Prácticas - Recuperación de Información

Práctica 2.2 - Tika

Iván Calle Gil Daniel López García Lothar Soto Palma
José Carlos Entrena Jiménez

Índice

1. Introducción

En esta práctica vamos a usar la herramienta *Tika*, que nos permite obtener numerosa información de todo tipo de archivos (PDF, HTML, CSV). Podremos obtener todos los metadatos y el texto que contienen en distintos formatos, datos que podremos utilizar para saber qué tipo de archivo estamos manejando y realizar operaciones específicas sobre ellos.

Por ejemplo, podemos obtener todo el texto que hay en un archivo HTML eliminando todas las cabeceras, y a partir de aquí trabajar con ese texto, o también podemos extrae información importante directamente, como los links que el archivo contiene, cosa que haremos en esta práctica.

Combinaremos las acciones de *Tika* y *Snowball* para poder lexificar el texto que encontremos en los archivos, tarea que se nos pide en esta práctica. Para ello, primero extraeremos el texto de un archivo e identificaremos su idioma utilizando *Tika*, para seguidamente utilizar un *stemmer* adecuado para lexificar dicho texto según el idioma correspondiente.

Además de lo mencionado anteriormente, en esta práctica extraeremos el conjunto de links de cada archivo que analicemos y realizaremos una tabla que contiene el tipo de cada archivo analizado y su codificación. Los resultados obtenidos serán almacenados de la siguiente forma:

- Para los enlaces extraídos, se creará un archivo Links.txt en la carpeta data/, que se sobrescribirá en cada ejecución del programa. Este archivo contendrá los nombres de los archivos analizados y, bajo ellos, todos los enlaces encontrados.
- La tabla que contiene el nombre de los ficheros analizados, su tipo y su codificación, será almacenada en un archivo CSV con el nombre Tabla_archivos.csv, también bajo la carpeta data/. Nuevamente, esta tabla será sobrescrita si se vuelve a ejecutar el programa.
- La lexificación del texto de los archivos se almacena bajo una subcarpeta stems/ dentro de data/, en el caso en el que tengamos más de un archivo. El resultado será un archivo con el mismo nombre que aquel que se ha lexificado. Para esto, como hemos explicado anteriormente, eliminaremos los datos irrelevantes del archivo (etiquetas HTML o similar) y extraeremos el texto plano, del cual comprobaremos el idioma y usaremos Snowball para el proceso de lexificación.
- La carpeta data/, donde se guardan los resultados, la crea el programa a no ser que ya exista.

2. Desarrollo de la práctica

2.1. Implementaciones realizadas

Para llevar a cabo esta práctica, hemos trabajado sobre el código que realizamos en la parte anterior con *Snowball*, de forma que el grueso del proceso de lexificación se encontraba ya implementado, teniendo únicamente que distinguir entre idiomas para usar el *stemmer* aducuado.

Además, hemos añadido una clase *TextParser* que contiene los métodos principales de *Tika* para la obtención de los datos necesarios. Estos son los siguientes:

- Método para la obtención del idioma de un texto, de nombre *identifyLanguage*, que usa la clase *LanguageDetector*.
- Métodos para la extracción del texto y de los metadatos de un archivo genérico, usando *BodyContentHandler*.
- Métodos para la obtención y escritura de los links que hay en un archivo, usando *LinkContentHandler*. Primero, obtenemos la lista de links para después escribirlos. Hemos separado ambos métodos para una mayor flexibilidad de cara al futuro.
- Hemos modificado la indexación del texto para aplicar el *stemmer* correspondiente, añadiendo una comprobación del lenguaje. Además, ahora creamos un archivo donde se guarda el texto lexificado.

- Hemos implementado distintos métodos para la creación de archivos de resultados o directorios que contengan estos archivos, como el método *generarResultados* en la clase *textIndexer*. o el método *writeFileTable* que genera la tabla con el archivo, el tipo y la codificación en la clase *FileIO*. No es necesario crear ningún tipo de archivo o directorio para el correcto funcionamiento del código y la obtención de resultados.

2.2. Proceso

Para realizar esta práctica, hemos buscado los datos que necesitábamos en la documentación de *Tika* para saber qué necesitábamos y cómo conseguirlo, buscando familiarizarnos con las clases necesarias y los métodos que estas poseen. Los miembros del grupo hemos elegido partes de las tareas a realizar y hemos implementado el código correspondiente, tratando de hacerlo legible y reutilizable de cara a futuras prácticas.

Hemos tenido alguna complicación en la obtención de rutas y creación de ficheros/directorios, debido a las diferencias que existen entre distintos sistemas operativos. Para esto, hemos incluído algunos métodos *replace* que sustituyen las barras '\' y '/', haciendo que todo funcione correctamente sea cual sea el sistema operativo utilizado.

2.3. Resultados

Creemos que hemos cumplido los objetivos de esta práctica de forma satisfactoria, aprovechando los recursos que *Tika* pone a nuestra disposición para realizar un análisis de documentos que, si bien no es excesivamente profundo, si nos permite trabajar de forma más eficiente y productiva con los archivos que tratemos.

3. Trabajo en grupo

- Lothar Soto: Ha sido el encargado de la obtención del tipo de archivo y la codificación, de extraer el texto plano de un archivo y de la creación de la tabla correspondiente, además de realizar cambios en la creación de archivos de resultados y solucionar errores varios.
- Iván Calle: Ha modificado el proceso de lexificación, adaptando todo el proceso, además de implementar los métodos de creación de archivos de resultados para esta parte y adaptar los *path* para que funcionen en cualquier SO.
- Daniel García: Se ha encargado de la obtención de los links de un fichero, de la implementación de los métodos para recopilar metadatos y la detección de idioma.
- José Carlos Entrena: Ha llevado a cabo la implementación del método de escritura de los links en un fichero, la elaboración del documento de entrega, revisión y reestructuración del código y la gestión de los métodos de escritura de datos en archivos.

Cabe destacar que la mayoría de las dificultades que han surgido a lo largo de la implementación han sido discutidas en grupo y luego llevadas a cabo por uno de los miembros, junto a la discusión de entrada y salida de datos, estructura del código y demás aspectos relevantes de la práctica. Es por esto que consideramos que todos los miembros del grupo han tenido una aportación similar y, sobre todo, activa, en la realización de esta práctica.