**INSTITUTO TECNOLÓGICO DE COSTA RICA I SEMESTRE 2020**

**INGENIERÍA EN COMPUTACIÓN**

**IC-8020 ELECTIVA TEXTO**

**TAREA PROGRAMADA 2**

**GUÍA DE DOCUMENTACIÓN**

1. Introducción

Presentar brevemente el problema. Puede “reciclar” partes del enunciado de la tarea programada.

Se necesita un programa para indexar y buscar una colección de páginas web. Además usar la indexación en campos independientes para usarlos en las búsquedas, y también actualizar el índice al incluir nuevos documentos. Con cuatro colecciones que contienen archivos con un subconjunto de las páginas HTML de Wikipedia se puede hacer la depuración del sistema, pero el verdadero objetivo es indexar una colección llamado “wiki-g” que es mucho más grande. Para la indexación se debe separar el texto por etiquetas y limpiarlas de forma específica. Debe tener una interfaz que permita escoger un índice, leer consultas e invocar facilidades de Lucene. Debe mostrar los títulos de los mejores resultados las búsquedas. Además debe poder actualizar el índice sin comenzarlo desde el inicio.

1. 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) |  |  |
| Eliminación de stopwords | 100 |  |
| Extracción de raíces (stemming) |  |  |
| Eliminación de acentos, preserva eñe |  |  |
| **Campo "ref"** |  |  |
| Extrae adecuadamente del elemento <a> | 100 |  |
| Separación en palabras (letras incluyendo eñe) |  |  |
| Conversión a minúsculas |  |  |
| **NO** hay extracción de raíces (stemming) |  |  |
| Eliminación de acentos, preserva eñe |  |  |
| **Campo "encab"** |  |  |
| Extrae adecuadamente del elemento <h?> | 100 |  |
| Separación en palabras (letras incluyendo eñe) |  |  |
| Eliminación de stopwords | 100 |  |
| Extracción de raíces (stemming) |  |  |
| Eliminación de acentos, preserva eñe |  |  |
| **Campo "titulo"** |  |  |
| Extrae adecuadamente del elemento <title> | 100 |  |
| Separación en palabras (letras incluyendo eñe) |  |  |
| Conversión a minúsculas |  |  |
| **NO** hay extracción de raíces (stemming) |  |  |
| Eliminación de acentos, preserva eñe |  |  |
|  |  |  |
| **Indexado de la colecciones (wiki-[pg][12])** |  |  |
| ¿Cuánto tarda en indexar wiki-p1.txt? |  |  |
| ¿Cuánto tarda en indexar wiki-p2.txt? |  |  |
| ¿Cuánto tarda en actualizar índice wiki-p1.txt? |  |  |
| ¿Cuánto tarda en indexar wiki-g1.txt? |  |  |
| ¿Cuánto tarda en indexar wiki-g2.txt? |  |  |
| ¿Cuánto tarda en actualizar índice wiki-g1.txt? |  |  |
| **Consultas** |  |  |
| Permite usar lenguaje de consultas de Lucene |  |  |

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

| **Consulta** | **Colección** | **Resultado esperado** | **Obervaciones sobre el resultado obtenido** |
| --- | --- | --- | --- |
| magnoel | wiki-p1.txt | Éxito |  |
| titulo:magnoel | wiki-p1.txt | Éxito |  |
| encab:magnoel | wiki-p1.txt | Éxito |  |
| cartago | wiki-p1.txt | Fallo |  |
| cartago | wiki-g1.txt  wiki-g2.txt | Éxito |  |
| alejandro AND magno | wiki-p1.txt  wiki-p2.txt  wiki-g2.txt | Fallo |  |
| alejandro AND magno | wiki-g1.txt | Éxito |  |
| ref:alejandro AND ref:magno | wiki-g1.txt | Éxito |  |
| amintas | wiki-p1.txt  wiki-p2.txt  wiki-g2.txt | Fallo |  |
| amintas | wiki-g1.txt | Éxito |  |
| tenistas de macedonia | wiki-p1.txt  wiki-p2.txt  wiki-g2.txt | Fallo |  |
| tenistas de macedonia | wiki-g1.txt | Éxito |  |
| ref:"tenistas de macedonia" | wiki-g1.txt | Éxito |  |
| turrialba | wiki-p1.txt  wiki-p2.txt  wiki-g2.txt  wiki-g1.txt | Fallo |  |

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

1. Describa brevemente cómo usó las facilidades de Lucene y de otras herramientas para realizar las siguientes tareas:

* Indexar <body>.
* Extraer y mostrar el contenido de una página.

1. Comentarios finales (estado del programa)

Indicar el estado final en que quedó el programa, problemas encontrados y limitaciones adicionales.

1. Directorio con documentación

La tarea debe ser entregada presentando un archivo comprimido ZIP que incluya la siguiente estructura:  
 TP2-*ApellidosNombre1-ApellidosNombre2*  (directorio)  
 D*ocumentación*  
 *Programas*  
 *Archivos\_adicionales*

*Pruebas*

**¡NO INCLUIR EN LA DOCUMENTACIÓN LOS ÍNDICES GENERADOS!  
TRAER LOS ÍNDICES GENERADOS A LAS REVISIONES.**

1. Entrega

**Enviar la tarea y su documentacion a la direccion** [**josee.arayamonge@gmail.com**](mailto:josee.arayamonge@gmail.com)**.  
Usar asunto: "** TP2-*ApellidosNombre1-ApellidosNombre2***"**