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

**INGENIERÍA EN COMPUTACIÓN**

**IC-8060 RECUPERACIÓN DE INFORMACIÓN TEXTUAL**

**Tarea programada 1**

Desarrollar un sistema que permita realizar consultas de texto en archivos (documentos) que se encuentran almacenados recursivamente bajo un directorio dado. Se deben elaborar tres herramientas:

* una herramienta que índice la colección
* una herramienta que permita procesar consultas
* una herramienta que permita explorar los archivos generados por la primera herramienta

Puede integrar las tres herramientas en una sola, pero no es necesario.

**INDIZACIÓN**

La primera herramienta procesa los documentos de una colección dada (directorio) y crea un conjunto de archivos que permite realizar las búsquedas posteriores (índice). Esta herramienta debe calcular y almacenar toda la información necesaria para implementar el modelo vectorial con similitud coseno y el modelo BM25.

Los archivos creados deben ser almacenados en disco. Esto es, el resultado de la indización no debe perderse al terminar la ejecución de esta herramienta; no es aceptable que solo exista en memoria.

Se trabajará con la colección **xml-es** que se les adjunta. Los documentos pueden estar almacenados en subdirectorios y se caracterizan porque sus nombres tienen la extensiones ".xml". Debe asegurarse de no indizar archivos que no cumplan con dicha extensión.

Los términos se obtienen transformando las palabras que aparecen en los documentos por medio de las siguientes reglas:

* se considera como palabra a cualquier secuencia consecutiva de una o más letras o dígitos; se considerará el caracter '\_' como letra
* las letras deben ser convertidas a minúsculas y sus tildes deben quitarse; sin embargo, se debe preservar la virgulilla de la eñe sin cambios
* se debe consultar la siguiente lista de *stopwords* con el fin de eliminar palabras no significativas:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **a**  **ante**  **bajo**  **cabe**  **con**  **contra**  **de**  **desde** | **e**  **el**  **en**  **entre**  **hacia**  **hasta**  **ni**  **la** | **le**  **lo**  **los**  **las**  **o**  **para**  **pero**  **por** | **que**  **se**  **segun**  **sin**  **so**  **sobre**  **tras**  **u** | **un**  **una**  **unas**  **uno**  **unos**  **y** |

* se deben usar expresiones regulares para extraer los términos; no se pueden usar bibliotecas que manejan XML y que permiten extraer el texto

Para ejecutar la herramienta de indización se debe invocar el comando siguiente:

**indizar *Colección* *Stopwords* *Índice***

Cuyos parámetros se describen a continuación:

*Colección* ruta del directorio raíz en donde están almacenados los documentos de la colección

*Stopwords* ruta del archivo que contiene la lista de palabras que se deben omitir por no ser significativas, el formato de ese archivo puede ser una línea por palabra o una línea conteniendo todas las palabras separadas por un espacio en blanco

*Índice* ruta del directorio en que se almacenará el archivo o archivos que componen el índice generado para la colección; el archivo de stopwords debe copiarse en este directorio para usarlo al indizar consultas

Se sugiere que el índice contenga lo siguiente:

* información sobre la colección: ruta del directorio que la contiene, número de documentos, longitud promedio de los documentos
* copia del archivo de stopwords (para usarlo al indizar consultas)
* información sobre documentos: su id, su ruta relativa al directorio de la colección, su longitud y su norma
* información sobre los términos (diccionario): término, ni, idfs, e información sobre cómo localizar la lista de postings correspondiente
* información sobre los postings: lista de tuples (docid, frecuencia, peso)

Puede incluir más información en el índice si así lo requiere.

**BÚSQUEDA**

La segunda herramienta del sistema realiza consultas tomando como base algún índice generado anteriormente.

Para ejecutar la herramienta de búsqueda se debe invocar el comando siguiente:

**buscar *Índice* *Tipo* *Prefijo* *NumDocs* *Consulta***

Cuyos parámetros se describen a continuación:

*Índice* ruta del directorio en que se contiene los archivos que componen el índice usado;

*Tipo* bandera para escoger si se trata de una búsqueda vectorial (**vec**) o BM25 (**bm25**).

*Prefijo* prefijo usado con todos los archivos de salida producidos por este comando

*NumDocs* cantidad de los primeros documentos del escalafón que serán mostrados en un archivo HTML

*Consulta* texto de la consulta; incluye todas las palabras hasta el final del comando; los términos de la consulta se pueden repetir

Ejemplos de búsqueda

* **buscar indx1 vec q\_arch1 20 compresión y manejo de archivos**
* **buscar indx2 bm25 q\_arroz 30 arroz con leche y leche agria**

Las consultas deberán sufrir las mismas transformaciones que se le aplicaron a los documentos.

Similitudes

La herramienta debe permitir escoger entre dos modalidades de consultas: ***vectorial tf\*idf*** y ***BM25***.

Parar la modalidad vectorial, se procesan las consultas usando la fórmula de similitud *coseno*:

, dónde

Los pesos wij y wiq se calculan usando las fórmulas estándar:

La segunda modalidad de consulta es una variante muy exitosa del modelo probabilístico: *Okapi BM25*, donde BM significa *best matching*. En realidad BM25 es una familia de funciones de similitud con algunas variantes en las fórmulas y en los parámetros. Se usará la siguiente función:

Dada una consulta Q con términos q1, ..., qn, y un documento D

donde

* *f(qi,D)* es la frecuencia con que aparece el término qi en el documento D
* *|D|* es la longitud del documento D en palabras (suma de las frecuencias de sus términos)
* *avgdl* es la longitud promedio de los documentos de la colección
* *k* es un parámetro que calibra la escala de la frecuencia del término qi en el documento D, **se usará k1=1.2**.
* b es otro parámetro, 0≤b ≤1, determina hasta qué punto se normalizará la longitud de un documento; **se usará b=0.75**.

Para IDF(qi) se usará la siguiente fórmula:

donde

*N* es el número total de documentos de la colección

*n(qi)* es el número de documentos de la colección que contienen qi.

En caso de que ni > N/2 el valor es negativo; en ese caso se usará el valor cero.

Usando esta función de similitud, se debe producir un escalafón que ordene los documentos en orden decreciente por similitud.

Salida

Ambas modalidades de consulta deben producir dos archivos de salida:

**Escalafón**Un archivo con el escalafón completo del resultado de la consulta.  
  
Esto es, un archivo que contenga para todos los documentos con similitud mayor que cero, su posición en el escalafón, el identificador de cada documento y el valor de similitud obtenido por ese documento. Este archivo debe venir ordenado descendentemente por similitud.  
El nombre de este archivo se obtiene concatenando el parámetro ***Prefijo*** con la extensión "**.esca**".

**HTML**Un archivo HTML con los primeros ***NumDocs*** documentos del escalafón.  
  
Al inicio del archivo se muestra el texto de la consulta y la fecha y hora en que se realizó.  
Luego, para cada uno de esos documentos se debe mostrar: su posición en el escalafón, la similitud con la consulta y su ruta como texto caliente. Además, para cada documento se deben mostrar los primeros 200 caracteres extraídos de cada documento (eliminando todas las etiquetas XML).  
El nombre de este archivo se obtiene concatenando el parámetro ***Prefijo*** con la extensión "**.html**".

**HERRAMIENTA DE INSPECCIÓN**

La tercera herramienta permite permita inspeccionar el índice creado, con el fin de poder verificar su contenido.

Para ejecutar la herramienta de inspección se debe invocar el comando siguiente:

**mostrar *Índice* *Tipo* *Dato***

Cuyos parámetros se describen a continuación:

*Índice* ruta del directorio en que se contiene los archivos que componen el índice usado

*Tipo* bandera que escoge lo que se quiere revisar: término (**ter**) o documento (**doc**)

*Dato* es la ruta del archivo (relativa al directorio de la colección) o el término que se quiere buscar

Si se le da a la herramienta el nombre de un documento (por ejemplo, "**applets/mixer-ug.xml**" la herramienta mostraría toda la información almacenada para ese documento: id, longitud y norma. Por otro lado, si se le da a la herramienta un término (por ejemplo, ***memoria***), la herramienta mostraría toda la información almacenada para ese término: ni, idf en sus dos versiones y la lista de sus postings.

La salida debe ser adecuadamente formateada para que sea razonablemente legible.

Si los archivos con escalafón no son directamente legibles por una persona, esta herramienta además deberá listarlos en forma legible.

**CONSIDERACIONES ADICIONALES**

Se recomienda usar una herramienta de scripting como Python o Perl, pero se puede usar cualquier lenguaje que permita usar expresiones regulares para manipular el texto, Java por ejemplo. También puede usar varias herramientas/lenguajes y combinarlas para obtener el resultado deseado.

La falta de un argumento o un formato inválido en los comandos no debe provocar la caída de las herramientas. En caso de un error simplemente rechace el comando.

Se deben correr los programas en la colección que se adjunta, **xml-es**. Se deberán correr las consultas estándar que se incluyen al final de este documento.

La tarea puede ser desarrollada en grupos de dos personas. La fecha de entrega es lunes 13 de setiembre a las 11:55pm.

**Consultas estándar**

A continuación se muestran tres consultas y los resultados esperados.

1. Consulta: "**carga de cpu**"

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

1. Consulta "**La euro moneda**"

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

1. Consulta "**impuestos y depreciación**"

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