Buscador de información en archivos de texto

Curso Estructuras de datos
Especificación de Proyecto

por Aurelio Sanabria

Motivación

El desarrollo de software facilita la automatización de tareas, el manejo y administración de información al mismo tiempo que posee una función modeladora de nuestro entorno. Es por lo anterior que durante este proyecto trabajaremos la aplicación de estas características mediante la creación de una herramienta para acompañar buscar información en grandes volúmenes de datos, en particular dato extraídos del periódico universitario “Hoy en el TEC”.

Objetivos Formativos

El presente proyecto tiene correspondencia con los siguientes objetivos específicos establecidos en el programa del curso de *Estructuras de datos* (presentado al inicio del semestre):

- Implementar soluciones utilizando las diversas estructuras de datos aprendidas
- Elaborar diseños detallados de soluciones basados en estructuras de datos que le permitan resolver problemas de la vida real
- Determinar buenas prácticas de programación para la implementación de estructuras de datos
- Explicar distintas estructuras de datos (lineales, jerárquicas, generales) con el objetivo de la resolución de problemas de mediana complejidad en la implementación y diseño.
- Elaborar estructuras de datos alternas que permitan la resolución de problemas con mayor eficacia, haciendo híbridos de las mismas con el objetivo de crear soluciones más robustas

Especificación del proyecto

El proyecto corresponde con un ejemplo visto en clase y que vamos a ampliar y ejecutar de acuerdo al diagrama de la figura 1

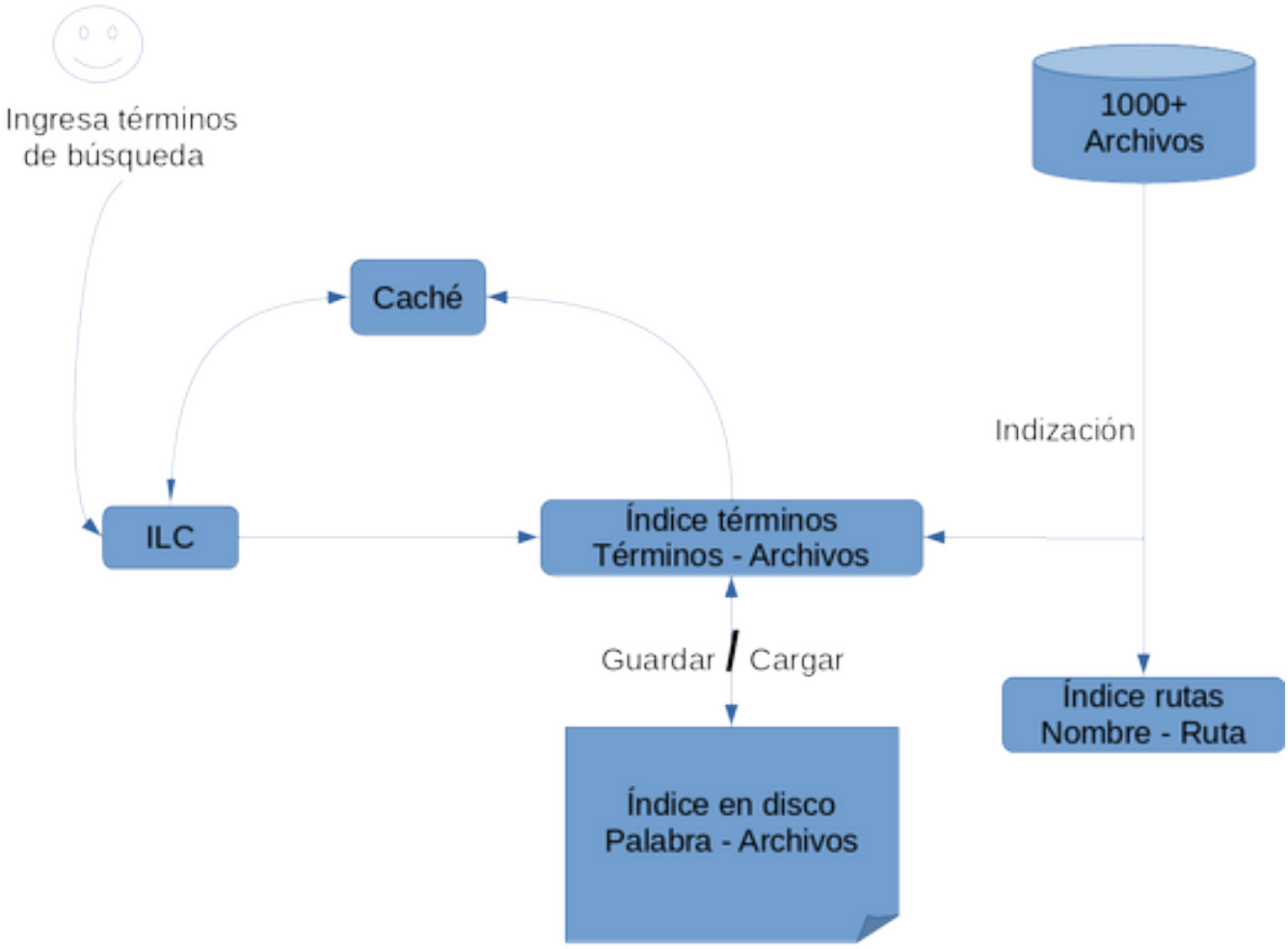

Figura 1: Diagrama de arquitectura de la progra

La persona usuaria introduce uno o más términos de búsqueda mediante la Interfaz de Línea de comandos (ILC). El sistema toma esos términos y los va a buscar el Caché (Árbol chaflaneado), si no hay resultados en el caché se debe ir a buscar en el índice de términos (Árbol B). Si se encuentran los términos en el índice de términos se toma el nombre del archivo y se va busca la ruta del archivo en el índice de rutas (Tabla de dispersión)

Los índices se crean mediante el proceso de indización que se ejecutará una única vez. El índice en disco permitirá guardar el resultado del proceso de indización y evitar repetirlo todas las veces.

A. ILC

Tendrá múltiples opciones dependiendo de la implementación pero como mínimo las siguientes opciones:

- Indizar: Crea el índice.
- Cargar índice: Carga el índice desde el archivo en disco.
- Guardar índice: Guarda el índice en un archivo en disco.
- Buscar: Permite ingresar los términos de búsqueda.
- Salir

B. Caché

El caché se implementará mediante un árbol chaflaneado (splay) en el cuál se debe almacenar, al menos la siguiente información: <***Términos de búsqueda, Lista de nombres de archivo***>

C. Índice de términos

Se implementará mediante un árbol B donde las llaves serán los términos y cada término tendrá asociado una lista de los nombres de los archivos donde aparecen.

D. Índice de rutas

Este se implementa en una tabla de dispersión donde la llave es el nombre del archivo y el valor corresponde a la ruta en disco.

E. Índice en disco

No tiene un formato específico pero debe almacenarse el índice de términos en un archivo para luego poder volver a cargarlo. El formato lo deben decidir dentro del equipo de trabajo.

F. Proceso de indización

Proceso que crea el índice de términos y debe hacer más o menos lo siguiente:

1. Recibir la ruta de donde deben estar los archivos.

2. Insertar en el índice de rutas el nombre del archivo y la ruta en disco.
3. Recorrer el archivo de texto e ir extrayendo los términos para luego insertarlos en índice de término.
    1. Si el término no está en el árbol, insertarlo y colocar la lista de archivos el nombre del archivo actual
    2. Si el término ya existe, agregar a la lista de archivos el nombre del archivo actual.

G. Proceso de búsqueda

1. La persona usuaria introduce uno o más términos de búsqueda mediante la Interfaz de Línea de comandos (ILC).
2. El sistema toma esos términos y los va a buscar el Caché (Árbol chaflaneado).
3. Si no hay resultados en el caché se debe ir a buscar en el índice de términos (Árbol B).
4. Busca los términos en el índice de términos se toma el nombre del archivo y se va busca la ruta del archivo en el índice de rutas (Tabla de dispersión)
5. Retorna una lista de dos columnas Nombre del archivo y ruta

Metodología

Se espera que lxs estudiantes tomen un tiempo para que, de forma creativa, se plantee como podría solucionar el problema utilizando para esto papel y lápiz, para dibujos, esquemas y diagramas que no debe entregar pero que es recomendable hacer.

Implementar las estructuras de datos de forma independiente y verificando que funcionen es altamente recomendable antes de integrarlas al resto del proyecto, adicionalmente se recomiendan pruebas extensivas de todos los componentes del proyecto. Por último es importante aclarar que las *tablas de dispersión* y las *listas* formarán parte del segundo proyecto programado por lo que deben hacerlas de forma que sean re-utilizables.

Es importante aclarar que la dificultad de este proyecto no reside solamente en la creación de estructuras sino también en cómo organizan la información dentro de estas de forma que:

1. Se pueda trabajar sobre estas con cierta facilidad.
2. Almacene todos los datos necesarios.
3. Permita convertir las estructuras de datos a una representación para la persona usuaria.

Se pueden discutir ideas, pros y contras, algoritmos, estrategias, recursos de internet u otros con lxs compañerxs pero NO SE PERMITE EL PLAGIO, si se detecta el plagio se asigna nota de 0 y se pierde la buena voluntad del votán.

Deben implementar la *tabla de dispersión* y al menos *un tipo de lista*

Rúbrica

El proyecto se calificará considerando los rubros expuestos en la tabla 1, cada rubro recibirá una puntuación entre 0 y 5 de acuerdo a lo designado en la Tabla 2. La siguiente tabla detalla algunos elementos esperados y que deben ser considerados para cada rubro.

Tabla 1: Rubros de calificación

Rubro	**Descripción**	**Puntos**
Interfaz de línea de comandos	*Sistema de menús con validaciones y que permita ver todas las opciones.* *Este sistema de menús debe incorporar opciones para imprimir las estructuras de datos en el estado actual (árbol chaflaneado, árbol B y tabla de hash)*	10
Caché	Implementa el árbol chaflaneado, permite: la impresión en preorden, hacer inserciones de nuevos elementos y se autoajusta de acuerdo a lo visto en clase.	15
Indice de términos	Implementa el árbol B, permite la impresión en preorden, la inserción de nuevos elementos y crece de acuerdo a lo visto en clase.	10
Indice de rutas	Re-utiliza la tabla de hash del primer proyecto. La cuál permite: redimencionarse, manejar colisiones, imprimirse, insertar elementos.	10
Indice en disco	*El programa permite escribir y leer un respaldo del índice de términos (Árbol B). El formato del archivo en disco es libre.* *Esto es para permitir que no se tenga que rehacer el árbol B cada vez que se arranca el programa.*	10
Proceso de indización	Utiliza los documentos provistos de texto por el canal de Telegram para la carga y procesamiento de archivos. Este proceso deja el árbol B del índice de términos cargado. Estos pueden estar en una sola carpeta, no hace falta respetar la estructura de carpetas presente en el conjunto de datos que les pasé.	15
Proceso de búsqueda	Permite realizar búsquedas (todo funciona junto) respetando el algoritmo de este documento o alguna variante justificada.	10
Discusión de resultados	Cada persona integrante del equipo debe aportar a la discusión, tener una visión crítica del proyecto y conocer de todas las partes del mismo (aún cuando no las programó es persona).	10
Calidad de código	Reglas de calidad de código y de modularización.	10
		100

Tabla 2: Distribución de puntos por calificación

Excelente	**Bueno**	**Regular**	**Malo**	**Deficiente**	**No existe**
5 puntos	4 puntos	3 puntos	2 puntos	1 punto	0 puntos

Estimación de tiempo

- Este proyecto será en grupos de 1 persona

1. Fecha de entrega: 11 de Noviembre 2022
    - 10:00 p.m GMT-6
    - Al correo aurelio.itcr@gmail.com deben enviar un link a repositorio privado en https://gitlab.com donde tienen el código de la progra.

Aspectos Generales

- Canal de entrega: Se enviarán al correo aurelio.itcr@gmail.com con el asunto el el formato de clase → [2022 IIS][Datos][Proyecto 1] Nombre Apellido
- Los trabajos de entrega tardía no se calificarán.
- Se aclararán dudas sobre la progra vía Telegram.
- Pueden solicitar que en clase sincrónica se reserve un espacio para hablar de la progra.
- Pueden debatir con otros grupos sobre posibles formas de resolver los problemas pero las implementaciones deben ser realizadas de forma independiente entre grupos pues deben estar preparados para defender su trabajo durante una revisión