UNIVERSIDAD DE GRANADA

Recuperación de Información

Práctica 1

Extracción de contenido con Tika

David Carrasco Chicharro

1 Introducción

En la entrega se presenta un directorio principal (`p1_david_carrasco_chicharro`) que contiene un directorio con los archivos del código fuente (`/extractorinformacion`) y un script (`ejecucionP1.sh`) para ejecutar el programa, además del archivo .jar de Tika. La carpeta con los documentos a procesar ha de situarse en el directorio principal, que en este caso se entrega como `/docs`, y que a su vez contiene la carpeta `/contador` con los archivos .csv de frecuencias.

2 Ejecución

Para ejecutar el programa hay que introducir el siguiente comando:

`./ejecucionP1.sh argumento(s) directorio`

Argumentos:

- `-d`: metadatos del documento
- `-l`: enlaces del documento
- `-t`: imprime CSV con ocurrencias de cada término

Ejemplo de uso:

- Mostrar metadatos y enlaces: `./ejecucionP1.sh -d -l dir`
- Mostrar metadatos y contar frecuencias: `./ejecucionP1.sh -d -t directorio`

Se puede mostrar una ayuda ejecutando: `./ejecucionP1.sh --help`

```
david@david:~/Universidad/RI/Prácticas/P1/p1_david_carrasco_chicharro$ ./ejecucionP1.sh -d -l -t docs
Nombre: de_la_tierra_a_la_luna-Julio_Verne.pdf
Tipo: application/pdf
Codificación: UTF-8
Idioma: es

Enlaces del documento: de_la_tierra_a_la_luna-Julio_Verne.pdf
http://www.biblioteca.org.ar/
http://www.biblioteca.org.ar/voluntariosform.htm
http://www.biblioteca.org.ar/donac.htm
http://www.biblioteca.org.ar/
http://www.biblioteca.org.ar/comentario/
de_la_tierra_a_la_luna-Julio_Verne.pdf -> ./largos/contador/contador_8.csv

--------------------------------------

Nombre: the_tragedy_of_romeo_and_juliet.pdf
Tipo: application/pdf
Codificación: UTF-8
Idioma: en

Enlaces del documento: the_tragedy_of_romeo_and_juliet.pdf
El documento no contiene enlaces.
the_tragedy_of_romeo_and_juliet.pdf -> ./largos/contador/contador_9.csv
```

3 Salidas

Si se desean obtener los metadatos y/o los enlaces de los documentos, estos se mostrarán en el terminal, como se aprecia en la imagen de la ejecución. En el caso del contador de frecuencias se creará una carpeta dentro del directorio que contiene los documentos a analizar llamada contador en cuyo interior se almacenarán los ficheros CSV; para saber a qué documento pertenece cada una de las salidas se mostrará la información en el terminal con un mensaje del tipo:

```
quijote.epub -> ./docs/contador/contador_1.csv
```

	A	B
1	Text	Size
2	le	456
3	de	432
4	je	321
5	et	296
6	il	263
7	les	254
8	un	233
9	la	222
10	petit	200
11	pas	181
12	à	178
13	prince	172
14	ne	169
15	que	165
16	mais	136
17	tu	136
18	des	131
19	c'est	130
20	dit	125
21	une	123
22	me	103
23	qui	102

	A	B
1	Text	Size
2	and	708
3	the	679
4	i	575
5	to	538
6	a	460
7	of	401
8	my	359
9	that	347
10	is	344
11	in	319
12	you	290
13	thou	277
14	me	263
15	not	258
16	with	252
17	it	226
18	this	226
19	for	223
20	be	215
21	but	183
22	what	165
23	thy	164

	A	B
1	Text	Size
2	und	2893
3	die	1606
4	sie	1438
5	er	1365
6	der	1253
7	in	1183
8	sich	1134
9	mit	947
10	zu	898
11	das	884
12	es	866
13	den	797
14	nicht	774
15	auf	738
16	ein	716
17	ich	625
18	wie	593
19	an	518
20	ihm	517
21	lande	517
22	daß	516
23	dem	481

Frecuencia de términos. De izquierda a derecha: “*Le Petit Prince*”, “*The Tragicomedy of Romeo and Juliet*” y “*Der Tod des Iwan Lande*”

Figura 1: Nube de palabras de "El Quijote" generada con https://wordart.com/