**PDF**

PyPDF2 es un módulo de Python que podemos usar para extraer la información de un documento PDF, fusionar documentos, dividir un documento, recortar páginas, cifrar o descifrar un archivo PDF, y más.

Abrimos el documento PDF en modo lectura binario usando open ('document\_path.PDF', 'rb'). PDFFileReader() se utiliza para crear un objeto lector de PDF para leer el documento. Podemos extraer texto de las páginas del documento PDF usando los métodos getPage() y extractText(). Para obtener el número de páginas en el documento PDF dado, usamos.numPages.

**DOCX**

Es necesario instalar el paquete python-docx, aunque existe un paquete que se llama docx este último es una versión legacy de python-docx.

from docx import Document

**Especifica la ruta del archivo .docx que deseas abrir**

archivo\_docx = "mi\_archivo.docx"

**Abre el archivo .docx**

documento = Document(archivo\_docx)

**Lee y muestra el contenido del archivo**

for paragraph in documento.paragraphs:

print(paragraph.text)

Prueba:

import PyPDF2

from docx import Document

archivo\_pdf = "mi\_archivo.pdf"

archivo\_docx = "mi\_archivo.docx"

with open(archivo\_pdf, 'rb') as pdf\_file, Document(archivo\_docx) as doc:

pdf\_reader = PyPDF2.PdfFileReader(pdf\_file)

for page\_num in range(pdf\_reader.numPages):

print(f"Contenido del PDF - Página {page\_num + 1}:\n{pdf\_reader.getPage(page\_num).extractText()}")

print("\nContenido del DOCX:")

for paragraph in doc.paragraphs:

print(paragraph.text)