TAN, K.L.; LEE, C.P.; LIM, K.M. A survey of sentiment analysis: approaches, datasets, and future research. **Applied Sciences.** 2023, v. 13, n. 7, p. 4550. doi: 10.3390/app13074550. Acesso em: 16 jul. 2024.

TECHNOLOGY AND ARTIFICIAL INTELLIGENCE LEAGUE - TAIL. Word Embeddings - Representação vetorial de textos para Machine Learning [Internet]. **Medium.com**. 2020. Universidade Federal da Paraíba. Disponível em: https://tail-ufpb.medium.com/word-embeddings-representação-vetorial-de-textos-para-machine-learning-74a227e18478. Acesso em: 16 jul. 2024.

VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZKOREIT, J.; JONES, L.; GOMEZ, A. N. *et al.*. Attention is all you need. *In*: **Advances in Neural Information Processing Systems 30 (NIPS 2017).** 2017, p. 5998-6008. Disponível em: https://papers.nips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf. Acesso em: 16 jul. 2024.

ZERBINATI, M.. Introdução a Recuperação de Informação (Information Retrieval) [Internet]. **Medium.com**. 2019. Disponível em: https://michel-zerbinati.medium.com/introdu%C3%A7%C3%A3o-a-recupera%C3%A7%C3%A3o-de-informa%C3%A7%C3%A3o-information-retrieval-463023294d7d. Acesso em: 16 jul. 2024.



Saiba mais...

» Competências Imersivas, uma parceria entre a Embrapii e a Universidade Federal de Goiás (UFG).

AKCIT

CENTRO DE COMPETÊNCIA EMBRAPII
EM TECNOLOGIAS IMERSIVAS

SOBRE O E-BOOK

Tipografia: Montserrat

Publicação: Cegraf UFG
Câmpus Samambaia, Goiânia - Goiás. Brasil. CEP 74690-900
Fone: (62) 3521-1358
https://cegraf.ufg.br