Extração de Texto de Imagens - OCR

https://www.kaggle.com/datasets/neuronlab/texto-ocr/data

O Objetivo da análise desse dataset, é utilizar a tecnologia tesseract que faz a leitura da imagem através do OCR, para extrair da imagem apenas o texto.

Serão realizados os seguintes pontos:

- Upload de 2 imagens;
- Tranformação das imagens para RGB;
- Extração de texto das imagens;
- Gravando os textos em arquivos csv;
- Fazer a contagem da quantidade de palavras existentes nos 2 textos;
- Medir a distância entre um texto e o outro com a medida de "LEVENSHTEIN";
- Ao final será feito o comparativo se um texto é iglau ao outro.

In [8]:
```
pip install pytesseract opencv-python-headless
```

```
Requirement already satisfied: pytesseract in c:\users\gisle\anaconda3\lib\site-pa
ckages (0.3.10)
Requirement already satisfied: opencv-python-headless in c:\users\gisle\anaconda3
\lib\site-packages (4.8.1.78)
Requirement already satisfied: packaging>=21.3 in c:\users\gisle\anaconda3\lib\sit
e-packages (from pytesseract) (23.0)
Requirement already satisfied: Pillow>=8.0.0 in c:\users\gisle\anaconda3\lib\site-
packages (from pytesseract) (9.4.0)
Requirement already satisfied: numpy>=1.21.2 in c:\users\gisle\anaconda3\lib\site-
packages (from opencv-python-headless) (1.24.3)
Note: you may need to restart the kernel to use updated packages.
```

In [54]:
```
pip install tensorflow
```

```
Requirement already satisfied: tensorflow in c:\users\gisle\anaconda3\lib\site-pac
kages (2.14.0)
Requirement already satisfied: tensorflow-intel==2.14.0 in c:\users\gisle\anaconda
3\lib\site-packages (from tensorflow) (2.14.0)
Requirement already satisfied: absl-py>=1.0.0 in c:\users\gisle\anaconda3\lib\site
-packages (from tensorflow-intel==2.14.0->tensorflow) (2.0.0)
Requirement already satisfied: astunparse>=1.6.0 in c:\users\gisle\anaconda3\lib\s
ite-packages (from tensorflow-intel==2.14.0->tensorflow) (1.6.3)
Requirement already satisfied: flatbuffers>=23.5.26 in c:\users\gisle\anaconda3\li
b\site-packages (from tensorflow-intel==2.14.0->tensorflow) (23.5.26)
Requirement already satisfied: gast!=0.5.0,!=0.5.1,!=0.5.2,>=0.2.1 in c:\users\gis
le\anaconda3\lib\site-packages (from tensorflow-intel==2.14.0->tensorflow) (0.5.4)
Requirement already satisfied: google-pasta>=0.1.1 in c:\users\gisle\anaconda3\lib
\site-packages (from tensorflow-intel==2.14.0->tensorflow) (0.2.0)
Requirement already satisfied: h5py>=2.9.0 in c:\users\gisle\anaconda3\lib\site-pa
ckages (from tensorflow-intel==2.14.0->tensorflow) (3.7.0)
Requirement already satisfied: libclang>=13.0.0 in c:\users\gisle\anaconda3\lib\si
te-packages (from tensorflow-intel==2.14.0->tensorflow) (16.0.6)
Requirement already satisfied: ml-dtypes==0.2.0 in c:\users\gisle\anaconda3\lib\si
te-packages (from tensorflow-intel==2.14.0->tensorflow) (0.2.0)
Requirement already satisfied: numpy>=1.23.5 in c:\users\gisle\anaconda3\lib\site-
packages (from tensorflow-intel==2.14.0->tensorflow) (1.24.3)
Requirement already satisfied: opt-einsum>=2.3.2 in c:\users\gisle\anaconda3\lib\s
ite-packages (from tensorflow-intel==2.14.0->tensorflow) (3.3.0)
Requirement already satisfied: packaging in c:\users\gisle\anaconda3\lib\site-pack
ages (from tensorflow-intel==2.14.0->tensorflow) (23.0)
Requirement already satisfied: protobuf!=4.21.0,!=4.21.1,!=4.21.2,!=4.21.3,!=4.21.
4,!=4.21.5,<5.0.0dev,>=3.20.3 in c:\users\gisle\anaconda3\lib\site-packages (from
tensorflow-intel==2.14.0->tensorflow) (4.24.4)
Requirement already satisfied: setuptools in c:\users\gisle\anaconda3\lib\site-pac
kages (from tensorflow-intel==2.14.0->tensorflow) (68.0.0)
Requirement already satisfied: six>=1.12.0 in c:\users\gisle\anaconda3\lib\site-pa
ckages (from tensorflow-intel==2.14.0->tensorflow) (1.16.0)
Requirement already satisfied: termcolor>=1.1.0 in c:\users\gisle\anaconda3\lib\si
te-packages (from tensorflow-intel==2.14.0->tensorflow) (2.3.0)
Requirement already satisfied: typing-extensions>=3.6.6 in c:\users\gisle\anaconda
3\lib\site-packages (from tensorflow-intel==2.14.0->tensorflow) (4.7.1)
Requirement already satisfied: wrapt<1.15,>=1.11.0 in c:\users\gisle\anaconda3\lib
\site-packages (from tensorflow-intel==2.14.0->tensorflow) (1.14.1)
Requirement already satisfied: tensorflow-io-gcs-filesystem>=0.23.1 in c:\users\gi
sle\anaconda3\lib\site-packages (from tensorflow-intel==2.14.0->tensorflow) (0.31.
0)
Requirement already satisfied: grpcio<2.0,>=1.24.3 in c:\users\gisle\anaconda3\lib
\site-packages (from tensorflow-intel==2.14.0->tensorflow) (1.59.0)
Requirement already satisfied: tensorboard<2.15,>=2.14 in c:\users\gisle\anaconda3
\lib\site-packages (from tensorflow-intel==2.14.0->tensorflow) (2.14.1)
Requirement already satisfied: tensorflow-estimator<2.15,>=2.14.0 in c:\users\gisl
e\anaconda3\lib\site-packages (from tensorflow-intel==2.14.0->tensorflow) (2.14.0)
Requirement already satisfied: keras<2.15,>=2.14.0 in c:\users\gisle\anaconda3\lib
\site-packages (from tensorflow-intel==2.14.0->tensorflow) (2.14.0)
Requirement already satisfied: wheel<1.0,>=0.23.0 in c:\users\gisle\anaconda3\lib
\site-packages (from astunparse>=1.6.0->tensorflow-intel==2.14.0->tensorflow) (0.3
8.4)
Requirement already satisfied: google-auth<3,>=1.6.3 in c:\users\gisle\anaconda3\l
ib\site-packages (from tensorboard<2.15,>=2.14->tensorflow-intel==2.14.0->tensorfl
ow) (2.23.2)
Requirement already satisfied: google-auth-oauthlib<1.1,>=0.5 in c:\users\gisle\an
aconda3\lib\site-packages (from tensorboard<2.15,>=2.14->tensorflow-intel==2.14.0-
>tensorflow) (1.0.0)
Requirement already satisfied: markdown>=2.6.8 in c:\users\gisle\anaconda3\lib\sit
e-packages (from tensorboard<2.15,>=2.14->tensorflow-intel==2.14.0->tensorflow)
(3.4.1)
Requirement already satisfied: requests<3,>=2.21.0 in c:\users\gisle\anaconda3\lib
\site-packages (from tensorboard<2.15,>=2.14->tensorflow-intel==2.14.0->tensorflo
```

```
w) (2.31.0)
Requirement already satisfied: tensorboard-data-server<0.8.0,>=0.7.0 in c:\users\g
isle\anaconda3\lib\site-packages (from tensorboard<2.15,>=2.14->tensorflow-intel==
2.14.0->tensorflow) (0.7.1)
Requirement already satisfied: werkzeug>=1.0.1 in c:\users\gisle\anaconda3\lib\sit
e-packages (from tensorboard<2.15,>=2.14->tensorflow-intel==2.14.0->tensorflow)
(2.2.3)
Requirement already satisfied: cachetools<6.0,>=2.0.0 in c:\users\gisle\anaconda3
\lib\site-packages (from google-auth<3,>=1.6.3->tensorboard<2.15,>=2.14->tensorflo
w-intel==2.14.0->tensorflow) (5.3.1)
Requirement already satisfied: pyasn1-modules>=0.2.1 in c:\users\gisle\anaconda3\l
ib\site-packages (from google-auth<3,>=1.6.3->tensorboard<2.15,>=2.14->tensorflow-
intel==2.14.0->tensorflow) (0.2.8)
Requirement already satisfied: rsa<5,>=3.1.4 in c:\users\gisle\anaconda3\lib\site-
packages (from google-auth<3,>=1.6.3->tensorboard<2.15,>=2.14->tensorflow-intel==
2.14.0->tensorflow) (4.9)
Requirement already satisfied: requests-oauthlib>=0.7.0 in c:\users\gisle\anaconda
3\lib\site-packages (from google-auth-oauthlib<1.1,>=0.5->tensorboard<2.15,>=2.14-
>tensorflow-intel==2.14.0->tensorflow) (1.3.1)
Requirement already satisfied: charset-normalizer<4,>=2 in c:\users\gisle\anaconda
3\lib\site-packages (from requests<3,>=2.21.0->tensorboard<2.15,>=2.14->tensorflow
-intel==2.14.0->tensorflow) (2.0.4)
Requirement already satisfied: idna<4,>=2.5 in c:\users\gisle\anaconda3\lib\site-p
ackages (from requests<3,>=2.21.0->tensorboard<2.15,>=2.14->tensorflow-intel==2.1
4.0->tensorflow) (3.4)
Requirement already satisfied: urllib3<3,>=1.21.1 in c:\users\gisle\anaconda3\lib
\site-packages (from requests<3,>=2.21.0->tensorboard<2.15,>=2.14->tensorflow-inte
l==2.14.0->tensorflow) (1.26.16)
Requirement already satisfied: certifi>=2017.4.17 in c:\users\gisle\anaconda3\lib
\site-packages (from requests<3,>=2.21.0->tensorboard<2.15,>=2.14->tensorflow-inte
l==2.14.0->tensorflow) (2023.7.22)
Requirement already satisfied: MarkupSafe>=2.1.1 in c:\users\gisle\anaconda3\lib\s
ite-packages (from werkzeug>=1.0.1->tensorboard<2.15,>=2.14->tensorflow-intel==2.1
4.0->tensorflow) (2.1.1)
Requirement already satisfied: pyasn1<0.5.0,>=0.4.6 in c:\users\gisle\anaconda3\li
b\site-packages (from pyasn1-modules>=0.2.1->google-auth<3,>=1.6.3->tensorboard<2.
15,>=2.14->tensorflow-intel==2.14.0->tensorflow) (0.4.8)
Requirement already satisfied: oauthlib>=3.0.0 in c:\users\gisle\anaconda3\lib\sit
e-packages (from requests-oauthlib>=0.7.0->google-auth-oauthlib<1.1,>=0.5->tensorb
oard<2.15,>=2.14->tensorflow-intel==2.14.0->tensorflow) (3.2.2)
Note: you may need to restart the kernel to use updated packages.
```

Importando as Bibliotecas

In [52]:
```
import pytesseract
import cv2
from PIL import Image
import matplotlib.pyplot as plt
```

Configurando o caminho para o executável do Tesseract

In [10]:
```
pytesseract.pytesseract.tesseract_cmd = r'C:\Program Files\Tesseract-OCR\tesseract.
```

Carregando a imagem usando o Pillow (PIL)

In [11]:
```
imagem = cv2.imread(r'C:\Users\gisle\Documentos\GitHub\OCR\ExtracaoTextoImagensOCR\
```

In [12]:
```
imagem2 = cv2.imread(r'C:\Users\gisle\Documentos\GitHub\OCR\ExtracaoTextoImagensOCR
```

In [13]:
```
print(cv2.__version__)
```

```
4.8.1
```

In [14]:
```
print(imagem)
print(imagem2)
```

```
[[[238 238 238]
  [237 237 237]
  [237 237 237]
  ...
  [247 247 247]
  [244 244 244]
  [239 239 239]]

 [[238 238 238]
  [236 236 236]
  [235 235 235]
  ...
  [246 246 246]
  [246 246 246]
  [243 243 243]]

 [[240 240 240]
  [238 238 238]
  [235 235 235]
  ...
  [242 242 242]
  [245 245 245]
  [245 245 245]]

 ...

 [[241 241 241]
  [241 241 241]
  [241 241 241]
  ...
  [240 240 240]
  [240 240 240]
  [239 239 239]]

 [[238 238 238]
  [239 239 239]
  [240 240 240]
  ...
  [242 242 242]
  [240 240 240]
  [237 237 237]]

 [[234 234 234]
  [235 235 235]
  [236 236 236]
  ...
  [244 244 244]
  [241 241 241]
  [237 237 237]]]
[[[238 238 238]
  [237 237 237]
  [237 237 237]
  ...
  [247 247 247]
  [244 244 244]
  [239 239 239]]

 [[238 238 238]
  [236 236 236]
  [235 235 235]
  ...
  [246 246 246]
  [246 246 246]
  [243 243 243]]
```

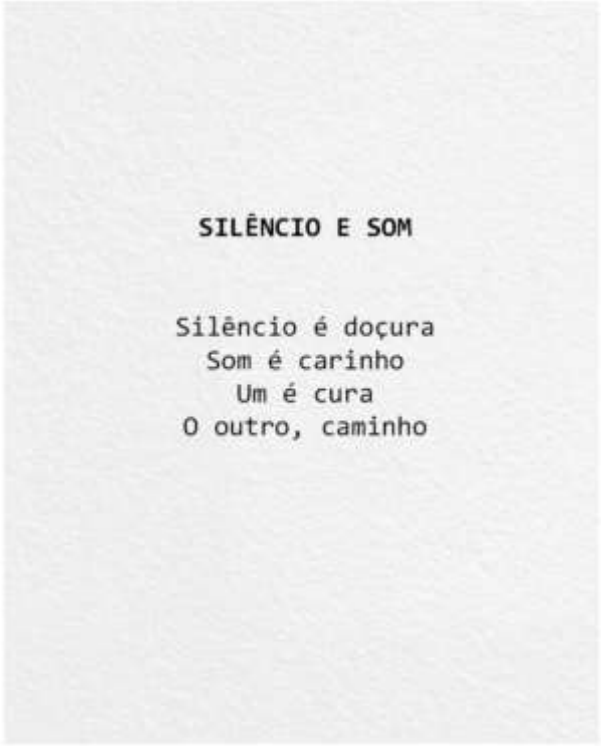
```
[[240 240 240]
 [238 238 238]
 [235 235 235]
 ...
 [242 242 242]
 [245 245 245]
 [245 245 245]]

...

[[241 241 241]
 [241 241 241]
 [241 241 241]
 ...
 [240 240 240]
 [240 240 240]
 [239 239 239]]

[[238 238 238]
 [239 239 239]
 [240 240 240]
 ...
 [242 242 242]
 [240 240 240]
 [237 237 237]]

[[234 234 234]
 [235 235 235]
 [236 236 236]
 ...
 [244 244 244]
 [241 241 241]
 [237 237 237]]]
```

Convertendo a imagem de BGR para RGB (matplotlib usa RGB)

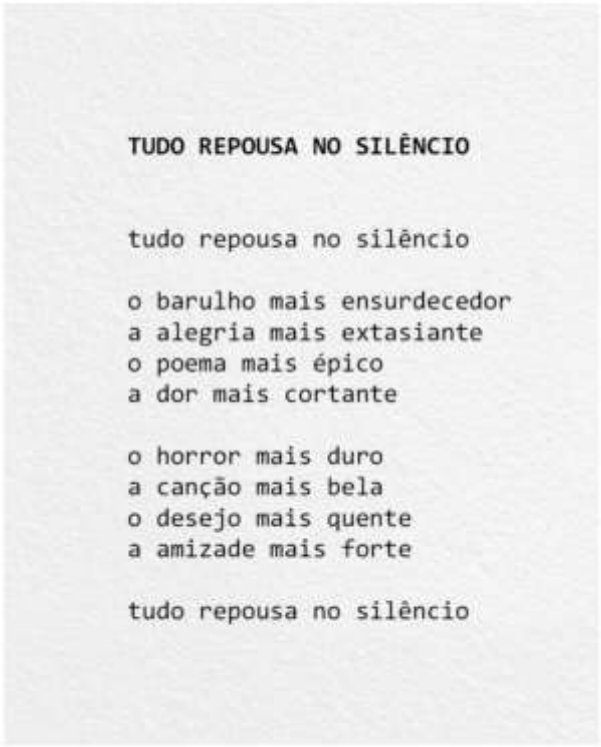
In [15]:
```
imagem_rgb = cv2.cvtColor(imagem, cv2.COLOR_BGR2RGB)
imagem_rgb2 = cv2.cvtColor(imagem2, cv2.COLOR_BGR2RGB)
```

Exibindo a imagem usando o matplotlib

In [16]:
```
plt.imshow(imagem_rgb)
plt.axis('off')  # Desligue os eixos
plt.show()
```



In [17]:
```
plt.imshow(imagem_rgb2)
plt.axis('off')  # Desligue os eixos
plt.show()
```



Fazendo o OCR na imagem

In [18]:
```
imagem_cinza = cv2.cvtColor(imagem, cv2.COLOR_BGR2GRAY)
print (imagem_cinza)
```

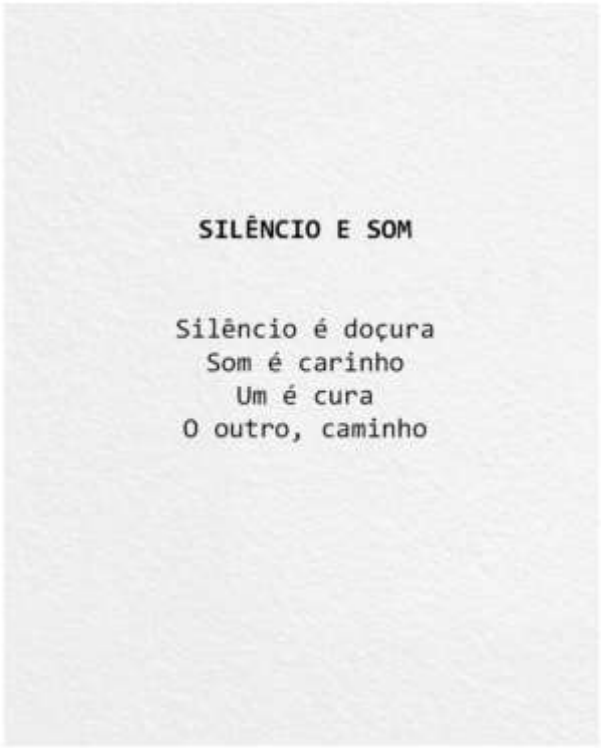
```
[[238 237 237 ... 247 244 239]
 [238 236 235 ... 246 246 243]
 [240 238 235 ... 242 245 245]
 ...
 [241 241 241 ... 240 240 239]
 [238 239 240 ... 242 240 237]
 [234 235 236 ... 244 241 237]]
```

In [19]:
```
imagem_cinza2 = cv2.cvtColor(imagem2, cv2.COLOR_BGR2GRAY)
print (imagem_cinza2)
```

```
[[238 237 237 ... 247 244 239]
 [238 236 235 ... 246 246 243]
 [240 238 235 ... 242 245 245]
 ...
 [241 241 241 ... 240 240 239]
 [238 239 240 ... 242 240 237]
 [234 235 236 ... 244 241 237]]
```

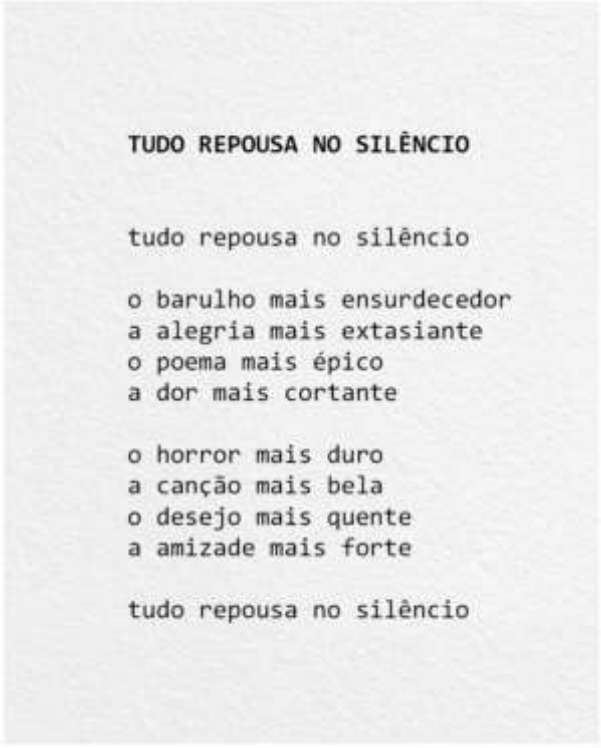
In [20]:
```
imagem_rgb_cinza = cv2.cvtColor(imagem_cinza, cv2.COLOR_BGR2RGB)
```

In [21]:
```
plt.imshow(imagem_rgb_cinza)
plt.axis('off')  # Desligue os eixos
plt.show()
```



In [22]:
```
imagem_rgb_cinza2 = cv2.cvtColor(imagem_cinza2, cv2.COLOR_BGR2RGB)
```

In [23]:
```
plt.imshow(imagem_rgb_cinza2)
plt.axis('off')  # Desligue os eixos
plt.show()
```



Tranformando imagem em texto

In [24]:
```
texto = pytesseract.image_to_string(imagem)
```

In [25]:
```
texto2 = pytesseract.image_to_string(imagem2)
```

Imprimindo o Texto identificado na imagem

In [26]:
```
print(texto)
print(texto2)
```

```
SILENCIO E SOM

Siléncio é docura
Som é carinho
Um 6 cura
O outro, caminho

TUDO REPOUSA NO SILENCIO

tudo repousa no siléncio

barulho mais ensurdecedor
alegria mais extasiante
poema mais épico

dor mais cortante

9owvo

horror mais duro
cancao mais bela
desejo mais quente
amizade mais forte

ooo

tudo repousa no siléncio
```

In [27]:
```
texto_extraido = pytesseract.image_to_string(imagem_cinza)
```

In [28]:
```
texto_extraido2 = pytesseract.image_to_string(imagem_cinza2)
```

Salvar o texto em um arquivo de texto

In [29]:
```
with open('texto_extraido.txt', 'w', encoding='utf-8') as arquivo:
    arquivo.write(texto_extraido)

with open('texto_extraido2.txt', 'w', encoding='utf-8') as arquivo:
    arquivo.write(texto_extraido2)
```

Ler o texto do arquivo

In [30]:
```
with open('texto_extraido.txt', 'r', encoding='utf-8') as arquivo:
    texto = arquivo.read()

with open('texto_extraido2.txt', 'r', encoding='utf-8') as arquivo:
    texto2 = arquivo.read()
```

Dividir o texto em palavras

In [31]:
```
palavras = texto.split()
```

In [32]:
```
palavras2 = texto2.split()
```

Contar o número de palavras

In [33]:
```
numero_de_palavras = len(palavras)
numero_de_palavras2 = len(palavras2)


print("Número de palavras no texto:", numero_de_palavras)
print("Número de palavras no texto2:", numero_de_palavras2)
```

```
Número de palavras no texto: 15
Número de palavras no texto2: 38
```

Levenshtein é conhecida como a distância de edição, que é utilizada para medir a similaridade entre as strings de um texto para o outro.

Sendo assim, será preciso instalar o pacote com o comando a seguir

In [34]:
```
pip install python-Levenshtein
```

```
Requirement already satisfied: python-Levenshtein in c:\users\gisle\anaconda3\lib
\site-packages (0.22.0)
Requirement already satisfied: Levenshtein==0.22.0 in c:\users\gisle\anaconda3\lib
\site-packages (from python-Levenshtein) (0.22.0)
Requirement already satisfied: rapidfuzz<4.0.0,>=2.3.0 in c:\users\gisle\anaconda3
\lib\site-packages (from Levenshtein==0.22.0->python-Levenshtein) (3.3.1)
Note: you may need to restart the kernel to use updated packages.
```

Importar o pacote Levenshtein

In [35]:
```
import Levenshtein
```

Calcular a distância de Levenshtein entre os dois textos

In [36]:
```
distancia = Levenshtein.distance(texto_extraido, texto_extraido2)
print("Distância de Levenshtein entre os textos:", distancia)
```

```
Distância de Levenshtein entre os textos: 199
```

Comparativo entre um texto e outro

In [37]:
```
limite = 10  # Definindo um limite
if distancia <= limite:
    print("Os textos são semelhantes.")
else:
    print("Os textos são diferentes.")
```

```
Os textos são diferentes.
```

FIM