Objetivo de la práctica:
Hacer un programa completo que permita poner en práctica los conceptos aprendidos en las sesiones anteriores.

NOTA 1: Durante la práctica todos los ejercicios deberán ser guardados **temporalmente** en el directorio \tmp. Una vez finalizada la misma y transferidos los ficheros a un disquete, se deberá eliminar dicho directorio.

OBJETIVO

Cuando se plantea el problema de buscar en un almacén de documentos (por ejemplo, una base de datos documental) aquellos documentos que contengan una palabra determinada, existen diversas aproximaciones para resolverlo. Una primera aproximación sería buscar directamente en todos los documentos las palabras utilizando algoritmos de búsqueda de textos, pero esta solución sería altamente ineficiente y demasiado lenta cuando el conjunto de documentos es muy elevado.

Lo más habitual es realizar un proceso de indexación sobre los documentos para extraer las palabras que contiene cada documento y crear una estructura de apoyo a las búsquedas (un índice), de forma que cuando se desea encontrar los documentos que contienen una palabra, se realiza una búsqueda sobre la estructura de apoyo y se obtiene directamente el conjunto de documentos que contienen la palabra.

En la estructura de apoyo se suele almacenar, entre otra información, las palabras que aparecen en los documentos, los documentos que contienen las palabras, y la frecuencia de aparición de las palabras en dichos documentos.

Por tanto, indexar un conjunto de documentos significa extraer las palabras que aparecen en dichos documentos y almacenarlas en una estructura de apoyo (índice) junto con cierta información estadística que permita acelerar los procesos de búsqueda.

El objetivo del proyecto que se plantea es crear un programa para indexar un conjunto de documentos y poder buscar palabras y cadenas de texto que aparezcan dentro de los documentos indexados.

DESARROLLO

El alumno dispondrá de un conjunto de ficheros de texto plano, cada uno correspondiente a un cuento infantil, que son los documentos que hay que indexar. Los ficheros son: cenicienta.txt, simbad.txt, los3cerditos.txt, caperucita.txt, elflautistadehamelin.txt, peterpan.txt, blancanieves.txt, merlin.txt, y elpatitofeo.txt.

El programa debe extraer todas las palabras de todos los ficheros de texto, y crear una estructura de memoria (por ejemplo, se puede utilizar un vector de estructuras de tamaño 10000) que contenga la siguiente información:

- Palabra que aparece en el fichero.
- Fichero donde aparece la palabra.
- Número de veces que aparece la palabra en el fichero.

El programa funcionará con un menú que indicará las siguientes opciones:

1. Indexar la colección de documentos.
2. Guardar el índice en un fichero.
3. Buscar una palabra completa.
4. Buscar una cadena de texto.
5. Visualizar todos los índices creados.
6. Salir de la aplicación.

1. Indexar la colección de documentos

Para realizar la indexación se utilizará un fichero de apoyo que contendrá la lista de documentos que se van a indexar, de modo que aparezca un nombre de fichero en cada línea. Dicho fichero se llamará *lista.dat* y lo debe crear el alumno. Este fichero es el equivalente a un índice de documentos que indica como está compuesta la colección.

Indexar los ficheros, como ya se ha mencionado, significa extraer todas las palabras que aparecen en los mismos, y guardarlas en la estructura de datos (índice) definida para permitir la realización de búsquedas de forma más rápida y eficiente. En el proceso de extracción de las palabras hay que tener en cuenta las siguientes consideraciones:

- Se considera una palabra a toda secuencia de caracteres del archivo delimitada por principios de línea, espacios en blanco, signos de puntuación y/o finales de línea. Para el proyecto que se plantea, se puede considerar que los signos de puntuación no van a influir en la determinación de los límites de una palabra, puesto que siempre estarán precedidos o se encontrarán seguidos de un principio de línea, espacio en blanco o un final de línea. No obstante, esto puede suponer que las palabras obtenidas en el proceso contengan signos de puntuación que no son válidos para identificar las palabras. En la práctica, esto significa que se puede utilizar *cin* para obtener palabras individuales, con un tratamiento posterior de los signos de puntuación.
- Las palabras se deben guardar en mayúsculas, puesto que de otro modo las búsquedas serían dependientes de si las palabras estaban en mayúsculas, minúsculas, o combinaciones de ambas (ayuda: usar la función touper de la librería cctype).
- Hay que eliminar signos de puntuación que puedan formar parte de las palabras. En concreto, tras obtener cada palabra del fichero correspondiente, habrá que eliminar de la misma los caracteres siguientes:

  - + * / . , ? ¿ ¡ ! " ' \ : ; ( )

- Por otra parte, los caracteres acentuados, diéresis y eñes se deben convertir a mayúsculas puesto que no son convertidos por la función touper. La tabla de conversión es la siguiente

Carácter	Convertir a
á	Á
é	É
í	Í
ó	Ó
ú	Ú
ü	Ü
ñ	Ñ

2. Guardar el índice en un fichero

Una vez realizada la indexación de documentos, se debe poder guardar el índice en un fichero de texto plano, donde cada línea corresponderá con una entrada del índice.

El fichero de índices se llamará *indice.dat*, y cada entrada del índice tendrá el formato:

<palabra>###<fichero>###<numero de apariciones>

Este fichero podría utilizarse en sesiones posteriores para generar en memoria la estructura de apoyo (índice) de los documentos contenidos en la base de datos documental, aunque no es el objetivo de este proyecto implementar esta funcionalidad.

3. Buscar una palabra completa

En esta opción del menú el usuario del programa introducirá una palabra por teclado, y el programa buscará dicha palabra en el índice. El resultado de la búsqueda serán todas las entradas del índice para dicha palabra, y se mostrará la palabra, los ficheros donde aparece, y el número de veces que aparece en cada fichero.

Hay que tener en cuenta que la palabra de consulta se debe formatear exactamente igual que las palabras que se han indexado.

4. Buscar una cadena de texto

En esta opción del menú el usuario del programa introducirá una cadena de caracteres por teclado, y el programa buscará dicha cadena como subcadena de entradas en el índice, es decir, se buscarán palabras que contengan a la cadena. El resultado de la búsqueda serán todas las entradas del índice para dicha palabra, y se mostrará la palabra, los ficheros donde aparece, y el número de veces que aparece en cada fichero.

Hay que tener en cuenta que la cadena de consulta se debe formatear exactamente igual que las palabras que se han indexado.

5. Visualizar todos los índices creados

Esta opción sacará por la pantalla del usuario, en conjuntos de 20 elementos, todos los índices obtenidos en el proceso de indexación.

NOTA 2: Copiar en disquete los ejercicios realizados y eliminar el directorio temporal de trabajo (\tmp).