



PROCESSAMENTO DE LÍNGUA NATURAL



Aplicações

Tradução

Reconhecimento de discurso

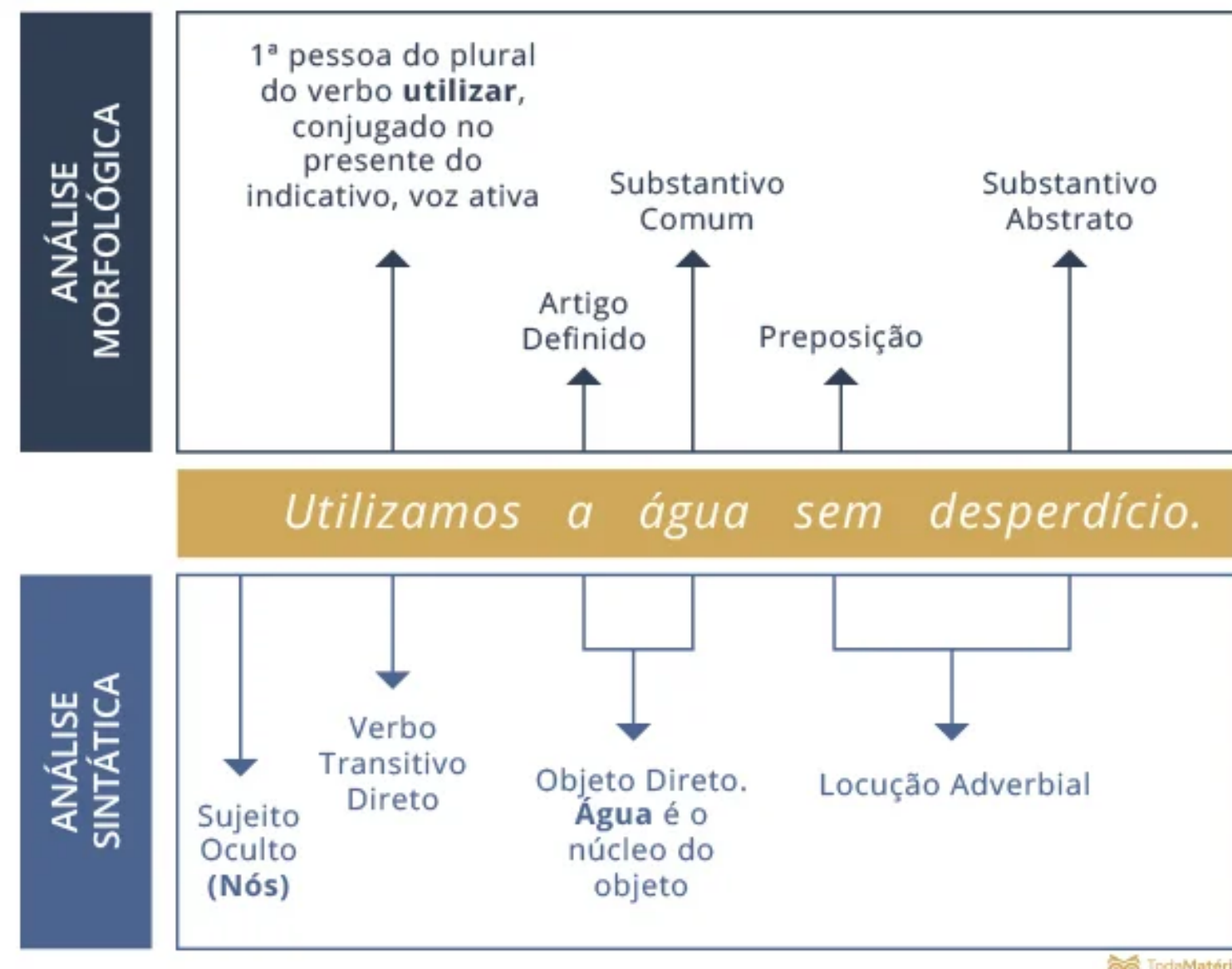
Identificação de notícias falsas

Análise de redes sociais

Geração/resumo de texto

Sintaxe

Estrutura de um texto



Semântica

Sentido de um texto



Toguro
@toguro

...

eu sonhei que o [@elonmusk](#) me pediu um pré treino

2:08 PM · 24 de set de 2022 de São Paulo, Brasil · Twitter for iPhone

1.485 Retweets **148** Tweets com comentário **24,6 mil** Curtidas

Imagem: [todamateria](#)

Problemas

Há informação **suficiente**?

Faz **sentido** resolver um problema com dados textuais?

Qual é a **saída**? Se aprendizado supervisionado, há **rótulos**?

Como são **capturados** os dados?



maia
@r_maiia

Ambiguidade

meta é ir pra academia de casal e treina juntos

5:21 PM · 5 de out de 2022 · Twitter for iOS

106 Retweets 29 Tweets com comentários



Clécio Nunes
@drcance



Quando eu começar a fazer academia vou até trancar algumas disciplinas da universidade que é pra não me atrapalhar.....

11:32 PM · 7 de out de 2022 · Twitter for Android



Allyson Silva, Ph.D.
@lalobackjung

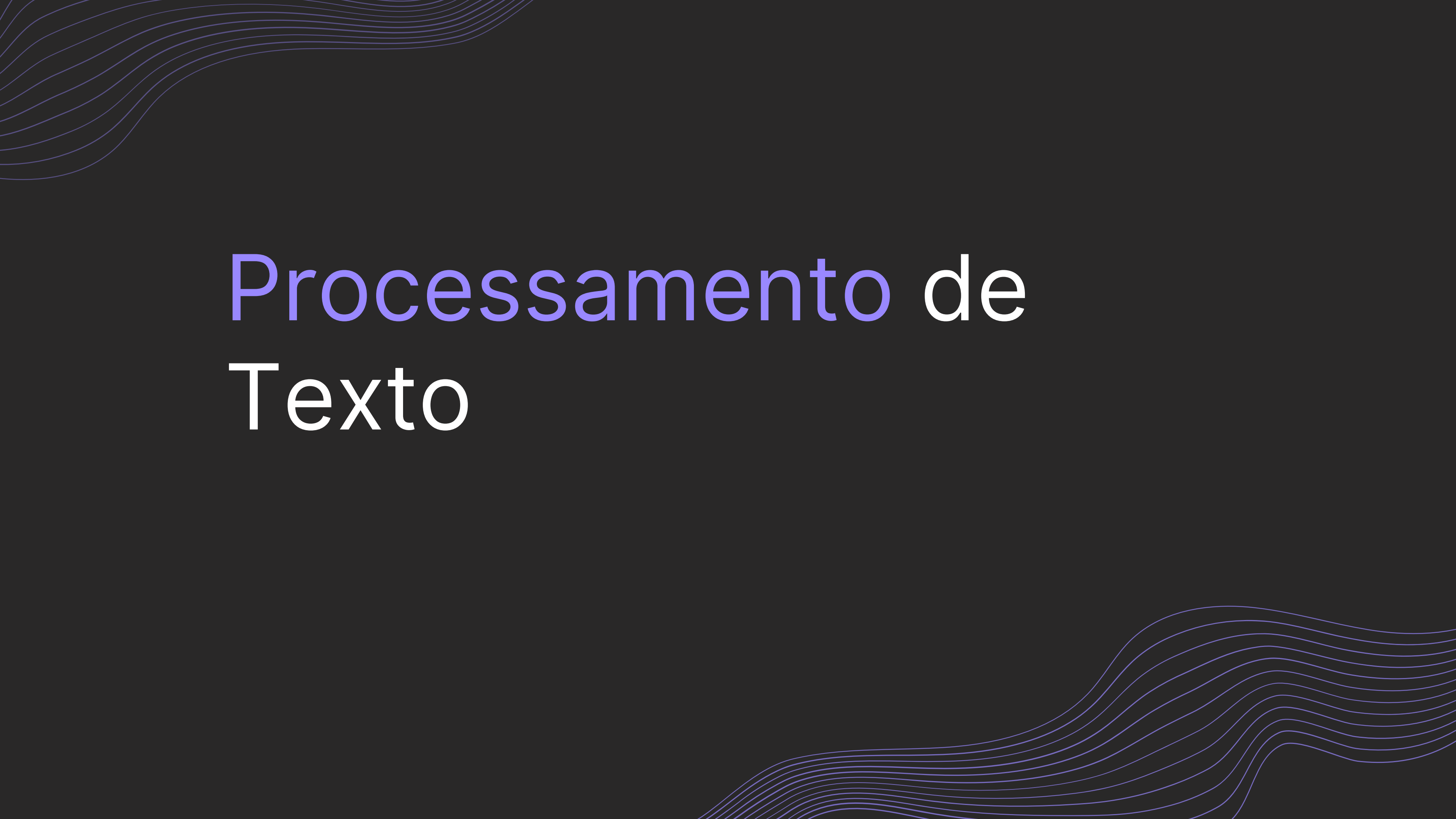


Eu não me conformo que só exista a possibilidade de seguir uma carreira na ciência numa universidade sendo também professor.

Estava conversando essa semana com ótimos cientistas da minha área que disseram, unanimemente, que deixaram a academia pq não queriam ser professores.

1:05 AM · 10 de out de 2022 · Twitter for Android



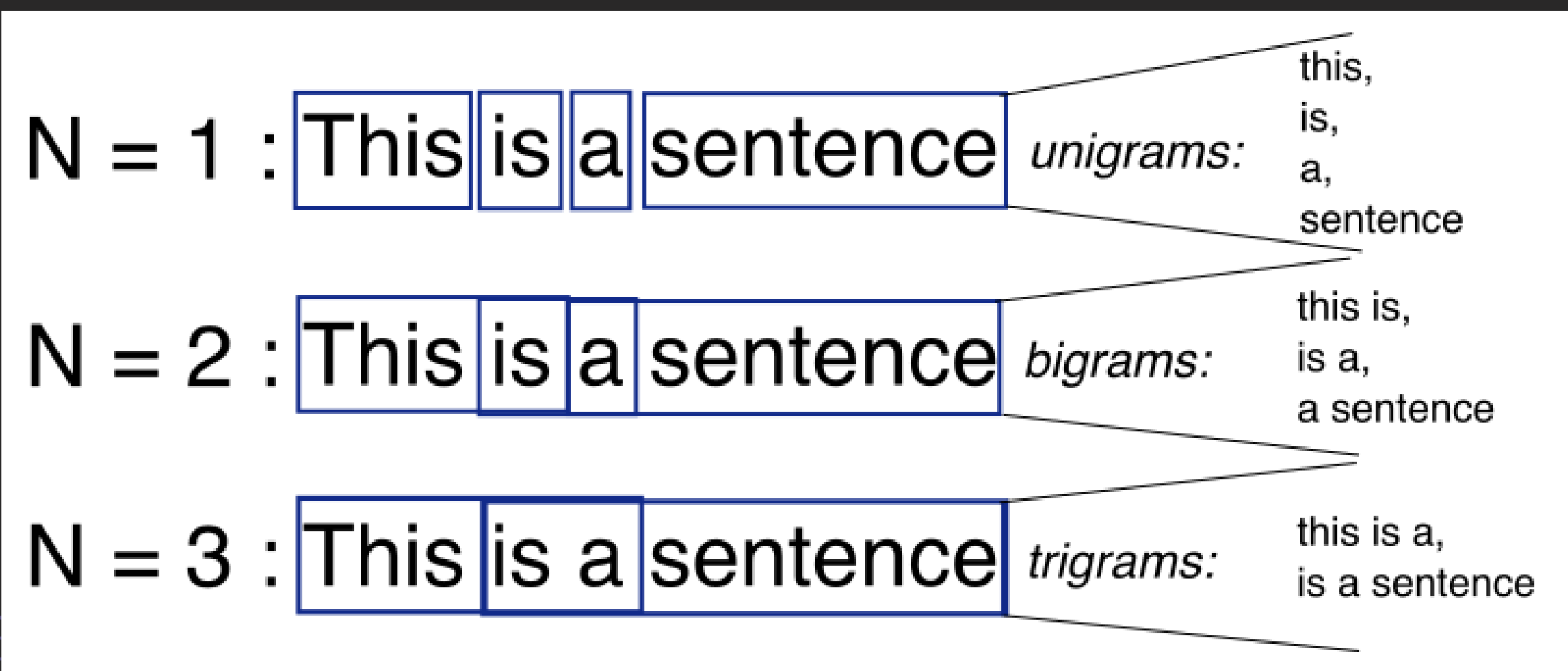


Processamento de Texto

Tokenização

SEPARAÇÃO DOS TERMOS DE UM TEXTO

- Palavras e caracteres
- n-gramas
- Expressões regulares



Redução dos termos

PARECIDOS FICAM IGUAIS

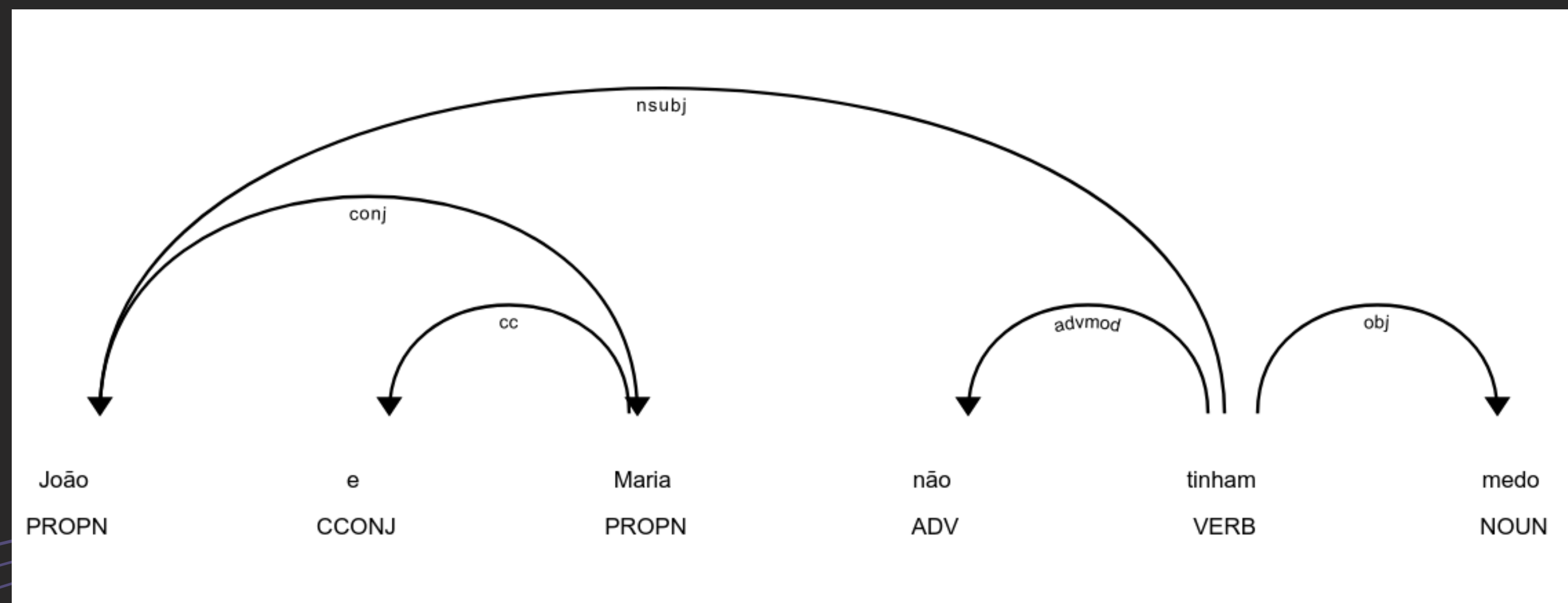
- Lematização
- Stemização

Word	Stemming	Lemmatization
information	inform	information
informative	inform	informative
computers	comput	computer
feet	feet	foot

POS-tagging

INFORMAÇÃO ESTRUTURAL

- Classes gramaticais
- Modelos para inferência





NLTK

spacy

PLN para Aprendizizado de Máquina

Bag of Words Contagem

BoW: Cada palavra vira uma **coluna**. Se um texto tem uma palavra, essa coluna terá como valor a **quantidade** de vezes que o termo aparece.

Muitos valores zero

CountVectorizer: Frequência de tokens (não somente palavras)

Data = ['The', 'quick', 'brown', 'fox', 'jumps', 'over', 'the', 'lazy', 'dog']



	The	quick	brown	fox	jumps	over	lazy	dog
Data	2	1	1	1	1	1	1	1

Extração de atributos por contagem disponível na biblioteca scikit-learn

Imagem: educative.io

Código



Código

Tf-idf

Substitui a abordagem de contagem por um valor dado por uma fórmula

Da **menos** valor a tokens muito frequentes

$$w_{x,y} = tf_{x,y} \times \log \left(\frac{N}{df_x} \right)$$

TF-IDF

Term x within document y

$tf_{x,y}$ = frequency of x in y

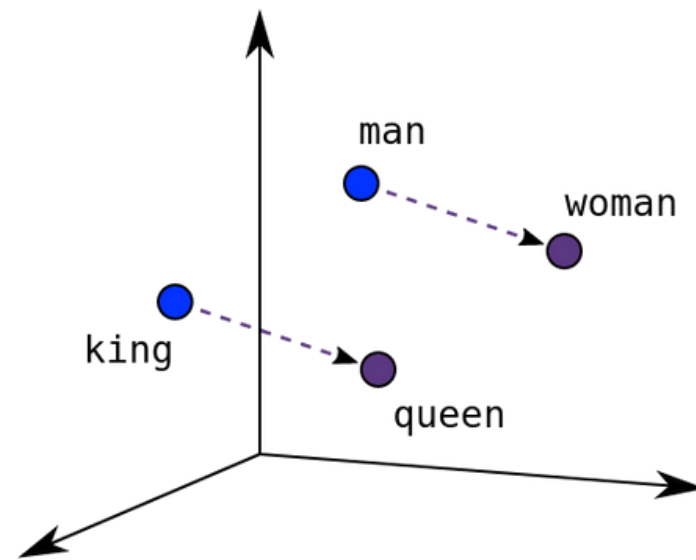
df_x = number of documents containing x

N = total number of documents

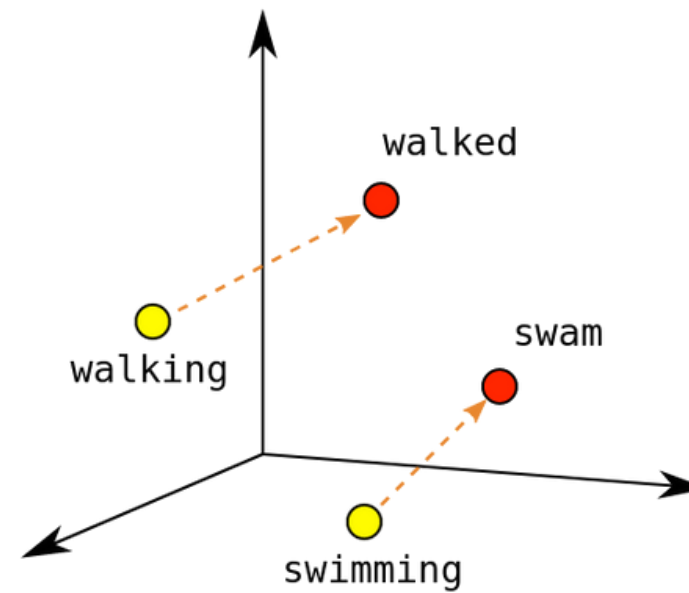
Tf-idf disponível na biblioteca scikit-learn

Imagem: [ted-mei.medium](https://ted-mei.medium.com)

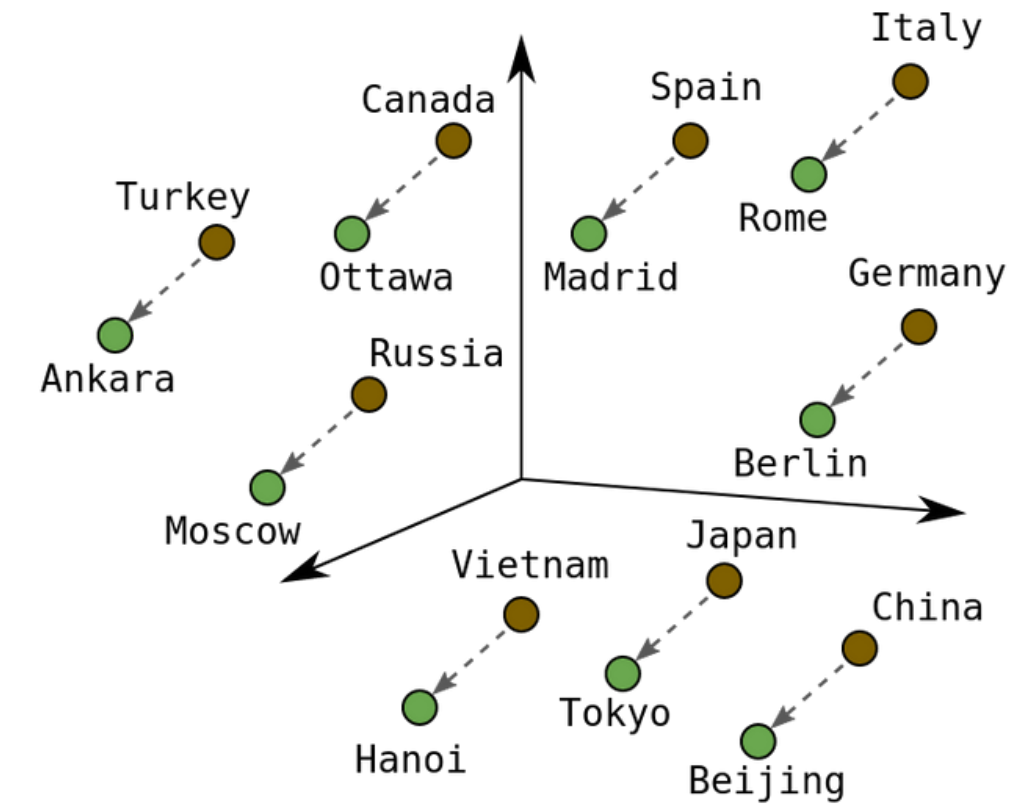
Word embeddings



Male-Female



Verb Tense



Country-Capital

Modelos para **transformar**
palavras em vetores
numéricos de **semelhança**

Algoritmo word2vec disponível na biblioteca [gensim](#)

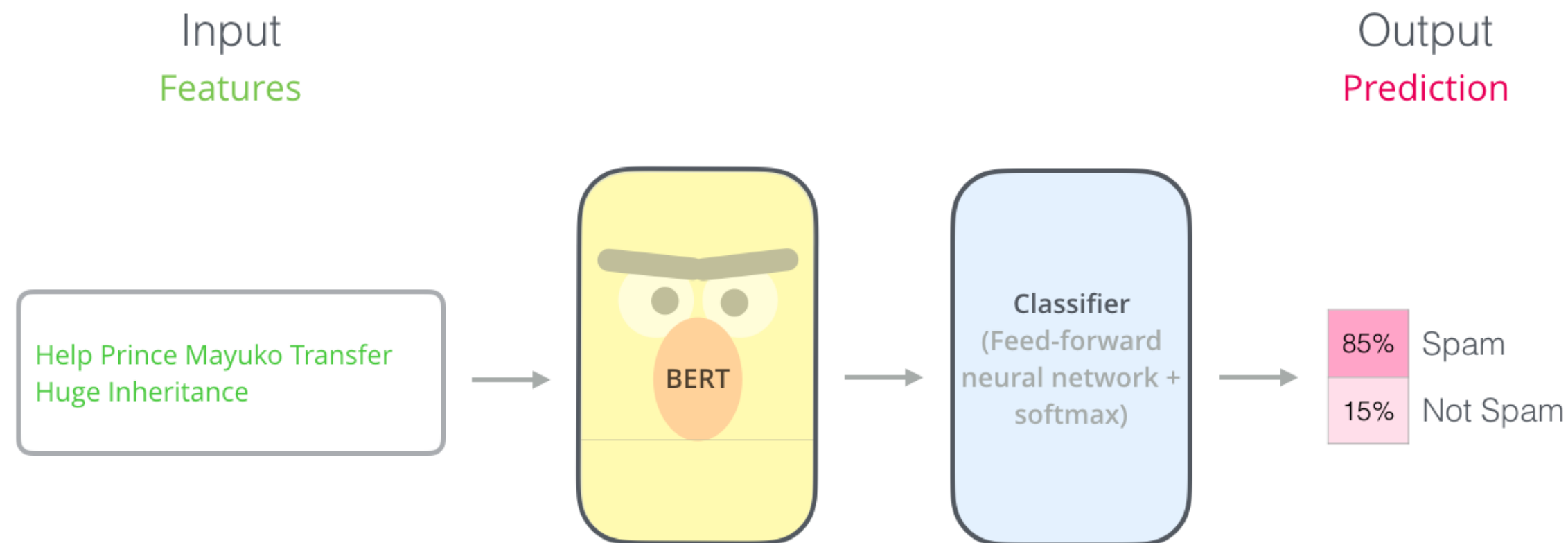
Modelos pré-treinados

BERT

Redes neurais com arquiteturas robustas treinadas com **muchos** dados

Geração de **embeddings**

Fine tuning na saída da rede neural



BERT (pt): Pesos na [huggingface](https://huggingface.co) e implementação por [Pytorch](https://pytorch.org/)

BERT (en): Implementação por [tensorflow](https://www.tensorflow.org/)



PROCESSAMENTO DE LÍNGUA NATURAL

