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

**INGENIERÍA EN COMPUTACIÓN**

**IC-8020 RECUPERACIÓN DE INFORMACIÓN TEXTUAL**

**Tarea programada 1**

Desarrollar un sistema que permita buscar consultas de texto en archivos (documentos) que se encuentran almacenados recursivamente bajo un directorio dado.

Indexación

El sistema consta de una primera herramienta que procesa los documentos del directorio y crea un conjunto de archivos que permite las búsquedas posteriores.

Los archivos de documentos se identifican porque tienen la extensión ‘**.txt**’.

A continuación se describe el contenido de los archivos producidos por la herramienta de indexación:

|  |  |
| --- | --- |
| **Archivo** | **Contenido** |
| ~~Colección~~ | ~~Contiene una única línea con la siguiente información:~~   * ~~nombre de la colección~~ * ~~ruta completa del directorio que almacena la colección~~ * ~~número de documentos de la colección~~ |
| ~~Documentos~~ | ~~Para cada documento contiene una línea con la siguiente información:~~   * ~~número consecutivo único generado automáticamente (docid)~~ * ~~ruta relativa del documento con respecto al directorio de la colección~~ |
| ~~Frecuencias~~ | ~~Para cada documento contiene una línea con la siguiente información:~~   * ~~docid~~ * ~~número de términos distintos del documento~~ * ~~lista de pares (término, frecuencia) para los términos que aparecen en el documento; la lista debe venir ordenada alfabéticamente~~ |
| ~~Pesos~~ | ~~Para cada documento contiene una línea con la siguiente información:~~   * ~~docid~~ * ~~número de términos distintos del documento~~ * ~~norma del documento: raíz cuadrada de la suma de los cuadrados de los pesos~~ * ~~lista de pares (término, peso) para los términos que aparecen en el documento; los pesos deben calcularse usando la fórmula tf \* idf:~~   ~~la lista debe venir ordenada alfabéticamente.~~ |
| ~~Vocabulario~~ | ~~Para cada término de la colección contiene una línea con la siguiente información:~~   * ~~término~~ * ~~n~~~~i~~ ~~(cantidad de documentos distintos en los que aparece).~~ |

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

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

* ~~las palabras de los documentos son transformadas para quitarles las tildes y convertirlas a minúsculas~~
* ~~se debe preservar la eñe~~
* ~~se debe consultar un archivo de~~ *~~stopwords~~* ~~con el fin de eliminar el siguiente conjunto de palabras no significativas (stopwords):~~

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

* ~~DEBEN USAR EXACTAMENTE ESTE CONJUNTO DE STOPWORDS.~~
* ~~los términos consisten de una secuencia de uno o más de los siguientes caracteres~~**~~[a-z0-9\_.]~~**~~; con las siguientes restricciones~~
  + ~~los números son aceptados como términos~~
  + ~~el caracter ‘.’ solo se acepta si separa fragmentos que por sí mismos son secuencias de u 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~~**~~”, “~~**~~2.1~~**~~” y “~~**~~B.1.1.1.3~~**~~”~~~~son términos válidos~~
    - ~~“~~**~~.hola~~**~~”, “~~**~~adios.~~**~~” y “~~**~~hola..adios~~**~~” no son términos válidos.~~
  + ~~Además del término completo que incluye puntos, se deben incluir los fragmentos que lo forman como términos por aparte; siempre y cuando no contengan solo dígitos:~~
    - ~~“~~**~~libnss\_nisplus.so.1~~**~~” → “~~**~~libnss\_nisplus~~**~~” “~~**~~so~~**~~”~~
    - ~~“~~**~~POSIX.1b~~**~~” → “~~**~~POSIX~~**~~” “~~**~~1b~~**~~”~~
    - ~~“~~**~~B.1.1.1.3~~**~~” → “~~**~~B~~**~~”~~
* ~~Al procesar las palabras hay que tomar en cuenta algunas de ellas están separadas entre líneas con guiones al final de la línea. Por ejemplo:~~ **~~# Se expande al número en base 10 de parámetros posi­  
   cionales.~~**~~Se deben unificar esas palabras dividas; 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.~~

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

## Duda con si de fijo se necesita hacer iteración igual a la de los documentos

Habrá dos modalidades de consultas. En la primera, se procesan las consultas usando la fórmula de similitud *coseno* del modelo vectorial:



Usando esa similitud, producir un escalafón que ordene los documentos en orden decreciente por similitud y mostrar dicho escalafón al usuario.

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)

*avgl* 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 k1=1.5**.

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=1.

Para IDF(qi) se usará la siguiente fórmula:

dónde

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

*n(qi)* es el número de documentos de la colección que contienen 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 eliminar los términos que aparezcan en más de la mitad de los documentos de la colección. Eso debe hacerse antes de calcular los pesos.**

El archivo **TP1 2017- BM25.xlsx** muestra un ejemplo de este modelo.

Salida

~~En ambas modalidades de consulta, el resultado es un archivo HTML que para cada documento recuperado incluye su posición en el escalafón, la similitud con la consulta, la ruta del archivo, fecha de creación y tamaño en bytes. Además deben listarse los primeros 200 caracteres del archivo que aparecen después de la línea que dice “~~**~~DESCRIPCIÓN~~**~~”, eso sí, consolidando dos o más espacios en blanco en uno solo, y reemplazando los cambios de línea por un espacio en blanco.~~

~~Se recomienda usar una herramienta de scripting como Python o Perl, pero en se puede usar cualquier lenguaje que permita usar expresiones regulares para manipular el texto. El lenguaje usado debe estar disponible eso sí en los laboratorios de la Escuela. También puede usar varias herramientas/lenguajes y combinarlas para obtener el resultado deseado.~~

Formato de los comandos de las herramientas

**~~generar~~***~~Stopwords~~**~~Directorio~~**~~Prefijo~~*

*~~Stopwords~~* ~~archivo que contiene la lista de palabras que se deben omitir por no ser significativas~~

*~~Directorio~~* ~~Ruta del directorio raíz a partir del cual están almacenados los documentos de la colección en subdirectorios.~~

*~~Prefijo~~* ~~Prefijo que será usado para los archivos que son generados por este comando:  
 Colección:~~ *~~Prefijo~~*~~\_~~**~~CO~~** ~~Documentos:~~ *~~Prefijo~~***~~\_DO~~** ~~Frecuencias:~~ *~~Prefijo~~***~~\_FC~~** ~~Pesos:~~ *~~Prefijo~~***~~\_PE~~** ~~Vocabulario:~~ *~~Prefijo~~***~~\_VO~~**

**~~buscar~~***~~Modalidad~~**~~NumInicio NumFin Prefijo ArchivoHTML Consulta~~*

*~~Modalidad~~* ~~señala el tipo de consulta que se usará para crear el escalafón; se tienen las siguientes alternativas:~~ **~~vec~~**~~: consulta vectorial con coseno como similitud~~ **~~bm25~~**~~: consulta probabilística usando BM25.~~

*~~NumInicio NumFin~~*~~rango de posiciones del escalafón que serán incluidas en el documento HTML de salida.~~

*~~Prefijo~~* ~~Prefijo para los archivos generados antes que son usados por este comando.~~

*~~ArchivoHTML~~* ~~Nombre del archivo html con el resultado de esta consulta.~~

*~~Consulta~~* ~~Lista de términos que forman la consulta.~~

~~Ejemplos~~

generar ./Stopwords.txt ./man.es TEST

buscar vec 1 20 PRU Resultado1.html redes seguras comunicacion datos

buscar bm25 1 30 PRU Resultado2.html compresion de archivos y manejo de archivos comprimidos

Herramientas de inspección

Si guarda la información en forma binaria, por ejemplo exportando un objeto, debe incluir herramientas que permita visualizar adecuadamente (imprimir) ese contenido binario. La salida debe ser adecuadamente formateada (no debe ser una línea de 100k de largo).

Consideraciones adicionales

* 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 miércoles 6 de setiembre.

**Consultas estándar**

Se tienen dos consultas para la coleccion man.es, cuyos enunciados de relevancia se adjuntan.

El tema de la primera consulta es:

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

Los documentos relevantes de la coleccion 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 coleccion 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**