Nuevos enfoques del OCR para los libros impresos antiguos

New Approaches to OCR for Early Printed Books

Nikolaus Weichselbaumer, y otros

DigItalia, ISSN 1972-6201, Vol. 2, 2020, p. 74-87

Se presenta el proyecto OCR-D, financiado por la Deutsche Forschungsgemeinschaft (DFG) y puesto en marcha en 2014, a partir de la organización de un taller en el que expertos historiadores del libro e informáticos evaluaron los nuevos avances en el reconocimiento óptico de caracteres (OCR). En las últimas décadas, muchas bibliotecas han empezado a digitalizar sus fondos de impresos antiguos, sin embargo, quedaba por desarrollar la estrategia tecnológica que permitiera que los documentos digitalizados estén disponibles para la búsqueda a texto completo y su posterior procesamiento con herramientas de las Humanidades Digitales. El reconocimiento del texto completo de los documentos históricos es especialmente complicado debido a su gran variabilidad en cuento a letra, diseño, idioma y ortografía. Además, los motores OCR suelen estar entrenados con los tipos de letra actuales por lo que se ignora la gran variedad regional y estilística de la tipografía de los impresos anteriores a 1800. El proyecto OCR-D, llevado a cabo por la Academia de Ciencias y Humanidades de Berlín-Brandeburgo, la Biblioteca Herzog-August de Wolfenbüttel, la Biblioteca Estatal de Berlín y el Instituto Tecnológico de Karlsruhe, tiene como objetivo la creación de un marco conceptual y técnico que permita la transformación del texto completo de cualquier copia digital. El proyecto se organiza por fases, siendo la primera de ellas la creación de una herramienta que identifique automáticamente grupos de fuentes en imágenes de documentos antiguos, centrados en los grupos de fuentes góticas que se utilizaban

habitualmente en los textos alemanes impresos en los siglos XV y XVI: Fraktur, Bastarda, Rotunda, Textura y Schwabacher. La herramienta fue entrenada con 35.000 imágenes y alcanza un nivel de precisión del 98%. No sólo puede diferenciar entre los grupos de fuentes antes mencionados, sino también entre las hebreas, griegas, anticuadas e itálicas. Desde que las tecnologías OCR empezaron a adoptar las redes neuronales profundas, existen varios motores OCR de código abierto que pueden adaptarse a diferentes tipos de documentos. En este sentido se ha desarrollado la segunda fase del proyecto, creando "okralact", una infraestructura que permite utilizar varios de estos motores OCR de software libre como Tesseract, OCRopus, Kraken y Calamari. Al mismo tiempo facilita el entrenamiento para modelos específicos de grupos de fuentes. Se trabaja, así mismo en la posibilidad de diferenciar la tipografía de los diferentes talleres impresores, lo cual podría aportar mucha luz sobre algunas lagunas en la investigación histórica. Para finaliza, se trabaja en la puesta a disposición del software, para el público de forma gratuita.

DigItalia, ISSN 1972-6201, Vol. 2, 2020, p. 74-87

Se presenta el proyecto OCR-D, financiado por la Deutsche Forschungsgemeinschaft (DFG) y puesto en marcha en 2014, a partir de la organización de un taller en el que expertos historiadores del libro e informáticos evaluaron los nuevos avances en el reconocimiento óptico de caracteres (OCR). En las últimas décadas, muchas bibliotecas han empezado a digitalizar sus fondos de impresos antiguos, sin embargo, quedaba por desarrollar la estrategia tecnológica que permitiera que los documentos digitalizados estén disponibles para la búsqueda a texto completo y su posterior procesamiento con herramientas de las Humanidades Digitales. El reconocimiento del texto completo de los documentos históricos es especialmente complicado debido a su gran variabilidad en

cuento a letra, diseño, idioma y ortografía. Además, los motores OCR suelen estar entrenados con los tipos de letra actuales por lo que se ignora la gran variedad regional y estilística de la tipografía de los impresos anteriores a 1800. El proyecto OCR-D, llevado a cabo por la Academia de Ciencias y Humanidades de Berlín-Brandeburgo, la Biblioteca Herzog-August de Wolfenbüttel, la Biblioteca Estatal de Berlín y el Instituto Tecnológico de Karlsruhe, tiene como objetivo la creación de un marco conceptual y técnico que permita la transformación del texto completo de cualquier copia digital. El proyecto se organiza por fases, siendo la primera de ellas la creación de una herramienta que identifique automáticamente grupos de fuentes en imágenes de documentos antiguos, centrados en los grupos de fuentes góticas que se utilizaban habitualmente en los textos alemanes impresos en los siglos XV y XVI: Fraktur, Bastarda, Rotunda, Textura y Schwabacher. La herramienta fue entrenada con 35.000 imágenes y alcanza un nivel de precisión del 98%. No sólo puede diferenciar entre los grupos de fuentes antes mencionados, sino también entre las hebreas, griegas, anticuadas e itálicas. Desde que las tecnologías OCR empezaron a adoptar las redes neuronales profundas, existen varios motores OCR de código abierto que pueden adaptarse a diferentes tipos de documentos. En este sentido se ha desarrollado la segunda fase del proyecto, creando “okralact”, una infraestructura que permite utilizar varios de estos motores OCR de software libre como Tesseract, OCRopus, Kraken y Calamari. Al mismo tiempo facilita el entrenamiento para modelos específicos de grupos de fuentes. Se trabaja, así mismo en la posibilidad de diferenciar la tipografía de los diferentes talleres impresores, lo cual podría aportar mucha luz sobre algunas lagunas en la investigación histórica. Para finalizar, se trabaja en la puesta a disposición del software para el público de forma gratuita.

http://digitalia.sbn.it/article/view/2630

Resumen elaborado por María Osuna González