

TIKA - EXTRACTING XML DOCUMENT

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

Copyright © tutorialspoint.com

Given below is the program to extract content and metadata from an XML 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.xml.XMLParser;
import org.apache.tika.sax.BodyContentHandler;

import org.xml.sax.SAXException;

public class XmlParse {

    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("pom.xml"));
        ParseContext pcontext = new ParseContext();

        //Xml parser
        XMLParser xmlparser = new XMLParser();
        xmlparser.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 **XmlParse.java**, and compile it from the command prompt by using the following commands:

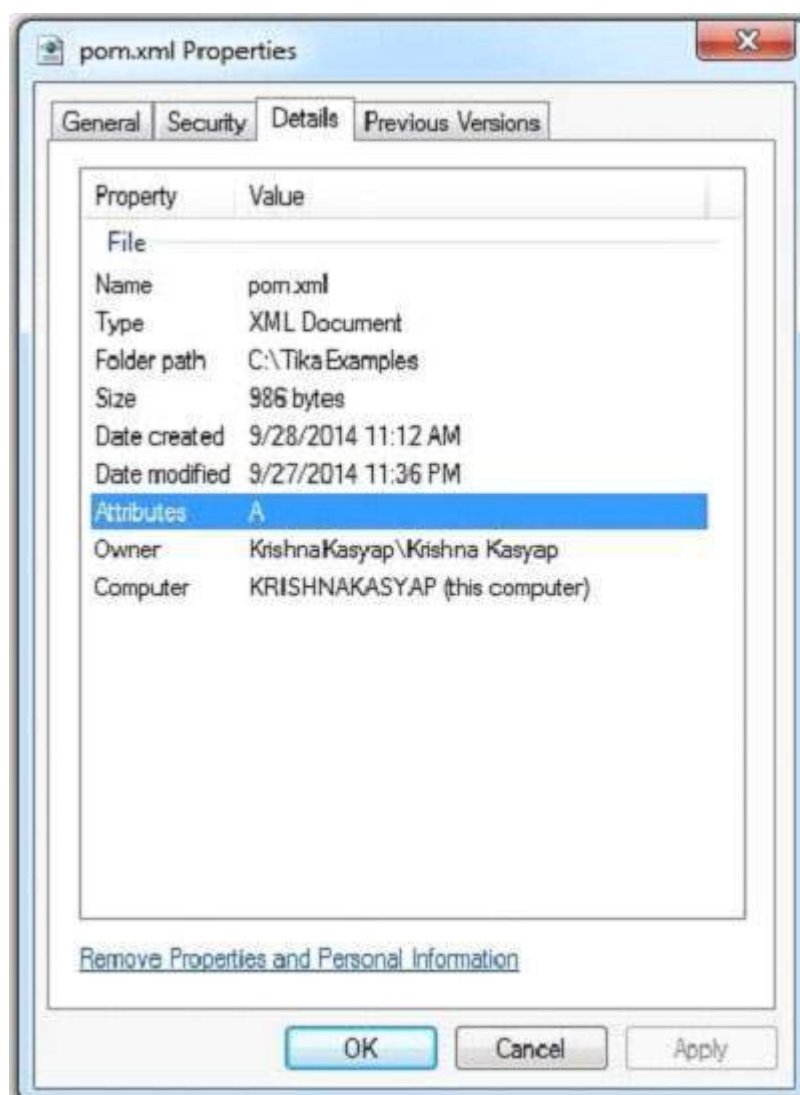
```
javac XmlParse.java
java XmlParse
```

Below given is the snapshot of pom.xml



```
<dependency>
<groupId>org.apache.tika</groupId>
<artifactId>tika-parsers</artifactId>
<version>1.6</version>
</dependency>
</dependencies>
<build>
  <sourceDirectory>src</sourceDirectory>
  <plugins>
    <plugin>
      <artifactId>maven-compiler-plugin</artifactId>
      <version>3.1</version>
      <configuration>
        <source>1.7</source>
        <target>1.7</target>
      </configuration>
    </plugin>
  </plugins>
</build>
```

This document has the following properties:



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

Output:

Contents of the document:
4.0.0
org.apache.tika
tika
1.6

```
org.apache.tika
tika-core
1.6
org.apache.tika
tika-parsers
1.6
src
maven-compiler-plugin
3.1
1.7
1.7
```

Metadata of the document:

Content-Type: application/xml