Classification of High-Quality PDF Text Layer Master Thesis

Contact

Robert Szabo
Medius AB
+46733378652
robert.szabo@medius.com

Introduction

Within the invoice workflow processes it is very common to send invoices as PDF files. These PDF files are many times electronically generated by financial systems and contain both an image representation as well as a textual representation including positional information about the text. In case the textual representation exists in the PDF, it does not have to be OCR'ed, which saves time and energy, which in turn is a benefit from a cost, efficiency, and environmental perspective.

One of the problems using the textual layer in PDF files is that they do not necessarily have to represent the same information as the image it is representing. This could happen due to a mistake in the financial system generating the PDF file, but it could also happen by purpose as a fraudulent act. In case there was a technique to classify the PDF file to have high quality in the text representation, it would be possible to fetch both textual and positional information from the PDF, without the need to run it through OCR.

Medius is a company working in the purchase to pay and document automation area and especially with invoice workflows. With many years of experience around the data capture area, as part of the invoice workflow, we are working with different machine learning techniques to extract information from scanned and electronic documents and to automate the workflows.

Problem formulation

One big part of the invoices that Medius processes are PDF files. These PDF files has many times been generated by a financial system, so they contain both an image representation and a textual representation, which includes positional information. In case the textual representation correlates with the image representation, there is no need to perform OCR, instead the textual representation including positional information could be used.

The problem to solve is to find a technique/algorithm that could determine if the PDF's textual representation correlates with the image representation. The accuracy of the classification step should be above 99% and it should be fast (below 0.1 second per page).

During the master thesis period, Medius will provide material that could be analyzed and be used for different types of algorithms.