



# INSIGHT

Data Science Laboratory  
Federal University of Ceará

# Processamento de Linguagem Natural via Aprendizagem Profunda

JORNADA DE ATUALIZAÇÃO EM INFORMÁTICA (JAI)



**42ºCSBC**  
CONGRESSO DA SOCIEDADE  
BRASILEIRA DE COMPUTAÇÃO

 DEPARTAMENTO  
DE COMPUTAÇÃO

 UNIVERSIDADE  
FEDERAL DO CEARÁ

# Olá!



## Bárbara Neves

- Estudante de mestrado @ufcinforma
- Pesquisadora @\_insightlab
- Cientista de Dados @iatlantico



## Gustavo Coutinho

- Estudante de doutorado @ufcinforma
- Pesquisador @\_insightlab
- Professor @ifceoficial



## José Antônio Macêdo

- Cientista Chefe de Transformação Digital @governodoceara
- Coordenador @\_insightlab
- Professor @ufcinforma

# INTRODUÇÃO

# INTRODUÇÃO



A cada 24 horas, **500 milhões de tweets** são publicados

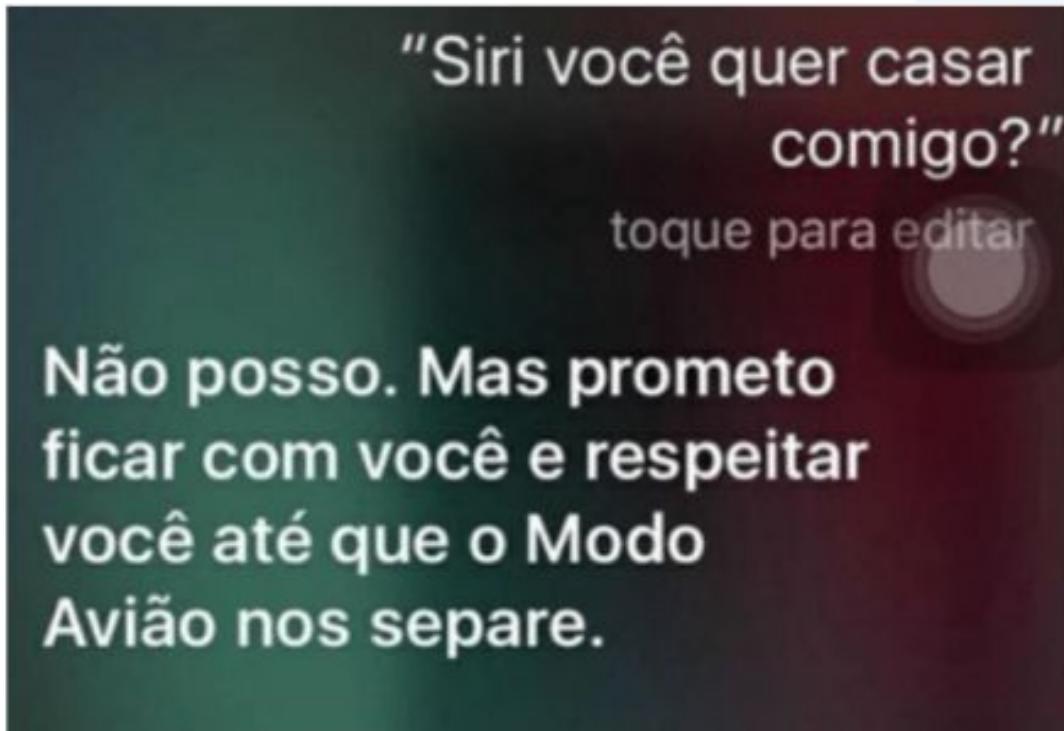


O Google processa **1.2 trilhões** de buscas todo ano

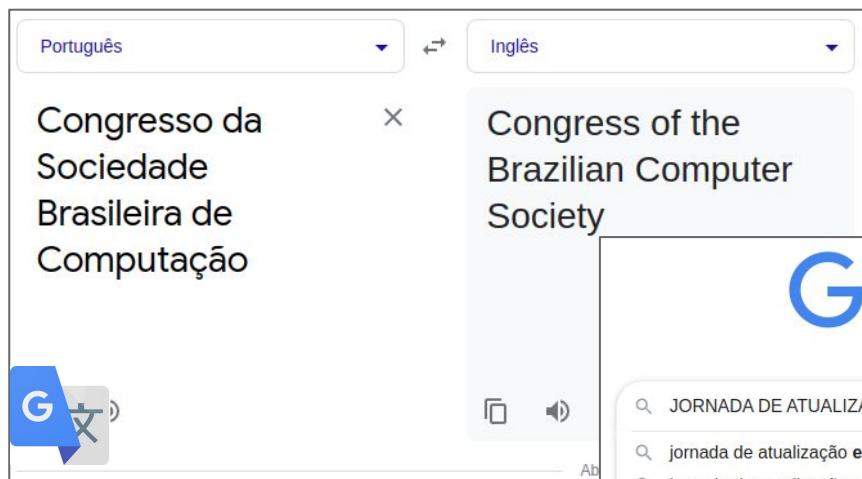


O Instagram possui mais de **1 bilhão** de usuários ativos

# INTRODUÇÃO



# INTRODUÇÃO



Congresso da Sociedade Brasileira de Computação

Congress of the Brazilian Computer Society

Google search results for "JORNADA DE ATUALIZAÇÃO":

- jornada de atualização em pediatria
- jornada de atualização em informática na educação
- jornada de atualização pedagógica - aulas remotas 2020
- jornadas de atualização cardiologia 2022
- jornadas de atualização cardiologia 2021
- jornadas de atualização em dermatologia para mgf
- jornadas de atualização cardiológica 2022
- jornadas de atualização cardiologica

Plantão Saúde Ceará

Precisa de ajuda? Clique aqui!

Estou com COVID

Olá! Bem-vindo(a) ao Assistente Virtual da Secretaria da Saúde do Estado do Ceará!

Antes de iniciarmos seu atendimento, em qual perfil você se enquadra:

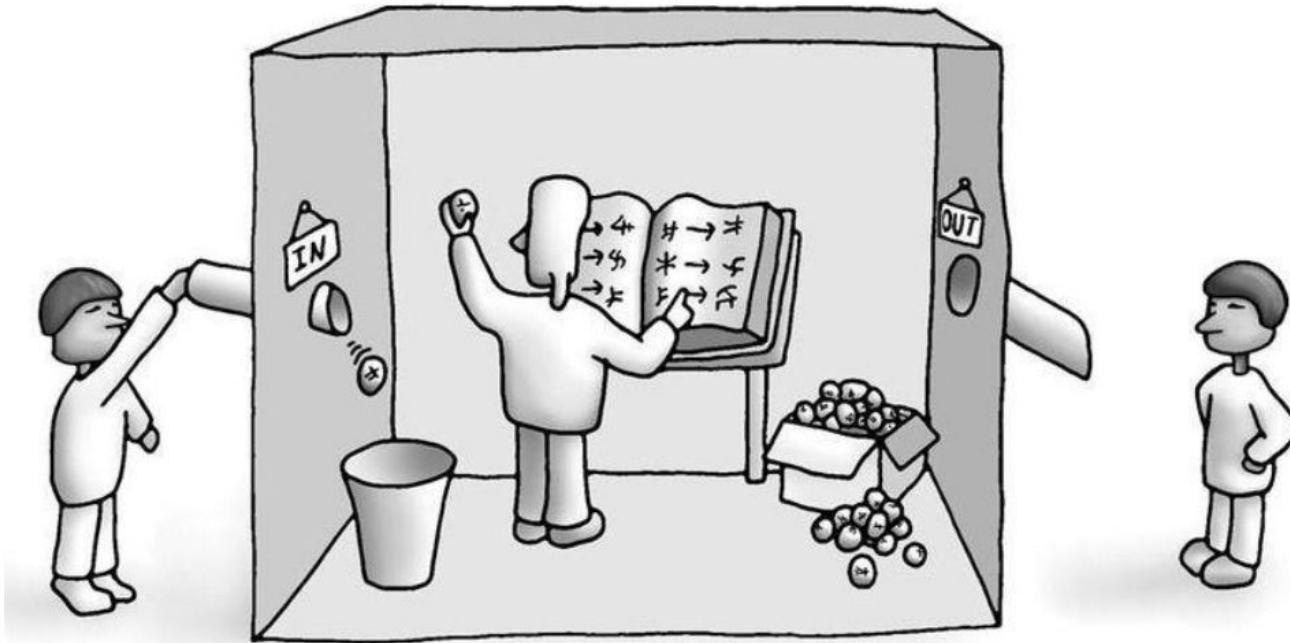
- Sou paciente
- Sou Profissional de Saúde
- Sou paciente SURDO V.D.A.
- Estou com reações adversas depois de tomar a vacina COVID

Type a message...

## INTRODUÇÃO - PLN

- ▶ Processamento da Linguagem Natural, ou PLN, é um campo em desenvolvimento que estuda as possibilidades de interação entre as áreas da linguagem e da computação
- ▶ Facilita a comunicação entre máquina e humano, sem que o usuário tenha que aprender a linguagem da máquina.

# INTRODUÇÃO - Quarto Chines



Fonte: <https://filosofianaescola.com/metafisica/quarto-chines/>



John Searle  
(Berkeley)

# INTRODUÇÃO - Teoria do Labirinto

Lidar com linguagem é um processo complexo, possuindo ruídos, erros gramaticais, dialetos, gírias e **ambiguidades**.

Lexical: causada por uma palavra

**“Me sugira um bom prato”**

Estrutural: causada pela posição de uma palavra na frase

**“Olhe meu irmão enquanto toca música!”**

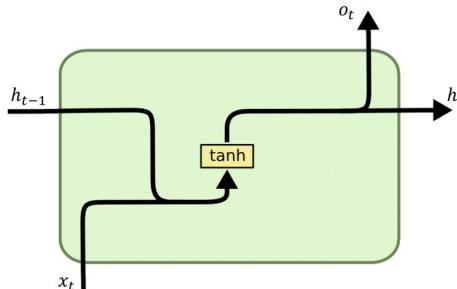
# INTRODUÇÃO

Modelos de Aprendizagem Profunda têm sido amplamente utilizados em tarefas de PLN:

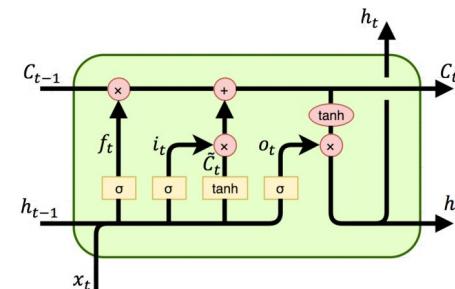
1. Exigem pouca **engenharia de features**
2. Produzem **representações vetoriais** que capturam similaridades de palavras
3. Permitem o aprendizado **não supervisionado** ou **semi-supervisionado**
4. Aprendem vários **níveis de representações**
5. Lidam com **recursividade da linguagem humana**, conseguindo capturar informações de forma sequencial

# INTRODUÇÃO

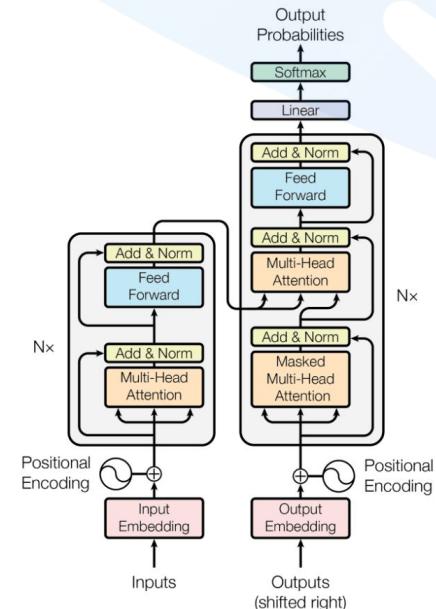
- ▷ Modelos de Deep Learning começaram a ser utilizados em PLN através de Redes Neurais Recorrentes (RNNs) e redes *Long Short Term Memory* (LSTMs)
- ▷ Mais recentemente, o uso de arquiteturas *Transformers* se popularizou



RNN



LSTM



Transformers



# Roteiro do Minicurso

O foco do presente minicurso é:

1. Apresentar as etapas de pré-processamento de texto
2. Familiarizar os(as) participantes com diferentes tipos de como representar textos
3. Aplicar os conceitos aprendidos em tarefas de Classificação de Textos e Sumarização de Sentenças

# TECNOLOGIAS



Numpy



Python



Keras



Google Colab



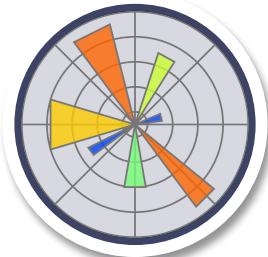
Scikit Learn



Seaborn



spaCy



Matplotlib

# PRELIMINARES

# BREVE HISTÓRICO



As linguagens naturais são **complexas, ambíguas** e estão em **constante mudança**

- ▶ Como aprendemos uma nova língua?
  - ▶ Como fazemos com que computadores processem diferentes linguagens?
  - ▶ Como *Alexa*, *Google Home* e muitos outros assistentes inteligentes entendem e conseguem nos responder?

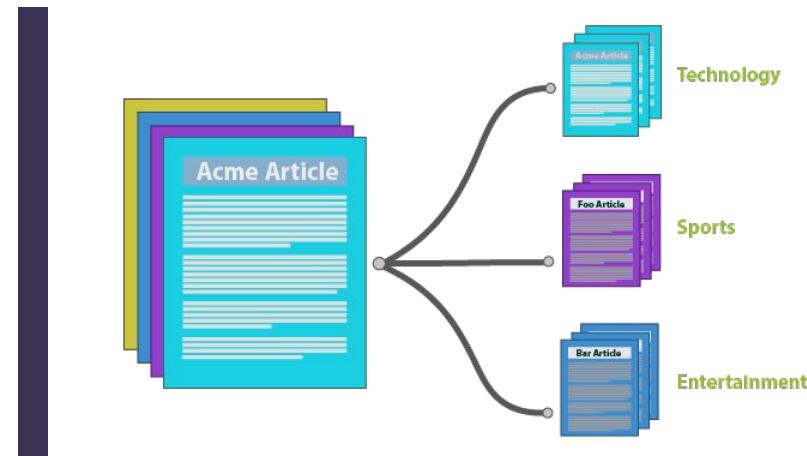


**Agravante:** não existe linguagem universal!

# BREVE HISTÓRICO

Com o avanço do **PLN**, sistemas são capazes de

- ▷ Classificar textos



## BREVE HISTÓRICO

Com o avanço do **PLN**, sistemas são capazes de

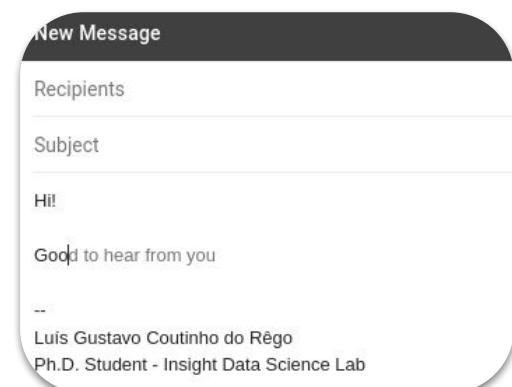
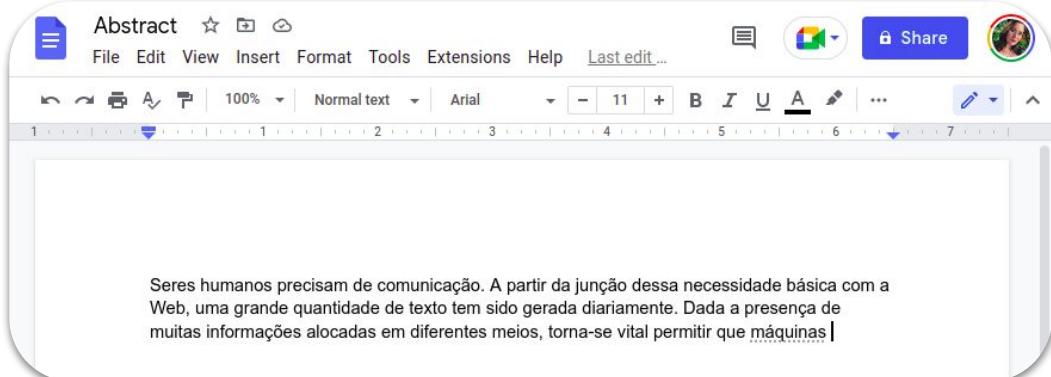
- ▷ Reconhecer entidades nomeadas

A **Jornada de Atualização em Informática** MISC é um dos mais importantes eventos acadêmicos de atualização científica e tecnológica da comunidade de computação do Brasil Loc. Sua 41<sup>a</sup> versão ocorrerá em Niterói Loc, de 31 de julho a 5 de agosto de 2022 como parte do **congresso da Sociedade Brasileira de Computação** MISC.

# BREVE HISTÓRICO

Com o avanço do **PLN**, sistemas são capazes de

- ▷ *Language Modeling*



# BREVE HISTÓRICO

Com o avanço do **PLN**, sistemas são capazes de

- ▷ Comprimir sentenças

Seres humanos precisam de comunicação. A partir da junção dessa necessidade básica com a Web, uma grande quantidade de texto tem sido gerada diariamente. Dada a presença de muitas informações alocadas em diferentes meios, torna-se vital permitir que máquinas compreendam textos falados e escritos. Este capítulo apresenta como técnicas de Aprendizagem Profunda podem ser utilizadas na resolução de tarefas de Processamento de Linguagem Natural (PLN), como Classificação e Sumarização de Sentenças, visando o benefício do poder computacional disponível atualmente e da baixa necessidade de engenharia de features na utilização destes modelos. Inicialmente, são apresentados alguns conceitos importantes sobre PLN e Aprendizagem Profunda. Em seguida, diferentes técnicas de pré-processamento e representação textuais são explicadas a fim de serem usadas como entrada em modelos de Aprendizagem Profunda. Por fim, é mostrado como aplicar os conhecimentos adquiridos em aplicações reais do PLN.

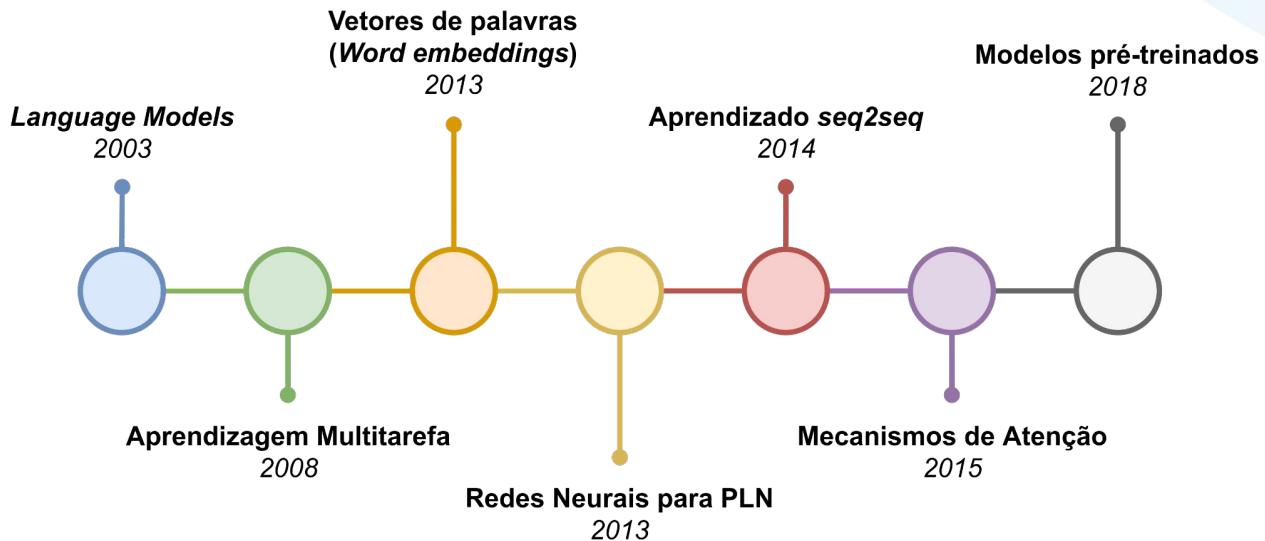
# BREVE HISTÓRICO

Com o avanço do **PLN**, sistemas são capazes de

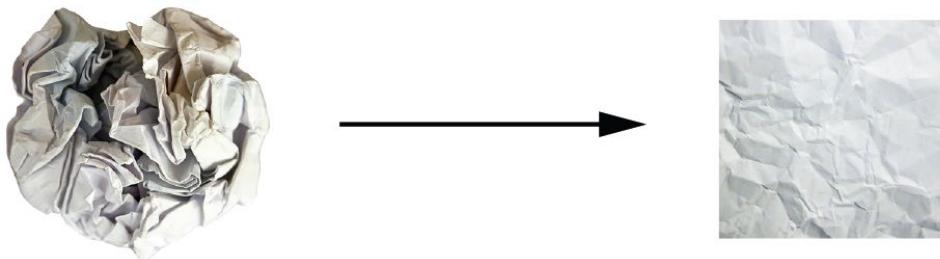
- ▷ Comprimir sentenças

Este capítulo apresenta como técnicas de Aprendizagem Profunda podem ser utilizadas na resolução de tarefas de Processamento de Linguagem Natural (PLN), como Classificação e Sumarização de Sentenças, visando o benefício do poder computacional disponível atualmente e da baixa necessidade de engenharia de features na utilização destes modelos. Inicialmente, são apresentados alguns conceitos importantes sobre PLN e Aprendizagem Profunda.

# BREVE HISTÓRICO



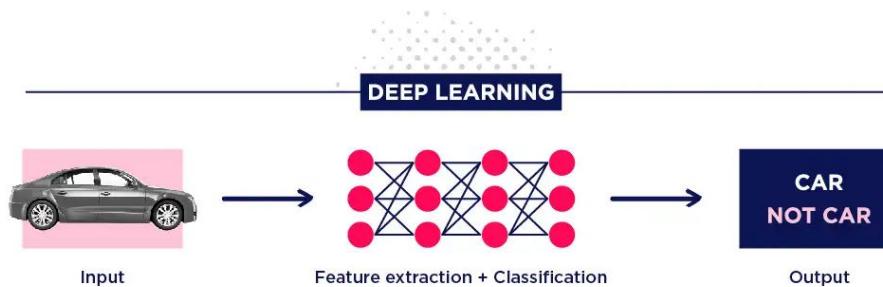
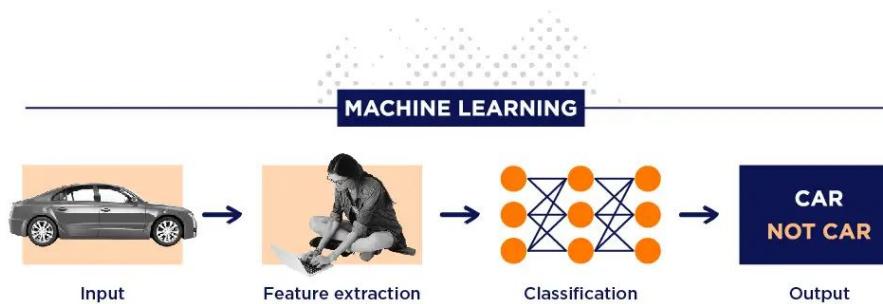
# REDES NEURAIS



**Figura 1.** Desamassando uma variedade complicada de dados

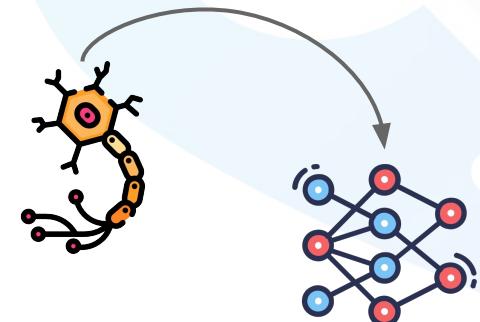
# CONCEITOS BÁSICOS

O que é a Aprendizagem Profunda?



# CONCEITOS BÁSICOS

- ▷ Redes neurais são:
  - ▶ Inspiradas em **neurônios biológicos**
  - ▶ Ideais para **tarefas complexas**
  - ▶ Tarefas com dados **não estruturados**

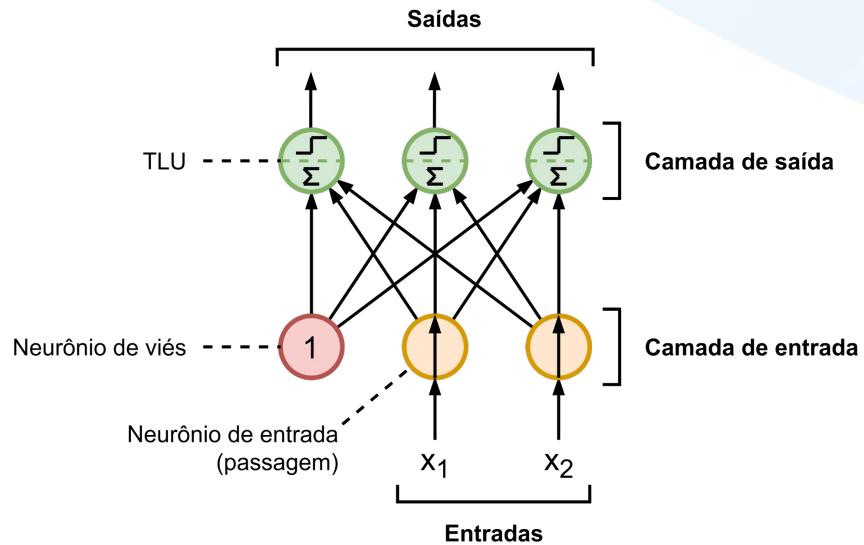
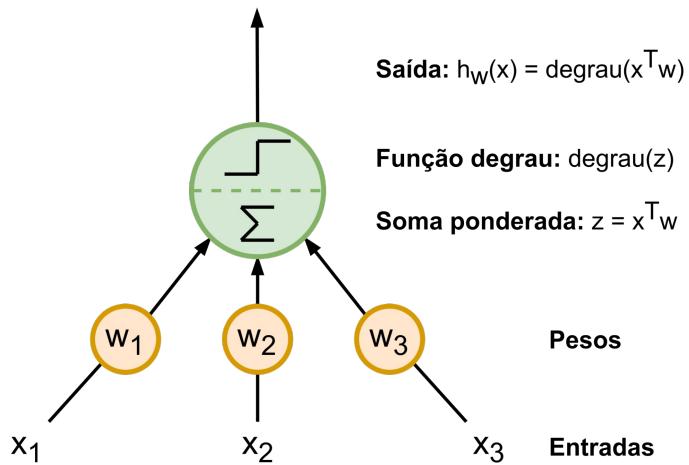


# PERCEPTRON

- ▶ Desenvolvido por McCulloch e Pitts em 1943
- ▶ Composto por:
  - ▶ Uma camada de entrada,
  - ▶ Uma camada de Unidades Lógicas de Limiar (TLUs)

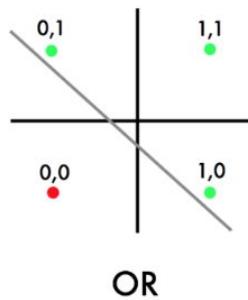


# PERCEPTRON

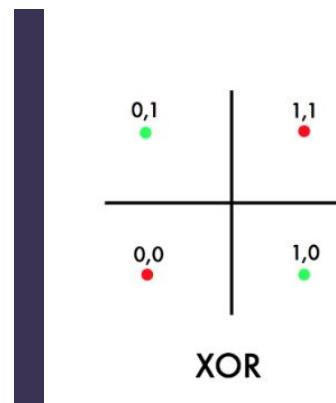


# PERCEPTRON

- ▷ Quais as suas **limitações**?
  - ▶ Problemas **um pouco** mais complexos, como o sistema XOR



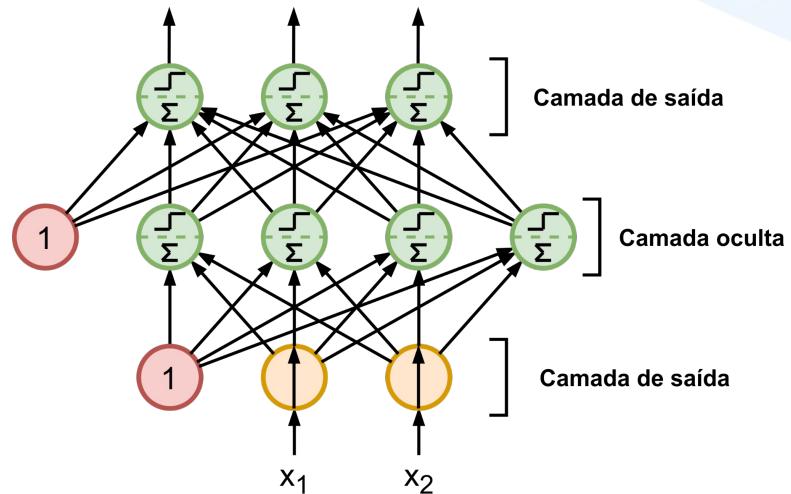
OR



XOR

# PERCEPTRON MULTICAMADAS

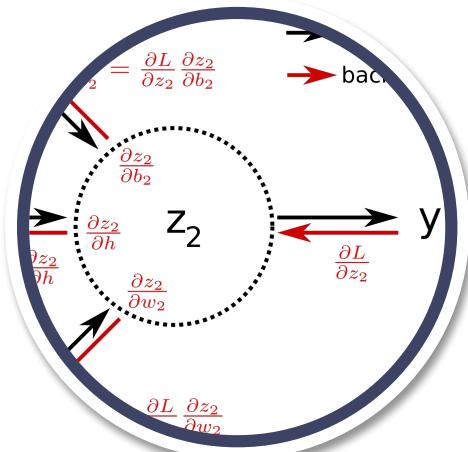
- ▷ Composto por:
  - ▶ Uma camada de entrada
  - ▶  $N$  camadas de TLUs
  - ▶ Uma camada de saída



# PERCEPTRON MULTICAMADAS

Quando um MLP possui mais de uma camada oculta, já podemos chamá-la de **Rede Neural Profunda**

- ▷ O treinamento é realizado por *backpropagation*
  - ▶ *Forward pass*
  - ▶ *Backward pass*
- ▷ Característica importante
  - ▶ Funções de ativação em MLPs são diferentes de um *perceptron* normal



# PERCEPTRON MULTICAMADAS

Quando utilizar uma MLP?

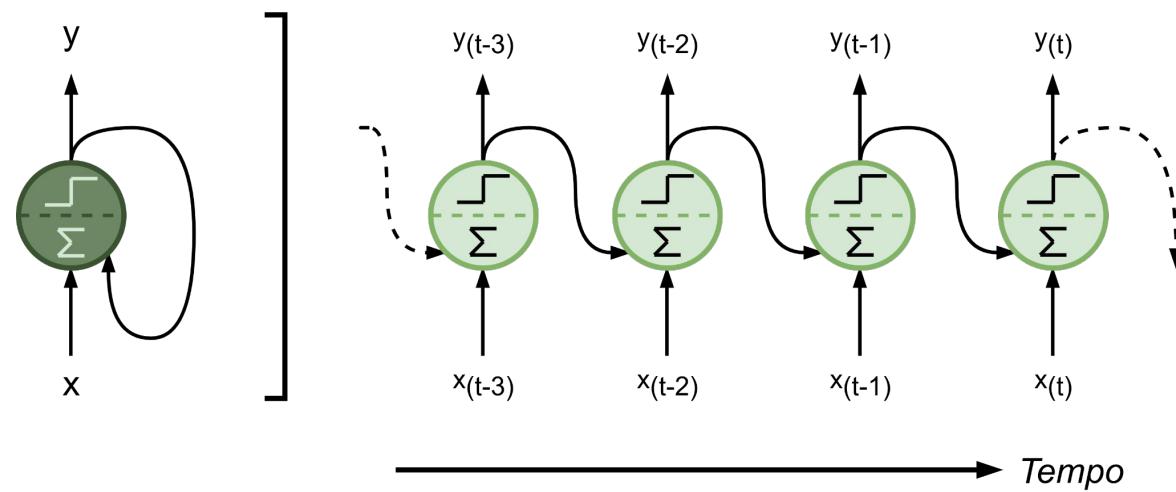
- ▷ Para entradas de tamanho fixo
- ▷ Para saídas discretas

Para problemas com formato de sequências, devemos utilizar redes mais **complexas**, como **Redes Neurais Recorrentes (RNNs)**

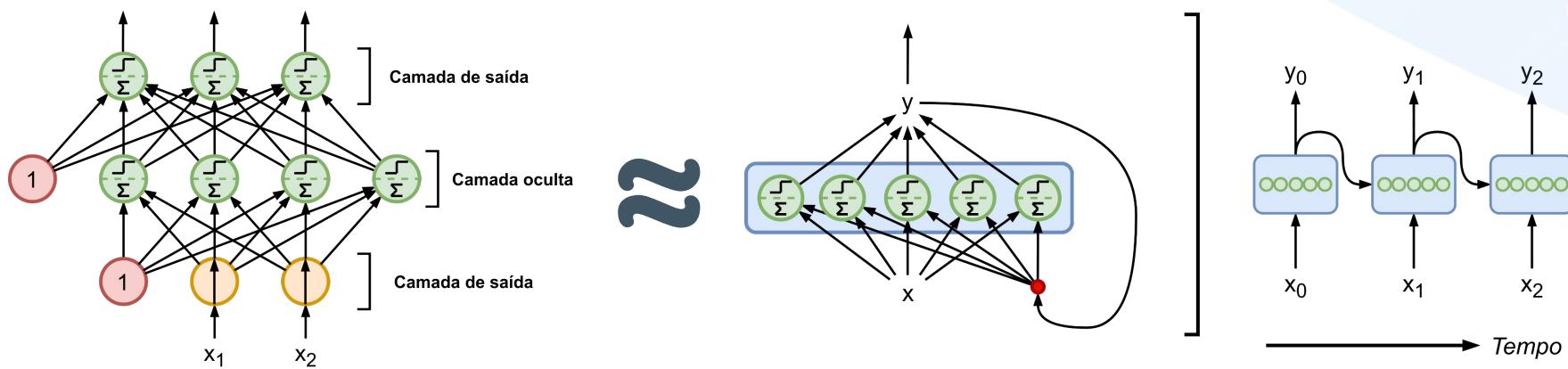
## MLPs x RNNs

Qual a **principal diferença** entre RNNs e MLPs?

- ▶ RNNs possuem ligações para neurônios de camadas anteriores



## MLPs x RNNs



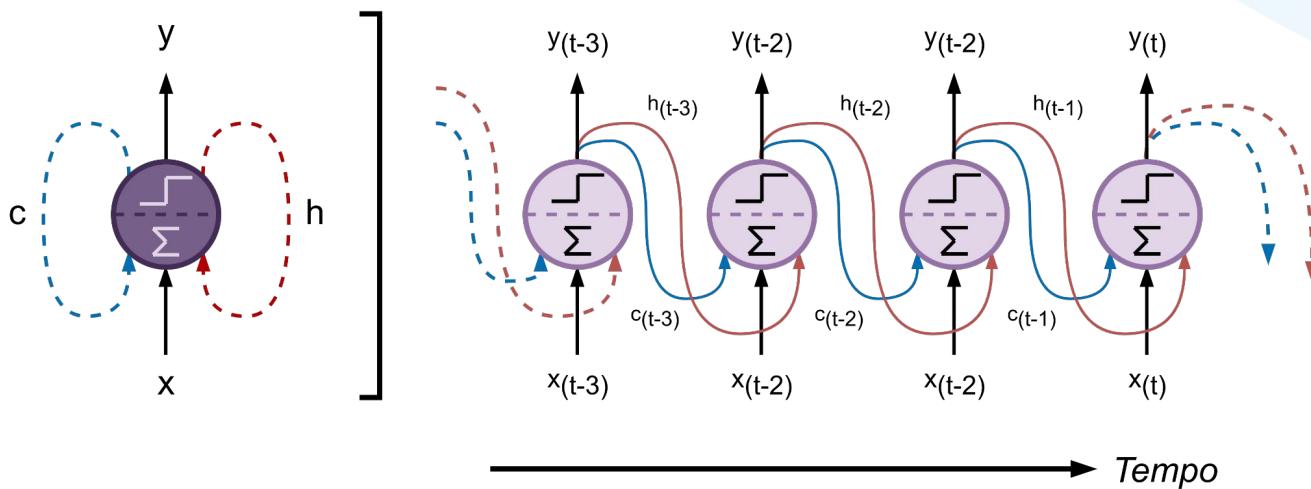
**Figura 2.** Representação de uma camada de unidades recorrentes à direita.

## VARIAÇÃO DAS RNNs

- ▶ **Sequências longas** são potencialmente **problemáticas** para RNNs
  - ▶ Essas redes tendem a crescer muito, tornando o treinamento computacionalmente custoso e demorado
  - ▶ Podemos ter perdas de informações do começo das sentenças

Para **mitigar** esses problemas, células de ***Long Short Term Memory (LSTM)*** foram desenvolvidas

# VARIAÇÃO DAS RNNs



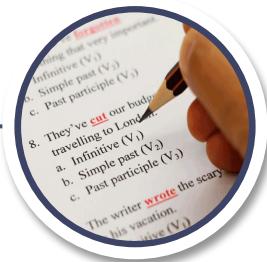
# PREPARAÇÃO DE DADOS TEXTUAIS



# PADRONIZAÇÃO



Frases com mesma informação podem possuir **estruturas diferentes**



A padronização tenta diminuir a diferença entre textos semelhantes



**Técnicas comuns** são remoção de pontuação e acentos

## PADRONIZAÇÃO

Sentenças podem possuir a **mesma informação**, mas serem **escritas de formas diferentes**

*"Olha só! Como PLN é legal! Estava estudando esses dias."*

*"Olha só, como PLN é legal! Estava estudando esses dias."*

# PADRONIZAÇÃO

"Olha só! Como PLN é legal!"

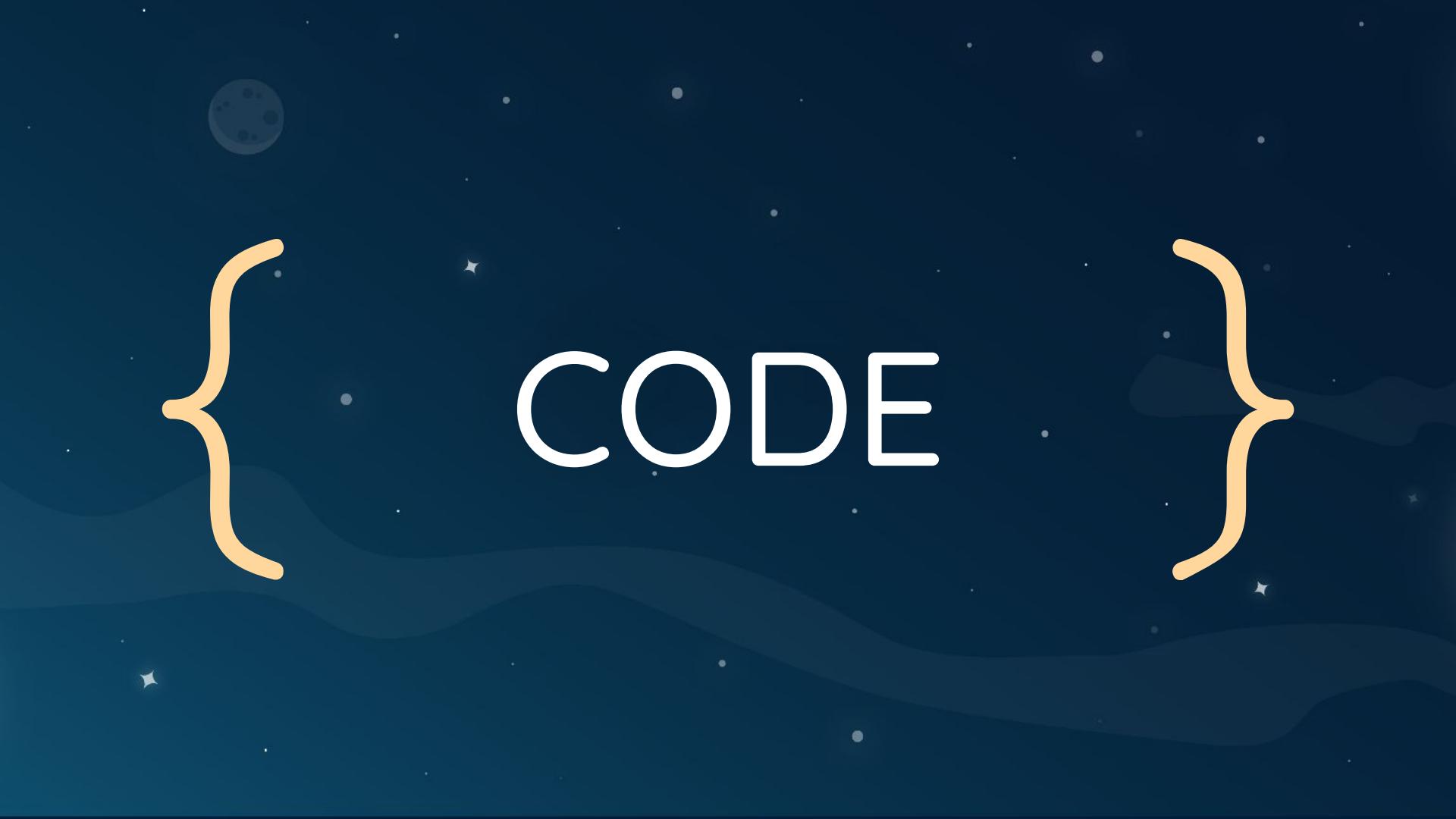
Estava estudando esses dias."

"Olha só, como PLN é legal!"

Estava estudando esses dias."

Padronização

"olha so como pln e legal estava estudando esses dias"

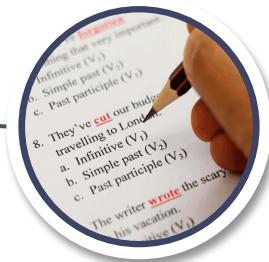


CODE

# TOKENIZAÇÃO



**Tokenização** é o ato de dividir o texto em **unidades vetorizáveis**



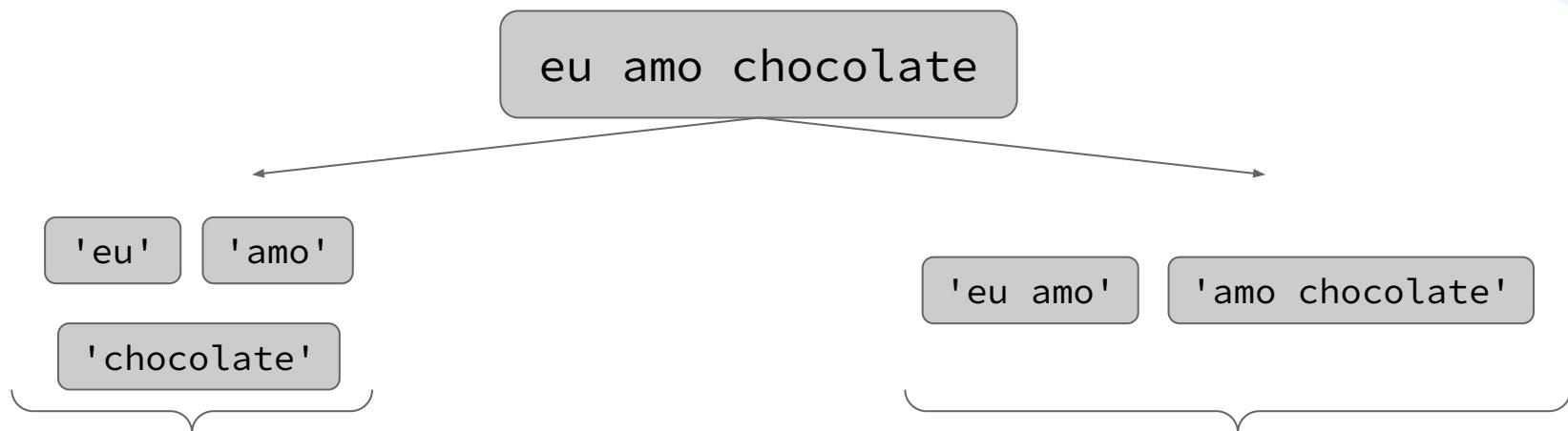
O processo de estemização pode ser necessário antes de uma **tokenização**



Um outro processo importante a ser considerado é o de **lematização**

# TOKENIZAÇÃO

Após a padronização, precisamos quebrar o texto em unidades a serem vetorizadas



Tokenização a nível de palavras.

Tokenização a nível de n-grams.

## TOKENIZAÇÃO

Em que as palavras seguintes diferem?

- ▷ presidente; presidenta
- ▷ estava, estive
- ▷ casa, casas

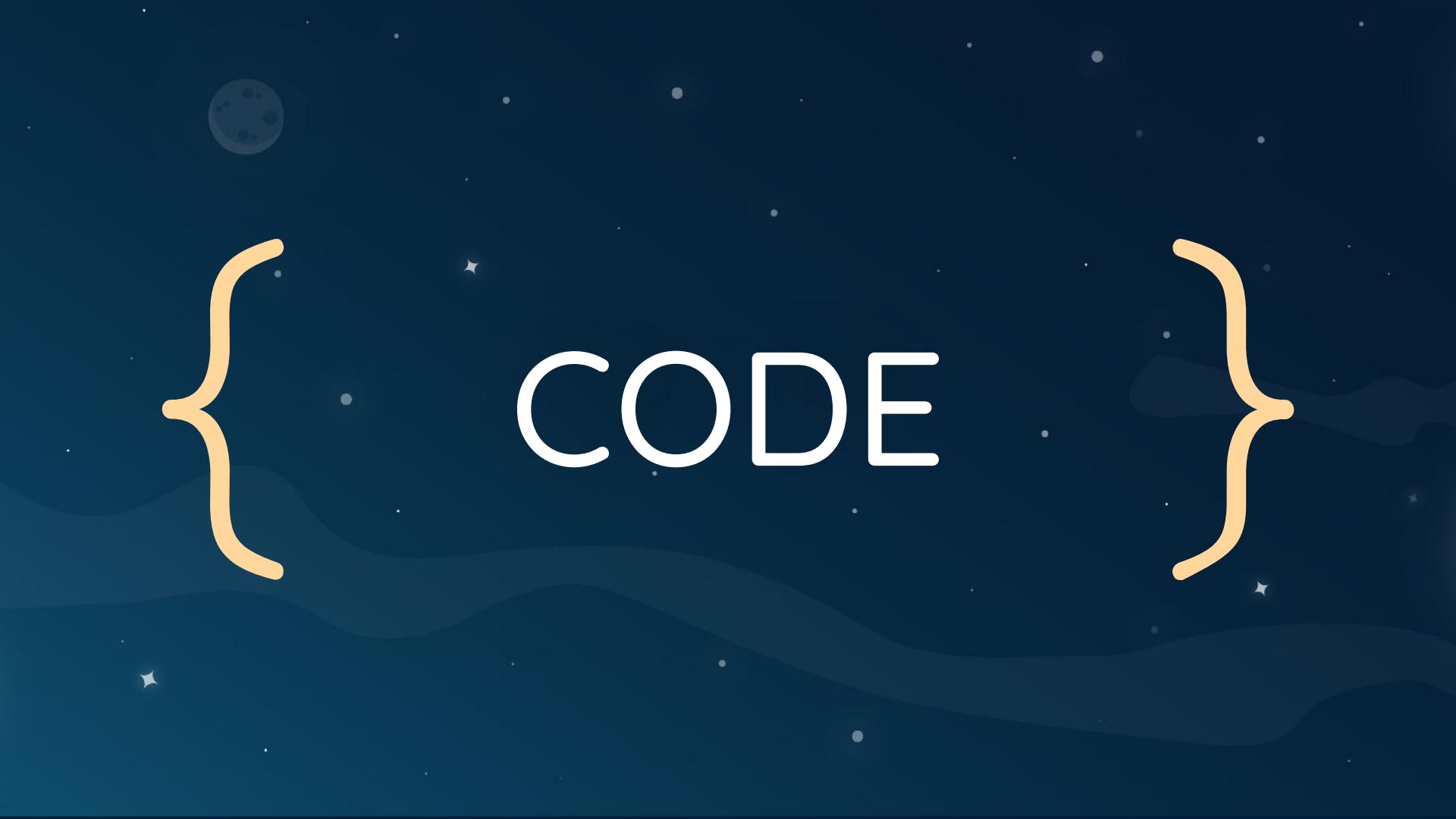
# TOKENIZAÇÃO

O processo de reduzir uma palavra aos seus radicais é chamado de **estemização**

## TOKENIZAÇÃO

A **lematização** é o processo de *flexionar* uma palavra para determinar o seu lema

**Exemplo:** “correr”, “corre” e “correu”: “correr” é o seu lema



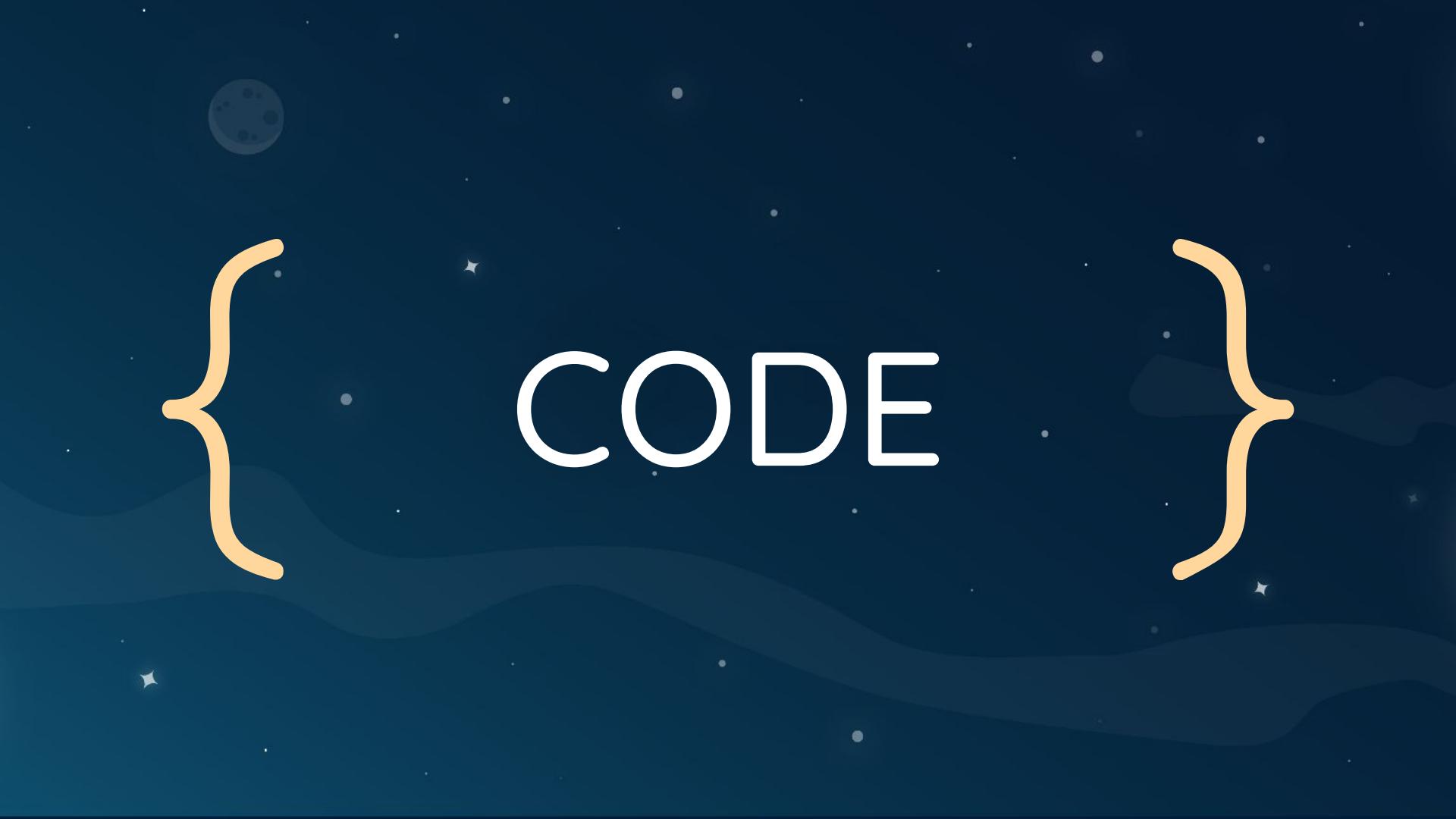
CODE



# INDEXAÇÃO DO VOCABULÁRIO

Com os *tokens* encontrados, precisamos fornecer uma representação numérica para eles

- ▷ Uma forma comum é construir um vocabulário de todos os termos presentes nos textos processados
- ▷ Com o vocabulário, podemos criar uma codificação *one-hot*



CODE

# REPRESENTAÇÃO TEXTUAL

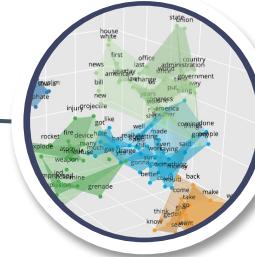


# REPRESENTANDO PALAVRAS



## COMO CONJUNTOS

Vetores que podem ser obtidos com modelos *bag-of-words*



## COMO SEQUÊNCIAS

Vetores comumente chamados de *word embeddings*



## COMO MODELOS

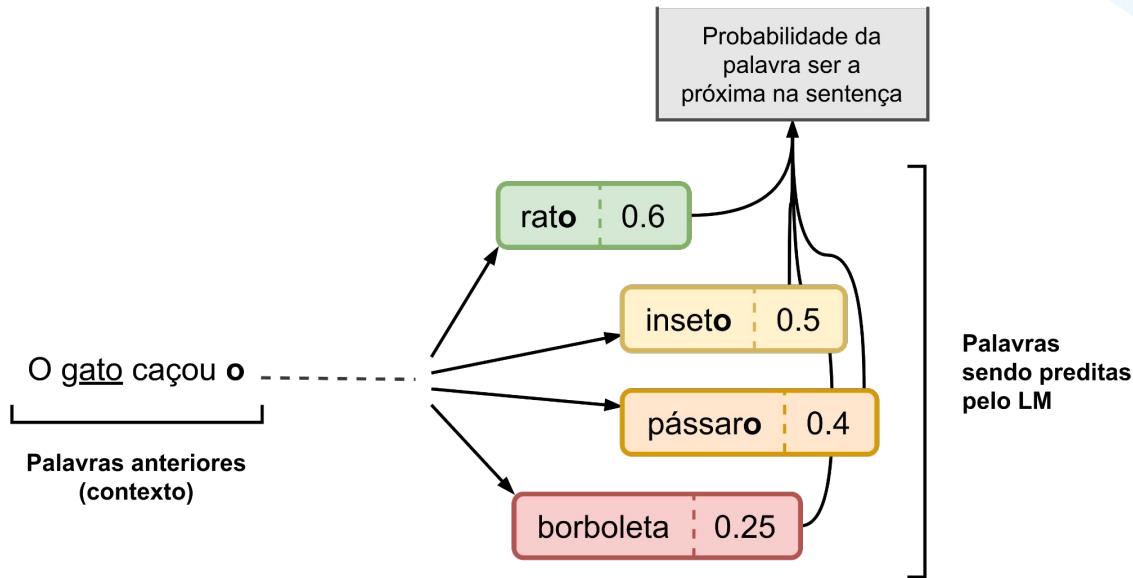
Vetores que podem ser obtidos através de *Language Models*

# O QUE SÃO LANGUAGE MODELS?

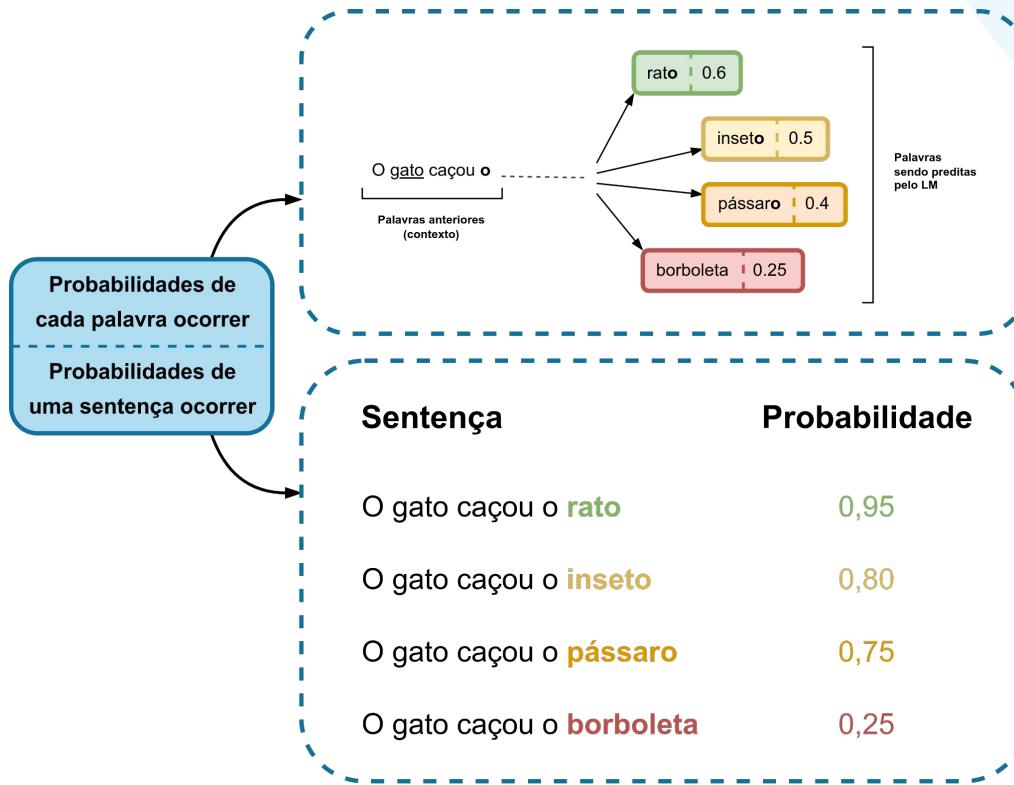
- ▶ Possuem um conceito um pouco abstrato: o de como **modelar uma linguagem**
- ▶ Os chamados *Language Models* (LMs) estimam a **probabilidade** de diferentes **unidades linguísticas**, como símbolos, *tokens*, e sequências de *tokens*
- ▶ Algumas arquiteturas que usam **Transformers**, por exemplo, são LMs



# O QUE SÃO LANGUAGE MODELS?



# O QUE SÃO LANGUAGE MODELS?





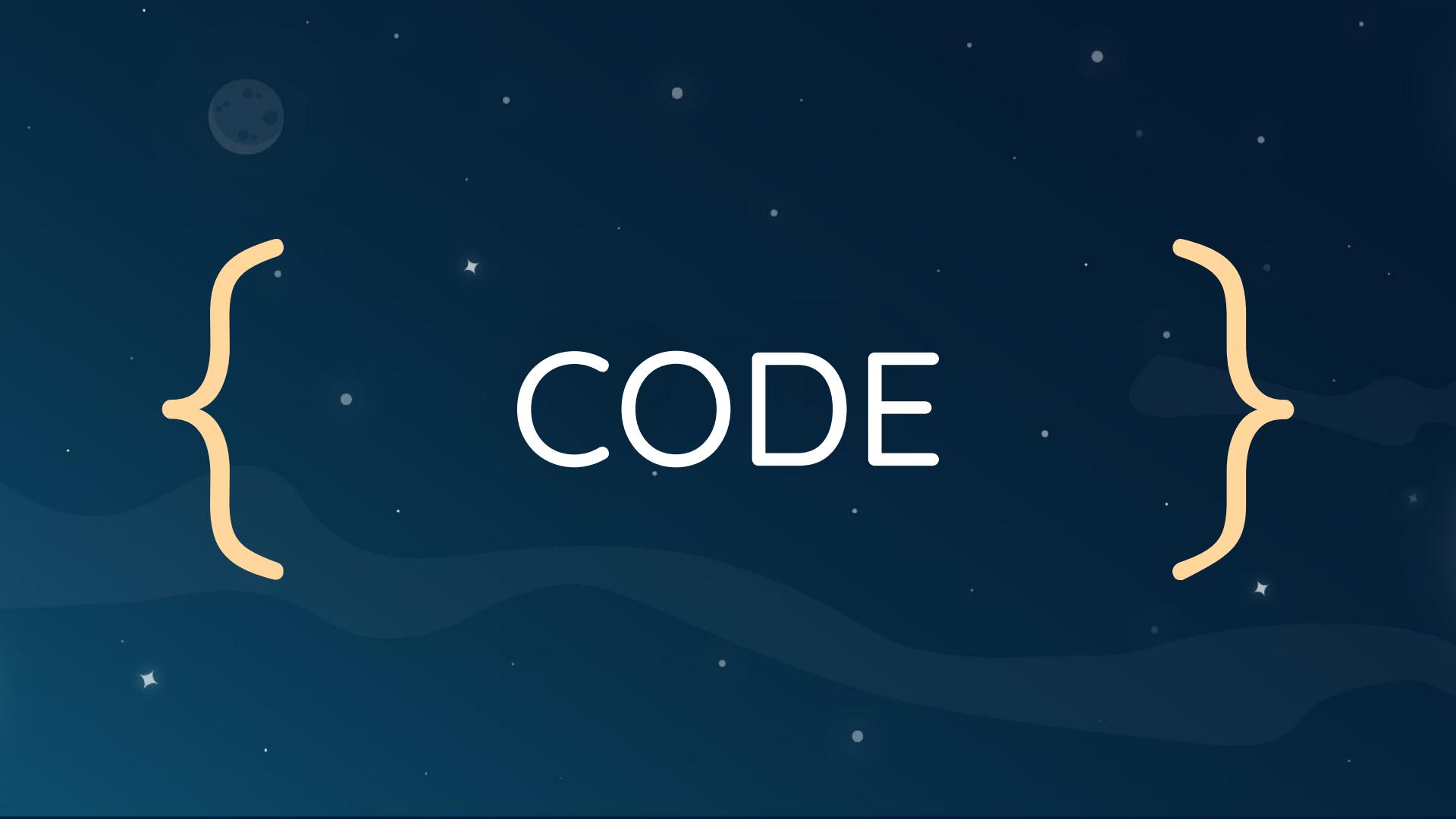
# BAG-OF-WORDS

A maneira mais simples de representar um texto é descartando a ordem e tratando como um conjunto de *tokens*

→ ***Unigrams com codificação multi-hot***

- *Unigrams* nada mais são do que palavras únicas sem nenhuma ordem
- A intenção é formar um conjunto apenas com as palavras distintas do texto





CODE

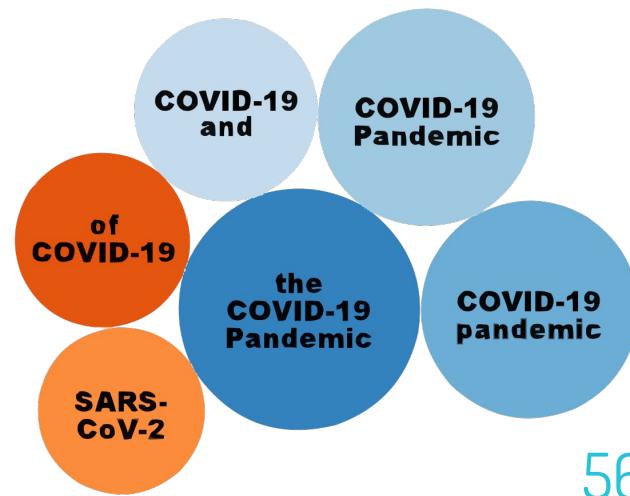


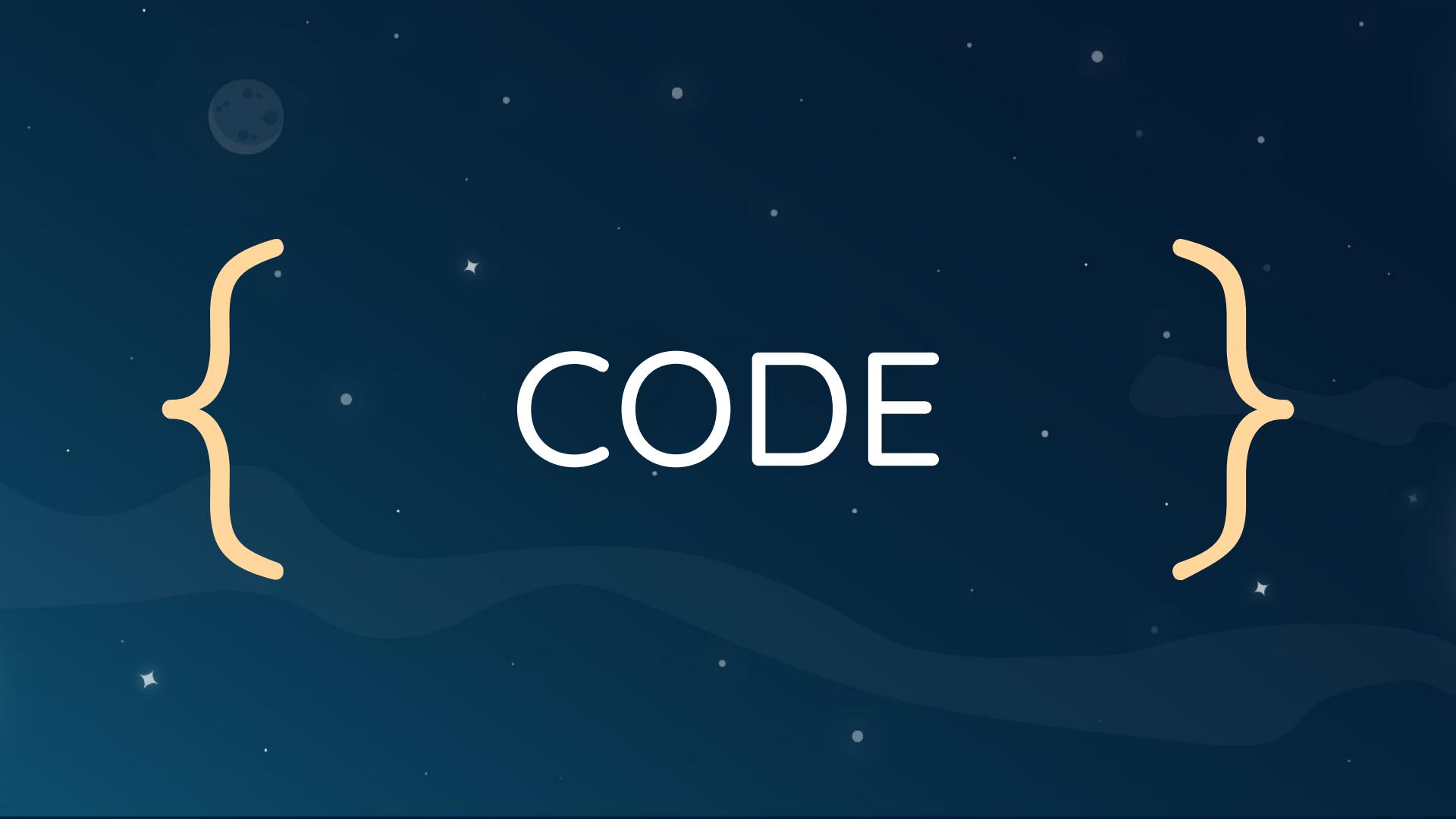
# BAG-OF-WORDS

A maneira mais simples de representar um texto é descartando a ordem e tratando como um conjunto de *tokens*

→ ***N-grams com codificação multi-hot***

- Dependendo do contexto, descartar a ordem das palavras pode não contribuir para o seu entendimento
- Talvez seja necessário inserir informações extras para uma representação de *bag-of-words* olhando para os *n-grams*





CODE

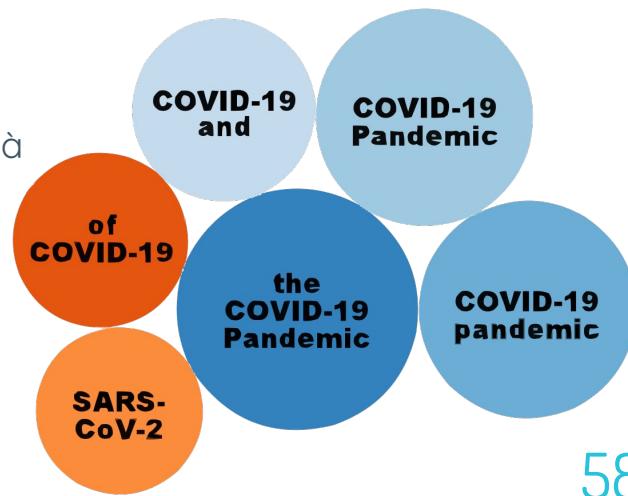


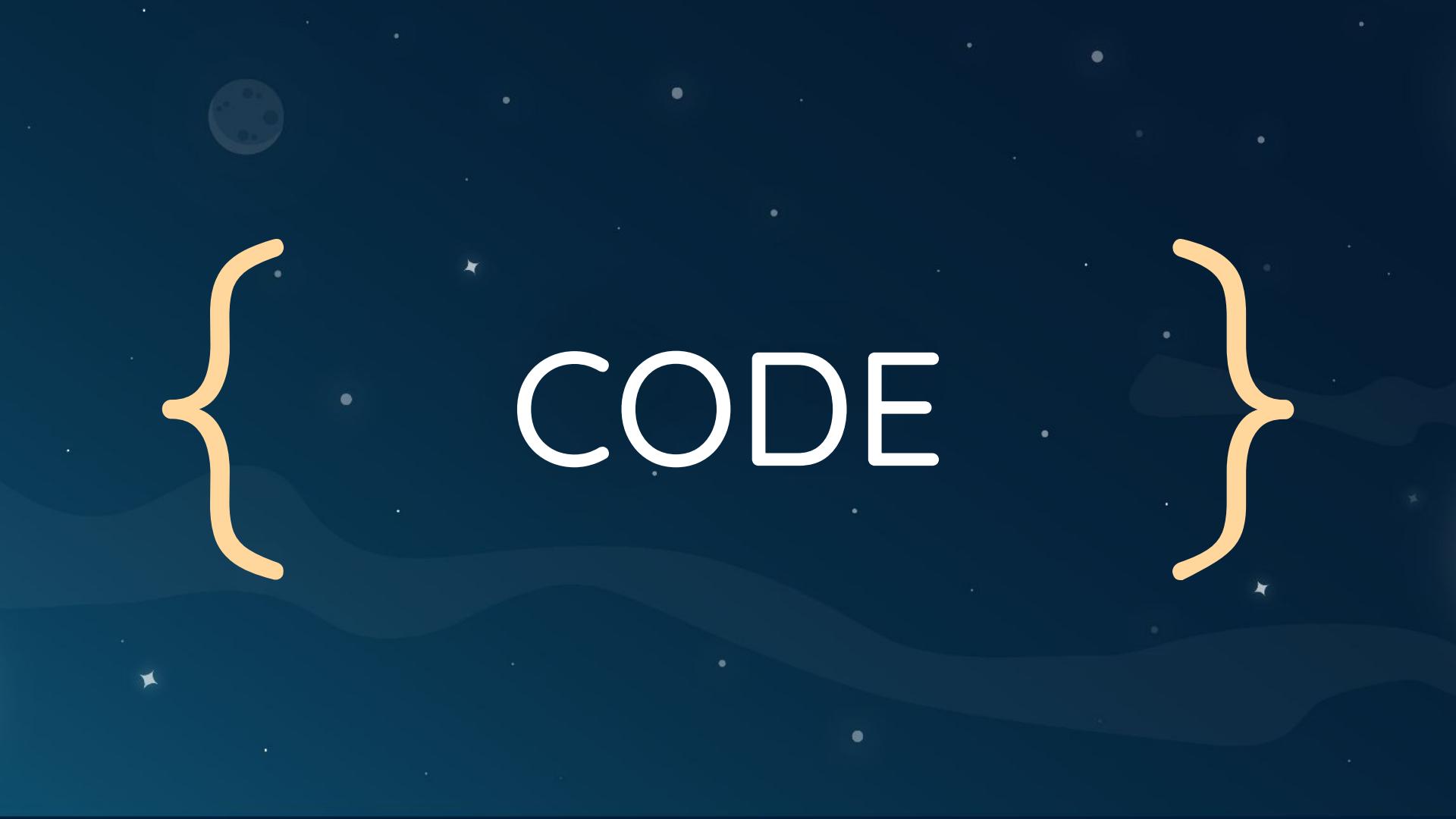
# BAG-OF-WORDS

A maneira mais simples de representar um texto é descartando a ordem e tratando como um conjunto de *tokens*

→ ***N-grams com codificação TF-IDF***

- Também é possível adicionar mais informações à representação por *bag-of-words* ao elencar quantas vezes cada palavra ou *n-grams* ocorrem no texto
- O TF-IDF (*term frequency, inverse document frequency*) é uma métrica que aplica essa ideia





CODE



CODE

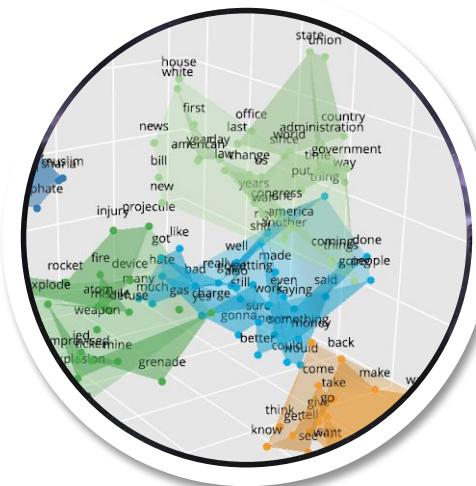
# WORD EMBEDDINGS



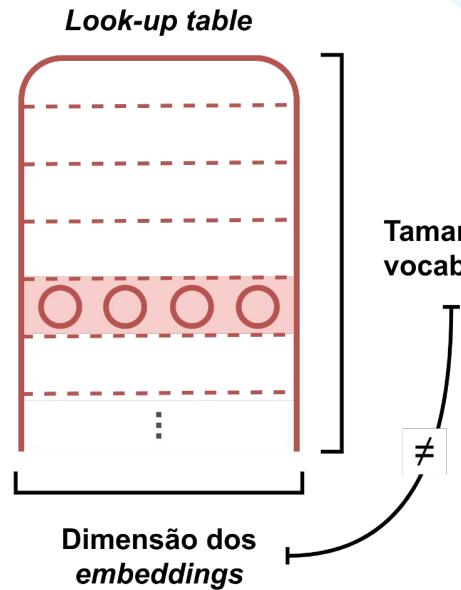
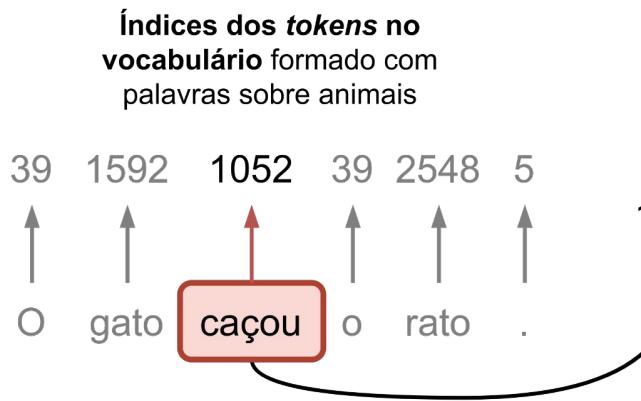
Diferentemente da representação por *bag-of-words*, ***word embeddings*** trabalham com **vocabulários** e consideram a **ordem das palavras**

- O que é um **vocabulário**?

  - Contém palavras pré-selecionadas
  - As palavras podem ser obtidas reunindo todas as palavras únicas de um corpus ou mesmo de um vocabulário preexistente
  - Possui uma espécie de tabela de consulta (uma matriz, chamada de *look-up table*) que contém vetores de *word embeddings*

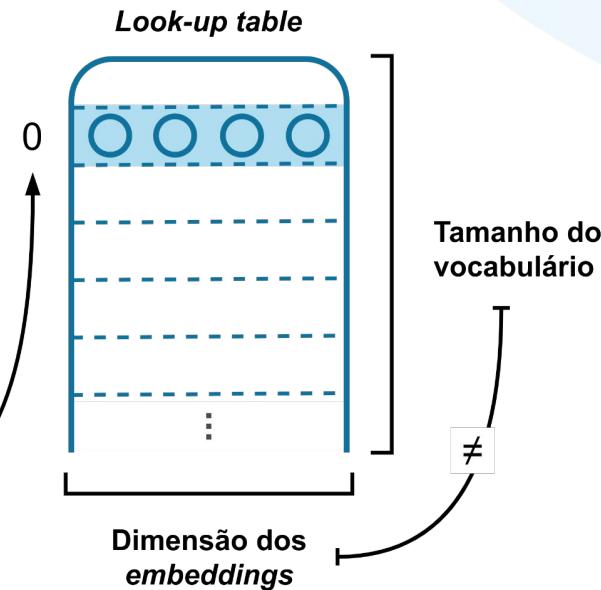


# VOCABULÁRIO



# OUT-OF-VOCABULARY

Alguns *tokens* podem ser “**desconhecidos**”, então é necessário um *token* especial UNK para estes casos



# SÍMBOLOS DISCRETOS

10.000  
PALAVRAS

Palavras	Índices
abelha	1
amor	2
...	...
é	2410
elefante	2411
...	...
galinha	3203
gato	3204
...	...
o	5282
ovelha	5283
...	...
pois	8677
preto	8678
...	...
zumbi	10000

# SÍMBOLOS DISCRETOS

10.000  
PALAVRAS

Palavras Índices	
abelha	1
amor	2
...	...
é	2410
elefante	2411
...	...
galinha	3203
gato	3204
...	...
o	5282
ovelha	5283
...	...
pois	8677
preto	8678
...	...
zumbi	10000

# SÍMBOLOS DISCRETOS

10.000  
PALAVRAS

Palavras	Índices
abelha	1
amor	2
...	...
é	2410
elefante	2411
...	...
galinha	3203
gato	3204
...	...
o	5282
ovelha	5283
...	...
pois	8677
preto	8678
...	...
zumbi	10000



## VOCABULÁRIO

O **vocabulário** em questão é formado com palavras que possuem nomes, características e comportamentos de **animais**

# SÍMBOLOS DISCRETOS

10.000  
PALAVRAS

Palavras	Índices
abelha	1
amor	2
...	...
é	2410
elefante	2411
...	...
galinha	3203
gato	3204
...	...
o	5282
ovelha	5283
...	...
pois	8677
preto	8678
...	...
zumbi	10000

# SÍMBOLOS DISCRETOS

10.000  
PALAVRAS

Palavras	Índices
abelha	1
amor	2
...	...
é	2410
elefante	2411
...	...
galinha	3203
gato	3204
...	...
o	5282
ovelha	5283
...	...
pois	8677
preto	8678
...	...
zumbi	10000

# SÍMBOLOS DISCRETOS

## Palavras Índices

abelha 1

amor 2

...

é 2410

elefante 2411

...

3203

3204

...

5282

5283

...

8677

8678

preto

...

zumbi 10000

→ Primeira palavra: abelha

[0, 1, 0, 0, 0, 0, 0, 0, ..., 0]

seguido de 9.998 zeros

→ Segunda palavra: amor

[0, 0, 1, 0, 0, 0, 0, 0, ..., 0]

seguido 9.997 zeros

# SÍMBOLOS DISCRETOS

## Palavras Índices

abelha 1

amor 2

...

é 2410

elefante 2411

...

3203

3204

...

5282

5283

...

8677

8678

...

zumbi 10000

- Primeira palavra: abelha

[0, 1, 0, 0, 0, 0, 0, 0, ..., 0]

seguido de 9.998 zeros

- Segunda palavra: amor

[0, 0, 1, 0, 0, 0, 0, 0, ..., 0]

seguido 9.997 zeros

preto

...

zumbi

## PROBLEMAS

Com a codificação one-hot os vetores são grandes, esparsos, e sabem nada sobre significado

# SÍMBOLOS DISCRETOS

## Palavras Índices

abelha	1
amor	2
...	...
é	2410
elefante	2411
...	...
	3203
	3204
...	...
	5282
	5283
...	...
	8677
	8678
...	...
zumbi	10000

→ Primeira palavra: abelha

[0, 1, 0, 0, 0, 0, 0, 0, ..., 0]

seguido de 9.998 zeros

→ Segunda palavra: amor

[0, 0, 1, 0, 0, 0, 0, 0, ..., 0]

seguido 9.997 zeros

preto

...

zumbi

## PROBLEMAS

Com a codificação one-hot os vetores são grandes, esparsos, e sabem nada sobre **significado**

## O QUE É “SIGNIFICADO”?

Vocês sabem o que a palavra **tezgüino** significa?

## O QUE É “SIGNIFICADO”?

Vocês sabem o que a palavra **tezgüino** significa?

- Agora veja como esta palavra é usada em **diferentes contextos**:
  - i. Uma garrafa de **tezgüino** está na mesa
  - ii. Todo mundo gosta de **tezgüino**
  - iii. **Tezgüino** deixa você bêbado
  - iv. Fazemos **tezgüino** de milho

# O QUE É “SIGNIFICADO”?

Vocês sabem o que a palavra **tezgüino** significa?

- Agora veja como esta palavra é usada em **diferentes contextos**:
- i. Uma garrafa de **tezgüino** está na mesa
  - ii. Todo mundo gosta de **tezgüino**
  - iii. **Tezgüino** deixa você bêbado
  - iv. Fazemos **tezgüino** de milho

Agora é possível entender o que **tezgüino** significa?



# O QUE É “SIGNIFICADO”?

Vocês sabem o que a palavra **tezgüino** significa?

- Agora veja como esta palavra é usada em **diferentes contextos**:
- i. Uma garrafa de **tezgüino** está na mesa
  - ii. Todo mundo gosta de **tezgüino**
  - iii. **Tezgüino** deixa você bêbado
  - iv. Fazemos **tezgüino** de milho

Agora é possível entender o que **tezgüino** significa? **Tezgüino** é uma espécie de bebida alcoólica feita de milho



## O QUE É “SIGNIFICADO”?

Como é possível obter contexto vetorialmente?

# O QUE É “SIGNIFICADO”?

Como é possível obter contexto vetorialmente?

- Que outras palavras se encaixam nesses contextos?
  - i. Uma garrafa de \_\_\_\_\_ está na mesa
  - ii. Todo mundo gosta de \_\_\_\_\_
  - iii. \_\_\_\_\_ deixa você bêbado
  - iv. Fazemos \_\_\_\_\_ de milho

# O QUE É “SIGNIFICADO”?

Como é possível obter contexto vetorialmente?

- Que outras palavras se encaixam nesses contextos?
- Uma garrafa de \_\_\_\_\_ está na mesa
  - Todo mundo gosta de \_\_\_\_\_
  - \_\_\_\_\_ deixa você bêbado
  - Fazemos \_\_\_\_\_ de milho

	FRASE i	FRASE ii	FRASE iii	FRASE iv
tezgüino	1	1	1	1
alto	0	0	0	0
óleo	1	0	0	1
cuscuz	0	1	0	1
vinho	1	1	1	0

# O QUE É “SIGNIFICADO”?

Como é possível obter contexto vetorialmente?

- Que outras palavras se encaixam nesses contextos?
  - i. Uma garrafa de \_\_\_\_\_ está na mesa
  - ii. Todo mundo gosta de \_\_\_\_\_
  - iii. \_\_\_\_\_ deixa você bêbado
  - iv. Fazemos \_\_\_\_\_ de milho

	FRASE i	FRASE ii	FRASE iii	FRASE iv
tezgüino	1	1	1	1
alto	0	0	0	0
óleo	1	0	0	1
cuscuz	0	1	0	1
vinho	1	1	1	0

# O QUE É “SIGNIFICADO”?

Como é possível obter contexto vetorialmente?

- Que outras palavras se encaixam nesses contextos?
  - i. Uma garrafa de \_\_\_\_\_ está na mesa
  - ii. Todo mundo gosta de \_\_\_\_\_
  - iii. \_\_\_\_\_ deixa você bêbado
  - iv. Fazemos \_\_\_\_\_ de milho

LINHAS SIMILARES

Significados das palavras são  
semelhantes = hipótese distributiva

	FRASE i	FRASE ii	FRASE iii	FRASE iv
tezgüino	1	1	1	1
alto	0	0	0	0
óleo	1	0	0	1
CUSCUZ	0	1	0	1
vinho	1	1	1	0

# DISTRIBUTIONAL HYPOTHESIS

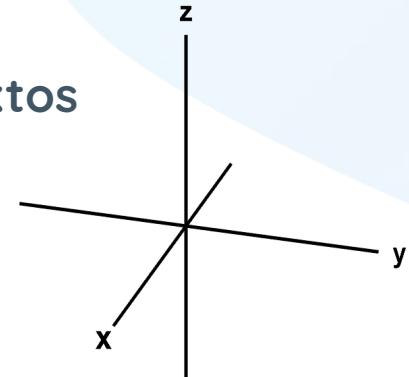
Palavras que **frequentemente** aparecem em **contextos** semelhantes têm **significado** semelhante

	FRASE i	FRASE ii	FRASE iii	FRASE iv
tezgüino	1	1	1	1
alto	0	0	0	0
óleo	1	0	0	1
cuscuz	0	1	0	1
vinho	1	1	1	0

# DISTRIBUTIONAL HYPOTHESIS

Palavras que **frequentemente** aparecem em **contextos** semelhantes têm **significado** semelhante

	FRASE i	FRASE ii	FRASE iii	FRASE iv
tezgüino	1	1	1	1
alto	0	0	0	0
óleo	1	0	0	1
cuscuz	0	1	0	1
vinho	1	1	1	0

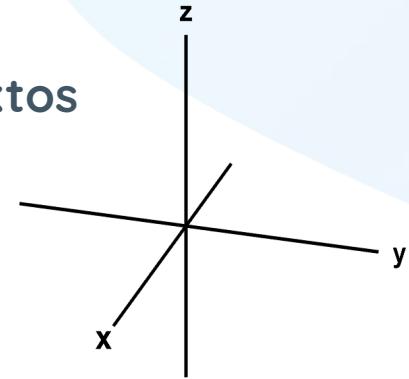


Cada **atributo semântico** pode ser considerado como uma **dimensão única** em um espaço semântico mais amplo e de maior dimensão

# DISTRIBUTIONAL HYPOTHESIS

Palavras que **frequentemente** aparecem em **contextos** semelhantes têm **significado** semelhante

	RECIPIENTE	AFINIDADE	BEBIDA	ALIMENTO
tezgüino	1	1	1	1
alto	0	0	0	0
óleo	1	0	0	1
cuscuz	0	1	0	1
vinho	1	1	1	0



Cada **atributo semântico** pode ser considerado como uma **dimensão única** em um espaço semântico mais amplo e de maior dimensão

# DISTRIBUTIONAL HYPOTHESIS

Palavras que **frequentemente** aparecem em **contextos** semelhantes têm **significado** semelhante

	RECIPIENTE	AFINIDADE	BEBIDA	ALIMENTO
tezgüino	1	1	1	1
alto	0	0	0	0
óleo	1	0	0	1
cuscuz	0	1	0	1
vinho	1	1	1	0

# DISTRIBUTIONAL HYPOTHESIS

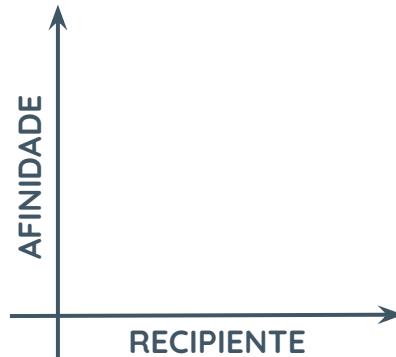
Palavras que **frequentemente** aparecem em **contextos** semelhantes têm **significado** semelhante

	RECIPIENTE	AFINIDADE	BEBIDA	ALIMENTO
tezgüino	1	1	1	1
alto	0	0	0	0
óleo	1	0	0	1
cuscuz	0	1	0	1
vinho	1	1	1	0

# DISTRIBUTIONAL HYPOTHESIS

Palavras que **frequentemente** aparecem em **contextos** semelhantes têm **significado** semelhante

	RECIPIENTE	AFINIDADE	BEBIDA	ALIMENTO
tezgüino	1	1	1	1
alto	0	0	0	0
óleo	1	0	0	1
CUSCUZ	0	1	0	1
vinho	1	1	1	0



# DISTRIBUTIONAL HYPOTHESIS

Palavras que **frequentemente** aparecem em **contextos** semelhantes têm **significado** semelhante

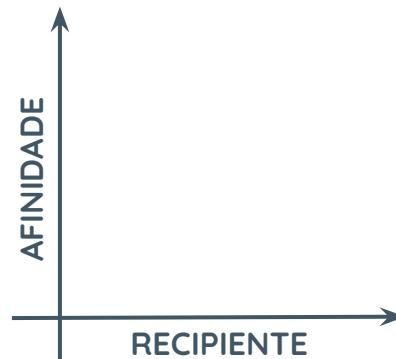
	RECIPIENTE	AFINIDADE	BEBIDA	ALIMENTO
tezgüino	1	1	1	1
alto	0	0	0	0
óleo	1	0	0	1
CUSCUZ	0	1	0	1
vinho	1	1	1	0



# DISTRIBUTIONAL HYPOTHESIS

Palavras que **frequentemente** aparecem em **contextos** semelhantes têm **significado** semelhante

	RECIPIENTE	AFINIDADE	BEBIDA	ALIMENTO
tezgüino	0.92	0.85	0.99	0.98
alto	0.01	0.01	0.01	0.01
óleo	0.75	0.01	0.20	0.45
cuscuz	0.07	0.99	0.01	0.99
vinho	0.97	0.95	0.98	0.05



# DISTRIBUTIONAL HYPOTHESIS

Palavras que **frequentemente** aparecem em **contextos** semelhantes têm **significado** semelhante

	RECIPIENTE	AFINIDADE	BEBIDA	ALIMENTO
tezgüino	0.92	0.85	0.99	0.98
alto	0.01	0.01	0.01	0.01
óleo	0.75	0.01	0.20	0.45
cuscuz	0.07	0.99	0.01	0.99
vinho	0.97	0.95	0.98	0.05



# DISTRIBUTIONAL HYPOTHESIS

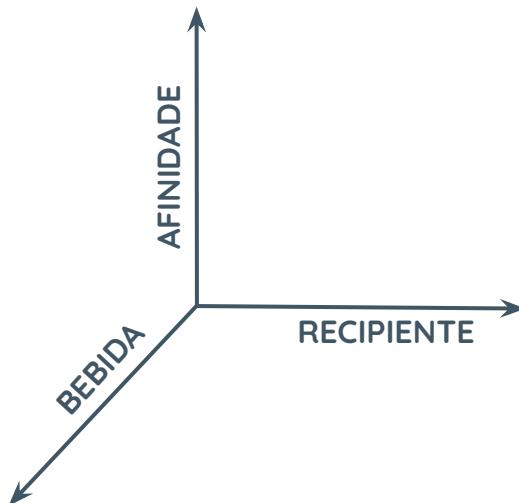
Palavras que **frequentemente** aparecem em **contextos** semelhantes têm **significado** semelhante

	RECIPIENTE	AFINIDADE	BEBIDA	ALIMENTO
tezgüino	0.92	0.85	0.99	0.98
alto	0.01	0.01	0.01	0.01
óleo	0.75	0.01	0.20	0.45
cuscuz	0.07	0.99	0.01	0.99
vinho	0.97	0.95	0.98	0.05

# DISTRIBUTIONAL HYPOTHESIS

Palavras que **frequentemente** aparecem em **contextos** semelhantes têm **significado** semelhante

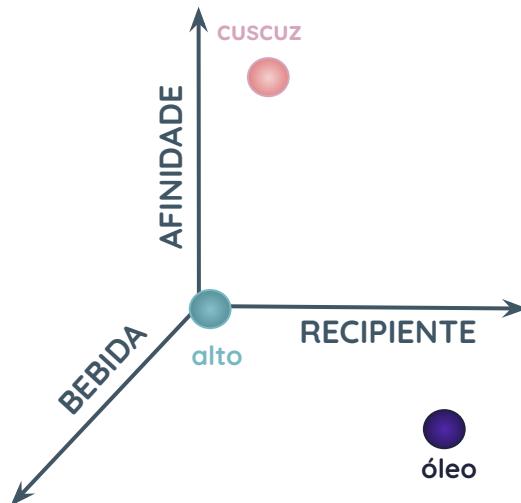
	RECIPIENTE	AFINIDADE	BEBIDA	ALIMENTO
tezgüino	0.92	0.85	0.99	0.98
alto	0.01	0.01	0.01	0.01
óleo	0.75	0.01	0.20	0.45
cuscuz	0.07	0.99	0.01	0.99
vinho	0.97	0.95	0.98	0.05



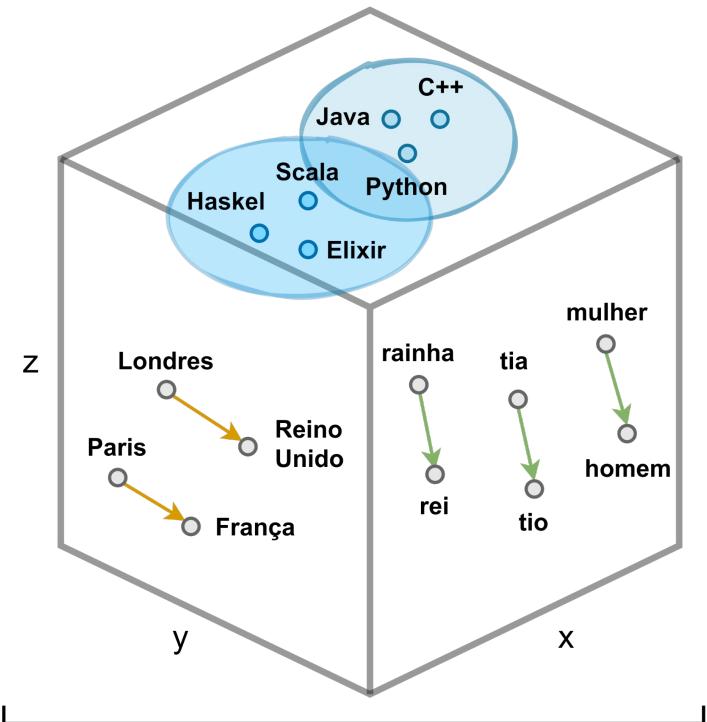
# DISTRIBUTIONAL HYPOTHESIS

Palavras que **frequentemente** aparecem em **contextos** semelhantes têm **significado** semelhante

	RECIPIENTE	AFINIDADE	BEBIDA	ALIMENTO
tezgüino	0.92	0.85	0.99	0.98
alto	0.01	0.01	0.01	0.01
óleo	0.75	0.01	0.20	0.45
cuscuz	0.07	0.99	0.01	0.99
vinho	0.97	0.95	0.98	0.05



# DISTRIBUTIONAL HYPOTHESIS

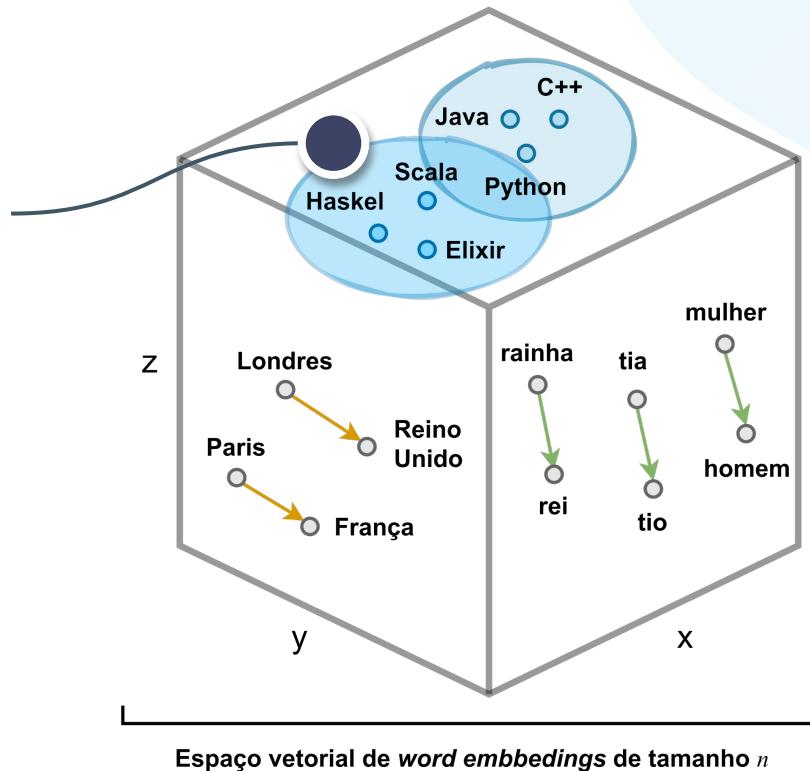


Espaço vetorial de *word embeddings* de tamanho  $n$

# DISTRIBUTIONAL HYPOTHESIS

## CLUSTERS

Podem conter grupos de palavras semelhantes



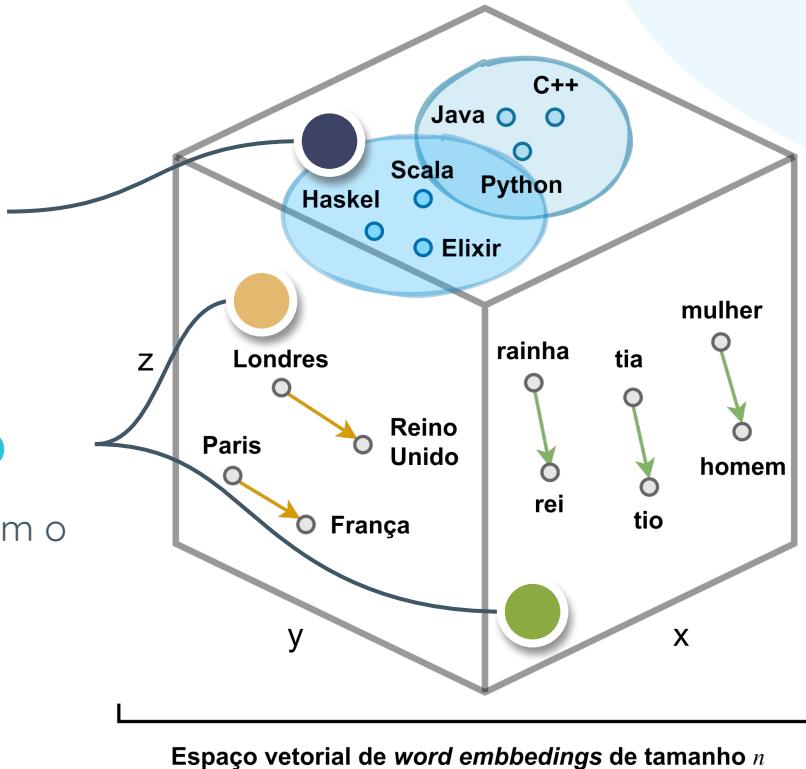
# DISTRIBUTIONAL HYPOTHESIS

## CLUSTERS

Podem conter grupos de palavras semelhantes

## DISTÂNCIA E DIREÇÃO

Essas características mapeiam o significado das palavras no espaço vetorial



# COMO OBTER WORD EMBEDDINGS?

- ▷ Existem **duas maneiras** principais de obter word embeddings
  - 1. Aprender word embeddings em conjunto com a tarefa principal
  - 2. Carregar word embeddings pré-treinadas
- ▷ E existem **dois grandes grupos** de word embeddings
  - ▶ Estáticas
  - ▶ Contextuais

# WORD EMBEDDINGS PRÉ-TREINADAS

- ▶ A principal **lógica** por trás do uso de *word embeddings* pré-treinadas é de que o espaço pré-formado é
  - ▶ Altamente estruturado
  - ▶ Compõe-se de aspectos genéricos da estrutura da linguagem
- ▶ Tais *word embeddings* são geralmente computadas usando **estatísticas das ocorrências** das palavras, usando uma variedade de técnicas
- ▶ Podem ser encontradas gratuitamente para **carregamento e/ou download** em diversos idiomas



CODE

# WORD EMBEDDINGS ESTÁTICAS

- ▷ Também conhecidas como ***word embeddings clássicas***
- ▷ Uma *word embedding* estática trata-se de uma representação de palavra que não varia entre contextos

# WORD EMBEDDINGS ESTÁTICAS

- ▷ Também conhecidas como ***word embeddings clássicas***
- ▷ Uma *word embedding* estática trata-se de uma representação de palavra que não varia entre contextos

Como assim?

# WORD EMBEDDINGS ESTÁTICAS

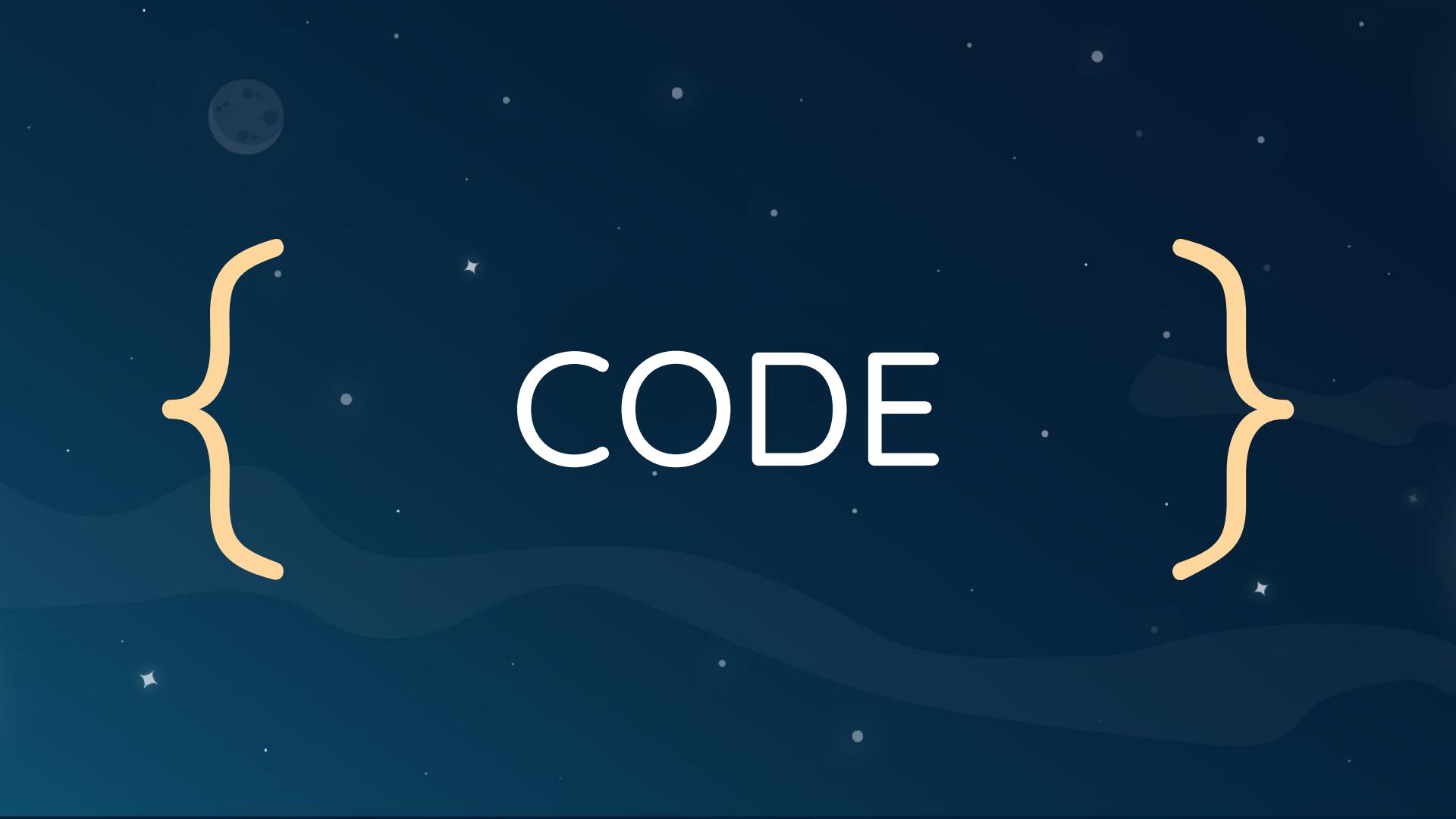
## CONTEXTO 1

“Pesquisei sobre receitas de **bolo** de laranja”



## CONTEXTO 2

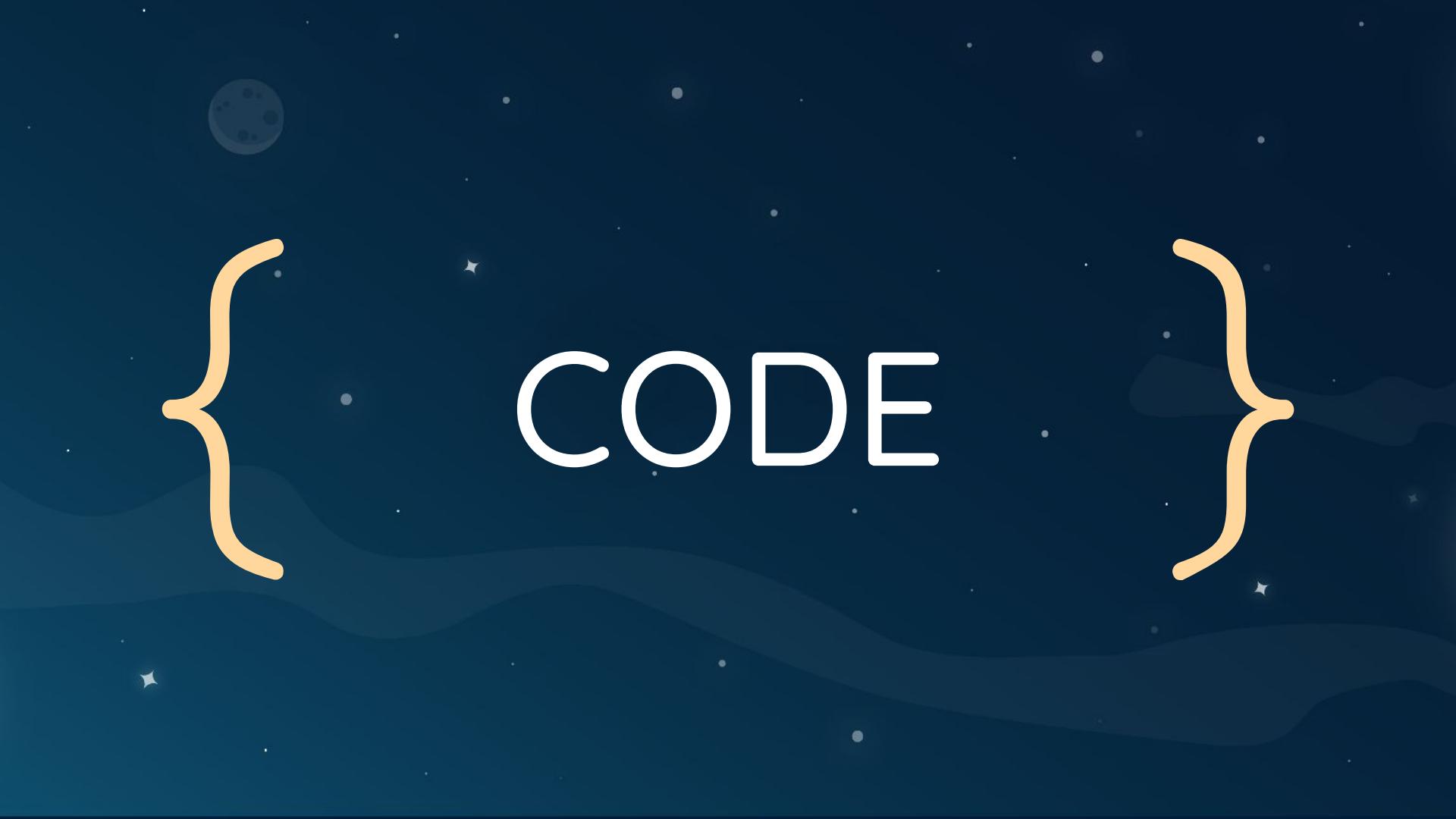
“Há um **bolo** de coisas sobre a mesa”



CODE

# WORD EMBEDDINGS CONTEXTUAIS

- ▷ Possuem propriedades variáveis e são diretamente obtidas a partir do **contexto**
- ▷ *Word embeddings* contextuais, em geral, são extraídas a partir de ***Language Models***
- ▷ Lidam com os **três problemas principais** das estáticas:
  1. *Word embeddings* estáticas não tratam a polissemia
  2. *Word embeddings* contextuais permitem o uso de representação por *n-grams*
  3. Não assumem que o significado de uma palavra depende de contextos similares



CODE



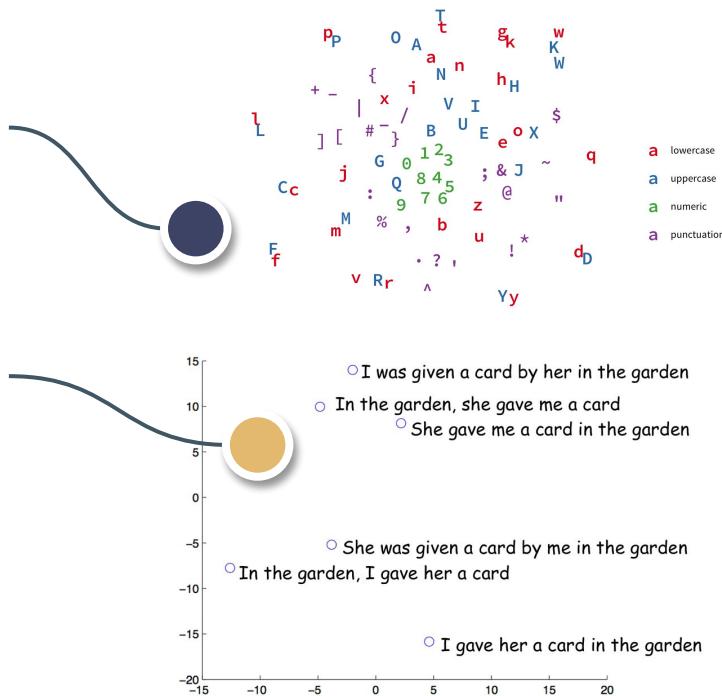
# OUTROS TIPOS

## CHARACTER

*Character embeddings* geralmente são úteis para obter subpalavras

## SENTENCE

*Sentence embeddings* representa frases inteiras e suas informações semânticas como vetores





# O QUE FAZ UMA BOA REPRESENTAÇÃO TEXTUAL?

## BOA REPRESENTAÇÃO

- ▶ O fato de existirem dois grandes grupos principais de *word embeddings* trazem o seguinte questionamento:
  - ▶ Existe algum algoritmo para criação de *word embeddings* ideal que mapeie perfeitamente a linguagem humana e que possa ser usado para qualquer tarefa do Processamento de Linguagem Natural?

## BOA REPRESENTAÇÃO

- ▷ O fato de existirem dois grandes grupos principais de *word embeddings* trazem o seguinte questionamento:
  - ▶ Existe algum algoritmo para criação de *word embeddings* ideal que mapeie perfeitamente a linguagem humana e que possa ser usado para qualquer tarefa do Processamento de Linguagem Natural?



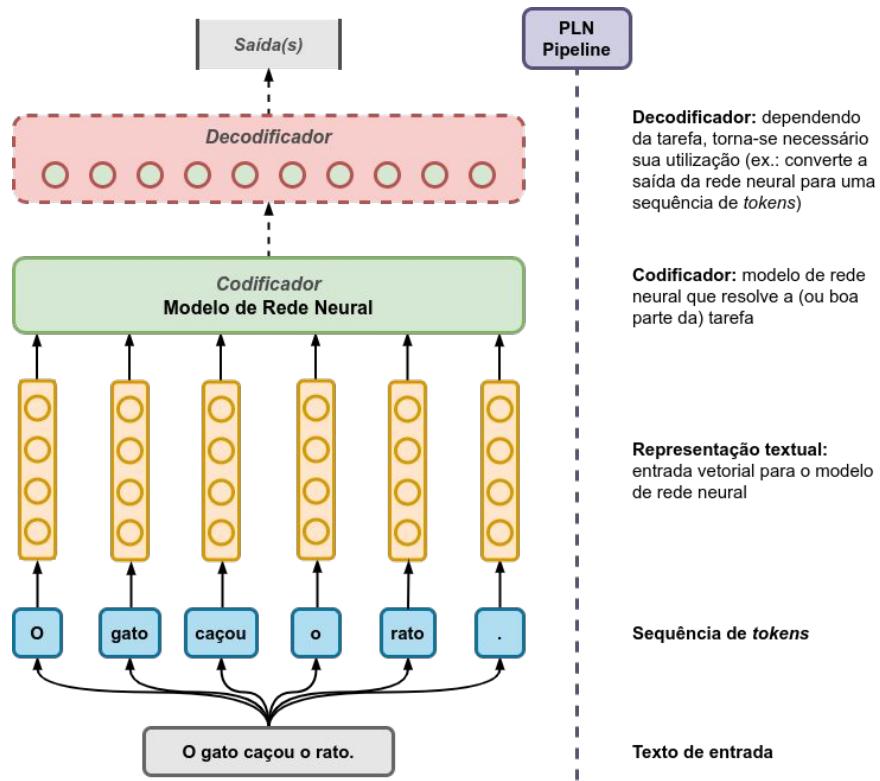
**RESPOSTA (?)**

Sempre depende de muita coisa!



ALÉM DA  
REPRESENTAÇÃO TEXTUAL

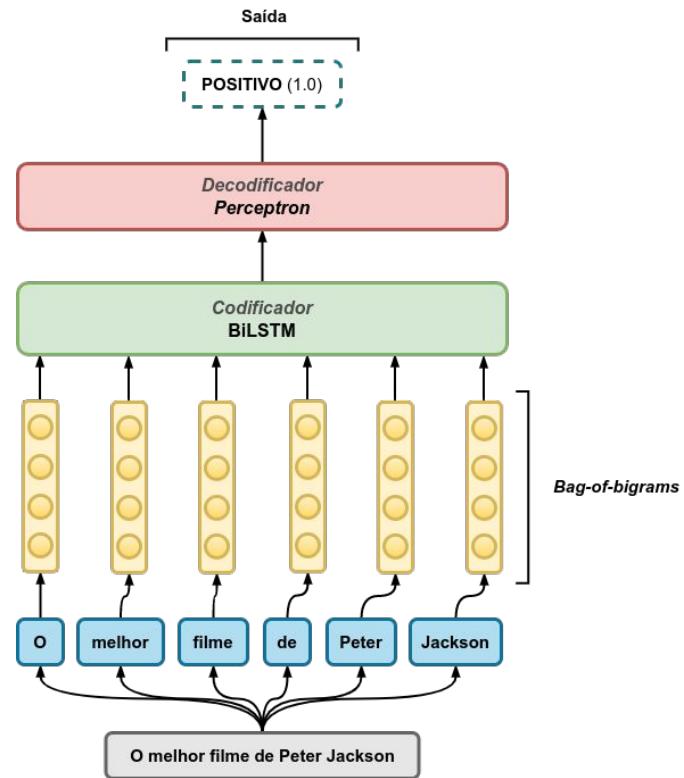
# PIPELINE DE PLN





CODE

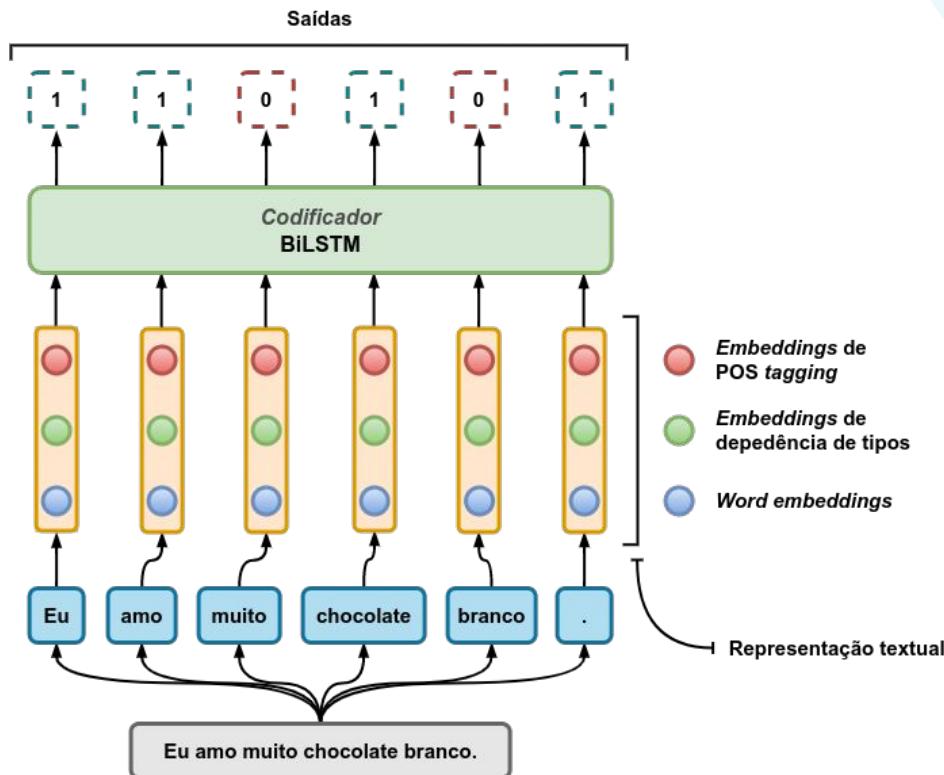
# CLASSIFICAÇÃO DE TEXTOS

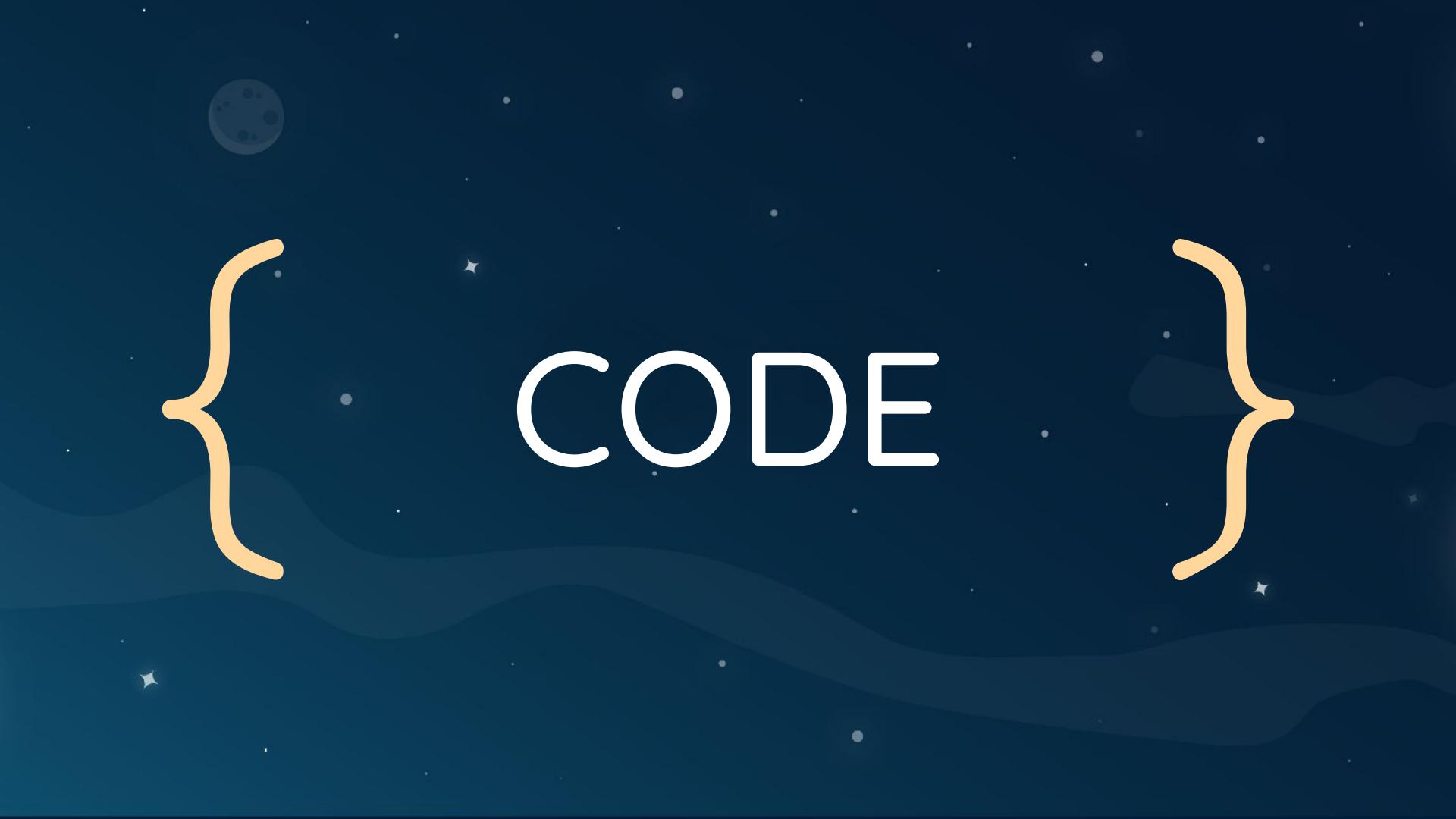




CODE

# SUMARIZAÇÃO DE SENTENÇAS





CODE

# DESAFIOS DE PESQUISA

## DESAFIOS DE PESQUISA

O problema do alto volume de dados gerados não está presente apenas na Web. Também podemos encontrar em áreas como:

- ▷ Setor público
- ▷ Saúde e bem-estar
- ▷ Ação social

# DESAFIOS DE PESQUISA: Setor público

Análise de Boletins de Ocorrência policiais:

- ▷ NER - Extração de informações relevantes dos BOs
- ▷ Classificação de textos - Classificar BOs como roubo, furto, homicídio
- ▷ Identificação de similaridade entre BOs (modus operandi)

# DESAFIOS DE PESQUISA: Saúde e Bem-estar

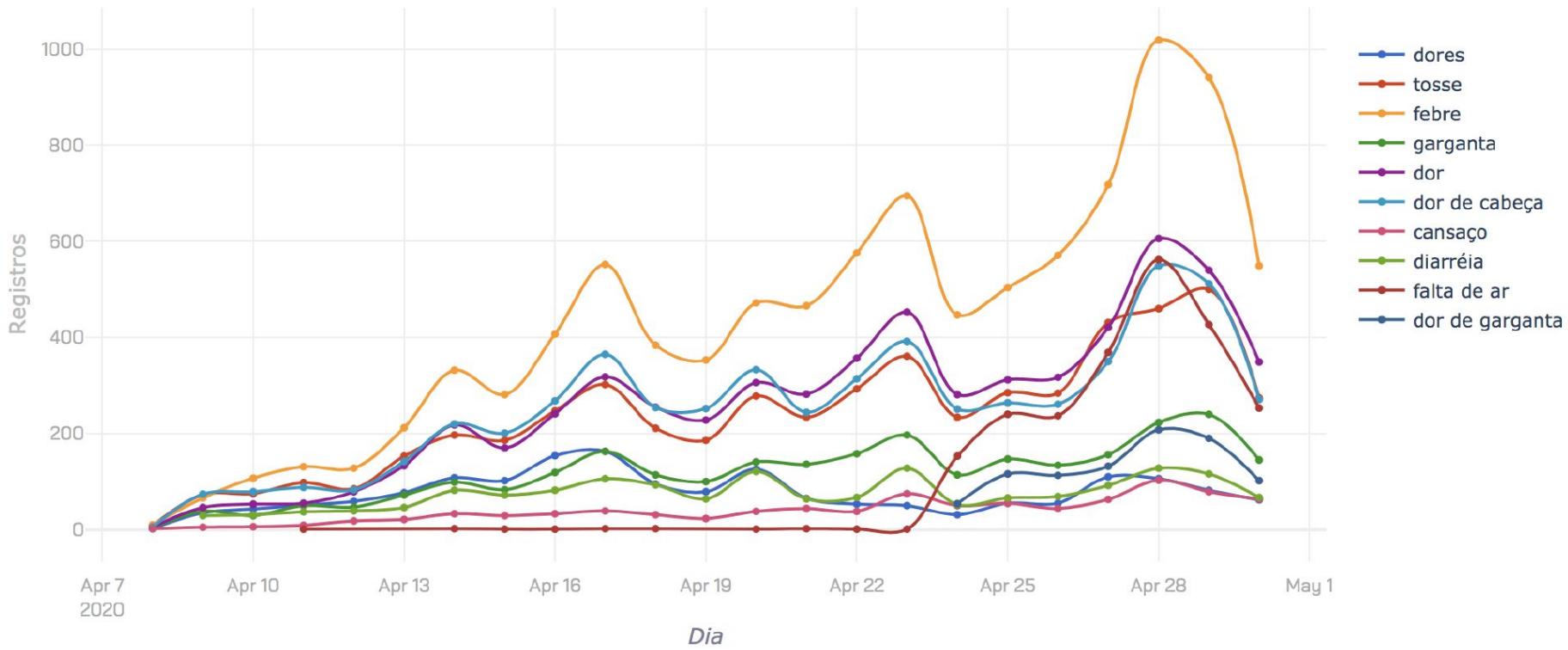
A pandemia de COVID-19 exigiou soluções ágeis para mitigar os efeitos da doença

The screenshot shows two instances of a mobile application interface for 'Plantão Coronavírus'. The first instance on the left shows a message from the bot: 'protegidas e ajudam o Governo do Estado a combater o novo Coronavírus (COVID-19.)' followed by a warning message: '⚠ É muito importante responder todas as perguntas e completar o atendimento. Ao final, se for necessário, uma equipe de saúde estará pronta para conversar e cuidar de você.' Below these are two input fields: 'Você está com algum desses sintomas? 😷' and 'Selecione uma ou mais das opções abaixo.' A red button at the bottom says 'Mal-estar intenso, Tosse incontrolável'. The second instance on the right shows a message from the bot: 'Você é homem, mulher ou prefere não declarar?' with a 'Mulher' button. Below it is another message: 'Você está grávida?' with a 'Não' button. At the bottom, there is a text input field with placeholder 'Escreva aqui...' and a red button with a right arrow.

## PLANTÃO CORONAVÍRUS

*Chatbot* para detecção de sintomas através do modelo *Sintomatic* e redirecionamento para atendimento com profissionais da saúde

## Notificação de Sintomas por dia



## DESAFIOS DE PESQUISA: Ação Social

- ▶ Integração e análise de dados através da plataforma Big Data Social
- ▶ Integração de dados textuais de saúde através da plataforma IntegraSUS
- ▶ Criação de ChatBot com uso de voz para coletar dados das famílias vulneráveis

# CONCLUSÃO

# CONCLUSÃO

Este minicurso apresentou, de forma geral, o processo utilizado no desenvolvimento de aplicações de PLN utilizando técnicas de Aprendizagem Profunda

Alguns assuntos não foram abordados, como:

- ▷ Criação de modelos com pouco volume de dados
- ▷ Uso de arquiteturas mais modernas para tarefas de PLN, como BERT, ELMo e Flair
- ▷ Outras tarefas como Tradução de Textos e NER

# OBRIGADO!

## Dúvidas?

Vocês podem nos encontrar em

- ▷ [barbaraneves@insightlab.ufc.br](mailto:barbaraneves@insightlab.ufc.br)
- ▷ [gustavo.coutinho@insightlab.ufc.br](mailto:gustavo.coutinho@insightlab.ufc.br)
- ▷ [jose.macedo@insightlab.ufc.br](mailto:jose.macedo@insightlab.ufc.br)



# INSIGHT

Data Science Laboratory  
Federal University of Ceará