POI Text Extraction

by Nick Burch

1. Overview

POI provides text extraction for all the supported file formats. In addition, it provides access to the metadata associated with a given file, such as title and author.

In addition to providing direct text extraction classes, POI works closely with the Apache Tika text extraction library. Users may wish to simply utilise the functionality provided by Tika.

2. Common functionality

All of the POI text extractors extend from *org.apache.poi.POITextExtractor*. This provides a common method across all extractors, getText(). For many cases, the text returned will be all you need. However, many extractors do provide more targetted text extraction methods, so you may wish to use these in some cases.

All POIFS / OLE 2 based text extractors also extend from *org.apache.poi.POIOLE2TextExtractor*. This additionally provides common methods to get at the HPFS document metadata.

All OOXML based text extractors (available in POI 3.5 and later) also extend from *org.apache.poi.POIOOXMLTextExtractor*. This additionally provides common methods to get at the OOXML metadata.

3. Text Extractor Factory - POI 3.5 or later

A new class in POI 3.5, *org.apache.poi.extractor.ExtractorFactory* provides a similar function to WorkbookFactory. You simply pass it an InputStream, a file, a POIFSFileSystem or a OOXML Package. It figures out the correct text extractor for you, and returns it.

4. Excel

For .xls files, there is *org.apache.poi.hssf.extractor.ExcelExtractor*, which will return text, optionally with formulas instead of their contents. Those using POI 3.5 can also use *org.apache.poi.xssf.extractor.XSSFExcelExtractor*, to perform a similar task for .xlsx files.

In addition, there is a second text extractor for .xls files, *org.apache.poi.hssf.extractor.EventBasedExcelExtractor*. This is based on the streaming EventUserModel code, and will generally deliver a lower memory footprint for extraction. However, it will have problems correctly outputting more complex formulas, as it works with records as they pass, and so doesn't have access to all parts of complex and shared formulas.

5. Word

For .doc files, in scratchpad there is *org.apache.poi.hwpf.extractor.WordExtractor*, which will return text for your document. Those using POI 3.5 can also use *org.apache.poi.xwpf.extractor.XPFFWordExtractor*, to perform a similar task for .docx files.

6. PowerPoint

For .ppt files, in scratchpad there is *org.apache.poi.hslf.extractor.PowerPointExtractor*, which will return text for your slideshow, optionally restricted to just slides text or notes text. Those using POI 3.5 can also use *org.apache.poi.xslf.extractor.XSLFPowerPointExtractor*, to perform a similar task for .pptx files.

7. Visio

For .vsd files, in scratchpad there is *org.apache.poi.hdgf.extractor.VisioTextExtractor*, which will return text for your file.