**INSTITUTO TECNOLÓGICO DE COSTA RICA II SEMESTRE 2021**

**INGENIERÍA EN COMPUTACIÓN**

**IC-8060 RECUPERACIÓN DE INFORMACIÓN TEXTUAL**

**TAREA PROGRAMADA 1**

**GUÍA DE DOCUMENTACIÓN**

1. Introducción

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

1. Describa la estructura del índice:

* qué información se almacena sobre toda la colección
* qué información se almacena sobre cada uno de los documentos
* qué información se almacena sobre cada uno de los términos
* qué información se almacena en cada posting
* qué información adicional se almacena en algún otro archivo

1. Completar la siguiente tabla para describir el estado en que quedó el proyecto.

|  |  |  |
| --- | --- | --- |
| **Etapa** | **% de complet.** | **Comentario o aclaración** |
| **INDIZACIÓN** |  |  |
| **Transformaciones implementadas** |  |  |
| Separación en palabras (letras, dígitos y '\_') |  |  |
| Conversión a minúsculas |  |  |
| Eliminación de acentos, preservación eñe |  |  |
| Eliminación de stopwords |  |  |
| Eliminación de etiquetas XML usando ExpReg |  |  |
| **Estructura del índice** |  |  |
| Archivo general colección |  |  |
| Archivo documentos |  |  |
| Diccionario |  |  |
| Postings |  |  |
| Algún otro archivo |  |  |
| **BÚSQUEDA** |  |  |
| **Implementa consultas vectoriales** |  |  |
| Crea archivo escalafón correctamente |  |  |
| **Implementa consultas BM25** |  |  |
| Crea archivo escalafón correctamente |  |  |
| **Documento HTML** |  |  |
| Genera archivo HTML de salida |  |  |
| Incluye: texto consulta, fecha y hora; posición, similitud, ruta documento y primeros 200 caracteres. |  |  |
| **INSPECCIÓN** |  |  |
| Muestra adecuadamente la información pedida para un documento: id, longitud, norma |  |  |
| Muestra adecuadamente lainformación pedida para un término: ni, idfs, lista postings |  |  |
| **COMANDOS** |  |  |
| Implementan comandos para indización, búsqueda e inspección |  |  |
| Los errores de entrada no provocan la caída todal de la herramienta |  |  |

1. Pruebas.

* Indexar la colección **xml-es** usando los stopwords dados.  
  Llamar al índice generado, **indx1**.  
  Adjuntar el directorio generado en la documentación.
* Correr las tres consultas estándar:  
   **carga de cpu**  
   **la euro moneda**  
   **impuestos y depreciación**  
  usando el modelo vectorial y el modelo BM25.  
  Adjuntar en la documentación los escalafones y HTMLs obtenidos con dichas pruebas.  
  Ver la sección de PRUEBAS más adelante para información adicional.
* Correr la siguientes consultas a la herramienta de inspección:  
   mostrar el documento **xacc-pnl.xml**  
   mostrar el término **impuestos**  
  Ver en la sección de PRUEBAS los resultados esperados.

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 (**formato ZIP**) que incluya la siguiente estructura:  
 TP2-*ApellidosNombre1-ApellidosNombre2*  (directorio)  
 *Archivo\_de\_documentacion*  
 *Programas*  
 *Archivos\_adicionales*

*Pruebas*

**Asegúrese de que todos los archivos estén bajo un directorio que también está dentro del ZIP.**

1. Entrega

La entrega es este lunes 13 de setiembre a las 11:55pm.  
Usar TecDigital para cargar el proyecto y la documentación . Solo en caso de tener inconvenientes serios enviar el proyecto y documentación a josee.arayamonge@gmail.com.

**PRUEBAS**

Los resultados que aquí se muestran se obtuvieron usando la facilidad de “Find in files” de Notepad++. Se utilizaron expresiones regulares en búsquedas que incluían recorrer los subdirectorios. ***Debido a limitaciones como no poder omitir el texto en las etiquetas XML, los resultados presentados son una buena aproximación, pero en algunos casos podrían aparecer documentos escogidos por texto en etiquetas XML que su proyecto debe omitir.***

**Prueba 1**

Consulta: "**carga de cpu**".   
Usar los prefijos **qvec\_cpu** y **qbm25\_cpu** para las corridas vectorial y BM25 respectivamente.  
El resultado esperado es:

|  |  |
| --- | --- |
| **applets\cpuload-ug.xml** | **Esto documentos contienen ambos términos por lo que deberían aparecer en los primeros puestos del escalafón.** |
| **applets\memload-ug.xml** |
| **applets\netload-ug.xml** |
| **applets\swapload-ug.xml** |
| **gnome-utils-1.xml** |
| **ug-applets-1.xml** |
| **ug-applets-2.xml** |
| **ug-applets.xml** |
|  |  |
| applets\cpumemusage-ug.xml | Estos documentos solo contienen "cpu" que es un término más raro que "carga". |
| applets\drivemount-ug.xml |
|  |  |
| applets\battery-ug.xml | Estos documentos solo contienen "carga" que es un término más frecuence que "cpu", por lo que probablemente aparezcan más abajo en el escalafón. |
| gmenu.xml |
| gnome-libs-tutorial.xml |
| gnome-utils-4a.xml |
| guia\_usuario.xml |
| introduction-to-gnome.xml |
| nautilus-user-manual.xml |
| nautilus.xml |
| xacc-qif-import.xml |
| xacc-ticker.xml |

**Prueba 2**

Consulta "**La euro moneda**".   
Usar los prefijos **qvec\_euro** y **qbm25\_euro** para las corridas vectorial y BM25 respectivamente.  
El resultado esperado es:

|  |  |
| --- | --- |
| **xacc-currencyhandling.xml** | Esto documentos contienen ambos términos por lo que deberían aparecer en los primeros puestos del escalafón. |
| **xacc-euro.xml** |
| **xacc-userdocs.xml** |
|  |  |
| panel.xml | Estos documentos solo contienen "moneda" por lo que deberían aparecer más abajo en el escalafón. |
| xacc-acctypes.xml |
| xacc-newacctwin.xml |
| xacc-qif-import.xml |
| xacc-ticker.xml |

**Prueba 3**

Consulta "**impuestos y depreciación**". Usar los prefijos **qvec\_imp** y **qbm25\_imp** para las corridas vectorial y BM25 respectivamente.  
El resultado esperado es:

|  |  |
| --- | --- |
| **xacc-acctypes.xml** | **Esto documentos contienen ambos términos por lo que deberían aparecer en los primeros puestos del escalafón.** |
| **xacc-apar.xml** |
| **xacc-depreciation.xml** |
| **xacc-doubleentry.xml** |
|  |  |
| xacc-userdocs.xml | Solo contiene "depreciación" |
|  |  |
| preface.xml | Estos documentos solo contienen "impuestos". |
| xacc-chartofaccts.xml |
| xacc-incomeexpense.xml |

**Prueba 4**

Mostrar el documento **xacc-pnl.xml**

Longitud esperada del documento: 20.

Si bien no debe mostrar los términos de ese documento; verifique que dicho documento aparece en las listas invertidas de algunos de sus términos. Todos tienen frecuencia 1, con excepción de **listado** que aparece dos veces.

|  |  |  |  |
| --- | --- | --- | --- |
| barra | entradas | herramientas | puede |
| beneficios | especificar | listado (2) | sus |
| boton | estado | parametros | totaliza |
| contable | este | perdidas | utilizando |
| datos | gastos | periodo |  |

**Prueba 5**

Mostrar la información del término **impuestos**

Valores esperados:

|  |  |  |  |
| --- | --- | --- | --- |
| **ni** | **idf1 *log2(N/ni)*** | **idf2 *log2((N-ni+0.5)/(ni+0.5))*** | **postings *peso = log2(1+frecuencia)\*idf1*** |
| 7 | log2(95/7) =  3,762500686 | log2((95-7+0.5)/(7+0.5) =  3,560714954 | |  |  |  | | --- | --- | --- | | **documento** | **frecuencia** | **peso** | | preface.xml | 1 | 3.76250 | | xacc-acctypes.xml | 2 | 5.96342 | | xacc-apar.xml | 3 | 7.52500 | | xacc-chartofaccts.xml | 3 | 7.52500 | | xacc-depreciation.xml | 11 | 13.48842 | | xacc-doubleentry.xml | 1 | 3.76250 | | xacc-incomeexpense.xml | 2 | 5.96342 | |