PRÁCTICA OBLIGATORIA EVALUABLE

METADATOS CON APACHE TIKA

Objetivo

El objetivo es poner en práctica diferentes conceptos aprendidos durante el módulo y relacionados con la unidad de metadatos.

Normativa de entrega

La entrega de esta práctica es obligatoria se realiza individualmente. A continuación, se detallan otros datos de interés relacionados con la normativa de entrega:

- La fecha límite de entrega de la práctica será el día indicado en el campus virtual.
- La entrega de la práctica se hará a través de Aula Virtual, empleando la actividad habilitada para ello.
- Se deberá subir un único archivo comprimido con todo el código fuente que se haya implementado y se necesite para ejecutar la práctica, así como una memoria explicando qué se ha hecho y por qué.
- El archivo debe ser nombrado de la siguiente manera: UAXAYZNombre/s. Donde:
  - X es el número de la Unidad
  - Y equivale a "I" si se ha hecho Individual o "C" si se ha hecho Colectiva
  - Z equivale al número de práctica de la unidad. Nombre/s es vuestro nombre/s completo/s. Ejemplo: UA1AI1RafaelMuñoz

Enunciado

El/ La alumno/a tiene que realizar una **extracción** de metadatos, **tratarlos** y **mostrarlos** haciendo uso de una serie de funciones de la librería Tika.

Objetivos Generales

Se proponen los siguientes ejercicios para el procesamiento de documentos con Tika.

1. Crear una función (se recomienda el nombre: TikaReader) que permita llevar a cabo las cuatro operaciones básicas de Tika en Python. Estas cuatro operaciones deben ir encapsuladas en un método propio de la clase. Hacer una función donde se debe pasar como parámetro el documento a procesar (se recomienda utilizar un pdf a vuestra elección). Hacer otra función que debe cargar el fichero jar del cliente de Tika.
   Ayuda:
   Librería necesaria para este desarrollo
   from tikapp import TikaApp
   Operaciones básicas de Tika
   detect_content_type(<fichero>)

detect_language(<fichero>)
extract_all_content((<fichero>, convert_to_obj=<True/False>)
extract_only_content((<fichero>)

2. Crear una función (se recomienda el nombre: ProcessJSONTika) para procesar el JSON proporcionado por Tika en el método:
   extract_all_content(<fichero>, convert_to_obj=<True/False>)
   Esta clase debe obtener los siguientes métodos:
   A) Dado el JSON de procesamiento Tika escribir por consola los metadatos sobre autores, fecha de creación y fecha de modificación.
   B) Obtener el contenido (solo texto) del fichero JSON proporcionado por Tika.

   Ayuda: Se recomienda reutilizar el lector de Tika creado en el ejercicio anterior. El atributo **convert_to_obj=True** es la manera de obtener el resultado en formato JSON. Los campos del fichero JSON para el apartado A son: 'Author', 'Creation-Date' y 'Last-Modified'. El campo del fichero JSON para el apartado B es: 'X-TIKA:content'.

Memoria.

La resolución de la siguiente práctica consiste en la utilización de Python para extraer los metadatos a un documento y mostrarlos. El/La alumno/a realizará una labor de investigación y aprendizaje sobre la librería, poniendo en práctica de una asignatura de proyectos los conocimientos adquiridos en la asignatura de POD 1 y Fundamentos de Programación.
La memoria deberá tener la estructura de Introducción, Desarrollo y Conclusiones. El nivel de explicación y profundidad en cada uno de dichos apartados determinará la nota de la memoria. La nota de la práctica la determinará el manejo que se demuestre de la librería Tika integrada con Python. La estructuración del código Python (uso de funciones y recursos aprendidos en POD 1), limpieza del código y comentarios explicativos también son evaluables.

Instalación de software necesario

Para la realización de la práctica vamos a hacer uso de la librería Tika. Realizar la instalación siguiendo el tutorial que proporciona la documentación en la teoría.