

TIKA - EXTRACTING HTML DOCUMENT

http://www.tutorialspoint.com/tika/tika_extracting_html_document.htm

Copyright © tutorialspoint.com

Given below is the program to extract content and metadata from an HTML document.

```
import java.io.File;
import java.io.FileInputStream;
import java.io.IOException;

import org.apache.tika.exception.TikaException;
import org.apache.tika.metadata.Metadata;
import org.apache.tika.parser.ParseContext;
import org.apache.tika.parser.html.HtmlParser;
import org.apache.tika.sax.BodyContentHandler;

import org.xml.sax.SAXException;

public class HtmlParse {

    public static void main(final String[] args) throws IOException, SAXException,
    TikaException {

        //detecting the file type
        BodyContentHandler handler = new BodyContentHandler();
        Metadata metadata = new Metadata();
        FileInputStream inputstream = new FileInputStream(new File("example.html"));
        ParseContext pcontext = new ParseContext();

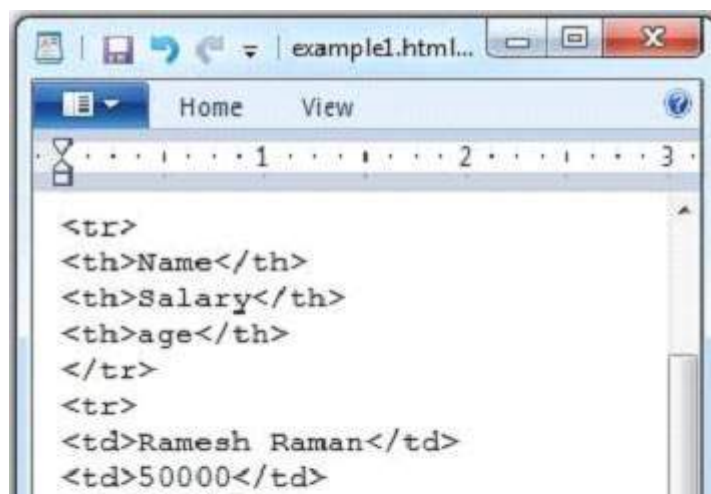
        //Html parser
        HtmlParser htmlparser = new HtmlParser();
        htmlparser.parse(inputstream, handler, metadata, pcontext);
        System.out.println("Contents of the document:" + handler.toString());
        System.out.println("Metadata of the document:");
        String[] metadataNames = metadata.names();

        for(String name : metadataNames) {
            System.out.println(name + ": " + metadata.get(name));
        }
    }
}
```

Save the above code as **HtmlParse.java**, and compile it from the command prompt by using the following commands:

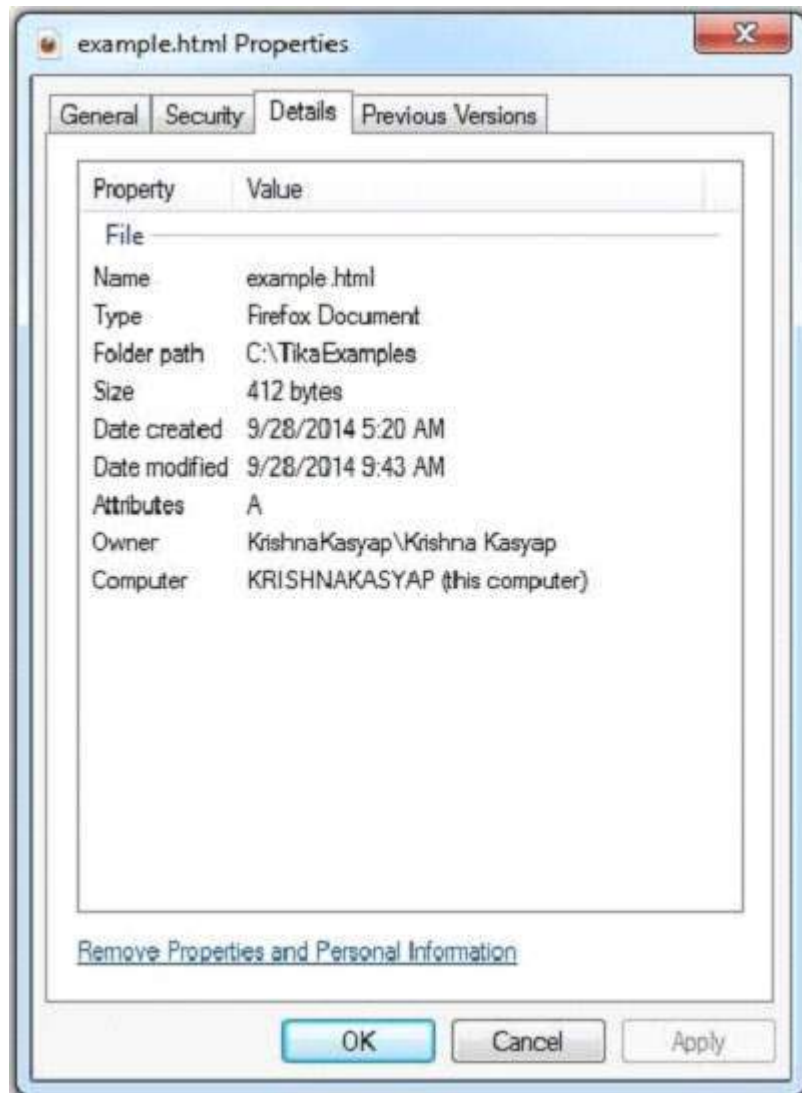
```
javac HtmlParse.java
java HtmlParse
```

Below given is the snapshot of example.html document.



```
<th>20</th>
</tr>
<tr>
<td>Shabbir Hussein</td>
<td>70000</td>
<th>25</th>
</tr>
<tr>
<td>umesh Raman</td>
<td>50000</td>
<th>30</th>
</tr>
</table>
</body>
</html>
```

The HTML document has the following properties:



After executing the above program you will get the following output.

Output:

Contents of the document:

Name	Salary	age
Ramesh Raman	50000	20
Shabbir Hussein	70000	25

Umesh Raman	50000	30
Somesh	50000	35

Metadata of the document:

title: HTML Table Header

Content-Encoding: windows-1252

Content-Type: text/html; charset=windows-1252

dc:title: HTML Table Header