Tecnológico de Costa Rica

Escuela de Ingeniería en Computación

Recuperación de Información Textual
Primera tarea programada

Profesor: José Enrique Araya Monge

Estudiantes:
Joel Barrantes Garro
Josué Villalobos Schmidt

Miércoles 28 de Septiembre
II Semestre del 2016

Introducción

Este documento presenta los principales objetivos acerca del desarrollo de esta tarea programada, los resultados de la tarea, el grado de completitud y documentación de las pruebas realizadas a la tarea programada.

La tarea en cuestión consta de tres partes:

- Implementar la creación de archivos invertidos.
- Implementar las búsquedas vectoriales usando el archivo invertido.
- Implementar búsquedas estructuradas para refinar el resultado de las búsquedas vectoriales.

Estos objetivos se desarrollan en base a una colección dada con anterioridad. Esta colección está compuesta por archivos en formato XML que contienen información acerca de distintas especies de árboles y plantas.

Tabla de completitud de las distintas funcionalidades de la tarea programada

Etapa	**% de complet.**	**Comentario o aclaración**
Indexa toda la colección	100	
Transformaciones implementadas	80	
Indexa palabras (conversión a minúsculas, eliminación de acentos, preservaión de eñe)	100	
Usa lista de stopwords indicada	100	
Indexa números	100	
Indexa frases pedidas	0	No indexa las frases, las separa en palabras(taxon_name)
Estructura archivo invertido	100	
Archivo documentos (dar # total docs.)	100	
Diccionario (dar # total términos)	100	
Postings (tamaño fijo, dar # total de entradas)	100	
Implementa consultas vectoriales	70	Las consultas por frase solo funciona por si solas, no funcionan con otros terrminos ni con otras frases
Permite términos simples	100	
Permite ajustes +/- en consultas	100	
Permite prefijos (*)	100	

Permite frases (“)	50	Las consultas por frase solo funciona por si solas, no funcionan con otros terrminos ni con otras frases
Crea archivo escalafón correctamente	100	
Implementa consultas avanzadas	0	No se implemento
Procesa una cláusula: <entidad> <caracter> <valor>	0	No se implemento
Procesa más de una cláusula (AND)	0	No se implemento
Crea archivo escalafón correctamente	0	No se implemento
Documento salida	100	
Genera archivo HTML de salida	100	
Incluye todos los datos pedidos	100	
Interfaz Indicar estado de la implementación de los siguientes aspectos:	80	No es posible hacer consulta estructurada
Indexar una colección	100	
Realizar consulta vectorial	100	
Realizar consulta estructurada	0	No es posible hacer consulta estructurada
Mostrar resultados	100	

Pruebas de consultas

Pruebas	Comentarios sobre resultados obtenidos vs resultados esperados
Prueba 1	La consulta devuelve 4 archivos(n. 201, 178, 67, 240) al igual que la consulta de prueba
Prueba 2	La consulta devuelve 15 archivos, que son los mismos de la consulta de prueba excepto 2 archivos, que fueron incluidos en la consulta de prueba porque el termino es un substring de otro termino en esos documentos excluidos
Prueba 3	Recupera 18 archivos, por la misma razon de la consulta anterior. Se comporta de la misma manera que los resultados esperados.
Prueba 4	Retorna 32 archivos, que son los mismos que los resultados esperados
Prueba 5	Retorna 1 archivo(doc n. 6) al igual que la consulta de prueba
Prueba 6	Retorna los 2 archivos esperados(n. 178, 683)
Prueba 7	**No se puede realizar esta consulta(Referirse al apartado de completitud)**
Prueba 8	Recupera 10 archivos, mismo resultado que los documentos esperados
Prueba 9	**No se puede realizar esta consulta(Referirse al apartado de completitud)**
Prueba 10	**No se puede realizar esta consulta(Referirse al apartado de completitud)**

Comentarios finales

El estado final de esta tarea programada es: **incompleto.**

Si fue posible indexar la colección, sin embargo no se pudo indexar las frases solicitadas. El archivo de postings es de tamaño fijo, y para leer de este archivo no es necesario cargar todo el archivo a memoria, basta con darle la dirección del archivo en disco, el desplazamiento dentro del archivo y la cantidad de bytes que se deben traer del archivo. El diccionario contiene todas las palabras encontradas en los archivos y el archivo de documentos contiene la dirección en disco de estos archivos.

La parte de consulta vectorial fue implementada siguiendo el formato recomendado, excepto que no es posible realizar una consulta por frase junto con otro tipo de consultas. Esto se debió a falta de tiempo para implementar esta característica.

La parte de consulta estructurada no fue implementada del todo por falta de planificación en el tiempo requerido.

Los documentos de salida son generados exitosamente con el formato requerido. La tarea programada no cuenta con una interfaz gráfica. Es necesario usar comandos para poder acceder a las funciones del programa. No obstante, si se puede mostrar el escalafón de manera gráfica.