Proyecto Semestral, Algoritmos II

Título: Generación de una base de datos de documentos.

(equipos de 4 estudiantes solamente)

Objetivos Generales

- Desarrollar una aplicación para manejar un conjunto de documentos utilizando una estructura de datos previamente definida en la misma aplicación.
- Desarrollar mecanismos eficientes de consultas sobre los documentos de la base de datos.

Requerimientos

- Lograr el cumplimiento de los objetivos a través de una aplicación (script) utilizando el lenguaje de programación python3.
- A través de la aplicación desarrollada, permitir la creación de una estructura de datos en una dirección local (un directorio del propio ordenador) el cual contiene conjunto de documentos.
- Para la creación de la estructura se utilizará el siguiente comando: **`python document_db.py -create <local_path>`**
- Una vez cargados los documentos en la aplicación, permitir realizar consultas sobre el contenido de los documentos.
- Para la generación de consultas se utilizará el siguiente comando: **`python document_db.py -search <text>`**
- Para el desarrollo de la aplicación **solamente queda permitido el uso** de algunas bibliotecas o funciones discutidas en clases, y en las diapositivas que explican la lectura y escritura de datos en disco. El resto de las estructuras utilizadas deben ser exclusivamente implementadas por el equipo de trabajo.
- Garantizar la persistencia de los datos. Esto significa que la estructura que compone la base de datos de documentos tiene que ser recuperable a través de consultas en todo momento.
- Los equipos de trabajo deben estar compuestos por 4 estudiantes. No se permiten trabajos individuales y en caso de que el número total de estudiantes sea mayor o menor es necesario avisar con previo inicio del proyecto para su consideración.

Evaluación del proyecto

- Para la evaluación del proyecto entra en consideración los siguientes factores:
    1. Perfecto entendimiento de cada integrante del equipo de todo el código del proyecto.
    2. Perfecto entendimiento de cada integrante del equipo de los problemas surgidos y soluciones generadas durante toda la fase de desarrollo de la aplicación.

3. Correcto funcionamiento de la aplicación acorde a los objetivos planteados.
4. Claridad y documentación del código.
5. Correcta elección de las estructuras de datos y algoritmos utilizados.
6. Eficiencia de la aplicación relacionada al costo temporal y espacial.

Creación de la Base de Datos de Documentos

- Para la creación de la base de datos se utilizará el siguiente comando: **`python document_db.py -create <local_path>`**
- **`<local_path>`** representa la dirección local de la carpeta que contiene los documentos de la base de datos que deberán ser procesados. En otras palabras, los documentos que componen la base de datos ya se encuentran en una carpeta, y **`<local_path>`** es la ruta para llegar a dicha carpeta.
- Una vez finalizado el proceso de creación de la base de datos de documentos la aplicación devolverá el texto ***"document data-base created successfully"***. A partir de este momento se pueden iniciar las búsquedas.
- La base de datos deberá persistir la información de manera que se pueda acceder a su información en todo momento. Esto significa que no se deberá volver a crear tal índice en cada búsqueda, sino que se realizará sobre una estructura persistente en disco, que se levantará a memoria cada vez que se requiera hacer una consulta.

Búsquedas de documentos

- La búsqueda de documentos se va a realizar a través de texto. Para ellos se utilizará el comando: **`python document_db.py -search <text>.`**
- El resultado de una búsqueda a través de un texto (**`<text>`**) va a devolver todos los títulos de los documentos (nombre del archivo, o ruta) que hablan del mismo tema al cual se refiere el texto. Esto quiere decir, el contenido interno del archivo correspondiente a un documento de la base de datos (no solo su nombre). Los resultados deberán devolverse ordenados por relevancia.
- La relevancia se calcula por el valor de cercanía del texto al documento. Los documentos con mayor relevancia irán primero en el resultado de la búsqueda.
- En caso de no existir ningún documento en la base de datos que contenga relación con el texto devolverá la salida: ***"document not found".***

Estructura de la Aplicación a realizar

- Se implementará un script en python utilizando la versión 3. El script tendrá el nombre **`document_db.py.`** Sobre ese script se realizarán las operaciones de creación y búsqueda. El manejo de errores, excepciones y posibles valores de entrada corren a cargo de los desarrolladores de la aplicación. Dicho script será utilizado para realizar las pruebas para evaluar el desempeño de la aplicación.