*Recuperación de Información*

Practica 2: Parser de documentos con TIKA



**Autores: García Maldonado, Fco Javier**

**Correo: franelas@correo.ugr.es**

**DNI: 76654015Y**

**1.- Introducción**

En esta práctica hemos realizado un trabajo de análisis de documentos básico. Este análisis consistía en procesar varios archivos ubicados en un directorio de los cuales se extraen los metadatos en primer lugar y en segundo y último lugar se procesan el número de términos en el archivo y la cantidad que hay de cada uno de ellos. Cabe destacar que a estos términos les hemos aplicado un pre-procesado antes de iniciar el conteo, ya que si no realizamos este pre-procesado el documento generado con los términos y la cantidad de veces que aparecen contendría errores al contar como términos independientes aquellas palabras que tengan signos de puntuación antes, después y en mitad de la palabra.

**2.- Procesado de documentos con TIKA**

Tal y como hemos indicado en el apartado anterior, en esta práctica hemos realizado un procesamiento de archivos con TIKA. En primer lugar, nos hemos centrado en extraer el idioma en el que está escrito el texto. A continuación, una vez extraído el idioma, pasamos a obtener los metadatos del archivo tales como: codificación, nombre del documento y tipo de documento. Por último, pero no por ello menos importante, lo que hemos realizado ha sido un conteo de las diferentes palabras que hay en cada documento y el número de veces que aparece cada una de ellas, todo ello apoyándonos en la estructura de datos HashMap. Una vez que tenemos todas las palabras almacenadas en nuestra ED, necesitamos que estas estén en orden decreciente en función de la frecuencia de aparición de las palabras. Por lo que una palabra que aparezca más que ninguna debe estar la primera.

El problema llegado a este punto es que el HashMap no tiene función de ordenación, ni se almacena ordenado, por lo que debemos recurrir a otra ED diferente, esta ED elegida ha sido el ArrayList. Con ella podemos volver a introducir los datos y ordenarlos, para así poder imprimir en un documento de texto las palabras junto con su frecuencia por orden decreciente.

Cabe destacar que para nuestro proyecto hemos utilizado como librería TIKA. Y que los libros son todos obtenidos del proyecto Gutenberg, los cuales están en diferentes idiomas, desde italiano, inglés y francés hasta islandés o incluso húngaro.

**3.- Verificación de la Ley de Zipf**

En este apartado trataremos de verificar que se cumple esta ley. Esta ley viene dada por la ecuación de Booth y Federowicz, y se expresa de la siguiente forma:

Donde F representa la frecuencia, R la posición en la ordenación que hemos realizaco y por último k y m son constantes. Para obtener dichas constantes podemos hacerlo a partir del grafico log-log teniendo en cuenta que:

Por tanto, si realizamos sobre el grafico log-log un ajuste lineal, podremos obtener dichas constantes k y m de forma sencilla.