Dirección de Economía Naranja y Desarrollo Digital

Unidad de Científicos de Datos

El futuro es de todos

DNP
Departamento
Nacional de Planeación

ACERVO DIGITAL: Procesamiento de texto

INFORME FINAL

Dependencias y entidades involucradas	Departamento Nacional de Planeación • Dirección de Economía Naranja y Desarrollo Digital - Unidad de Científicos de Datos • Grupo CONPES
Sector	Planeación
Tecnologías utilizadas	Python
Fuentes de datos	File server Acervo Digital

Contenido

1. Presentación

El proyecto de Acervo Digital del Grupo CONPES se enmarca en la necesidad de contar con una biblioteca de documentos digitales en formato PDF que contienen información de los diferentes sectores del gobierno, que apoyen la elaboración de informes de gestión y faciliten el proceso de empalme con el gobierno entrante. Por lo anterior, se hace necesario contar con información estructurada y legible por máquina que permitan tener herramientas de indexación, búsqueda y generación de estadísticas de los documentos digitales. Atendiendo a esta necesidad, la Unidad de Cientificos de Datos desarrolló una metodología de analisis de texto para transformar los documentos digitales en texto plano que pueda ser leído por máquina, para facilitar los procesos de indexación y búsqueda sobre los documentos; y cálculo de rankings de unigramas y rankings de bigramas que puedan dar estadísticas y una visión general de los documentos. Además de la integración de la metodología dentro de un API que pueda ser usado para estructurar información de nuevos documentos digital que no fueron procesados en el marco de este proyecto.

The Acervo Digital project of the CONPES Group is framed in the need to have a library of digital documents in PDF format containing information from different spheres of the government, to support the preparation of management reports and facilitate the process of joining the new government. Therefore, it is necessary to have structured and machine-readable information that allows indexing, search, and statistics generation of digital documents. In response to this need, the Data Scientists Team developed a text analysis methodology to transform digital documents into machine-readable texts, to facilitate the indexing and search processes on the documents; and calculate unigram and bigram rankings that can provide statistics and an overview of the documents. In addition, the methodology was put together within an API that can be used to structure information from new documents.

2. Objetivos del proyecto

2.1. General

Implementar algoritmos para lectura y procesamiento de texto que faciliten el análisis de documentos de acervo digital en formato PDF, y estructurarlos en formatos legibles por máquina e implementar un API para integrar los algoritmos y estos puedan ser usados por el sistema final.

2.2. Específicos

1. Implementar script para descarga automática de documentos en formato PDF almacenados en un servidor de documentos.
2. Implementar scripts de lectura y consolidación automática de archivos en formato PDF y su posterior transformación a archivos leíbles por máquina.
3. Implementar script para el procesamiento y limpieza de los textos planos.
4. Implementar un API que integre todos los algoritmos de lectura y procesamiento de texto para analizar nuevos documentos PDF.

3. Metodología

El desarrollo de este proyecto abarcó 3 etapas: i) Descarga y lectura automática de documentos PDF almacenados en un *file server*, ii) procesamiento de texto e iii) integración de la metodología en un API de consulta.

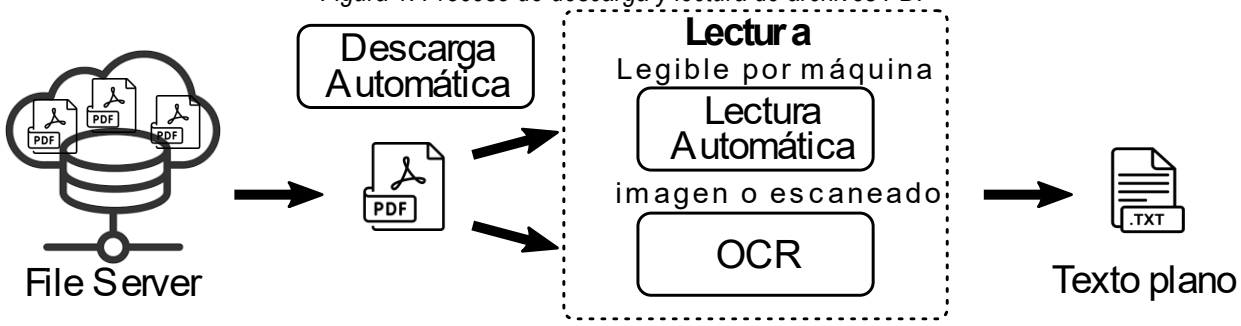
3.1. Descarga y lectura automática de documentos

En la Figura 1 se muestra el proceso general de descarga y lectura de archivos PDF. Primero, una rutina en Python descarga el archivo PDF dada la ruta de almacenamiento en el file server, luego este archivo es pasado por una etapa de lectura en 2 pasos: Si el archivo es legible por máquina, se lee directamente con Python (Esto rescata el contenido fiel al original, es decir no se pierde información o se confunden palabras); Si el archivo no se puede leer[1], pasa por una etapa de reconocimiento de caracteres (OCR, en inglés) pero dependiendo de la calidad de la imagen del

[1] Todos los documentos se asumen legibles por máquina, si después de intentar la lectura se rescatan menos de 12 palabras, el documento pasa por el proceso de OCR.

documento, el contenido puede variar del original (se pueden confundir palabras por similares, signos de puntuación, o no reconocer algunas palabras). Por último, se devuelve el texto plano de cada documento para ser analizado por la etapa de procesamiento de texto.

Figura 1: Proceso de descarga y lectura de archivos PDF



Fuente: Elaboración propia.

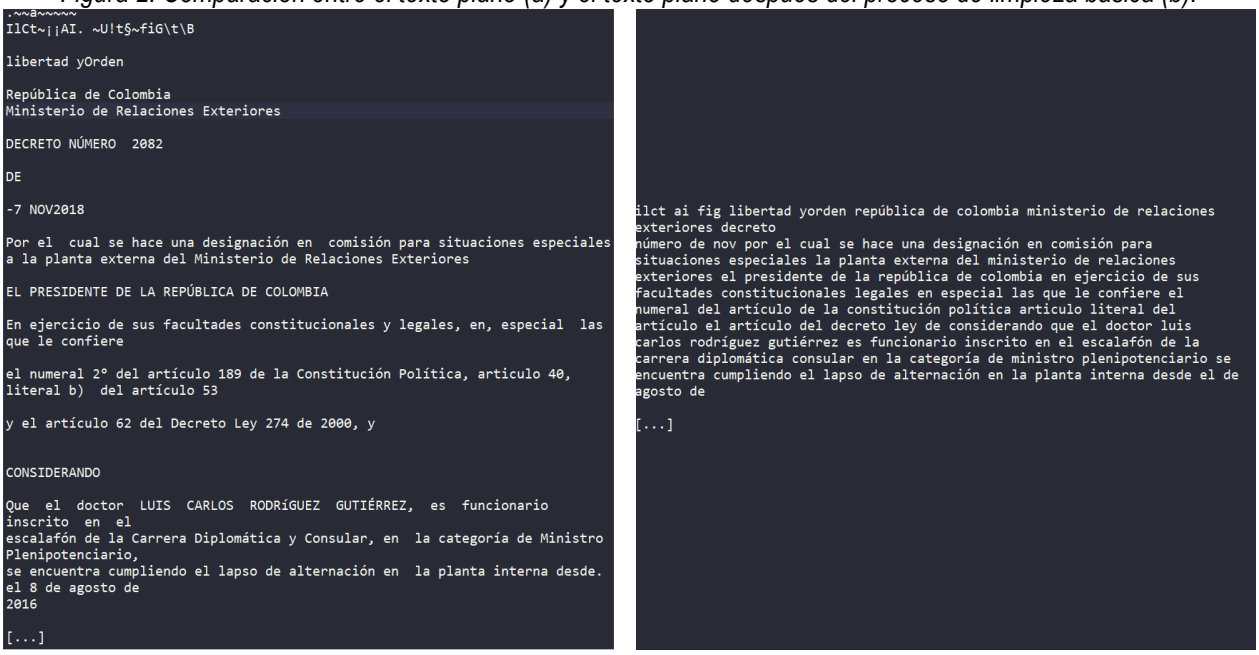
3.2. Procesamiento de texto

En esta etapa se consideran todos los procesos al texto plano, resultado de la primera etapa. De este proceso se obtienen tres salidas diferentes: i) texto plano con limpieza básica de texto, ii) Ranking de unigramas del texto y iii) Ranking de bigramas del texto.

Para la primera salida, la limpieza básica del texto plano pasa por los siguientes pasos:

- Se convierte todo el texto a minúsculas.
- Se remueven todos los espacios múltiples, saltos de línea, números, signos de puntuación y caracteres especiales tales como @, #, entre otros.

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

[...]
```

(b)

```
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

[...]
```

Fuente: Elaboración propia.

Para las dos salidas restantes, el ranking de unigramas y bigramas, el texto plano después de la limpieza básica pasa por los siguientes pasos:

- Se eliminan las stopwords o palabras vacías. Estas hacen referencia a palabras que carecen de sentido cuando se escriben solas. Estas corresponden a conjunciones, artículos, preposiciones y adverbios. Ejemplos de estas palabras son: ante, antes, aún, aunque, aquí, arriba, atrás así, bajo, bastante, cabe, conmigo, bien, casi, cierto, como, debajo, ahí, ajeno, algo, algún, ambos, aquello.
- Luego pasa por una etapa de lematización, que implica, dada una forma flexiva (es decir, plural, femenino, conjugación, etc.), encontrar el lema correspondiente. Un lema es una forma que se acepta convencionalmente como la representación de todas las formas de la misma palabra. Es decir, los lemas de palabras son las entradas que podemos encontrar en los diccionarios tradicionales: sustantivo singular, adjetivo masculino singular, verbo infinitivo. Por ejemplo, decir es el lema de dije, pero también de diré o dijéramos; guapo es el lema de guapas; mesa es el lema de mesas. Este proceso se hace con el fin de reducir las palabras que aparecen en el ranking y que palabras que pertenezcan al mismo lema no se cuente su ocurrencia por separado.
- Para el caso de unigramas, el texto lematizado, se divide en palabras individuales y se hace el conteo de su ocurrencia en todo el texto. Con la información resultante se pueden armar nubes de palabras como la que se muestra en la Figura 3, donde palabras como artículo, decreto y relación son las palabras más frecuentes en el texto.

Figura 3: Nube de palabras por unigramas. Nótese que aparece colombio en lugar de colombia, esto sucede por temas de lematización.

Fuente: elaboración propia.

- Para el caso de los bigramas, el texto lematizado es agrupado en pares de palabras consecutivas, es decir, el texto “república colombio ministerio relación exterior decreto” se convierte en “(república colombio), (colombio ministerio), (ministerio relación), (relación exterior), (exterior decreto)”. Luego, se hace el conteo de apariciones de estas parejas en el texto. Con la información resultante se pueden armar nubes de palabras como la que se muestra en la Figura 4. Esta representación puede ser de mayor utilidad que la de unigramas porque da más información en contexto.

Figura 4: Nube de palabras por bigramas.

Fuente: Elaboración propia.

Después del procesamiento de texto, la información de texto plano con limpieza básica, el ranking de unigramas y el ranking de bigramas se guardan en una estructura *json* para ser almacenada en base de datos. Esta información será usada por el aplicativo de acervo digital que está siendo desarrollada por el equipo técnico del Grupo Conpes.

3.3. Implementación de API

Como etapa final del proyecto se procedió a implementar una API (interfaz de programación de aplicaciones) de tal manera que permita al equipo técnico del Grupo CONPES realizar el procesamiento descrito anteriormente de cualquier documento adicional que sea requerido mediante peticiones *GET*.

Para la implementación se utilizó la librería FastAPI, librería de Python diseñada específicamente para la implementación de APIs en ese lenguaje, y se definieron dos parámetros de entrada, *file_id* el cual corresponde al identificador único que se le desea asignar al archivo y *url* correspondiente al enlace de acceso al archivo de interés en el servidor. Al Grupo CONPES le fue entregado tanto el código fuente del API al igual que los archivos de procesamiento y de configuración para hacer el despliegue en el servidor usando Docker.

Para verificar el correcto funcionamiento de los scripts y el diseño del API se realizó el despliegue de estos en el servidor de la UCD, disponible únicamente desde la intranet del DNP. El API se encuentra disponible para pruebas mediante el enlace http://vdatascience:8043/, se puede acceder a la documentación de esta y realizar pruebas usando la ruta *docs* http://vdatascience:8043/docs y para hacer una petición de procesamiento sería mediante la ruta *get-info*.

Al hacer peticiones al API este retornará el JSON resultado del procesamiento y la respuesta tendrá el código 200 de "*Successful Response*". En caso de presentarse un error en el procesamiento la respuesta tendrá el código 500 de "*Error: Internal Server Error"* y se obtendrá el siguiente JSON "*{"detail":"Se presentó un error de procesamiento"}*", en este último caso se deberá verificar manualmente cuál es la causa del error.

A continuación, se presenta un ejemplo con los valores *prueba* y *https://jairoruizsaenz.github.io/pruebas_acervo/prueba.pdf* para los parámetros *file_id* y *enlace* respectivamente.

http://vdatascience:8043/get-info?file_id=prueba&url=https://jairoruizsaenz.github.io/pruebas_acervo/prueba.pdf

4. Resultados

A través del desarrollo metodológico descrito en la Sección 3, se obtuvieron los resultados que se presentan a continuación. Toda retroalimentación desde un punto de vista experto o de usuario por parte del grupo CONPES es bienvenida. Este insumo será de gran ayuda para mejorar la calidad y utilidad de los resultados obtenidos, de manera que agreguen mayor valor.

1. Se logró realizar el procesamiento de 2599 documentos de acuerdo con lo solicitado por el Grupo CONPES
2. Se logró implementar scripts de procesamiento y realizar la implementación de un API para acceder a estos mediante el enlace http://vdatascience:8043/
3. 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.

5. Socialización

Este proyecto se socializó con la Unidad de Cientificos de Datos, la directora y asesores de la Dirección de Economía Naranja y Desarrollo Digital, y el equipo técnico del Grupo CONPES.

Contacto

Si tiene alguna duda, comentario o sugerencia sobre este proyecto, o si le gustaría conversar con la Unidad de Científicos de Datos sobre la posibilidad de una nueva fase para el mismo, puede comunicarse con nosotros a través del correo electrónico ucd@dnp.gov.co.