**INSTITUTO TECNOLÓGICO DE COSTA RICA II SEMESTRE 2018**

**INGENIERÍA EN COMPUTACIÓN**

**IC-8020 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 las siguientes 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 BM25.

Los archivos creados den 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.

A continuación, se describen los diferentes archivos producidos por la indización:

|  |  |
| --- | --- |
| **Archivo** | **Contenido** |
| Colección | Contiene una única línea o registro con la siguiente información:   * nombre de la colección * ruta completa del directorio que almacena la colección * número de documentos de la colección * longitud promedio de los documentos |
| Documentos | Para cada documento contiene una línea o registro con la siguiente información:   * número consecutivo único generado automáticamente (docid) * ruta relativa del documento con respecto al directorio de la colección * primeras 200 palabras de la sección DESCRIPCIÓN del documento (ver más adelante) * número de términos distintos del documento * longitud del documento: suma de las frecuencias de todos los términos del documento * lista de pares (término, frecuencia) para los términos que aparecen en el documento; la lista debe venir ordenada alfabéticamente por término |
| Vocabulario | Para cada término de la colección contiene una línea con la siguiente información:   * término * ni (cantidad de documentos distintos en los que aparece) * valor idfi usado en BM-25:   si ni aparece en más de la mitad de los documentos este valor sería negativo, poner un cero en su lugar. |

Los archivos anteriores deben ser explícitamente creados y preservados para ser usados por las otras herramientas. Puede incluir información adicional en los mismos, pero no se puede omitir la información detallada arriba.

Se trabajará con la colección **man-es** que se les adjunta. Los documentos se encuentran almacenados en subdirectorios y se caracterizan porque su nombre tiene una extensión que es un dígito en el rango 1-8. Debe asegurarse de no indizar archivos que no cumplan con dicha extensión.

Los términos se extraen del texto siguiendo las reglas que se detallan a continuación:

* Los términos consisten de una secuencia de uno o más de los caracteres **[a-z0-9\_.]**. Además de esto, los términos deben cumplir las otras reglas.
* Las letras deben ser convertidas a minúsculas y se deben eliminar las tildes y diéresis. La letra eñe debe ser preservada y **no** ser convertida a una ene.
* 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**  **segun**  **sin**  **so**  **uno**  **unas**  **unos**  **y** | **sobre**  **tras**  **u**  **un**  **una** |

**NO MODIFIQUE ESTA CONJUNTO DE STOPWORDS. SE DEBE USAR TAL COMO ESTÁ.**

* Los números también son considerados términos.
* El carácter punto (‘**.**’) solo se acepta si separa fragmentos que por sí mismos son secuencias de uno o más caracteres **[a-z0-9\_]**; los puntos no pueden estar ni al inicio ni al final, y no pueden estar juntos:
  + “**libnss\_nisplus.so.1**”, “**POSIX.1b**” y “**B.1.1.1.3**” son términos válidos
  + “**.hola**”, “**adios.**” y “**hola..adios**” no son términos válidos
  + “**1.2**” y “**1.1.2**” son términos válidos
  + “**1.**”, “**.2**” y “**1..2**” no son términos válidos
* Además del término completo que incluye puntos, se deben incluir como términos aparte los fragmentos que lo forman; siempre y cuando no contengan solo dígitos:
  + “**libnss\_nisplus.so.1**” 🡪 “**libnss\_nisplus**” “**so**”
  + “**POSIX.1b**” 🡪 “**POSIX**” “**1b**”
  + “**B.1.1.1.3**” 🡪 “**B**”
  + “**a1.2b.3**” 🡪 “**a1**” “**2b**”
* Si una palabra es precedida por la tira "**--**" entonces se además de la palabra como tal, se debe incluir un término que le antepone el carácter "**@**". Por ejemplo   
   "**--version**"  
  produce los términos  
   "**version**" y "**@version**";  
  la idea es permitir buscar palabras que son usadas como parámetros
* Al procesar las palabras hay que tomar en cuenta algunas de ellas están separadas entre líneas con guiones ***especiales*** al final de la línea. Por ejemplo:

**# Se expande al número en base 10 de parámetros posi­  
 cionales.**

Se deben unificar esas palabras dividas (**posicionales**); ignorar los espacios en blanco y el cambio de línea que hay entre el guion final de una línea y el resto de la palabra en la línea siguiente. Obviamente también se debe eliminar el guion al unificar la palabra. Tomar en cuenta que este guion es un carácter especial.

Como entrada a la herramienta de indización se debe proveer la siguiente información:

*stopwords* ruta del archivo que contiene la lista de palabras que se deben omitir por no ser significativas

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

*índice* ruta del directorio en que se almacenarán los archivos que componen el índice generado para la colección

**La colección debe ser considerada como de lectura solamente. No se deben modificar los archivos de la colección o agregar archivos dentro de ningún directorio de la colección.**

Búsqueda

La segunda herramienta del sistema realiza consultas tomando como base los archivos generados anteriormente. Las consultas consisten de una lista de términos (los cuales pueden repetirse); por ejemplo:  
 **compresión de archivos y manejo de archivos comprimidos**

Las consultas serán procesadas usando el modelo probabilístico *Okapi BM25*. Esto es, se usará la siguiente función de similitud.

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

k=0, sería frecuencia binaria (0 si no aparece, 1 si aparece sin importar las veces)

k grande, se usa la frecuencia cruda sin ninguna escala

Usualmente se toma k ∈ [1.2, 2.0], **se usará k=1.2**.

* b es otro parámetro, 0≤b ≤1, determina la escala de la longitud del documento

b=1 corresponde a normalizar completamente usando la longitud del documento

b=0 corresponde a no normalizar, no tomar en cuenta la longitud del documento

Se tomará b=0.75.

Para IDF(qi) se usará la siguiente fórmula:

donde

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

*ni* es el número de documentos de la colección que contienen el término qi.

Esta fórmula de problemas si un término aparece en más de la mitad de los documentos de la colección (el logaritmo se vuelve negativo). **Por eso, se deben descartar de la consulta aquellos términos que aparezcan en más de la mitad de los documentos de la colección.**

Salida

Cada consulta debe producir dos archivos de salida:

* Un archivo con el escalafón completo del resultado de la consulta. Esto es, un archivo de texto que para todos los documentos con similitud mayor que cero, contenga una línea con la posición en el escalafón, el identificador del documento y el valor de similitud obtenido por ese documento. Este archivo debe venir ordenado descendentemente por similitud.
* Un archivo HTML que para cada uno de los primeros documentos del escalafón incluya su posición en el escalafón, la similitud con la consulta y su ruta. Además, para cada documento se deben listar los primeros 200 caracteres que aparecen después de la línea que dice “**DESCRIPCIÓN**”; antes de contar los 200 caracteres, se deben reemplazar los cambios de línea por un espacio en blanco y se deben consolidar dos o más espacios en blanco en uno solo. El usuario debe tener la posibilidad de poder la cantidad de documentos que se muestran en la página HTML.

Como entrada a la herramienta de búsqueda se debe proveer la siguiente información:

*índice* ruta del directorio en que se almacena el archivo o archivos que componen el índice

*numDocs* número de documentos que serán incluidos en la página HTML de salida;  
serán los documentos en las posiciones **1..numDocs** del escalafón

*resultado* prefijo que será usado para obtener los nombres de los dos archivos de respuesta;  
  
por ejemplo si resultado es “**q1-disco-virtual**”, los archivos resultantes serán “**q1-disco-virtual.esc.txt**” “**q1-disco-virtual.html**” para el escalafón y página html resultantes

*consulta* lista de términos que forman la consulta.

Herramienta de inspección

Se debe desarrollar una herramienta que permita inspeccionar los archivos creados. En particular, debe proveer la siguiente funcionalidad:

* Si se le da a la herramienta un índice y el nombre de un documento (por ejemplo, **chown.2**) la herramienta mostraría toda la información almacenada en ese índice para ese documento: términos, frecuencias, pesos, norma, longitud, etc.
* Por otro lado, si se le da a la herramienta un índice y un término (por ejemplo, ***memoria***), la herramienta mostraría toda la información almacenada en ese índice para ese término: ni, idfi. y para sus términos vecinos: los 5 anteriores y los 5 posteriores. (Los términos deben estar ordenados alfabéticamente)

La salida debe ser adecuadamente formateada para que sea razonablemente legible, no debe haber líneas con 1000 caracteres de largo, por ejemplo.

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.

* Se deben correr los programas en la colección que se adjunta, **man.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 el lunes 12 de abril a las 8 am.

**Consultas estándar**

Se tienen tres consultas para la colección man.es, cuyos enunciados de relevancia se adjuntan.

El tema de la primera consulta es:

**compresión de archivos y manejo de archivos comprimidos**.

Tener en cuenta que la colección usa fichero en vez de archivo.

Los documentos relevantes de la colección man.es son:

|  |  |
| --- | --- |
| **bunzip2.1**  **bzcat.1**  **bzip2.1**  **gunzip.1**  **gzexe.1**  **gzip.1** | **zcat.1**  **zcmp.1**  **zdiff.1**  **zgrep.1**  **zmore.1**  **znew.1** |

El tema de la segunda consulta es:

**colas de mensajes para la comunicación entre procesos**.

Los documentos relevantes de la colección man.es son:

|  |  |
| --- | --- |
| **ipc.5**  **ipcs.8**  **msgctl.2**  **msgget.2**  **msgop.2**  **msgrcv.2**  **msgsnd.2** | **recv.2**  **recvfrom.2**  **recvmsg.2**  **sendmsg.2**  **sendto.2**  **socket.2**  **socket.7**  **tcp.7** |

El tema de la tercera consulta es:

**segmentos de memoria compartida**

Los documentos relevantes de la colección man.es son:

|  |
| --- |
| **shmat.2**  **shmctl.2**  **shmdt.2**  **shmget.2**  **shmop.2**  **ipc.5**  **ipcs.8** |