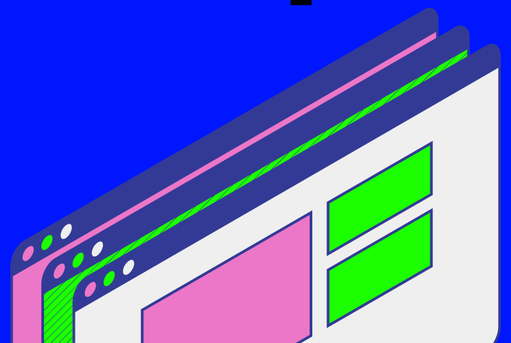
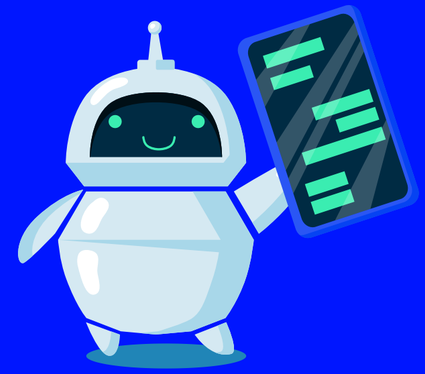
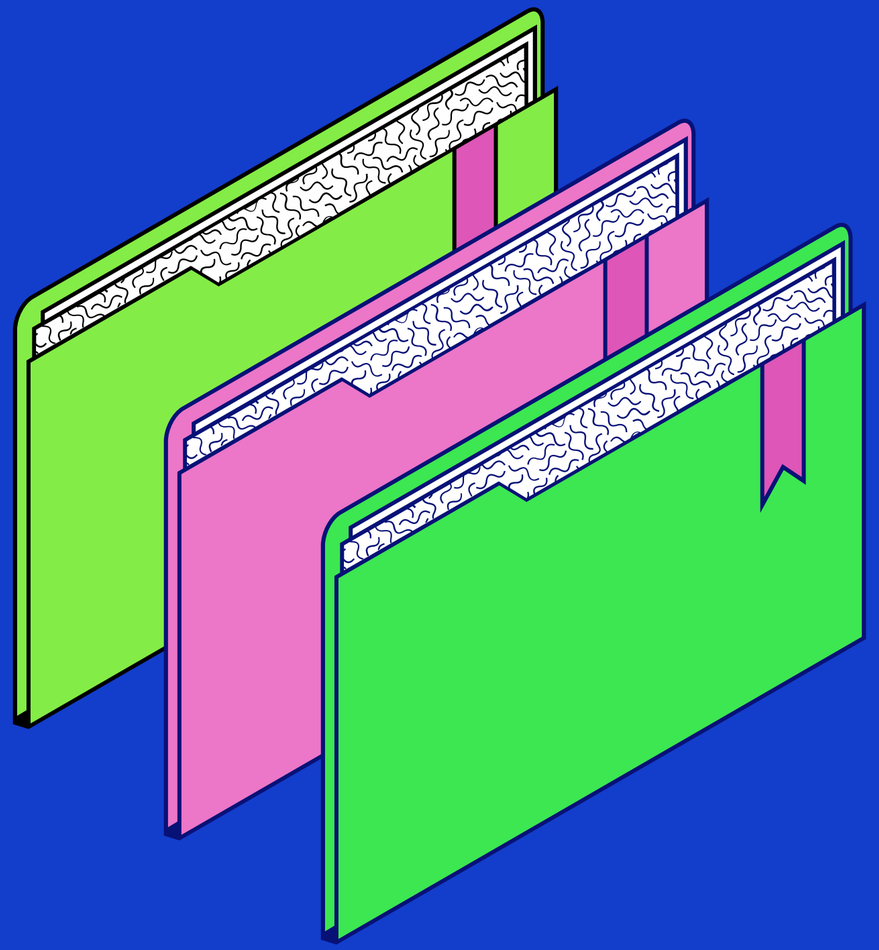
AULA INTRODUTÓRIA

Processamento de Linguagem Natural

Bárbara Dib Oliveira
Igor Kenji Kawai Ueno
Leticia Bossatto Marchezi
Vinícius Gonçalves Perillo







Agenda

PRINCIPAIS TÓPICOS DISCUTIDOS
NESTA APRESENTAÇÃO

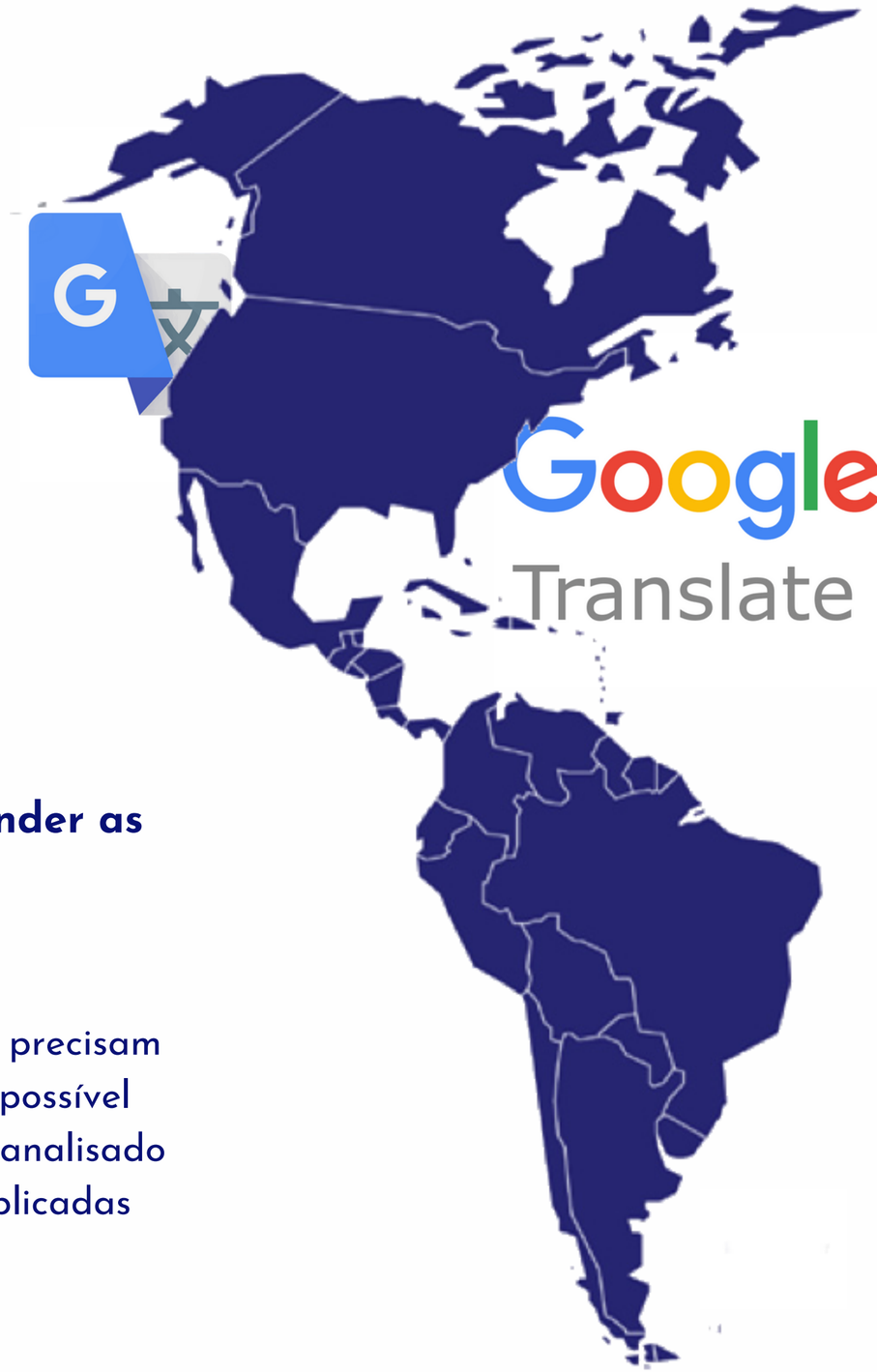
- Roteiro de estudos
- Aplicações práticas
- Bias, ética e imperialismo em modelos linguísticos
- Linguística computacional
- Pré-processamento
- Modelos de AM
- Exemplo prático

Aplicações práticas

1. Filtros de e-mail
2. Assistentes virtuais inteligentes
3. Resultados de pesquisa
4. Texto preditivo
5. Tradução de idiomas
6. Chamadas telefônicas digitais
7. Copilot
8. Análise de textos

Bias, ética e imperialismo em modelos linguísticos

O IMPACTO QUE A TECNOLOGIA CAUSA SOCIEDADE: UMA VIA DE MÃO DUPLA.



Bias: transpondo preconceitos estrututrais

A IAs acabam transpondo preconceitos para os modelos através dos dados, já que o dados é um reflexo da sociedade imperfeita que vivemos

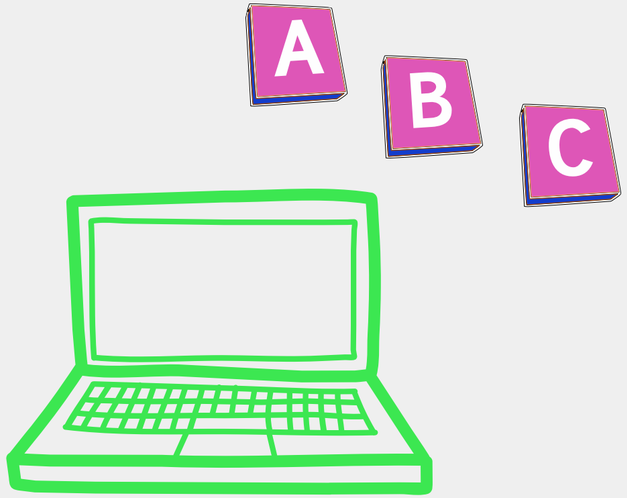
Imperialismo: primeiro e terceiro mundo ainda existem

Atualmente as melhores tecnologias de PLN são produzidas em paises que falam inglês, negligenciando o acesso de linguas com menos falantes

Ética: compreender as limitações da tecnologia

Certas aplicações precisam ser revistas e seu possível impacto deve ser analisado antes de serem aplicadas

Linguística computacional

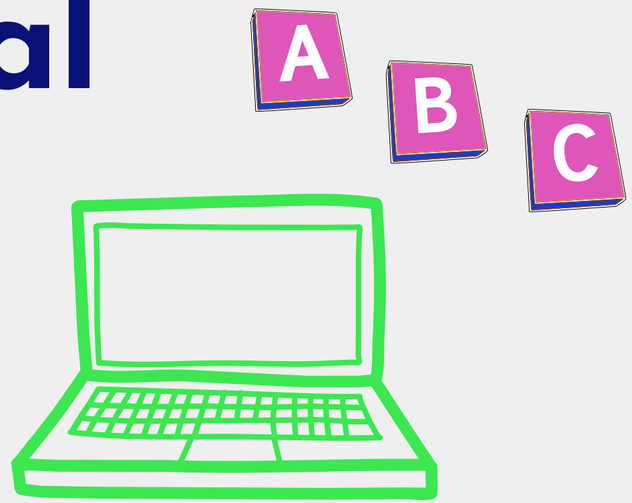


Princípios da Linguística Teórica e da Ciência da Computação

Técnicas computacionais para análise, compreensão e geração da linguagem humana

PLN, análise linguística (sintática, morfológica, semântica), linguística de corpus (construção e análise)

Linguística computacional



Expressões regulares

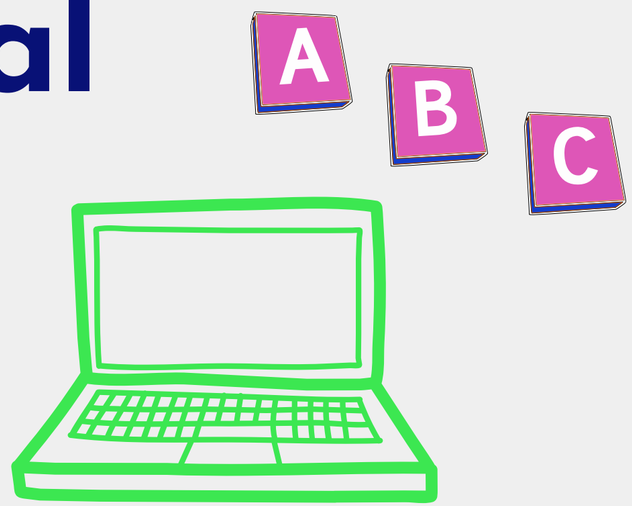
- método formal de se especificar um padrão de texto;
- tarefas de busca, correspondência e manipulação de texto.

```
palavras = ["cachorro", "gato", "carro", "banana", "computador"]

padrao = r'^c\w+'

for palavra in palavras:
    if re.match(padrao, palavra):
        print(palavra)
```

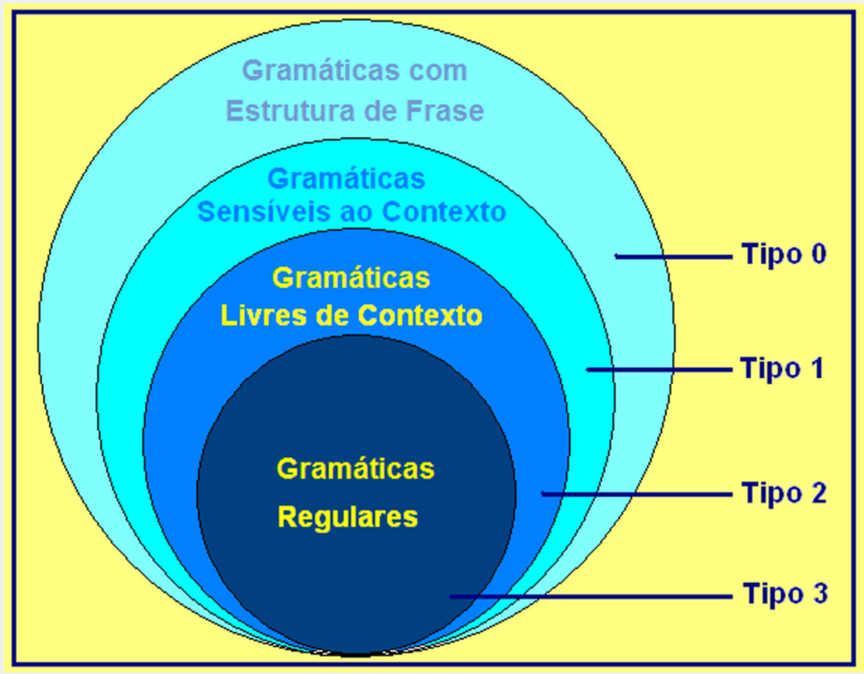
```
cachorro
carro
computador
```

Exemplo de expressão regular

Linguística computacional



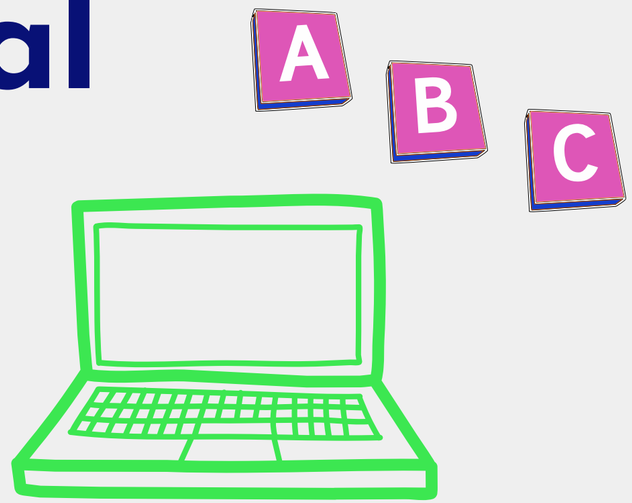
Gramática sensível ao contexto



Hierarquia de Chomsky

- \+ poderosa que a regular;
- expressa regras gramaticais que dependem do **contexto**, considerando **relações**;
- relações sintáticas + complexas (concordância de número e gênero).

Linguística computacional



Chomsky vs Norvig

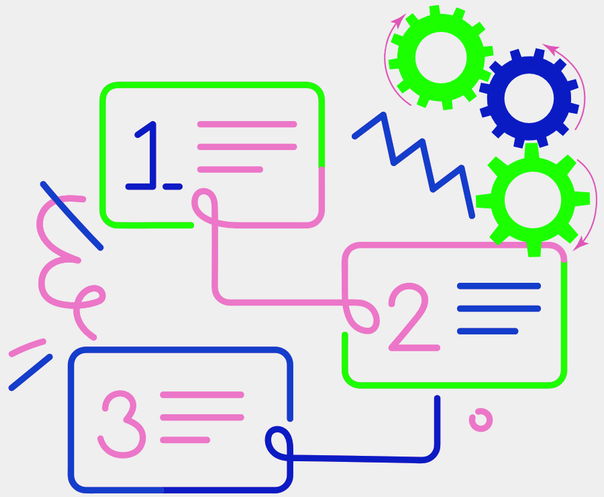
- “features linguísticas ou aprendizado de máquina?”
- **Chomsky**: estruturas **linguísticas** teóricas e regras para entender a linguagem;
- **Norvig**: técnicas de **AM** e grandes quantidades de dados;
- na prática, sistemas modernos de PLN usam uma **combinação**.

Pré-processamento

Normalização

Tokenização

- tokens são unidades atômicas;
- tokenização lexical e sentencial;
- processos seguintes atuam sobre tokens.



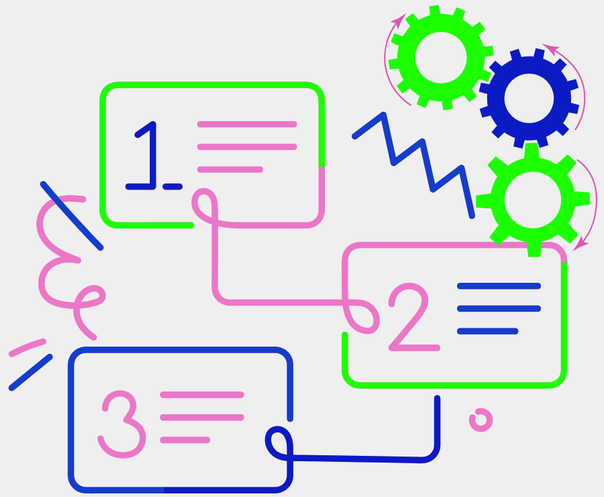
Esta é uma sentença.

['esta', 'é', 'uma', sentença', '.']

Exemplo de Tokenização

Pré-processamento

Normalização



Remoção de tags para web scraping

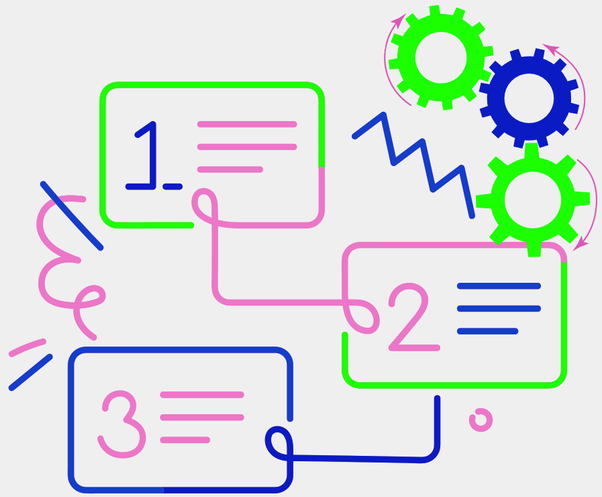
- tags HTML/JavaScript/CSS trazem excesso de formatação e marcação;
- geralmente o foco é o conteúdo textual dos sites.


```
'<!DOCTYPE html>\n<html class="client-nojs" lang="en" dir="ltr">\n<head>\n<meta charset="UTF-8"/>\n<title>Forbes list of Indi
an billionaires - Wikipedia</title>\n<script>document.documentElement.className = document.documentElement.className.replace(
/(^|\\s)client-nojs(\\s|$)/, "$1client-js$2" );</script>\n<script>(window.RLQ=window.RLQ||[]).push(function(){mw.config.set
({"wgCanonicalNamespace":"","wgCanonicalSpecialPageName":false,"wgNamespaceNumber":0,"wgPageName":"Forbes_list_of_Indian_bill
ionaires","wgTitle":"Forbes list of Indian billionaires","wgCurRevisionId":873046063,"wgRevisionId":873046063,"wgArticleId":3
693912,"wgIsArticle":true,"wgIsRedirect":false,"wgAction":"view","wgUserName":null,"wgUserGroups":["*"],"wgCategories":["Wiki
pedia articles in need of updating from August 2018","All Wikipedia articles in need of updating","Wikipedia semi-protected p
ages","Indian billionaires","Lists of Indian people","Wealth in India","Forbes lists","Lists of people by wealth","Economy of
India lists"],"wgBreakFrames":false,"wgPageContentLanguage":"en","wgPageContentModel":"wikitext","wgSeparatorTransformTable":
["",""],"wgDigitTransformTable":["",""],"wgDefaultDateFormat":"dmy","wgMonthNames":["","January","February","March","Apri
l","May","June","July","August","September","October","November","December"],"wgMonthNamesShort":["","Jan","Feb","Mar","Ap
r","May","Jun","Jul","Aug","Sep","Oct","Nov","Dec"],"wgRelevantPageName":"Forbes_list_of_Indian_billionaires","wgRelevantArti
cleId":3693912,"wgRequestId":"XHybiApAICsAAGndxsgAAAAF","wgCSPNonce":false,"wgIsProbablyEditable":false,"wgRelevantPageIsProb
ablyEditable":false,"wgRestrictionEdit":["autoconfirmed"],"wgRestrictionMove":[],"wgFlaggedRevsParams":{"tags":{}},"wgStableR
evisionId":null,"wgCategoryTreePageCategoryOptions":"{\"mode\":0,\"hideprefix\":20,\"showcount\":true,\"namespaces\":
false}","wgWikiEditorEnabledModules":[],"wgBetaFeaturesFeatures":[],"wgMediaViewerOnClick":true,"wgMediaViewerEnabledByDefaul
t":true,"wgPopupsReferencePreviews":false,"wgPopupsShouldSendModuleToUser":true,"wgPopupsConflictsWithNavPopupGadget":fals
e,"wgVisualEditor":{"pageLanguageCode":"en","pageLanguageDir":"ltr","pageVariantFallbacks":"en","usePageImages":true,"usePage
Descriptions":true},"wgMFIsPageContentModelEditable":true,"wgMFEnableFontChanger":true,"wgMFDisplayWikibaseDescriptions":{"se
arch":true,"nearby":true,"watchlist":true,"tagline":false},"wgRelatedArticles":null,"wgRelatedArticlesUseCirrusSearch":tru
```


Exemplo de Web Scraping

Pré-processamento

Normalização

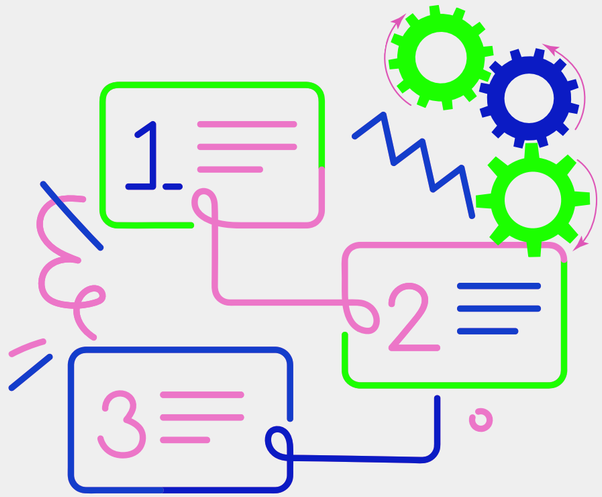


Remoção de stopwords

- palavras muito frequentes sem muita importância;
- “a”, “de”, “o”, “da”, “que”, “e”, “do”;
- depende do contexto e do objetivo.

Pré-processamento

Normalização

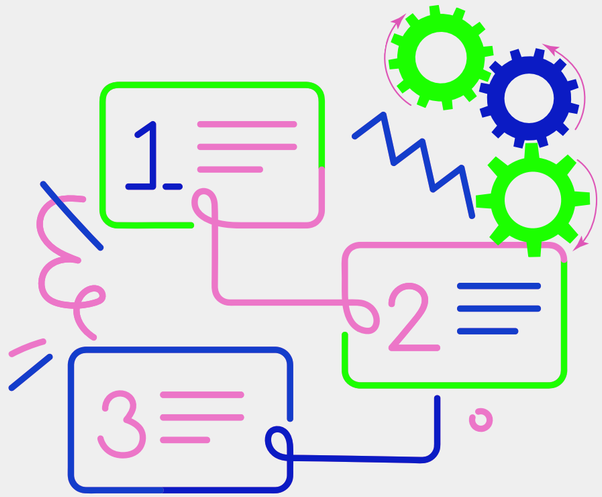


Correção ortográfica e slangs

- tratar erros de digitação, abreviação e vocabulário informal;
- erros prejudiciais por gerarem novos tokens (aumenta a esparsidade dos dados).

Pré-processamento

Análise léxica



Lematização

- reduz uma palavra ao seu lema;
- geração morfológica;
- “gato”, “gata”, “gatos” e “gatas” = "gato".

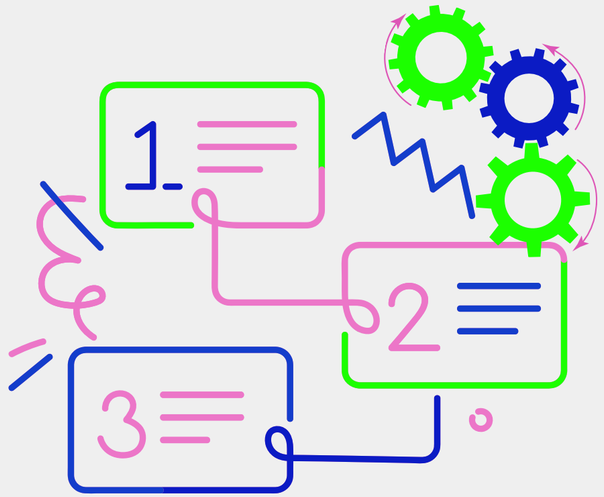
Stemização

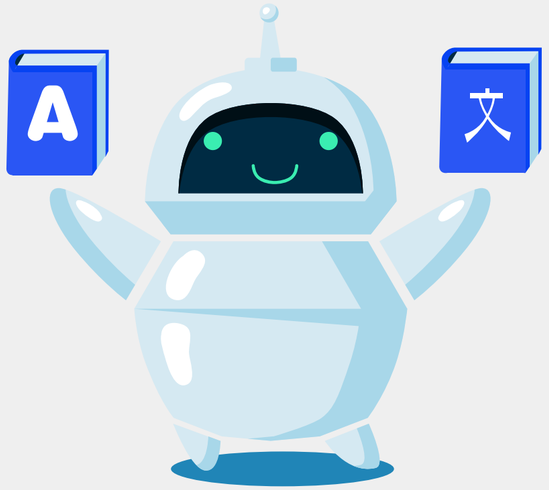
- reduz uma palavra ao seu radical (stem);
- ex.: “gato”, “gata”, “gatos” e “gatas” = “gat”.

Pré-processamento

Análise semântica

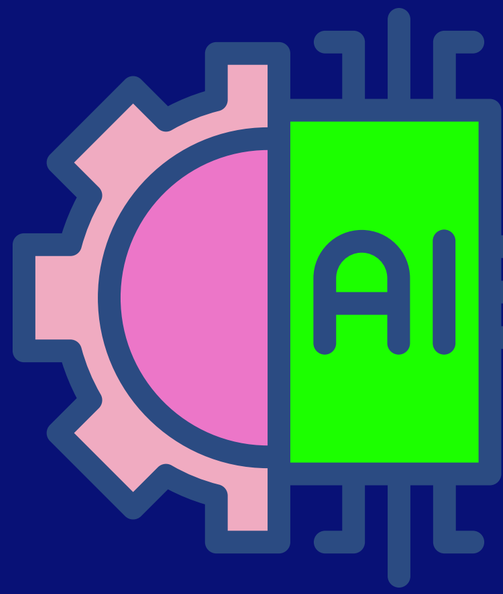
- **significado** das palavras, expressões, sentenças inteiras e enunciados **no contexto**;
- resolução de anáfora e ambiguidades.

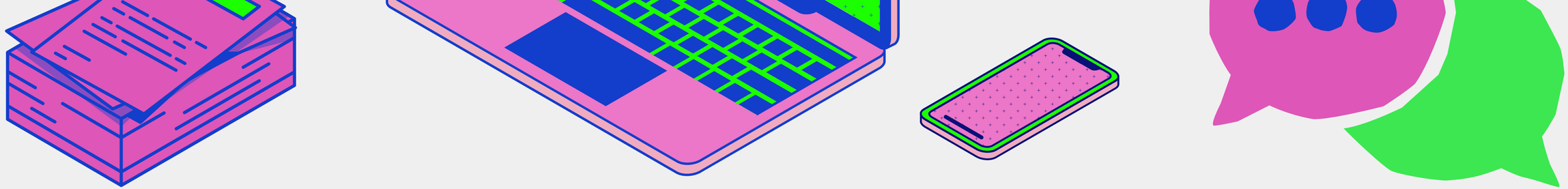




Aprendizado de Máquina

- **MÉTODOS PARA CRIAÇÃO DE FEATURE CONVERTER STRINGS PARA VETORES NUMÉRICOS**
  - BAG OF WORDS
  - TF-IDF
  - EMBEDDINGS
- **MODELOS DE APRENDIZADO DE MÁQUINA**
  - SHALLOW LEARNING
  - DEEP LEARNING





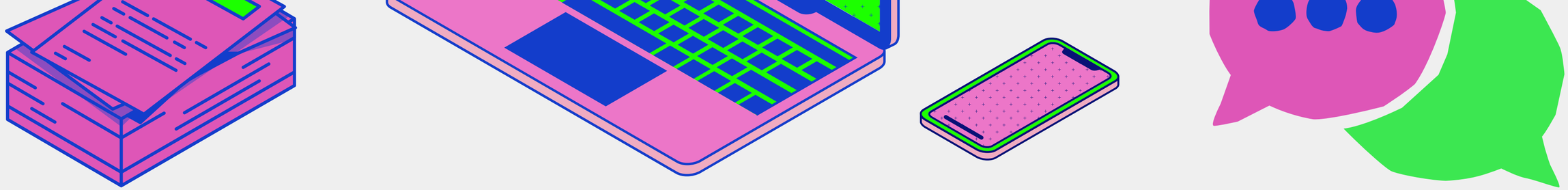
Bag-of-Words(BOW)

word-count

*__documento__ = sequência de strings, arquivos, etc

- Lista de tokens não repetidos em cada documento* com contangem de ocorrência
- **sklearn CountVectorizer()**

	about	bird	heard	is	the	word	you
About the bird, the bird, bird bird bird	1	5	0	0	2	0	0
You heard about the bird	1	1	1	0	1	0	1
The bird is the word	0	1	0	1	2	1	0



TF-IDF

term-frequency times inverse document-frequency

Balanceia o tamanho dos documentos e a frequência
É aplicado em tokens
Retorna um peso para o token (feature)

- **sklearn TfidfVectorizer()**

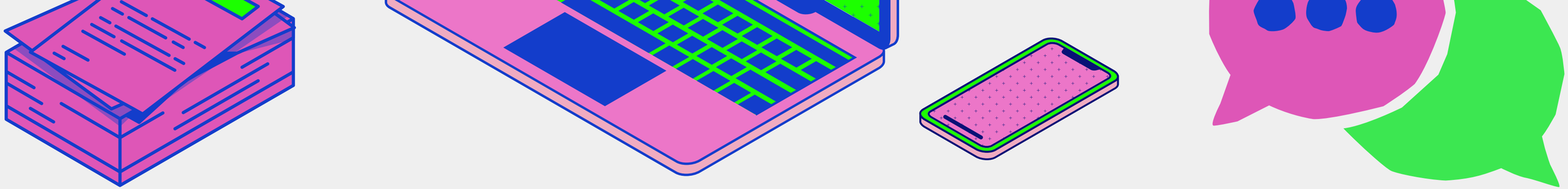
Para um termo x em um documento y:

term-frequency:
TF=(ocorrências de x em y)/(quantidade de termos em y)

inverse document frequency: quão rara a palavra x é nos documentos
IDF=log_10(quantidade de documentos) / (quantidade de documentos que possuem a palavra x)

TF-IDF = TF * IDF

- A presença do log amortece o impacto do IDF!



Word-Embeddings

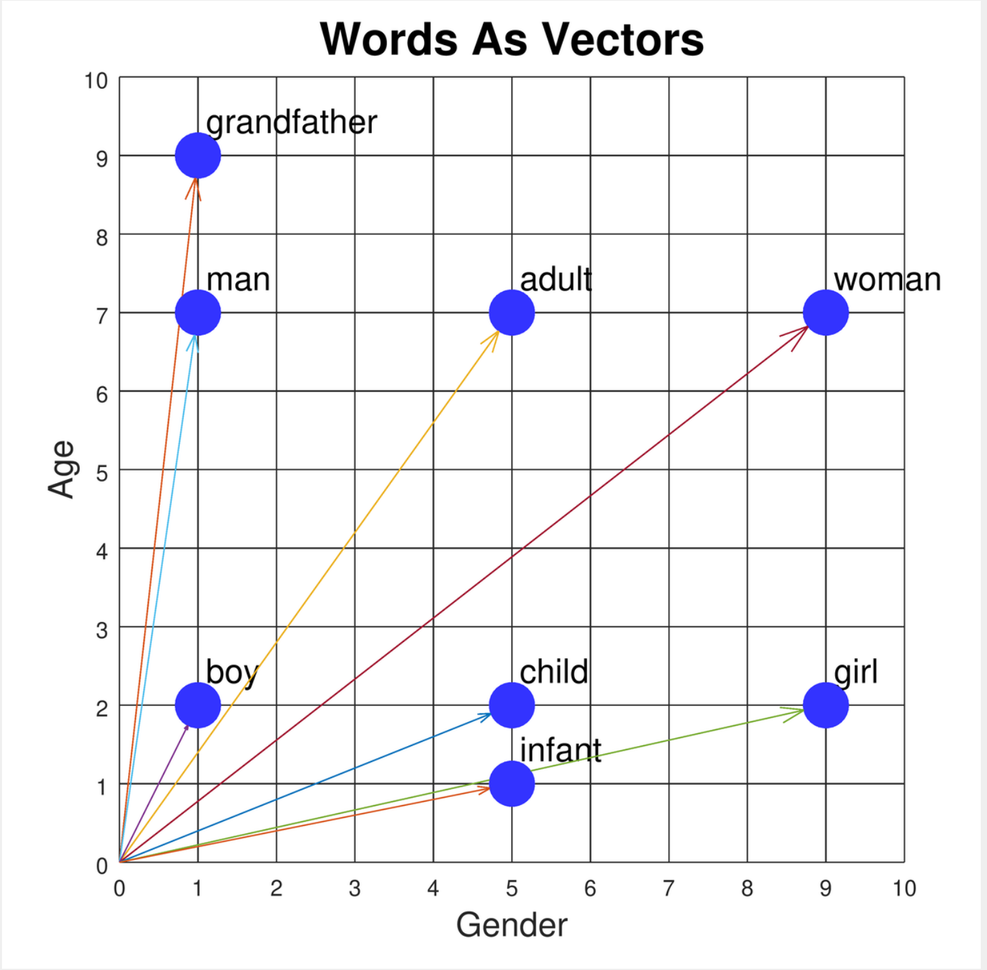
Métodos anteriores não captam semântica!
Este método define palavras por escalas de similaridade de significado
Como calcular similaridade?

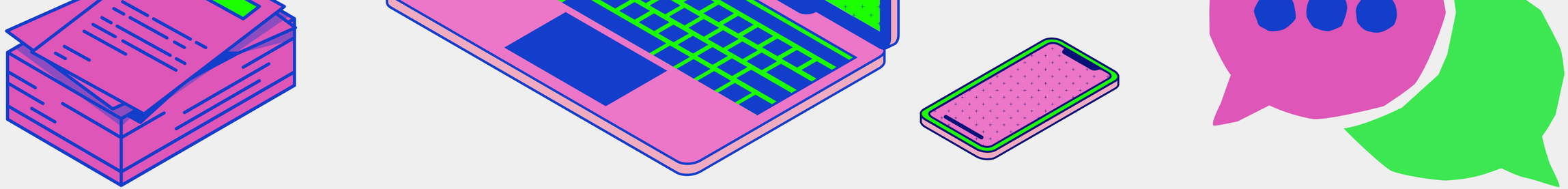
- Distância Euclidiana
- Similaridade de cossenos

Como calcular novos embeddings?

- Aprendizado de máquina!

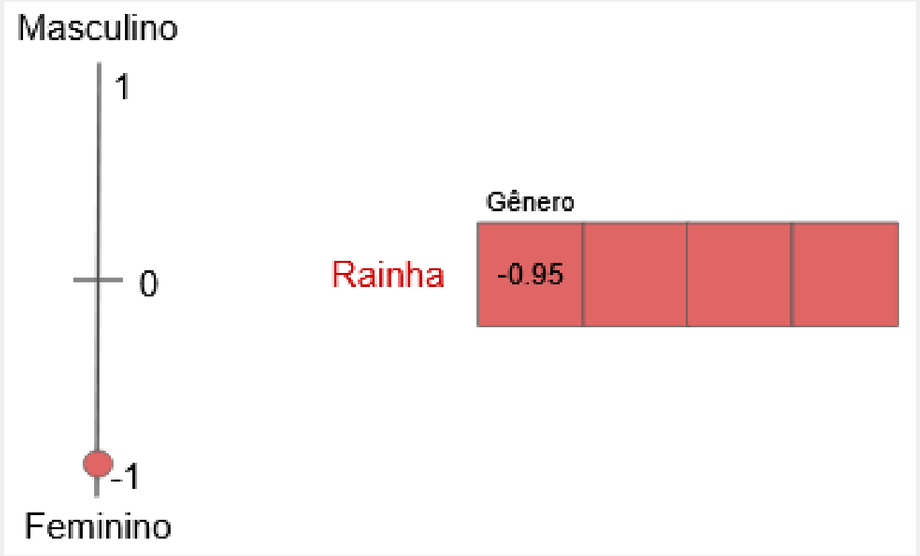
Exemplo de implementação: **word2vec**





Word-Embeddings

Rainha em escala de gênero (1D)

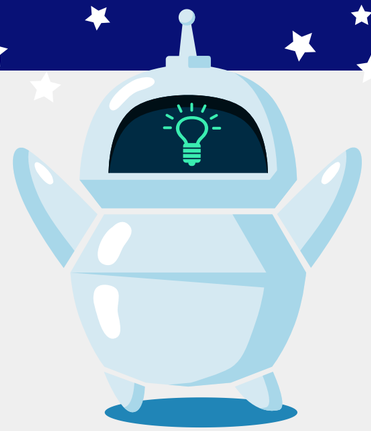


Rainha em escala de gênero e realeza (2D)

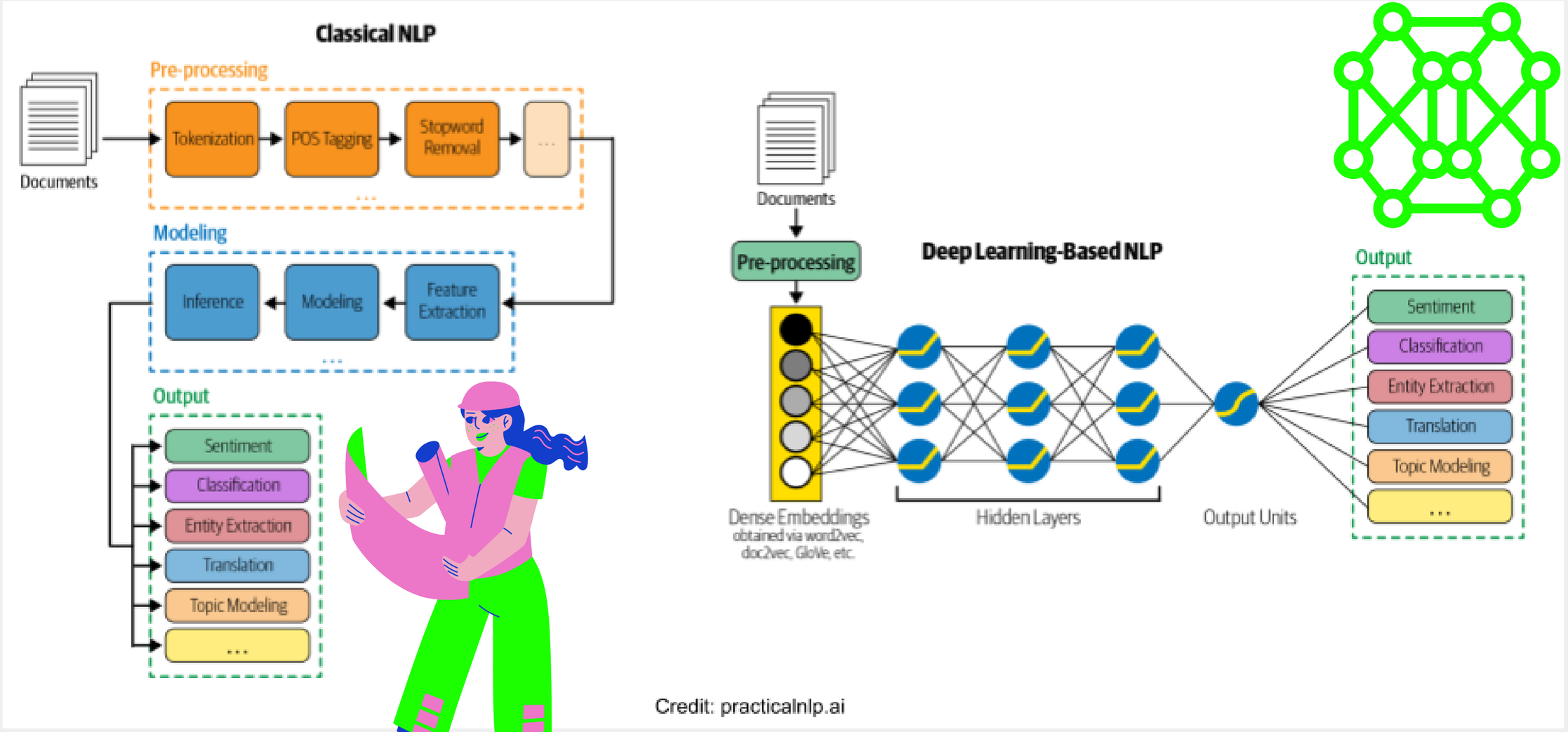
	Rainha	Rei
Gênero	-0.95	0.789
Realeza	0.89	0.96
...	...	...
Fruta	0.015	-0.05
Violência	0.56	0.8

Cada dimensão do vetor representa uma informação do significado!

Mulher - Homem + Rei = Rainha



Shallow learning vs Deep Learning



Outros modelos:

BERT;
GPT;
ROBERTA,
BERTIMBAU...

Obrigado pela atenção!