Processamento de Linguagem Natural

Episode II : NLTK Básico

Alex Marino

14 de novembro de 2024

Conteúdo

PROCESSAMENTO DE LINGUAGEM NATURAL

EPISODE II : NLTK BÁSICO

Alex Marino

14 de novembro de 2024

Introdução ao Pipeline de PLN

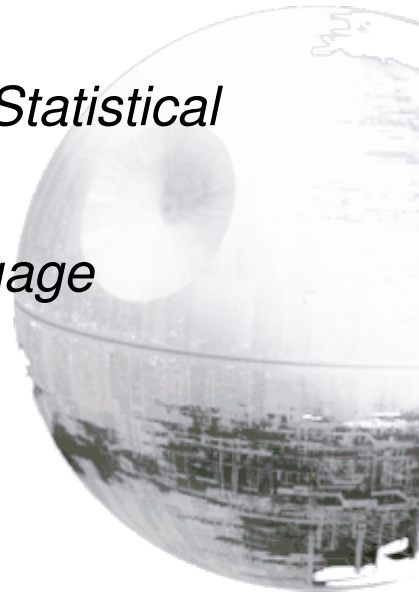
- O Processamento de Linguagem Natural (PLN) é uma subárea da inteligência artificial que lida com a interação entre computadores e a linguagem humana.
- O pipeline de PLN inclui diversas etapas para transformar texto bruto em uma forma que possa ser compreendida e processada por máquinas.

Objetivos do Pipeline de PLN

- Extrair informações úteis de textos.
- Entender o contexto, estrutura e significado das palavras.
- Facilitar a análise de grandes volumes de dados textuais.

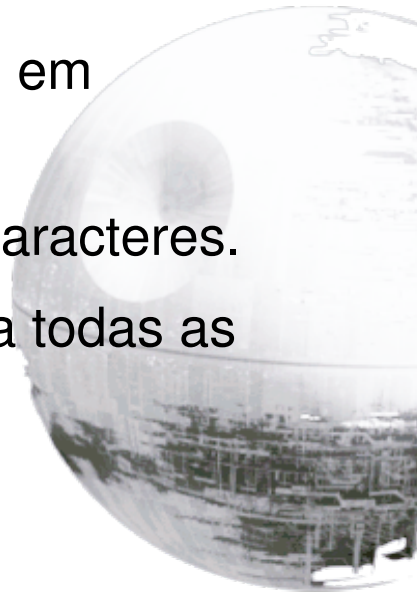
Referências Fundamentais

- Manning, C. D., & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press.
- Jurafsky, D., & Martin, J. H. (2009). *Speech and Language Processing*. Pearson.



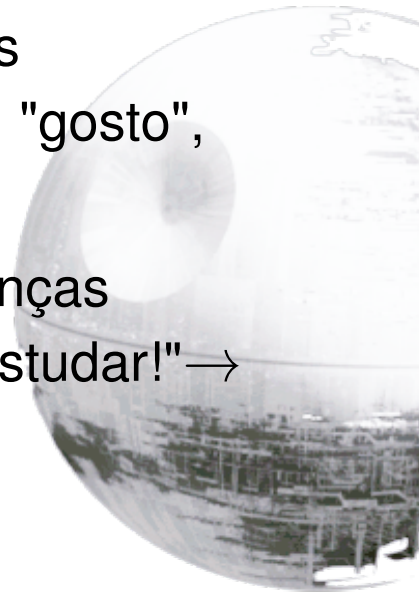
Tokenização

- **Definição**: Tokenização é o processo de dividir o texto em unidades menores, chamadas tokens.
- Os tokens podem ser palavras, frases ou até mesmo caracteres.
- É a primeira etapa do pipeline de PLN e essencial para todas as tarefas subsequentes.



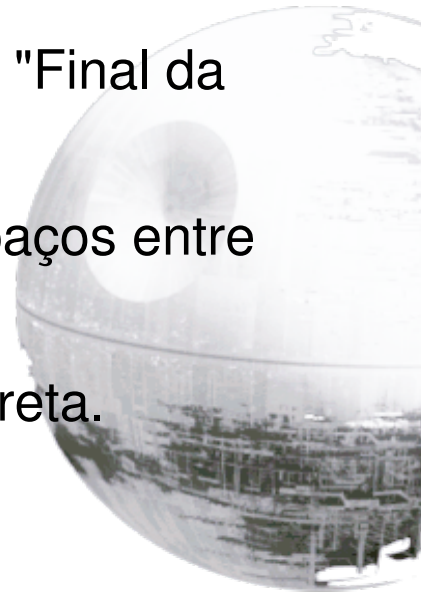
Tipos de Tokenização

- **Tokenização por palavras**: Divide o texto em palavras individuais. Exemplo: "Eu gosto de aprender."→ ["Eu", "gosto", "de", "aprender", "."]
- **Tokenização por sentenças**: Divide o texto em sentenças completas. Exemplo: "Eu gosto de aprender. Vamos estudar!"→ ["Eu gosto de aprender.", "Vamos estudar!"]



Desafios na Tokenização

- Ambiguidade no uso de pontuações (ex: "Dr. João"vs. "Final da frase.").
- Línguas com morfologia complexa ou ausência de espaços entre palavras (ex: chinês).
- Diferenças linguísticas que afetam a segmentação correta.



Referências sobre Tokenização

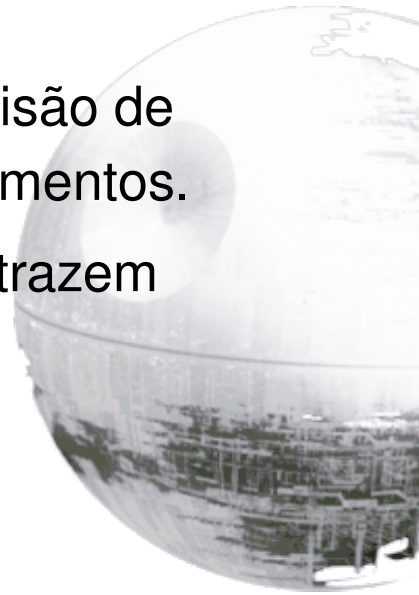
- Grefenstette, G. (1999). *Tokenization*. In *Text-Based Information Retrieval* (pp. 26-32).
- Jurafsky, D., & Martin, J. H. (2009). *Speech and Language Processing*. Pearson, Cap. 2.

Stopwords

- **Definição**: Stopwords são palavras comuns, como artigos e preposições, que geralmente são removidas durante a análise textual.
- Essas palavras têm baixo valor semântico e são filtradas para reduzir o ruído na análise.

Importância de Remover Stopwords

- A remoção de stopwords melhora a eficiência e a precisão de tarefas como classificação de textos e análise de sentimentos.
- Sem as stopwords, o algoritmo foca nas palavras que trazem mais significado.

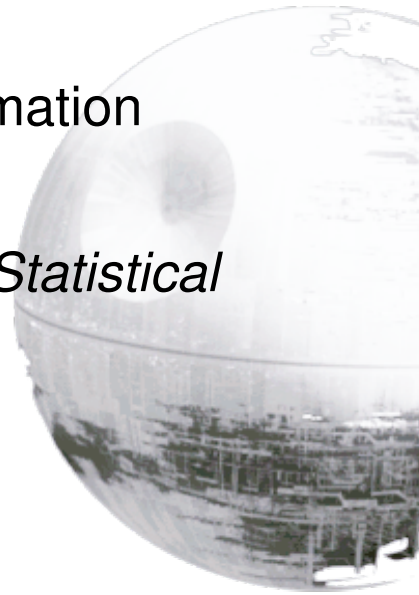


Desafios na Definição de Stopwords

- Nem todas as palavras comuns devem ser removidas, pois o contexto pode torná-las relevantes.
- Diferenças entre domínios e tipos de texto podem exigir listas de stopwords personalizadas.

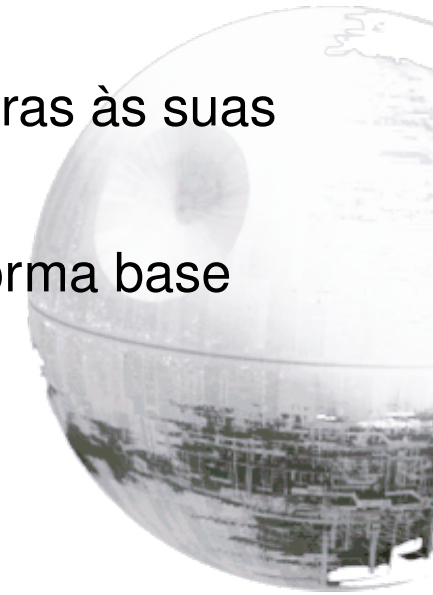
Referências sobre Stopwords

- Fox, C. (1992). *Lexical Analysis and Stoplists*. In Information Retrieval: Data Structures and Algorithms.
- Manning, C. D., & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*, Cap. 4.



Stemming

- **Definição**: O stemming é o processo de reduzir palavras às suas raízes ou radicais.
- A ideia é remover sufixos e prefixos para obter uma forma base da palavra.



Tipos de Algoritmos de Stemming

- **Porter Stemmer**: Um dos algoritmos de stemming mais utilizados.
- **Snowball Stemmer**: Uma versão aprimorada do Porter Stemmer.
- **Stemmer RSLP**: Um algoritmo de stemming desenvolvido especificamente para o português.

Limitações do Stemming

- O stemming pode produzir radicais que não são palavras válidas no idioma.
- Algoritmos de stemming podem não lidar bem com palavras compostas ou sufixos complexos.

Referências sobre Stemming

- Porter, M. F. (1980). *An algorithm for suffix stripping*. Program, 14(3), 130-137.
- Lovins, J. B. (1968). *Development of a stemming algorithm*. Mechanical Translation and Computational Linguistics.

Lematização

- **Definição**: A lematização é o processo de transformar uma palavra em sua forma base ou "lema".
- Diferente do stemming, a lematização leva em consideração o contexto e a morfologia da palavra.

Exemplos de Lematização

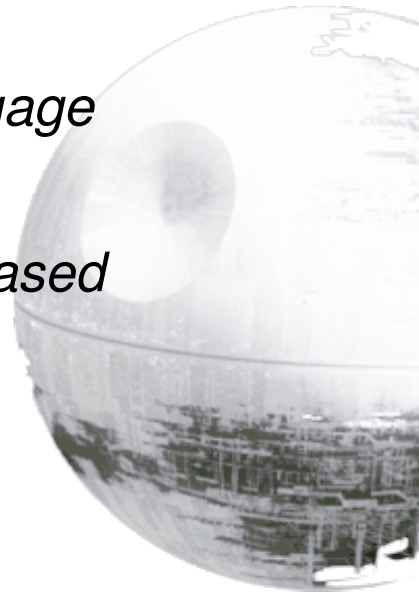
- Palavras: "correndo", "correu", "correrão"
- Lema: "correr"
- A lematização retorna a forma base da palavra respeitando suas variações gramaticais.

Benefícios da Lematização

- Produz resultados mais precisos do que o stemming.
- Ajuda a melhorar a precisão em tarefas de classificação de textos e recuperação de informação.

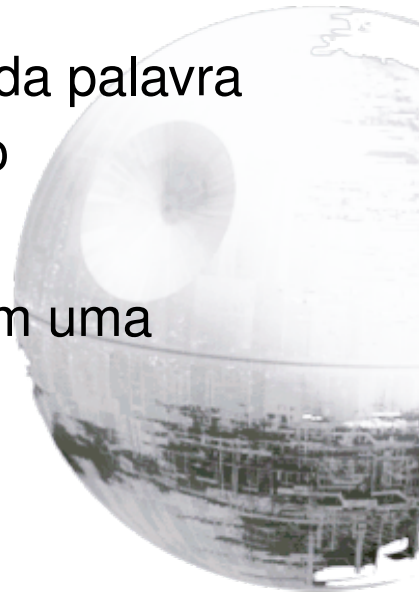
Referências sobre Lematização

- Jurafsky, D., & Martin, J. H. (2009). *Speech and Language Processing*. Pearson, Cap. 3.
- Plisson, J., Lavrac, N., & Mladenic, D. (2004). *A rule-based approach to word lemmatization*. Proceedings of IS.



POS-Tagging (Part-of-Speech Tagging)

- **Definição**: POS-Tagging é o processo de etiquetar cada palavra em uma sentença com sua categoria gramatical, como substantivo, verbo, adjetivo, etc.
- Isso ajuda a compreender a função de cada palavra em uma sentença.

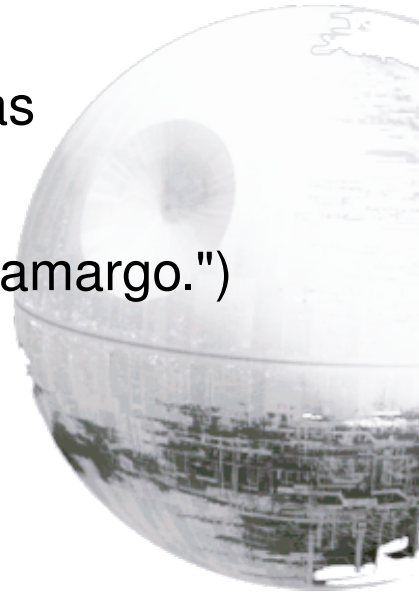


Categorias Gramaticais

- **Substantivo (NOUN)**: Nomeia seres, objetos, lugares, etc.
- **Verbo (VERB)**: Expressa ações ou estados.
- **Adjetivo (ADJ)**: Atribui características aos substantivos.
- E outras categorias, como pronomes, preposições, advérbios, etc.

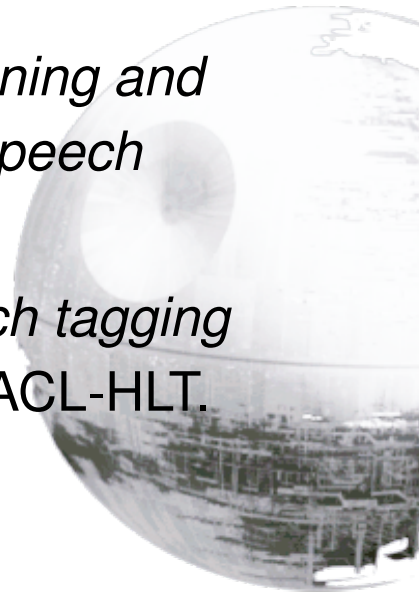
Desafios no POS-Tagging

- Palavras ambíguas que podem ter diferentes categorias dependendo do contexto.
- Exemplo: "gosto"pode ser um substantivo ("O gosto é amargo.") ou um verbo ("Eu gosto de café.").



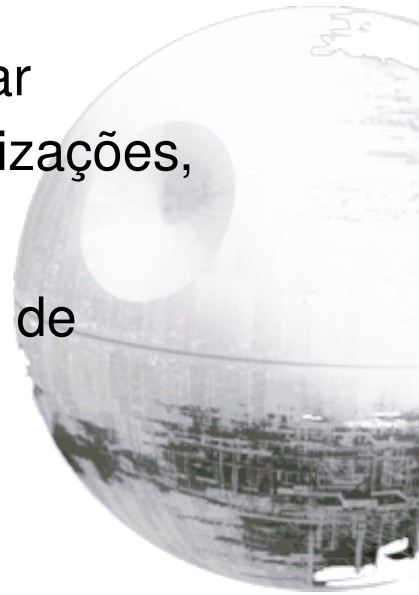
Referências sobre POS-Tagging

- Brill, E. (1995). *Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging*. Computational linguistics.
- Toutanova, K., et al. (2003). *Feature-rich part-of-speech tagging with a cyclic dependency network*. Proceedings of NAACL-HLT.



Reconhecimento de Entidades Nomeadas (NER)

- **Definição**: NER é o processo de identificar e classificar entidades nomeadas, como nomes de pessoas, organizações, locais, datas, etc., dentro de um texto.
- O NER é amplamente usado em sistemas de extração de informações.



Tipos de Entidades Nomeadas

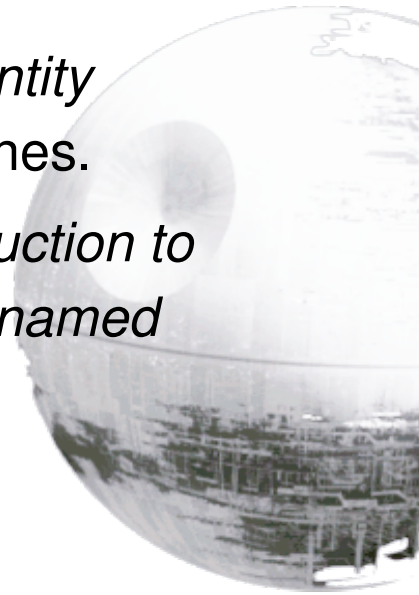
- **Pessoa**: Identifica nomes de indivíduos.
- **Organização**: Empresas, instituições, etc.
- **Local**: Países, cidades, continentes, etc.
- **Datas e Períodos**: Identificação de expressões temporais.

Aplicações de NER

- Extração de informações relevantes em grandes volumes de texto.
- Melhoria de sistemas de busca e classificação de documentos.
- Análise de notícias e redes sociais.

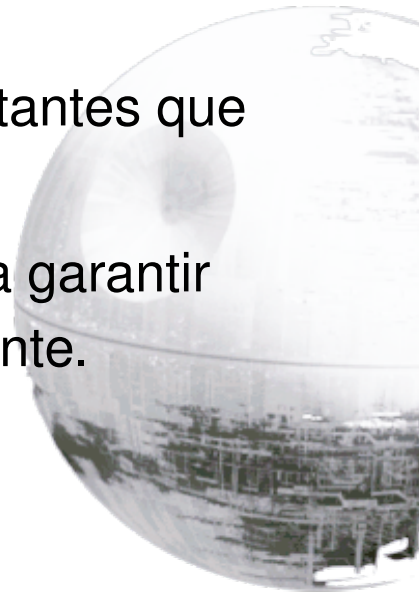
Referências sobre NER

- Nadeau, D., & Sekine, S. (2007). *A survey of named entity recognition and classification*. Lingvisticae Investigationes.
- Tjong Kim Sang, E. F., & De Meulder, F. (2003). *Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition*. Proceedings of CoNLL.



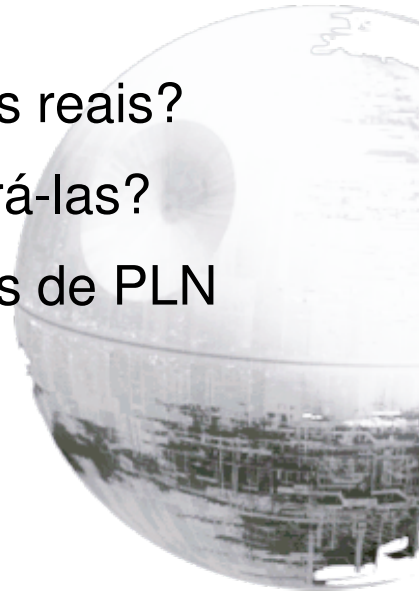
Conclusão

- O pipeline básico de PLN envolve várias etapas importantes que ajudam a estruturar e processar texto para análise.
- Cada técnica desempenha um papel fundamental para garantir que o texto seja entendido e processado adequadamente.



Discussão

- Como essas técnicas podem ser aplicadas em projetos reais?
- Quais são as limitações de cada técnica e como superá-las?
- Como combinar essas técnicas para construir sistemas de PLN mais robustos?



MAY THE

SOURCE

BE WITH

YOU