| Tópico 1

O que é o Processamento de Linguagem Natural?

> https://medium.com/botsbrasil/o-que-é-o-processamento-de-linguagem-natural-49ece9371cff

📈 **Pré-processamento textual**

Caráter morfossíntático, atuando sobre itens lexicais, i.e., palavras.

- Normalização.
- Remoção de *stopwords*.
- Remoção de numerais.
- Correção ortográfica.
- Stemização e lemtização.

What is Natural Language Processing?

> https://www.oracle.com/artificial-intelligence/what-is-natural-language-processing/#definingdb
>
> https://www.oracle.com/br/artificial-intelligence/what-is-natural-language-processing/

Processamento de Linguagem Natural (PLN) Definido

O Processamento de Linguagem Natural é um ramo da Inteligência Artificial (IA) que permite aos computadores compreender, produzir e manipular a linguagem humana.

- Compreensão da Linguagem Natural (NLU - Natural Language Understanding) — uso de computadores para compreender linguagem humana.
- Geração de Linguagem Natural (NLG - Natural Language Generation) — uso de computadores para produzir linguagem humana.
- Linguística Computacional (CL - Computational Linguistics) — campo científico que estuda aspectos computacionais da linguagem humana.

Aplicações de Processamento de Linguagem Natural

- Automatização de tarefas rotineiras.
- Aperfeiçoamento de pesquisa.
- Otimização de mecanismos de pesquisa.
- Análise e organização de grandes coleções de documentos.
- Análise de redes sociais.
- Insights de mercado.
- Moderação de conteúdo.

Setores que usam o Processamento de Linguagem Natural

- Saúde.
- Legal.
- Financeiro.
- Atendimento ao cliente.
- Seguros.

Capítulo 5 — Introdução ao Processamento de Linguagem Natural usando Python

> https://www.facom.ufu.br/~wendelmelo/terceiros/tutorial_nltk.pdf

5.2. A abordagem clássica do Processamento de Linguagem Natural

Estágios da análise no PLN

Texto → Tokenização → Análise Léxica → Análise Sintática → Análise Semântica → Análise Pragmática → **Significado pretendido do falante**.

1. Tokenização

- Segmentação de palavras.
- Quebra da sequência de caracteres de um texto por meio da identificação do limite de cada palavra.
    - Na linguística computacional, essas palavras são chamadas de *tokens*.
- Bem estabelecida para linguagens artificiais, como linguagens de programação.
    - Entretanto, em linguagens naturais, os mesmo caracteres podem ter significados diferentes e a sintaxe não é estritamente definida.

- Abordagens diferentes para:
  - Linguagens delimitadas por espaço.
    - Ambiguidade devido ao uso de sinais de pontuação.
  - Linguagens não segmentadas.
    - Necessidade de informação léxica e morfológica adicional.

2. Análise Léxica

- Análise a nível de palavras.
- Essas palavras podem ser pensadas como:
  - Uma sequência de caracteres no texto em execução.
  - Um objeto mais abstrato (lemma), i.e., o termo principal para um conjunto palavras.
- Lematização — relacionamento entre variantes morfológicas a seus *lemmas*.
  - *Lemma*: forma canônica das palavras (como são encontradas nos dicionários).
- Dois lados:
  - Mapeamento da palavra até seu *lemma* (*parsing side*).
  - Mapeamento do *lemma* para a palavra (*morphological generation* - geração morfológica).
- Na recuperação de informação, o parsing e o generation servem diferentes propósitos.
  - Stemming — operação de pré-processamento de textos em que identificam-se palavras morfologicamente complexas, então decompõe-as em seu stem invariante, i.e., forma canônica do lemma e seu afixo, por fim, deletam-se os afixos.
    - Stem: radical da palavra.

3. Análise Sintática

- Frase — unidade básica de análise de significado.

    - Expressa uma proposição, uma ideia ou um pensamento.
    - Diz algo sobre o mundo real ou imaginário.
- Questão crucial: extração de significa das frases.
    - Frases não são apenas sequências lineares de palavras.
    - Deve-se fazer uma análise fiel de cada frase, determinando suas estruturas.
        - Exemplo: na linguística generativa determina-se a estrutura sintática ou gramatical de cada frase.
- Representação da análise sintática:
    - Gramáticas.
    - Árvores sintáticas (*syntax tree* ou *parse tree*).

4. Análise Semântica

- Útil ao considerar-se textos longos.
    - Recuperação de informação.
    - Extração de informação.
    - Sumarização de textos.
    - Mineração de dados.
    - Tradução para linguagem de máquina e auxiliares de tradução.
- Também é útil em textos pequenos.
- Objetivo final: entendimento do enunciado.
    - Dependente das circunstâncias.
        - Incorporação de informações providas pelo enunciado dentro da base de conhecimento do sistema de PLN.
        - Execução de alguma ação em resposta ao enunciado.
    - Complexo, sendo dependente dos resultados das etapas anteriores(léxica e sintática).
        - Além de precisar de informações léxicas, contexto e do raciocínio comum.

- Análise do significado das palavras, expressões fixadas, sentenças inteiras e enunciados no contexto.
    - Tradução das expressões originais em um tipo de metalinguagem.
    - Evidência primária → interpretações do orador nativo sobre o uso das expressões em contexto.
- Resolução de ambiguidade.
    - Do ponto de vista da máquina, enunciados humanos estão abertos à múltiplas interpretações.
        - Ambiguidade léxica.
            - Homônimos e polissemia.
        - Ambiguidade escopal.
        - Ambiguidade referencial.

~~**5. Análise Pragmática**~~

What is NLP? (NLP video 1)

> https://www.youtube.com/watch?v=cce8ntxP_XI
>
> https://github.com/fastai/course-nlp/blob/master/1-what-is-nlp.ipynb

What is NLP

PLN é uma ampla área, englobando uma grande variedade de tarefas que incluem:

- **Marcação de classes gramaticais** — identificação da classe gramatical de cada palavra, i.e., se ela é um pronome, verbo, adjetivo, etc.
- **NER (Named Entity Recognition)** — o Reconhecimento de Entidade Nomeada consiste na identificação e categorização de informações-chave (entidades) em textos, sendo considerada uma entidade qualquer palavra ou série de palavras que se referem ao mesmo tema, e.g., nomes de pessoas, organizações, localizações, quantidades, valores monetários, etc.

    - Cada entidade detectada é classificada em uma categoria predeterminada.
- Resposta a perguntas.
- Reconhecimento da fala.
- Text-to-speech e Speech-to-text — conversão de texto em fala e de fala em texto, respectivamente.
- Modelagem de tópicos.
- Classificação de sentimento.
- Modelagem da língua.
- Tradução.

NLP techniques

Muitas técnicas de PLN mostram-se úteis em diversos campos, por exemplo, pode-se ter texto dentro de dados tabulares.

- Há técnicas que permitem alternar entre texto e imagem.

Top-down teaching approach

Será utilizada a metodologia *top-down* que em contrapartida com a abordagem *bottom-top* (estudo inicial de todos os componentes que serão usados, seguido da construção gradual de estruturas mais complexas com base nesses elementos).

Isto é, primeiro será visto o quadro geral, trabalhando com aplicações interessantes que usam macro partes como “caixas pretas” que ainda não foram explicadas, então serão estudados os detalhes de mais baixo nível debaixo de cada uma desses componentes.

- Inicialmente, foca-se em **o que** as coisas fazem, então sobre elas **são**.

NLP overview

Historicamente, PLN baseava-se em regras *hard-coded*, i.e., utilizava-se valores constantes na codificação das aplicações. Na década de 90, houve uma mudança em direção à abordagem utilizando estatística e aprendizado de máquina, mas a

complexidade de linguagem natural fazia a abordagem com estatística simples não ser considerada o estado da arte da época.

Atualmente, estamos no meio de uma grande mudança em direção ao movimento das redes neurais. Pelo fato de *deep learning* permite alcançar uma complexidade muito maior, está se tornando o estado da arte de diversas áreas.

Spell checkers

Historicamente, corretores ortográficos exigiam milhares de linhas de código *hard-coded* (mais de 2 mil linhas).

Por sua vez, uma versão que utiliza informações históricas e probabilidade pode ser escrita em pouquíssimas linhas (cerca de 17 linhas).

Um campo em contínua mudança

O PLN ainda encontra-se me um estado de mudança contínua, com suas melhores práticas sofrendo alterações.

- Pode-se ponderar em relação à remoção de *stopwords*, visto que há casos em que a acurácia de modelos aumenta ao pular essa etapa.

Norvig vs. Chomsky

Esse “debate” resume a tensão entre duas abordagens distintas:

- Modelagem dos mecanismos fundamentais por trás de um fenômeno → Chomsky.
- Utilização de aprendizado de máquina para predição de saídas, sem, necessariamente, entender os mecanismos responsáveis por elas → Norvig.

Essa tensão ainda está muito presente em PLN, além de diversos outros campos em que o aprendizado de máquina está sendo adotado, assim como onde há a abordagem de “inteligência artificial” em geral.

Yann LeCun vs. Chris Manning

Outra discussão interessante sobre o tópico trata sobre quanto estruturas linguísticas devem ser incorporadas em modelos de PLN.

- Incorporação de mais estruturas linguísticas em sistemas de *deep learning* → Manning.
- Capacidade de arquiteturas neurais simples mas poderosas para executar tarefas sofisticadas sem a engenharia de recursos específica a essa tarefa → *LeCun*.

NLP tools

- Regex — encontrar dados em diferentes formatos, e.g., números de telefone.
- Tokenização — separação do texto em unidades significativas.
- Incorporação de palavras.
- Álgebra linear/Matriz de decomposição.
- Redes neurais.
- Modelos ocultos de Markov.
- Árvores de análise.

NLP libraries

- nltk, spaCy e gensim — bibliotecas de PLN.
- PyText e fastText — bibliotecas especializadas.
- sklearn e fastai — bibliotecas de *machine learning* com recursos de texto.

Algumas aplicações de PLN

- Uso de deep learning em pequenos dados.
    - Baseando-se em uma base de dados de descrições de companhias, identificou-se quais descrições apresentavam baixa qualidade, i.e., usavam uma linguagem de marketing muito genérica, assim, promoviam as empresas enquanto traziam pouca informação realmente útil sobre elas.

- Categorização de documentos legais.
    - Utilizando o Universal Language Model Fine-Tuning.
- Análise de sentimentos dos tweets feitos por políticos.
    - Relação entre negatividade das postagens e o número de votos do político que a fez.

Metadata enchanced fit

- Classificação de citações de artigos.
    - Utilização de metadados (como publicação, país e fonte) junto ao texto da citação para melhorar a acurácia da classificação.
- Junção de dados estruturados e não estruturados para obter melhor acurácia.

Bias

- Viés produzido devido ao aprendizado com base em dados históricos.
    - Como os modelos se baseiam no histórico de dados, línguas que apresentam gênero em suas palavras podem acabar tendo estereótipos passados aos modelos.
    - Assim, o viés presente nos dados da língua é passado aos modelos.

Language models

- Modelos que geram texto podem apresentar viés naquilo que produzem.
    - O viés presente nos dados da língua também é passado para esse tipo de modelo.
- Possível uso desses modelos por pessoas com má intenção.
    - Exemplos: aumento do atrito entre líderes de estados; manipulação de campanhas eleitorais; incapacidade de diferenciação entre emails reais e spam.

Spam

- Modelos têm a capacidade de escrever, por exemplo, emails, com tom e escrita apropriados a contextos específicos, assim, tornam-se impossíveis de serem filtrados como spam.

forgery

- Deve-se considerar as implicações de modelos terem a capacidade de gerar textos convincentes, i.e., muito semelhantes àquilo que pessoas reais escreveriam.
    - Exemplo: competição entre humano e máquina em corridas eleitorais.