



# Quién soy yo

- Senior Data Scientist en Cognizant
- Especializado en Deep Learning y Computer Vision
- [LinkedIn](#)
- [GitHub](#)

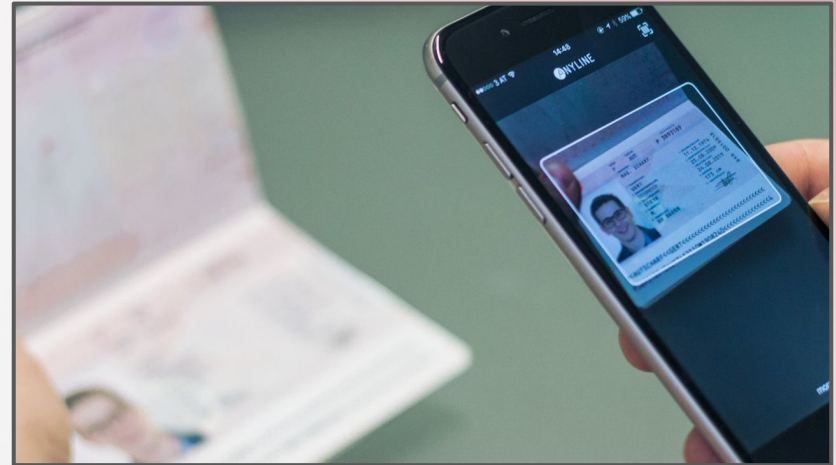
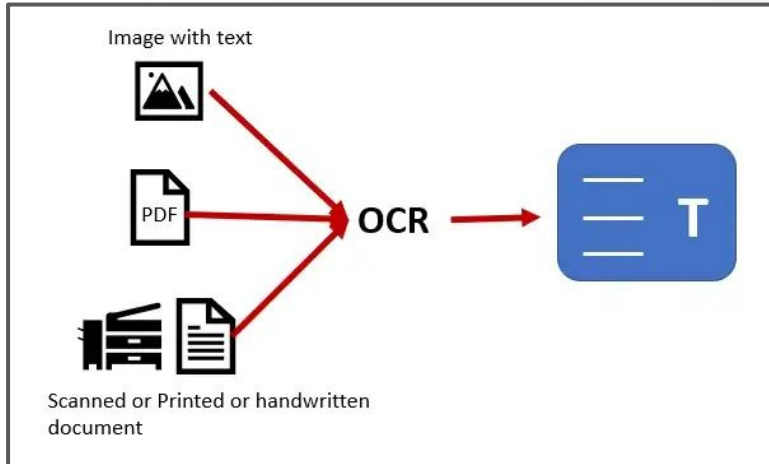


# Índice de la sesión

- Reconocimiento Óptico de Caracteres
- Tesseract OCR
- Casos Prácticos
- Práctica

# 1. Reconocimiento Óptico de Caracteres

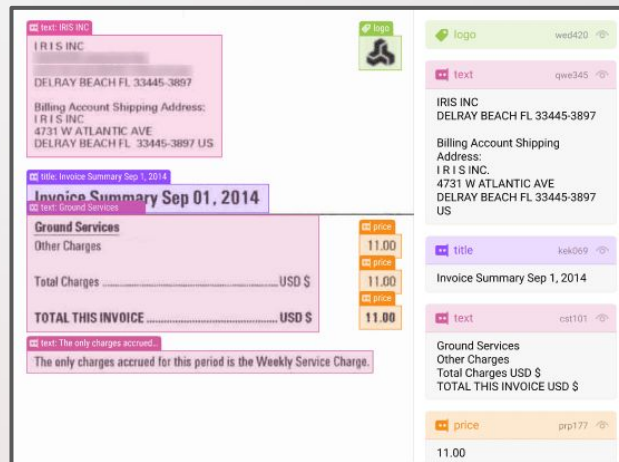
El reconocimiento óptico de caracteres (OCR), es una tecnología que convierte texto escrito a mano o a máquina a texto procesable.



## 2. Reconocimiento Óptico de Caracteres

Un sistema de OCR tiene muchos usos incluyendo transporte, retail y banca.

Puede ir apoyado de otros sistemas como reconocimiento de entidades, reconstrucción o clasificación.

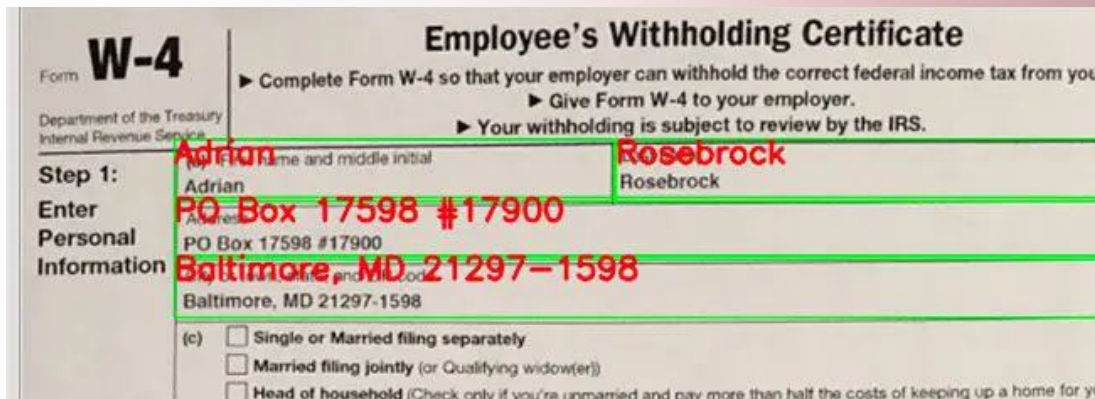


### 3. Tesseract OCR



## Tesseract OCR

Sistema Open-Source patrocinado por Google.  
 Soporta 116 idiomas, y 37 alfabetos.  
 Detecta la jerarquía del texto y puede leer en muchos formatos de texto (columnas, tablas, circular, etc).



**Form W-4**  
 Department of the Treasury  
 Internal Revenue Service

**Employee's Withholding Certificate**  
 ▶ Complete Form W-4 so that your employer can withhold the correct federal income tax from you.  
 ▶ Give Form W-4 to your employer.  
 ▶ Your withholding is subject to review by the IRS.

**Step 1:**  
**Enter Personal Information**

**First name and middle initial**  
 Adrian Rosebrock

**Address**  
 PO Box 17598 #17900  
 Baltimore, MD 21297-1598

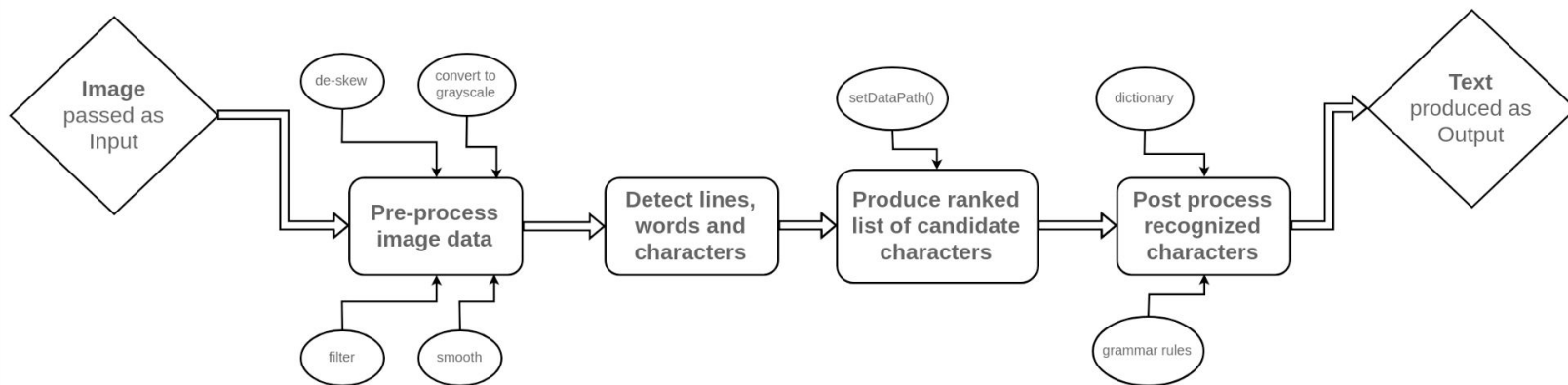
**City, state, and ZIP code**  
 Baltimore, MD 21297-1598

(c) ☐ Single or Married filing separately  
☐ Married filing jointly (or Qualifying widow(er))  
☐ Head of household (Check only if you're unmarried and pay more than half the costs of keeping up a home for you)

## 4. Tesseract OCR

Tesseract necesita un preprocesamiento de imagen para aumentar su robustez.

Detección y cambio de perspectiva, rotación, mejora de contraste, binarización, reducción de ruido, etc.





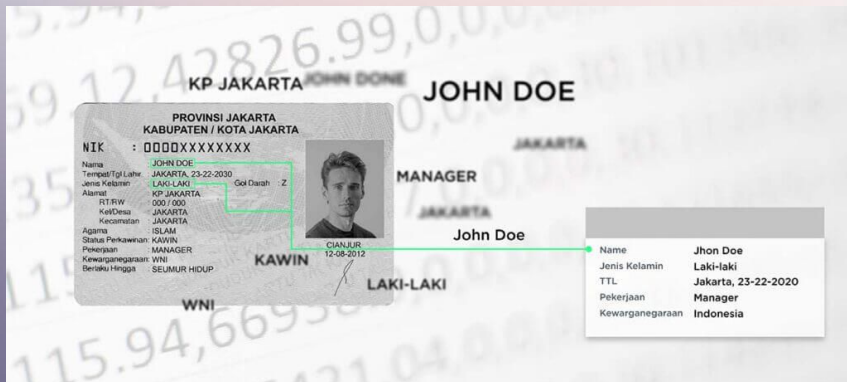
## Page segmentation modes:

- 0 Orientation and script detection (OSD) only.
- 1 Automatic page segmentation with OSD.
- 2 Automatic page segmentation, but no OSD, or OCR. (not implemented)
- 3 Fully automatic page segmentation, but no OSD. (Default)
- 4 Assume a single column of text of variable sizes.
- 5 Assume a single uniform block of vertically aligned text.
- 6 Assume a single uniform block of text.
- 7 Treat the image as a single text line.
- 8 Treat the image as a single word.
- 9 Treat the image as a single word in a circle.
- 10 Treat the image as a single character.
- 11 Sparse text. Find as much text as possible in no particular order.
- 12 Sparse text with OSD.
- 13 Raw line. Treat the image as a single text line, bypassing hacks that are Tesseract-specific.

## 5. Casos Prácticos



OCR para reconocimiento de documentos para un banco español.





## 6. Práctica

Lectura de PDFs

Lectura de documentos escaneados

Extracción de entidades

LINK



colab

¡Muchas gracias por  
uniros a esta W3 Drop!