DEPARTAMENTO
INGENIERÍA INFORMÁTICA
Y CIENCIAS DE LA COMPUTACIÓN
FACULTAD DE INGENIERÍA UNIVERSIDAD DE CONCEPCIÓN

Estructuras de Datos y Algoritmos Avanzados (2021-2)

Laboratorio 5

Profesor: Diego Seco

Ayudante: Alexander Irribarra

Objetivos

Los objetivos del laboratorio son:

- Implementar y entender el código de estructuras de datos de complejidad media para IR.
- Evaluar experimentalmente estructuras de datos y algoritmos sobre IR.

Ejercicios

1. Implementar un índice invertido orientado a documentos, es decir, la lista de ocurrencia de una palabra W almacena $\{x, y, z, ...\}$ cuando la palabra W aparece en los documentos x, y, z, etc.
2. Las listas de ocurrencias se codificarán utilizando la biblioteca `SDSL`, en concreto con un `int_vector` y con un `enc_vector`. Como codificador[1] para este último, se puede utilizar `elias_gamma` o `elias_delta`. Describir brevemente cómo funcionan.
3. Implementar el vocabulario[2] utilizando un `map<string, TIPO_LISTA>` de la librería estandar y resolver búsquedas utilizando *merge*, es decir, recorriendo secuencialmente ambas listas para generar la intersección. Las búsquedas corresponden al caso particular de la intersección de 2 palabras, la que retorna una lista de documentos que contiene a ambas palabras.

[1] Notar que en la template de `enc_vector` el primer argumento correponde a la clase del codificador a utilizar.

[2] Dado que la implementación del índice invertido va a ser independiente de la estructura utilizada para codificar la lista, es recomendable implementar su clase como una *template*.

4. Evaluar utilizando el documento **english.100MB** del corpus Pizza&Chili, dividiendo el texto en 1000 partes de tal manera que cada parte corresponda a un documento. Para simplificar esto, se puede contar la cantidad total de caracteres C y dividir de tal manera que cada documento tenga $\lfloor C/1000 \rfloor$ caracteres.

5. Medir y tabular el tiempo promedio de la consulta de intersección de las 3 implementaciones.

Normas de entrega

Antes del viernes 3 de diciembre se deben enviar todos los ejercicios resueltos a través de CANVAS.

Se deben subir **dos archivos separados**:

- Archivo PDF con el nombre completo, las respuestas a las preguntas que correspondan y capturas de pantalla mostrando brevemente la ejecución de sus códigos.
- Un archivo comprimido que contenga los ficheros del código fuente (formato .zip, .gz, etc.).
- **IMPORTANTE:** el archivo debe llamarse *apellido1_nombre_05*.