
DocStreet
DAVINCI

DocStreet

WHITEPAPER

Op zoek naar een manier om automatisch dossiers te ordenen

Met DocStreet kunt u op een veilige manier, vanuit de cloud, automatisch grote aantallen documenten analyseren en verwerken. De oplossing is uitermate geschikt in het aanvraagproces van hypotheken en consumptief krediet. Daarbij worden grote hoeveelheden documenten verwerkt en is er behoefte aan het verkorten van doorlooptijden.

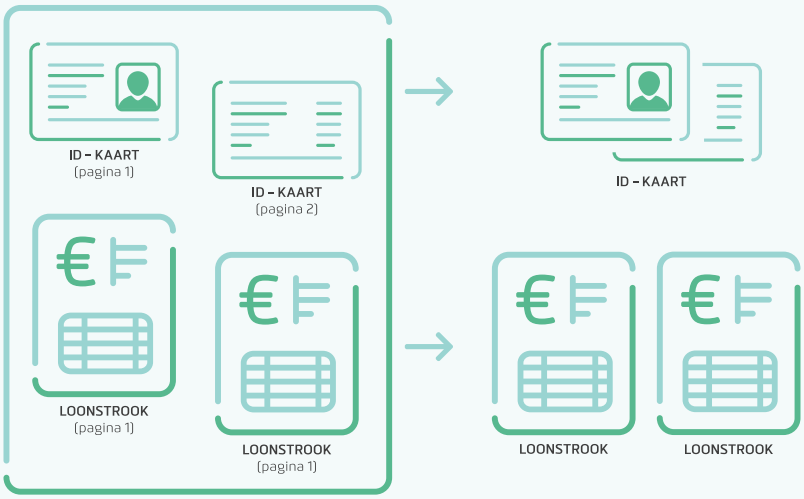
"Op basis van machine learning kunnen we ongeorganiseerde dossiers van klanten automatisch laten ordenen. Dat voorkomt fouten in de volgende stappen van het proces, en een hoop handwerk!"

Wessel Stoop
Machine learning-specialist bij Davinci
ook verbonden aan de Radboud Universiteit Nijmegen

DOCSTREET

DocStreet zorgt er door middel van data herkennings- en extractietechnieken voor dat ongestructureerde informatie uit afbeeldingen (pdf, tiff, jpg, etc. van gescande documenten) en searchable pdf's, omgezet wordt in gestructureerde informatie. Dit maakt het mogelijk om dossiers automatisch te verwerken, inclusief het automatisch invullen van de aanvraag en het aanvullen en verifiëren van de data. Een uitdaging voor DocStreet die Davinci onlangs aangepakt heeft is het verwerken van bestanden waarin meerdere documenten zitten. Denk aan een hypotheekadviseur die een loonstrook en een rijbewijs van een klant in één pdf bewaart, en die ook graag in één keer wil verwerken.

Tot voor kort betekende dat veel handwerk, nu kan het grotendeels automatisch.



Het hele proces: van ongestructureerde dossiers naar losse bestanden per document.

MACHINE LEARNING IN DOCSTREET

Om DocStreet zo intelligent mogelijk te maken, wordt gebruik gemaakt van machine learning. Eén van die machine learning-modules betreft het automatisch herkennen van het documenttype. Als er bijvoorbeeld een nieuwe scan binnenkomt, is het de vraag of dit een salarisstrook of een paspoort is.

DocStreet kan dit herkennen omdat het veel voorbeelden heeft gezien van deze documenttypes, en zichzelf heeft geleerd welke woorden nuttig zijn om de verschillende documenten uit elkaar te houden. Dit principe wilden we graag uitbreiden naar het automatisch ordenen van binnenkomende dossiers. In dit whitepaper een beschrijving van onze experimenten, en hoe ver we daar nu mee zijn.

EEN EERSTE POGING

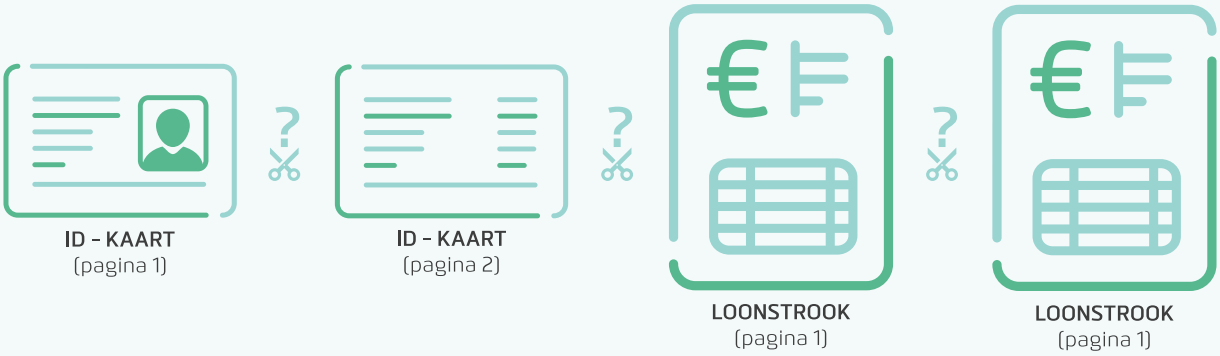
De eerste poging was om van alle pagina's één voor één te bepalen welk documenttype het is, en aan te nemen dat groepjes pagina's van hetzelfde type bij hetzelfde document horen.

Dit gaf al enige richting, maar had twee grote nadelen: 1) pagina's in de verkeerde volgorde werden er niet mee opgelost en (2) als toevallig twee documenten van hetzelfde type achter elkaar zaten, werd dit gezien als één document. In het plaatje hierboven bijvoorbeeld zouden de twee loonstroken ten onrechte aan elkaar worden geplakt.

OVER DAVINCI

Davinci is een Europese, ervaren softwareleverancier en ICT-adviesorganisatie met vestigingen in Nederland, België en Slowakije. Al 25 jaar helpen wij onze klanten door middel van onze businesskennis met complexe vraagstukken op het gebied van consumptieve en hypothecaire kredietverstrekking, kredietbeheer en bijbehorende businessprocessen.

Daardoor beschikken wij inmiddels over een bewezen track record en onderscheidende resultaten en oplossingen. Sinds enige tijd zijn wij gestart met het leveren van diensten vanuit de cloud om op die manier onze klanten maximaal te kunnen ontzorgen.



Het systeem evalueren: waar is er geknipt, en waar had geknipt moeten worden?

PAGINANUMMERHERKENNING

De volgende poging was om de machine learning applicatie te leren het paginanummer te herkennen. Dit is gedaan door de applicatie voorbeelden te laten zien van eerste pagina's, tweede pagina's, derde pagina's, etc, zodat de applicatie zichzelf kon leren hoe bijvoorbeeld 'een tweede pagina' er nu typisch uitziet. Dat ging beter dan verwacht; de applicatie deed natuurlijk veel met de paginanummers die vaak onderaan of in de hoekjes van pagina's ziet, maar er blijken ook veel andere clues in documenten te staan die iets kunnen zeggen over het paginanummer. Denk bijvoorbeeld aan de vishaken (<<<<) die achterop een identiteitskaart staan. De paginanummeraanpak bleek accurater dan de vorige, maar had een andere uitdaging. Zodra, zoals in het voorbeeld hierboven, een scan bestaat uit een pagina 1, een pagina 2, en dan weer een pagina 1 is het de vraag of pagina 2 bij de pagina ervoor of erachter hoort. Uiteindelijk bleek de combinatie van de twee methodes het beste te werken. Van iedere pagina wordt dus zowel het documenttype als het paginanummer bepaald. Vervolgens worden de pagina's met verschillende paginanummers maar hetzelfde documenttype aan elkaar geplakt. Als dan bijvoorbeeld geteld wordt op hoeveel van de plekken waar een knip had moeten staan ook daadwerkelijk een knip is gezet (zie het plaatje bovenaan deze pagina), komt daar een hitratio van 86% uit. Daarbij is van alle knippen die gezet worden 98% correct. Uit deze experimentele resultaten blijkt dat er een veelbelovende methode is gevonden voor het automatisch herkennen en ordenen van dossiers, waardoor zowel fouten in de vervolgstappen van het proces als handwerk voorkomen wordt.

Er is dus zeker nog ruimte voor verbetering, bijvoorbeeld door er meer trainingsmateriaal in te stoppen, maar de methode is al heel erg bruikbaar!

NEEM CONTACT MET ONS OP VOOR EEN DEMO:

+31 (0)20-5503750
info@davincigroep.nl
www.davincigroep.nl

Davinci, Atlas Arena gebouw Azië
Hoogoorddreef 5, 1101 BA
Amsterdam, Nederland

v.11.4.17

DAVINCI