**INSTITUTO TECNOLÓGICO DE COSTA RICA II SEMESTRE 2017**

**INGENIERÍA EN COMPUTACIÓN**

**IC-8020 ELECTIVA TEXTO**

**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](https://lucene.apache.org/) tanto para crear el índice de la colección como para realizar las búsquedasCompletar 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.

|  |  |  |
| --- | --- | --- |
| **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 |  |

1. 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 Egipto Estonia |  | Primera consulta exitosa con america y asia indexados al solo mostrar omán    La segunda consulta devuelve los nuevos paises en el escalafón. |
| ref:esfinge | Egipto Estonia |  | Resultado esperado en la primera consulta, sin elementos en el escalafón. |
| ref:guiza | Egipto |  | Resultado esperado en la primera consulta, sin elementos en el escalafón.  Segunda consulta exitosa |
| texto:tikal | Guatemala Nicaragua |  | Los documentos esperados aparecen en los primeros puestos del escalafón. |
| ref:Uaxactún | Guatemala.htm |  | Falla la consulta |
| texto:Alajuela | Costa\_Rica Panama |  | La consulta sólo incluye a Costa Rica dentro del escalafón |
| **nilo** | |  |  | | --- | --- | | Bolivia  Israel  Jordania  Burundi  Egipto  Etiopía | Kenia  Nigeria  Níger  Sudán  Sudán\_del\_Sur | |  | En la primera consulta devuelv dos de los 3 documetnos deseados.    La segunda consulta si incluye los nuevos países en el escalafón. |
| **"río nilo"** | Egipto Sudán |  | Muestra los resultados esperados para la segunda consulta pero no en el tope del escalafón |
| **fiebre^1 amarilla^4** | |  |  | | --- | --- | | Cuba  Panamá  República\_Dominicana | Venezuela  Ghana  Togo | |  | La primera consulta devuelve los paíse deseados.    La segunda consulta resulta exitosa al agregar los nuevos documentos al escalafón.  Muestra todos los documentos esperados, al aumentar el indice muestra exitosamente los nuevo documentos dentro del escalafón. |
| **fiebre^4 amarilla^1** | |  |  | | --- | --- | | Cuba  Panamá  República\_Dominicana | Venezuela  Ghana  Togo | |  | Primera consulta exitosa, muestra todos los documentos deseado más otros documentos, sin embargo los deseado están en la cima del escalafón.    La segunda consulta muestra exitosamente los nuevos documentos en el escalafón. |
| **cuanacaste~0.8** | Costa\_Rica |  | Consulta exitosa. |
| **"fuga cerebros"~2** | |  |  | | --- | --- | | Argentina  Colombia | Haití  Nueva\_Zelanda | |  | Consulta fallida |
| **ref:"fiebre oro"~2** | |  |  | | --- | --- | | Aruba  Canadá  Estados\_Unidos | Panamá  Australia  Sudáfrica | |  | Consulta fallida |

1. 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.

1. 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*

1. Entrega

**Enviar la tarea y su documentacion a la direccion** [**josee.arayamonge@gmail.com**](mailto:josee.arayamonge@gmail.com)**.**