LECTURA DE UN .DOC/.DOCX MEDIANTE CODIGO JAVA

Para leer un contenido de un .doc o un .docx con java usaremos una la librería POI de Apache que nos permitirá extraer el texto y guardarlo en un String.

Comenzamos descargando la librería desde la siguiente pagina

https://poi.apache.org/download.html

Una vez descargado lo metemos en una carpeta de librerías dentro de nuestro proyecto y abrimos eclipse. Hacemos click derecho > configure build path y en la pestaña de librerías seleccionamos classpath y add external Jar, y tendríamos que añadir todos los jar que tiene la carpeta que habíamos descargado.

Una vez añadidos todos los jar aplicaríamos los cambios.

UTILIZACION

Los imports que deberíamos utilizar para este trabajo serían los siguientes:

```
5
6  import org.apache.poi.hwpf.HWPFDocument;
7  import org.apache.poi.xwpf.extractor.*;
8  import org.apache.poi.xwpf.usermodel.XWPFDocument;
9  import org.apache.poi.hwpf.extractor.WordExtractor;
```

Lo primero que tendríamos que hacer sería crear el FileInputStream.

Después dependiendo si el documento es un .doc o un .docx a partir de aquí cambiaria un poco el proceso.

Si es un .docx tendríamos que crear un objeto de tipo XWPFDocument pasándole el FileInputStream como parámetro, y este nuevo objeto podríamos crear un XWPFWordExtractor que ya podríamos guardar en un String con el método getText().

Si es un .doc tendríamos que crear un objeto de tipo HWPFDocument pasándole el FileInputStream como parámetro, y este nuevo objeto podríamos crear un WordExtractor que ya podríamos guardar en un String con el método getText().