PROCESSAMENTO DE LINGUAGEM NATURAL

Episode I : Apresentação do Curso

Alex Marino

14 de novembro de 2024

Conteúdo

Apresentação do Curso

Histórico

Estratégias

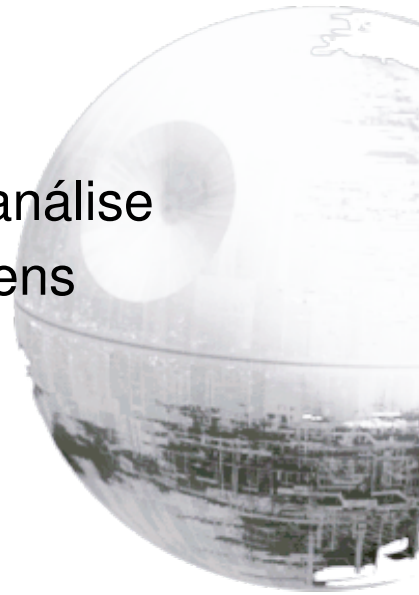
Dupla vitoriosa

Matemática e PLN

Resumo

Objetivo

Conhecer, implementar e aplicar técnicas e estratégias de análise léxica, sintática e semântica do reconhecimento de linguagens humanas.

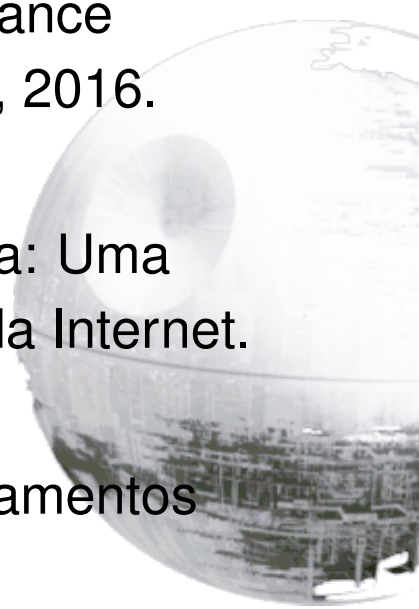


Ementa

Linguagem Natural com Estilo de Interface. Problemas Linguísticos da Linguagem Natural. Análises Léxico-Morfológica, Sintática, Semântica e Pragmática do Processamento de Linguagem Natural.

Bibliografia Básica

- MARTINS, Ana Maria; CARRILHO, Ernestina. Manual de linguística portuguesa: Volume 16 de Manuals of Romance Linguistics. Editora Walter de Gruyter GmbH & Co KG, 2016. ISBN: 3110368846.
- SANTOS; Emilson Moreira dos. Engenharia Lingüística: Uma tecnologia para apoiar as decisões gerenciais na era da Internet. Editora E-papers. ISBN: 8576501554.
- TRAMUNT IBAÑOS, Ana; BATISTA PAIL, Daisy. Fundamentos linguísticos e computação. EDIPUCRS, 2017. ISBN: 853970661X.



Bibliografia de referência

- BIRD, Steven; KLEIN, Ewan; LOPER, Edward. Natural Language Processing with Python. O'ReillyMedia, 2009. ISBN: 978-0-596-51649-9.
- JURAFSKY, Daniel; MARTIN, James H. Speech and Language Processing, 2nd edition. Pearson Prentice Hall, 2008. ISBN: 978-0-13-187321-6.
- MANNING, Christopher D; RAGHAVAN, Prabhakar; SCHÜTZE, Hinrich. Introduction to Information Retrieval. Cambridge University Press, 2008. ISBN 978-0-521-86571-5. Disponível em: https://nlp.stanford.edu/IR-book/pdf/irbookprint.pdf. Acesso em 23 de junho de 2017.
- MANNING, Christopher D; SCHÜTZE, Hinrich. Foundations of Statistical Natural Language Processing, MIT Press, 1999. ISBN 978-0-262-13360-9.

Plano de Ensino

- Curso de 80 horas;
- 20 aulas semanais;
- Aulas expositivas dialógicas;
- Seminários;

Plano de Ensino

Aula	Assunto	Aula	Assunto
01	Apresentação do Curso	11	Seminário II
02	Pre processamento	12	Trabalho II
03	Pré Processamento	13	NER
04	Exploração de Dados	14	POSTAG
05	Redução de dimensionalidade	15	Sumarização
06	Seminário I	16	BERT
07	Trabalho I	16	LLM
08	TFIDF e Word2Vec	18	Seminario III
09	Classificação de Texto	19	Trabalho III
10	Análise de Sentimentos	20	Seminário IV

Avaliação

- Trabalhos em grupos:
  - T1 -
  - T2 -
  - T3 -
  - T4 -
- Trabalho Final individual:
  - Formato artigo científico.

Introdução

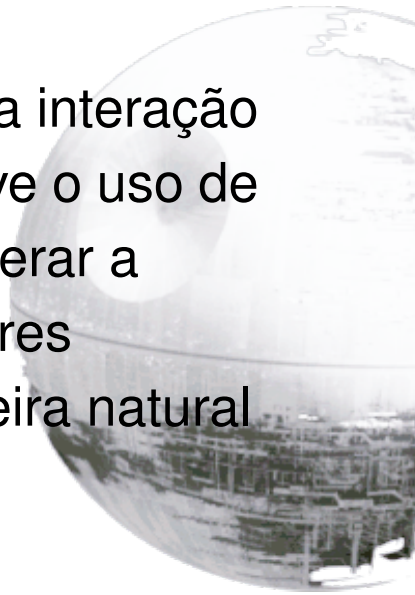
- **O que é processamento de linguagem natural?**

  PLN é um campo da inteligência artificial (IA) focado na interação entre computadores e linguagens humanas. Ele envolve o uso de técnicas computacionais para entender, interpretar e gerar a linguagem humana, tornando possível que computadores compreendam e respondam ao input humano de maneira natural e significativa.

Introdução

- **O que é processamento de linguagem natural?**

  PLN é um campo da inteligência artificial (IA) focado na interação entre computadores e linguagens humanas. Ele envolve o uso de técnicas computacionais para entender, interpretar e gerar a linguagem humana, tornando possível que computadores compreendam e respondam ao input humano de maneira natural e significativa.

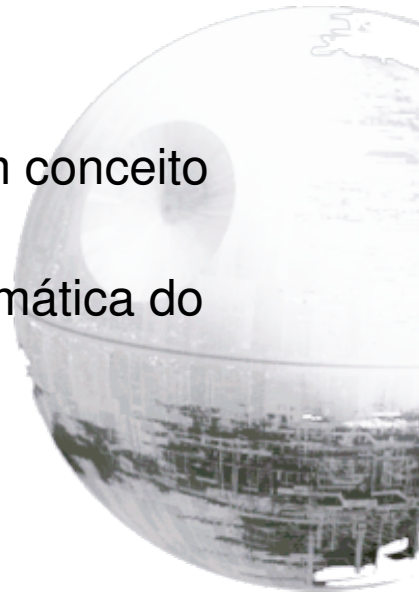


Introdução ao PLN

- **Definição:** O Processamento de Linguagem Natural (PLN) é uma subárea da IA focada na interação entre computadores e linguagens humanas.
- **Objetivo:** Permitir que computadores compreendam, interpretem e gerem linguagem humana.

Primeiros Passos do PLN

- **Décadas de 1940 - 1950:**
  - **Alan Turing:** Introduz o "Teste de Turing"em 1950, um conceito fundamental para IA e processamento de linguagem.
  - **Experimento de Georgetown (1954):** Tradução automática do russo para o inglês, um marco inicial no PLN.

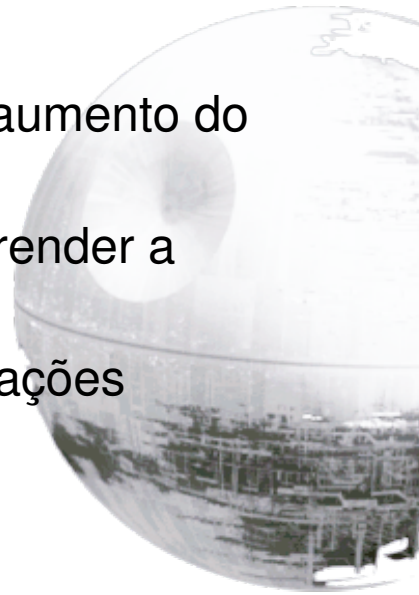


Era dos Sistemas Baseados em Regras

- **Décadas de 1960 - 1970:**
  - Desenvolvimento de sistemas de PLN utilizando **abordagens simbólicas**.
  - **PLN Baseado em Regras:** Sistemas projetados para seguir regras linguísticas predefinidas.
  - **Limitações:** Inflexibilidade e incapacidade de lidar com a complexidade da linguagem humana.

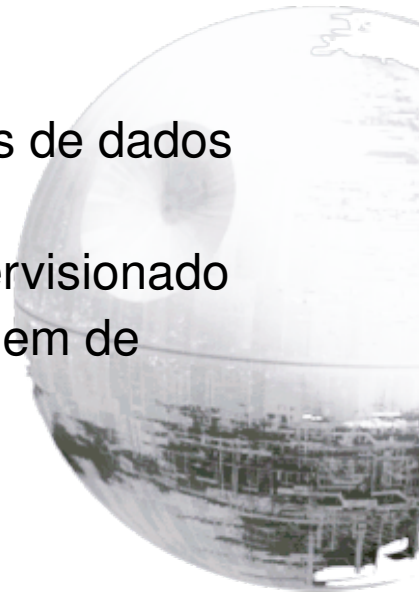
Revolução do PLN Estatístico

- **Décadas de 1980 - 1990:**
  - Mudança para o **PLN Estatístico**, impulsionado pelo aumento do poder computacional e grandes volumes de dados.
  - **Linguística de Corpus:** As máquinas começam a aprender a partir de grandes corpora de texto.
  - **Tradução Automática e Tagging Morfológico:** Aplicações iniciais utilizando modelos probabilísticos.



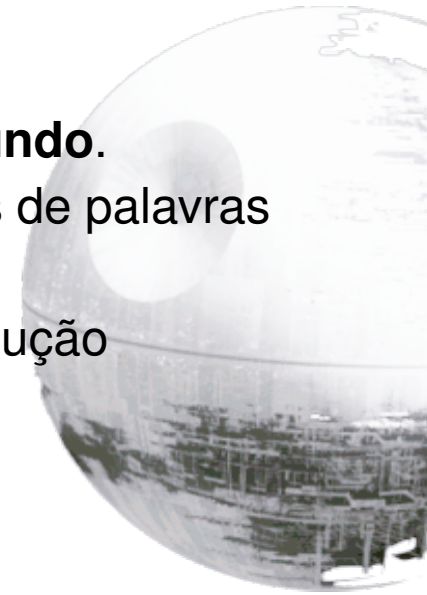
Aprendizado Não-Supervisionado e Semi-Supervisionado

- **Início dos anos 2000:**
  - Surgimento de dados da web oferece vastos conjuntos de dados para tarefas de PLN.
  - Crescimento dos algoritmos de aprendizado não-supervisionado e semi-supervisionado, incluindo clustering e modelagem de tópicos.



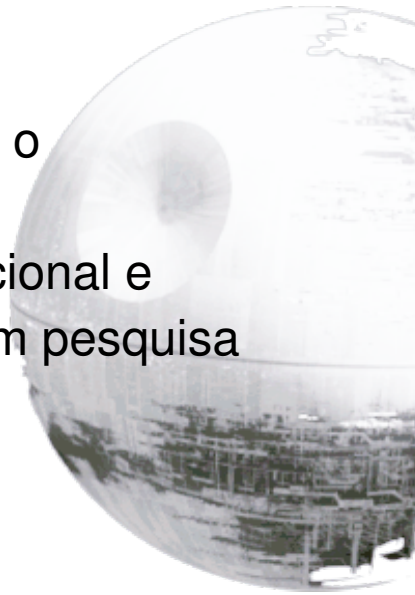
Era do Aprendizado Profundo

- **Década de 2010:**
  - Introdução das **Redes Neurais e Aprendizado Profundo**.
  - **Word2Vec (2013):** Revoluciona o uso de embeddings de palavras no PLN.
  - **RNNs (Redes Neurais Recorrentes):** Melhora a tradução automática e o modelamento de linguagem.



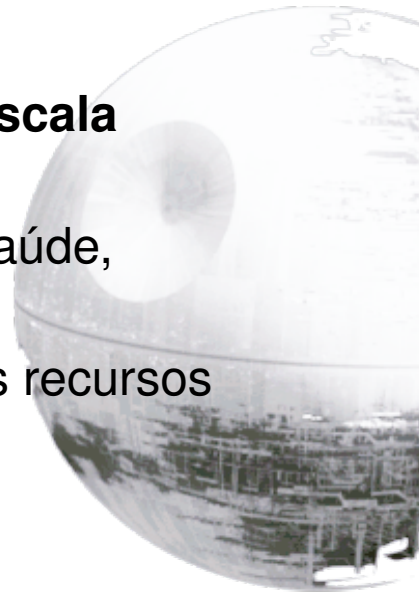
PLN Moderno - Transformers

- **Transformers (2017):**
  - **Inovação-chave:** Mecanismo de atenção que permite o processamento paralelo de palavras.
  - **Modelos BERT e GPT:** BERT para codificação bidirecional e GPT para tarefas generativas tornaram-se o padrão em pesquisa e aplicações de PLN.

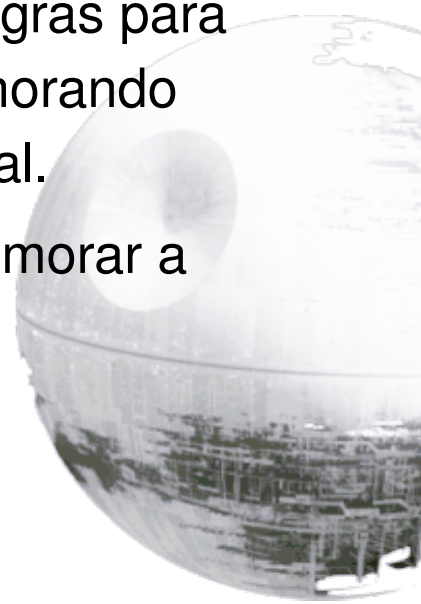


Modelos de Linguagem de Grande Escala e o Futuro

- **Anos 2020 em diante:**
  - Ascensão dos **Modelos de Linguagem de Grande Escala (LLMs)** como o GPT-3 e GPT-4.
  - **Aplicações:** Utilizados em domínios diversos como saúde, finanças e assistentes pessoais.
  - **Desafios:** Viés nos modelos, necessidade de grandes recursos computacionais e impacto ambiental.

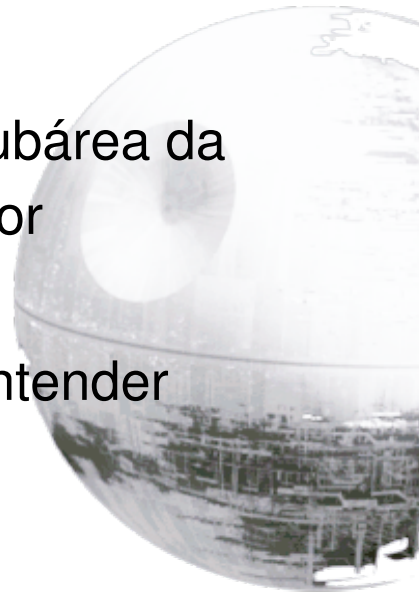


- **Resumo:** O PLN evoluiu de sistemas baseados em regras para abordagens baseadas em aprendizado profundo, melhorando significativamente a compreensão da linguagem natural.
- **Futuro:** Avanços contínuos no PLN continuarão a aprimorar a interação humano-computador em diversos setores.



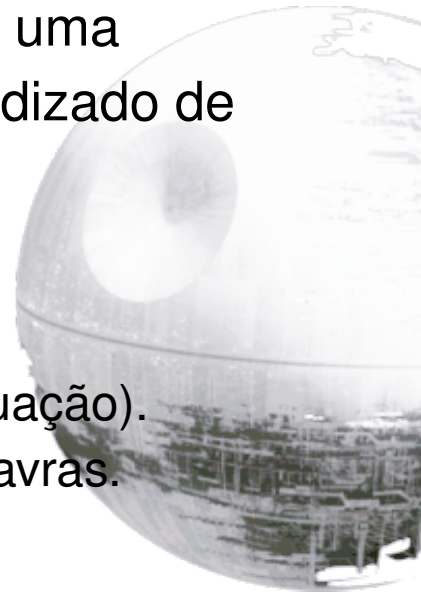
Estratégias Iniciais no Processamento de Linguagem Natural por Máquinas

- Processamento de Linguagem Natural (PLN) é uma subárea da IA dedicada ao entendimento da linguagem humana por máquinas.
- O objetivo é criar sistemas que possam processar e entender textos de maneira eficaz.



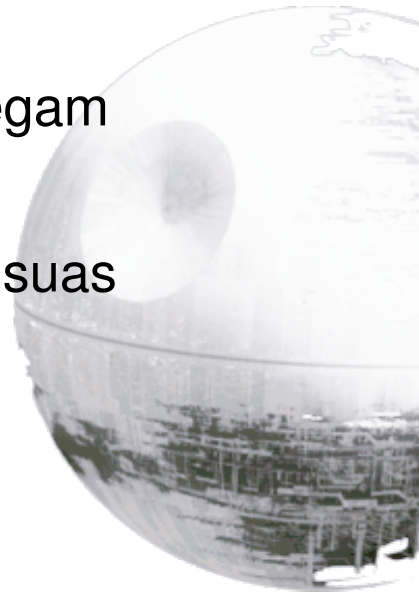
Pré-processamento de Texto

- **Importância:** Limpar e transformar o texto original em uma forma que pode ser analisada por algoritmos de aprendizado de máquina (AM).
- **Técnicas:**
  - Remoção de ruídos e inconsistências.
  - Normalização do texto (minúsculas, remoção de pontuação).
  - Tokenização: divisão do texto em palavras ou sub-palavras.



Redução da Dimensionalidade

- **Stop Words:** Remoção de palavras comuns que carregam pouco significado (ex.: "a", "de", "o").
- **Stemming e Lematização:** Redução das palavras às suas formas base (ex.: "correu"$\rightarrow$ "correr").



Normalização e Limpeza de Dados

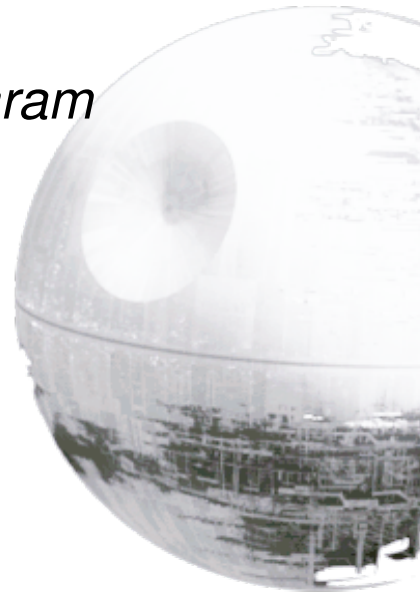
- **Normalização:** Converte todo o texto para um formato padronizado, facilitando a análise.
- **Limpeza de Dados:** Remoção de dados duplicados, irrelevantes ou erros.
- **Benefícios:** Melhor desempenho dos algoritmos de aprendizado de máquina.

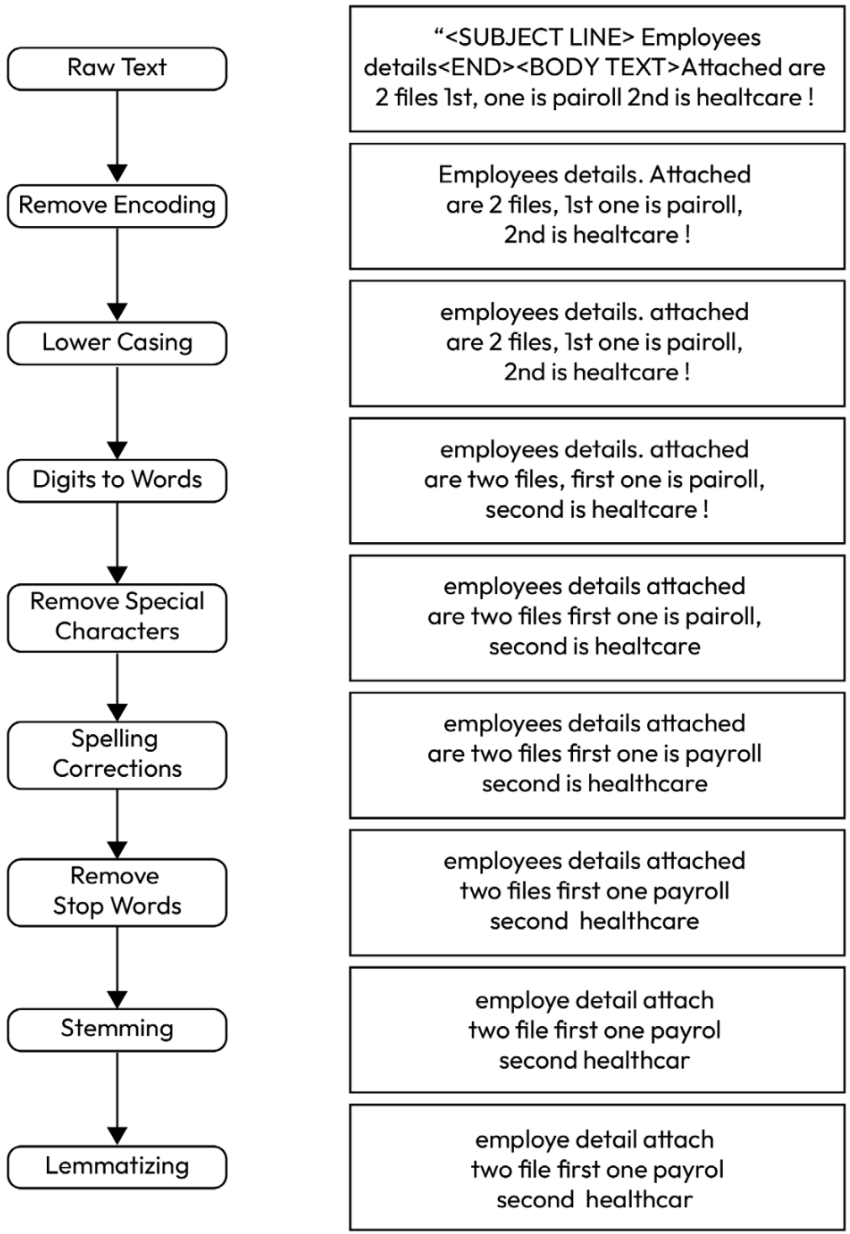
Stemming e Lematização

- **Stemming:** Reduz palavras às suas raízes morfológicas, ignorando afixos e tempos verbais.
- **Lematização:** Reduz palavras à sua forma base gramaticalmente correta (ex.: "correndo"$\rightarrow$ "correr").

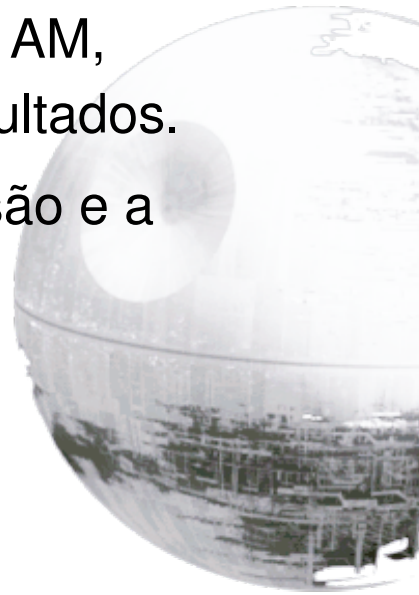
Exemplo de Pré-processamento de Texto

- Frase original: *"Os meninos correram, pularam e nadaram rapidamente."*
- Após Stemming: *"Menino correr, pular, nadar rápido."*
- Após Lematização: *"Menino corre, pula, nada rápido."*



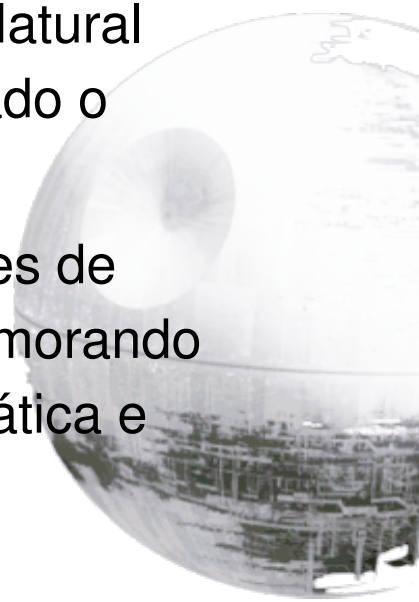

Raw Text
"<SUBJECT LINE> Employees details<END><BODY TEXT>Attached are 2 files 1st, one is pairoll 2nd is healtcare !
Remove Encoding
Employees details. Attached are 2 files, 1st one is pairoll, 2nd is healtcare !
Lower Casing
employees details. attached are 2 files, 1st one is pairoll, 2nd is healtcare !
Digits to Words
employees details. attached are two files, first one is pairoll, second is healtcare !
Remove Special Characters
employees details attached are two files first one is pairoll, second is healtcare
Spelling Corrections
employees details attached are two files first one is payroll second is healthcare
Remove Stop Words
employees details attached two files first one payroll second healthcare
Stemming
employe detail attach two file first one payrol second healthcar
Lemmatizing
employe detail attach two file first one payrol second healthcar

- O pré-processamento é um passo essencial no PLN e AM, garantindo dados de alta qualidade para melhores resultados.
- Investir tempo no pré-processamento melhora a precisão e a confiabilidade dos modelos.



Introdução

- A convergência entre Processamento de Linguagem Natural (PLN) e Aprendizado de Máquina (ML) tem transformado o campo da inteligência artificial.
- Essa sinergia permite que máquinas aprendam padrões de linguagem a partir de grandes volumes de dados, aprimorando tarefas como reconhecimento de fala, tradução automática e geração de texto.

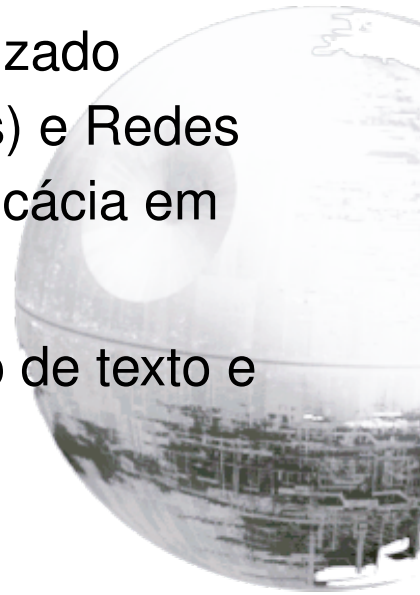


Modelos Estatísticos de Linguagem

- **Modelagem Estatística de Linguagem:** Envolve treinar algoritmos em grandes corpora de texto para prever a probabilidade de uma sequência de palavras.
- **Aplicações:** Reconhecimento de fala, tradução automática, geração de texto.

Aprendizado Profundo em PLN

- **Redes Neurais Profundas (DL):** Técnicas de Aprendizado Profundo, como Redes Neurais Convolucionais (CNNs) e Redes Neurais Recorrentes (RNNs), têm mostrado grande eficácia em tarefas de PLN.
- **Exemplos:** Compreensão de linguagem, sumarização de texto e análise de sentimentos.



- O PLN está intimamente relacionado ao Aprendizado de Máquina e ao Aprendizado Profundo.

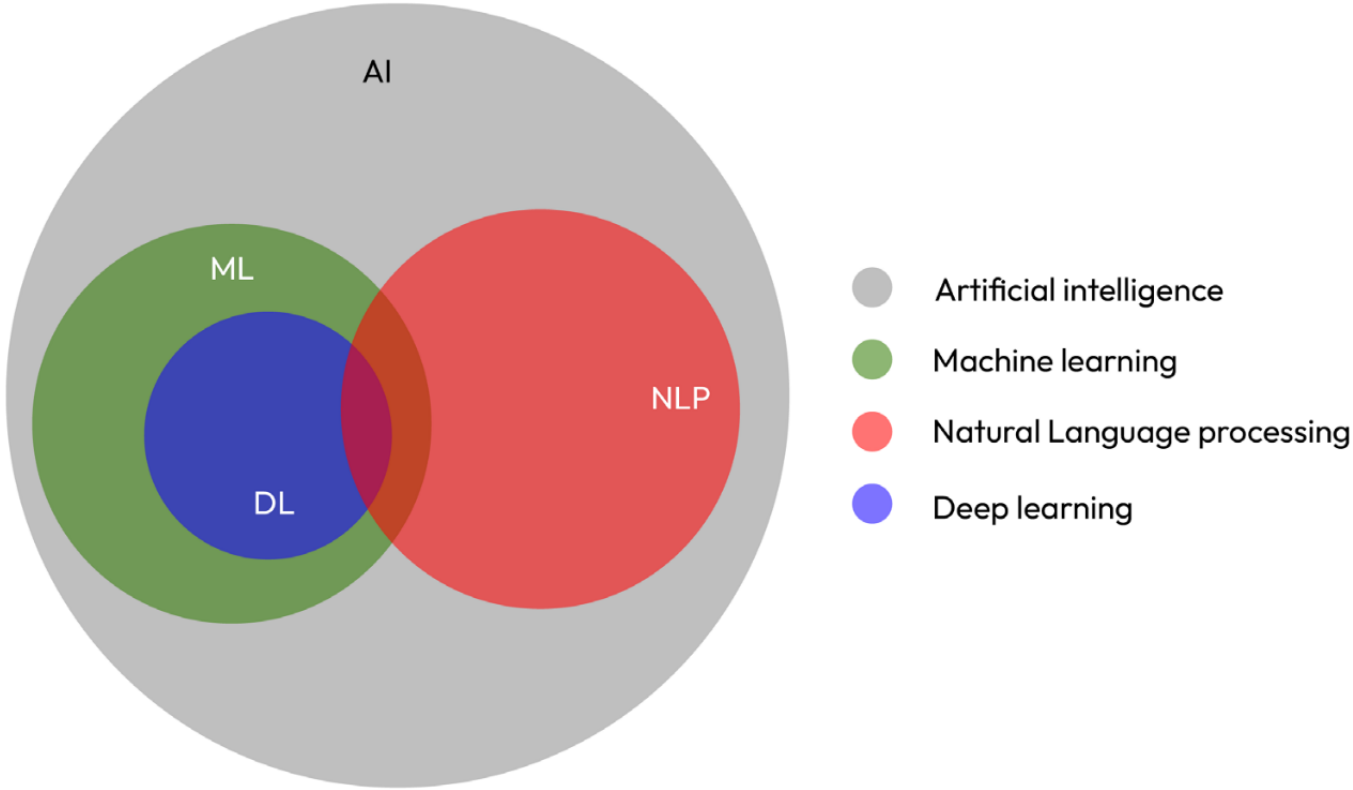
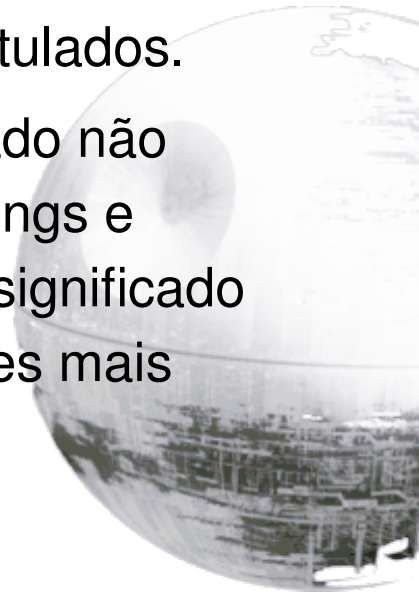


Figura: Relação entre IA, ML, DL e PLN (inserir imagem de uma pirâmide ou gráfico mostrando a hierarquia entre as áreas).

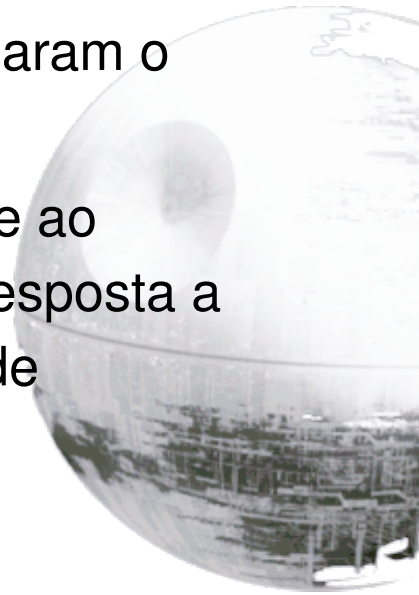
Desafios e Soluções no PLN

- **Desafio:** Lidar com grandes volumes de dados não rotulados.
- **Soluções:** Desenvolvimento de técnicas de aprendizado não supervisionado e semi-supervisionado, como embeddings e mecanismos de atenção, que ajudam a representar o significado das palavras de forma numérica e a identificar as partes mais importantes do texto.



Impacto dos Modelos Pré-Treinados

- Modelos pré-treinados, como **BERT** e **GPT**, revolucionaram o campo do PLN.
- Esses modelos são capazes de gerar texto semelhante ao humano e melhorar o desempenho em tarefas como resposta a perguntas, análise de sentimentos e reconhecimento de entidades nomeadas.



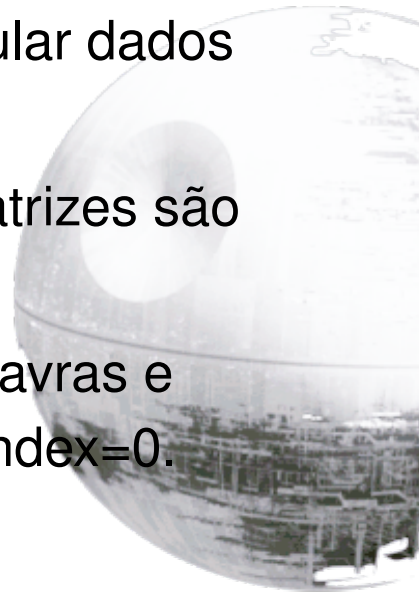
- A integração de PLN e ML continua a impulsionar avanços significativos em IA.
- Essa sinergia permite que sistemas automatizados realizem tarefas linguísticas complexas com maior precisão, transformando setores como saúde, finanças e atendimento ao cliente.

Introdução

- PLN e aprendizado de máquina são construídos sobre bases matemáticas sólidas.
- Conceitos chave incluem Álgebra Linear, Estatística, Probabilidade e Otimização.
- Recentes avanços incluem modelos pré-treinados como BERT e GPT, que revolucionaram o campo do PLN.

Álgebra Linear no PLN

- Vetores e Matrizes: Usados para representar e manipular dados textuais.
- Operações: Soma, multiplicação e transposição de matrizes são fundamentais.
- Aplicações: Modelagem de dados, embeddings de palavras e análise semântica&8203;:contentReference[oaicite:0]index=0.

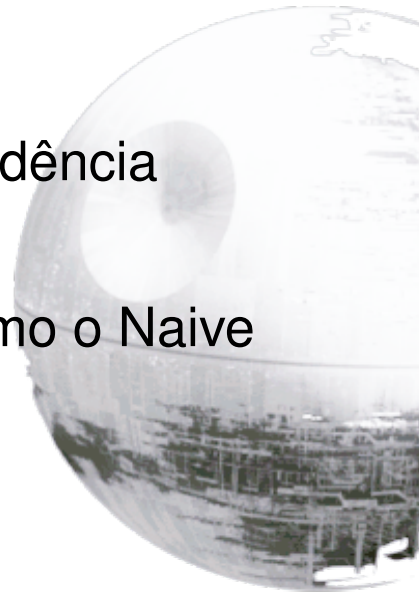


Autovalores e Autovetores

- **Autovalores e Autovetores**: Utilizados em técnicas de redução de dimensionalidade, como PCA.
- Importante para análise de grandes corpora textuais&8203;:contentReference[oaicite:1]index=1.

Probabilidade Básica para PLN

- A probabilidade mede a chance de um evento ocorrer.
- Conceitos chave: Espaço amostral, eventos e independência estatística.
- Aplicações: Modelos de linguagem probabilísticos, como o Naive Bayes8203;:contentReference[oaicite:2]index=2.

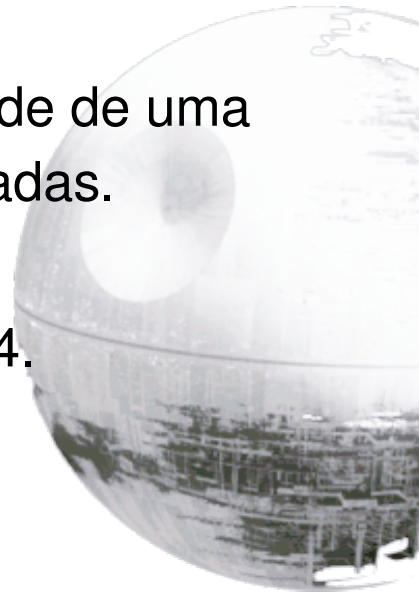


Distribuições de Probabilidade

- Distribuições discretas e contínuas são usadas para modelar variações na linguagem.
- Exemplo: A distribuição normal é frequentemente utilizada para representar a variação de parâmetros em modelos de linguagem&8203;:contentReference[oaicite:3]index=3.

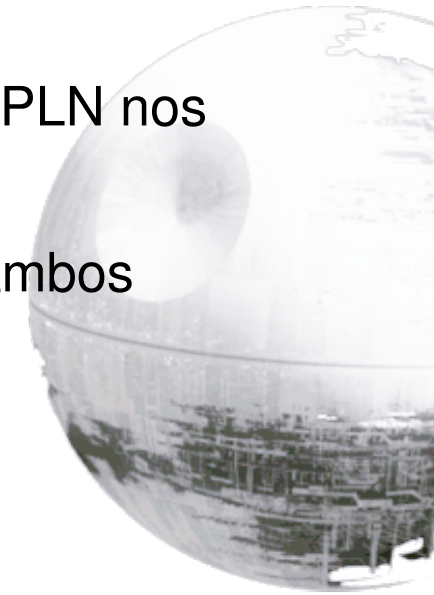
Estimativa Bayesiana

- A estimativa Bayesiana permite atualizar a probabilidade de uma hipótese à medida que novas evidências são incorporadas.
- Importante para classificação de texto e análise de sentimentos&8203;:contentReference[oaicite:4]index=4.



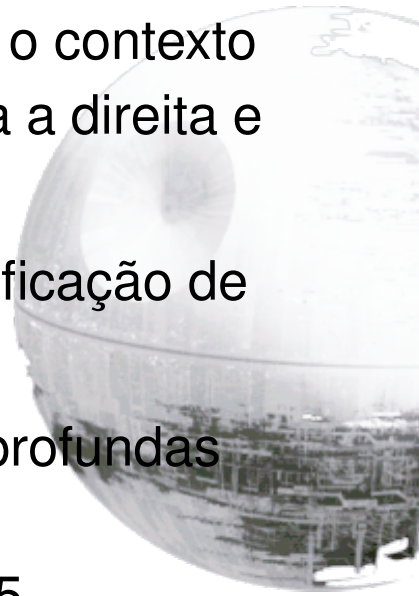
Modelos Pré-Treinados no PLN

- Avanços em modelos pré-treinados revolucionaram o PLN nos últimos anos.
- Dois dos mais influentes modelos são BERT e GPT, ambos baseados na arquitetura Transformer.



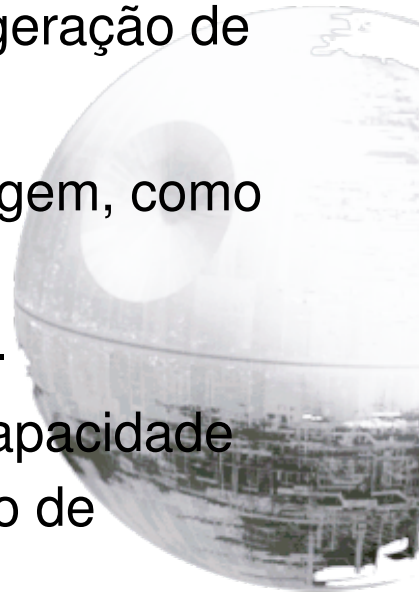
BERT: Bidirectional Encoder Representations from Transformers

- **BERT**: Modelo de linguagem bidirecional que entende o contexto das palavras em ambas as direções (da esquerda para a direita e da direita para a esquerda).
- Aplicações: Tarefas como resposta a perguntas, classificação de texto e reconhecimento de entidades nomeadas.
- Vantagem: Captura de informações contextuais mais profundas em comparação com modelos unidirecionais8203;:contentReference[oaicite:5]index=5.



GPT: Generative Pre-trained Transformer

- **GPT**: Modelo unidirecional que gera texto com base em uma sequência anterior, tornando-o eficaz para tarefas de geração de texto, como diálogos e completamento de frases.
- Vantagem: Excelente em tarefas de geração de linguagem, como chatbots e escrita automática&8203;:contentReference[oaicite:6]index=6.
- Evolução: GPT-2 e GPT-3 ampliaram o tamanho e a capacidade dos modelos, levando a uma maior fluência na geração de texto&8203;:contentReference[oaicite:7]index=7.



- A compreensão dos fundamentos matemáticos e o uso de modelos pré-treinados são essenciais para o sucesso no PLN moderno.
- BERT e GPT demonstram o poder das arquiteturas Transformer em revolucionar o campo da linguagem natural.
- O PLN continuará a evoluir, impulsionado por inovações matemáticas e de modelos de aprendizado profundo.

Exemplo de Modelos de Linguagem - ChatGPT

- ChatGPT é um exemplo popular de modelo de linguagem baseado em transformers.
- Ele é pré-treinado em grandes quantidades de dados textuais para entender nuances da linguagem humana.
- Aplicações: chatbots, assistentes virtuais, sistemas de diálogo, sumarização de texto, entre outros.

Popularidade do ChatGPT

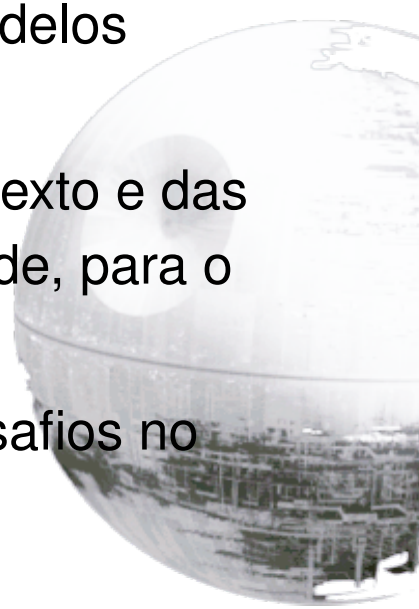
- Sua popularidade se deve à capacidade de gerar texto semelhante ao humano e de ser adaptado para casos de uso específicos.
- Pode ser ajustado para tarefas como análise de sentimentos, resposta a perguntas, entre outras.
- Usado amplamente na indústria e em pesquisas .

Capacidades do ChatGPT

- Gera texto com qualidade elevada e respostas naturais.
- Suas saídas são difíceis de distinguir de texto escrito por humanos.
- Adequado para tarefas que requerem compreensão profunda da linguagem .

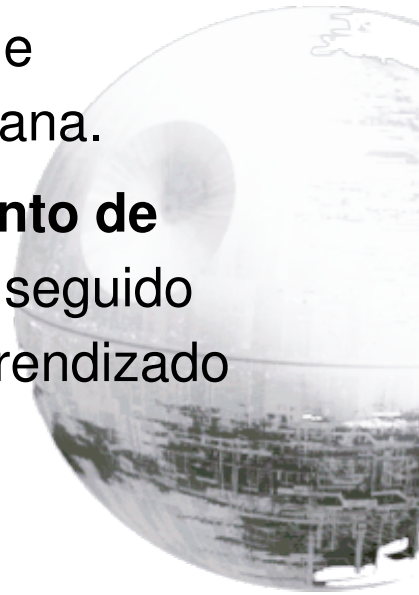
Resumo

- Discutimos os avanços recentes no PLN, incluindo modelos pré-treinados como BERT e GPT.
- Destacamos a importância do pré-processamento de texto e das bases matemáticas, como álgebra linear e probabilidade, para o PLN.
- Abordamos a importância de dados rotulados e os desafios no entendimento do contexto e significado das palavras .

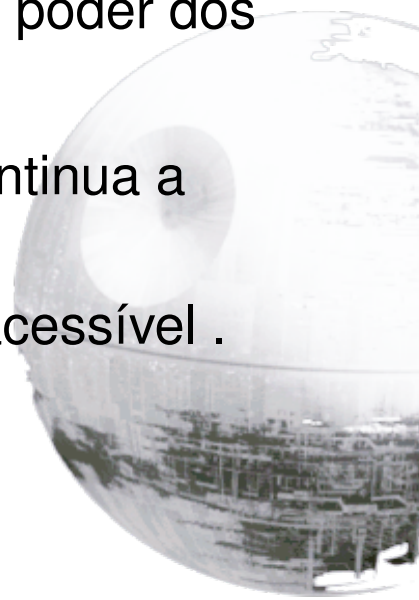


Perguntas e Respostas

- **O que é PLN?** Subcampo da IA focado em permitir que computadores compreendam e gerem linguagem humana.
- **Quais foram as estratégias iniciais no processamento de linguagem?** Uso de regras baseadas em gramáticas, seguido pelo desenvolvimento de modelos estatísticos e de aprendizado profundo .



- Modelos de linguagem como ChatGPT exemplificam o poder dos transformers no PLN moderno.
- A integração de PLN com aprendizado de máquina continua a impulsionar avanços em várias aplicações, tornando o entendimento da linguagem cada vez mais preciso e acessível .



MAY THE

SOURCE

BE WITH

YOU