



UNIVERSIDAD
NACIONAL
DE COLOMBIA

Universidad Nacional de Colombia - sede Bogotá
Facultad de
Ingeniería Departamento de Sistemas e
Industrial Curso: Ingeniería de Software 1
(2016701)

EXTRAER TEXTO DESDE UN ARCHIVO PDF

ACTORES

Sistema



REQUERIMIENTO

RF_010 - Extracción de texto de PDF: El sistema debe extraer únicamente el contenido textual de los archivos PDF, ignorando imágenes y elementos gráficos.






DESCRIPCIÓN

Se extrae únicamente el contenido en texto del archivo PDF cargado previamente, dejando de un lado imágenes y gráficos. El texto se muestra al usuario para validar el envío de este a la IA.


PRECONDICIONES

- El usuario debe haber cargado un archivo PDF válido.

FLUJO NORMAL

1. El sistema accede al archivo PDF previamente cargado y validado.
2. El sistema ejecuta el proceso de lectura del documento  ignorando imágenes y gráficos.
3. El sistema verifica que:
 -  Si el archivo PDF tiene solamente imágenes sin texto legible se muestra un mensaje “El documento no contiene texto extraíble” → se retorna al panel para cargar un nuevo archivo.
 -  Si ocurre algún error técnico durante la lectura del archivo PDF, se muestra un mensaje de error “No se pudo procesar el archivo, intente nuevamente”. → se vuelve al paso 1.
4. Si resultó texto extraíble del archivo PDF el sistema muestra al usuario el texto extraído .
5. El usuario revisa el contenido para su consentimiento de envío a la IA , ó puede cancelar el proceso de extracción de texto volviendo al paso 1.

POSTCONDICIONES

- El texto queda disponible para el módulo de generación automática de preguntas con uso de IA 

NOTAS

- Solo se procesa texto plano, sin utilizar reconocimiento óptico de caracteres.
- El sistema maneja la aparición de error en el proceso de extracción, como por la mal codificación de archivos.
- El usuario confirma si el texto extraído corresponde al contenido del archivo PDF cargado