



# Processamento de Linguagem Natural - PLN

Stefane Adna dos Santos

# Sumário

1. Introdução ao PLN.
2. Tarefas de PLN.
3. Os principais desafios de PLN.
4. Pré-processamento de texto.
5. Bibliotecas.

# O que é Processamento de Linguagem Natural?

# Introdução ao PLN

- Processamento de Linguagem Natural (PLN) é uma área da ciência da computação e da inteligência artificial que se preocupa com o desenvolvimento de sistemas computacionais capazes de entender, interpretar e gerar linguagem natural humana.

# Introdução ao PLN

- Alguns exemplos de aplicações de PLN incluem:
  - Assistentes virtuais e chatbots que podem interagir com usuários em linguagem natural e responder perguntas.
  - Sistemas de tradução automática que podem traduzir textos entre diferentes idiomas.

# Introdução ao PLN

- Alguns exemplos de aplicações de PLN incluem:
  - Sistemas de recomendação de conteúdo que utilizam análise de sentimentos para recomendar produtos ou serviços para usuários.
  - Análise de dados de redes sociais para identificar tendências e opiniões públicas.

# Introdução ao PLN

- Alguns exemplos de aplicações de PLN incluem:
  - Reconhecimento de fala para interação com dispositivos e sistemas. Sumarização automática de textos para auxiliar na compreensão de grandes volumes de informação.

# Tarefas de PLN

- **Classificação de Texto:** Identificar o tipo de texto ou classificar documentos em categorias específicas. Exemplo: classificar se um e-mail é spam ou não spam.
- **Sumarização de Texto:** Criar um resumo ou síntese de um texto longo, ou documento. Exemplo: Resumir um artigo de notícias em alguns parágrafos.



# Tarefas de PLN

- **Tradução Automática:** Converter um texto escrito em uma língua natural para outra. Exemplo: traduzir um texto em inglês para o português.
- **Reconhecimento de Entidades Nomeadas:** Identificar e extrair informações de entidades nomeadas, como nomes de pessoas, lugares e organizações. Exemplo: Identificar os nomes de empresas mencionadas em um artigo.

# Tarefas de PLN

- **Análise de Sentimentos:** Determinar a atitude emocional de um texto, geralmente positiva, negativa ou neutra. Exemplo: Analisar as avaliações de produtos em uma loja virtual para determinar se os clientes estão satisfeitos ou insatisfeitos com os produtos.
- **Preenchimento de lacunas ou Infilling:** prever ou gerar palavras, frases ou sentenças em branco em um texto. Exemplo: Preencher as lacunas em uma frase com palavras apropriadas ou corrigir erros gramaticais.

# Tarefas de PLN

- **Resposta Automática:** Gerar uma resposta para uma pergunta ou mensagem. Exemplo: respostas automáticas para perguntas frequentes em um website.
- **Análise de Tópicos:** Identificar os tópicos principais discutidos em um texto ou conjunto de textos. Exemplo: Analisar uma coleção de documentos para identificar os temas mais discutidos.

# Os principais desafios de PLN

- Ambiguidade.
- Variedade de linguagens e dialetos.
- Diferentes formas de expressão.
- Ambiente online.
- Aprendizado de máquina.
- Falta de compreensão profunda da linguagem.

# Pré-processamento de texto

- O pré-processamento de texto é uma etapa crucial no processamento de linguagem natural (NLP). Envolve a limpeza e a transformação do texto de entrada para torná-lo adequado para ser analisado por algoritmos de machine learning.

# Pré-processamento de texto

- Ao preparar o texto adequadamente antes de analisá-lo, é possível garantir que os algoritmos de NLP possam extrair informações precisas e úteis a partir dos dados.

# Pré-processamento de texto

- Algumas das tarefas de pré-processamento de texto comuns incluem:
  - Tokenização
  - Remoção de stopwords
  - Normalização de texto
  - Remoção de caracteres especiais
  - Stemming e lematização
  - Vetorização de texto

# Pré-processamento de texto

- **Tokenização:** é o processo de dividir o texto em unidades menores chamadas tokens, que geralmente são palavras ou símbolos. Isso é necessário porque os algoritmos de NLP trabalham ao nível de token para analisar a linguagem.



# Pré-processamento de texto

- Remoção de stopwords: são palavras que aparecem com frequência na linguagem, mas que não carregam significado, como "a", "e" e "o". Remover essas palavras pode reduzir o tamanho do vocabulário e melhorar o desempenho do algoritmo de NLP.

# Pré-processamento de texto

- Normalização de texto: envolve a padronização do texto para remover variações desnecessárias. Isso pode incluir converter todas as letras para minúsculas, remover a pontuação ou substituir abreviações por suas formas completas.

# Pré-processamento de texto

- Remoção de caracteres especiais: caracteres especiais, como emojis e emoticons, podem não ter nenhum significado em algumas tarefas de NLP e, portanto, podem ser removidos.

# Pré-processamento de texto

- Stemming e lematização: são técnicas para reduzir as palavras a sua forma base ou raiz. O stemming envolve a remoção dos sufixos das palavras, enquanto a lematização envolve a identificação do lema ou forma base da palavra.

Original	Stemização	Lematização
Amigas	amig	amigo
Amigos	amig	amigo
Amizade	Amizad	Amizade

# Pré-processamento de texto

- Vetorização de texto: é o processo de converter o texto em um vetor numérico para ser utilizado como entrada em modelos de NLP. Existem várias técnicas de vetorização, incluindo a contagem de frequência de palavras e o uso de modelos pré-treinados de word embeddings.

# Bibliotecas

- Spacy
  - `pip install spacy`
- NLTK
  - `pip install nltk`
- Gensim
  - `pip install gensim`
- Re
  - `pip install regex`

# Referencias

- Spacy. Industrial-Strength Natural Language Processing. Disponível em:< <https://spacy.io/>>. Acesso em 16 de fevereiro de 2023.
- NLTK. Natural Language Toolkit. Disponível em:< <https://www.nltk.org/>. Acesso em 16 fevereiro de 2023.
- Wagner, W. (2010). Steven bird, Ewan Klein and Edward Loper: Natural language processing with python, analyzing text with the natural language toolkit: O'Reilly media, Beijing, 2009, ISBN 978-0-596-51649-9.