

Recuperación de la Información  
Curso 2020/2021



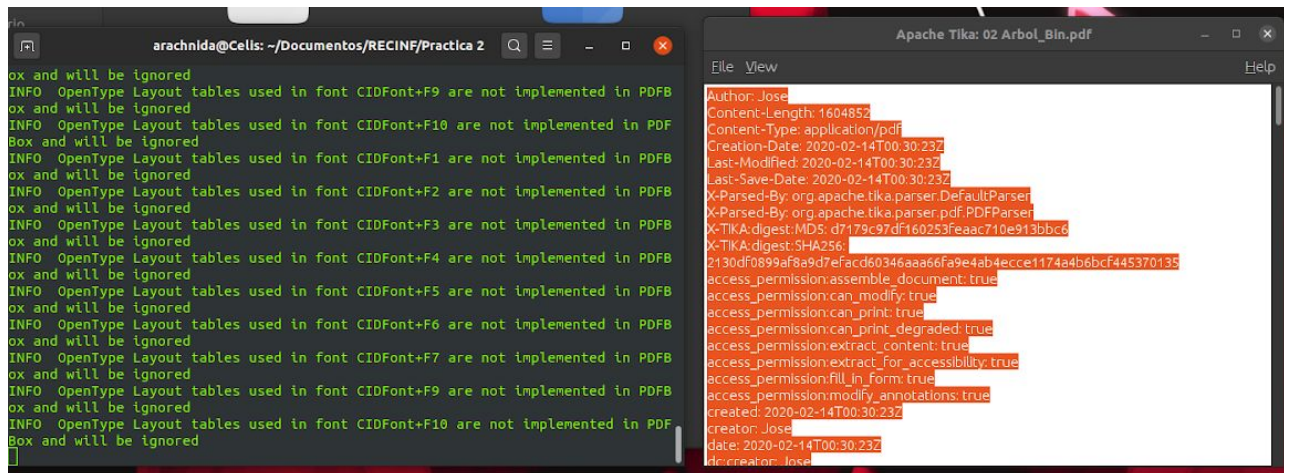
Práctica 2: Tika  
Carmen del Mar Ruiz de Celis

## 1. Abrir un archivo pdf en el modo ventana y decir quién es el autor del archivo, si tiene.

Primero, lanzo la aplicación con el comando:

```
java -jar tika-app-1.22.jar
```

A continuación, pulso sobre: File > Open y selecciono un documento .pdf cualquiera. En este caso, he elegido uno sobre árboles binarios de la asignatura EDNL.



Como se puede observar, en la primera obtenemos el autor del archivo, el cual es “Jose”.

## 2. Almacenar el contenido de un archivo .pdf en un archivo .doc.

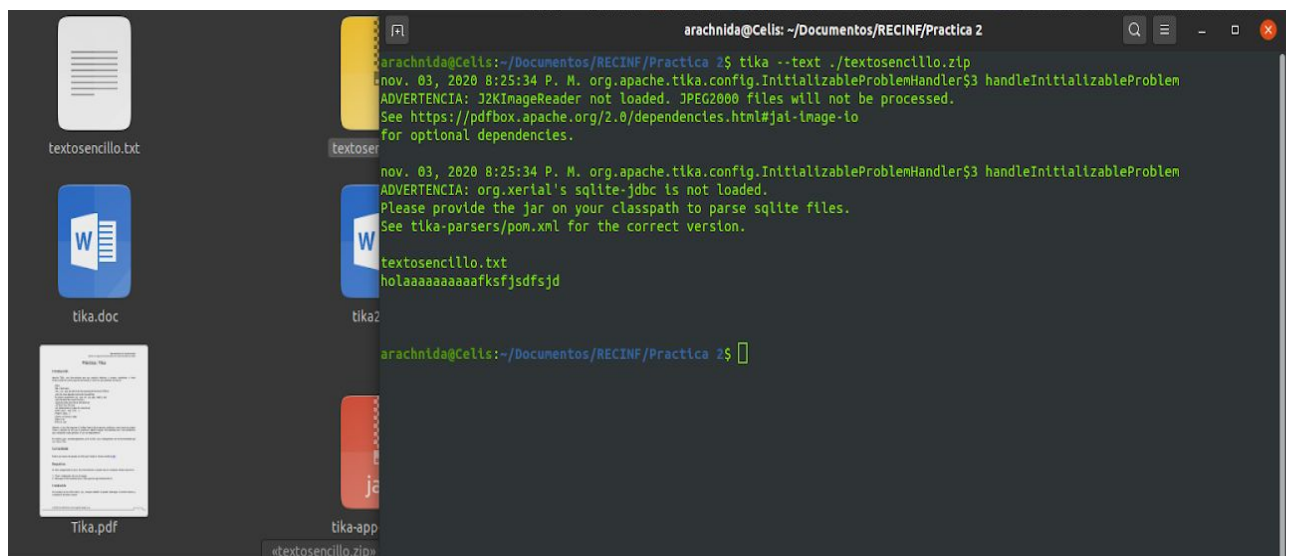
```
tika ./Tika.pdf > tika.doc
```

### 3. Ver los metadatos de un archivo que esté subido en una página web.

```
tika --metadata  
https://esingenieria.uca.es/wp-content/uploads/2014/03/Pagina-Principal-Trabajos-Fin-de-Grado-y-Master.jpg
```

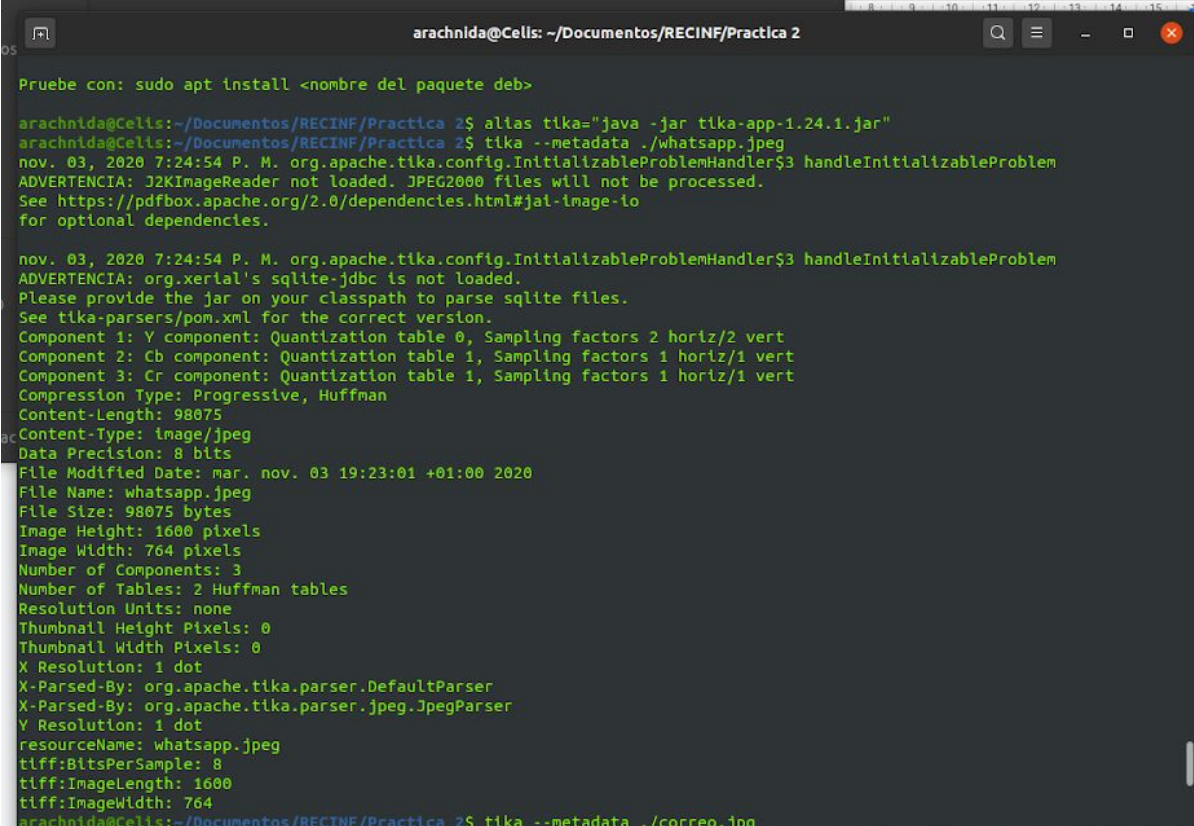
### 4. Comprimir un archivo de texto sencillo y abrir con Tika el archivo comprimido.

```
tika --text ./textosencillo.zip
```



## 5. Pasar por correo y Whatsapp una foto y comparar los metadatos comprobando las diferencias.

- Metadatos de la imagen de whatsapp:



```
arachnida@Celis: ~/Documentos/RECINF/Practica 2
Pruebe con: sudo apt install <nombre del paquete deb>

arachnida@Celis:~/Documentos/RECINF/Practica 2$ alias tika="java -jar tika-app-1.24.1.jar"
arachnida@Celis:~/Documentos/RECINF/Practica 2$ tika --metadata ./whatsapp.jpeg
nov. 03, 2020 7:24:54 P. M. org.apache.tika.config.InitializableProblemHandler$3 handleInitializableProblem
ADVERTENCIA: J2KImageReader not loaded. JPEG2000 files will not be processed.
See https://pdfbox.apache.org/2.0/dependencies.html#jai-image-io
for optional dependencies.

nov. 03, 2020 7:24:54 P. M. org.apache.tika.config.InitializableProblemHandler$3 handleInitializableProblem
ADVERTENCIA: org.xerial's sqlite-jdbc is not loaded.
Please provide the jar on your classpath to parse sqlite files.
See tika-parsers/pom.xml for the correct version.
Component 1: Y component: Quantization table 0, Sampling factors 2 horiz/2 vert
Component 2: Cb component: Quantization table 1, Sampling factors 1 horiz/1 vert
Component 3: Cr component: Quantization table 1, Sampling factors 1 horiz/1 vert
Compression Type: Progressive, Huffman
Content-Length: 98075
Content-Type: image/jpeg
Data Precision: 8 bits
File Modified Date: mar. nov. 03 19:23:01 +01:00 2020
File Name: whatsapp.jpeg
File Size: 98075 bytes
Image Height: 1600 pixels
Image Width: 764 pixels
Number of Components: 3
Number of Tables: 2 Huffman tables
Resolution Units: none
Thumbnail Height Pixels: 0
Thumbnail Width Pixels: 0
X Resolution: 1 dot
X-Parsed-By: org.apache.tika.parser.DefaultParser
X-Parsed-By: org.apache.tika.parser.jpeg.JpegParser
Y Resolution: 1 dot
resourceName: whatsapp.jpeg
tiff:BitsPerSample: 8
tiff:ImageLength: 1600
tiff:ImageWidth: 764
arachnida@Celis:~/Documentos/RECINF/Practica 2$ tika --metadata ./correo.jpg
```

- Metadatos de la imagen por correo:

```
arachnid@Cells: ~/Documentos/RECINF/Practica 2
resourceName: whatsapp.jpeg
tiff:BitsPerSample: 8
tiff:ImageLength: 1600
tiff:ImageWidth: 764
arachnid@Cells:~/Documentos/RECINF/Practica 2$ tika --metadata ./correo.jpg
nov. 03, 2020 7:25:58 P. M. org.apache.tika.config.InitializableProblemHandler$3 handleInitializableProblem
ADVERTENCIA: J2KImageReader not loaded. JPEG2000 files will not be processed.
See https://pdfbox.apache.org/2.0/dependencies.html#jai-image-io
for optional dependencies.

nov. 03, 2020 7:25:58 P. M. org.apache.tika.config.InitializableProblemHandler$3 handleInitializableProblem
ADVERTENCIA: org.xerial's sqlite-jdbc is not loaded.
Please provide the jar on your classpath to parse sqlite files.
See tika-parsers/pom.xml for the correct version.
Component 1: Y component: Quantization table 0, Sampling factors 2 horiz/2 vert
Component 2: Cb component: Quantization table 1, Sampling factors 1 horiz/1 vert
Component 3: Cr component: Quantization table 1, Sampling factors 1 horiz/1 vert
Compression Type: Progressive, Huffman
Content-Length: 556639
Content-Type: image/jpeg
Data Precision: 8 bits
File Modified Date: mar. nov. 03 19:22:26 +01:00 2020
File Name: correo.jpg
File Size: 556639 bytes
Image Height: 2261 pixels
Image Width: 1080 pixels
Number of Components: 3
Number of Tables: 2 Huffman tables
Resolution Units: none
Thumbnail Height Pixels: 0
Thumbnail Width Pixels: 0
X Resolution: 1 dot
X-Parsed-By: org.apache.tika.parser.DefaultParser
X-Parsed-By: org.apache.tika.parser.jpeg.JpegParser
Y Resolution: 1 dot
resourceName: correo.jpg
tiff:BitsPerSample: 8
tiff:ImageLength: 2261
tiff:ImageWidth: 1080
arachnid@Cells:~/Documentos/RECINF/Practica 2$
```

- Diferencias:

Content-length/File size, Image-length e ImageWidth son menores en la imagen de whatsapp, ya que este las comprime con objeto de ahorrar datos móviles, pero en el correo se envía la foto intacta.

## 6. Ver los metadatos de <http://www.uca.es/es/> y guardarlo en un archivo.txt.

La URL es incorrecta y me devuelve una excepción, así que lo he sustituido por https.

```
tika --metadata https://www.uca.es/es/ >
ejercicio6.txt
```



De esta forma no hay ningún error, pero en los metadatos indica como título “302 Found”, lo que indica que la URL ha sido movida a la indicada en la cabecera Location. Se puede comprobar al introducirla en un navegador: nos redirigirá a <https://www.uca.es/>.

## 7. Pasar un archivo .rdf a .doc. ¿Pasar el archivo .doc o .rdf al formato .pdf dará error?

Todas las conversiones se realizan correctamente en la terminal:

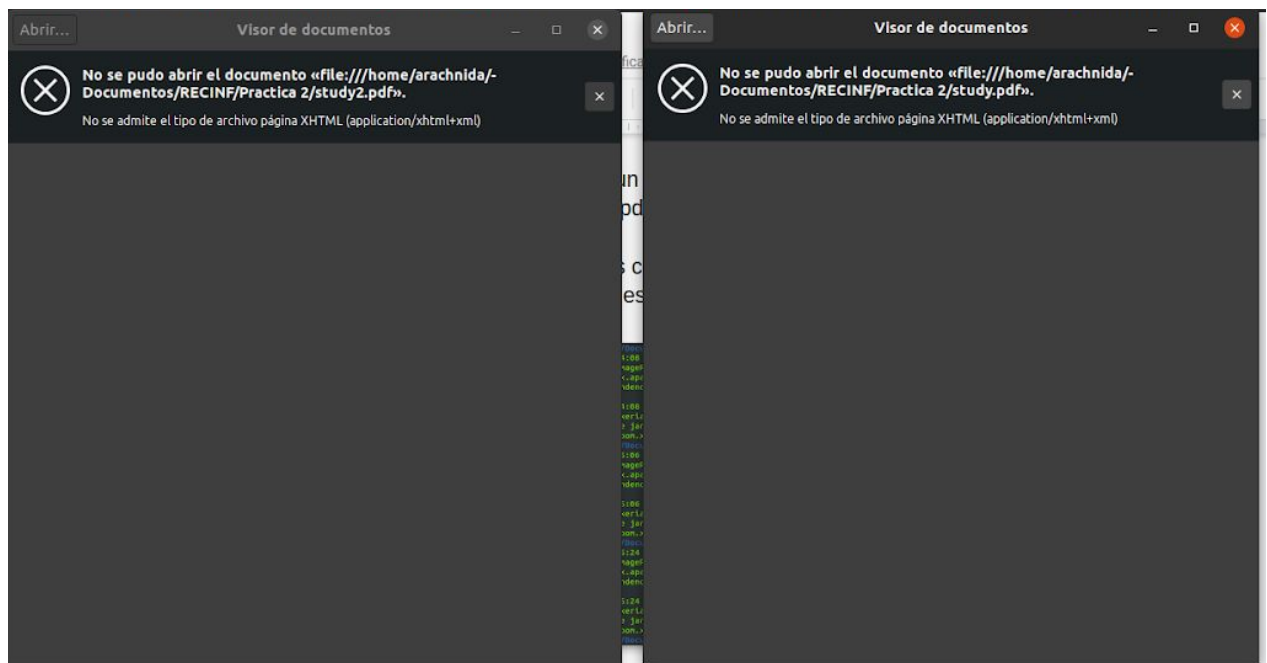
```
arachnida@Celis:~/Documentos/RECINF/Practica 2$ tika ./study-288.rdf > study.doc
nov. 03, 2020 7:44:08 P. M. org.apache.tika.config.InitializableProblemHandler$3 handleInitializableProblem
ADVERTENCIA: J2KImageReader not loaded. JPEG2000 files will not be processed.
See https://pdfbox.apache.org/2.0/dependencies.html#jai-image-io
for optional dependencies.

nov. 03, 2020 7:44:08 P. M. org.apache.tika.config.InitializableProblemHandler$3 handleInitializableProblem
ADVERTENCIA: org.xerial's sqlite-jdbc is not loaded.
Please provide the jar on your classpath to parse sqlite files.
See tika-parsers/pom.xml for the correct version.
arachnida@Celis:~/Documentos/RECINF/Practica 2$ tika ./study-288.rdf > study.pdf
nov. 03, 2020 7:45:06 P. M. org.apache.tika.config.InitializableProblemHandler$3 handleInitializableProblem
ADVERTENCIA: J2KImageReader not loaded. JPEG2000 files will not be processed.
See https://pdfbox.apache.org/2.0/dependencies.html#jai-image-io
for optional dependencies.

nov. 03, 2020 7:45:06 P. M. org.apache.tika.config.InitializableProblemHandler$3 handleInitializableProblem
ADVERTENCIA: org.xerial's sqlite-jdbc is not loaded.
Please provide the jar on your classpath to parse sqlite files.
See tika-parsers/pom.xml for the correct version.
arachnida@Celis:~/Documentos/RECINF/Practica 2$ tika ./study.doc > study2.pdf
nov. 03, 2020 7:45:24 P. M. org.apache.tika.config.InitializableProblemHandler$3 handleInitializableProblem
ADVERTENCIA: J2KImageReader not loaded. JPEG2000 files will not be processed.
See https://pdfbox.apache.org/2.0/dependencies.html#jai-image-io
for optional dependencies.

nov. 03, 2020 7:45:24 P. M. org.apache.tika.config.InitializableProblemHandler$3 handleInitializableProblem
ADVERTENCIA: org.xerial's sqlite-jdbc is not loaded.
Please provide the jar on your classpath to parse sqlite files.
See tika-parsers/pom.xml for the correct version.
arachnida@Celis:~/Documentos/RECINF/Practica 2$
```

La conversión .rdf -> .doc funciona correctamente al abrirlo. Sin embargo, las conversiones .doc -> -pdf y .rdf -> .pdf dan error:



**8.Descargar las 3 imágenes que se proporcionan en la carpeta “Material Tika” y decir qué imagen (o imágenes) se sacó (sacaron) con un producto perteneciente a apple computer inc.**

Las fotos “q” y “r” se hicieron con un dispositivo Apple. Se puede encontrar en la propiedad “Primary Platform”.

```
tika --metadata ./Material\ Tika-20201030/q.jpg  
tika --metadata ./Material\ Tika-20201030/r.jpg
```

Sin embargo, la foto “s” se hizo con una cámara Fujifilm FinePix A500. Se puede observar en los parámetros Exif IFD0:Make y Exif IFD0:Model.

```
tika --metadata ./Material\ Tika-20201030/s.JPG
```

**9. Describir el procedimiento seguido para guardar el contenido de una web cualquiera en un archivo html, este convertirlo en doc y comprobar los metadatos de este último.**

Primero, obtengo el html de una URL dada y lo guardo en un documento .html localmente:

```
tika --html https://www.20minutos.es/ > 20minutos.html
```

Luego, convierto el archivo .html a .doc:

```
tika ./20minutos.html > 20minutos.doc
```

Finalmente, obtengo los metadatos del archivo .doc

```
tika --metadata 20minutos.doc
```

**10. Descargar tres imágenes, a elección del alumno, de tres sitios diferentes donde los usuarios compartan imágenes como pueden ser: facebook, instagram, flickr, twitter... y comentar las diferencias que encontramos en los metadatos.**

- Twitter:



```

arachnid@Cells:~/Documentos/RECINF/Practica 2$ tika --metadata ./twitter.jpeg
nov. 03, 2020 8:31:38 P. M. org.apache.tika.config.InitializableProblemHandler$3 handleInitializableProblem
ADVERTENCIA: J2KImageReader not loaded. JPEG2000 files will not be processed.
See https://pdfbox.apache.org/2.0/dependencies.html#jai-image-io
for optional dependencies.

nov. 03, 2020 8:31:38 P. M. org.apache.tika.config.InitializableProblemHandler$3 handleInitializableProblem
ADVERTENCIA: org.xerial's sqlite-jdbc is not loaded.
Please provide the jar on your classpath to parse sqlite files.
See tika-parsers/pom.xml for the correct version.
Component 1: Y component: Quantization table 0, Sampling factors 2 horiz/2 vert
Component 2: Cb component: Quantization table 1, Sampling factors 1 horiz/1 vert
Component 3: Cr component: Quantization table 1, Sampling factors 1 horiz/1 vert
Compression Type: Progressive, Huffman
Content-Length: 305384
Content-Type: image/jpeg
Data Precision: 8 bits
File Modified Date: mar. nov. 03 20:07:42 +01:00 2020
File Name: twitter.jpeg
File Size: 305384 bytes
Image Height: 1027 pixels
Image Width: 1918 pixels
Number of Components: 3
Number of Tables: 2 Huffman tables
Resolution Units: none
Thumbnail Height Pixels: 0
Thumbnail Width Pixels: 0
X Resolution: 1 dot
X-Parsed-By: org.apache.tika.parser.DefaultParser
X-Parsed-By: org.apache.tika.parser.jpeg.JpegParser
Y Resolution: 1 dot
resourceName: twitter.jpeg
tiff:BitsPerSample: 8
tiff:ImageLength: 1027
tiff:ImageWidth: 1918
arachnid@Cells:~/Documentos/RECINF/Practica 2$

```

- Facebook:

```

arachnid@Cells:~/Documentos/RECINF/Practica 2$ tika --metadata facebook.jpg
nov. 03, 2020 8:31:51 P. M. org.apache.tika.config.InitializableProblemHandler$3 handleInitializableProblem
ADVERTENCIA: J2KImageReader not loaded. JPEG2000 files will not be processed.
See https://pdfbox.apache.org/2.0/dependencies.html#jai-image-io
for optional dependencies.

nov. 03, 2020 8:31:52 P. M. org.apache.tika.config.InitializableProblemHandler$3 handleInitializableProblem
ADVERTENCIA: org.xerial's sqlite-jdbc is not loaded.
Please provide the jar on your classpath to parse sqlite files.
See tika-parsers/pom.xml for the correct version.
Component 1: Y component: Quantization table 0, Sampling factors 2 horiz/2 vert
Component 2: Cb component: Quantization table 1, Sampling factors 1 horiz/1 vert
Component 3: Cr component: Quantization table 1, Sampling factors 1 horiz/1 vert
Compression Type: Progressive, Huffman
Content-Length: 197677
Content-Type: image/jpeg
Data Precision: 8 bits
File Modified Date: mar. nov. 03 20:08:28 +01:00 2020
File Name: facebook.jpg
File Size: 197677 bytes
Image Height: 1030 pixels
Image Width: 1440 pixels
Number of Components: 3
Number of Tables: 2 Huffman tables
Original Transmission Reference: tts5PYoqBaQIn-mKTmke
Resolution Units: none
Thumbnail Height Pixels: 0
Thumbnail Width Pixels: 0
X Resolution: 1 dot
X-Parsed-By: org.apache.tika.parser.DefaultParser
X-Parsed-By: org.apache.tika.parser.jpeg.JpegParser
Y Resolution: 1 dot
resourceName: facebook.jpg
tiff:BitsPerSample: 8
tiff:ImageLength: 1030
tiff:ImageWidth: 1440
arachnid@Cells:~/Documentos/RECINF/Practica 2$

```

- Flickr:

```

brachini@cells:~/Documents/RSCTM/Practica_25$ tika --metadata flickr.jpg
nov. 03, 2020 8:32:02 P. M. org.apache.tika.config.InitializableProblemHandler$3 handleInitializableProblem
ADVERTENCIA: 22ImageHeader not loaded. 2PCC000 files will not be processed.
See https://pdfbox.apache.org/2.0/dependencies.html#jal-image-io
for optional dependencies.

nov. 03, 2020 8:32:02 P. M. org.apache.tika.config.InitializableProblemHandler$3 handleInitializableProblem
ADVERTENCIA: org.xerial's sqlite-jdbc is not loaded.
Please provide the jar on your classpath to parse sqlite files.
See tika-parsers/pom.xml for the correct version.
Application Record Version: 4
Author: Adelheid Smitt
Blue Colorant: (0,1492, 0,8632, 0,7446)
Blue TRC: 0.0085908
By-Line: Adelheid Smitt
CM Type: AD8E
Class: Display Device
Coded Character Set: UTF-8
Color Transform: YCbCr
Color space: RGB
Component 1: Unknown (0) component: Quantization table 0, Sampling factors 2 horiz/2 vert
Component 2: Y component: Quantization table 1, Sampling factors 1 horiz/1 vert
Component 3: Cb component: Quantization table 1, Sampling factors 1 horiz/1 vert
Compression type: baseline
Content-Length: 235103
Content-Type: image/jpeg
Copyright: Copyright 1999 Adobe Systems Incorporated
DCT Encode Version: 25000
Data Precision: 8 bits
Device manufacturer: none
Enveloped Record Version: 4
Exif IFD0:Artist: Adelheid Smitt
File Modified Date: Mon. Nov. 03 20:09:20 +01:00 2020
File Name: flickr.jpg
File Size: 235103 bytes
Flags 0: 04
Flags 1: 0
Green Colorant: (0,2053, 0,6257, 0,0609)
Green TRC: 0.0085908
Image Height: 900 pixels
Image Width: 1000 pixels
Media Black Point: (0, 0, 0)
Media White Point: (0,9505, 1, 1,0891)
Number of Components: 3
Number of Tables: 4 Huffman tables
Primary Platform: Apple Computer, Inc.
Profile Connection Space: XYZ
Profile Date/Time: 1999:06:03 00:00:00
Profile Description: Adobe RGB (1998)
Profile Size: 500
Red Colorant: (0,6997, 0,3111, 0,0195)
Red TRC: 0.0085908

Signature: acsp
Tag Count: 10
Unknown tag (0x02e0): https://flickr.com/e/m0QTE1yVa2vrlDN2FNxhaQELZdZ4ETX28IpeAMgPCPhuIUMK30
X-Parsed-By: org.apache.tika.parser.DefaultParser
X-Parsed-By: org.apache.tika.parser.jpeg.JpegParser
XYZ values: 0,964 1 0,825
X-creator: Adelheid Smitt
X-metadata: Adelheid Smitt
ResourceName: flickr.jpg
tiff:BitsPerSample: 8
tiff:ImageLength: 900
tiff:ImageWidth: 1000
brachini@cells:~/Documents/RSCTM/Practica_25$

```

Los metadatos de Twitter y Facebook son prácticamente iguales. Lo único en lo que se diferencian es en el atributo “Original Transmission Reference”. Sin embargo, Flickr cuenta con muchos más metadatos que los anteriores, ya que la plataforma está destinada a un público profesional.

## 11. Se puede trabajar con Tika desde Eclipse. Describe los pasos que han de realizarse para poder crear un proyecto Tika en Eclipse en el que se extraiga el contenido de un fichero PDF.

Primero, debemos crear un proyecto Java en eclipse de la manera en que lo haríamos regularmente.

Tras esto, creamos una carpeta “lib” en la raíz del proyecto y copiamos allí el .jar de Tika.

Una vez hecho esto, hacemos click derecho sobre el proyecto y nos dirigimos a Properties -> Java Build Path -> Libraries -> Add JARs y lo añadimos.

A continuación creamos una clase nueva, importamos el paquete org.apache.tika.Tika y añadimos el siguiente código:

```
import java.io.File;
import org.apache.tika.Tika;
public class prueba {
    public static void main(String[] args) throws
Exception
    {
        try {
            File file = new File("ruta del
archivo");
            String content = new
Tika().parseToString(file);
            System.out.println(content);
        } catch (final Exception e) {
            e.printStackTrace();
        }
    }
}
```

