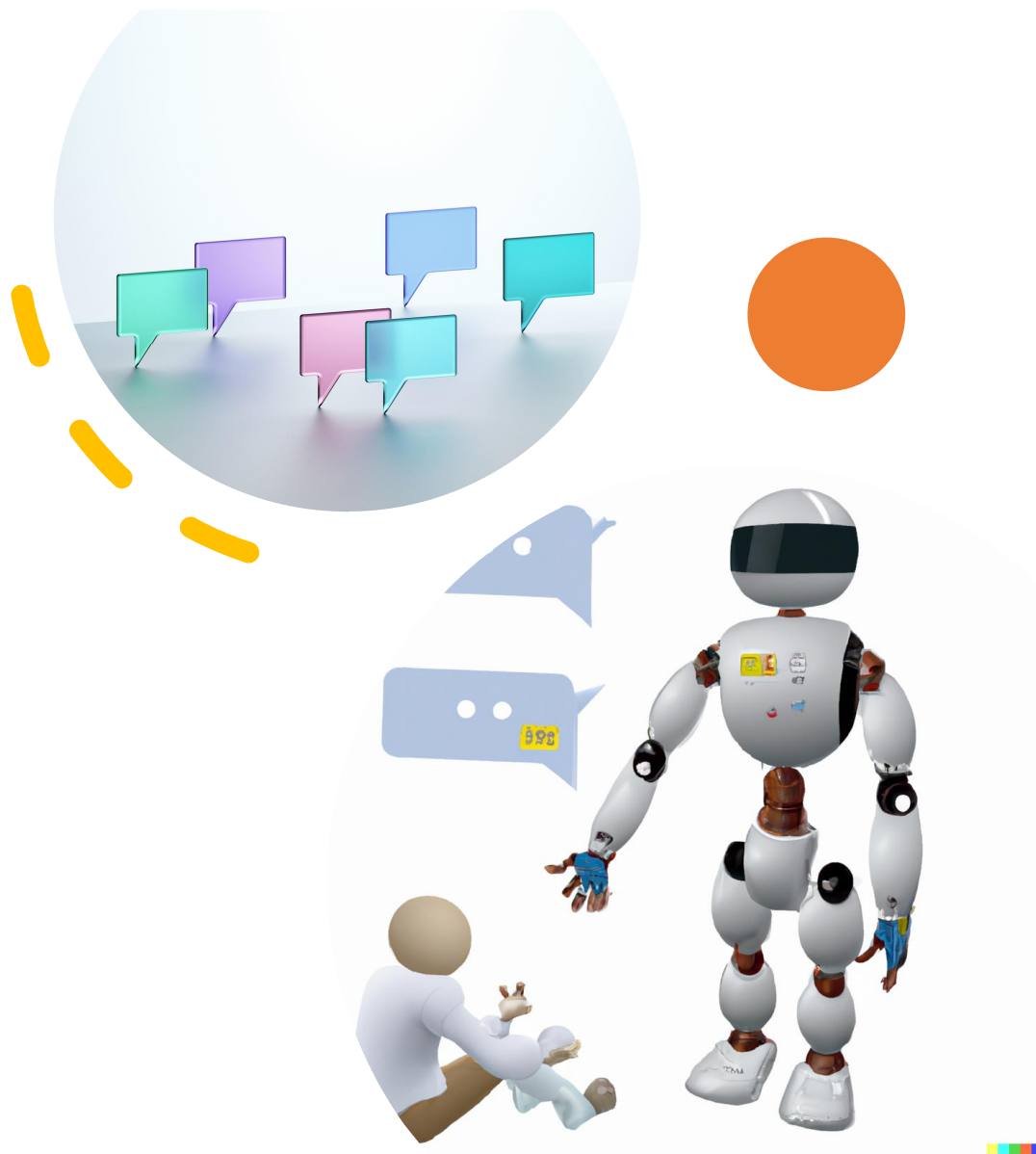


# NLP

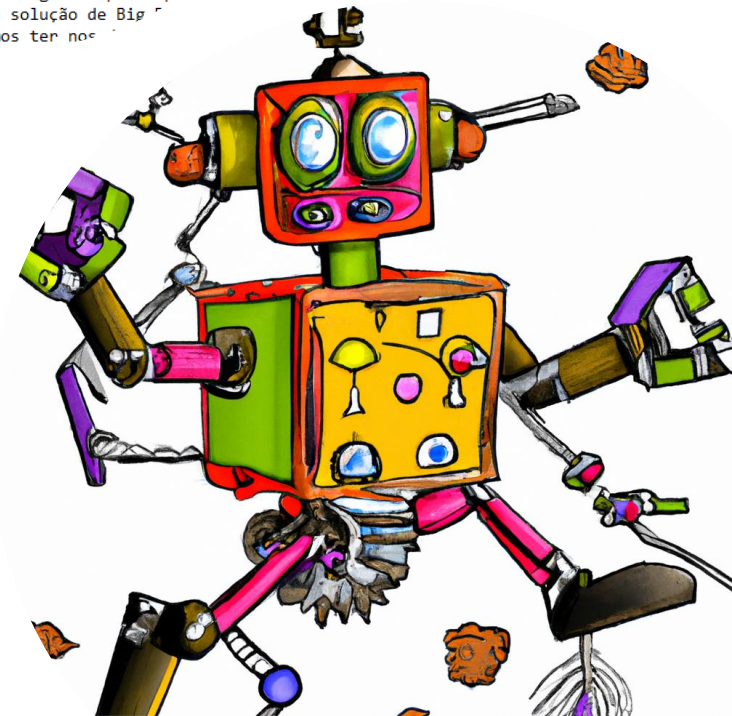
- Tradução
- Geração de Texto
- Produção de Resumos
- Análise de Sentimentos
- Assistentes Virtuais
- Chatbots
- Classificação de Texto
- Reconhecimento de Voz



# Corpus

- Conjunto de documentos (texto não estruturado) em linguagem natural

Informação e conhecimento, analisar dados não é extrair informação e conhecimento de dados? Implementar conhecimento relevante para a tomada de decisão. Isso para dizer que, de uma maneira geral, um projeto de Big Data vamos usar alguns termos que precisamos definir antes. Primeiro de análise de dados ou de Big Data: Nesta obra vamos usar termos de origem: analisar dados requer coletar dados de algum lugar. Segundo de Staging: Muitos processos de análises de dados possuem um fluxo de destino: aqui estaremos sempre nos referindo ao resultado. O MBOK nos ensina que nem todos os seus processos são obrigatórios: você está lendo esta obra provavelmente já ouviu e leu muito sobre isso em seção anterior mas vamos repetir: cabe ao gerente de projeto de a pré-história o homem analisa dados. A análise de dados e a análise de dados só começou a tomar força na década de 90, o que diferencia um projeto de análise de dados tradicional, a velocidade: a velocidade diz respeito não somente a da produção. O gráfico da Figura 1-1 abaixo mostra a relação inversa entre projetos tradicionais eram construídos em armazéns de dados: projetos tradicionais carregavam dados estruturados. As "Vs" existem outras diferenças significativas. Do ponto de vista de arquitetura: projetos tradicionais, existe uma grande preocupação. Vamos olhar uma solução de Big Data. Podemos ter nos



# Anotações / Annotations

- Localizar e classificar elementos específicos no texto
- Exemplos:
  - Anotar sentimentos para treinar um modelo de IA
- Pode ser específico do domínio: Ex: medicina
- Existem empresas especializadas em anotar
- Existem ferramentas especializadas: Doccano, brat etc.
- Alguns tipos podem ser feitos por máquina



Destacados representantes del **ORG** Parlamento y la prensa rusos criticaron hoy el "belicismo ha definido como posible blanco de su lucha antiterrorista.

El presidente de la Duma (cámara baja), **ORG** Guennadi Selezniiov, calificó de "claramente ap **PER**

del Kremlin para Chechenia, **ORG** Serguéi Yastrzhembski. **LOC** **PER**

El asesor presidencial dijo que **LOC** Rusia puede lanzar un ataque preventivo contra los camp

1\n# newpar\n# sent\_id = 1\n# text = Nossa vida é controlada por algoritmos, disse artista e professor de artes digitais de uma universidade

americana\n1\tNossa\t\_\tDET\tDET\t\_\t2\tdet:poss\t\_\t\_\tn2\tvida\t\_\tNOUN\tNOUN\t\_\t4\ttnsubj:pass\t\_\t\_\tn3\té\t\_\tAUX\tAUX\t\_\t4\taux:pass\t\_\t\_\tn4\tcontrolada\t\_\tVERB\tVERB\t\_\t8\tccomp\t\_\t\_\tn5\tpor\t\_\tADP\tADP\t\_\t6\tcase\t\_\t\_\tn6\talgoritmos\t\_\tNOUN\tNOUN\t\_\t

## Tokenization

- Processo de separar a sentença em suas partes: palavras, pontos, símbolos etc.

Nossa vida é controlada por algoritmos,

Nossa

vida

é

controlada

por

algoritmos

,

## Parts-of-Speech Tagging (POS)

- Adiciona tags a cada token, como por exemplo, se é verbo, substantivo, adjetivo etc.

Nossa vida é controlada por algoritmos,

Nossa	vida	é	controlada	por	algoritmos	,
Pron /Interj	Subst.	Verb	Adj	Prep/ LOC. ADVL	Subst.	Pont

# POS Tagging

ABREVIACÃO	SIGNIFICADO	EXEMPLO
PROPN	Nome Próprio	José, Maria
VERB	Verbo	Andar, Dirigir
ADP	Adposição	De, em, durante
DET	Determinante	A, Aquela, muitas
NOUN	Substantivo	Casa, carro
PUNCT	Pontuação	,,;
ADJ	Adjetivo	Infeliz, apavorado, brasileiro
CCONJ	Conjunções Coordenativas	E, nem, mas, entretanto
SCONJ	Conjunções Subordinativas	Embora, mesmo que, uma vez que
AUX	Verbos Auxiliares	Ser, estar, ter
PART	Funções de Partícula	Se, que
PRON	Pronomes	Meu, minha, meus, os quais
NUM	Números	10, vinte
ADV	Advérbios	Tarde, aqui, mal
SYM	Sinais Gráficos	~, ", '
INTJ	Interjeição	Ah, droga, psiu, hum
X	Outros	

# Parts-of-Speech Tagging

- Vamos comer **porco** SUBS
- Ele não toma banho, é um **porco**! ADJ





# Lemmatizing (Lemma)

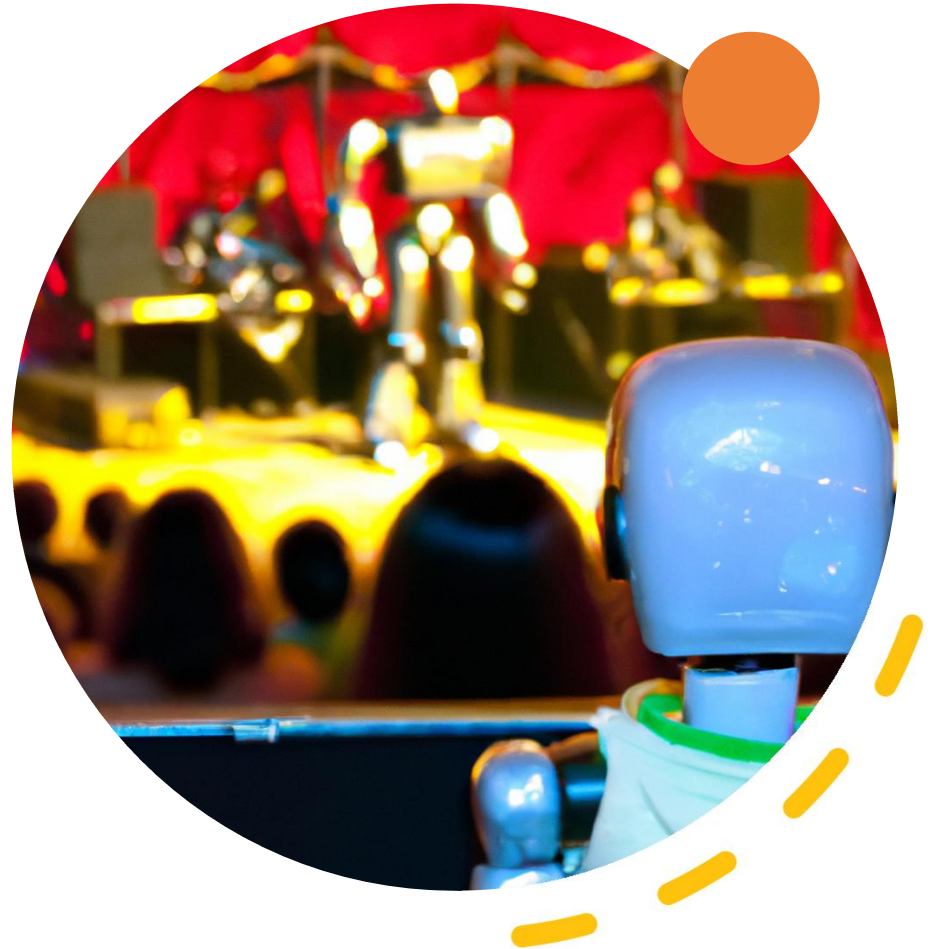
Traz a palavra na sua flexão, de modo que possam ser analisadas juntas

Nossa vida é controlada por algoritmos,

Nossa	vida	é	controlada	por	algoritmos	,
Pron /Interj	Subst.	Verb	Adj	Prep/ LOC. ADVL	Subst.	Pont
meu	vida	ser	controlado	por	algoritmo	,

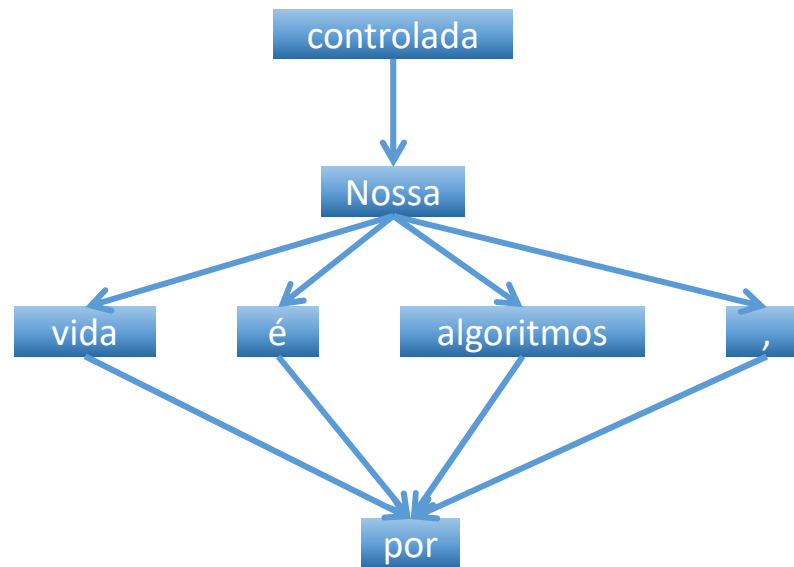
# Stemming

- Cortas palavras, buscando ter uma representação raiz e única
- Diferentes técnicas
- Lemmatization é mais sofisticado
- Amigo, amigos, amiga, amigas => amig



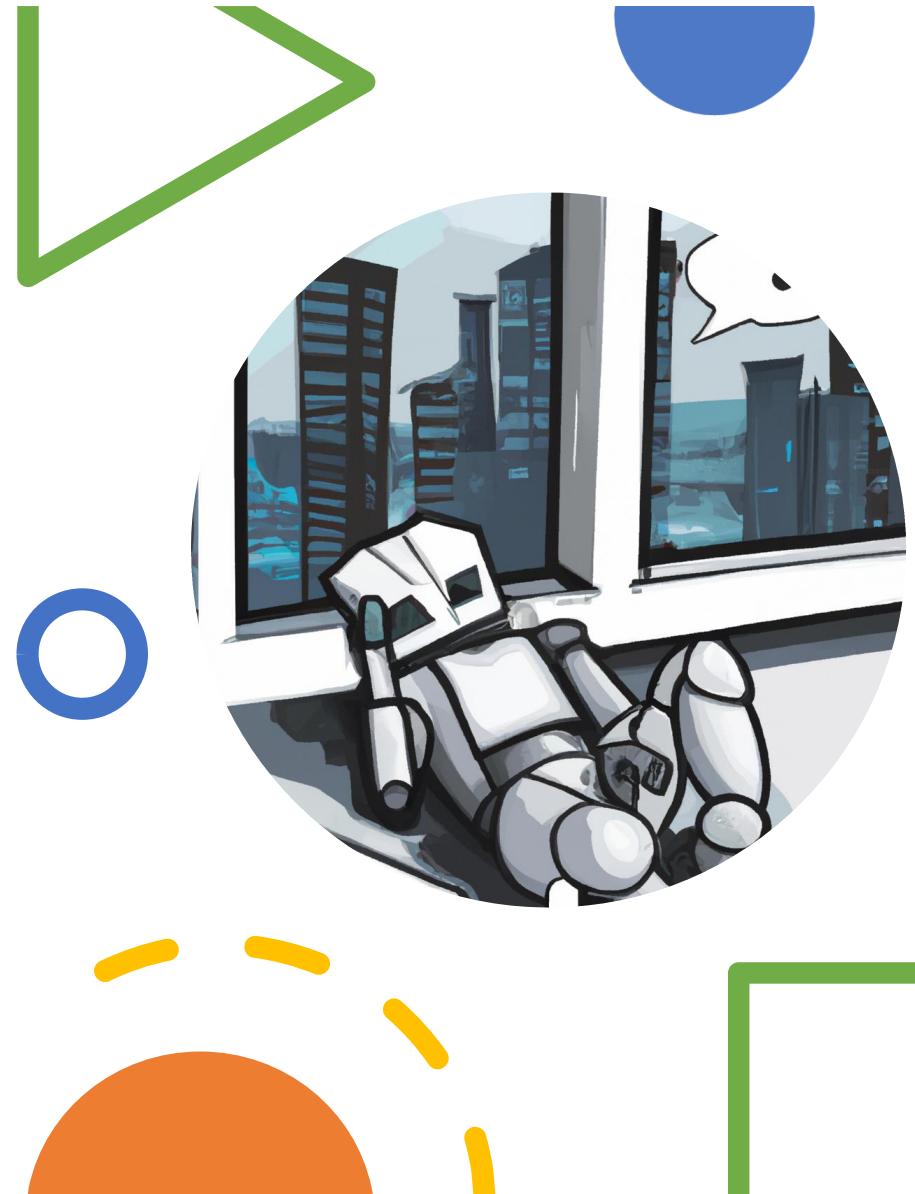
# Dependency Parsing

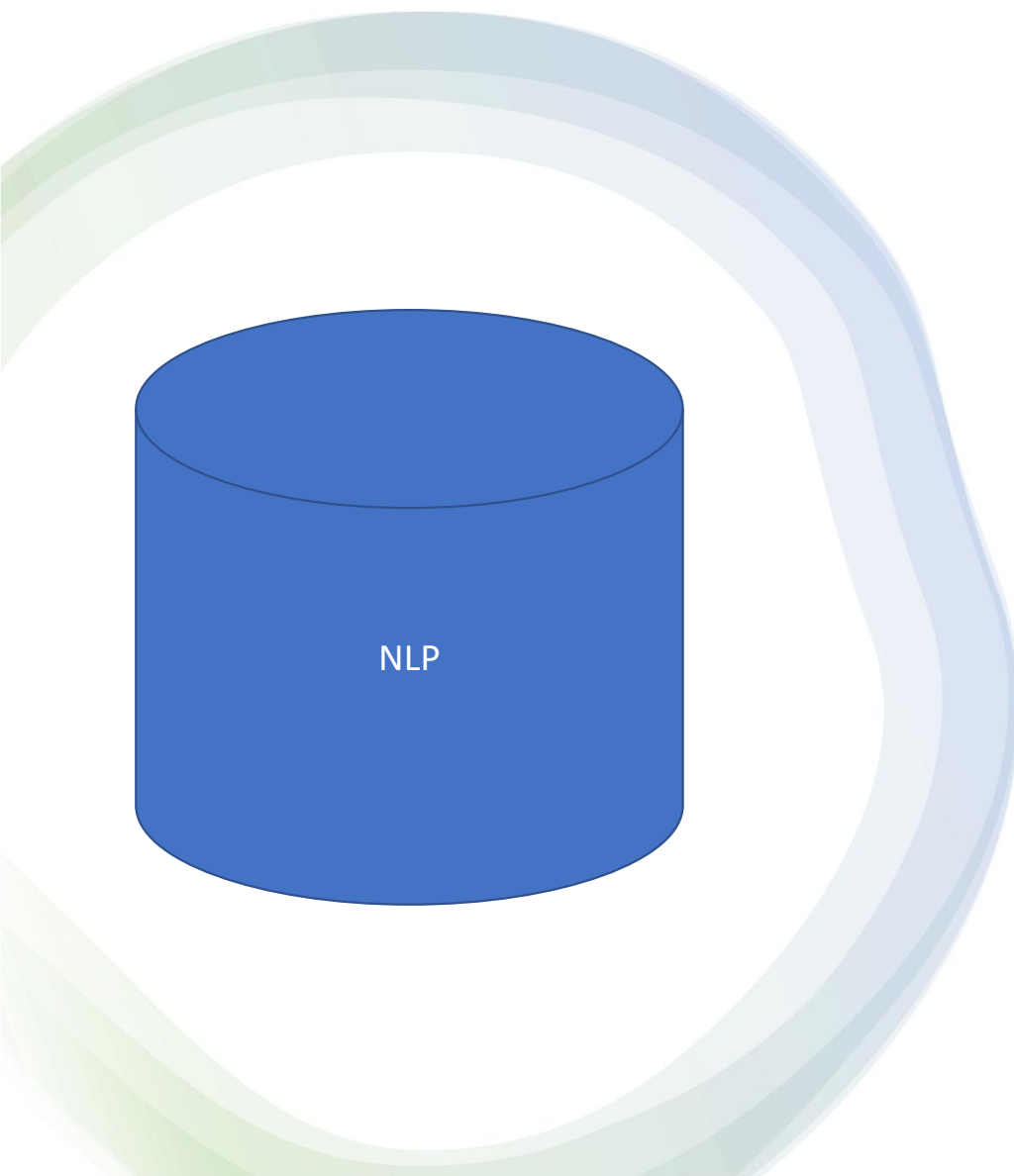
Encontra relação entre palavras “pais” e “filhos”



# NGRAM

- N palavras consecutivas
- Bigrams e trigrams
- 4 ou mais não usado para palavras devido a esparsidade
- Pode também ser aplicado a letras





## Modelo

- Análise
  - Verbo? Substantivo? Quais são as flexões?  
Quais as dependências?
- Um modelo é um banco de dados linguístico
- Específico de cada idioma
- Maioria das plataformas de NLP tem seus próprios modelos (ou usam de terceiros)
- Você pode criar o seu!