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

ASIGNAR PALABRAS CLAVE SEGÚN EL CONTENIDO DEL ARCHIVO	
ACTORES Sistema	**REQUERIMIENTO** **RF_21** El sistema debe ser capaz de leer el contenido interno de ciertos tipos de archivo (TXT, PDF, DOCX, archivos de texto simple) y obtener palabras clave del contenido.

DESCRIPCIÓN

El sistema identificará palabras clave de los archivos de texto a partir de su contenido según la frecuencia de las palabras en este, ignorando palabras conectoras y artículos.

PRECONDICIONES

- El sistema debe estar haciendo la indexación de un archivo e identificarlo como uno de texto simple según su extensión, e inicia a hacer el proceso descrito en este caso de uso después de haber terminado de realizar los demás pasos de la indexación de dicho archivo.
- Si al intentar indexar el archivo no logra completar el proceso, no iniciará con el flujo del caso de uso actual.

FLUJO NORMAL

1. El sistema abre el archivo.
  1.1. El sistema no puede abrir el archivo.
  1.2. El sistema termina el proceso de identificación de palabras clave.
2. El sistema lee el texto que se encuentra en el archivo.
3. El sistema calcula la frecuencia con la que aparece cada palabra.
4. El sistema toma las 3 palabras más frecuentes en el archivo y las asigna como palabras clave de este.
  4.1. A. Si hay 3 o menos palabras distintas en el archivo, toma solo la más frecuente como palabra clave.
  4.1. B. Si hay varias palabras que empaten en frecuencia en el tercer lugar, el sistema toma todas las que empataron en este puesto como palabras clave.
5. {Añadir palabras clave al índice}

POSTCONDICIONES

- El sistema debe seguir con la indexación de los demás archivos que estén en la fila de indexación.

NOTAS

RNP_2 Cada archivo puede tener múltiples palabras clave, pero ninguna de ellas debe repetirse dentro del mismo archivo.
RNP_6 No todos los archivos deben tener una categoría, una etiqueta o palabras claves, por lo tanto, si se elimina una categoría, no es necesario que los archivos en esta se reasignen a otras categorías.
RNP_22 Las palabras clave no pueden contener espacios; deben ser un token continuo de 1–50 caracteres válidos (a–z, 0–9, _, -). Para conceptos de varias palabras, usar etiquetas.

Se debe definir una lista de palabras que el sistema debe ignorar al usar este proceso, estas palabras incluyen conectores, artículos, y demás palabras que no tengan una función semántica por sí mismas o que sean muy frecuentes en el lenguaje (inicialmente solo para el español).