



DocStreet

DAVINCI

DocStreet

WHITEPAPER

Op zoek naar een manier om te controleren op handtekeningen.

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 hypotheek en consumptief krediet. Daarbij worden grote hoeveelheden documenten verwerkt en is er behoefte aan het verkorten van doorlooptijden.



“Er staan veel dingen op een document die voor een computer op een handtekening lijken. We gebruiken zelflerende algoritmes om de juiste eruit te pikken en zo in de toekomst handwerk te besparen.”



Wessel Stoop

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

DOCSTREET

DocStreet zorgt er door middel van data herkenning- 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.

DocStreet heeft zich tot nu toe gefocust op de tekst van documenten, omdat daar 90% van de benodigde info uit te halen is. Er is echter een toepassing waar ook de grafische informatie van een document geanalyseerd moet worden: het vinden van handtekeningen. Voor een aantal van onze klanten is het namelijk essentieel dat op sommige documenten zoals een leningaanvraag een handtekening van de aanvrager aanwezig is; tot nu toe was dat een handmatige check.

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.

WAT WIL DAVINCI PRECIES KUNNEN?

Kort gezegd: weten of er een handtekening op een binnenkomend document staat. De software die we gebruiken voor het herkennen van tekst in binnenkomende documenten, trekt automatisch het grafische materiaal al los. 'Grafisch materiaal' kan van alles zijn: handtekeningen, maar ook logo's, foto's, stempels, barcodes, grafieken, kaartjes, etc. De vraag is dan of een van de gevonden plaatjes wel of geen handtekening is.

Wat deze taak moeilijk maakt, is dat er bijzonder veel randgevallen zijn: denk aan krabbels in de marge (geen handtekening) of handtekeningen die door de gedrukte tekst heen gaan. Door dit soort gevallen, die soms zelfs door mensen lastig te beoordelen zijn, is al op voorhand te zeggen dat automatische mechanismen nooit 100% van de gevallen correct zullen beoordelen; de vraag is of er desondanks een oplossing gevonden kan worden die goed genoeg is.

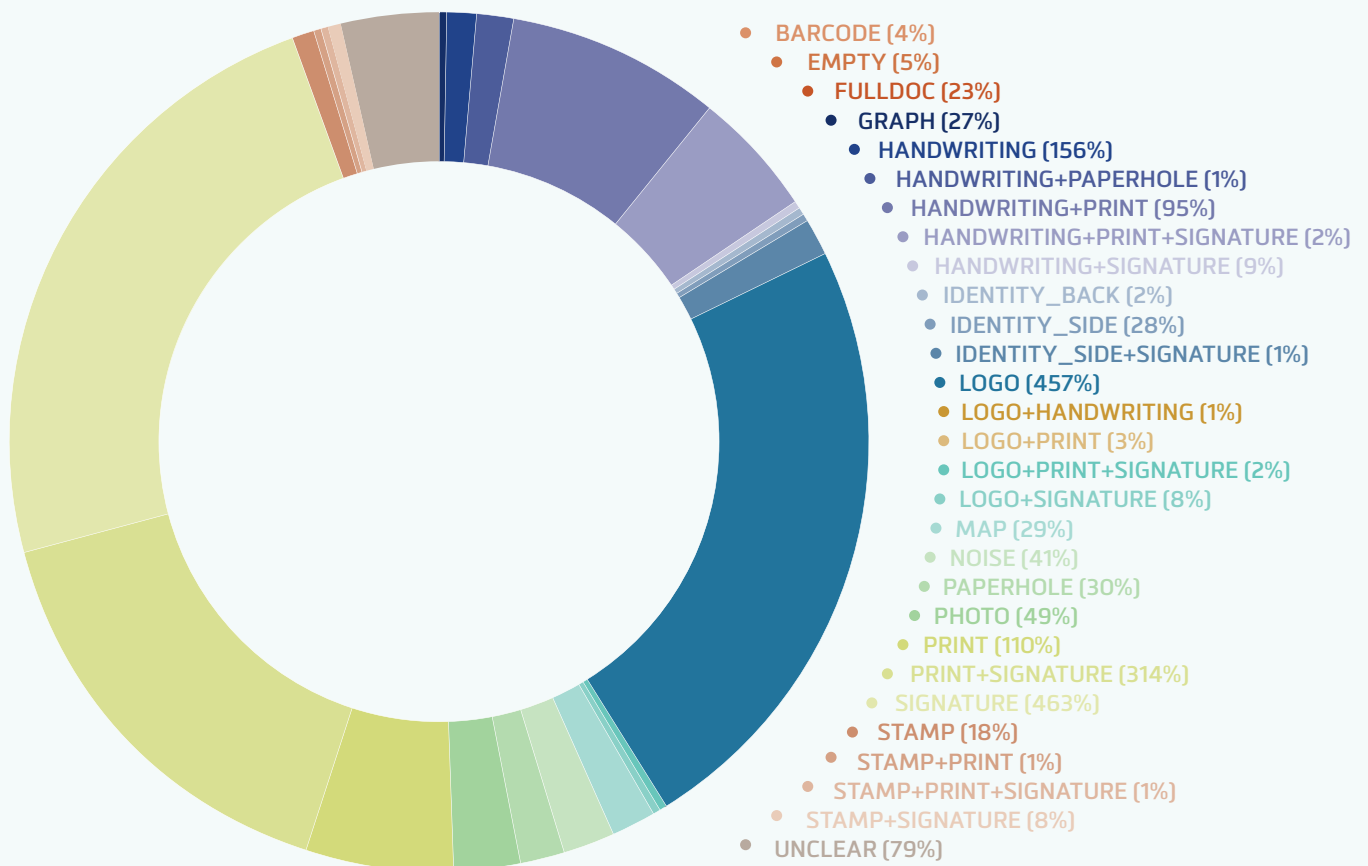
ONZE AANPAK

Als eerste stap is een dataset van 2000 plaatjes aangelegd, geknipt uit een grote verzameling realistische documenten en voor elk plaatje bijgehouden of het een handtekening, logo, stempel, etc. was. Vervolgens is een literatuurstudie (dat wil zeggen: de academische machine learning-literatuur) gedaan om erachter te komen hoe een systeem dat dit soort

classificaties automatisch kan maken zou kunnen werken. Hoe goed zo'n systeem presteert kan getest worden door te kijken hoe goed het de handmatige classificaties kan reproduceren.

Het idee van (supervised) machine learning is dat de computer niet verteld wordt hoe hij zijn werk moet doen, maar in plaats daarvan veel voorbeelden van het gewenste resultaat te zien krijgt; de computer leert dan zelf hoe de taak het beste aangepakt kan worden. Er zijn veel verschillende leermechanismes (algoritmes) beschikbaar die voor deze taak gebruikt zouden kunnen worden en het staat nooit van tevoren vast welke het beste zal werken. Uit de literatuur bleek al snel welke leermechanismes de beste kans maken voor een taak als deze; ze zijn allemaal uitgeprobeerd en telkens is geteld hoeveel handtekeningen correct als zodanig werden herkend.

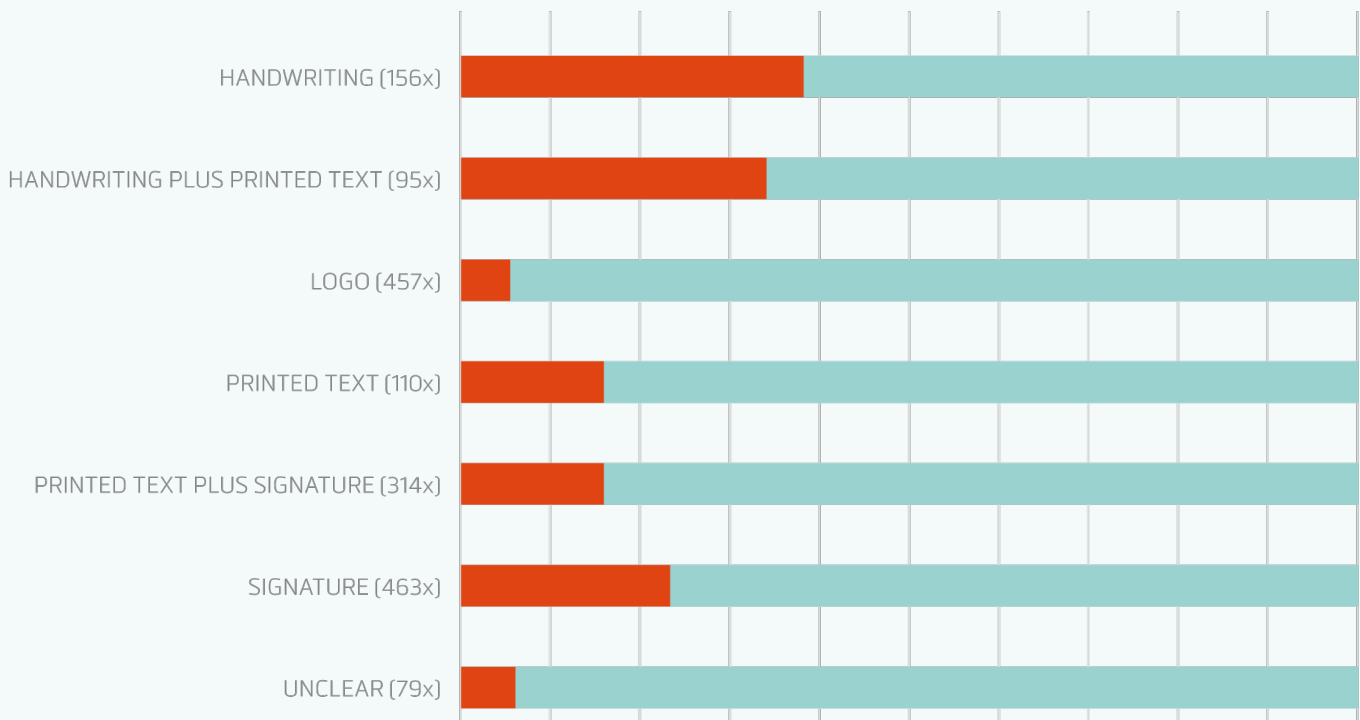
Een ander aspect van de taak waarmee gevarieerd kan worden is welke informatie het systeem als input krijgt. Dat kan de ruwe pixeldata van de plaatjes zijn, maar ook informatie over herkende contouren, de hoeveelheid zwarte vlakken, de aanwezigheid van aaneengesloten lijnen, etc. Ook hier is de literatuur gevolgd en zijn meerdere technieken die een goede kans maakten uitgeprobeerd en daarbij is telkens geteld hoeveel handtekeningen correct als zodanig werden herkend.



De verschillende soorten grafisch materiaal op documenten die DocStreet verwerkt.

DE RESULTATEN

Bij de best presterende combinatie (voor de liefhebber: het algoritme k-Nearest Neighbors met Local Binary Patterns en Fast Fourier Transform data als input) was het systeem correct in 82% van de keren dat het een plaatje als handtekening markeerde; daarnaast markeerde het 87% van de handtekeningen inderdaad als handtekening. Om meer inzicht te krijgen in waar het systeem goed en slecht presteert, is per categorie bekeken hoeveel plaatjes het juiste label kregen. Dit zijn de scores van de 7 categorieën die het meest voorkwamen:



Het percentage correct geïdentificeerde plaatjes per categorie.

We zien dat het systeem weinig moeite heeft met het herkennen van gedrukte tekst en logo's. Relatief veel fouten worden gemaakt met door mensen gemaakte krabbels, bijvoorbeeld in de marge – deze bleken later ook voor mensen niet altijd even makkelijk te onderscheiden.

Met andere woorden: ondanks dat er veel zaken op documenten staan die onze taak moeilijk maken, kan de handtekeningherkenner van Davinci al met grote zekerheid aangeven of er een handtekening op een document staat. Nog mooier: des te meer er aan het systeem gesleuteld wordt, des te meer de scores stijgen en des te geschikter het systeem wordt om de menselijke check te vervangen.

We werken bijvoorbeeld aan de functionaliteit waarbij de gebruiker zelf kan aangeven wat voor soort vergissingen hij of zij minder erg vindt: de meeste klanten van Davinci zullen het bijvoorbeeld minder erg vinden als er een handtekening gemist wordt (want dan gaat er toch nog een medewerker naar kijken) en in plaats daarvan 99.99% zeker willen weten dat er ook echt een handtekening staat als het systeem dat aangeeft. Het punt waarbij de menselijke check helemaal overbodig wordt is in zicht!

NEEM CONTACT MET ONS OP VOOR EEN DEMO:

v.9.8.17

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

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

DAVINCI