

Clasificación, detección y análisis de letras escritas a mano y por ordenador

Jordi Morales (1564921), Marina Vázquez
(1563735), Mireia Fernández (1562636)

01

INTRODUCCIÓN

■ Objetivos y datos utilizados

OBJETIVOS

CLASIFICACIÓN DE DOCUMENTOS

Documentos manuscritos contra documentos digitales



SEGMENTACIÓN DE TEXTO

Encontrar palabras en los documentos según su origen



CONVERSIÓN DE IMÁGENES A TEXTO

Dada una imagen cualquiera, extraer el texto que contiene



DATOS DE ENTRENAMIENTO Y VALIDACIÓN

- Construcción de un dataset propio
 - Banco de 500 imágenes separadas en:
 - 400 imágenes de entrenamiento
 - 100 imágenes de validación
 - 50% imágenes digitales, 50% manuscritos

**Imagen de documento
manuscrito**

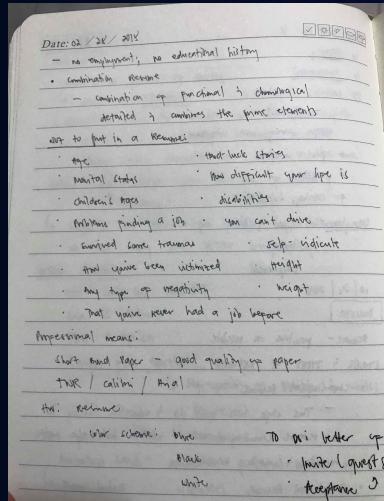
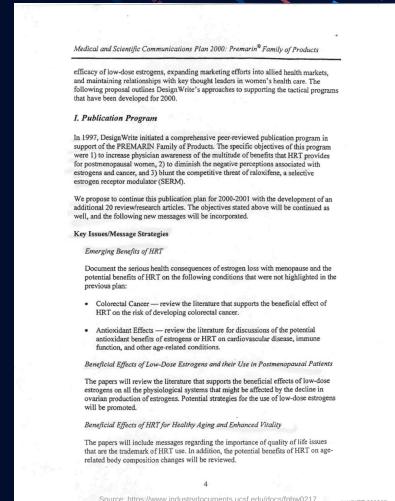


Imagen de documento digital





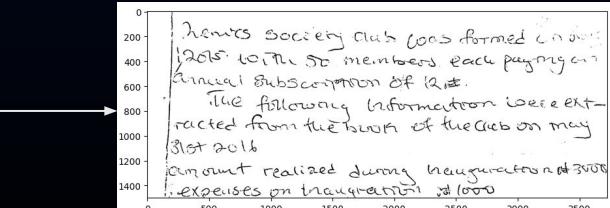
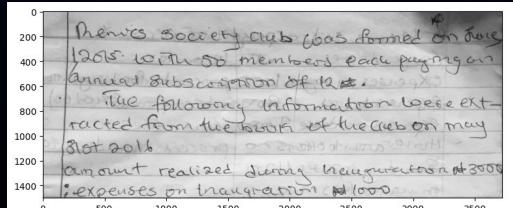
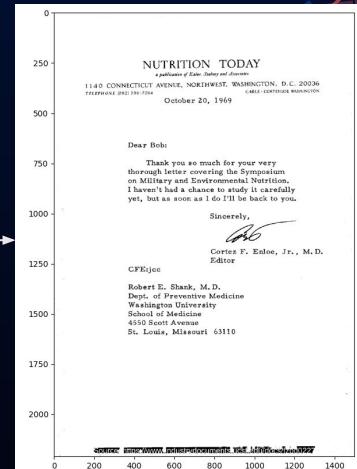
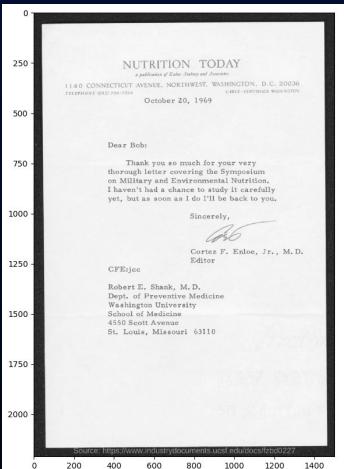
PREPROCESAMIENTO DE LAS IMÁGENES

Escalado

- Escalado dinámico manteniendo la proporciones. Mínimo 1500.

Black Hat

- Resaltamos primero y después binarizamos para destacar el texto.



02

MÉTODOS Y RESULTADOS

Soluciones implementadas para cumplir
nuestros objetivos



CLASIFICACIÓN SEGÚN ORIGEN

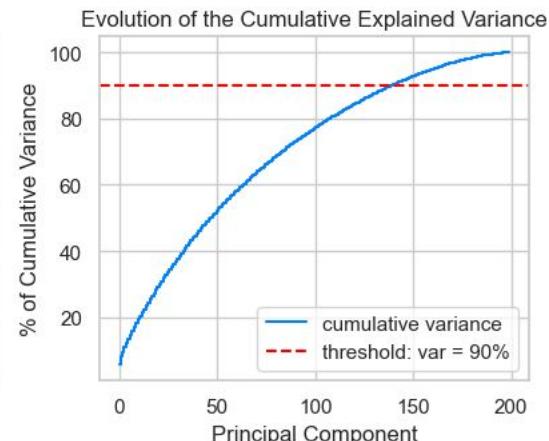
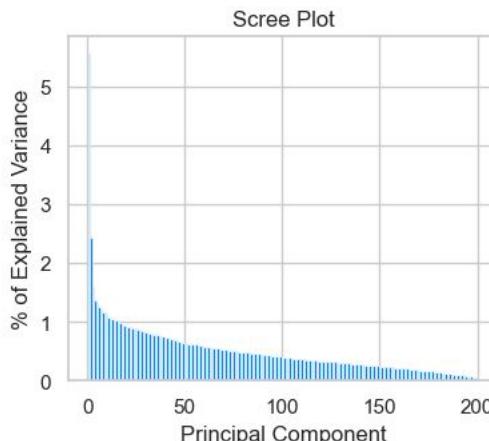
Preprocesado

- Escalado a mínimos. Necesitamos que todas las imágenes tengan las mismas dimensiones: (647, 670).

PCA

- Reducción a vectores de 140 valores utilizando un PCA sobre las imágenes vectorizadas.
- Hemos seleccionado un umbral de variancia explicada del 90%.

Reducción de dimensionalidad utilizando un PCA





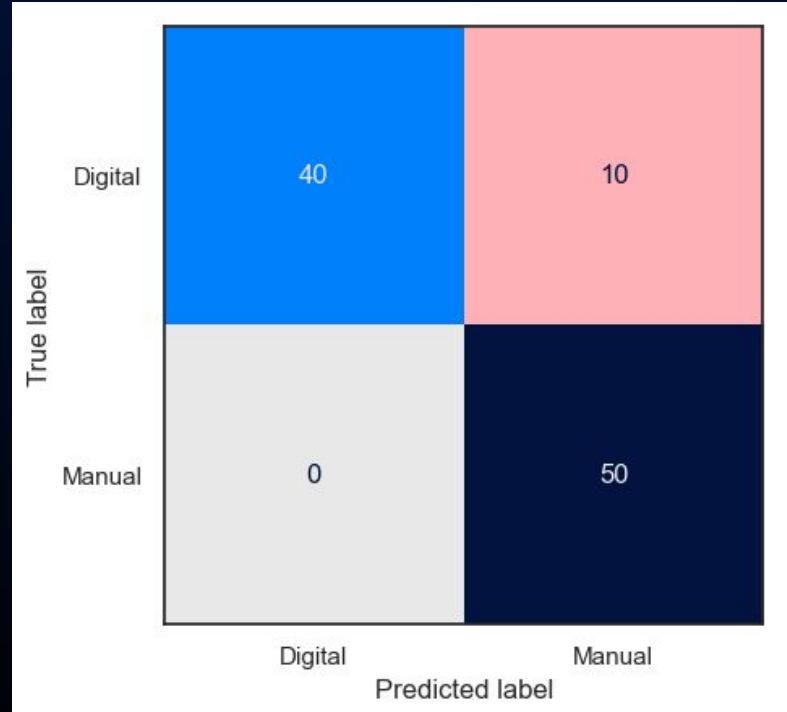
CLASIFICACIÓN SEGÚN ORIGEN: Resultado

Modelo

- Random Forest Classifier con 150 árboles de decisión.

Resultados

- Precisión obtenida:
92% en entrenamiento
90% en validación
- Tendencia a clasificar documentos digitales como manuscritos.



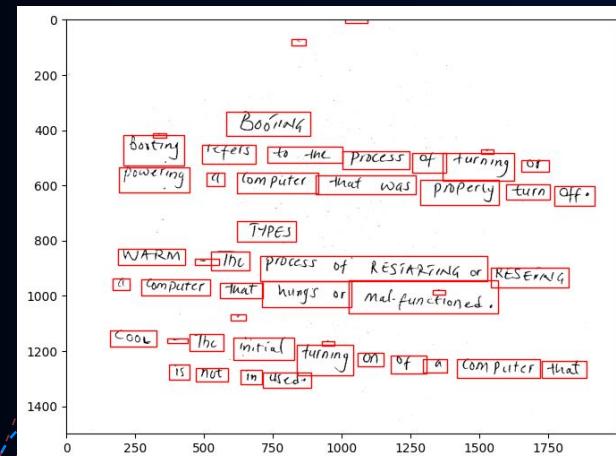
SEGMENTACIÓN DE TEXTO

Paso previo

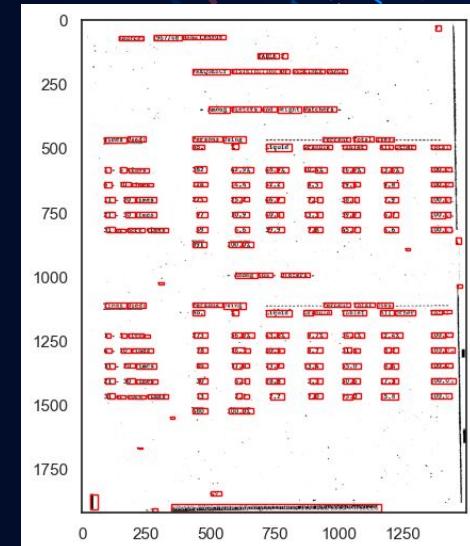
- Aplicamos una dilatación para agrupar letras en palabras.

Segmentación de palabras

- Buscamos contornos en la imagen.
- Filtramos para obtener propuestas de palabras.

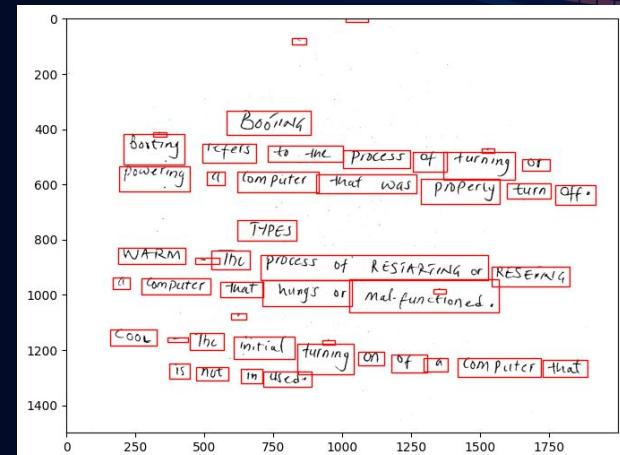
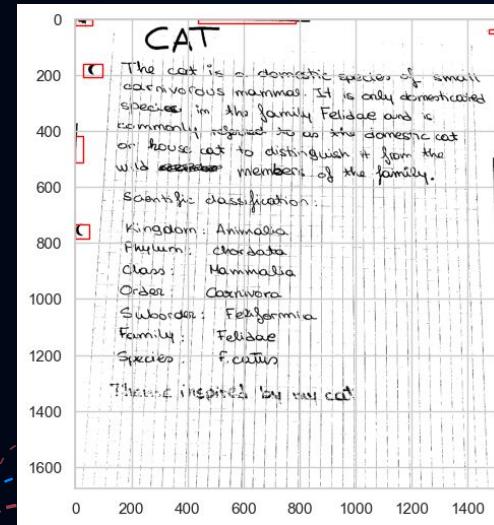
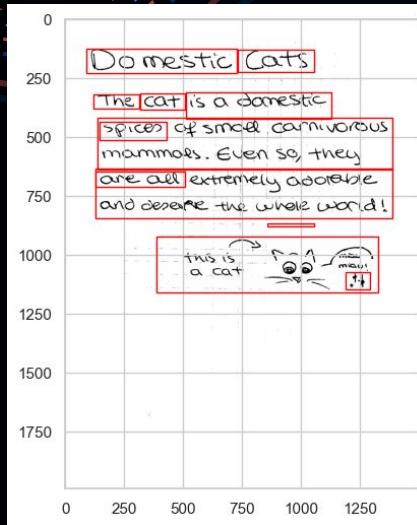


Documento manuscrito



Documento digital

SEGMENTACIÓN DE TEXTO: Ejemplos de Resultado



Ejemplo de documento manuscrito mal segmentado

Ejemplo de documento manuscrito mal segmentado

Ejemplo de documento manuscrito bien segmentado

CLASIFICADOR DE CARACTERES

FINETUNING DE UNA RED EXISTENTE

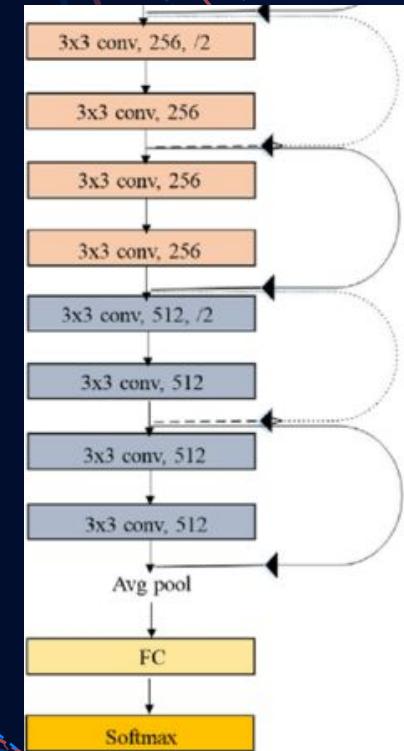
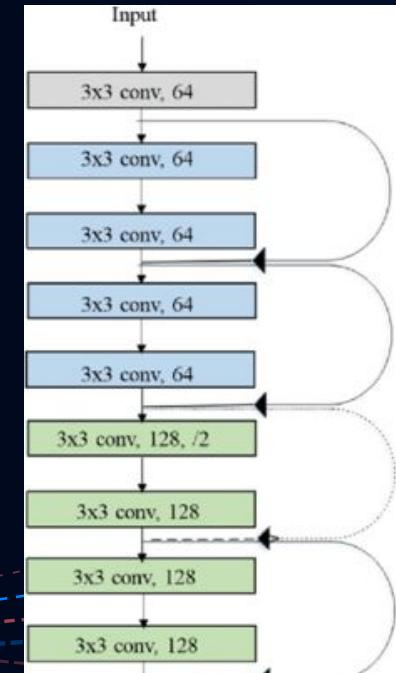
- Uso de la ResNet-18 pre-entrenada en ImageNet.

CHARS74K

- ~66.000 imágenes
- 62 clases:
(0-9, A-Z, a-z)

ENTRENAMIENTO

- 84.56% precisión durante validación



CLASIFICADOR DE CARACTERES: Resultados

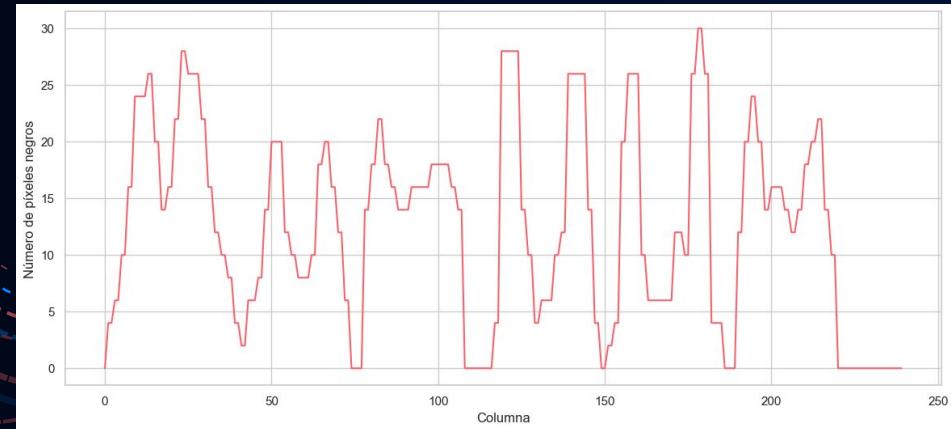
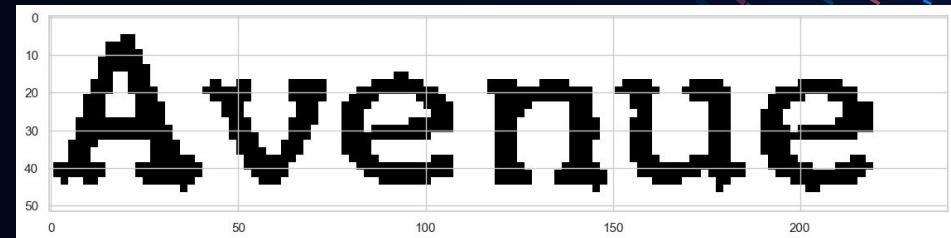
RECONOCIMIENTO ÓPTICO DE CARACTERES



OCR: DOCUMENTOS DIGITALES

PROCEDIMIENTO

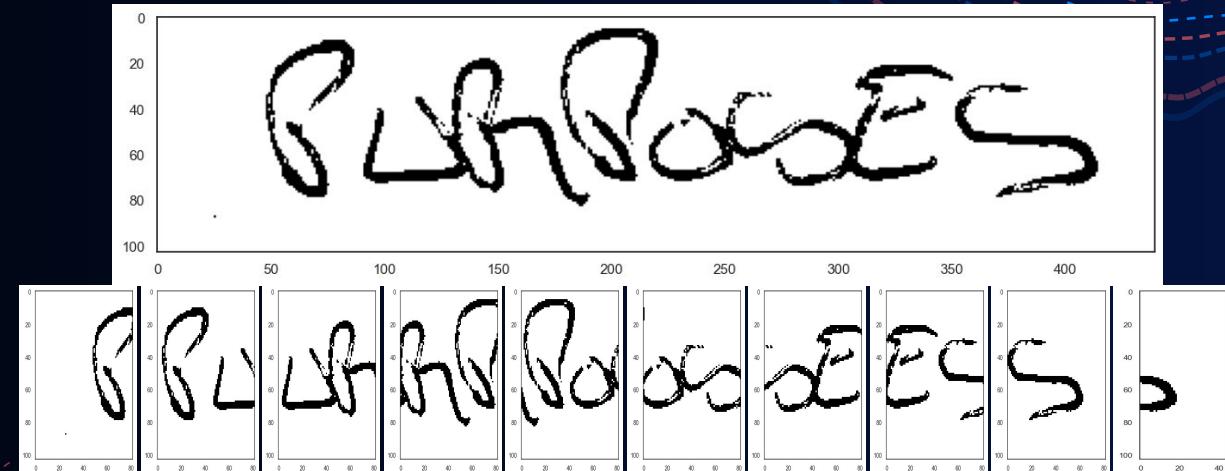
1. Separamos cada palabra en posibles caracteres.
2. Filtramos signos de puntuación y anomalías.
3. Preprocesamos la imagen de los caracteres.
4. Clasificamos.



OCR: DOCUMENTOS MANUSCRITOS

PROCEDIMIENTO

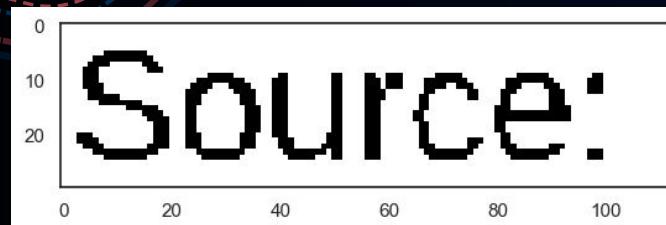
1. Utilizamos una ventana deslizante para obtener subimagenes.
2. Filtramos signos de puntuación y anomalías.
3. Preprocesamos las subimagenes restantes.
4. Clasificamos.
5. Utilizamos una función de similitud para encontrar la palabra real más parecida



EJEMPLOS DE TEXTO EXTRAÍDO

Extrae:

souce

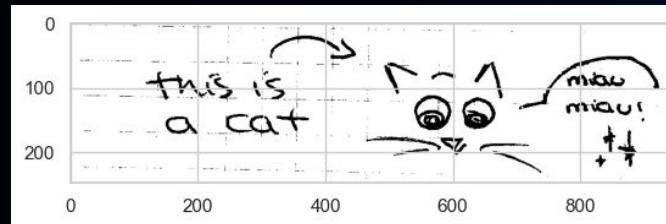


Extrae:

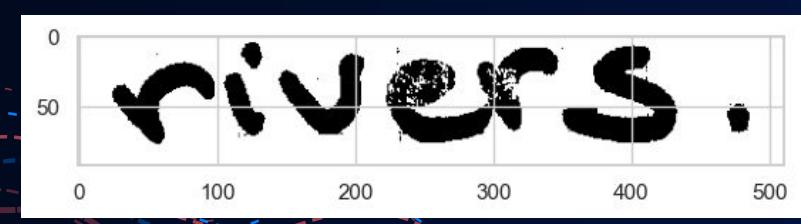
ofshares6



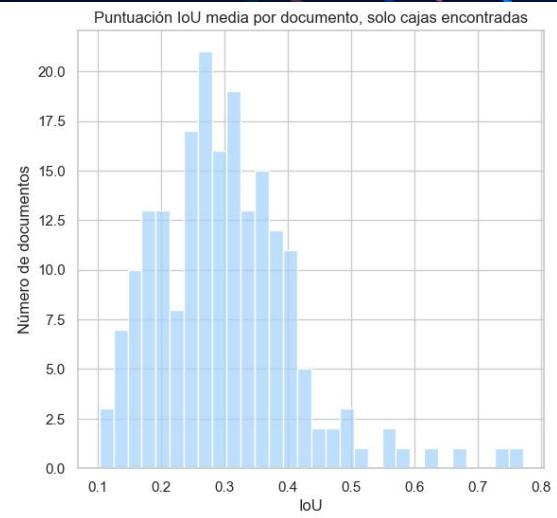
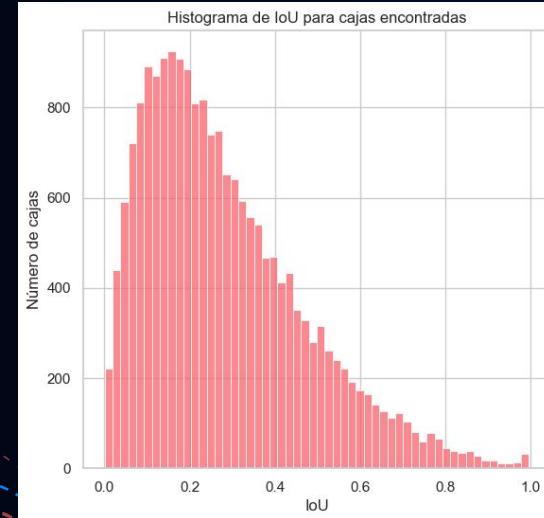
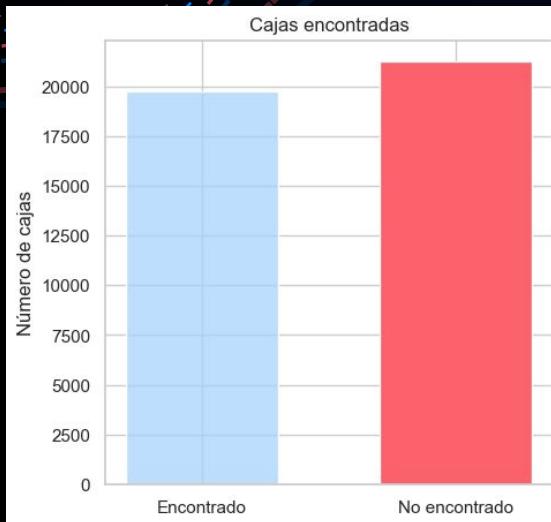
Extrae: mmmm
Convierte: common



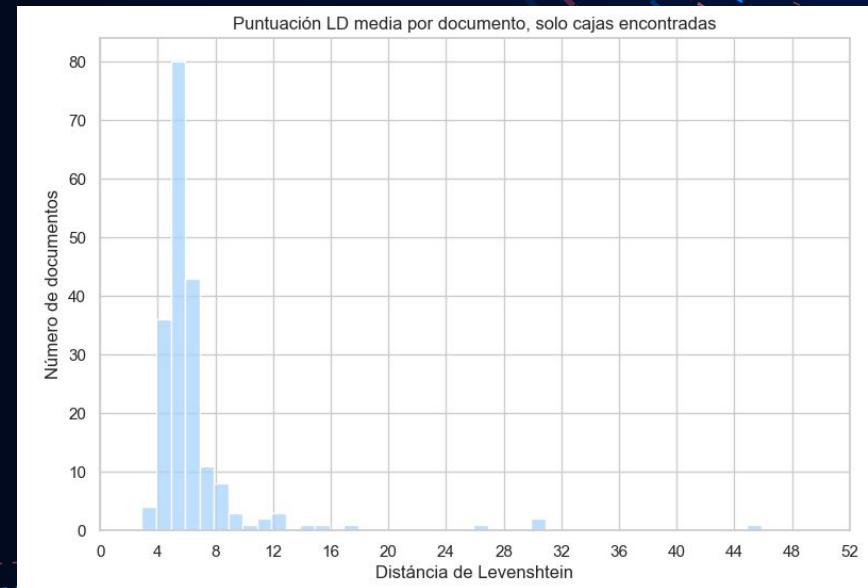
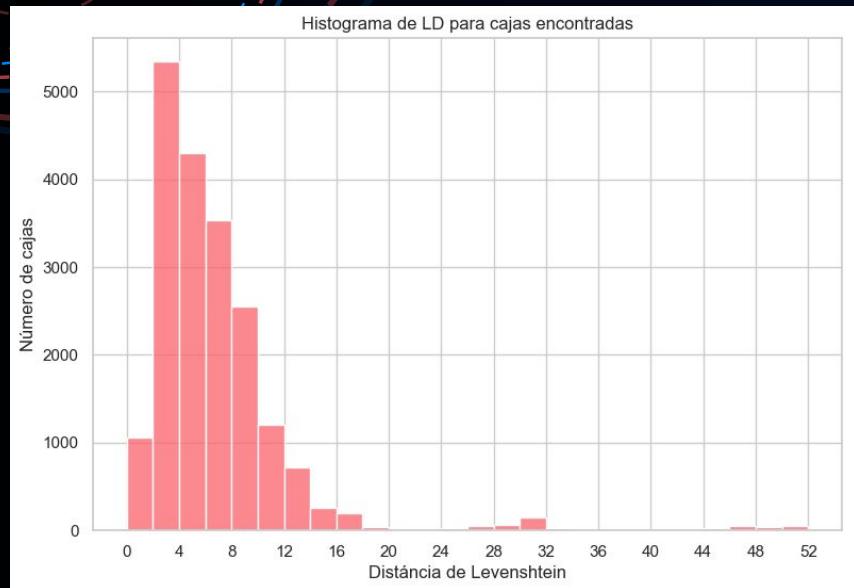
Extrae: mwwwn
Convierte: man



ANÁLISIS CUANTITATIVO



ANÁLISIS CUANTITATIVO



03. CONCLUSIONES

- Clasificar los documentos según su origen utilizando un modelo RF ha sido un éxito.
- Nuestro programa de segmentación no es robusto: algunos documentos no han segmentado nada bien pero en otros rozaba la perfección.
- Utilizar nuestro OCR para documentos digitales sí que ha funcionado bastante bien.
- Nuestro OCR para documentos manuscritos no se ha acercado a la realidad en ningun caso.



The background features a dark navy blue gradient with a subtle, organic pattern of wavy, undulating lines in shades of blue, teal, and light purple. These lines create a sense of depth and movement across the slide.

FIN!

Muchas gracias por su atención. Alguna pregunta?