El futuro es de todos

DNP
Departamento
Nacional de Planeación

El futuro es de todos

DNP
Departamento Nacional de Planeación

Acervo Digital: Procesamiento de texto

Unidad de Científicos de Datos
Dirección de Economía Naranja y Desarrollo Digital

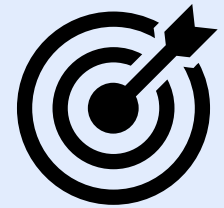
Julio, 2022

Agenda

1. Introducción
2. Metodología
3. Resultados

1. Introducción

Acervo Digital: Procesamiento de Textos

¿Qué?	Implementar algoritmos para la lectura, procesamiento y análisis documentos PDF dentro del marco de acervo documental digital del gobierno.	
¿Para qué?	Para facilitar la indexación, búsqueda y generación de estadísticas de documentos digitales. Con el propósito de contar con una biblioteca de documentos que apoyen la elaboración de informes de gestión y faciliten el empalme con el gobierno entrante.	
Resultados	1. Implementación de scripts para la descarga, lectura, limpieza y procesamiento de documentos PDF. 2. Implementación de API que integró todos los algoritmos de procesamiento y que se conectó con el Datálogo Colombia (https://datalogo.dnp.gov.co/)	

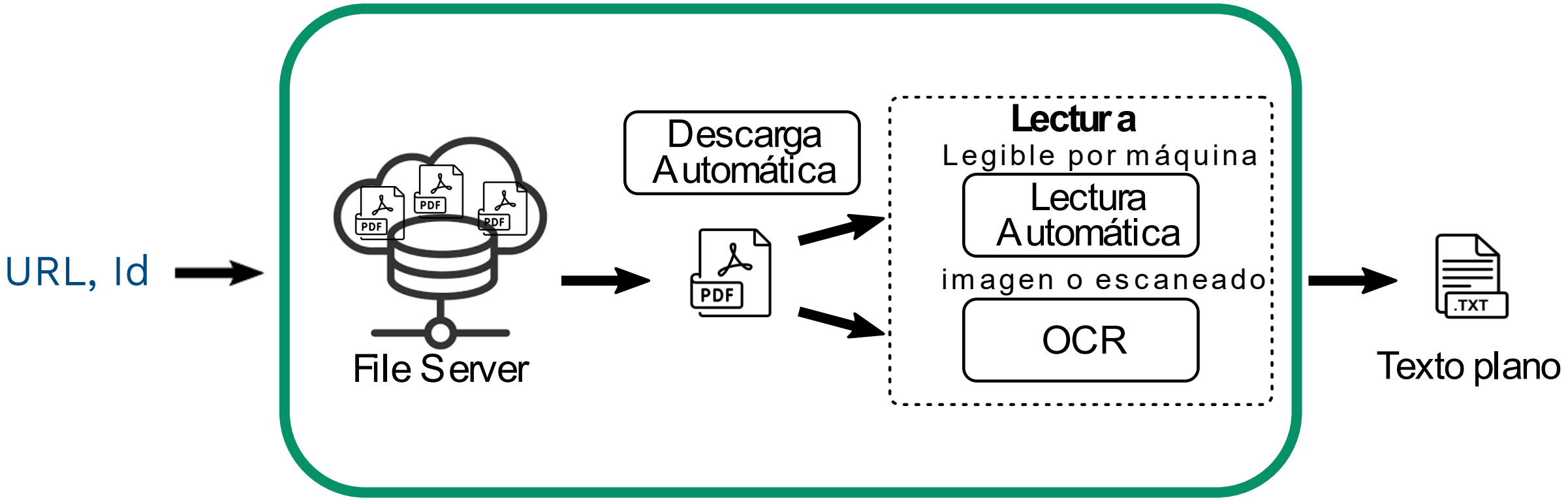
2. Metodología

Acervo Digital: Procesamiento de Textos

Descarga y lectura automática de documentos

Figura 1: Proceso de descarga y lectura de archivos PDF

API – Interfaz de Programación de Aplicaciones



Fuente: Elaboración propia.

Acervo Digital: Procesamiento de Textos

Limpieza básica del texto

- ✓ Texto a Minúsculas
- ✓ Remoción de caracteres especiales
- ✓ Remoción Stopwords.
- ✓ Lematización

Figura 2: Comparación entre el texto plano (a) y el texto plano después del proceso de limpieza básica (b).

```
.^^a^^^^^
IlCt~¡¡AI. ~U!t§~fiG\t\B

libertad yOrden

República de Colombia
Ministerio de Relaciones Exteriores

DECRETO NÚMERO  2082

DE

-7 NOV2018

Por el  cual se hace una designación en  comisión para situaciones especiales
a la planta externa del Ministerio de Relaciones Exteriores

EL PRESIDENTE DE LA REPÚBLICA DE COLOMBIA

En ejercicio de sus facultades constitucionales y legales, en, especial  las
que le confiere

el numeral 2° del artículo 189 de la Constitución Política, articulo 40,
literal b)  del artículo 53

y el artículo 62 del Decreto Ley 274 de 2000, y


CONSIDERANDO

Que  el  doctor  LUIS  CARLOS  RODRíGUEZ  GUTIÉRREZ,  es  funcionario
inscrito  en  el
escalafón de la Carrera Diplomática y Consular, en  la categoría de Ministro
Plenipotenciario,
se encuentra cumpliendo el lapso de alternación en  la planta interna desde.
el 8 de agosto de
2016

[...]
```

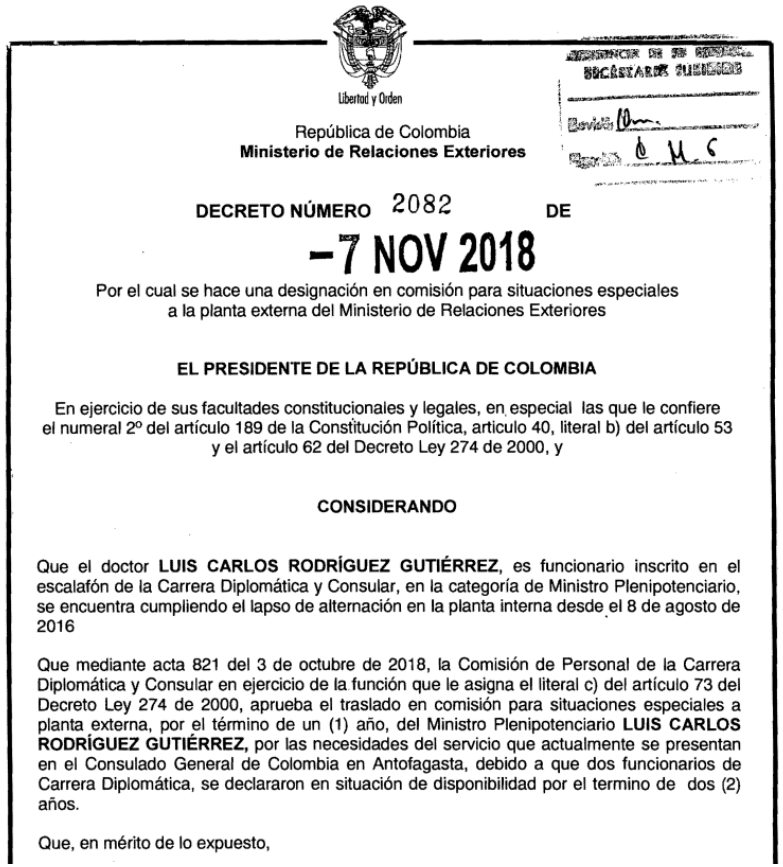
```
ilct ai fig libertad yorden república de colombia ministerio de relaciones
exteriores decreto
número de nov por el cual se hace una designación en comisión para
situaciones especiales la planta externa del ministerio de relaciones
exteriores el presidente de la república de colombia en ejercicio de sus
facultades constitucionales legales en especial las que le confiere el
numeral del artículo de la constitución política articulo literal del
artículo el artículo del decreto ley de considerando que el doctor luis
carlos rodríguez gutiérrez es funcionario inscrito en el escalafón de la
carrera diplomática consular en la categoría de ministro plenipotenciario se
encuentra cumpliendo el lapso de alternación en la planta interna desde el de
agosto de

[...]
```

(a) (b)

Fuente: Elaboración propia.

Acervo Digital: Procesamiento de Textos

Implementación del API



Libertad y Orden

República de Colombia
Ministerio de Relaciones Exteriores

DECRETO NÚMERO 2082 **DE**

-7 NOV 2018

Por el cual se hace una designación en comisión para situaciones especiales a la planta externa del Ministerio de Relaciones Exteriores

EL PRESIDENTE DE LA REPÚBLICA DE COLOMBIA

En ejercicio de sus facultades constitucionales y legales, en especial las que le confiere el numeral 2º del artículo 189 de la Constitución Política, artículo 40, literal b) del artículo 53 y el artículo 62 del Decreto Ley 274 de 2000, y

CONSIDERANDO

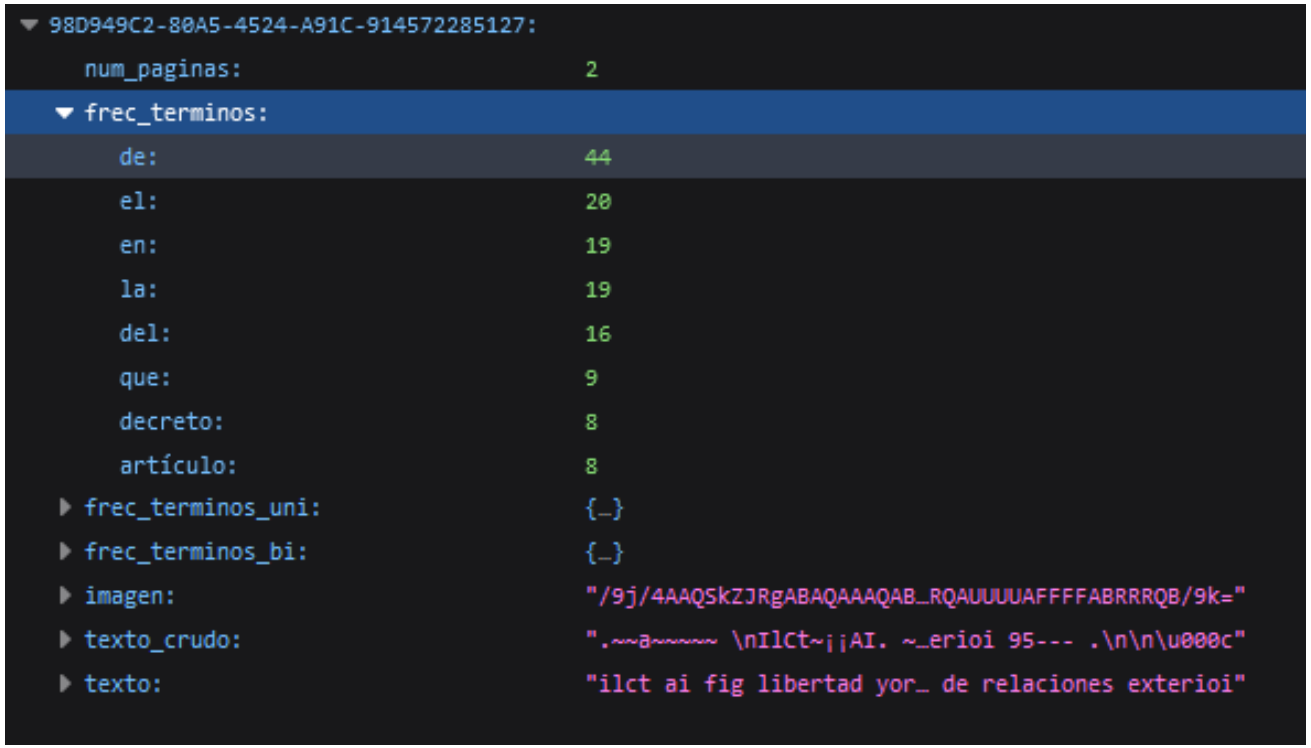
Que el doctor **LUIS CARLOS RODRÍGUEZ GUTIÉRREZ**, es funcionario inscrito en el escalafón de la Carrera Diplomática y Consular, en la categoría de Ministro Plenipotenciario, se encuentra cumpliendo el lapso de alternación en la planta interna desde el 8 de agosto de 2016

Que mediante acta 821 del 3 de octubre de 2018, la Comisión de Personal de la Carrera Diplomática y Consular en ejercicio de la función que le asigna el literal c) del artículo 73 del Decreto Ley 274 de 2000, aprueba el traslado en comisión para situaciones especiales a planta externa, por el término de un (1) año, del Ministro Plenipotenciario **LUIS CARLOS RODRÍGUEZ GUTIÉRREZ,** por las necesidades del servicio que actualmente se presentan en el Consulado General de Colombia en Antofagasta, debido a que dos funcionarios de Carrera Diplomática, se declararon en situación de disponibilidad por el termino de dos (2) años.

Que, en mérito de lo expuesto,

API

Archivo JSON



```
▼ 98D949C2-80A5-4524-A91C-914572285127:
    num_paginas:          2
  ▼ frec_terminos:
      de:                 44
      el:                 20
      en:                 19
      la:                 19
      del:                16
      que:                9
      decreto:            8
      artículo:           8
  ▶ frec_terminos_uni:    {…}
  ▶ frec_terminos_bi:     {…}
  ▶ imagen:               "/9j/4AAQSkZJRgABAQAAAQAB…RQAUUUUAFFFFABRRRQB/9k="
  ▶ texto_crudo:          ".~~a~~~~~ \nIlCt~¡¡AI. ~…erioi 95--- .\n\n\u000c"
  ▶ texto:                "ilct ai fig libertad yor… de relaciones exterioi"
```

Proyecto solicitado por Grupo CONPES

3. Resultados

Resultados

1. Se logró realizar el procesamiento de 2599 documentos de acuerdo con lo solicitado por el Grupo CONPES

1. Se implementó un API con los algoritmos de procesamiento de texto que se integró al Datálogo Colombia (https://datalogo.dnp.gov.co/) para el procesamiento de nuevos documentos de forma automática.

2. El tiempo de procesamiento promedio de un documento PDF es de 12 segundos (y se tiene una desviación estándar de 25 segundos, se tomó como referencia una muestra de 100 documentos. El procesamiento se realizó en un equipo con procesador Intel(R) Xeon(R) W-2145 CPU @ 3.75GHz y 64GB de memoria RAM), vale la pena mencionar que el tiempo de procesamiento dependerá de la calidad del documento PDF, si se requiere aplicar OCR (Reconocimiento Óptico de Caracteres), la longitud del documento y la capacidad de procesamiento del computador en el que se realice el análisis.

El futuro es de todos

DNP
Departamento
Nacional de Planeación