Implementación de una herramienta basada en PLN para la detección y anonimización de datos personales en documentos

José Manuel Simón Ramos

Máster en Inteligencia de Negocio y Big Data en Entornos Seguros

jmsimonramos@gmail.com

Jueves, 09 de septiembre de 2021

Índice

Índice

Introducción I

- En los últimos años, el avance en el campo del Aprendizaje Automático, unido a las mejoras hardware, y al aumento de datos, ha motivado la utilización de técnicas de aprendizaje para automatizar procesos o extraer conocimiento a partir de los datos.

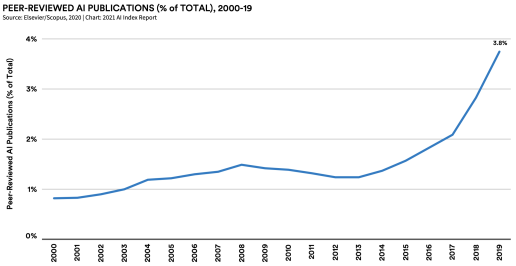


Figura: Evolución de las revisiones de artículos de Inteligencia Artificial. Fuente: *AI Index Report*.

Introducción II

- Desde el punto de vista del Procesamiento del Lenguaje Natural (PLN), la utilización de estos datos se encuentra más limitada:
  - Aparición de información de carácter personal en los datos.
- Esta característica, unida a la fuerte legislación que existe actualmente sobre la Protección de Datos (RGPD), hace que las administraciones hayan de tener un mayor control y cuidado a la hora de utilizar y/o compartir documentos con información personal.
- Técnicas para evitar estos problemas:
  - Eliminar las apariciones de las menciones personales.
  - **Anonimización de las menciones**.

Motivación

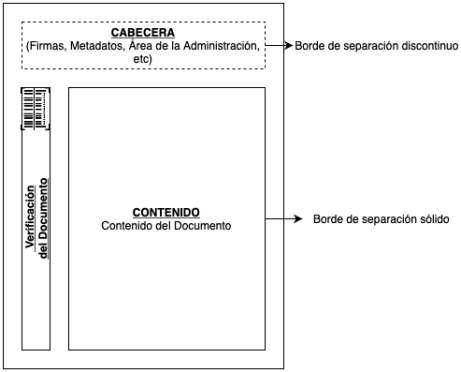
- Debido a ese control y privacidad sobre los datos, una pérdida en los mismos podría suponer grandes sanciones, además del problema que supone el filtrado de datos personales.
- En este proyecto se va a presentar una propuesta **genérica**, que permita tanto detectar distintos tipos de menciones personales en los datos, así como anonimizarlas.
- Además, la propuesta presentada abordará todas las etapas a la hora de anonimizar un documento:
  1) Obtención del texto del documento origen → Aplicando técnicas de OCR.
  2) Detección de las menciones personales → Utilizando modelos de Deep Learning para PLN.
  3) Anonimización de las menciones → Implementando mecanismos de reemplazos eficientes.
  4) Generación del nuevo documento anonimizado → Creando y modificando archivos PDF.

Índice

Conjunto de Datos

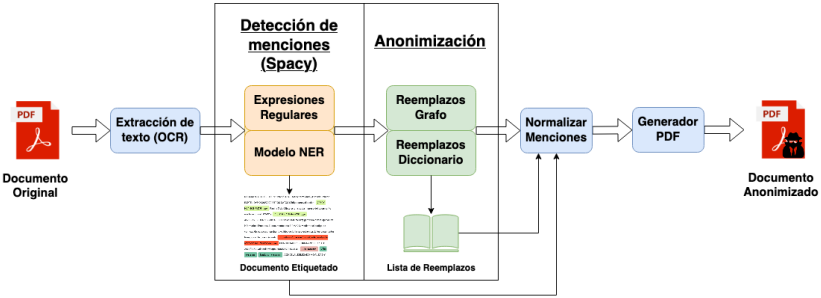
- El conjunto de datos se encuentra formado por 309 documentos administrativos de distintas clases: multas, permisos, pagos, etc.

ENTIDAD	DOCS	DOCS (%)	APARICIONES
NOMBRE	309	100.00	2.263
APELLIDO	309	100.00	4.526
DNI	79	25,57	710
DIRECCIÓN	160	51,68	323
CIUDAD	309	100.00	1.994
PROVINCIA	5	1,62	7
CP	119	38,51	141
TELÉFONO	117	37,86	118
REFERENCIA CATASTRAL	10	3,24	36
SEGURIDAD SOCIAL	-	-	-
CUENTA BANCARIA	17	5,50	56
EMAIL	3	0,97	4
MATRÍCULA	6	1,94	7
CSV	309	100.00	309
URL	282	91,26	282



Índice

Diseño del *Pipeline*

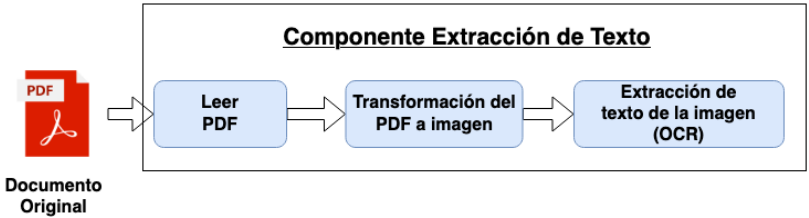


- El *pipeline* desarrollado se encuentra formado por los siguientes componentes:

1. Extracción de Texto.
2. Detección de menciones personales.
3. Anonimización.
4. Normalización de las menciones.
5. Generador de PDF.

Extracción del texto de los documentos

- Este componente se encarga de obtener el texto del documento PDF de entrada.



Detección de las menciones personales I

- Componente encargado de extraer las menciones personales del texto.
- Las menciones son detectadas utilizando modelos de Deep Learning generados utilizando la librería de PLN Spacy.
- Se ha implementado un componente de soporte basado en expresiones regulares para detectar menciones con una estructura predecible.

Tipo de detección	Precision	Recall	F1
RegEx	0,028	0,059	0,038
NER	0,953 ▲0,925	0,930 ▲0,871	0,942 ▲0,904
RegEx + NER	0,936 ▼0,017	0,949 ▲0,019	0,945 ▲0,003

Detección de las menciones personales II

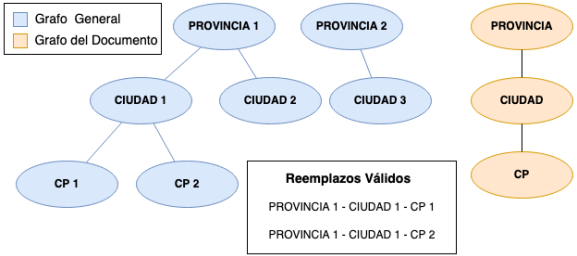
Entidad	NER			RegEx + NER		
	Precision	Recall	F1	Precision	Recall	F1
DNI	0,871	0,90	0,885	0,915 ▲0,044	0,973 ▲0,073	0,942 ▲0,057
REFERENCIA CATASTRAL	0	0	0	1 ▲1	0,812 ▲0,812	0,891 ▲0,891
CUENTA BANCARIA	0,056	0,067	0,061	1 ▲0,944	0,800 ▲0,733	0,889 ▲0,828
EMAIL	0	0	0	1 ▲1	1 ▲1	1 ▲1
MATRICULA	0	0	0	1 ▲1	1 ▲1	1 ▲1

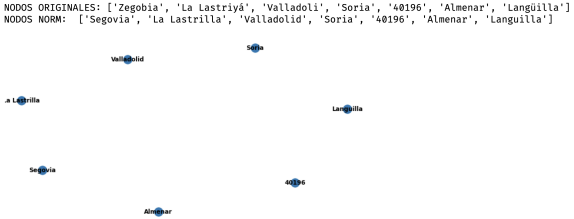
Anonimización de las menciones

- El componente de anonimización se ha desarrollado siguiendo un enfoque híbrido basado en grafos y diccionarios:
  - **Enfoque Basado en Grafos:** Permite generar reemplazos para los tipos de menciones que guardan relación entre sí (provincia, ciudad y código postal).
  - **Enfoque Basado en Diccionarios:** Permite generar reemplazos para el resto de las entidades.
- Alguna de las propiedades que cumplen los nuevos reemplazos son:
  - **Robustez** $\rightarrow$ Se generan reemplazos para cualquier tipo de mención.
  - **Integridad** $\rightarrow$ Mantienen la estructura y tipo del original.
  - **Concordancia** $\rightarrow$ Los nuevos reemplazos mantienen relaciones entre sí.

Generación de Reemplazos Mediante Grafos

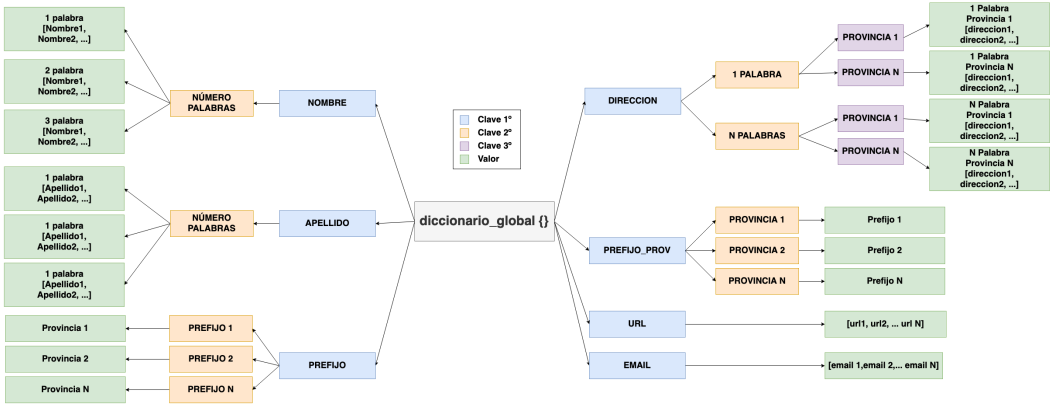
- Permite generar reemplazos entre distintos tipos de entidades manteniendo su jerarquía.
- La jerarquía implementada es **Provincia** → **Ciudad** → **Código Postal**.





Generación de Reemplazos Mediante Diccionarios

- Permite generar reemplazos para el resto de las entidades manteniendo su estructura.

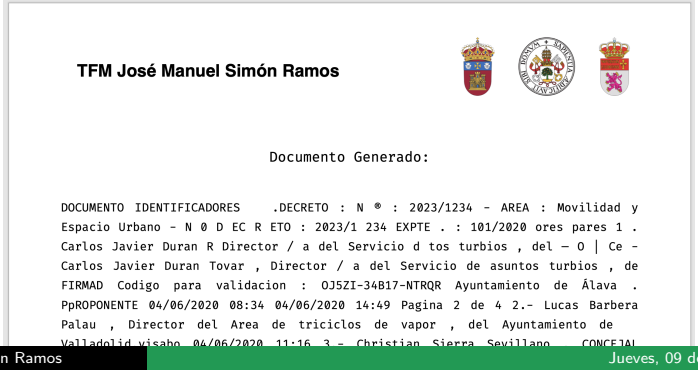


Normalización de las Menciones

- Algunos tipos de documentos tienen partes cuyo texto se escribe todo en mayúsculas o en minúsculas.
- Para evitar que sea evidente que el documento ha sido modificado, se normalizan los reemplazos para que estos tengan el formato de la mención original: todo mayúsculas, todo minúsculas, etc.

Generación del nuevo documento PDF

- Este componente se encarga de tomar el texto del documento original, y de generar un nuevo documento PDF con dicho texto y reemplazando las menciones por sus reemplazos normalizados.



TFM José Manuel Simón Ramos

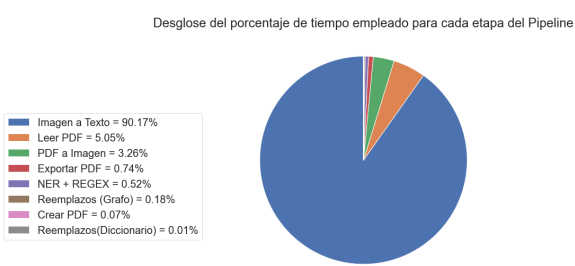
Documento Generado:

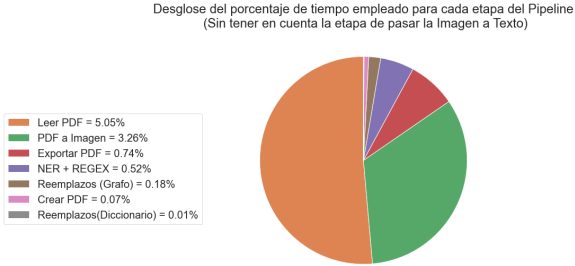
DOCUMENTO IDENTIFICADORES .DECRETO : N ® : 2023/1234 - AREA : Movilidad y Espacio Urbano - N 0 D EC R ETO : 2023/1 234 EXPTE . : 101/2020 ores pares 1 . Carlos Javier Duran R Director / a del Servicio d tos turbios , del — O | Ce - Carlos Javier Duran Tovar , Director / a del Servicio de asuntos turbios , de FIRMAD Codigo para validacion : OJ5ZI-34B17-NTRQR Ayuntamiento de Álava . PpROPONENTE 04/06/2020 08:34 04/06/2020 14:49 Pagina 2 de 4 2.- Lucas Barbera Palau , Director del Area de triciclos de vapor , del Ayuntamiento de Valladolid visado 04/06/2020 11:16 3.- Christian Sierra Sevillano , CONCEJAL

Índice

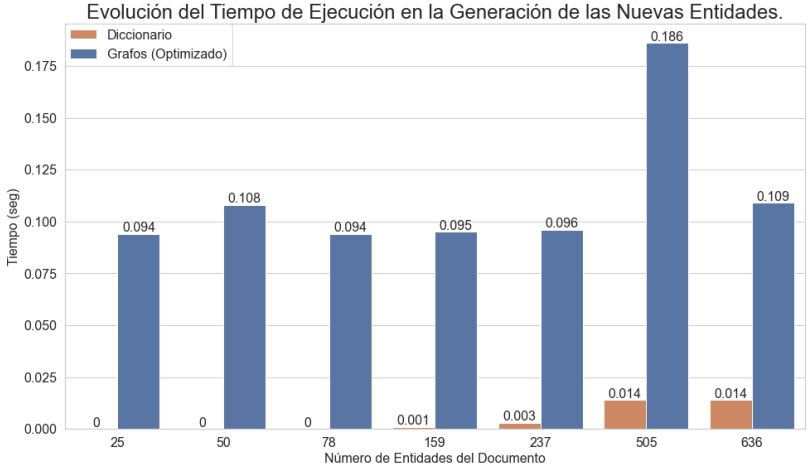
Rendimiento I

- Tras la realización de varias pruebas con distintos documentos de diferente tamaño, número de páginas, menciones, palabras, etc, los resultados (en términos de rendimiento) obtenidos son los siguientes:





Rendimiento II



Índice

Conclusiones

- Los resultados obtenidos durante la creación y optimización del modelo son muy buenos a pesar de la variedad de las menciones a detectar (en torno a un 94 % de detección).
- La utilización de expresiones regulares ha mejorado de forma sustancial los resultados para menciones con una estructura predecible (▲57,59 % de media).
- Los tiempos de ejecución de los componentes desarrollados en el proyecto son bajos (0,19 segundos de media para documentos entre 25 y 630 menciones $\approx$ 240 menciones de media), y se mantienen constantes (o aumentan muy ligeramente) a medida que aumentan las palabras o las menciones: **Escalabilidad**.
- A pesar de las limitaciones de la herramienta, esta puede considerarse genérica ya que es posible adaptarla a cualquier contexto modificando el modelo o la base de conocimiento para generar los reemplazos.

Trabajo Futuro

- Optimización o implementación de un mecanismo eficiente de búsqueda en grafos.
- Implementación de diccionarios dinámicos para realizar los reemplazos.
- Aplicación de la herramienta a otro tipo de documentos del mismo tipo, o a los mismos documentos pero en otra administración distinta.
- Mejora de la gestión y manejo de los PDF.
- Implementación de nuevos componentes en el *pipeline* (introducción de metadatos para volver a las menciones originales, cifrado de los metadatos, etc).