KULTURA 2.0 - Píldoras de formación:
OCR: tecnología para el reconocimiento óptico de caracteres en una imagen

abril 2011

Desarrollada en el marco de la iniciativa kultura 2.0, el Observatorio Vasco de la Cultura ha puesto en marcha la elaboración y difusión de «kultura 2.0. píldoras de formación», cuyo objetivo es dar una respuesta concreta a demandas de las empresas y profesionales del sector de la cultura en la CAE.

¿Qué es el OCR?

El OCR (Optical Character Recognition) es una tecnología que trata de emular la capacidad del ojo humano para reconocer objetos. Concretamente es un software que permite el reconocimiento óptico de los caracteres contenidos en una imagen (documento escaneado o fotografía), de forma que estos se vuelven comprensibles o reconocibles para un ordenador, obteniendo como resultado final un archivo en un formato de texto editable. El formato del archivo de salida (txt, pdf, etc.) dependerá de las posibilidades que ofrezca el software.

¿Cómo funciona el OCR?

Para reconocer los caracteres, el software inspecciona la imagen pixel[1] a pixel, buscando formas que coincidan con los rasgos de los caracteres. En función del nivel de complejidad o grado de desarrollo del software, éste buscará coincidencias con los caracteres y fuentes disponibles en el programa, o tratará de identificar los caracteres a través del análisis de sus características, de forma que el reconocimiento de los mismos no se limite exclusivamente a un determinado número de fuentes.

El OCR puede analizar los elementos del documento (bloques de texto, imágenes, tablas…), examinando los espacios en blanco y descomponiendo el texto en líneas, palabras y caracteres, de forma que el programa puede formular distintas hipótesis y cotejarlas con los diccionarios contenidos por el mismo (actualmente los programas contienen diccionarios en distintos idiomas), para formar palabras y textos completos.

Aunque actualmente el OCR puede llegar a mantener la estructura de los documentos originales en el archivo de salida, e incluso reconocer caracteres contenidos en documentos manuscritos, diagramas, partituras, etc., no hay que olvidar que su nivel de efectividad sigue siendo limitado, lo que hace necesaria una posterior revisión y corrección manual del texto generado. Actualmente existe en el mercado una amplia oferta de software OCR, cuyo nivel de sofisticación y precio está directamente relacionado con su precisión y efectividad en el reconocimiento de caracteres.

1 Unidad menor o elemento más pequeño que forma parte de una imagen digital.

¿Para qué se utiliza el OCR?

El OCR es una tecnología transversal, aplicable en distintos ámbitos y sectores para la digitalización de formularios, documentos administrativos, informes, etc., ya que las ventajas que ofrece son comunes para todos ellos.

En el sector de la cultura, por ejemplo en el ámbito de la preservación del patrimonio, el OCR se aplica principalmente en los procesos de digitalización de documentos históricos, en soporte papel o microformas.

La tipología de documentos sobre los que se recomienda aplicar el OCR es diversa. Identificamos a continuación algunas posibilidades extraídas del documento *«Proceso de Digitalización en la Biblioteca Nacional de España»*:

Tipo de documento	Archivo de preservación	Archivo de difusión
Texto impreso	X (texto impreso sin imágenes)	X
Prensa	X	X
Páginas mecanografiadas	X	X
Panfletos	X	X
Partituras impresas	–	X

Buenas prácticas – PROYECTO IMPACT

A raíz de los problemas detectados en los proyectos de digitalización surgidos alrededor de la iniciativa europea i2010, se ha puesto en marcha el proyecto IMPACT, en el que participan 11 bibliotecas nacionales y regionales, 13 entidades de investigación y 2 empresas. Todas ellas comparten su conocimiento con el objetivo de desarrollar una tecnología OCR que ayude a mejorar los procesos y resolver los problemas de los proyectos de digitalización masiva, facilitando la plena integración de los contenidos digitalizados en las nuevas tecnologías de la información y comunicación.

¿Qué ventajas puede tener la aplicación del OCR?

Búsqueda y recuperación de documentos

1. La aplicación del OCR permite realizar búsquedas de texto libre sobre la totalidad del documento.
2. En el proceso de creación de los metadatos, el OCR se puede utilizar para generar índices de palabras clave del texto reconocido de forma automática.

Explotación de los documentos

3. El OCR permite convertir el texto de los documentos digitalizados a formatos editables.
4. Aunque el OCR no es una herramienta para hacer los documentos accesibles para personas con discapacidades visuales, su aplicación combinada con otras tecnologías permite que el texto resultante se sintetice en líneas de braille o archivos de audio.

Perspectiva económica

5. Ahorro de tiempo respecto a la inserción manual de datos (el OCR puede alcanzar una velocidad de lectura de hasta 1.200 caracteres por segundo).

6. El almacenamiento en formato de texto puede suponer un ahorro de espacio respecto del almacenamiento como imagen (el archivo de texto necesita aproximadamente 1/3 del espacio que ocupa la imagen).

¿Cuáles son los inconvenientes del OCR?

En el marco de la iniciativa i2010 promovida por la Comisión Europea, se han desarrollado numerosos proyectos de digitalización masiva que han puesto de manifiesto los siguientes problemas en relación con el OCR:

1. Carencia de conocimiento y expertos en las instituciones.

2. Elevado coste de generar texto electrónico (no confundir con imagen digital) con todas sus funciones (este proceso puede realizarse tecleando el texto o a través de OCR y posterior revisión y corrección del texto).

3. Nivel de efectividad insatisfactorio del OCR en el reconocimiento de documentos históricos, anteriores al inicio de la edición industrial de libros a mediados del siglo XIX.

Respecto al limitado nivel de efectividad del OCR, además de la calidad o grado de desarrollo del propio software, existen factores extrínsecos asociados al estado físico del documento original o a la calidad de la imagen digital, que pueden resultar determinantes en el resultado del proceso:

Factores relativos al estado del documento original:

- Deterioro de los documentos originales
- Letra borrosa o poco nítida
- Manchas o transparencias en el papel
- Letras fragmentadas o solapadas
- Tipografías extrañas o fuera de uso
- Dimensiones del documento original (aunque no afecte directamente a la efectividad del OCR, las dimensiones del documento original –por ejemplo periódicos–, puede hacer que resulte complejo escanearlos en equipos convencionales)

Factores relativos a la calidad de la imagen digital:

- Baja resolución de la imagen
- Incorrecta configuración del escáner

Recomendaciones técnicas para la aplicación del OCR

Las recomendaciones técnicas que hacen las distintas instituciones implicadas en procesos de digitalización en relación con el OCR se refieren principalmente a la resolución mínima de la imagen escaneada ya que, como se ha citado, es un factor determinante para obtener un resultado satisfactorio: a mayor resolución de escaneo mayor precisión del OCR.

Con carácter general, se establece una resolución mínima de 300 ppp[2] para que el reconocimiento de los caracteres sea efectivo, aunque dependiendo de las características del documento se aconseja una resolución mínima superior.

Tipo de documento	Resolución mínima
Textos con tipos de letra claros	300 ppp
Tipos de letra pequeña u originales de poca calidad (prensa)	600 ppp

¿Qué otros temas sobre kultura 2.0 consideras podríamos tratar en próximas píldoras de formación? (concretar lo mejor posible)

Envíanos tus sugerencias al mail: kulturabehatokia@ej-gv.es

2 Puntos por pulgada.