

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
Instituto de Informática
Departamento de Informática Aplicada

Aula 1: Introdução, Histórico e Modelos Baseados em Regras

Prof. Dennis Giovani Balreira



INF01221 - Tópicos Especiais em Computação XXXVI:
Processamento de Linguagem Natural



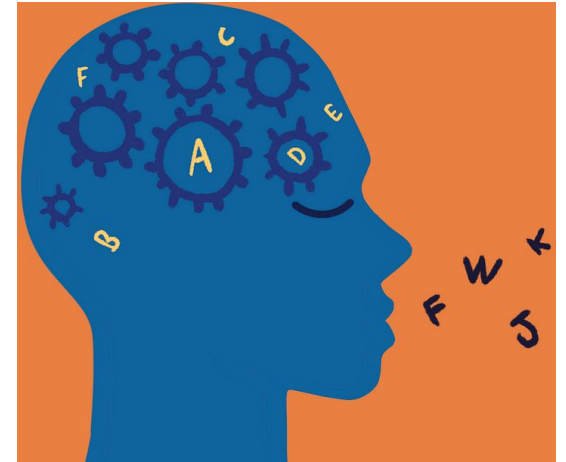
Conteúdo

- Introdução e motivação
- Histórico
- Modelos baseados em regras
- Avanços de PLN
- Fatores de crescimento de PLN
- Aplicações

Introdução e motivação

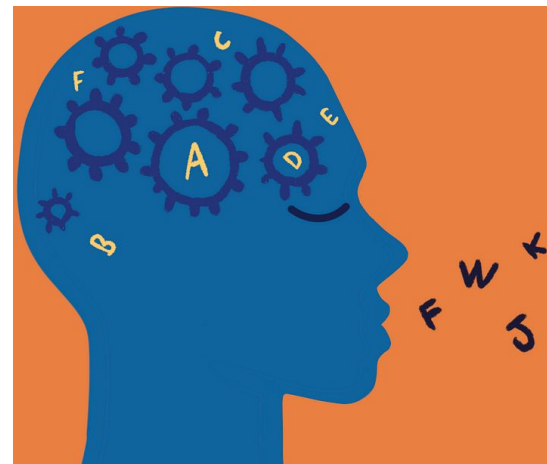
Processamento de Linguagem Natural

- O que é **Processamento de Linguagem Natural (PLN)**?
 - **Processamento:?**
 - **Linguagem Natural: ?**



Processamento de Linguagem Natural

- O que é **Processamento de Linguagem Natural (PLN)**?
 - **Processamento**: manipulação ou tratamento de informações de alguma forma
 - **Linguagem Natural**: forma de comunicação que os seres humanos usam de maneira orgânica e espontânea para expressar pensamentos, sentimentos e ideias
- **Definição**: “Campo de estudo que se concentra na capacidade dos computadores de compreender, interpretar e gerar a linguagem que os seres humanos usam naturalmente, como a fala e o texto”
- A linguagem natural se desenvolveu naturalmente como
 - Texto
 - Fala (*speech*)



Processamento de Linguagem Natural (PLN)

- Quais “áreas” envolvidas?

- Computação!
- Biologia?
- História?
- Letras?

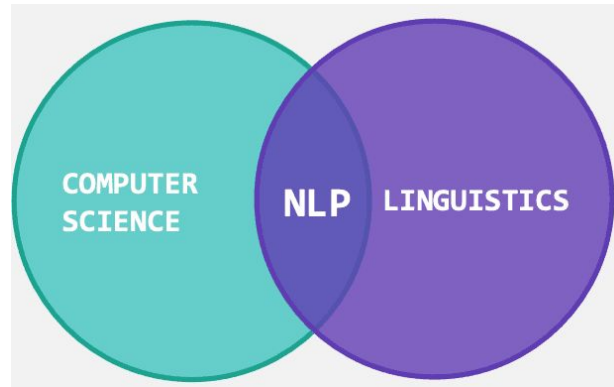
- Qual “subárea” da computação?

- Engenharia de software?
- Computação gráfica?
- Inteligência artificial?
- Computação de alto desempenho?

Processamento de Linguagem Natural (PLN)

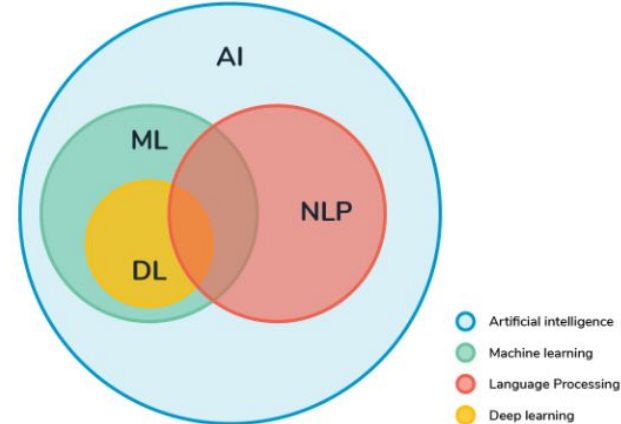
- Quais “áreas” envolvidas?

- Computação!
- Biologia?
- História?
- Letras?



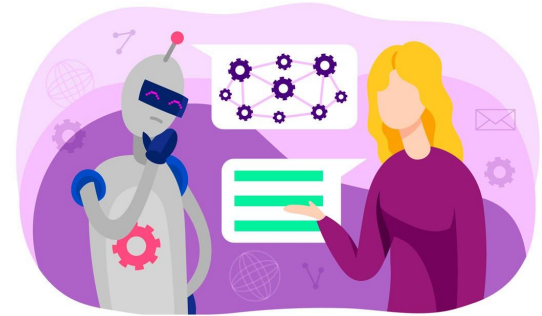
- Qual “subárea” da computação?

- Engenharia de software?
- Computação gráfica?
- Inteligência artificial?
- Computação de alto desempenho?



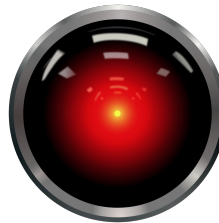
Subdivisões

- PLN possui duas subdivisões:
 - **Compreensão** (Natural Language Understanding - NLU)
 - Ex.: Classificação de texto, perguntas e respostas, reconhecimento de entidades nomeadas
 - **Geração** (Natural Language Generation - NLG)
 - Ex.: Sumarização, tradução, agentes conversacionais



Origens

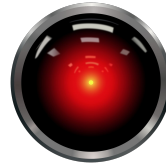
- Ideia de fazer com que computadores consigam compreender a linguagem humana
 - Desde o início da computação
- Exemplos na ficção científica:
 - HAL 9000 do filme “2001: Uma Odisseia no Espaço” de Stanley Kubrick (1968)
 - Conversa entre HAL e Bowman:
 - <https://www.youtube.com/watch?v=hL10-aNnZZ0>
- Quais habilidades do HAL 9000?



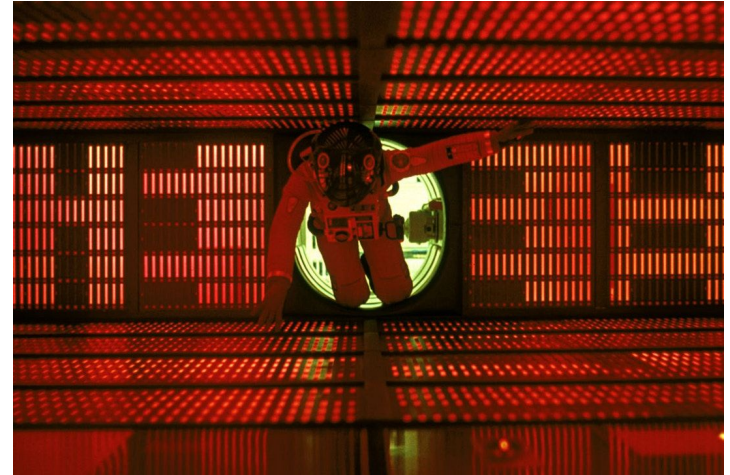
**An epic drama of
adventure and exploration**



Habilidades do HAL 9000



- Quais habilidades do HAL 9000?
- HAL = Heuristically programmed ALgorithmic computer era um agente conversacional que:
 - Compreendia a linguagem (reconhecimento de fala)
 - Gerava fala
 - Fazia inferências (tirava conclusões a partir de fatos)
 - Desambiguava sentidos de palavras
 - Conhecia morfologia (formação das palavras)
 - Sumarizava informações de diversas fontes
 - Reconhecia imagens
 - Dominava fonética e fonologia
 - Tinha conhecimento pragmático (contexto)
 - Demonstrava sentimentos
 - Fazia leitura labial



Teste de Turing

- A capacidade dos computadores de processar a linguagem com a mesma habilidade dos humanos pode sinalizar a chegada de máquinas verdadeiramente inteligentes
 - O “uso da linguagem” é humanamente fundamental
- **Alan Turing** propôs em 1950 o Teste de Turing (*imitation game*) que avalia se a máquina consegue ser indistinguível (equivalente) ao ser humano

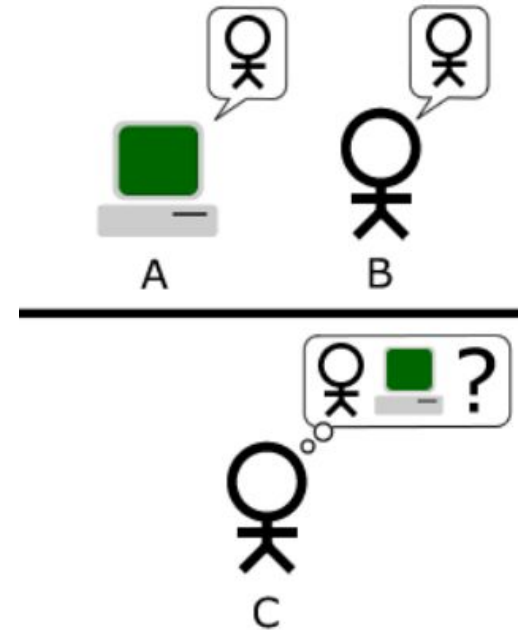


Alan Turing

https://en.wikipedia.org/wiki/Turing_test

Teste de Turing

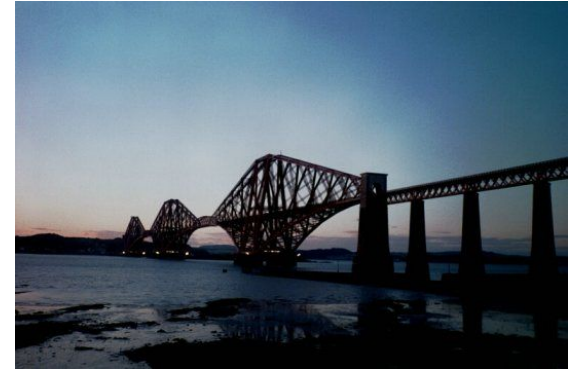
- O teste supõe três participantes e duas salas
 - Sala 1:
 - Máquina (A)
 - Pessoa (B)
 - Sala 2:
 - Juiz (C)
- O juiz (C):
 - Pode fazer perguntas para A e para B
 - Vence se acertar quem é a máquina entre A e B
- A máquina A precisa enganar o juiz dizendo ser pessoa
- A pessoa B precisa convencer o juiz que a máquina é A



https://en.wikipedia.org/wiki/Turing_test

Teste de Turing

- Turing projetou perguntas como:
 - 1. P: Por favor, escreva-me um soneto sobre o tema da Ponte Forth.
R: Não conte comigo. Nunca consegui escrever poesia.
 - 2. P: Some 34957 e 70764.
R: (após uma pausa de 30 segundos, responde) 105621.



Chatbot ELIZA (1966)

- Sistema que simulava um psicoterapeuta
- Baseado apenas em casamento de padrões (sistemas de regras)
 - Ex: “Eu preciso de X” → “O que significaria para você se tivesse X?”

Usuário: Estou infeliz.
ELIZA: VOCÊ ACHA QUE VIR AQUI A AJUDARÁ A NÃO SER INFELIZ
Usuário: Preciso de ajuda, isso parece certo.
ELIZA: O QUE SIGNIFICARIA PARA VOCÊ SE VOCÊ RECEBESSE ALGUMA AJUDA
Usuário: Talvez eu pudesse aprender a me dar bem com minha mãe.
ELIZA: CONTE-ME MAIS SOBRE SUA FAMÍLIA
Usuário: Minha mãe cuida de mim.
ELIZA: QUEM MAIS EM VOCÊ A FAMÍLIA CUIDA DE VOCÊ
Usuário: Meu pai.
ELIZA: SEU PAI
Usuário: Você é como meu pai em alguns aspectos.

```
Welcome to
      EEEEE LL   IIII ZZZZZZZ AAAA
      EE   LL   II   ZZ   AA  AA
      EEEEE LL   II   ZZZ  AAAAAA
      EE   LL   II   ZZ   AA  AA
      EEEEE LLLLL IIII ZZZZZZZ AA  AA

Eliza is a mock Rogerian psychotherapist.
The original program was described by Joseph Weizenbaum in 1966.
This implementation by Norbert Landsteiner 2005.

ELIZA: Is something troubling you ?
YOU:   Men are all alike.
ELIZA: What is the connection, do you suppose ?
YOU:   They're always bugging us about something or other.
ELIZA: Can you think of a specific example ?
```

Weizenbaum, J. 1966. ELIZA – A computer program for the study of natural language communication between man and machine. CACM, 9(1):36–45.

Interação com simulador:

<https://web.njit.edu/~ronkowit/eliza.html>

Chatbot Eugene (2014)

- Sistema que simulava um menino ucraniano de 13 anos
 - Por ser estrangeiro e falar em inglês, erros de gramática ou frases sem sentido eram justificadas
- Convenceu 10 de 30 juízes de que era um menino e não um computador



<https://g1.globo.com/tecnologia/noticia/2014/06/computador-convence-juizes-que-e-garoto-de-13-anos-em-teste-de-turing.html>

Judge: what music do you like?

Eugene: I like Eminem – because he doesn't have a thing to do with "polite correctness".

Judge: OMG i love eminem, he's my favourite artist!

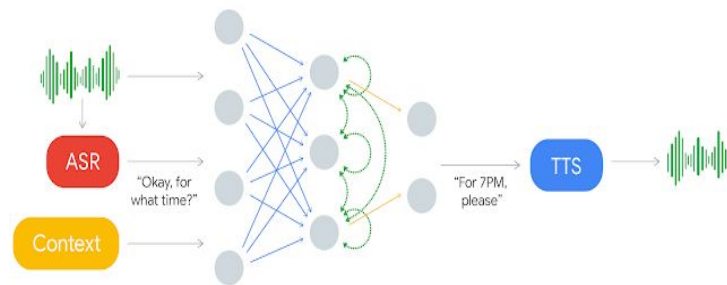
Eugene: I wonder what Eminem's nick mean. I associate it only with a word "enema". But he is a cool guy, I like his The Real Slim Shady.

Judge: i know what his nickname means, his real name is marshall Mathers. From the initials M&M phynetically sounding like Eminem.

Eugene: Eminem is a famous crapper. No, raper... ripper... Yeah, I mean – rapper. I like his Stan and The Real Slim Shady.

Google Duplex Voice AI (2018)

- Tecnologia que permite fazer chamadas telefônicas a fim de realizar tarefas práticas
 - Soa muito natural!
 - Entonação
 - Disfluências (hum, uh, etc)
 - Consegue compreender falas rápidas e com sotaques
- Sistema baseado em
 - Deep learning (Redes Neurais Recorrentes)
 - Reconhecimento de fala
 - Text to Speech (TTS)
- Ainda não disponível em português



<https://research.google/blog/google-duplex-an-ai-system-for-accomplishing-real-world-tasks-over-the-phone/>

Vídeo de demonstração:

<https://www.youtube.com/watch?v=D5VN56jQMWM>

Assistentes de voz (softwares e dispositivos)

- Google assistant (Google)
- Alexa (Amazon)
- Cortana (Windows)
- Siri (Apple)
- Bixby (Samsung)



Voice assistant comparison (video):

<https://www.youtube.com/watch?v=3gHgLiJFyw>

Google LaMDA (2021)

- LaMDA: Language Model for Dialogue Applications
 - Seu objetivo é se comunicar com humanos de maneira tão natural quanto possível
- Pré-treinada com 1,5 trilhões de palavras
- Ajustada (fine-tuned) em tarefas de geração de texto
- Gerou polêmica em 2022 quando um funcionário da Google disse que a tecnologia tinha “ganhado vida”*

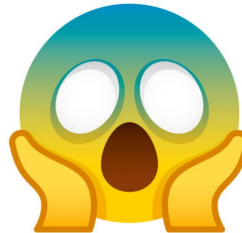
* <https://g1.globo.com/mundo/noticia/2022/06/17/como-funciona-o-lamda-cerebro-artificial-do-google-acusado-por-engenheiro-de-ter-consciencia-propria.ghtml>



<https://ai.googleblog.com/2022/01/lamdatowards-safe-grounded-and-high.html>

ChatGPT – Open AI (2022)

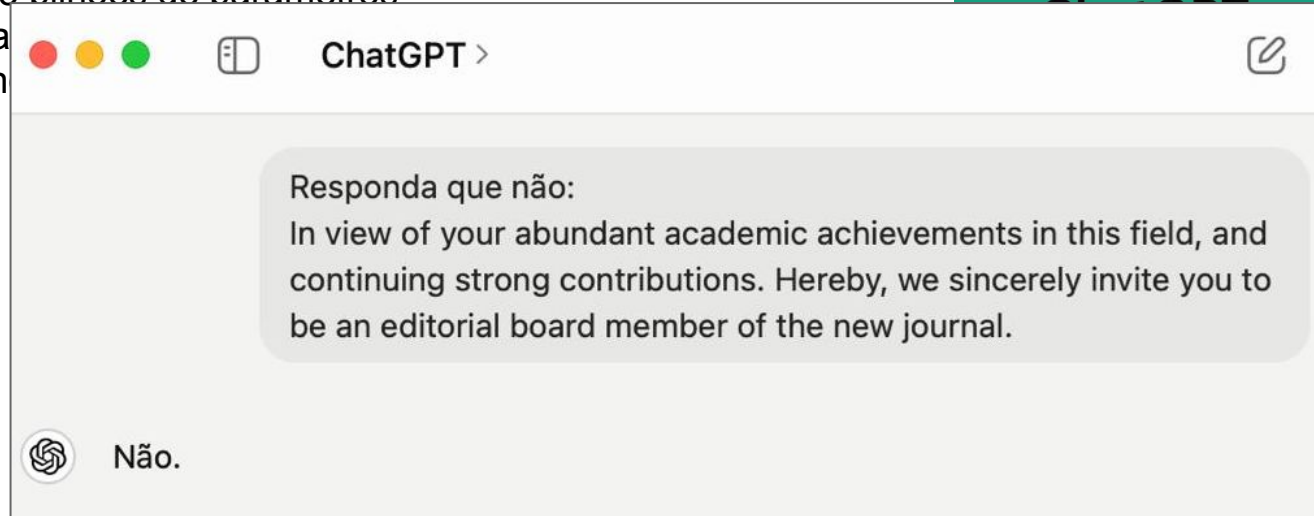
- Baseado no GPT-3.5 (GPT 3, ajustado para não gerar conteúdo tóxico)
 - GPT 3 foi pré-treinado com 410 bilhões de tokens e tem 175 bilhões de parâmetros
 - Otimizado para gerar diálogos usando Reinforcement Learning with Human Feedback (RLHF)
- “O ChatGPT dominará o mundo!”



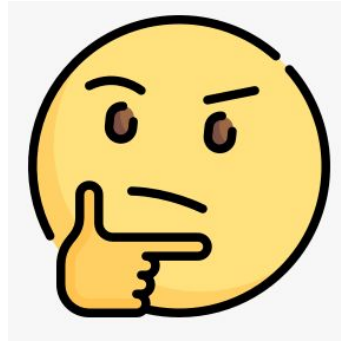
<https://openai.com/blog/chatgpt/>

ChatGPT – Open AI (2022)

- Baseado no GPT-3.5 (GPT 3, ajustado para não gerar conteúdo tóxico)
 - GPT 3 foi pré-treinado com 410 bilhões de tokens e tem 175 bilhões de parâmetros
 - Otimiza Learning
- O ChatGPT:



Passamos no Turing Test?



<https://bobhannahbob1.medium.com/how-and-why-chatgpt-failed-the-turing-test-a7f3e5c941d7>

Passamos no Turing Test?

- Considerando que algumas pessoas conversaram com máquinas achando que eram humanos, sim!
- Passar-se por humano em um domínio/tarefa específica já foi feito
- Mas o teste não era exatamente esse...



<https://bobhannahbob1.medium.com/how-and-why-chatgpt-failed-the-turing-test-a7f3e5c941d7>

Histórico

Histórico

1950 – 1984
Modelos baseados em
Regras

1985 – 2012
Modelos Estatísticos

2013 – 2017
Deep Learning &
Word Embeddings

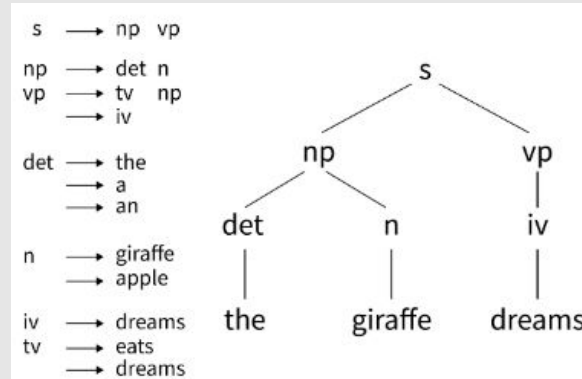
2018 – 2022
Modelos baseados
no Transformer -
NLU

2023 – ?
Modelos
Generativos
GPT e Prompting

Histórico

Sistemas que operam seguindo um conjunto predefinido de regras e padrões linguísticos para analisar, interpretar e gerar texto

1950 – 1984
Modelos baseados em Regras



1985 – 2012
Modelos Estatísticos

2023 – ?
Modelos
Generativos
GPT e Prompting

2018 – 2022
Modelos baseados
no Transformer -
NLU

2013 – 2017
Deep Learning &
Word Embeddings

Histórico

$$= P(\mathbf{x}^{(t+1)} | \overbrace{\mathbf{x}^{(t)}, \dots, \mathbf{x}^{(t-n+2)}}^{n-1 \text{ words}})$$
$$= \frac{P(\mathbf{x}^{(t+1)}, \mathbf{x}^{(t)}, \dots, \mathbf{x}^{(t-n+2)})}{P(\mathbf{x}^{(t)}, \dots, \mathbf{x}^{(t-n+2)})}$$

Sistemas que utilizam técnicas probabilísticas para aprender padrões e relações nos dados linguísticos. Envolve TF-IDF, n-grams, engenharia de features, ML tradicional (Naive Bayes, SVM, etc.) Hidden Markov Models, Conditional Random Fields

1985 – 2012
Modelos Estatísticos

1950 – 1984
Modelos baseados em
Regras

2023 – ?
Modelos
Generativos
GPT e Prompting

2018 – 2022
Modelos baseados
no Transformer -
NLU

2013 – 2017
Deep Learning &
Word Embeddings

Histórico

	living being	feline	human	gender	royalty	verb	plural
cat →	0.6	0.9	0.1	0.4	-0.7	-0.3	-0.2
kitten →	0.5	0.8	-0.1	0.2	-0.6	-0.5	-0.1
dog →	0.7	-0.1	0.4	0.3	-0.4	-0.1	-0.3
houses →	-0.8	-0.4	-0.5	0.1	-0.9	0.3	0.8

Sistemas que utilizam deep learning (redes neurais profundas capazes de aprender representações complexas e hierárquicas dos dados) e word embeddings (mapeamento de palavras para vetores de números reais capturando semântica e relacionamentos entre palavras). Envolve RNN (LSTM, GRU) Word2vec, gloVe

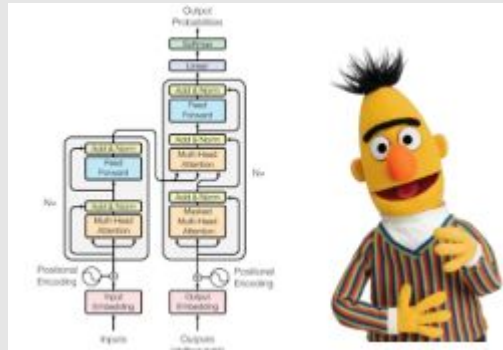
1950 – 1984
Modelos baseados em
Regras

1985 – 2012
Modelos Estatísticos

2013 – 2017
Deep Learning &
Word Embeddings

2023 – ?
Modelos
Generativos
GPT e Prompting
2018 – 2022
Modelos baseados
no Transformer -
NLU

Histórico



Sistemas baseados em transformers (modelos funcionam através do mecanismo de atenção, que permite que eles capturem relações complexas em sequências de texto), focados em Natural Language Understanding (compreensão de texto ou fala em linguagem humana). Envolve principalmente modelos como o BERT e suas variações

2023 – ?

Modelos
Generativos
GPT e Prompting

2018 – 2022

Modelos baseados
no Transformer -
NLU

2013 – 2017

Deep Learning &
Word Embeddings

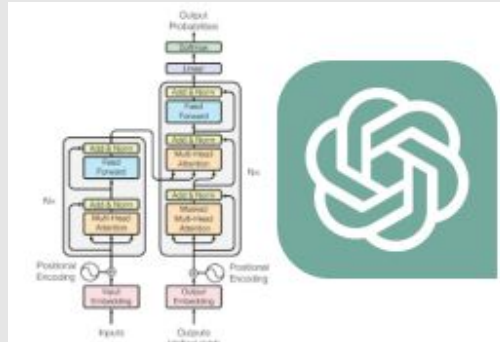
1985 – 2012

Modelos Estatísticos

1950 – 1984

Modelos baseados em
Regras

Histórico



1950 – 1984
Modelos baseados em
Regras

Sistemas baseados em transformers (modelos funcionam através do mecanismo de atenção, que permite que eles capturem relações complexas em sequências de texto), focados em Natural Language Generation (geração de texto ou fala em linguagem humana). Prompting envolve fornecer ao modelo uma entrada inicial (prompt) que orienta a geração do texto. Envolve principalmente modelos como o GPT e suas variações.

1985 – 2012
Modelos Estatísticos

2023 – ?
Modelos
Generativos
GPT e Prompting

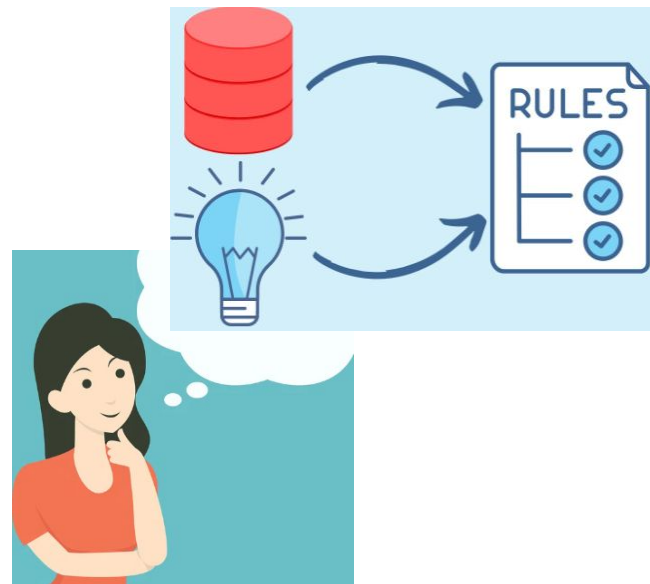
2018 – 2022
Modelos baseados
no Transformer -
NLU

2013 – 2017
Deep Learning &
Word Embeddings

Modelos baseados em regras

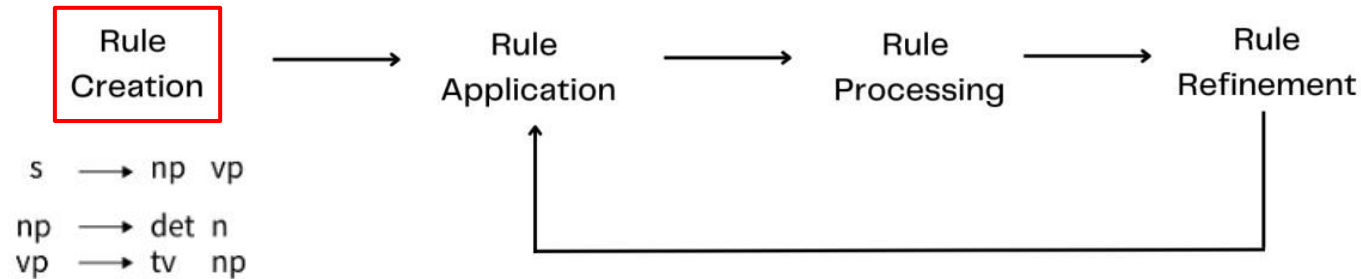
Modelos baseados em regras

- **Modelo baseado em regras** é um sistema que depende de um **conjunto de regras** para realizar uma **tarefa específica**
 - Muito comum até os anos 80
- As regras geralmente são criadas por **especialistas** humanos com conhecimento linguístico e no domínio
 - Foco em estruturas com “if-then-else”
 - Foco em programação com paradigma lógico
- Por exemplo, um modelo para **análise de sentimentos**:
 - Lista pré-definida de palavras ou frases **positivas** e **negativas** para classificar um texto em “+ ou -”
 - Pode criar uma regra que, toda vez que tiver adjetivos bons (como “feliz”) será positivo



Modelos baseados em regras

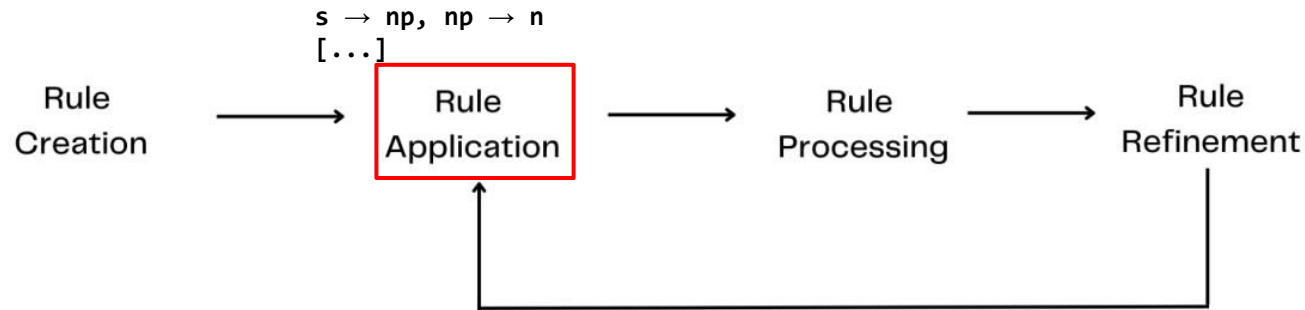
- Os modelos seguem o seguinte pipeline:



As regras são criadas, como regras gramaticais, padrões de sintaxe, regras semânticas ou expressões regulares.

Modelos baseados em regras

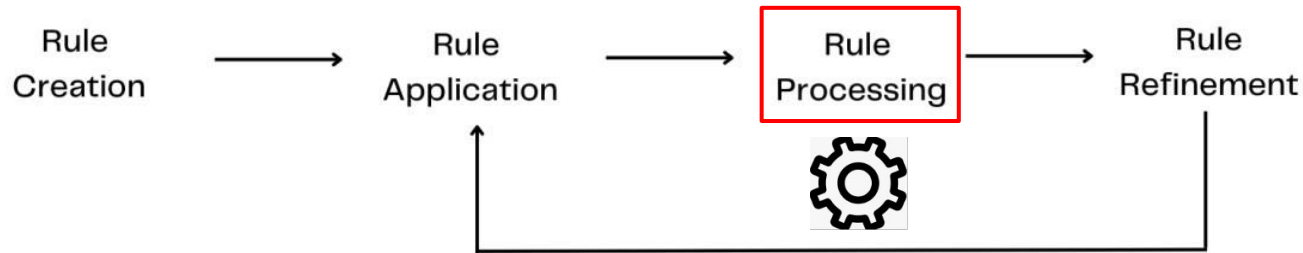
- Os modelos seguem o seguinte pipeline:



As regras predefinidas são aplicadas aos dados de entrada para capturar padrões correspondentes.

Modelos baseados em regras

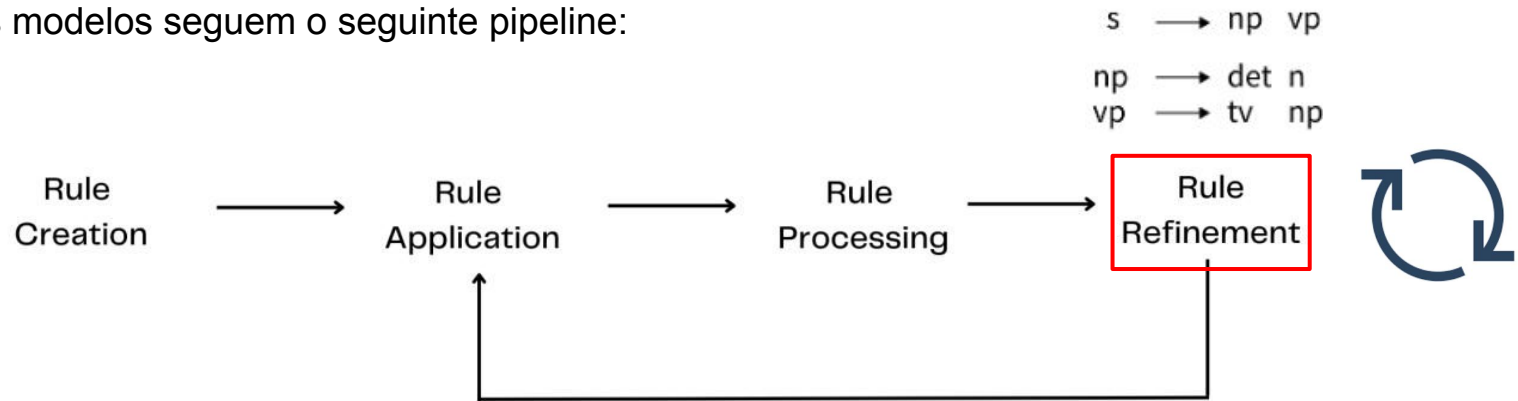
- Os modelos seguem o seguinte pipeline:



Os dados de texto são processados de acordo com os resultados das regras correspondentes para extrair informações, tomar decisões ou realizar outras tarefas.

Modelos baseados em regras

- Os modelos seguem o seguinte pipeline:



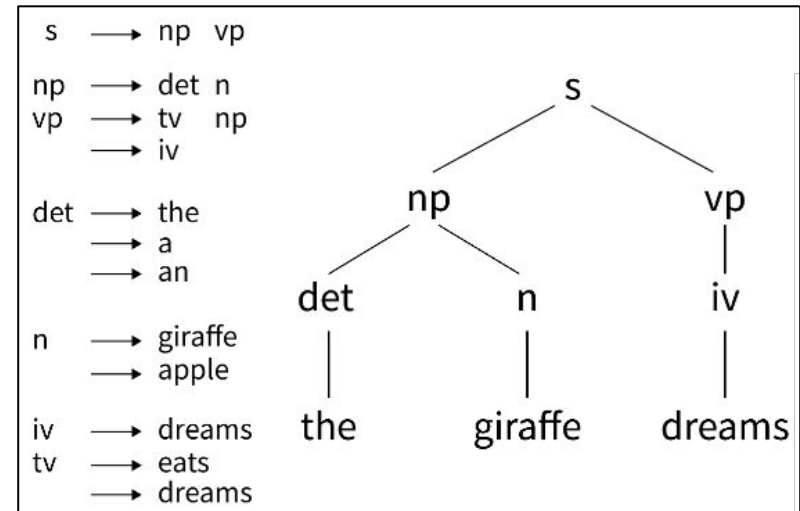
As regras criadas são refinadas iterativamente por meio de processamento repetitivo para melhorar a precisão e o desempenho. Com base no feedback anterior, as regras são modificadas e atualizadas quando necessário.

Modelos baseados em regras

- Formas principais de expressar modelos baseados em regras:
 - Gramáticas
 - Expressões Regulares
 - Regras de Substituição e Transformação

Modelos baseados em regras

- Formas principais de expressar modelos baseados em regras:
 - **Gramáticas:** formalismos que descrevem a estrutura sintática ou semântica de uma linguagem natural
 - Expressões Regulares
 - Regras de Substituição e Transformação



Modelos baseados em regras

- Formas principais de expressar modelos baseados em regras:
 - Gramáticas
 - **Expressões Regulares:** padrões de texto que descrevem conjuntos de strings de acordo com regras específicas. Elas são amplamente utilizadas para busca e manipulação de padrões em texto
 - Regras de Substituição e Transformação

Serão vistas nas próximas aulas!

União. $(r + s)$ é ER

* denota a linguagem $R \cup S$

Concatenação. (rs) é ER

* denota a linguagem $RS = \{uv \mid u \in R \text{ e } v \in S\}$

Concatenação Sucessiva. (r^*) é ER

* denota a linguagem R^*

Modelos baseados em regras

- Formas principais de expressar modelos baseados em regras:
 - Gramáticas
 - Expressões Regulares
 - **Regras de Substituição e Transformação:** descrevem como transformar ou substituir partes de texto com base em critérios específicos. São usadas em processos de normalização de texto e pré-processamento de dados

```
S(f)=substituir(f,"amigo","companheiro")
```

```
T(f)=transformar_voz(f,"ativa","passiva")
```

Exemplo: Gramáticas Livres de Contexto

- Uma gramática livre de contexto G para gerar uma linguagem L é

$G = (V, T, P, S)$, onde

V : conjunto de variáveis usadas

T : símbolos terminais

P : regras de produção

S : variável inicial

Regras de produção seguem a forma $A \rightarrow \alpha$, onde

A é variável de V (ou seja, apenas uma variável)

α é palavra de $(V \cup T)^*$

$GERA(G) = \{w \in T^* \mid S \Rightarrow^+ w\}$

MENEZES, Paulo Blauth. **Linguagens Formais e Autômatos: Volume 3 da Série Livros Didáticos Informática UFRGS**. Bookman Editora, 2009.

Exemplo: Gramáticas Livres de Contexto

- Ex. L1: subconjunto de palavras do inglês

$G1 = (V1, T1, P1, S)$, onde

$V1 = \{S, U, P, V, O, A, N\}$

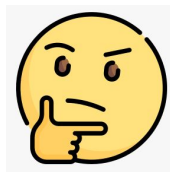
$T1 = \{a, the, boy, ball, rock, pumpkin, hit, threw, ate, girl, he, she, it, him, her, it\}$

$P1 = \{$

$S \rightarrow UP.$	$N \rightarrow boy_$	$N \rightarrow girl_$
$U \rightarrow AN$	$N \rightarrow ball_$	$U \rightarrow he_$
$P \rightarrow VO$	$N \rightarrow rock_$	$U \rightarrow she_$
$O \rightarrow AN$	$N \rightarrow pumpkin_$	$U \rightarrow it_$
$A \rightarrow a_$	$V \rightarrow hit_$	$O \rightarrow him_$
$A \rightarrow the_$	$V \rightarrow threw_$	$O \rightarrow her_$
	$V \rightarrow ate_$	$O \rightarrow it_$

$\}$

Exemplo de derivação para a frase “The boy hit a ball”:



Exemplo: Gramáticas Livres de Contexto

- Ex. L1: subconjunto de palavras do inglês

$G1 = (V1, T1, P1, S)$, onde

$V1 = \{S, U, P, V, O, A, N\}$

$T1 = \{a, the, boy, ball, rock, pumpkin, hit, threw, ate, girl, he, she, it, him, her, it\}$

$P1 = \{$

$S \rightarrow UP.$	$N \rightarrow boy_$	$N \rightarrow girl_$
$U \rightarrow AN$	$N \rightarrow ball_$	$U \rightarrow he_$
$P \rightarrow VO$	$N \rightarrow rock_$	$U \rightarrow she_$
$O \rightarrow AN$	$N \rightarrow pumpkin_$	$U \rightarrow it_$
$A \rightarrow a_$	$V \rightarrow hit_$	$O \rightarrow him_$
$A \rightarrow the_$	$V \rightarrow threw_$	$O \rightarrow her_$
	$V \rightarrow ate_$	$O \rightarrow it_$

$\}$

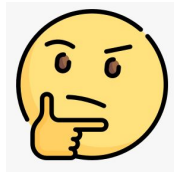
Exemplo de derivação para a frase “The boy hit a ball”:

$S \Rightarrow UP. \Rightarrow ANP. \Rightarrow the_NP. \Rightarrow the_boy_P. \Rightarrow the_boy_VO.$
 $\Rightarrow the_boy_hit_O \Rightarrow the_boy_hit_AN. \Rightarrow the_boy_hit_a_N. \Rightarrow$
 $the_boy_hit_a_ball_.$

Modelos baseados em regras: vantagens e desvantagens

- Vantagens:

- Desvantagens:



Modelos baseados em regras: vantagens e desvantagens

- **Vantagens:**
 - São explícitas e bem definidas, melhorando a confiabilidade do sistema
 - São facilmente entendidas e modificadas por humanos
 - Podem ser aplicadas sem depender de um grande número de dados, como técnicas de machine learning e deep learning
 - Regras permitem cobrir casos novos ou mais raros
- **Desvantagens:**
 - Difícil de escalar e manter, pois regras possuem muitos casos de exceção
 - Difícil (e demorado) de criar, testar e “debugar” as regras
 - Costuma refletir uma visão tendenciosa dos criadores das regras
 - Não conseguem lidar com ambiguidade e variabilidade das linguagens

Modelos baseados em regras: ainda usados hoje?

- Os modelos baseados em regras foram “estado-da-arte” por **várias décadas**
- Entretanto, perdeu relevância com o avanço de abordagens de **aprendizado de máquina e redes neurais**
- Ainda são relevantes para:
 - Casos que podem ser resolvidos em grande parte com modelos baseados em regras de **forma mais eficiente** do que com modelos complexos
 - Servem como **baselines** eficazes contra as quais o desempenho de modelos mais sofisticados pode ser avaliado

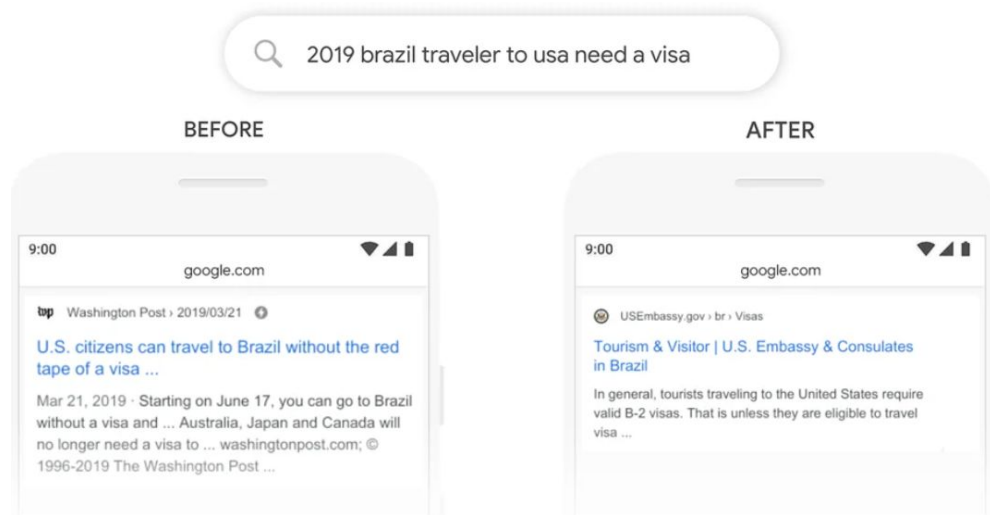
Avanços de PLN

Avanços de PLN

- PLN tem sido aplicado com sucesso em várias áreas, melhorando a eficiência, acessibilidade e personalização de serviços e produtos
- Algumas áreas de avanços:
 - Motores de busca
 - Tradução automática
 - Sugestão de escrita
 - Assistentes pessoais
 - Assistência à codificação
 - Moderação de conteúdo
 - ...

Avanços de PLN

- PLN tem sido aplicado com sucesso em várias áreas, melhorando a eficiência, acessibilidade e personalização de serviços e produtos
- Algumas áreas de avanços:
 - Motores de busca
 - Tradução automática
 - Sugestão de escrita
 - Assistentes pessoais
 - Assistência à codificação
 - Moderação de conteúdo
 - ...



<https://blog.google/products/search/search-language-understanding-bert/>

Avanços de PLN

- PLN tem sido aplicado com sucesso em várias áreas, melhorando a eficiência, acessibilidade e personalização de serviços e produtos
- Algumas áreas de avanços:
 - Motores de busca
 - Tradução automática
 - Sugestão de escrita
 - Assistentes pessoais
 - Assistência à codificação
 - Moderação de conteúdo
 - ...

Original em inglês	Tradução SYSTRAN (2003)
the reaction was skeptical	a reação era skeptical
Michael Jackson's fans	Os ventiladores de Michael Jackson
The Princess of Rock	A princesa da rocha
When the man was late picking Tikriti up	quando o homem era colheita atrasada Tikriti acima
Soviet rule	régua soviética
Kristen Pfaff found dead	Kristen Pfaff encontrado inoperante
United States and Britain	estados e Grã Bretanha unidos
make out	fazer para fora
New York	iork novo
Iraq back to square 1 on sanctions	Traseiro de Iraq ao quadrado 1 em sanctions

Avanços de PLN

- PLN tem sido aplicado com sucesso em várias áreas, melhorando a eficiência, acessibilidade e personalização de serviços e produtos
- Algumas áreas de avanços:
 - Motores de busca
 - Tradução automática
 - Sugestão de escrita
 - Assistentes pessoais
 - Assistência à codificação
 - Moderação de conteúdo
 - ...

Original em inglês	GOOGLE TRANSLATE (2024)
the reaction was skeptical	a reação foi cética
Michael Jackson's fans	fãs de michael jackson
The Princess of Rock	a princesa do rock
When the man was late picking Tikriti up	quando o homem se atrasou para pegar Tikriti
Soviet rule	regra soviética
Kristen Pfaff found dead	Kristen Pfaff é encontrada morta
United States and Britain make out	Estados Unidos e Grã-Bretanha decifrar
New York	Nova Iorque
Iraq back to square 1 on sanctions	Iraque volta à estaca zero em sanções

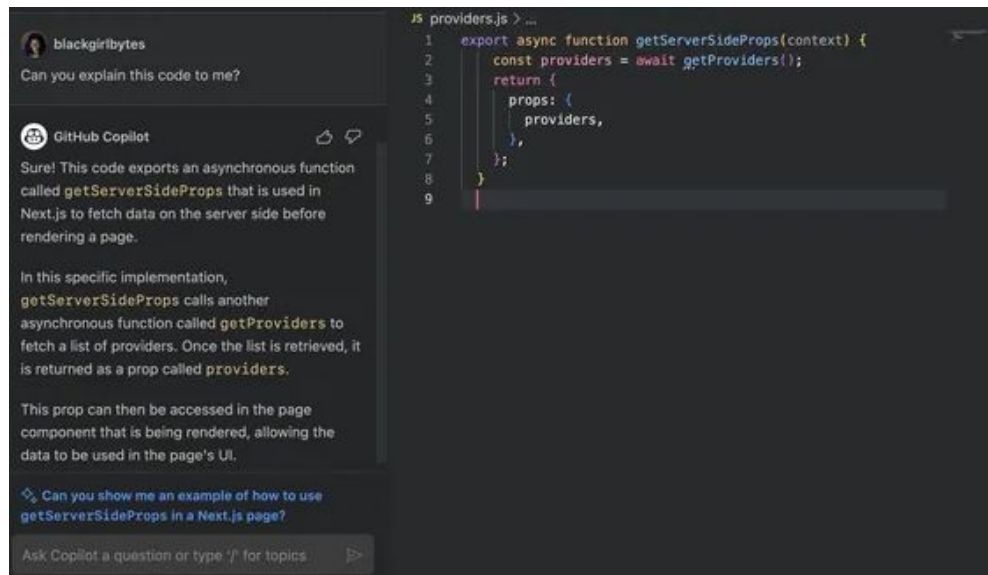
Avanços de PLN

- PLN tem sido aplicado com sucesso em várias áreas, melhorando a eficiência, acessibilidade e personalização de serviços e produtos
- Algumas áreas de avanços:
 - Motores de busca
 - Tradução automática
 - Sugestão de escrita
 - **Assistentes pessoais**
 - Assistência à codificação
 - Moderação de conteúdo
 - ...



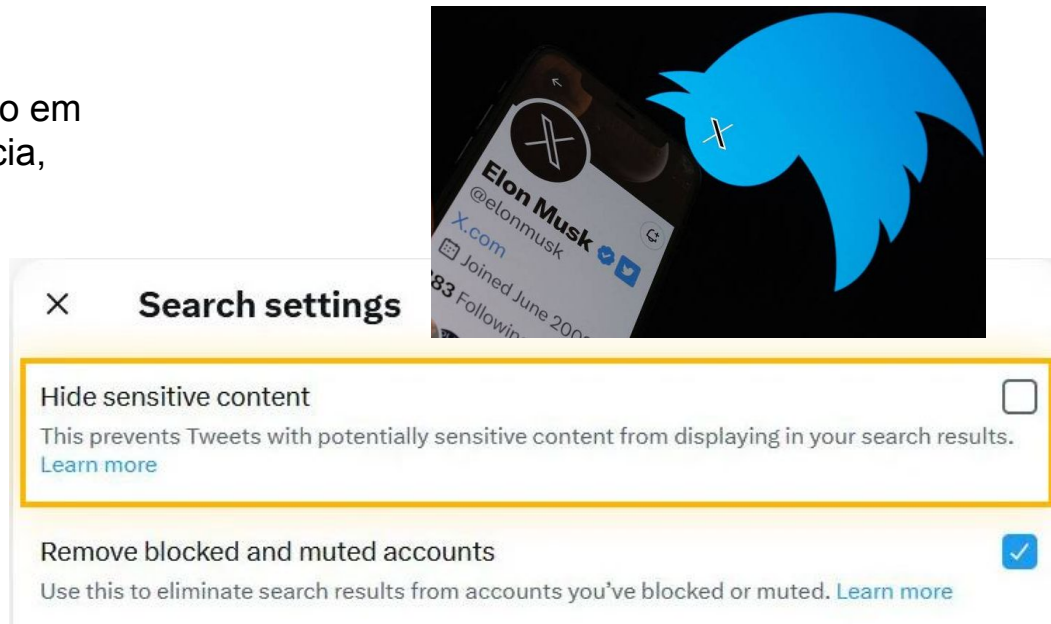
Avanços de PLN

- PLN tem sido aplicado com sucesso em várias áreas, melhorando a eficiência, acessibilidade e personalização de serviços e produtos
- Algumas áreas de avanços:
 - Motores de busca
 - Tradução automática
 - Sugestão de escrita
 - Assistentes pessoais
 - **Assistência à codificação**
 - Moderação de conteúdo
 - ...



Avanços de PLN

- PLN tem sido aplicado com sucesso em várias áreas, melhorando a eficiência, acessibilidade e personalização de serviços e produtos
- Algumas áreas de avanços:
 - Motores de busca
 - Tradução automática
 - Sugestão de escrita
 - Assistentes pessoais
 - Assistência à codificação
 - **Moderação de conteúdo**
 - ...



Fatores de crescimento de PLN: 7 motivos

Fatores de crescimento de PLN: 7 motivos

- **1. Grande disponibilidade de corpora**

- Em inglês e multilíngue:
 - Wikipedia (200 idiomas) [1]
 - Common Crawl (petabytes de dados, 40 idiomas) [2]
 - Colossal Clean Crawled (365 milhões de documentos) [3]
 - BookCorpus (11 mil livros) [4]
- Em português:
 - brWac (3,5 milhões de documentos) [5]
 - CHAVE (dados de jornais, artigos científicos, etc.) [6]



[1] <https://dumps.wikimedia.org/backup-index.html>

[2] <https://commoncrawl.org/>

[3] <https://github.com/allenai/c4-documentation>

[4] <https://github.com/soskek/bookcorpus>

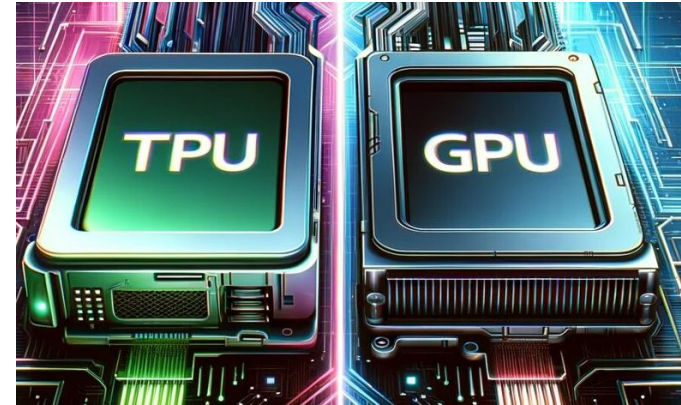
[5] <https://www.inf.ufrgs.br/pln/wiki/index.php?title=BrWaC>

[6] <https://www.linguateca.pt/CHAVE/>

Fatores de crescimento de PLN: 7 motivos

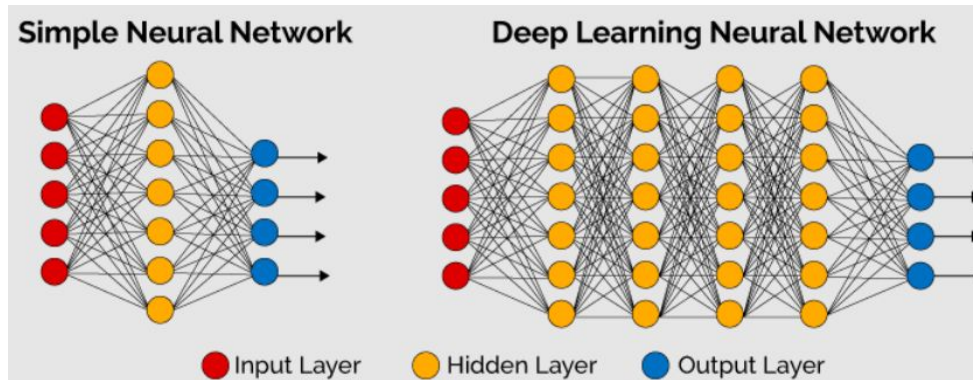
- **2. Avanços em poder computacional com GPUs e TPUs**

- Graphics Processing Units (GPUs) são usadas há vários anos para acelerar a renderização de imagens
- Passaram a ser usadas para problemas que requerem computação intensiva
- Operações por ciclo
 - CPU – dezenas
 - GPU – dezenas de milhares
 - TPU – 128 mil
 - Desenvolvidas para aprendizado de máquina com TensorFlow
- Grande capacidade de processamento paralelo



Fatores de crescimento de PLN: 7 motivos

- **3. Desenvolvimento de algoritmos de deep learning**
 - Algoritmos de deep learning revolucionaram PLN nos últimos anos, impulsionando um progresso significativo em diversas áreas



Pioneiros do DL: Yann LeCun, Geoffrey Hinton, Yoshua Bengio e Andrew Ng
<https://www.deeplearningbook.com.br/o-que-sao-redes-neurais-artificiais-prof-ondas/>

Fatores de crescimento de PLN: 7 motivos

- **4. Disponibilização de código**
 - Ferramentas de versionamento, controle de código-fonte, e um ambiente colaborativo onde desenvolvedores podem compartilhar, revisar e melhorar projetos de PLN de forma aberta e eficiente

**Over 100 million developers
call GitHub home³**

Whether you're scaling your startup or just learning how to code, GitHub is your home. Join the world's largest developer platform to build the innovations that empower humanity. Let's build from here.



Fatores de crescimento de PLN: 7 motivos

- **5. Acesso rápido e gratuito a artigos científicos**
 - ACL Anthology: artigos das melhores conferências de PLN disponíveis gratuitamente
 - arXiv: repositório online de acesso aberto para pré-publicações de artigos científicos
 - Não passam por revisão antes da publicação!



arXiv is a free distribution service and an open-access archive for nearly 2.4 million scholarly articles in the fields of physics, mathematics, computer science, quantitative biology, quantitative finance, statistics, electrical engineering and systems science, and economics. Materials on this site are not peer-reviewed by arXiv.

Subject search and browse:

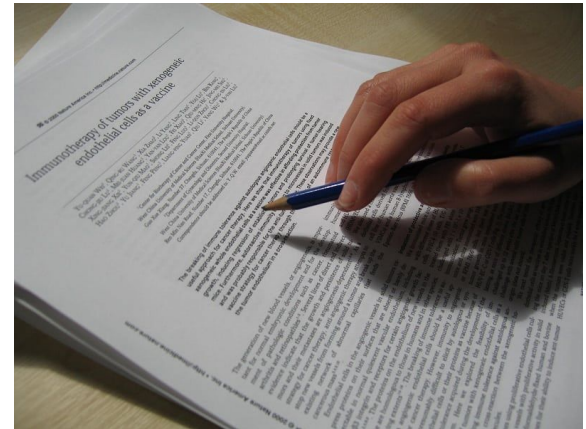
Physics



Search

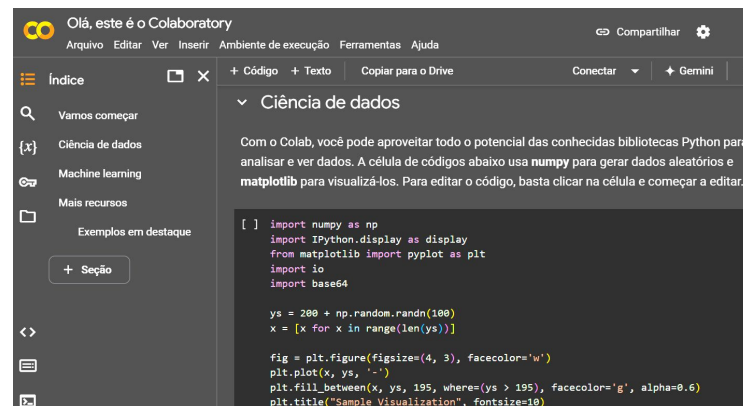
Form Interface

Catchup



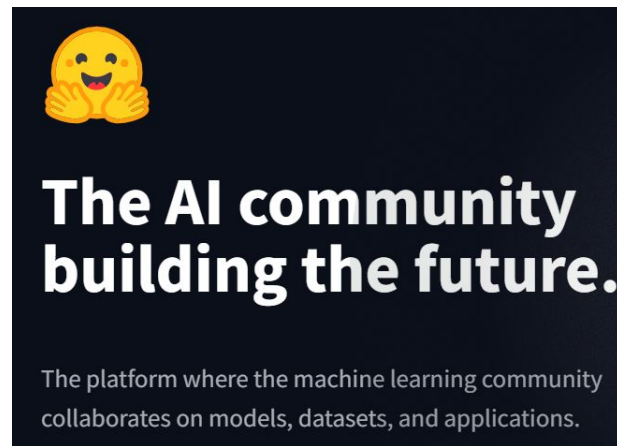
Fatores de crescimento de PLN: 7 motivos

- **6. Ambiente de desenvolvimento de código via browser**
 - Google Colab: facilitou a experimentação e desenvolvimento de projetos de Processamento de Linguagem Natural ao oferecer um ambiente gratuito e acessível para execução de código Python em notebooks Jupyter na nuvem



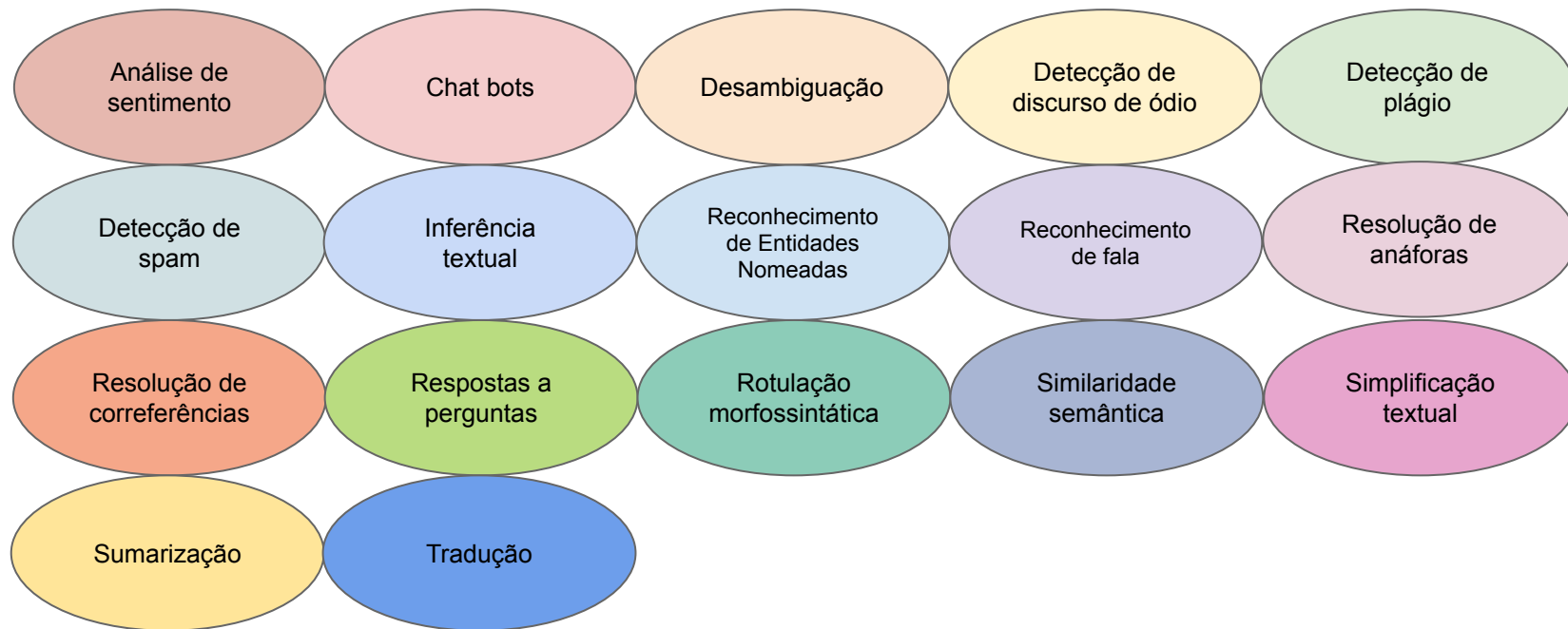
Fatores de crescimento de PLN: 7 motivos

- **7. Grandes modelos de linguagem pré-treinados disponíveis**
 - Fornecimento de bibliotecas e modelos pré-treinados de fácil acesso, como Transformers, facilitando a implementação e experimentação de avançadas técnicas de PLN
 - Hugging Face:
 - Fundada em 2016 com foco em chatbots
 - Tem uma grande comunidade de software livre
 - Objetivo – democratizar o acesso ao PLN
 - Disponibiliza milhares de datasets e modelos (Transformers) pré-treinados
 - Em maio de 2021 as bibliotecas de Transformers já tinham mais de 1 milhão de downloads

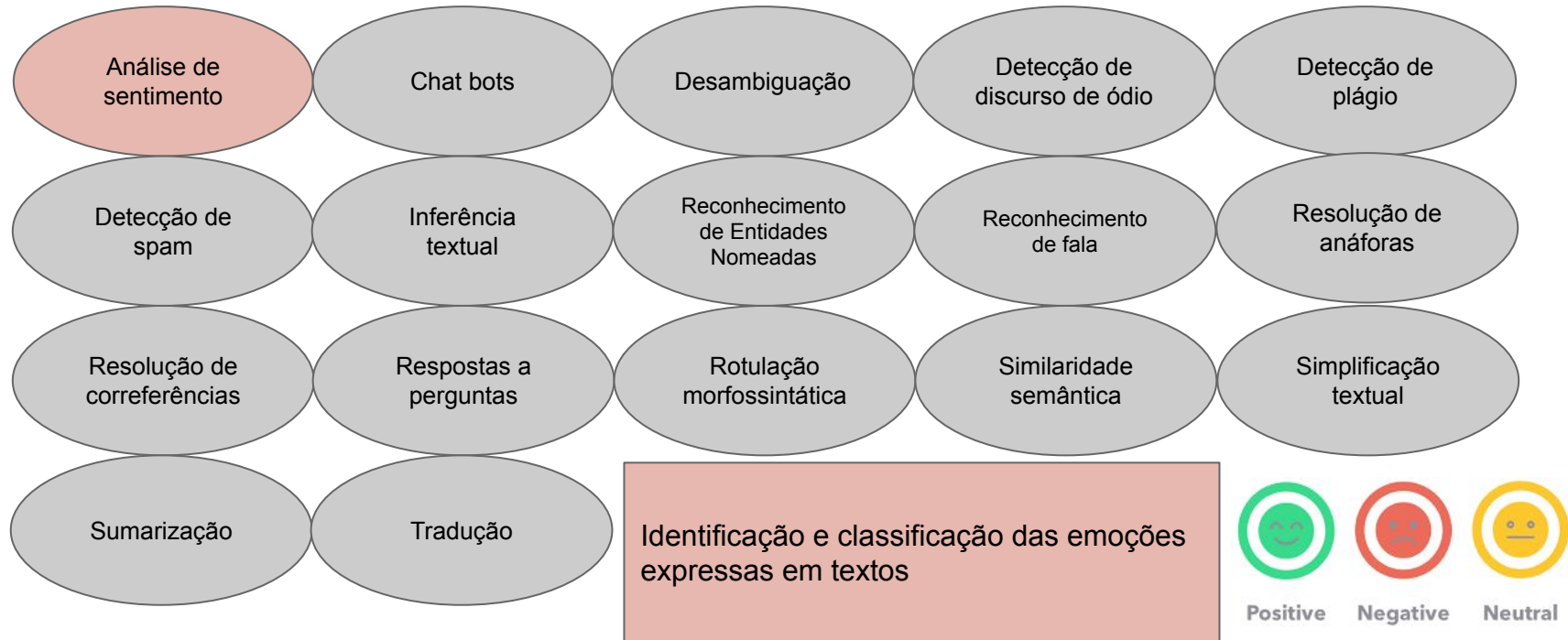


Aplicações

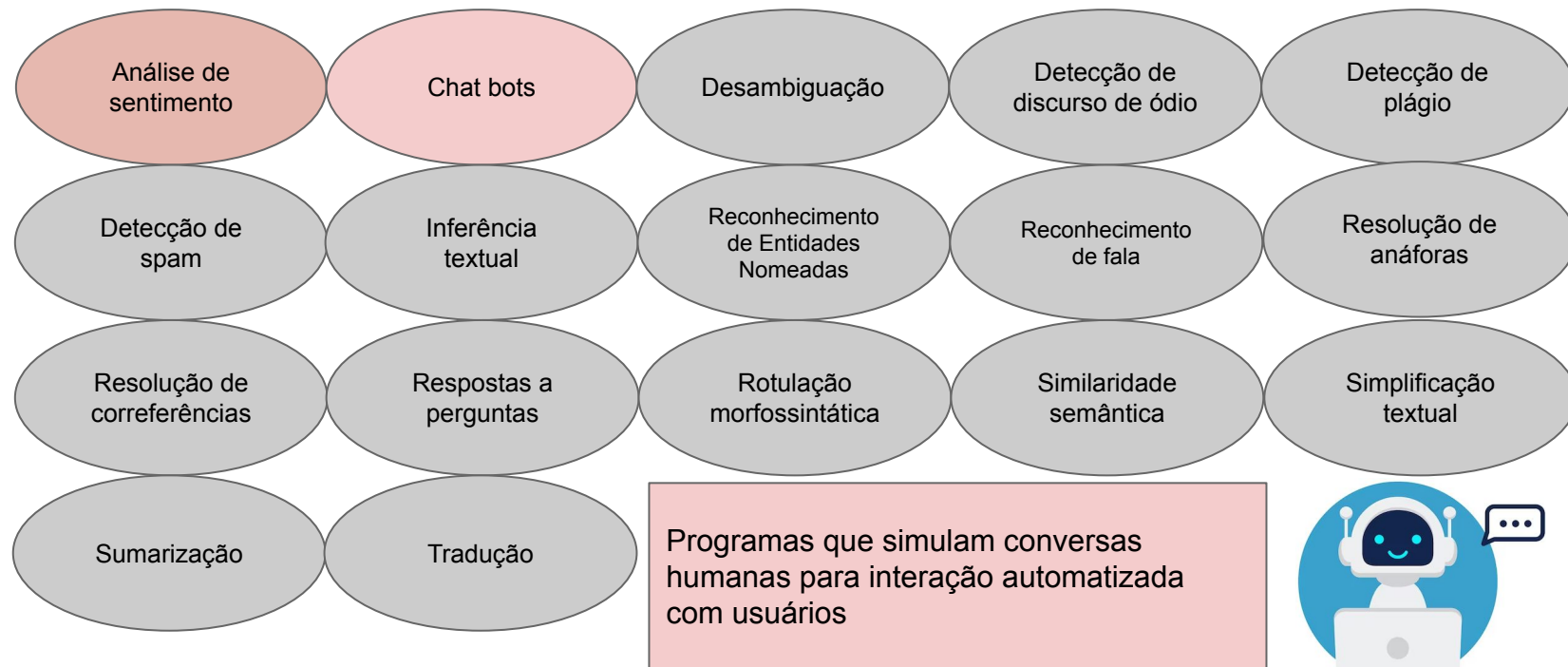
Aplicações



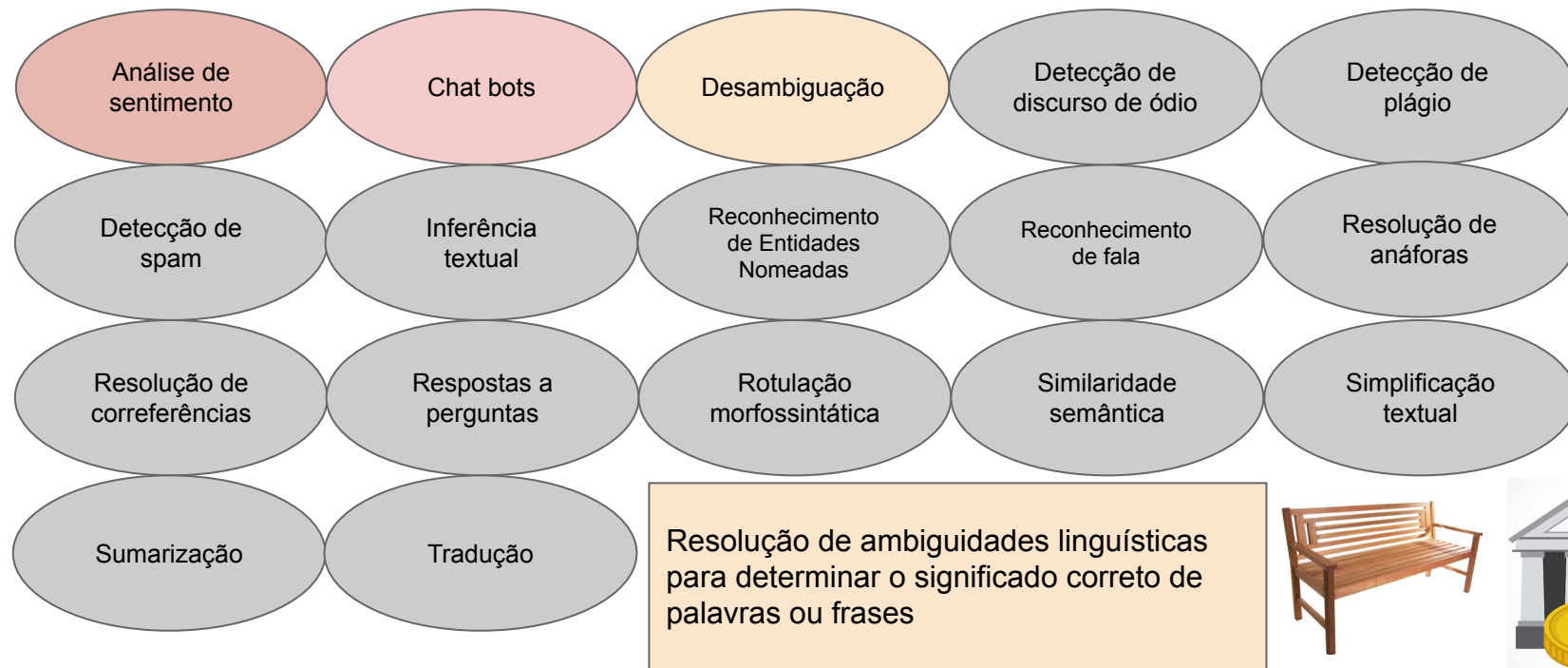
Aplicações



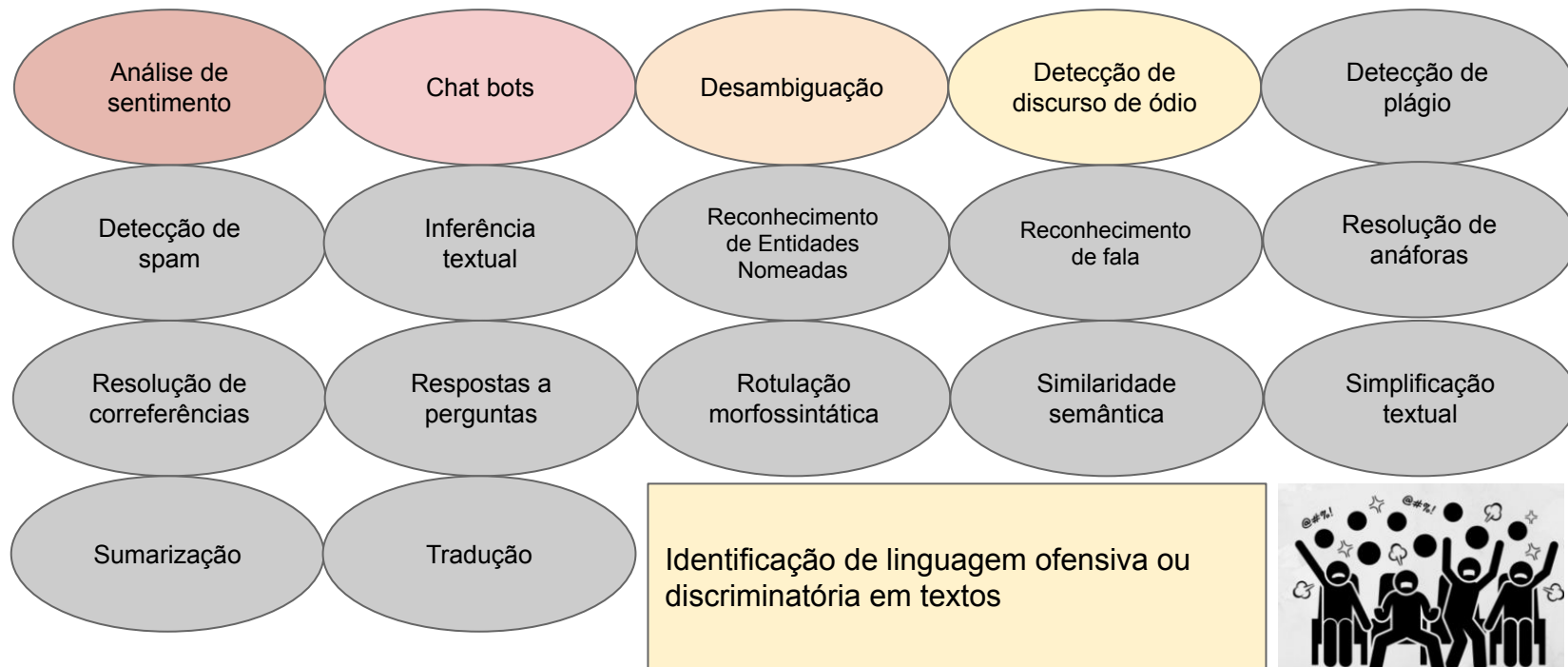
Aplicações



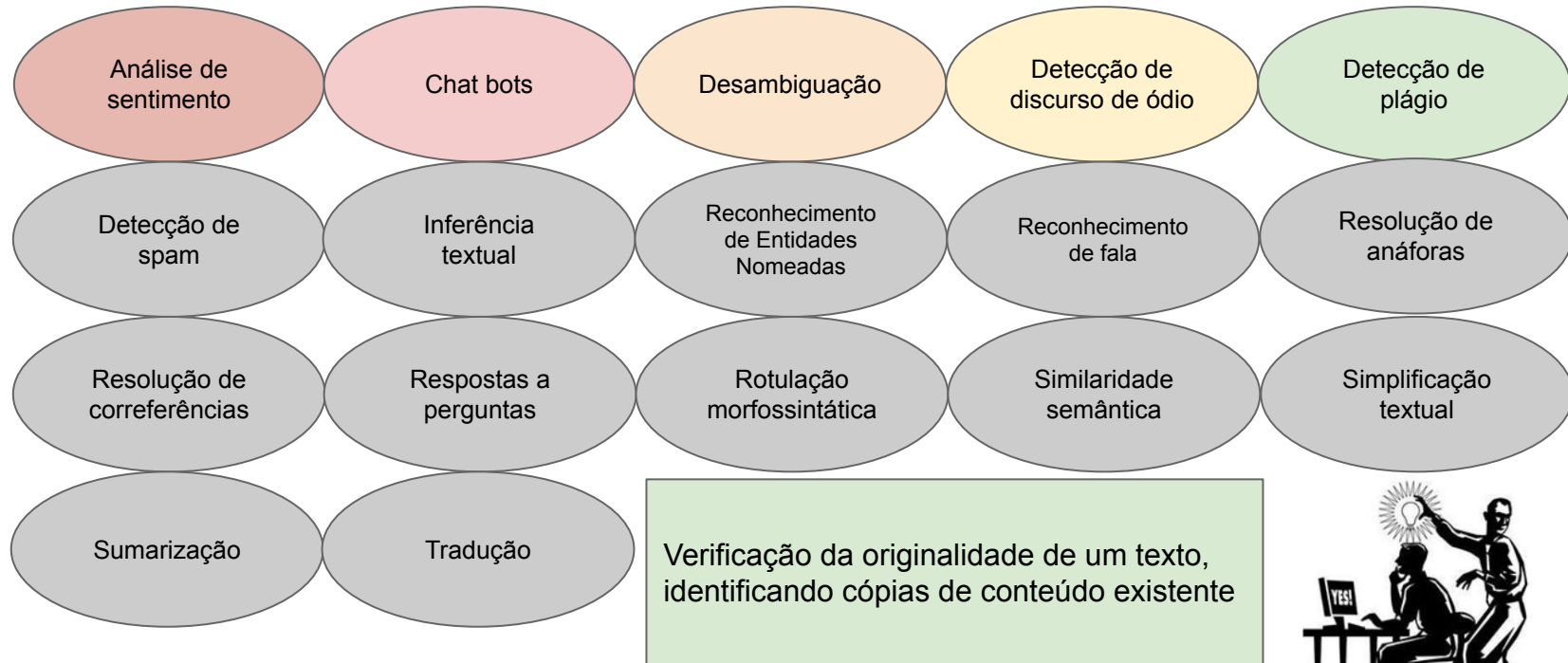
Aplicações



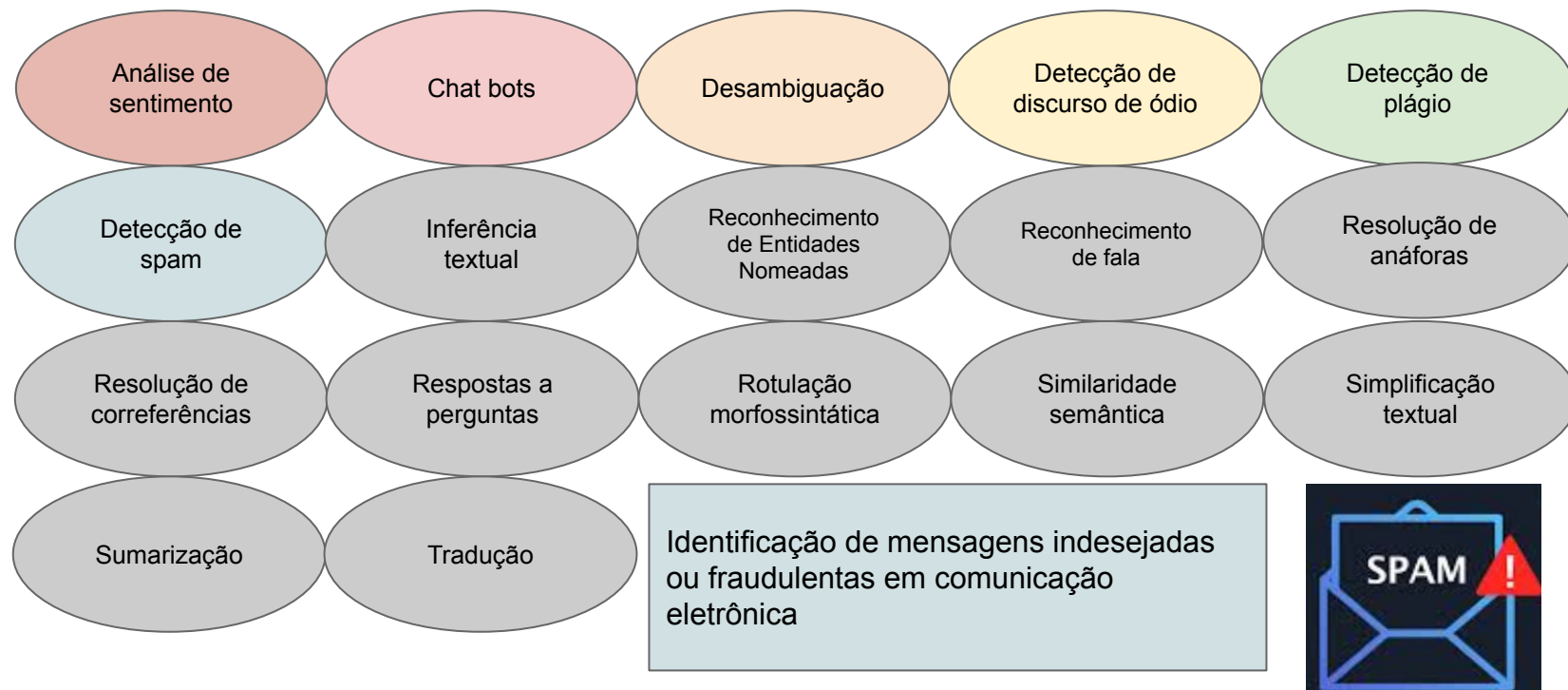
Aplicações



Aplicações



Aplicações



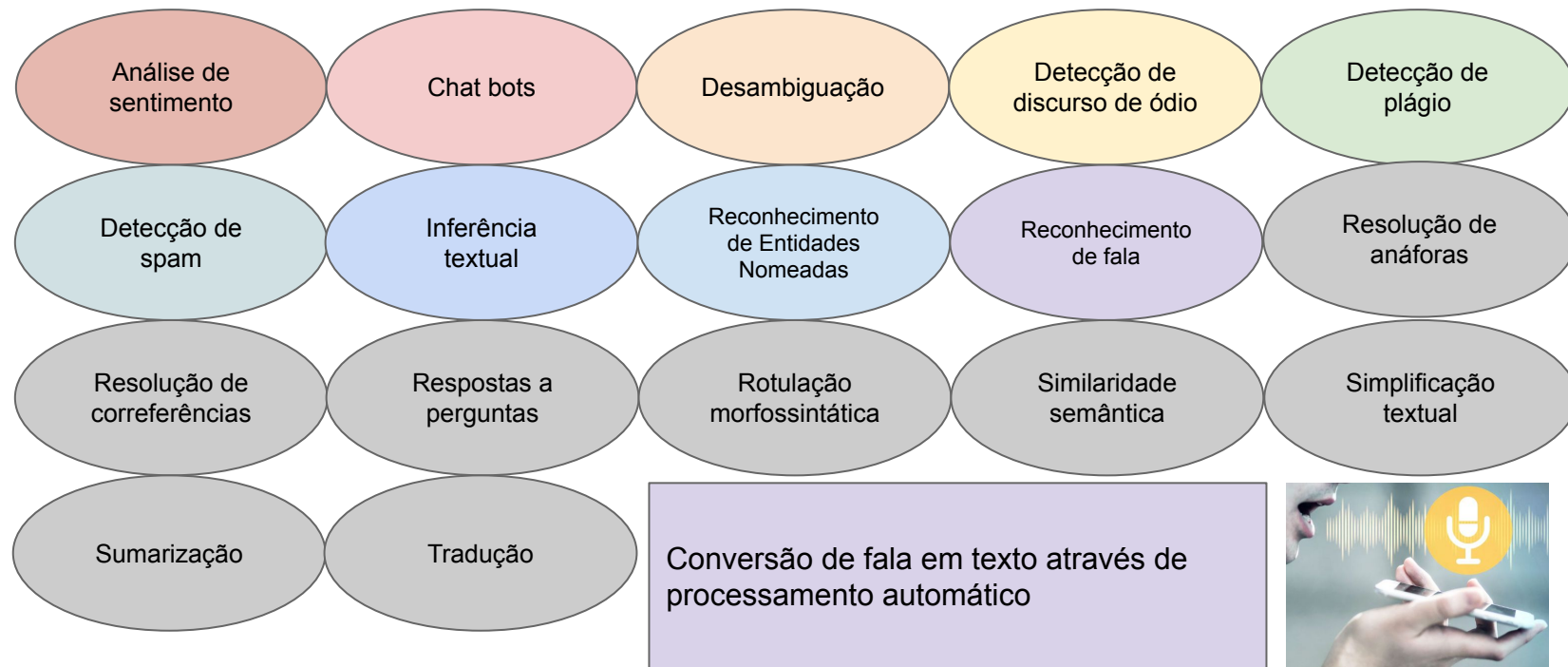
Aplicações



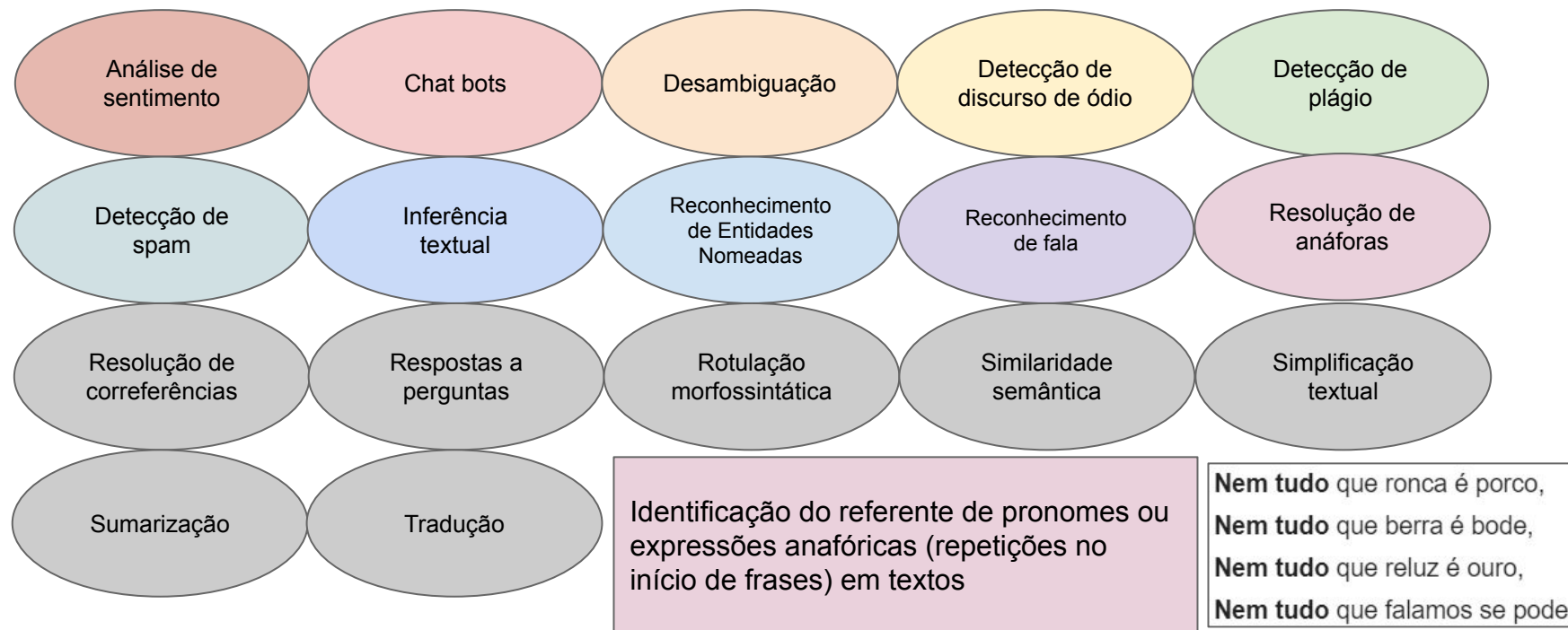
Aplicações



Aplicações



Aplicações



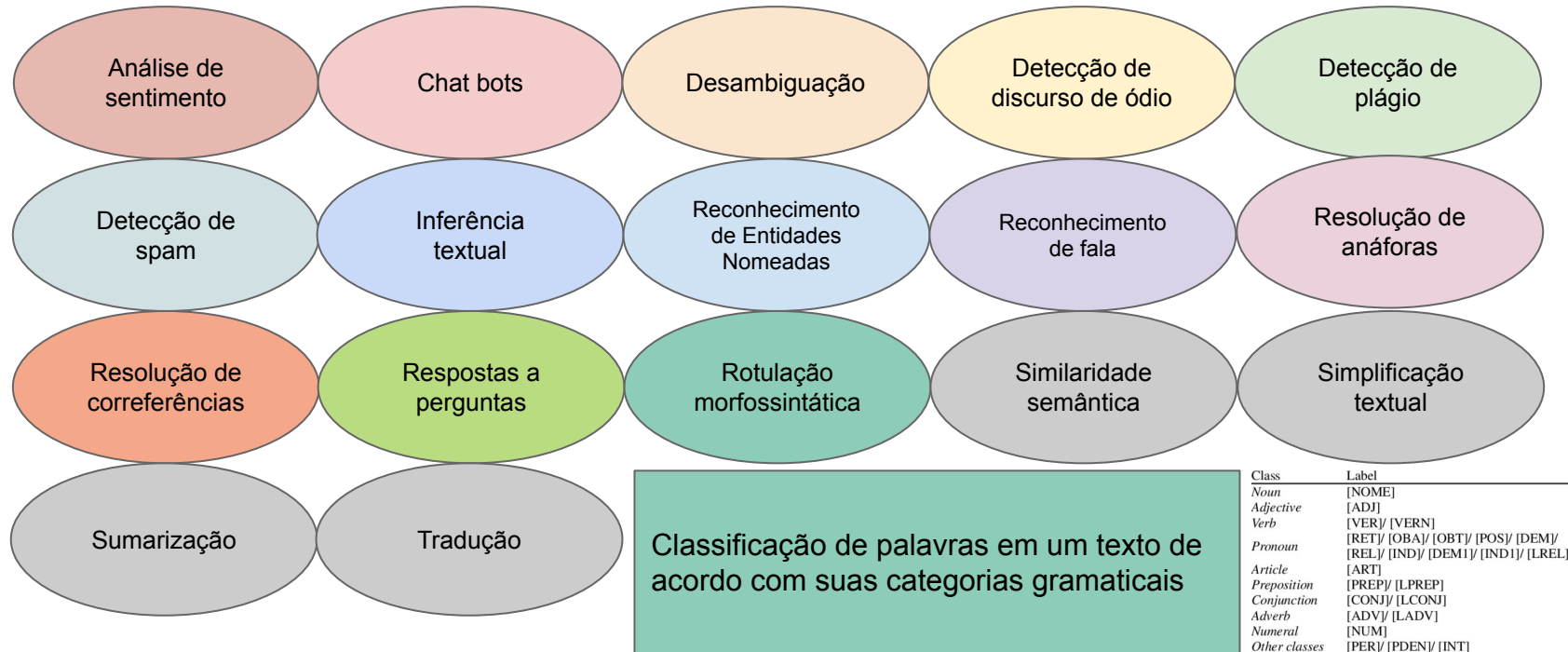
Aplicações



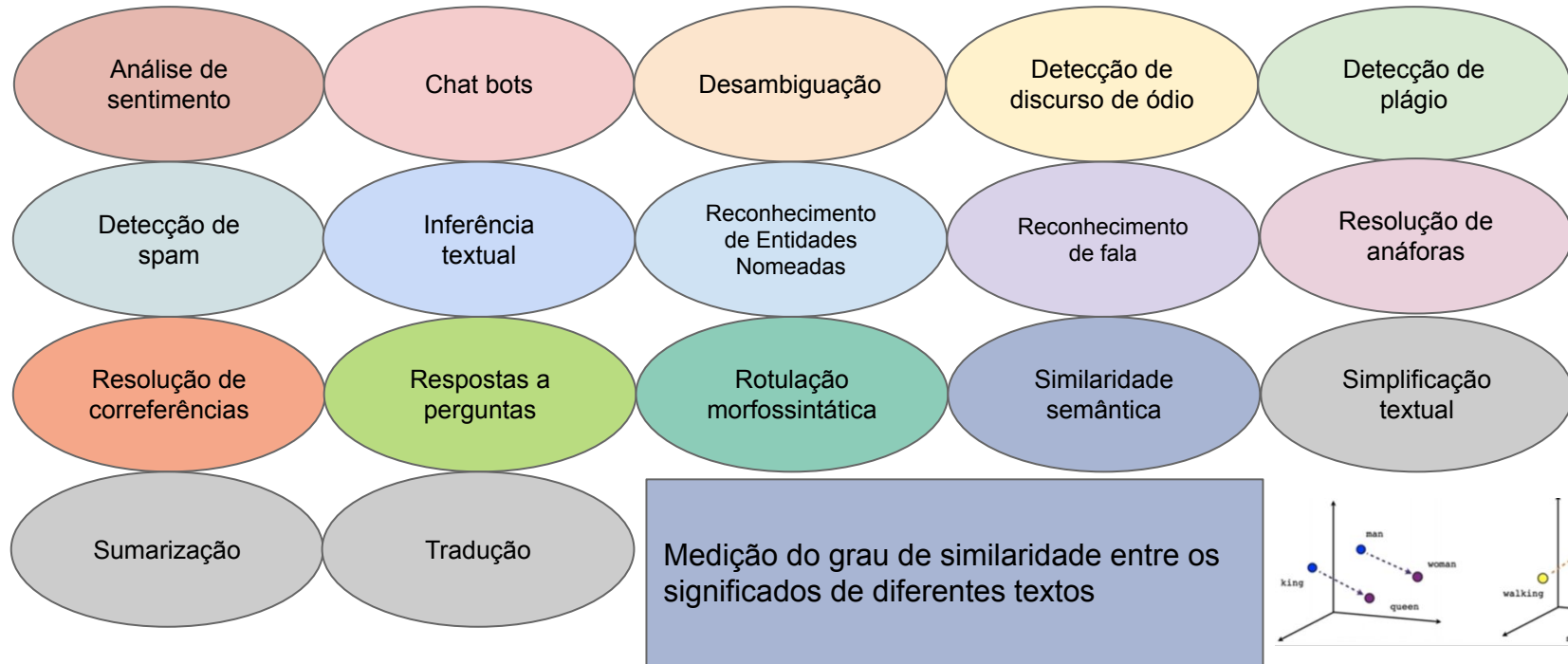
Aplicações



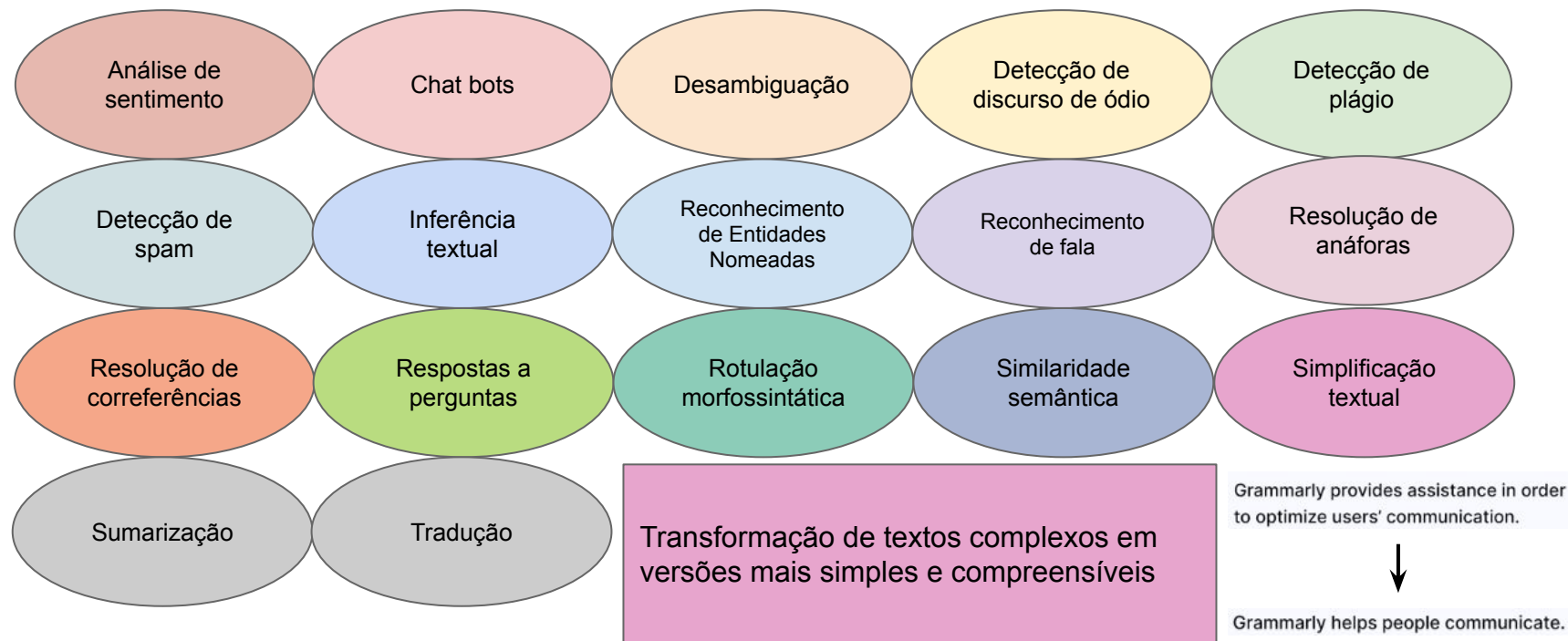
Aplicações



Aplicações



Aplicações



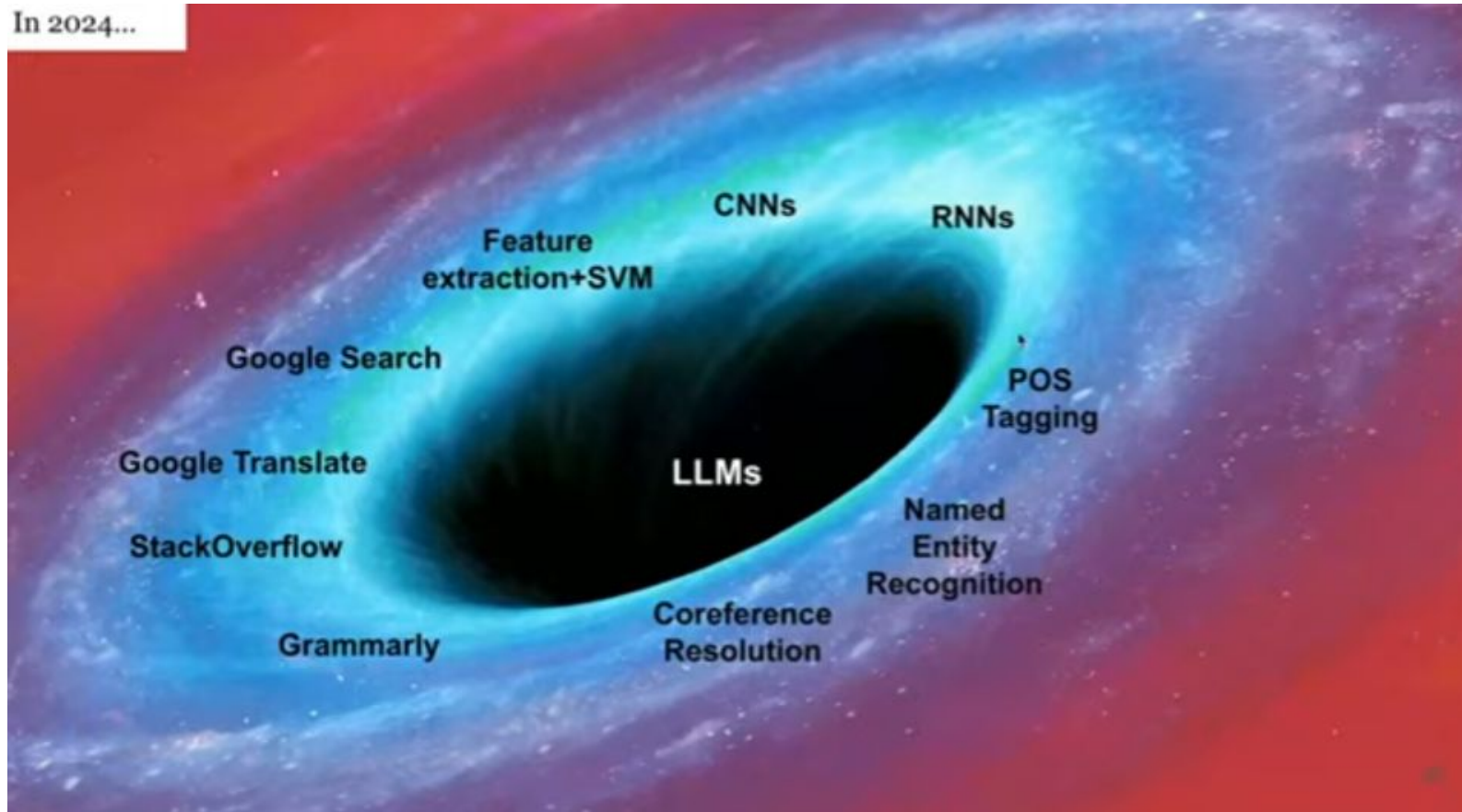
Aplicações



Aplicações



In 2024...



IA382 - Seminar in Computer Engineering: "Generalist vs Specialist Language Models" - Rodrigo Nogueira (<https://www.youtube.com/watch?v=HHT8n7HW6Eg>)

Materiais de PLN

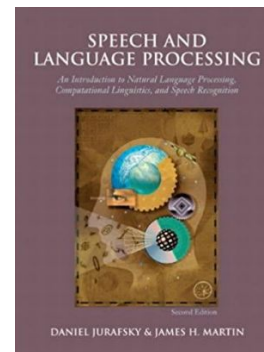
- Livro Speech and Language Processing:
 - Bastante abrangente (mais de 600 páginas)
 - Draft da versão 3 disponível gratuitamente em <https://web.stanford.edu/~jurafsky/slp3/>
- Livro Processamento de Linguagem Natural
 - Online e gratuito em português
 - Diversos autores (61 ao todo)
 - Incluindo a **Profa. Viviane do INF!**
- Sites:
 - Towards data science:
 - <https://towardsdatascience.com>
 - Hugging face:
 - <https://huggingface.co/learn/nlp-course>
 - Medium:
 - <https://medium.com/>



Daniel Jurafsky



James Martin



Helena Caseli



Viviane Moreira



Maria Nunes

■ ■ ■



Pesquisa em PLN

- Extremamente em alta!
- Vários tópicos sendo explorados
- Foco grande em LLMs
 - GPT, BERT, etc.
- Vários domínios interessantes:
 - Legal (legislativo, jurídico)
 - Político
 - Geológico
 - Saúde
- Está gostando do conteúdo?
 - Fale com o professor!

id	Conferência	Frequência	Deadline (aproximado)	Coferência (aproximado)	Qualis
	AAAI Conference on Web and Social Media	Anual	1 janeiro		A1
	SEMISH	Anual	5 abril		A4
	RANLP	Anual	30 abril		A3
3	BRACIS	Anual	5 maio	9 setembro	A4
	SBBB	Anual	13 maio		A4
6	EMNLP	Anual	6 junho	12 dezembro	A1
	webmedia	Anual	1 julho		A4
	ICTAI	Anual	1 julho		A3
	CoNLL	Anual	1 julho		A1
	WI-IAT	Anual	15 julho		A3
	ICMLA	Anual	31 julho		A2
2	COLING	Bianual	16 setembro	5 maio	A1
1	PROPOR	Bianual	10 outubro	3 março	A4
2	LREC	Bianual	10 outubro	5 maio	A1
7	ACL	Anual	11 novembro	7 julho	A1
4	ICEIS	Anual	12 dezembro	3 março	A3
5	NAACL	Anual	12 dezembro	6 junho	A1

Próxima aula

- Conceitos básicos de textos:
 - Dados e informação
 - Tipos de dados
 - Níveis de conhecimento da linguagem
 - Terminologia (palavra, corpus, texto, dataset, sentença)
 - Propriedades estatísticas dos textos
- Modelos de linguagem probabilísticos:
 - Introdução e definição formal
 - Modelo N-grama

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
Instituto de Informática
Departamento de Informática Aplicada

Obrigado pela atenção!
Dúvidas?

Prof. Dennis Giovani Balreira
(Material adaptado da Profa. Viviane Moreira e do Prof. Dan Jurafski)



INF01221 - Tópicos Especiais em Computação XXXVI:
Processamento de Linguagem Natural

