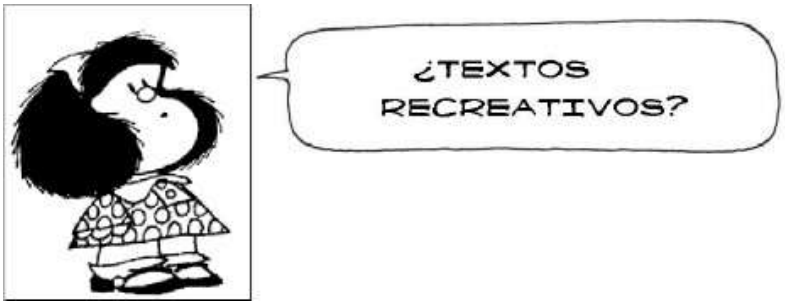
Curso: CO-4108 Taller I
Grupo 2, II-2017

Proyecto
Buscador de textos



Objetivo

Desarrollar una aplicación en Python 3 que permita buscar en un contenedor de documentos aquellos documentos que contengan una palabra determinada, con el fin de evitar buscar directamente en todos los documentos las palabras utilizando algoritmos de búsqueda de textos.

Descripción.

Se requiere que el sistema realice un proceso de indexación sobre los documentos para extraer las palabras que contiene cada documento y así crear una estructura que permita realizar las búsquedas (esto es lo que se conoce como un índice), de forma que cuando se desea encontrar los documentos que contienen una palabra, se realiza una búsqueda sobre la estructura de apoyo (índice) y se obtiene directamente de dicho índice el conjunto de documentos que contienen la palabra.

En el índice 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 lo que indexar un conjunto de documentos significa extraer las diferentes palabras que aparecen en dichos documentos y almacenarlas en el índice junto con cierta información estadística que permita acelerar los procesos de búsqueda.

El objetivo del proyecto es crear un programa para indexar un conjunto de documentos en formato “txt” proporcionado y poder buscar palabras y subcadenas de texto que aparezcan dentro de los documentos indexados, además de los documentos proporcionados el programa

deberá permitir agregar cualquier otro documento en formato texto("txt") para realizar la indexación respectiva.

Para esto usted dispondrá de un conjunto de archivos (contenidos en: *ArchivosProyecto-II-2017.rar* en la carpeta de proyectos del curso en el tecDigital) que deberá utilizar para verificar la correcta funcionalidad de acuerdo a lo que se pide en ésta especificación, el programa debe extraer todas las palabras de todos los archivos de texto, y crear una estructura de memoria (índice) (por ejemplo, se puede utilizar listas, diccionarios, arreglos etc.) que contengan al menos la siguiente información:

- Palabra que aparece en el archivo.
- Archivo donde aparece la palabra.
- Número de veces que aparece la palabra en el archivo.

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

1. Indexar la colección de documentos.
2. Guardar el índice en un archivo.
3. Buscar una palabra completa.
4. Buscar una subcadena de texto.
5. Visualizar todos los índices creados.
6. Crear / editar archivo de lista de documentos
7. Salir del programa.

Para realizar la indexación se deberá crear un archivo de apoyo que contendrá la lista o colección de documentos que se van a indexar, de modo que aparezca un nombre de archivo en cada línea.

Dicho archivo se llamará directorio.dat y se debe crear/modificar mediante la aplicación. Este archivo es el equivalente a una lista de documentos que indica cómo está compuesta la colección.

A continuación se describe la funcionalidad requerida por cada opción del menú indicado anteriormente:

Indexar la colección de documentos

Como ya se ha mencionado el proceso de indexación significa extraer todas las palabras que aparecen en los archivos, 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, 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.

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.

Hay que eliminar signos de puntuación que puedan formar parte de las palabras. En concreto, tras obtener cada palabra del archivo correspondiente, habrá que eliminar de la misma los caracteres siguientes: - + * / . , ? ¿ ¡ ! " ' \ : ; ( )

Por otra parte, los caracteres con acento (tildes), diéresis y eñes se deben convertir a mayúsculas. Para todos los efectos, la tabla de conversión es la siguiente:

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

Guardar el índice en un archivo

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

El archivo de **índice** se llamará indice.dat, y cada entrada del índice tendrá el formato:
<palabra>###<archivo>###<numero de apariciones>

Este archivo podría utilizarse en sesiones posteriores para cargar en memoria el **índice** con la estructura indicada de todos de los documentos contenidos en la carpeta de documentos, para realizar las búsquedas indicadas en las opciones del menú.

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 archivos donde aparece, y el número de veces que aparece en cada archivo.
Hay que tener en cuenta que la palabra de consulta se debe formatear exactamente igual que las palabras que se han indexado.

Buscar una subcadena 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 archivos donde aparece, y el número de veces que aparece en cada archivo.

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

Visualizar todos los índices creados

Esta opción listará por la pantalla, en conjuntos de 20 elementos, todos los índices obtenidos en el proceso de indexación de modo que se pueda ir visualizando en grupos de 20 desde el inicio hasta el final.

Crear / editar archivo de lista de documentos

Esta opción permitirá dar mantenimiento (incluir, modificar o borrar) la lista o colección de documentos a indexar.

Lo procesos de búsqueda de palabras o subcadenas deberán realizarse solamente mediante la utilización del índice.

Contenido de la aplicación.

Su programa deberá contar con lo siguiente:

- Adecuada interfase para entrada de datos y salida de datos.(Menus, pantallas de ingreso y datos y salida de datos estadísticos)
- En las pantallas deben manejarse reglas de validación que garanticen la consistencia de la información.
- Disponer de visualización de ayuda al usuario.
- Documentación interna y externa.
- Manual del usuario.

La aplicación deberá hacer uso de las estructuras de datos estudiadas en el curso, así como de sentencias apropiadas. Deberá contar con módulos. Los procedimientos y variables deberán contar con comentarios adecuados para su mejor lectura.

Metodología.

El proyecto se trabajará de forma individual, por lo que en caso de detectarse copias totales o parciales se otorgará cero puntos.

Evaluación

- Trabajo escrito (manual de usuario, documentación interna) 20%

- Entrega y funcionalidad del sistema 80%
    - Esta parte se evaluará según rúbrica.
    - ☞ *Eventualmente podrá realizarse una prueba de comprobación escrita la cual formaría parte de la nota % de este rubro.*

Entrega.

Cada estudiante deberá entregar el archivo del proyecto realizado a través del TEC DIGITAL en el espacio creado para tal fin en las fechas establecidas de acuerdo al programa del curso.

Los proyectos deberán ser de autoría del estudiante, por lo que en caso de presentarse trabajos con un alto grado de similitud presentados por otro estudiante del actual semestre o de trabajos presentados por alumnos de semestres anteriores, se otorgará cero puntos en la calificación de dicho proyecto, quedando a discreción del profesor, aplicar lo indicado en el artículo 75 del Reglamento del Régimen Enseñanza-Aprendizaje del Instituto Tecnológico de Costa Rica.

El contenido a entregar del proyecto consta de una carpeta comprimida con extensión .zip o .7z. El nombre de la carpeta debe ser: Número de carné y el nombre, PF y apellidos del estudiante Ejemplo: Juan Pérez Pérez, el nombre del archivo quedaría: 201600100101_PF_JuanPerezPerez.

Dentro de la carpeta se debe adjuntar:

Un documento de Word que lleve el nombre del estudiante y PF (ejemplo JuanPerez_PF), en él se debe incluir:

La portada, Introducción, Objetivos, Desarrollo de actividades, donde se incluya la descripción del planteamiento del problema y una descripción técnica del aplicativo, la documentación del código fuente, las imágenes de la salida en pantalla del programa, las conclusiones y la Bibliografía consultada, para todo el documento se debe aplicar las normas APA.

También se debe incluir en la carpeta el programa realizado (código fuente) y manual de usuario, debe estar todo el programa dentro de una sub carpeta llamada Trabajo_final_taller1.

La fecha y hora máxima de entrega del proyecto se deberá realizar de conformidad con lo indicado en el programa del curso (fecha de entrega del proyecto) y **solamente mediante el tecDigital** según se defina en dicha plataforma**.**