Classification of Non-Invoices
Master Thesis

Contact

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

Introduction

Many documents that are processed in a workflow are coming in the form of PDF and image files. Before data could be extracted, they are OCR'ed so the textual information is converted to a machine readable format. The OCR step itself is time and energy intensive so this part should only be applied if necessary.

It often happens that PDF and image files that are not invoices are sent to the invoice workflow. These documents are later in the flow deleted manually when the users are detecting that they are not representing an invoice.

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

To make the invoice process more efficient and environmentally friendly, it is unnecessary to perform the energy intensive OCR step on PDF and image files that later in the process will be deleted since they are not representing an invoice.

The problem to solve is to find a technique/algorithm that could classify if the PDF or image file is an invoice or not. The accuracy of the classification step should be above 90% and it should be fast (below 1 second per file).

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