Deloitte.

Enhancing the quality and accuracy of OCR

Deloitte's Optical Character Recognition tool DocuMend not only helps identify OCR issues, it delivers an easy fix for them.

The Need

Text mining tools powered by natural language processing (NLP) have become indispensable to modern businesses. Like all AI, the quality of NLP solutions heavily depends on the accuracy of the optical character recognition (OCR) engine used. OCR technology is based on a sophisticated computer vision algorithm for recognizing characters from printed books, handwritten papers, and images in all possible fonts, sizes, and orientations. It quickly transforms document images into electronic, machine-readable text - the input for many downstream processes, including text-mining solutions.

OCR has become more accurate and is a

mature technology. Nevertheless, errors still occur in practical application, which, if left undetected, can skew results – especially where machine learning NLP algorithms are in-volved. Sources with tiny font sizes, blurred copies or colored paper can trip up OCR algorithms. The resulting electronic text might annoy or even amuse human readers, who can infer something has gone wrong. NLP algorithms, however, could completely misinterpret the
text, leading to failure of the text mining tool and to manual correction of the failed conversion. ➤

MAKING AN
IMPACT THAT
MATTERS *since 1845*

Our Solution: DocuMend

Deloitte uses a sophisticated ensemble approach to countering OCR problems, all packed into an intuitive, no-code graphical interface. DocuMend identifies any snags and provides users a quick and easy way to correct them.

It identifies issues by assessing the accuracy of OCR layers both at document and individual word level. Users can choose from several OCR engines and set the confidence threshold for OCR quality. The accuracy assessment is overlayed onto the original text in the PDF as a sort of textual heat map – ranking the OCR-processed words from lowest to highest confidence in OCR quality.

Until now, OCR quality control has required human proof-reading for error detection. DocuMend employs multiple OCR engines to cross-validate and highlight discrepancies between the engines as likely sources of error. Users correct errors by navigating the textual heat-map document or by working directly from a list where identified errors are ranked from most to least certain, down to the user-specified threshold.

Advantages/Benefits of DocuMend

· Automated assessment of OCR quality across multi-page documents.

· Visual feedback via word-level confidence heat-maps

· Intuitive, no-code graphical interface allows business users to quickly navigate documents and gain confidence in the quality of their OCR layers.

· Direct user interaction to correct errors on the spot.

Example Use Cases

· Improve quality of CV screening for AI-supported HR recruiting processes

· Improve reliability of fit-and-proper screening for banks with regulators

· Increase confidence in digital document processing within the business

· Lay a strong foundation for countless AI applications making use of text-mining/NLP

Contact

David Thogmartin
Leader aiStudio
dthogmartin@deloitte.de

Deloitte.