Hackathon RIIAA 2021

"JusticIA para los desaparecidos"

Equipo: Pista latente ml

Reto: 1 y 2, Limpieza de imágenes y Extracción de texto

PROBLEMA

RETO ABORDADO

Reto 1 y Reto 2.

DESCRIPCIÓN DEL PROBLEMA

El Reto 1, tenía como objetivo la mejora de la calidad de la imágen y la clasificación de los objetos segmentados, tales como sellos, firmas, lineas, imágenes y letras escritas a mano o impresas.

El reto 2, tenía como objetivo la extracción de información de las imágenes y su identificación semantica, sobre todo para nombre de personas, lugares, fechas y organizaciones.

DIFICULTADES OBSERVADAS DEL PROBLEMA

Entre los problemas encontrados, están la baja resolución de las letras de las fichas al no estar a una distancia adecuada con la lente de la cámara con la que se capturó,
la iluminación en las imágenes, el bajo contraste entre el texto y el fondo de la imagen, la tinta que ya casi se caía del papel.

En el caso del reto 1, la poca información de sellos y firmas complicó su búsqueda y clasificación, lo mismo para los textos escritos a mano..

PLANTEAMIENTO DE LA SOLUCIÓN

COMO SE ABORDÓ EL PROBLEMA

De acuerdo al enfoque del reto, se utilizaron técnicas de procesamiento de imágenes para encontrar, segmentar y reconocer el texto dentro de la imagen. Al respecto de la extracción de entidades, lo primero fue revisar la calidad del texto extraído, para determinar pasos a seguir, como lo son la limpieza de caracteres y la aplicación de alguna etapa de comparación con palabras en un corpus. Al final se usó un procedimiento de etiquetado y búsqueda de expresiones gramaticales muy general, debido a que los textos pueden contener fechas y claves numéricas, además de sustantivos, las cuales pueden ser definitivas al momento de localizar direcciones o adscripciones de servidores públicos.

HERRAMIENTAS PRE-EXISTENTES USADAS

La técnica de extracción de entidades utilizada, tiene inspiración en los tutoriales publicados para la librería nltk.RegexpParser, utilizada para análisis gramatical, complementada con expresiones regulares para la búsqueda de claves numéricas. Para esto se ocupó un corpus en español y una función propia, de otros proyectos, que etiqueta las palabras de una frase.

Se utlizó pytesseract para reconocer el texto, OpenCV para realizar el procesamiento de las imágenes y Face_recognition para identificar personas en una imágen.

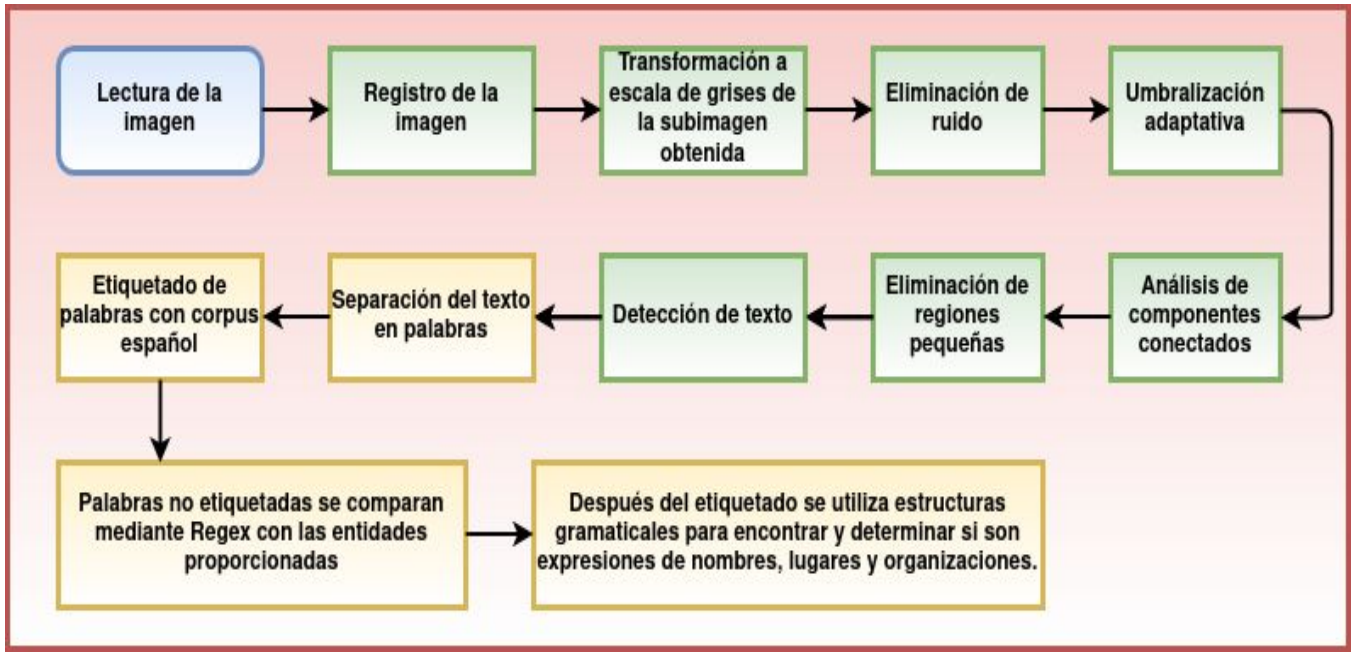
"JUSTICIA PARA LOS DESAPARECIDOS"

DESARROLLO DE LA SOLUCIÓN

TÉCNICAS QUE SE USARON PARA RESOLVER EL PROBLEMA

- Procesamiento de imágenes
- Procesamiento de lenguaje natural

FLUJO DEL ALGORITMO



PAQUETERÍAS USADAS

- OpenCV
- Scikit-Image
- Scikit-Learn
- PIL.Image
- Numpy
- Scipy
- Pytesseract
- Re
- NLTK
- Face_recognition
- Tensorflow/Keras

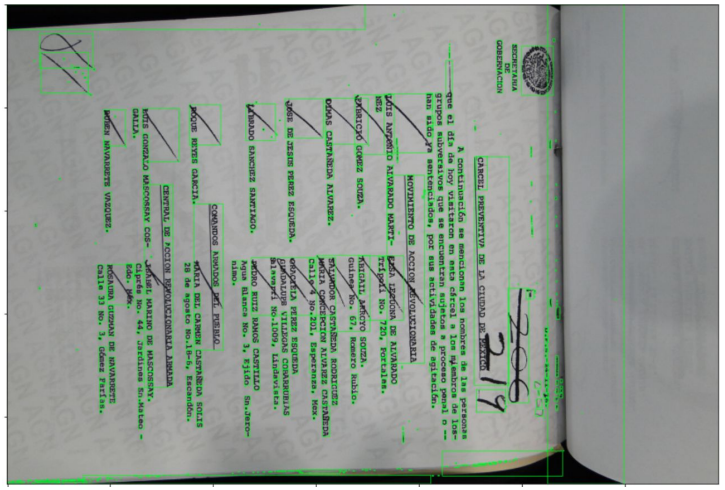
RESULTADOS

Resultados Reto 1:

El reto 1 y el reto 2 comparten el preprocesamiento de la imagen, ambos son recortados a la ficha u hoja y segmentados con el mismo método. En el caso del reto 1, consideramos que no es necesario hacer una orientación del documento.

Los pasos de segmentación son: filtro gaussiano para limpiar de ruido, blending para reforzar los bordes, umbralización adaptativa para binarizar y análisis de componentes conectados para limpiar de regiones que no cumplan el tamaño suficiente para ser letra u firma/sello. Como segunda alternativa se propone; corregir la orientación de la imagen; eliminar ruido mediante, un filtro gaussiano, umbralización adaptativa y análisis de componentes conectadas; aplicar una operación morfológica de dilatación empleando un kernel de mayor anchura que altura.

Los bounding box, de las regiones segmentadas, en la imagen a color determinarán los parches donde extraemos características a analizar.



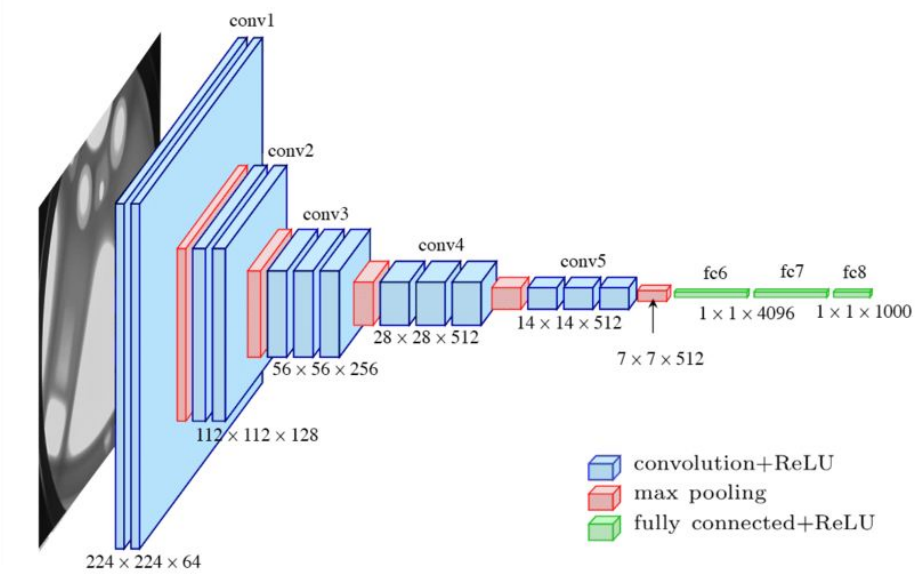
RESULTADOS

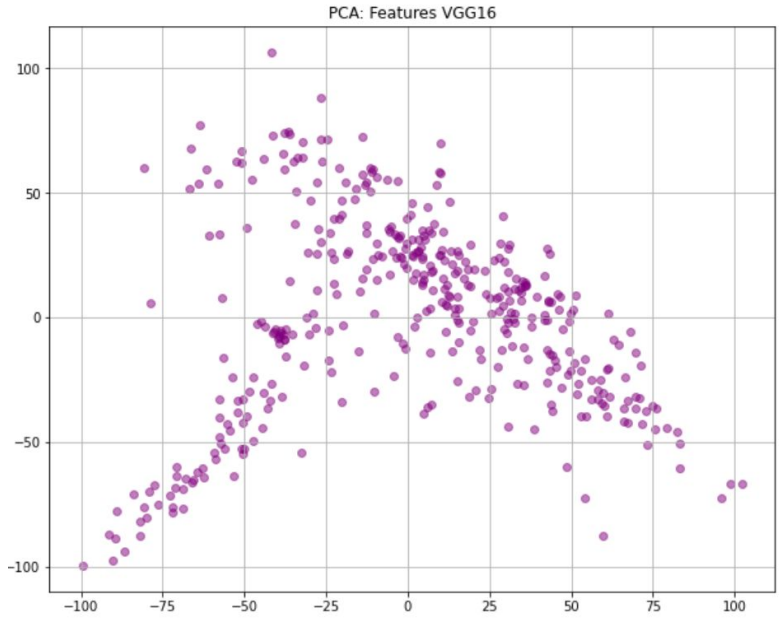
Resultados Reto 1:

Las características son extraídas mediante una red neuronal VGG16 (esto nos evita tener que hacer un preprocesamiento de los parches y una selección de características), la cual nos da un total de 512 características.

Eliminamos las columnas constantes de ceros y proyectamos utilizando PCA para quedarnos con las componentes que explican el 99% de la varianza.

Estás son pasadas al algoritmo de k-means para clusterizar los datos y ver si los objetos que nos interesan se agrupan de forma natural.





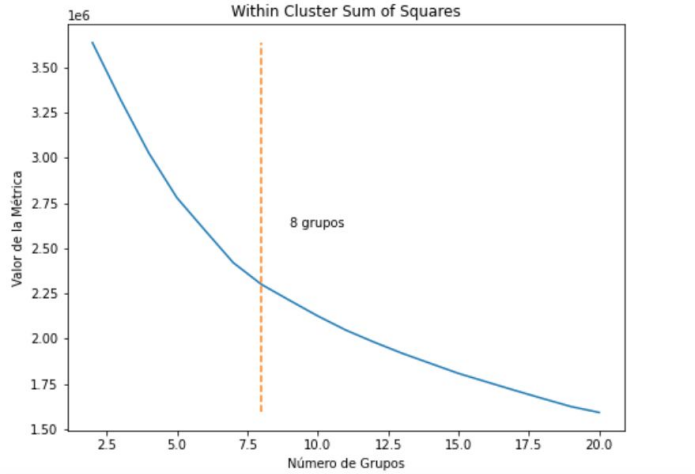
RESULTADOS

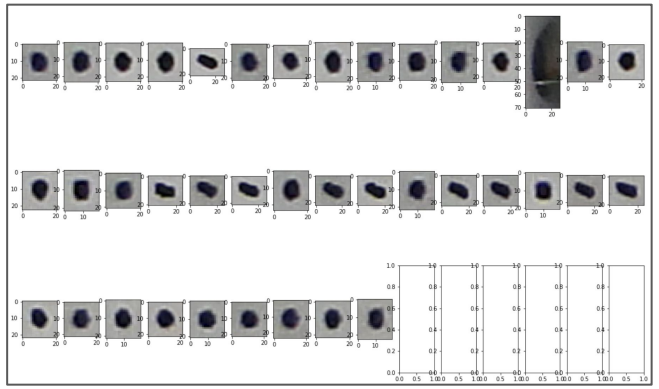
Resultados Reto 1:

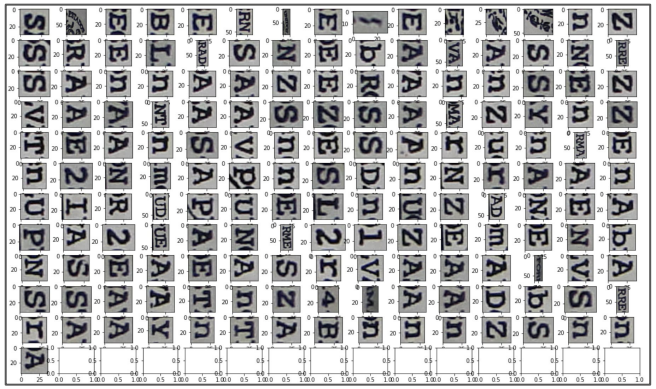
El número de clusters óptimo fue obtenido con la métrica WCSS, la cual mide la dispersión de los grupos con la suma de cuadrados.

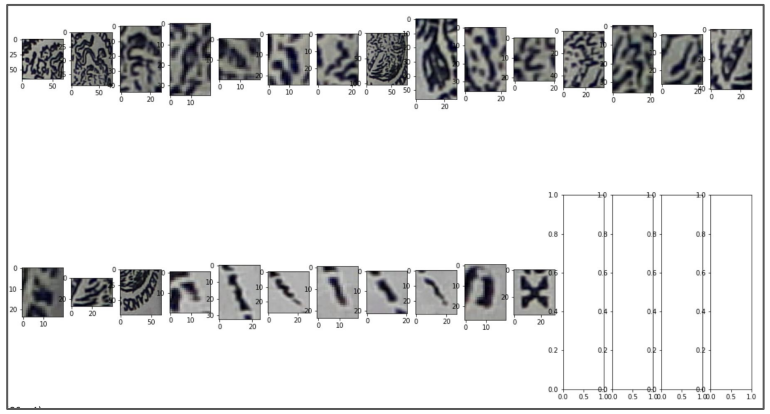
Algunos clúster obtenidos son los siguientes:

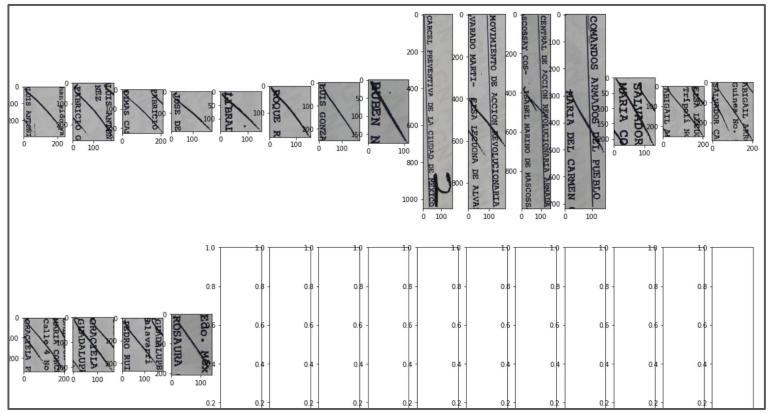
WCSS: Mide la dispersión de los grupos con la suma de cuadrados.











RESULTADOS

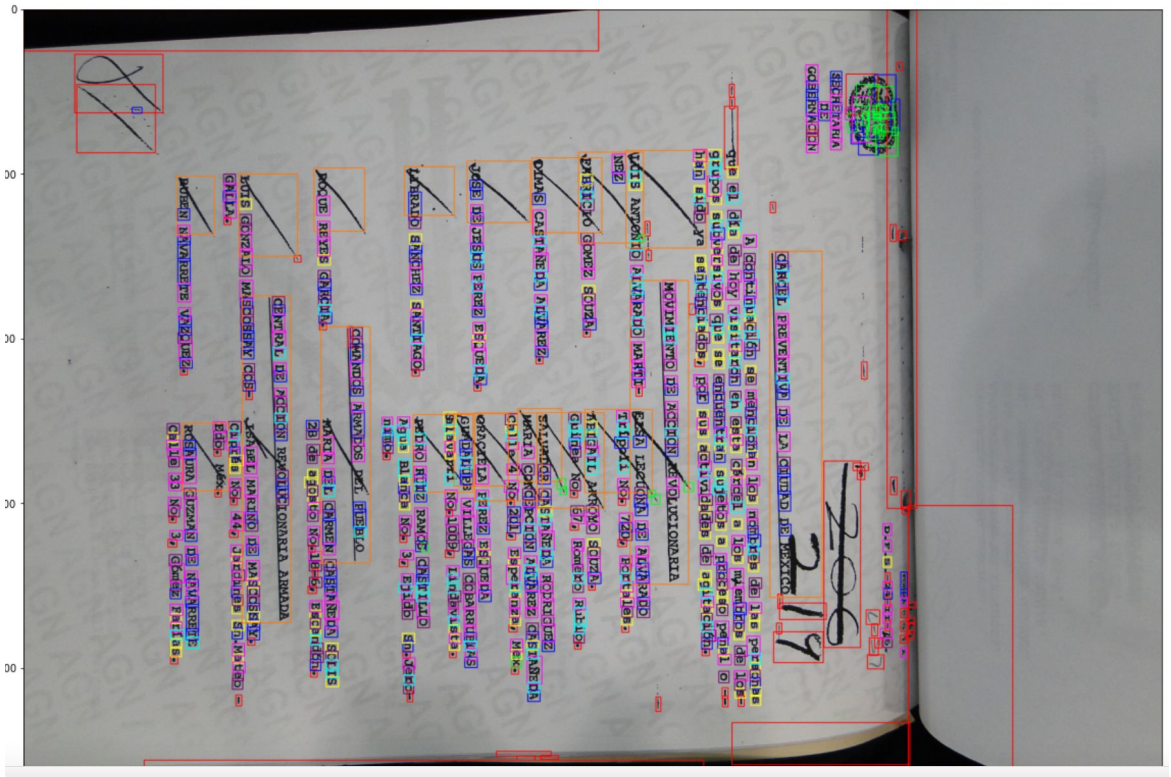
Resultados Reto 1:

La siguiente imagen muestra por colores cada cluster.

Como conclusión podemos decir que los objetos si se agrupan de forma natural con este conjunto de características.

Sin embargo, al tener pocos parches de algunos objetos, el agrupamiento puede ser difícil para estos, lo que no evita que tengamos parches incorrectos dentro de ciertos clusters.

Por lo anterior, resulta difícil etiquetar si el clúster pertenece a perches de sellos, firmas o texto escrito a mano. La asignación de etiqueta a cada cluster al final se hizo manual según lo observado en cada uno.



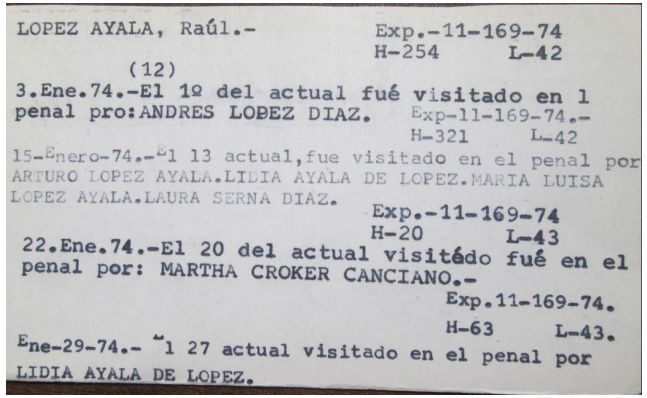
RESULTADOS

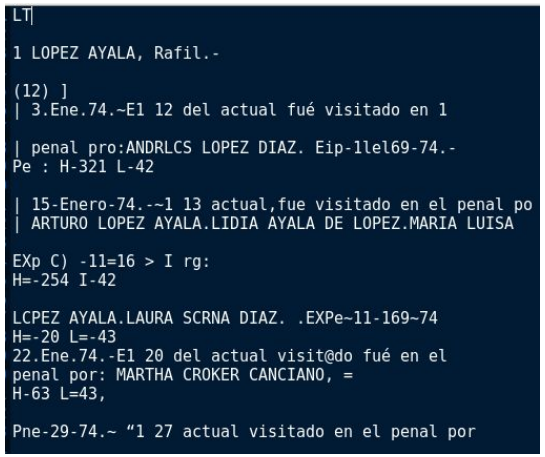
Resultados Reto 2:

Como se mencionó anteriormente, el registro de la imagen es similar al del reto 1. Sin embargo, aquí sí se utiliza el pytesseract para orientar la imágen. Identificamos que la orientación de pytesseract funciona mejor si primero recortamos la ficha.

Los pasos de segmentación son: filtro gaussiano para limpiar de ruido, blending para reforzar los bordes, umbralización adaptativa para binarizar y análisis de componentes conectados para limpiar de regiones que no cumplan el tamaño suficiente para ser letra, similar al reto 1.

Con lo anterior se obtiene una segmentación bastante limpia del texto, por lo que usamos directamente pytesseract para extraer el texto a un string.





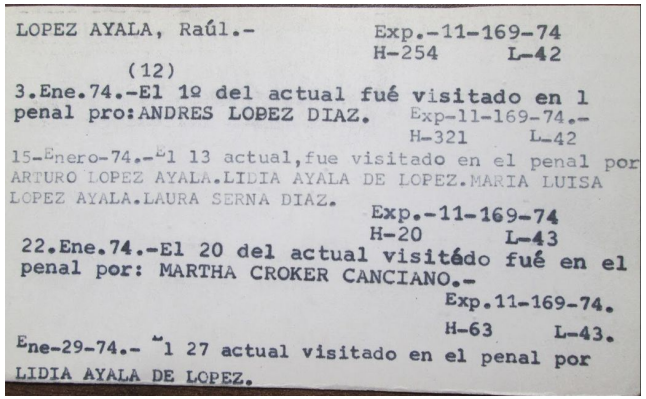
MainEnt	SecondEnt	PosiblesEnt
		12 n
		12
		fué
		1nn
		pro
		13
		C
		l rg
		20
		visit do fué
		1 27
LOPEZ AYALA		12 n1
LOPEZ DIAZ		12
pon ARTURO LOPEZ		fué
AYALA DE LOPEZ.MARIA		1nn
DIAZ		pro
MARTHA CROKER CANCIANO		13
AYALA DE LOPEZ		JI EEE
LOPEZ AYALA Rail		20
LOPEZ DIAZ		visit do fud
LOPEZ		L43
AYALA DE LOPEZ.MARIA		1 27
MARTHA CROKER OF V 0		ela
AYALA		pl ann

RESULTADOS

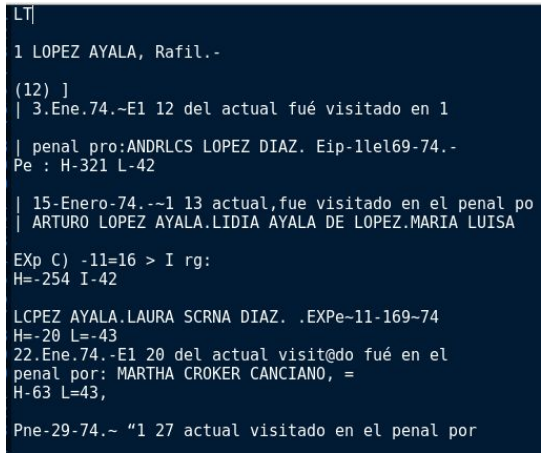
Resultados Reto 2:

La técnica de extracción de entidades utilizada, tiene inspiración en los tutoriales publicados para la librería nltk.RegexpParser, utilizada para análisis gramatical, complementada con expresiones regulares para la búsqueda de claves numéricas.

Para esto se ocupó un corpus en español y una función propia, de otros proyectos, que etiqueta las palabras de una frase



LOPEZ AYALA, Raúl.- Exp.-11-169-74
H-254 L-42
(12)
3.Ene.74.-El 1º del actual fué visitado en l
penal pro:ANDRES LOPEZ DIAZ. Exp-11-169-74.-
H-321 L-42
15-Enero-74.-El 13 actual,fue visitado en el penal por
ARTURO LOPEZ AYALA.LIDIA AYALA DE LOPEZ.MARIA LUISA
LOPEZ AYALA.LAURA SERNA DIAZ.
Exp.-11-169-74
H-20 L-43
22.Ene.74.-El 20 del actual visitado fué en el
penal por: MARTHA CROKER CANCIANO.-
Exp.11-169-74.
H-63 L-43.
Ene-29-74.- "1 27 actual visitado en el penal por
LIDIA AYALA DE LOPEZ.



```
LT

1 LOPEZ AYALA, Rafil.-

(12) ]
| 3.Ene.74.~E1 12 del actual fué visitado en 1

| penal pro:ANDRLCS LOPEZ DIAZ. Eip-1lel69-74.-
Pe : H-321 L-42

| 15-Enero-74.-~1 13 actual,fue visitado en el penal po
| ARTURO LOPEZ AYALA.LIDIA AYALA DE LOPEZ.MARIA LUISA

EXp C) -11=16 > I rg:
H=-254 I-42

LCPEZ AYALA.LAURA SCRNA DIAZ. .EXPe~11-169~74
H=-20 L=-43
22.Ene.74.-E1 20 del actual visit@do fué en el
penal por: MARTHA CROKER CANCIANO, =
H-63 L=43,

Pne-29-74.~ "1 27 actual visitado en el penal por
```

MainEnt	SecondEnt	PosiblesEnt
		12 n
		12
		fué
		1nn
		pro
		13
		C
		l rg
		20
		visit do fué
		1 27
LOPEZ AYALA		12 n1
LOPEZ DIAZ		12
pon ARTURO LOPEZ		fué
AYALA DE LOPEZ.MARIA		1nn
DIAZ		pro
MARTHA CROKER CANCIANO		13
AYALA DE LOPEZ		JI EEE
LOPEZ AYALA Rail		20
LOPEZ DIAZ		visit do fud
LOPEZ		L43
AYALA DE LOPEZ.MARIA		1 27
MARTHA CROKER OF V 0		ela
AYALA		pl ann

NIVEL DE COMPLETITUD

- **Oportunidades de mejora a corto y mediano plazo**

- Detectar la ficha para mejorar el reconocimiento de las cadenas de texto.
- Es necesario revisar con detalle los caracteres que deben ser excluidos o limpiados al inicio del procesamiento del texto, debido a que muchos símbolos de puntuación son introducidos de manera artificial por el OCR, que representan pérdida de información sobre todo durante la limpieza y separación en palabras.

- **¿Qué nos faltó?**

- Se extrajeron las entidades, sin embargo, falto agregar la etiqueta de clasificación correspondiente a sí es un lugar, una persona u organización específicamente. Se optó por crear tres arreglos de entidades, uno para entidades como nombres, lugares y cargos, otro que contiene sustantivos en general y un tercero con claves y datos numéricos.
- Con respecto al reto 1, falta etiquetar cada cluster de parches. Aunque se realizó la agrupación de los bounding boxes, falta asignar las etiquetas de rostro, sello, firma, etc.

- **¿Identificaron una dificultad particular que de momento no encontraron como resolver? ¿Tienen una propuesta para resolverlo, que por cuestiones de tiempo no pudieron implementar?**

- Lograr separar el texto escrito a máquina con el escrito a mano.
- Eliminar los rayones de las imágenes de expedientes.
- No fue posible atribuir etiquetas de si se habla de un posible servidor público, un lugar o una organización de forma específica, con el método empleado, debido a que los tres tipos de entidades pueden contener nombres propios o sustantivos de lugares y gremios. Una propuesta de solución sería implementar un clasificador de frases utilizando word-embeddings, u otro método que permita un análisis de contexto más robusto que el análisis sintáctico.
- La propuesta para mejorar el resultado del reto 1, es la implementación de un algoritmo de deep learning conocido

INTEGRANTES DEL EQUIPO

- Andrea Berenice Ek Hobak
- Gabriela Marali Mundo Cortés
- Mario Xavier Canche Uc
- Myrna Citlali Castillo Silva
- Ramón Sidonio Aparicio García

Equipo: Pista Latente ML