VALIDAÇÃO DE DOCUMENTOS RG, CPF E DIPLOMAS, COM AUXÍLIO DE OCR-TESSERACT

Gabriel Mancini

Pedro Marcelo Prado

O QUE É OCR ?

- OCR – OPTICAL CHARACTER RECOGNITION (Reconhecimento Óptico de Caracteres)
- É uma tecnologia que permite a conversão de texto apresentado em imagens ou documentos digitalizados em texto editável e pesquisável por máquinas.

COMO ELE FUNCIONA ?

O OCR utiliza algoritmos de processamento de imagem e inteligência artificial para identificar e interpretar caracteres (letras, números, símbolos) contidos em uma imagem.

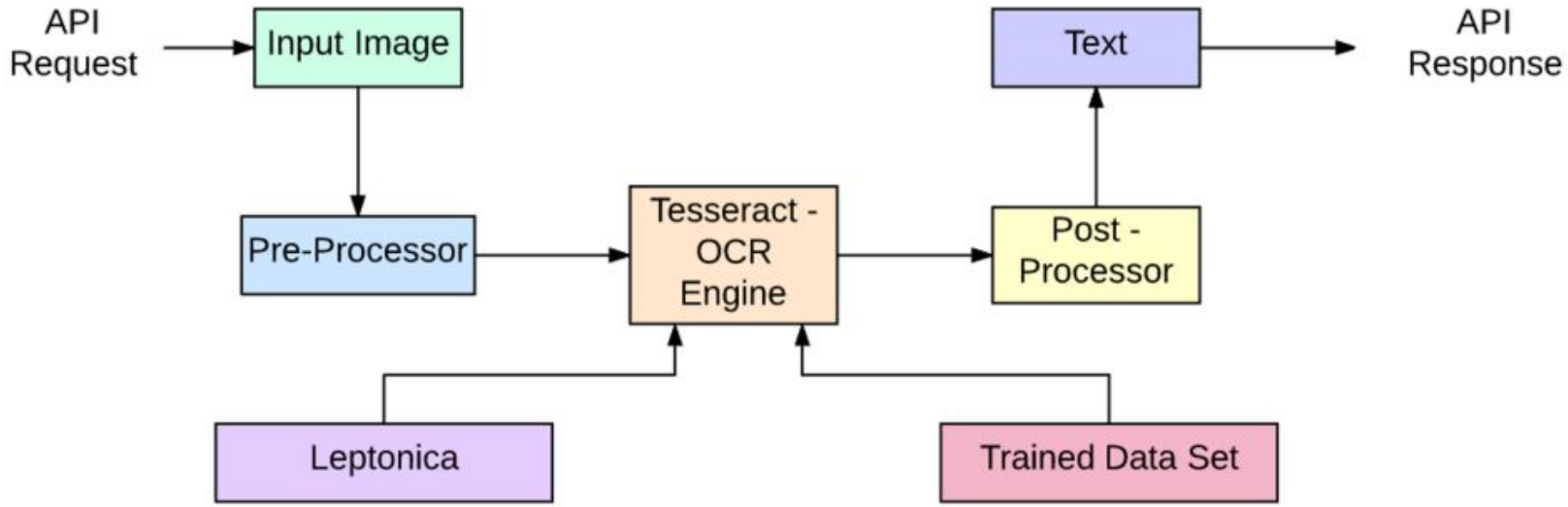
- **Pré-processamento da imagem** :
  Ajuste de brilho e contraste.
  Remoção de ruídos.
  Transformação para escala de cinza ou binarização.

- **Segmentação** :
  Identificação de linhas, palavras e caracteres individuais.

- **Reconhecimento de padrões** :
  Comparação de cada característica com padrões predefinidos ou treinamento em modelos baseados em aprendizado de máquina.

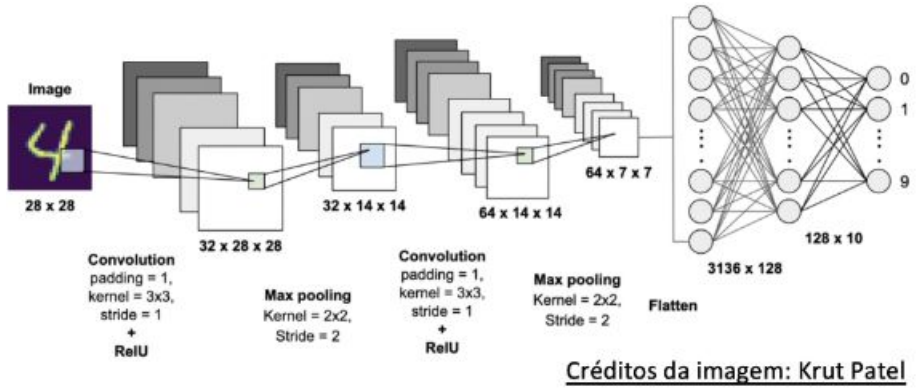
- **Pós-processamento** :
  Correção de erros e ajustes no texto reconhecido, utilizando dicionários ou regras gramaticais.

ESTRUTURA DO TESSERACT



Créditos da imagem: Balaaji Parthasarathy

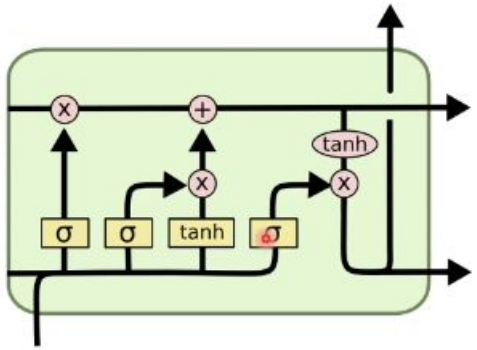
APRENDIZADO PROFUNDO - TESSERACT



Créditos da imagem: Krut Patel

Para reconhecer uma imagem contendo um único caractere, normalmente usamos uma Rede Neural Convolucional (CNN)

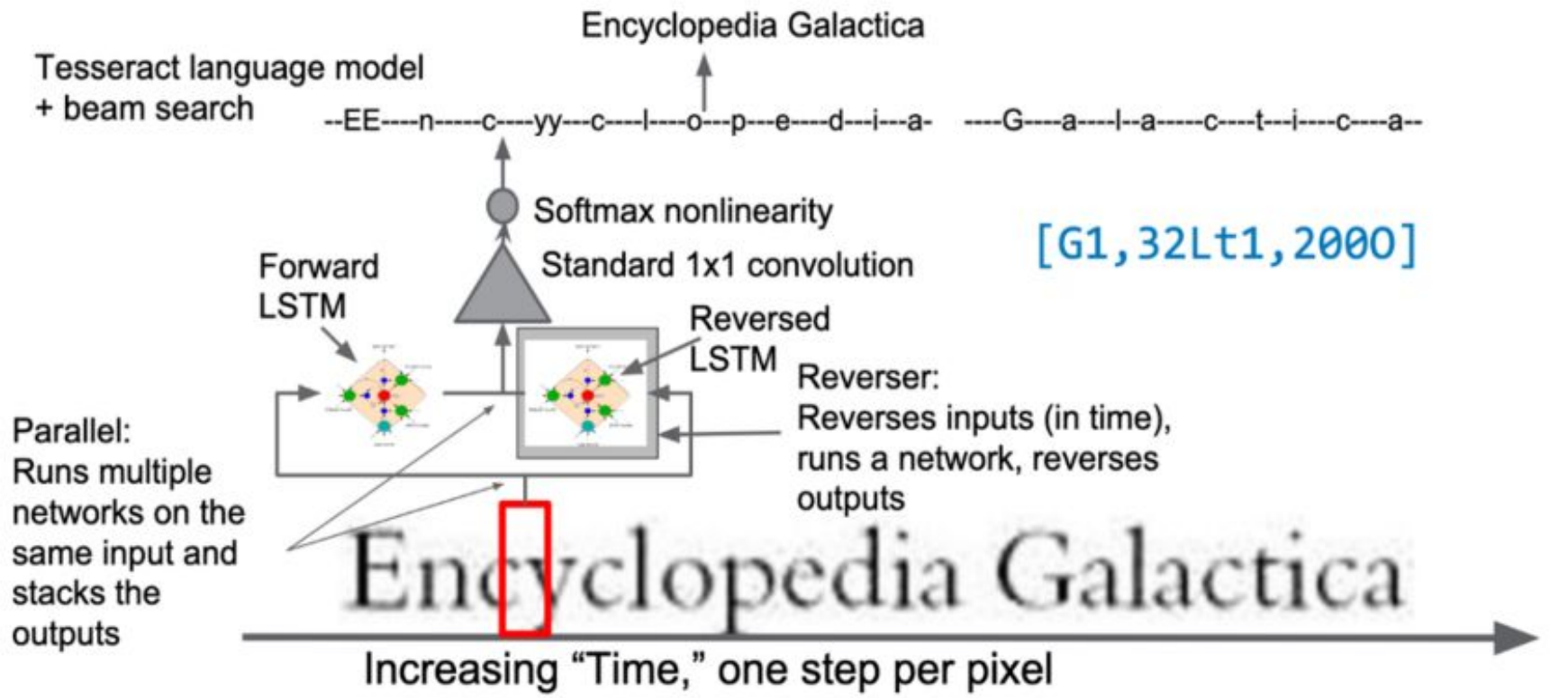
Tesseract - OCR Engine



Estrutura de uma LSTM.
Créditos da imagem: Colah's blog

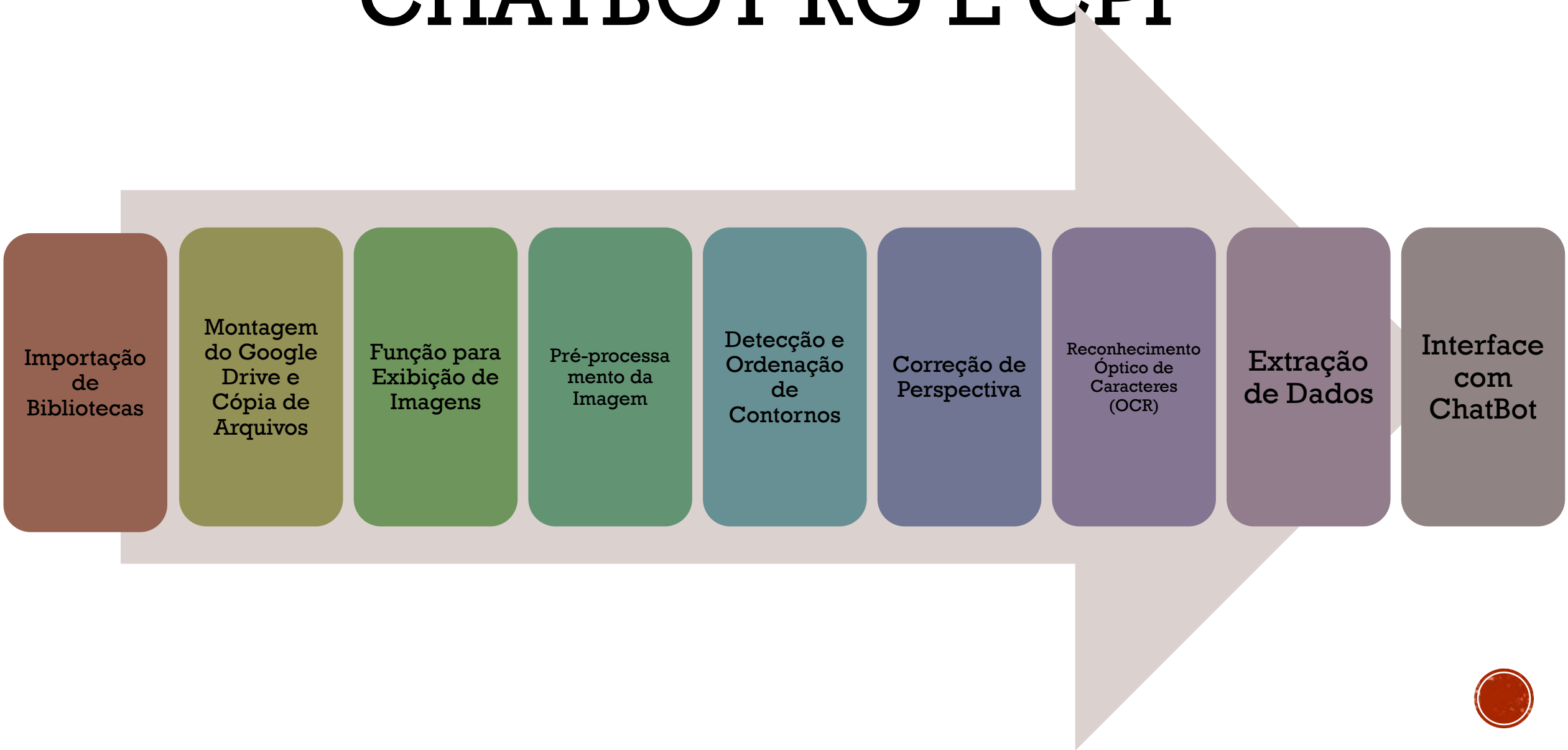
Um texto de comprimento arbitrário é uma sequência de caracteres e é mais interessante utilizar RNNs (Redes Neurais Recorrentes). LSTM (Long short-term memory) é uma forma popular de RNN.

FUNCIONAMENTO

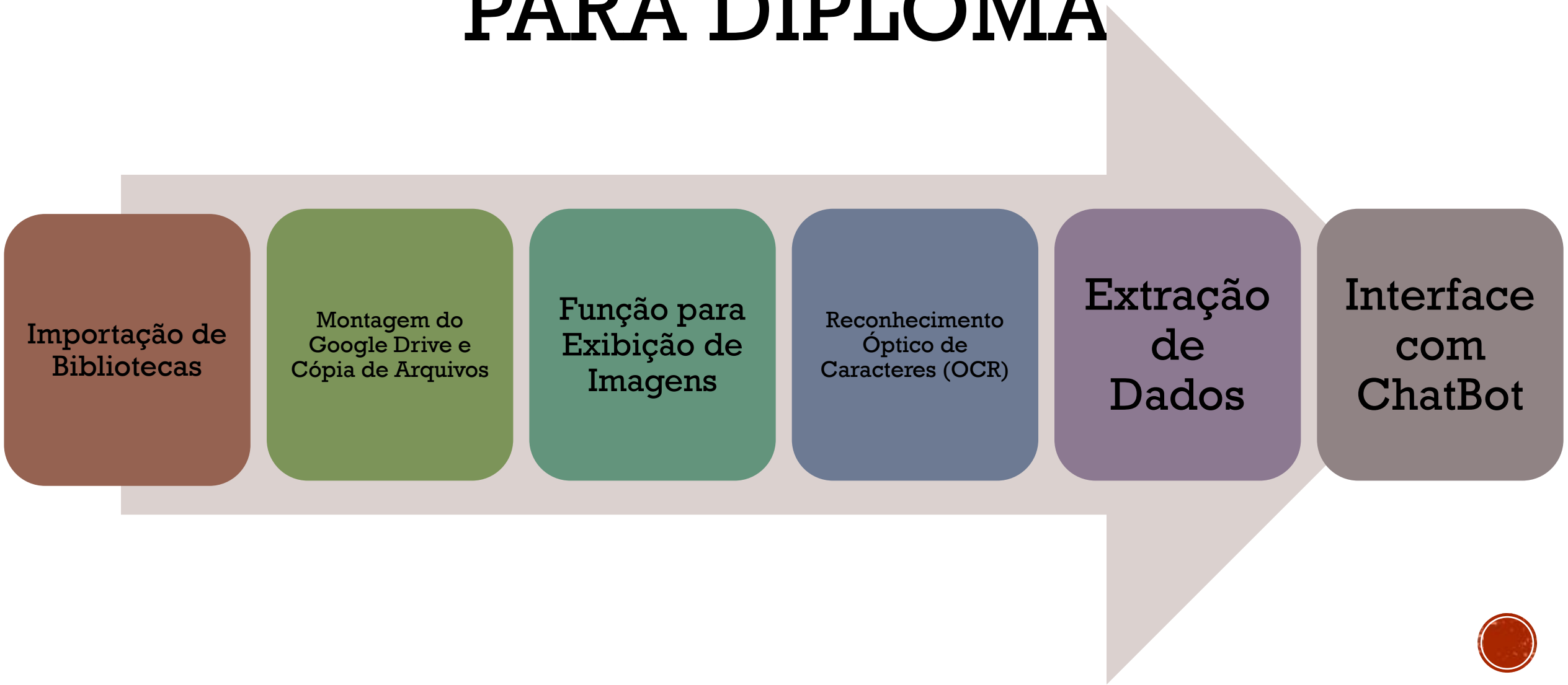


Créditos da imagem: Tesseract 3 OCR process paper

FUNCIONAMENTO DO CHATBOT RG E CPF



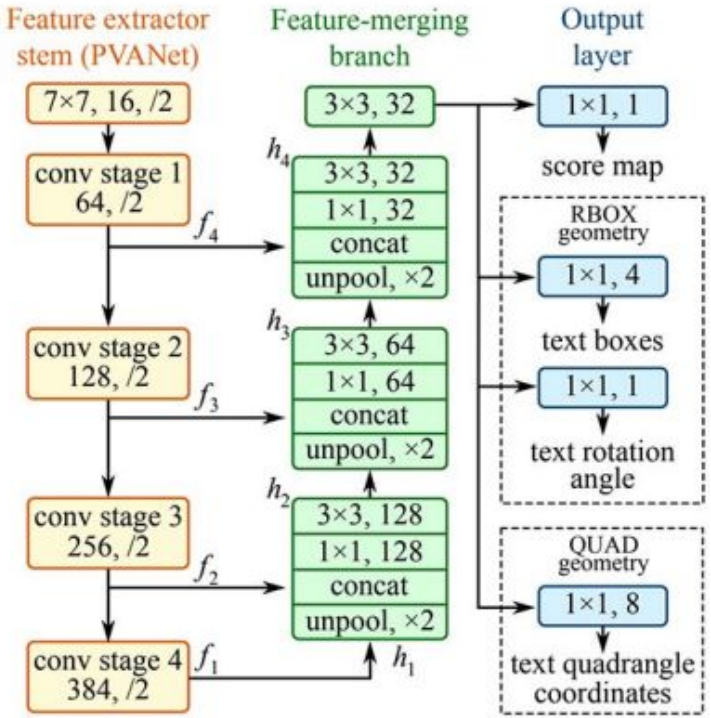
FUNCIONAMENTO DO CHATBOT PARA DIPLOMA



1° TENTATIVA EAST COM TESSERACT

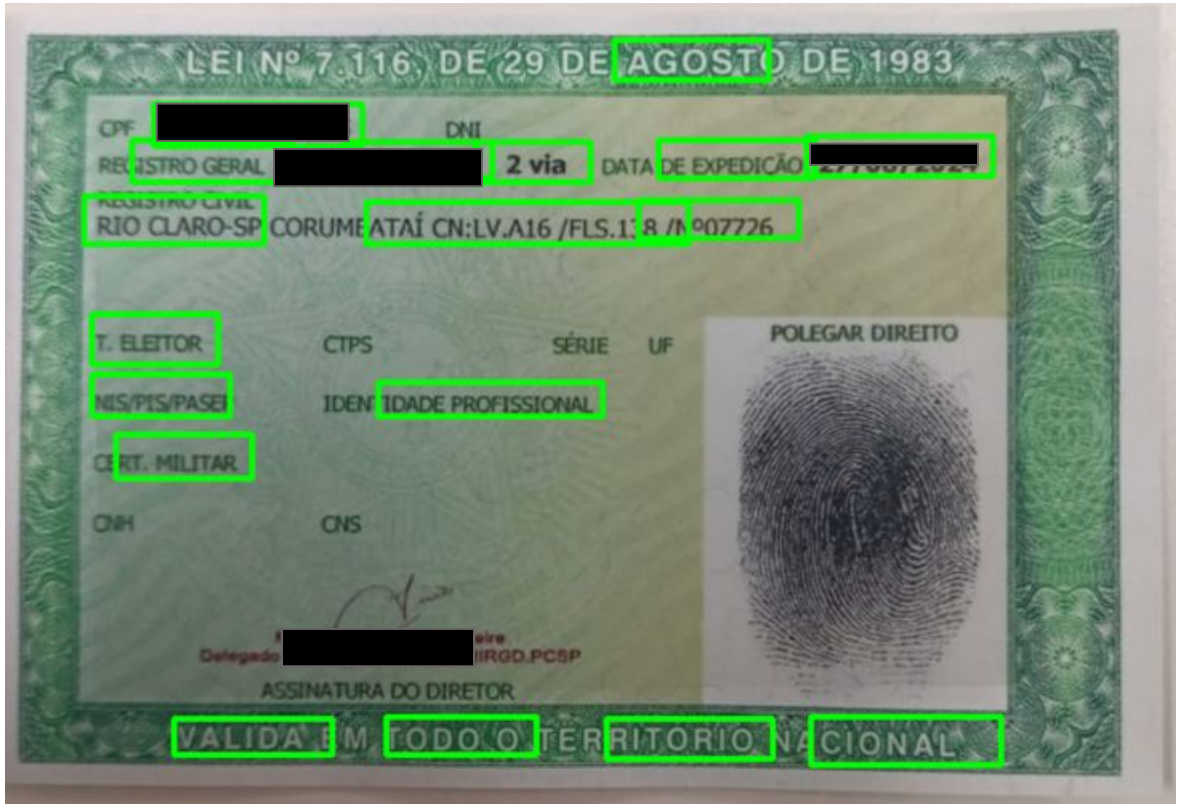
- O EAST (Efficient Accurate Scene Text detector) é um modelo de aprendizagem profunda (deep learning), publicado oficialmente em 2017 por Zhou et al.
  - Usa camadas convolucionais para extrair características das imagens e dessa forma detectar a presença de textos

Estrutura do EAST
(Fully-Convolutional Network)

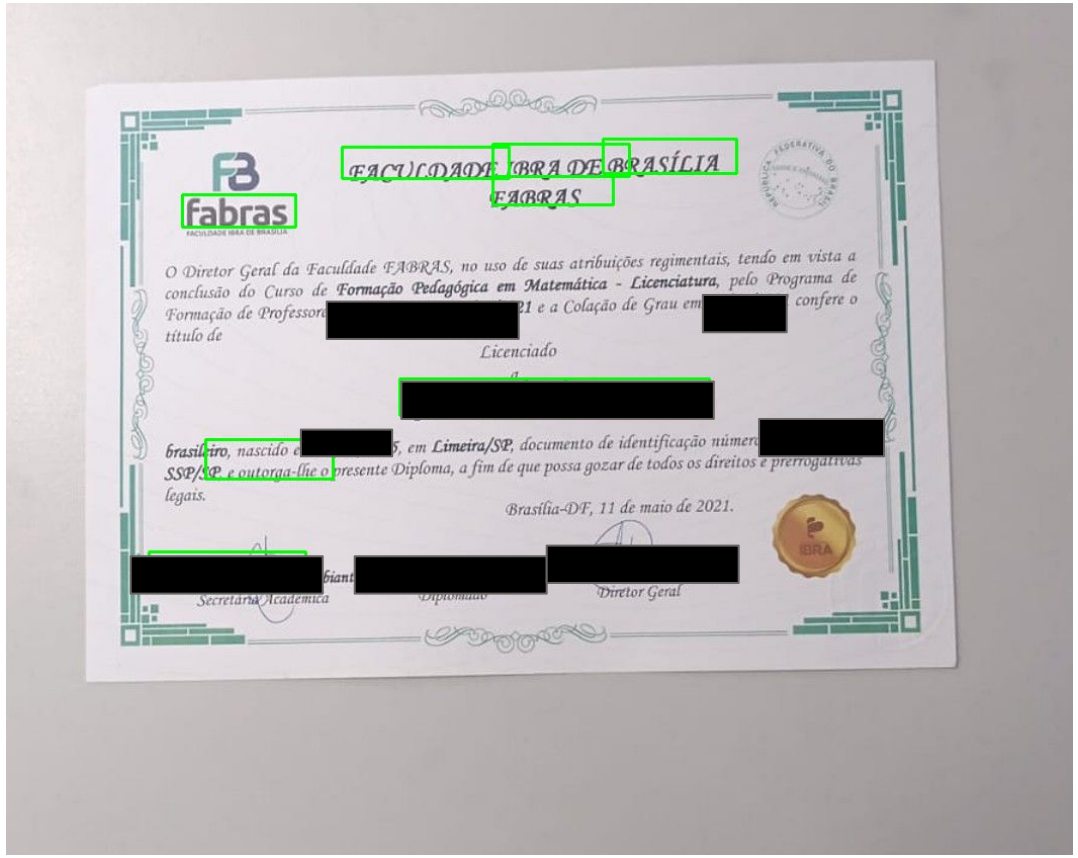


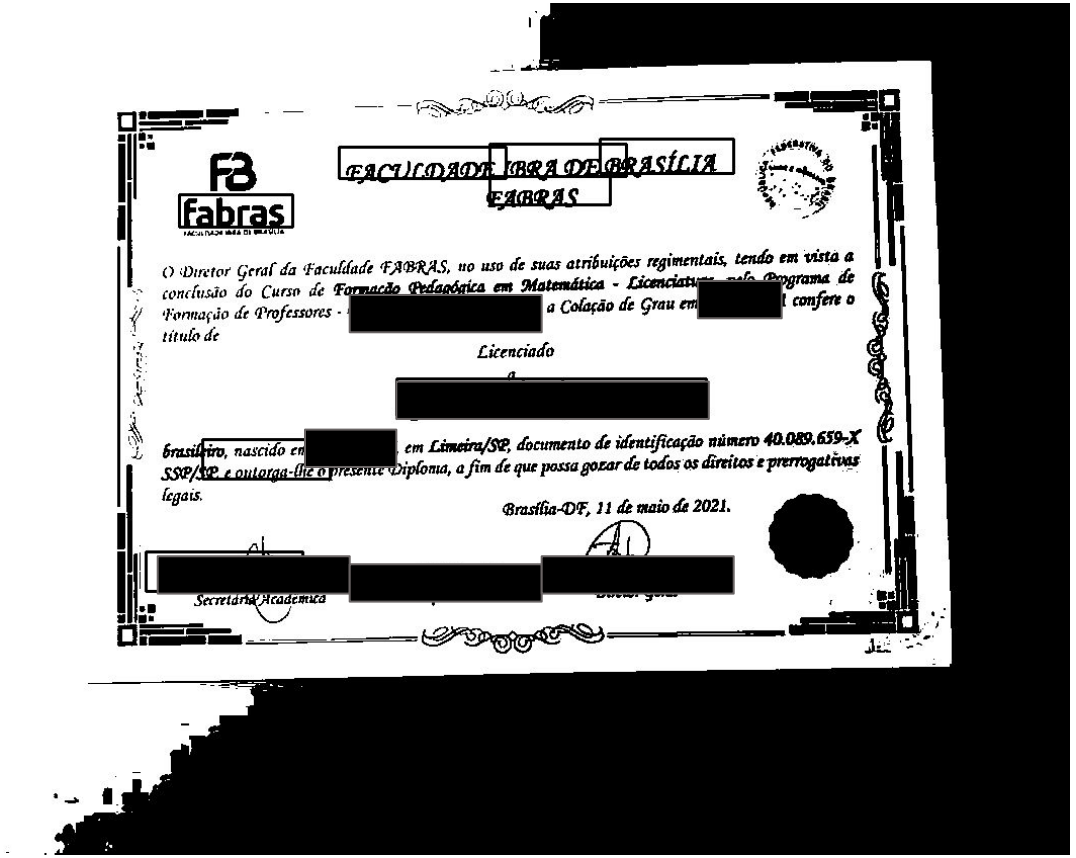
Créditos: Zhou et al.

RESULTADO DO EAST



RESULTADO DO EAST

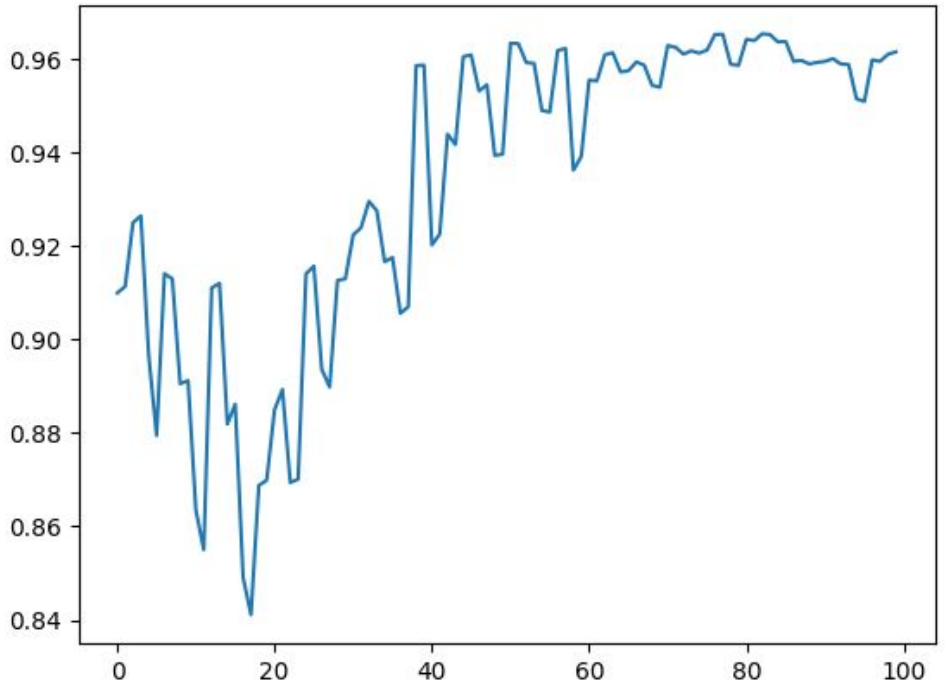




2° TENTATIVA OCR PERSONALIZADO

Layer (type)	Output Shape	Param #
conv2d (Conv2D)	(None, 26, 26, 32)	320
max_pooling2d (MaxPooling2D)	(None, 13, 13, 32)	0
conv2d_1 (Conv2D)	(None, 13, 13, 64)	18,496
max_pooling2d_1 (MaxPooling2D)	(None, 6, 6, 64)	0
conv2d_2 (Conv2D)	(None, 4, 4, 128)	73,856
max_pooling2d_2 (MaxPooling2D)	(None, 2, 2, 128)	0
flatten (Flatten)	(None, 512)	0
dense (Dense)	(None, 64)	32,832
dense_1 (Dense)	(None, 128)	8,320
dense_2 (Dense)	(None, 36)	4,644

Total params: 138,470 (540.90 KB)
Trainable params: 138,468 (540.89 KB)
Non-trainable params: 0 (0.00 B)
Optimizer params: 2 (12.00 B)

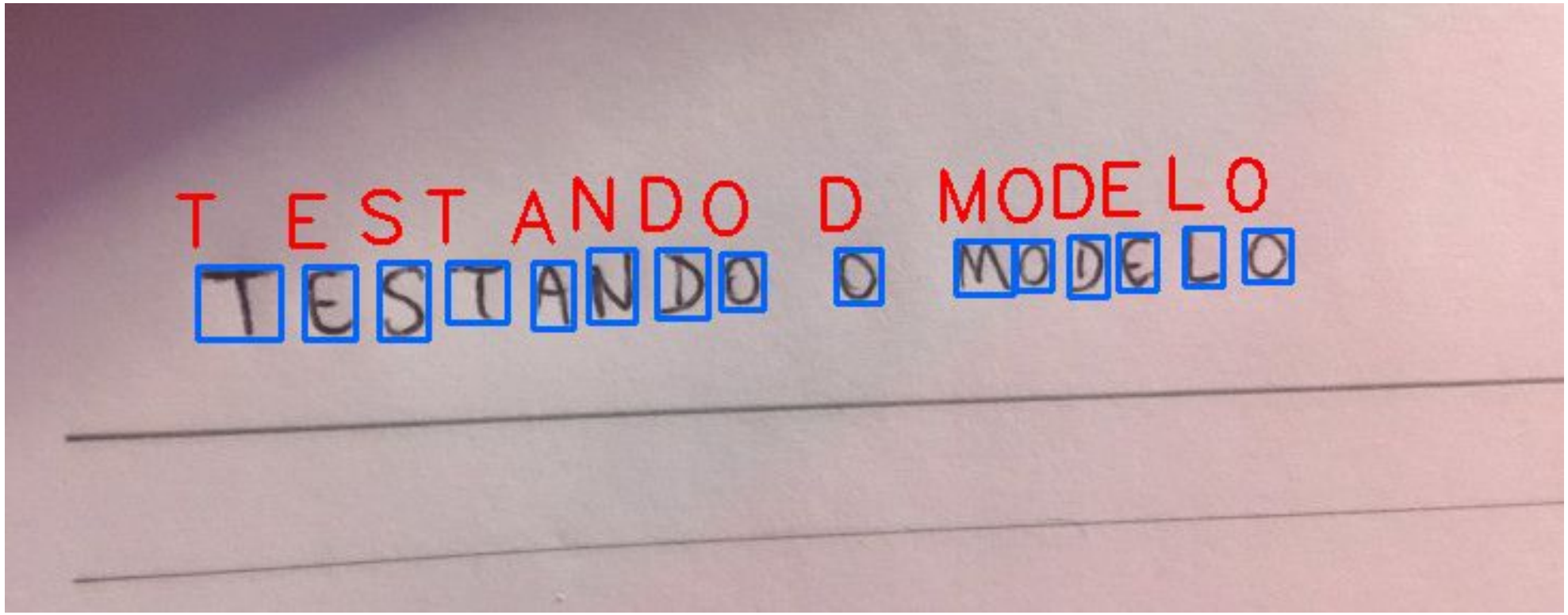


Banco de dados:

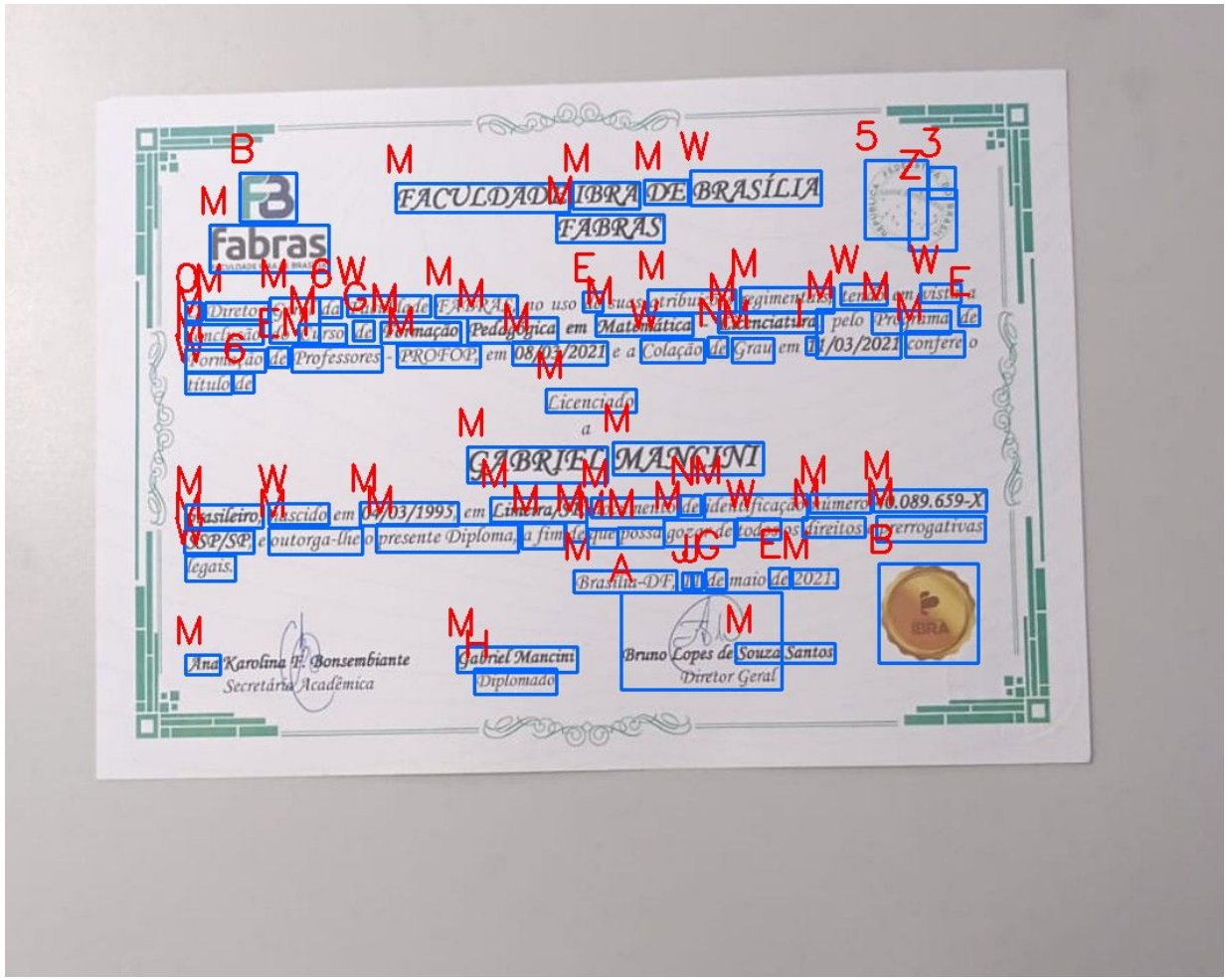
MNIST 0-9
Kaggle A-Z

2° TENTATIVA OCR PERSONALIZADO

```
T -> 99.89%
E -> 100.00%
S -> 98.41%
T -> 99.34%
A -> 100.00%
N -> 99.99%
D -> 100.00%
O -> 54.03%
D -> 93.65%
M -> 99.94%
O -> 44.77%
D -> 99.97%
E -> 99.98%
L -> 100.00%
O -> 77.17%
```



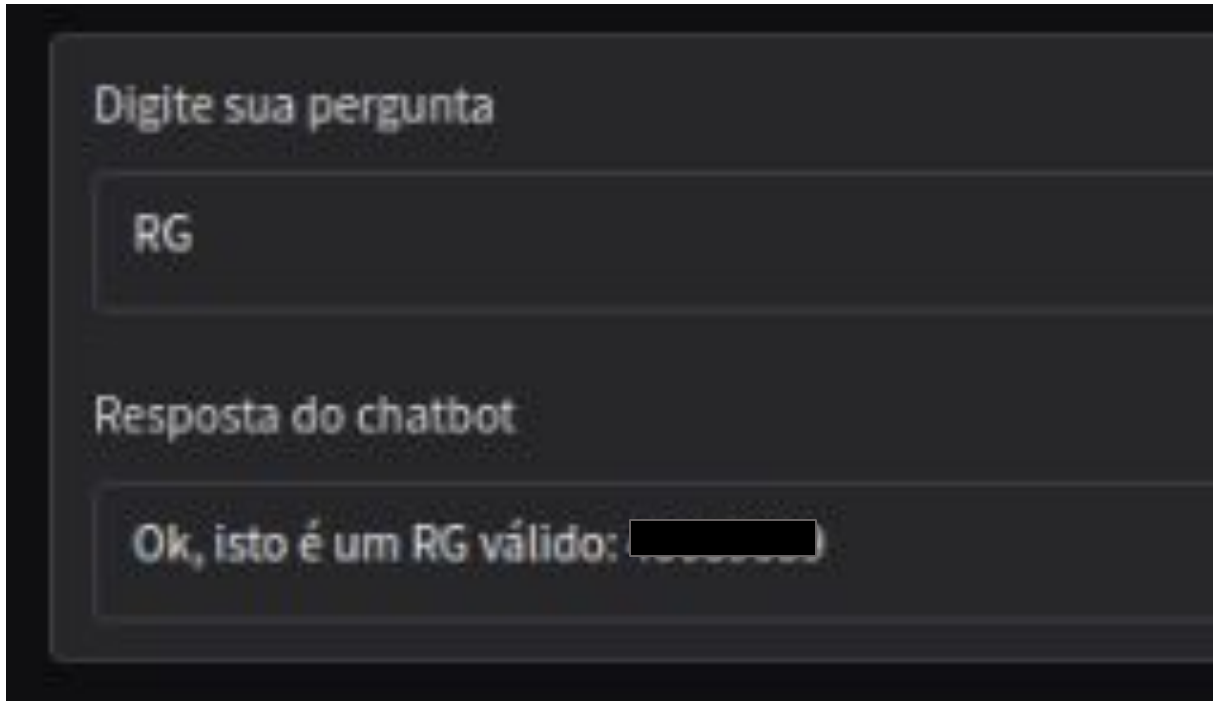
2° TENTATIVA OCR PERSONALIZADO



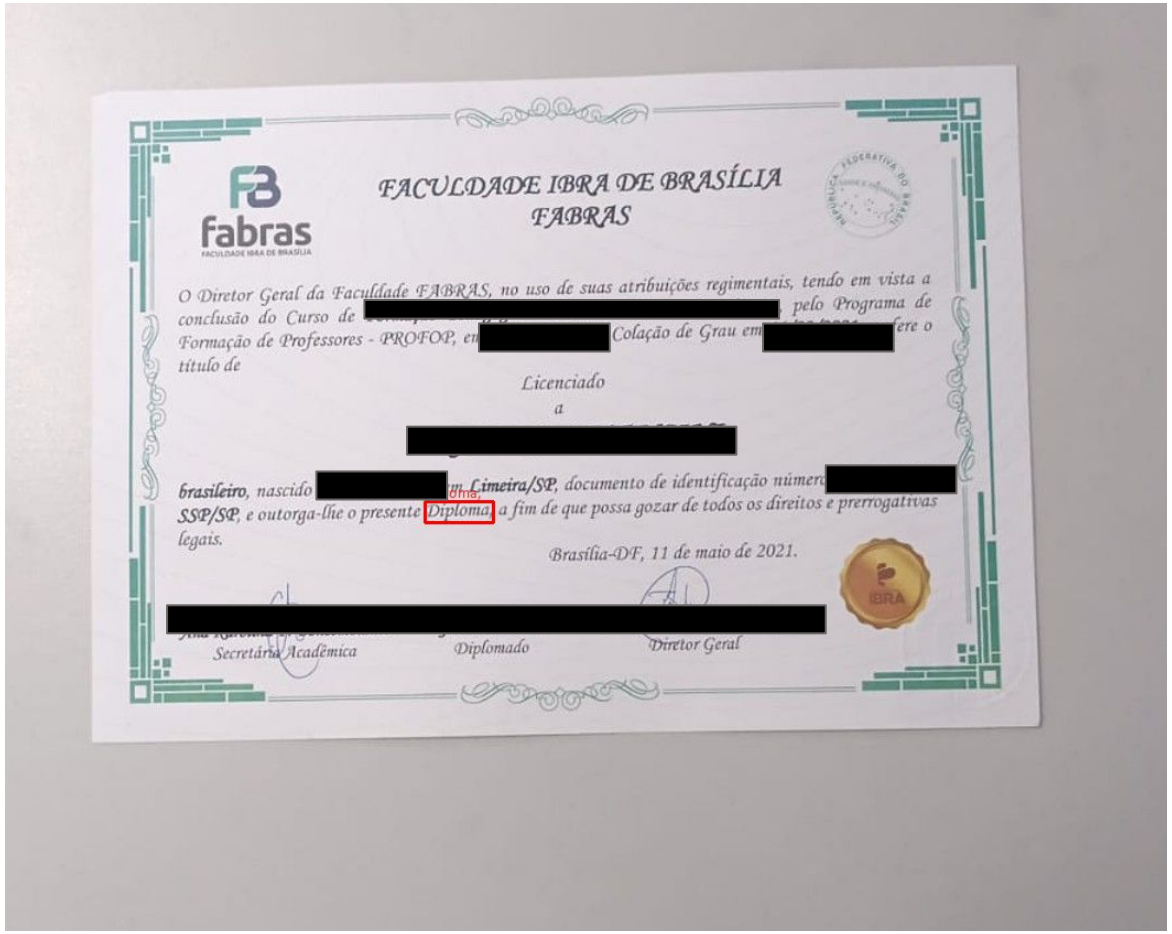
Texto reconhecido:

MMWMWMM0MM6BEMWMMM6WGMMMMMMMMMHMM
MMMMMMEMMMMMAWMMMNJWJMGNMMMWMEMMI
MMW5MMMBMZW3E

O QUE DEU CERTO



O QUE DEU CERTO



FACULDADE IBRA DE BRASÍLIA
FABRAS

O Diretor Geral da Faculdade FABRAS, no uso de suas atribuições regimentais, tendo em vista a conclusão do Curso de [redacted] pelo Programa de Formação de Professores - PROFOP, em [redacted] Colação de Grau em [redacted] confere o título de

Licenciado
a
[redacted]

brasileiro, nascido [redacted] em Limeira/SP, documento de identificação número [redacted] SSP/SP, e outorga-lhe o presente Diploma, a fim de que possa gozar de todos os direitos e prerrogativas legais.

Brasília-DF, 11 de maio de 2021.

[redacted]

Secretária Acadêmica | Diplomado | Diretor Geral

Implementação

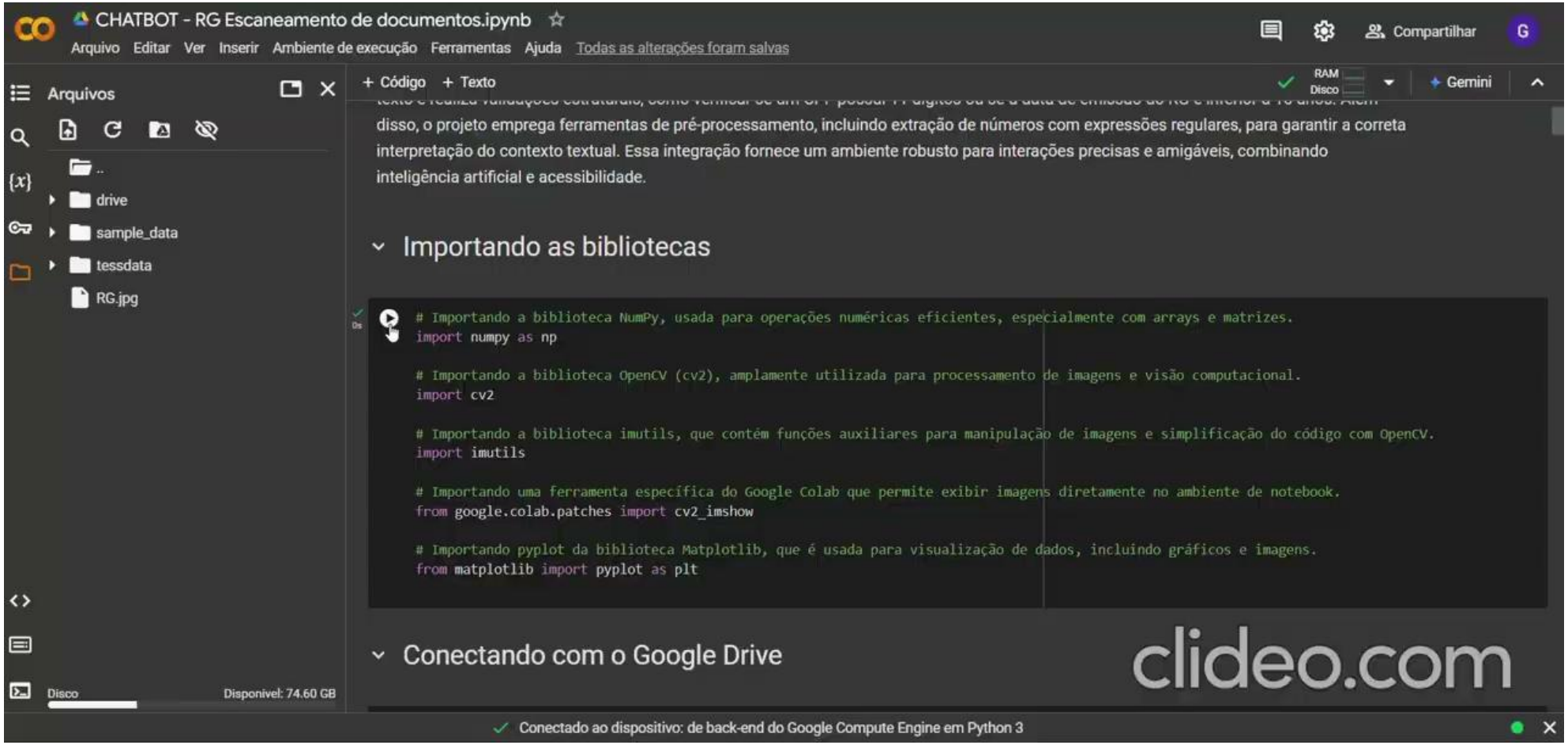
Diploma

O código usa OCR e ferramentas de processamento de imagens para detectar e destacar termos em documentos, validando e extraindo informações automaticamente.

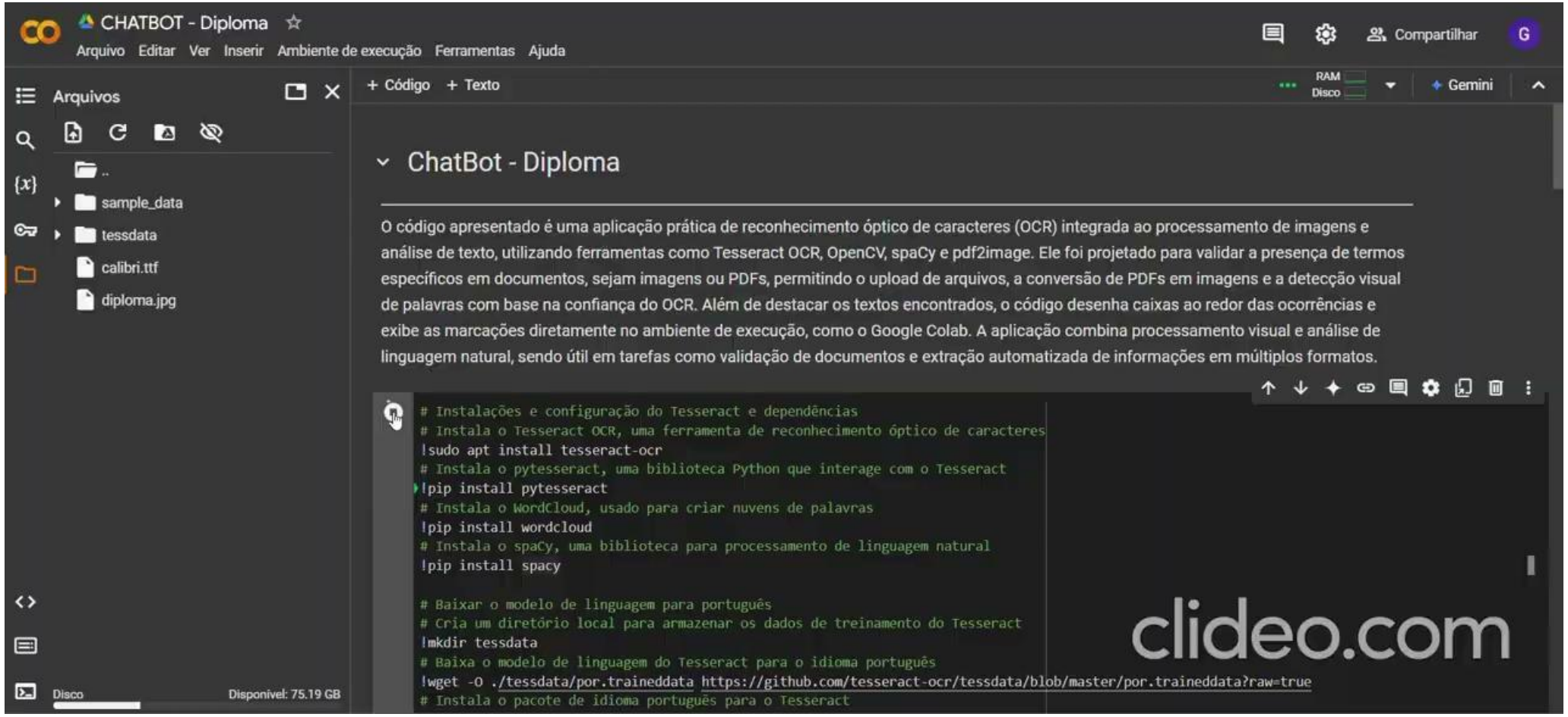
RG

O projeto usa NLP e **DistilBERT** para criar um chatbot que valida informações de documentos, como CPF e RG, com uma interface gráfica feita com **Gradio**.

RG e CPF


CHATBOT - RG Escaneamento de documentos.ipynb
Arquivo Editar Ver Inserir Ambiente de execução Ferramentas Ajuda Todas as alterações foram salvas
Compartilhar
+ Código + Texto
RAM
Disco
Gemini
Arquivos
drive
sample_data
tessdata
RG.jpg
Disco
Disponível: 74.60 GB
disso, o projeto emprega ferramentas de pré-processamento, incluindo extração de números com expressões regulares, para garantir a correta interpretação do contexto textual. Essa integração fornece um ambiente robusto para interações precisas e amigáveis, combinando inteligência artificial e acessibilidade.
Importando as bibliotecas
# Importando a biblioteca NumPy, usada para operações numéricas eficientes, especialmente com arrays e matrizes.
import numpy as np
# Importando a biblioteca OpenCV (cv2), amplamente utilizada para processamento de imagens e visão computacional.
import cv2
# Importando a biblioteca imutils, que contém funções auxiliares para manipulação de imagens e simplificação do código com OpenCV.
import imutils
# Importando uma ferramenta específica do Google Colab que permite exibir imagens diretamente no ambiente de notebook.
from google.colab.patches import cv2_imshow
# Importando pyplot da biblioteca Matplotlib, que é usada para visualização de dados, incluindo gráficos e imagens.
from matplotlib import pyplot as plt
Conectando com o Google Drive
clideo.com
Conectado ao dispositivo: de back-end do Google Compute Engine em Python 3

Diploma


CHATBOT - Diploma
Arquivo Editar Ver Inserir Ambiente de execução Ferramentas Ajuda
Compartilhar
Arquivos
+ Código + Texto
RAM
Disco
Gemini
sample_data
tessdata
calibri.ttf
diploma.jpg
ChatBot - Diploma
O código apresentado é uma aplicação prática de reconhecimento óptico de caracteres (OCR) integrada ao processamento de imagens e análise de texto, utilizando ferramentas como Tesseract OCR, OpenCV, spaCy e pdf2image. Ele foi projetado para validar a presença de termos específicos em documentos, sejam imagens ou PDFs, permitindo o upload de arquivos, a conversão de PDFs em imagens e a detecção visual de palavras com base na confiança do OCR. Além de destacar os textos encontrados, o código desenha caixas ao redor das ocorrências e exibe as marcações diretamente no ambiente de execução, como o Google Colab. A aplicação combina processamento visual e análise de linguagem natural, sendo útil em tarefas como validação de documentos e extração automatizada de informações em múltiplos formatos.
Instalações e configuração do Tesseract e dependências
Instala o Tesseract OCR, uma ferramenta de reconhecimento óptico de caracteres
!sudo apt install tesseract-ocr
Instala o pytesseract, uma biblioteca Python que interage com o Tesseract
!pip install pytesseract
Instala o WordCloud, usado para criar nuvens de palavras
!pip install wordcloud
Instala o spaCy, uma biblioteca para processamento de linguagem natural
!pip install spacy
Baixar o modelo de linguagem para português
Cria um diretório local para armazenar os dados de treinamento do Tesseract
!mkdir tessdata
Baixa o modelo de linguagem do Tesseract para o idioma português
!wget -O ./tessdata/por.traineddata https://github.com/tesseract-ocr/tessdata/blob/master/por.traineddata?raw=true
Instala o pacote de idioma português para o Tesseract
clideo.com
Disco
Disponível: 75.19 GB

Pontos positivos

A implementação do sistema na UNESP proporcionaria:

- Maior eficiência em processos administrativos.
- Redução de erros na validação de documentos.
- Atendimento ágil e acessível para a comunidade acadêmica.
- Redução da carga de trabalho dos funcionários, permitindo maior foco em atividades estratégicas.

PERSPECTIVAS FUTURAS

- Implantar o CHATBOT de RG e CPF para leitura de pdf.
- De posse de um banco de dados estruturado seria possível realizar o treinamento de uma rede yolo para detectar se a imagem inserida se trata de um RG ou não, e em seguida extrair as informações de acordo com o CHATBOT implementado.
- Com o banco de dados seria possível realizar o treinamento personalizado de um extrator de texto, específico para extrair informações de grande relevância nestes documentos.
- Melhor a detecção de contornos para que seja possível usar o CHATBOT de RG e CPF para o DIPLOMA.

Referências

- GOODFELLOW, Ian; BENGIO, Yoshua; COURVILLE, Aaron. *Deep Learning*. MIT Press, 2016.
- SMITH, Ray. An Overview of the Tesseract OCR Engine. *Proceedings of the Ninth International Conference on Document Analysis and Recognition*, IEEE, 2007.
- BROWN, Tom et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 2020.

VALIDAÇÃO DE DOCUMENTOS RG, CPF E DIPLOMAS, COM AUXÍLIO DE OCR-TESSERACT

Gabriel Mancini

Pedro Marcelo Prado