UNIVERSIDAD DE BUENOS AIRES

FACULTAD DE INGENIERIA

DEPARTAMENTO DE COMPUTACIÓN

75.06 – ORGANIZACIÓN DE DATOS

TRABAJO PRÁCTICO

BOOQUERIO

ETAPA 2

MARZO 2011

Índice General

1 Introducción

Este documento consiste en el enunciado del trabajo práctico de la asignatura. En el mismo se especifican los requerimientos de cada etapa de entrega, dejando de lado el cronograma de entregas que se encuentran en la página o grupo de correo de comunicación de la cátedra respectivamente.

Toda aclaración, indicación o respuesta a consultas (ofrecidas en clase o mediante el grupo yahoo) serán tomadas como extensión y parte explícita de este enunciado.

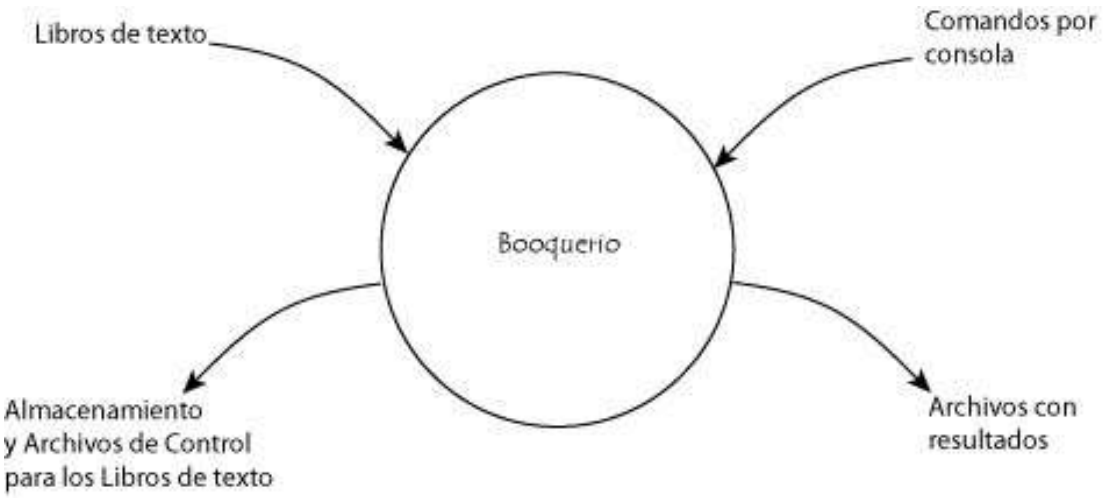
La forma de trabajo con los grupos es descripta en el Reglamento de Trabajos Prácticos de la Cátedra (http://materias.fi.uba.ar/7506C/blog/?page_id=9)

El presente enunciado describe el desarrollo de un sistema de Almacenamiento de Libros, para su consulta y lectura, llamado Booquerio.

2 Enunciado

El sistema se propone resolver el almacenamiento, consulta y obtención de libros de texto. Para lograr este objetivo se hará uso de los contenidos brindados en diferentes módulos de la materia.

La Primer Entrega se avocará a la aplicación de los conceptos de Organización de Archivos.



Libros de Texto = Son los Datos a almacenar y administrar. Se estructuran en formato texto plano, por ello cada grupo debe analizarlos para crear el interprete (parser) adecuado.

Almacenamiento y Archivos de Control para los Libros de Texto = Se encuentran todos dentro de un directorio, especificado a través de un archivo de configuración de la aplicación, y pueden tener jerarquía de subdirectorios interna. Es donde se guarda toda la información necesaria para poder funcionar.

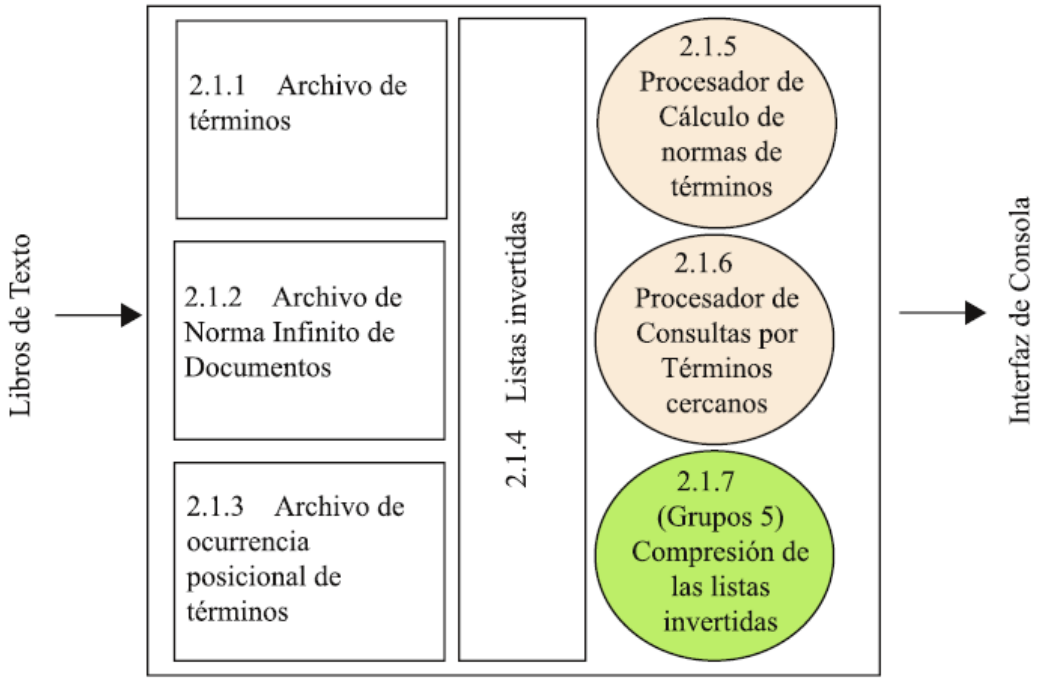
Comandos por consola = Toda la interacción del usuario con el sistema se realiza a través de comandos por consola.

Archivos con resultados = Como respuesta a toda interacción, el sistema generará archivos de respuesta en el directorio donde se llame a la aplicación.

La resolución del trabajo práctico debe ser realizada en plataforma Linux y lenguajes C o C++ (preferentemente respetando el estándar ANSI), y la entrega debe constar de un Makefile para su compilación. Además debe funcionar en calidad de Usuario (user) del sistema operativo.

Segunda Etapa

El objetivo de esta etapa es resolver la aplicación, utilizando las herramientas brindadas por el módulo Recuperación de Textos. Por ello se plantea la siguiente relación entre esos conceptos.



2.1.1 Archivo de términos

Se creará un archivo de términos por orden de aparición.

2.1.2 Archivo de Norma Infinito de Documentos

Un archivo donde se tienen todos los documentos con sus normas infinito.

2.1.3 Archivo de ocurrencia posicional de términos

Se almacenar las tríadas de Identificación de término, Identificación de Documento y posición relativa del término en el documento.

2.1.4 Listas invertidas

Se deberá implementar un modulo cuya funcionalidad sea la construcción de un archivo de listas invertidas apto para consultas por términos cercanos y rankeados.

2.1.5 Procesador de Cálculo de normas de términos

Se implementará un algoritmo para calcular los valores normalizados de los términos para poder comparar los documentos obtenidos. Que va utilizado por el Procesador de Consultas por Términos.

2.1.6 Procesador de Consultas por Términos cercanos

Se implementará un módulo algorítmico para procesar los datos del índice invertido y poder obtener un resultado de consultas por Términos Cercanos.

2.1.7 (Grupos 5) Compresión de las listas invertidas

Se implementará un módulo algorítmico para comprimir con método Gamma todos los valores en las listas invertidas (las distancias).

Entrega 20/06/2011 23:59hs

3 Criterio de aprobación

Como se especifica en el Reglamento de la materia existe un criterio mínimo para poder acceder a una re-entrega en cada etapa. A continuación se menciona una lista de requerimientos que forman parte de dicho criterio. No cumplir con alguna de ellas implica no cumplir el mínimo requerido. Pero no vale la inversa, es decir, cumplir con ellas no implica cumplir con el criterio mínimo.

3.1 Funcionalidad 1ra Etapa

- *Archivo de configuración (con directorio de almacenamiento)*
- *Tomar Texto: ./ejecutable -i "archivo de texto"*
- *Procesar Editorial: ./ejecutable -e (procesa los no procesados)*
- *Procesar Autor: ./ejecutable –a*
- *Procesa Título: ./ejecutable –t*
- *Procesa Palabras: ./ejecutable –p*
- *Listar Archivos Tomados: ./ejecutable –l (muestra identificador, Título, Autor, Editorial y cantidad de palabras registradas para ese libro).*
- *Obtener Archivo: ./ejecutable -o ID_Archivo*
- *Quita Archivo: ./ejecutable -q ID_Archivo (se elminan las entradas en los otros índices)*
- *Ver Estructura: Genera archivos en forma de texto plano, que describen las estructuras y contenidos de los archivos de almacenamiento y control del sistema.*
  - ./ejecutable -v [-e árbol de Editorial, -a árbol de Autor, -t hash de Título, -p hash de Palabra] "Nombre Archivo"
  - Nombre y estructuras para los archivos generados:
    - Archivo de Estructura de control: "Nombre Archivo"_Índice, "Nombre Archivo"_tabla.
    - Archivos de control de espacios libres: "Nombre Archivo"_libres.
    - Archivos de bloques de datos: "Nombre Archivo"_datos.
    - Estructura: para árboles la indicada en teórica. Separadores: Bloques con "|", Registros con ";", atributos con ",".

3.2 Funcionalidad 2da Etapa

- *Consultar Editorial: ./ejecutable -qe "Editorial"*
- *Consultar Autor: ./ejecutable –qa "Autor"*
- *Consultar Título: ./ejecutable –qt "Título"*
- *Consultar Palabras: ./ejecutable –qp "Palabras para búsqueda por cercania y rankeada"*
- *Ver Estructura: Genera archivos en forma de texto plano, que describen las estructuras y contenidos de los archivos de almacenamiento y control del sistema.*

- ./ejecutable -v [-at Archivo de Términos, -ani Archivo de Norma Infinito, -aop Archivo de ocurrencia posicional, -li Listas Invertidas] "Nombre Archivo"
- Nombre y estructuras para los archivos generados:
  - Archivo de Términos: "Nombre Archivo"_Terminos
  - Archivo de Norma Infinito: "Nombre Archivo"_NormaInfinito.
  - Archivo de ocurrencia posicional: "Nombre_Archivo"_OcurrenciaPosicional.
  - Listas Invertidas: "Nombre_Archivo"_ListasInvertidas.
  - Separadores: Bloques con "|", Registros con ";", atributos con ",".

3.3 Documentación

- *General*
  - Diagrama de clases o módulos (según corresponda)
  - Especificación de cada clase o módulo (según corresponda)
  - Diagramas de secuencia o intercambio de mensajes entre capas. Mostrar escenarios.
  - Planificación (identificación de tareas, estimación de duración y asignación)
  - Bugs conocidos
  - Manual de usuario. Indicaciones generales del trabajo práctico, modo de instalación y ejemplos de uso.

- *Física – Organización*

  Organización de registros

  - ¿Cómo delimitan la longitud de un registro y de un campo variable?. Mostrar los campos que posee y cuanto espacio ocupa cada uno.
  - Indicar que información administrativa se utiliza.

- *Índices – Búsqueda*

  Datos de control
  - Datos que se agregan a los archivos de índices para poder dar respuesta a las consultas.

  Árbol B
  - Pseudocódigo del proceso de creación del índice.
  - Pseudocódigo del proceso de consulta.

4 Referencias

Managing Gigabytes: Compressing and Indexing Documents and Images, Second Edition (The Morgan Kaufmann Series in Multimedia Information and Systems)