

TAREA PROGRAMADA 2
GUÍA DE DOCUMENTACIÓN

1. Introducción

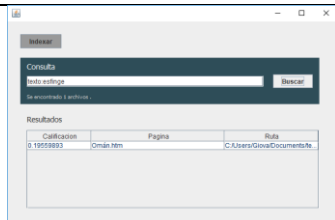
La tarea consiste en programar una aplicación que permita indexar una colección de páginas web geográficas así como hacer consultas a dicha colección. Se usará la biblioteca [Lucene de Apache](#) tanto para crear el índice de la colección como para realizar las búsquedas. Completar la siguiente tabla para describir el estado en que quedó la indización de la colección.

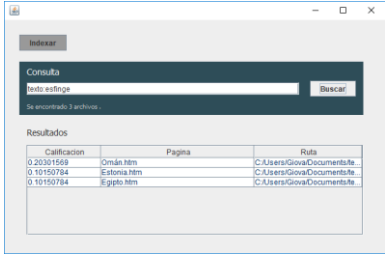
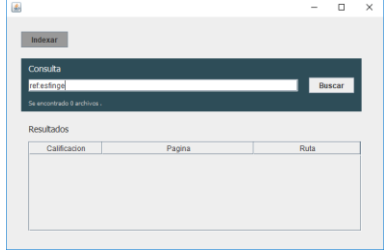
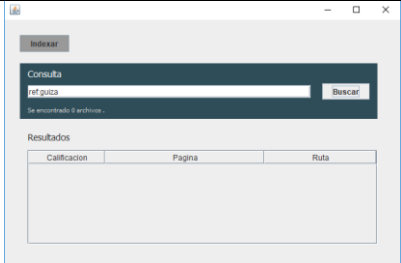
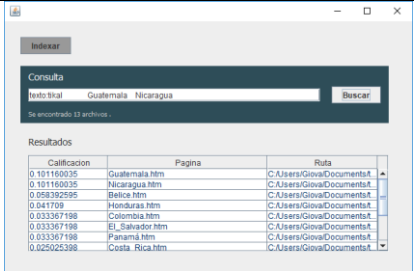
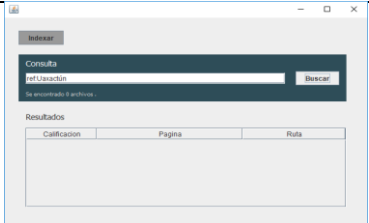
Para la solución se implementará el indexado con python y lucene.

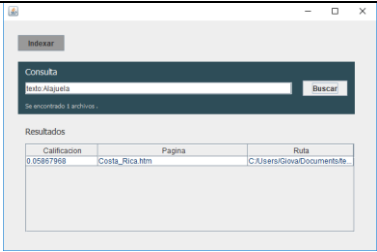
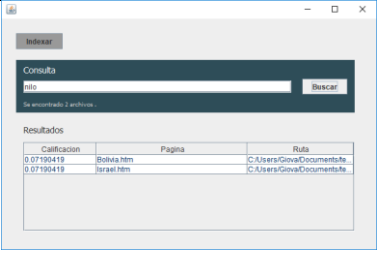
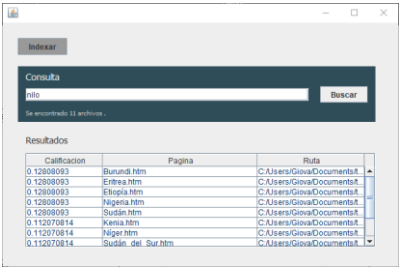
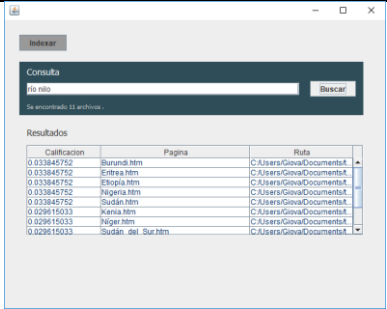
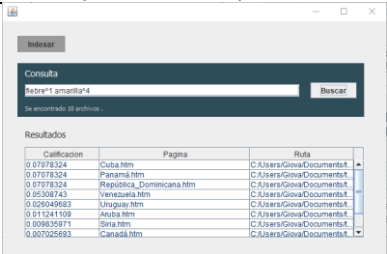
2.

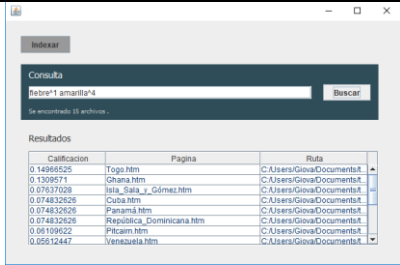
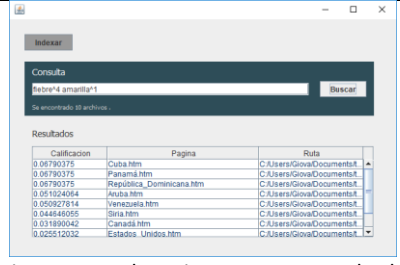
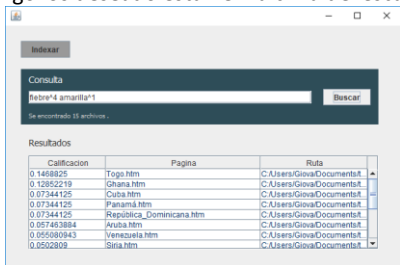
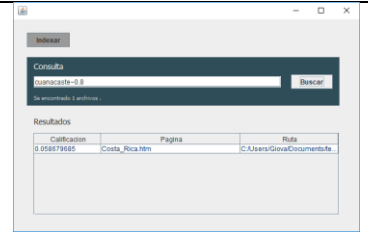
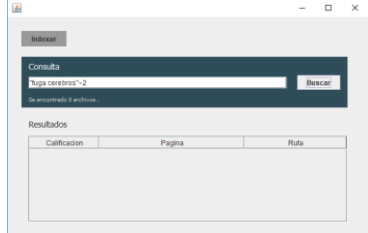
Etapa	% de complet.	Comentario o aclaración
INDIZACIÓN		
Campo "texto"		
Extrae adecuadamente del elemento <p>	100	
Extrae adecuadamente del elemento <a>	100	
Separación en palabras (letras incluyendo eñe)	100	
Eliminación de stopwords	100	
Extracción de raíces (stemming)	100	
Eliminación de acentos, preservación eñe	100	
Campo "ref"		
Extrae adecuadamente del elemento <a>	100	
Separación en palabras (letras incluyendo eñe)	100	
Conversión a minúsculas	100	
Eliminación de acentos (menos eñe=	100	
Indexado de la colección		
Indexa correctamente América y Asia	100	
Agrega correctamente África, Europa y Oceanía	100	
Consultas		
Permite usar lenguaje de consultas de Lucene	100	

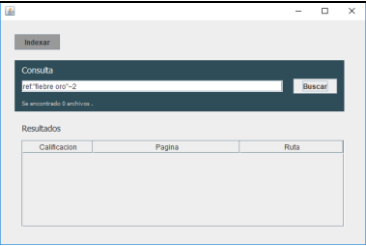
3. Completar la siguiente tabla para describir el resultado obtenido para algunas consultas de prueba. Los casos subrayados son de la segunda fase de indización (África, Asia y Oceanía) y no deben aparecer cuando se corren las consultas en la primera fase de indización (América y Asia).

Consulta	Resultado esperado	Comentarios sobre resultado obtenido
texto:esfinge	Omán <u>Egipto</u> <u>Estonia</u>	

Consulta	Resultado esperado	Comentarios sobre resultado obtenido
		<p>Primera consulta exitosa con america y asia indexados al solo mostrar omán</p>  <p>La segunda consulta devuelve los nuevos paises en el escalafón.</p>
ref:esfinge	<u>Egipto</u> <u>Estonia</u>	 <p>Resultado esperado en la primera consulta, sin elementos en el escalafón.</p>
ref:guiza	<u>Egipto</u>	 <p>Resultado esperado en la primera consulta, sin elementos en el escalafón. Segunda consulta exitosa</p>
texto:tikal	Guatemala Nicaragua	 <p>Los documentos esperados aparecen en los primeros puestos del escalafón.</p>
ref:Uaxactún	Guatemala.htm	 <p>Falla la consulta</p>

Consulta	Resultado esperado	Comentarios sobre resultado obtenido
texto:Alajuela	Costa_Rica Panama	 <p>La consulta sólo incluye a Costa Rica dentro del escalafón</p>
nilo	Bolivia Kenia Israel Nigeria Jordania Níger Burundi Sudán Egipto Sudán del Sur Etiopía	 <p>En la primera consulta devuelve dos de los 3 documentos deseados.</p>  <p>La segunda consulta si incluye los nuevos países en el escalafón.</p>
"río nilo"	Egipto Sudán	 <p>Muestra los resultados esperados para la segunda consulta pero no en el tope del escalafón</p>
fiebre^1 amarilla^4	Cuba Venezuela Panamá Ghana República_Dominicana Togo	 <p>La primera consulta devuelve los países deseados.</p>

Consulta	Resultado esperado	Comentarios sobre resultado obtenido
		 <p>La segunda consulta resulta exitosa al agregar los nuevos documentos al escalafón.</p> <p>Muestra todos los documentos esperados, al aumentar el índice muestra exitosamente los nuevo documentos dentro del escalafón.</p>
fiebre^4 amarilla^1	Cuba Panamá República_Dominicana Venezuela <u>Ghana</u> <u>Togo</u>	 <p>Primera consulta exitosa, muestra todos los documentos deseado más otros documentos, sin embargo los deseado están en la cima del escalafón.</p>  <p>La segunda consulta muestra exitosamente los nuevos documentos en el escalafón.</p>
cuanacaste~0.8	Costa_Rica	 <p>Consulta exitosa.</p>
"fuga cerebros"~2	Argentina Colombia Haití <u>Nueva_Zelanda</u>	 <p>Consulta fallida</p>

Consulta	Resultado esperado	Comentarios sobre resultado obtenido
ref:"fiebre oro"~2	Aruba Panamá Canadá <u>Australia</u> Estados_Unidos <u>Sudáfrica</u>	 <p>Consulta fallida</p>

4. Comentarios finales (estado del programa)

El programa funciona correctamente en la parte del indexado incremental, funciona con en dos partes, una parte en python donde procesa los textos en html y genera un JSON para luego indexarlos desde el segundo programa en java.

Desafortunadamente ciertas consultas con caracteres especiales generan problemas con las consultas, al no extraer solo los datos requeridos para la consulta. Sin embargo todas las demas funciones se encuentran al 100% desarrolladas.

5. Directorio con documentación

La tarea debe ser entregada presentando un archivo comprimido (ZIP, RAR, TGZ) que incluya la siguiente estructura:

TP1-ApellidosNombre1-ApellidosNombre2 (directorio)
Archivo_de_documentacion
Programas
Archivos_adicionales
Pruebas

8. Entrega

Enviar la tarea y su documentacion a la direccion josee.arayamonge@gmail.com.