Digitization of Health Records using Spark OCR

Alina Dia TRAMBITAS MIRON*, Andrei Marian FEIER*, Marius MĂRUŞTERI, and Andrei IOANOVICI

"George Emil Palade" University of Medicine, Pharmacy, Sciences and Technology, Faculty of Medicine, 540142 Târgu Mureş, Romania
E-mails: dia.trambitas-miron@umfst.ro; andreifeier@gmx.com; marius.marusteri@umfst.ro; andrei.ioanovici@umfst.ro

* Author to whom correspondence should be addressed

Abstract
Processing data in the healthcare domain often involves extracting information from documents with complex and heterogeneous formats such as forms, lab results, academic papers, receipts, genomic sequencing reports, signed legal agreements, clinical trial documents, application forms, invoices, etc. Those documents are usually available in paper format and their digitization and analysis in a secured, integrated and accurate manner remains a challenge. In this presentation we will explain our approach to patient charts digitization using Spark OCR library and the transformers it offers, with concrete code examples and obtained results. Spark OCR library enables the processing of documents privately without uploading them to a cloud service; and most importantly, provides state-of-the-art accuracy for a variety of common use cases. A primary method of maximizing accuracy is using a set of pre-built image pre-processing transformers - for noise reduction, skew correction, object removal, automated scaling, erosion, binarization, and dilation. These transformers can be combined into OCR pipelines that effectively resolve common 'document noise' issues that reduce OCR accuracy.

Keywords: Health record digitization; Spark OCR; Transformers