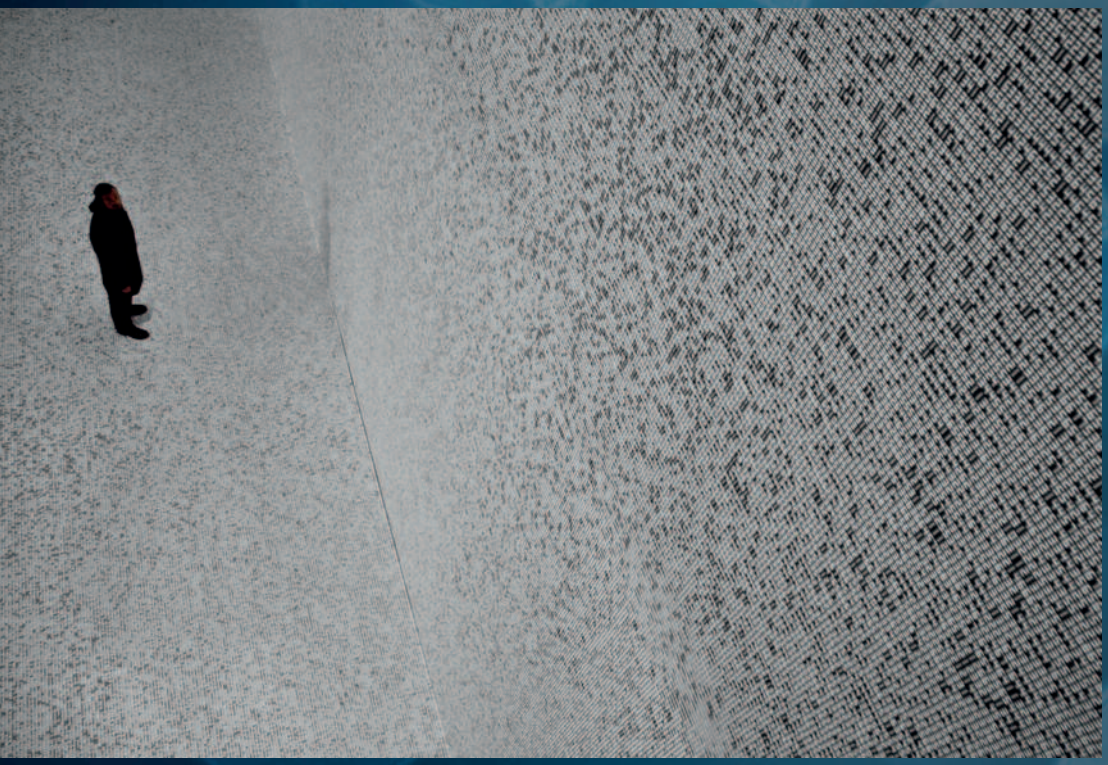
Gestión documental inteligente

Guillermo Castellano | @guillearch

Una de las escenas más famosas de El burgués gentilhombre de Molière es el diálogo entre Monsieur Jourdain y su profesor Eduardo, en el que el protagonista descubre que lleva más de cuarenta años hablando en prosa sin saberlo. ¿Y si nosotros lleváramos años empleando la IA (inteligencia artificial) en el ámbito de la gestión documental sin saberlo?

Al igual que le ocurre al protagonista de *El burgués gentilhombre* con la prosa, llevamos décadas utilizando la IA sin que muchos sean conscientes de ello. Como afirma Ray Kurzweil en *How to create a mind: The secret of human thought revealed*, "si todos los sistemas de IA decidieran ir a la huelga mañana, nuestra civilización se paralizaría: no podríamos sacar dinero del banco y, de hecho, nuestro dinero desaparecía; las comunicaciones, el transporte y la manufactura se detendrían por completo." Uno de los ámbitos donde se está aplicando la IA es en el de la gestión documental, donde ayuda a ahorrar tiempo en los procesos de captura, descripción, clasificación y recuperación.



Los OCR (programas de reconocimiento óptico de caracteres) se basan en algoritmos entrenados para reconocer los caracteres que contiene el documento electrónico. Gracias a esta tecnología, desarrollada inicialmente con el propósito de crear máquinas lectoras para ciegos, es posible automatizar el proceso de extracción de datos de documentos que hayan sido escritos con tipografías informáticas. La extracción funciona tanto si se trata de un documento en papel digitalizado como si se trata de un documento electrónico nativo.

La tecnología OCR se puede combinar con el uso de plantillas, de tal manera que el OCR reconozca el tipo de documento y solo extraiga metadatos en unos campos determinados. Esto es muy útil cuando se gestionan documentos normalizados (facturas, albaranes, pedidos, notas simples registrales, etc.), donde los metadatos que interesan se encuentran siempre en las mismas coordenadas.

La mayoría de gestores documentales se integran de forma nativa con programas OCR, lo que permite que los metadatos capturados queden asociados a cada documento en el momento del alta. Además, el gestor documental puede utilizar estos metadatos para clasificar automáticamente los documentos según las reglas que se hayan establecido. La intervención humana en este proceso se limita a supervisar el desarrollo y corregir manualmente los posibles errores.

Aunque el mercado de los OCR está dominado por soluciones privativas, existen alternativas libres como Cuneiform, GOCR y Tesseract, creado por HP y desarrollado desde 2006 por Google.

Otra aplicación de la IA en el ámbito archivístico es la transcripción automática de manuscritos. Estos suelen resultar difíciles de leer para el usuario sin conocimientos de paleografía, lo que está llevando a distintas instituciones a trabajar en proyectos de reconocimiento inteligente de caracteres (una evolución de la tecnología OCR) como READ, una iniciativa de la Unión Europea que busca "revolucionar el acceso a los documentos de archivo" y se ha materializado en Transkribus, una aplicación de software libre para automatizar la transcripción de documentos históricos.

Aunque esta tecnología todavía no está madura, dada la gran diferencia existente entre la escritura de cada persona, su aplicación en el ámbito de la transcripción de manuscritos es prometedora y sería interesante que proyectos de descripción y difusión de documentos históricos libres como AtoM se interesaran por ella.

Hasta el momento solo hemos hablado de extraer texto,

pero la IA se puede utilizar también para describir imágenes, audios o vídeos. Un ejemplo de ello es el proyecto de etiquetado automático y búsqueda semántica de contenidos audiovisuales en el que está trabajando nuestra compañera Virginia Bazán en RTVE.

Aparte de ahorrar tiempo en los procesos de captura, descripción y clasificación, la IA ayuda a que los resultados devueltos por los motores de búsqueda sean más precisos. Los motores de búsqueda más habituales, como Apache Lucene o Elasticsearch (ambos, de código abierto), utilizan algoritmos de radicación (*stemmers*) que permiten buscar por la raíz de la palabra en lugar de por la palabra completa. También existen métodos para que estos buscadores devuelvan palabras relacionadas semánticamente.

Quizá muchas personas no reconocen este tipo de tecnologías como IA debido al denominado "efecto IA": cuanto mejor funciona IA, más alto ponemos el listón de lo que consideramos inteligente. Esta paradoja, analizada por Pamela McCorduck en su ensayo *Máquinas que piensan: Una incursión personal en la historia y las perspectivas de la inteligencia artificial*, podría explicar que hoy en día solo se piense en IA cuando se habla de tecnologías de moda como el "*machine learning*" o el "*deep learning*", a pesar de la contribución de las tecnologías de las que hemos hablado a construir una gestión documental (más) inteligente.■