TIKA - DOCUMENT TYPE DETECTION

MIME Standards

Multipurpose Internet Mail Extensions *MIME* standards are the best available standards for identifying document types. The knowledge of these standards helps the browser during internal interactions.

Whenever the browser encounters a media file, it chooses a compatible software available with it to display its contents. In case it does not have any suitable application to run a particular media file, it recommends the user to get the suitable plugin software for it.

Type Detection in Tika

Tika supports all the Internet media document types provided in MIME. Whenever a file is passed through Tika, it detects the file and its document type. To detect media types, Tika internally uses the following mechanisms.

File extensions

Checking the file extensions is the simplest and most-widely used method to detect the format of a file. Many applications and operating systems provide support for these extensions. Shown below are the extension of a few known file types.

File name	Extention
image	.jpg
audio	.mp3
java archive file	.jar
java class file	.class

Content-type hints

Whenever you retrieve a file from a database or attach it to another document, you may lose the file's name or extension. In such cases, the metadata supplied with the file is used to detect the file extension.

Magic bytes

Observing the raw bytes of a file, you can find some unique character patterns for each file. Some files have special byte prefixes called **magic bytes** that are specially made and included in a file for the purpose of identifying the file type.

For example, you can find CA FE BA BE *hexadecimalformat* in a java file and %PDF *ASCIIformat* in a pdf file. Tika uses this information to identify the media type of a file.

Character encodings

Files with plain text are encoded using different types of character encoding. The main challenge here is to identify the type of character encoding used in the files. Tika follows character encoding techniques like **Bom markers** and **Byte Frequencies** to identify the encoding system used by the plain text content.

Xml root characters

To detect XML documents, Tika parses the xml documents and extracts the information such as

root elements, namespaces, and referenced schemas from where the true media type of the files can be found.

Type detection using Facade class

The **detect** method of facade class is used to detect the document type. This method accepts a file as input. Shown below is an example program for document type detection with Tika facade class.

```
import java.io.File;

import org.apache.tika.Tika;

public class Typedetection {

   public static void main(String[] args) throws Exception {

      //assume example.mp3 is in your current directory
      File file = new File("example.mp3");//

      //Instantiating tika facade class
      Tika tika = new Tika();

      //detecting the file type using detect method
      String filetype = tika.detect(file);
      System.out.println(filetype);
   }
}
```

Save the above code as TypeDetection.java and run it from the command prompt using the following commands:

```
javac TypeDetection.java
java TypeDetection

audio/mpeg
```