INSTITUTO TECNOLÓGICO DE COSTA RICA **II SEMESTRE 2018**
INGENIERÍA EN COMPUTACIÓN
IC-8020 ELECTIVA TEXTO

TAREA PROGRAMADA 2
GUÍA DE DOCUMENTACIÓN

1. Introducción

   En esta tarea programada por medio de herramientas tales como Lucene, Snowball y Java, se realizrán búsquedas en archivos HTML que se encuentran en un arhcivo de texto y son leidas por el programas lo mas rápidas posible, a través de las herramientas mencionadas anteriormente. Por medio de estas bpusquedas se pueden encontrar los textos que son abiertos finalmente a través de HTML para mostrar el arhcivo de texto que fue indexado.

2. Completar la siguiente tabla para describir el estado en que quedó la indización de la colección.

Etapa	% de complet.	Comentario o aclaración
INDIZACIÓN		
Campo "texto"		
Extrae adecuadamente del elemento <body>	**100%**	
Separación en palabras (letras incluyendo eñe)	**100%**	
Eliminación de stopwords	**100%**	
Extracción de raíces (stemming)	**100%**	
Eliminación de acentos, preserva eñe	**100%**	
Campo "ref"		
Extrae adecuadamente del elemento <a>	**100%**	
Separación en palabras (letras incluyendo eñe)	**100%**	
Conversión a minúsculas	**100%**	
NO hay extracción de raíces (stemming)	**100%**	
Eliminación de acentos, preserva eñe	**100%**	
Campo "encab"		
Extrae adecuadamente del elemento <h?>	**100%**	
Separación en palabras (letras incluyendo eñe)	**100%**	
Eliminación de stopwords	**100%**	
Extracción de raíces (stemming)	**100%**	
Eliminación de acentos, preserva eñe	**100%**	
Campo "titulo"		
Extrae adecuadamente del elemento <title>	**100%**	
Separación en palabras (letras incluyendo eñe)	**100%**	
Conversión a minúsculas	**100%**	
NO hay extracción de raíces (stemming)	**100%**	
Eliminación de acentos, preserva eñe	**100%**	
Indexado de la colección (h8, h7, ---, h1, h0)		
Colección más grande que puede indexar	**100%**	
¿En 5 minutos o menos?	**100%**	H8, H7, H6, H5
¿En 30 minutos o menos?	**100%**	H4
¿En una hora o menos?	**100%**	
¿En más de una hora? Dar tiempo.	**100%**	
Consultas		
Permite usar lenguaje de consultas de Lucene	**100%**	

3. Completar la siguiente tabla para describir el resultado obtenido para algunas consultas de prueba.

Consulta	Colección	Resultado esperado	Obervaciones sobre el resultado obtenido
`magnoel`	h8	Éxito	Éxito
`titulo:magnoel`	h8	Éxito	Éxito
`encab:magnoel`	h8	Éxito	Éxito
`cartago`	h8	Fallo	Fallo
`cartago`	h7	Éxito	Éxito
`alejandro AND magno`	h8	Fallo	Éxito debido a q ue hace stemming en magno
`alejandro AND magno`	h7	Éxito	Éxito
`ref:alejandro AND ref:magno`	h7	Éxito	**Éxito**
`amintas`	h7	Fallo	Fallo
`amintas`	h6	Éxito	Éxito
`tenistas de macedonia`	h6	Fallo	Éxito
`tenistas de macedonia`	h5	Éxito	Éxito
`ref:"tenistas de macedonia"`	h5	Éxito	**Éxito**
`turrialba`	h5	Fallo	Fallo
`turrialba`	h4	Éxito	Éxito

Éxito: encontró al menos un documento en la colección especificada.
Fallo: no encontró ningún documento en la colección especificada.

Tiempos:

H8: 22s

H7:43s

H6: 1:37

H5: 3:45

H4: 7:37

4. Comentarios finales (estado del programa)

   La tarea programada se encuentra en un estado de un 100%, en el que todas las funcionalidades y herramientas necesarias fueron establecidas e implementadas. Por lo tanto las búsquedas son realizadas de forma correcta. Se revisaron los resultados que se obtuvieron y se encuentran en estado esperado.