Eindrapport verrijkingsfase TRIADO

TRIADO-projectteam, mei 2019

Inhoudsopgave

Toelichting

In dit stuk worden de resultaten gepresenteerd van verschillende experimenten uitgevoerd in de 'verrijkingsfase' van het project TRIADO (februari 2018-december 2018). In deze fase stond de volgende vraagstelling centraal:

Welke digitale methoden zijn het meest geschikt (in termen van kwaliteit, efficiency, etc.) om grote corpora van ongestructureerde, imperfecte data, gebaseerd op analoge archiefcollecties, geschikt te maken als onderzoeksfaciliteit? Nadruk ligt hierbij op het toepasbaar maken van al beschikbare tools uit het digitale domein (van laboratorium naar reality check) en het verbeteren van de toegangen "wie, wat, waar en wanneer".

Uitgangspunt van de verrijking vormde 13,8 meter gedigitaliseerd CABR-materiaal (167.197 scans). Voor dit onderdeel bestond het TRIADO-projectteam uit:

- Lars Buitinck (Huygens ING/KNAW Humanities Cluster), computer engineer.
- Anne Gorter (Nationaal Archief), collectiespecialist.
- Edwin Klijn (Netwerk Oorlogsbronnen), projectleider.
- Rutger van Koert (Huygens ING/KNAW Humanities Cluster), computer engineer.
- Marielle Scherer (Huygens ING), datamanager.

Belangrijkste bevindingen

- De kwaliteit van de OCR is bij gebruik van Abbyy-software voldoende om het materiaal – met een zekere foutenmarge – doorzoekbaar te maken. Vooral goed leesbare, getypte stukken zoals processen-verbaal en besluiten worden goed herkend door de software (ca. 85% van de woorden worden correct omgezet). Dit zijn ook de meest informatierijke documenten in het archief met veel informatie over personen, plaatsen, gebeurtenissen.
- Abbyy en Tesseract hebben beide hun sterke kanten. Om de vindbaarheid te optimaliseren is het een goede strategie om meerdere 'lagen' met OCR-tekst te combineren.
- Met preprocessing van de images, machine-learning op basis van ground truth en post-correctie kan de OCR-score nog verder worden geoptimaliseerd.
- Automatische classificatie heeft potentie. Door de computer te trainen met voorbeelden kunnen soorten documenten met een acceptabele foutmarge (80% correct) worden herkend.
- Named entity recognition blijkt heel lastig. Het 'bottom-up' extraheren van personen, locaties of organisaties uit het CABR levert met de op dit moment beschikbare software matige resultaten op. Het matchen van bestaande databestanden met personen, locaties, organisaties etc. in de OCR van het CABR daarentegen lijkt goed te werken.
- Datumextractie is zinvol. Het verder aanscherpen van enkele scripts maakt het mogelijk datums er goed uit te halen (mits de OCR van voldoende kwaliteit is).
- Potentieel interessant: automatische clustering, topic modelling en SIFT matching (similarity searching).

Leeswijzer

In hoofdstuk 1 wordt de steekproef beschreven, die als uitgangspunt is genomen voor alle tests. Hoofdstuk 2 tot en met 5 beschrijven de bevindingen van de experimenten rondom automatische tekstherkenning, het herkennen van 'named entities', het koppelen aan externe databestanden en het automatisch classificeren van documenten. Hoofdstuk 6 gaat in op technologieën die potentieel interessant kunnen zijn, maar waar in beperkte mate tests mee zijn gedaan. Dit rapport sluit af met een glossarium waarin de belangrijkste technische termen worden uitgelegd. Alle uitdrukkingen die in de tekst ***vetgedrukt en cursief*** zijn, kunnen teruggevonden worden in het glossarium.

Dit rapport is gebaseerd op onderliggende gegevens, die als aparte bijlagen bij dit rapport horen:

- Bijlage A Scores testset A
- Bijlage B Scores testset B
- Bijlage C Uitgebreide beschrijving van de steekproef
- Bijlage D Interface TRIADO demonstrator

1. Beschrijving van steekproef

Voor het project is gewerkt met een steekproef bestaande uit 13,8 meter gedigitaliseerd materiaal (167.197 scans) uit het Centraal Archief Bijzondere Rechtspleging (CABR), berustende bij het Nationaal Archief (toegangsnummer: 2.09.09). Het CABR is samengesteld uit diverse kleinere archieven van onder andere gerechtshoven, openbare aanklagers en politiedepartementen. Vanwege de omvang van het gehele CABR (3,8 kilometer) en de grote diversiteit van het archiefmateriaal was het praktisch onmogelijk een representatieve selectie te maken. Bovendien bestond de noodzaak voor een inhoudelijke samenhang met het oog op de 'proof of concept' in het tweede deel van het project. Het onderzoek dat in deze fase centraal staat, richt zich op het in kaart brengen van de criminele carrières van personen met bijzondere aandacht voor sequentie en geografie. De collectiespecialist van het Nationaal Archief en de onderzoeker van het NIOD hebben gezamenlijk de selectie bepaald. De steekproef bestaat uit twee werksets:

Werkset 1:
Uit 13,8 meter gedigitaliseerde CABR-dossiers is een werkset samengesteld die de geografische spreiding van de dossiers over Nederland goed weergeeft. Uitgangspunt waren hierbij persoonsdossiers gevormd door politiedepartementen. Het CABR bevat honderd series persoonsdossiers samengesteld door tachtig politiedepartementen. Van elk van deze series is een dossier geselecteerd: uit de inventaris bij deze series is random een inventarisnummer gekozen. Van dit inventarisnummer werd het eerste dossier geselecteerd.[1] Van het dossier werd vervolgens vastgesteld wie de verdachte was. In de CABR-database werd nagegaan of er mogelijk meer dossiers over deze persoon aanwezig zijn. Indien dit het geval was, werden de dossiers opgezocht en toegevoegd aan de werkset. In totaal leverde deze wijze van steekproeftrekking 217 dossiers op en 1,2 meter archiefmateriaal.

Werkset 2:
Bovenstaande selectiemethodiek bleek enorm tijdrovend. Om tijd te besparen, maar wel geografische spreiding als uitgangspunt te houden, werd een nieuwe insteek bedacht. Voor de overdracht van het CABR naar het toenmalige Algemeen Rijksarchief (Nationaal Archief) is een aantal bewerkingen op het archief uitgevoerd. Hierdoor heeft een deel van de dossiers[2] een eigen inventarisnummer, waaronder de dossiers van de rechtsprekende instanties. In een database is opgenomen op wie deze dossiers betrekking hebben. Uit de database is een uitdraai gemaakt van alle rechtsprekende instanties met bijbehorende dossiers en inventarisnummers. Per instantie is vervolgens het aantal inventarisnummers vastgesteld. Daarna is berekend hoeveel inventarisnummers naar rato per instantie geselecteerd moesten worden. In totaal zijn voor deze werkset 1204 inventarisnummers aselect geselecteerd bestaande uit ongeveer 12,6 meter archief.

[1] In de meeste inventarisnummers zitten meerdere dossiers.
[2] Dossiers samengesteld door de rechtsprekende instanties (de Tribunalen, de Bijzondere Gerechtshoven, de Bijzondere Strafkamers en de Bijzondere Raad van Cassatie) die verspreid door Nederland gevestigd waren.

2. OCR-score

Methodiek

Van de 167.197 scans zijn er 167.122 ge-***OCR***'d.[3] 75 tiff-bestanden waren corrupt. Er is gebruik gemaakt van twee ***OCR***-programma's:

- Abbyy CLI OCR 11
- Tesseract 4.0.9 rc4

Voor de meting van de ***OCR***-kwaliteit zijn twee samples samengesteld waarvoor ***ground truth*** bestanden zijn aangemaakt:

- Set A bestaat uit een random sample van 100 documenten uit de 13,8 meter gedigitaliseerd materiaal. Set A bestaat uit verschillende typen documenten waaronder handgeschreven briefjes, foto's, formulieren en andere niet-getypte stukken. In de ***ground truth*** zijn 96 documenten getranscribeerd.[4]
- Set B bestaat uit een non-random sample van 150 documenten uit de 13,8 meter gedigitaliseerd materiaal. Geselecteerd zijn scans van processen-verbaal en besluiten, ofwel uitsluitend getypt materiaal. In de ***ground truth*** zijn handgeschreven marginalia niet getranscribeerd.

Er is gekozen voor een opsplitsing in twee datasets omdat enkel een random sample uit de gehele steekproef een vertekend beeld zou geven. De documenten die kunnen worden als het meest informatief en waardevol voor onderzoek - processen-verbaal, besluiten - zijn doorgaans getypt.

Scores

Set A

	Gemiddelde *WER*[5]	**Mediaan *WER***	**Gemiddelde *CER***	**Mediaan *CER***	**Gewogen *WER*[6]**
Abbyy	71.24	35.22	50.97	24.95	37.17
Tesseract	85.54	44.66	60.17	50.57	53.98
Tesseract darkened	272.95	69.54	132.11	48.55	52.23

[3] Vetgedrukt geeft aan dat de term in 7. Glossarium nader wordt toegelicht.
[4] De vier resterende documenten omvatten foto's en ander niet te transcriberen stukken.
[5] Alle WER-metingen in dit rapport betreffen WER independent ('bag of words').
[6] Percentage van het totaal aantal woorden van steekproef dat fout is omgezet.

Set B

	Gemiddelde WER	Mediaan WER	Gemiddelde CER	Mediaan CER	Gewogen WER
Abbyy	21.64	12.33	9.86	4.62	15.62
Tesseract	89.23	39.26	60.96	29.19	55.56

Er zijn verschillende zaken die opvallen aan deze scores:

- Buitensporig hoge ***WER***- en ***CER***-scores en grote onderlinge verschillen tussen documenten, in het bijzonder voor Tesseract. Over het algemeen leveren de documenten met veel getypte tekst de beste scores op. ***Hybride*** of handgeschreven documenten trekken de gemiddelde ***CER*** en ***WER*** flink omlaag. Dit komt vooral doordat dat Tesseract - en in mindere mate Abbyy - in het bijzonder bij lichte inkt op doorslagpapier woorden probeert te herkennen. Ook deelt Tesseract woorden vaak op in meerdere delen. Hierdoor ontstaan grote verschillen tussen het aantal woorden in de ***ground truth*** en het aantal woorden dat de software herkent. In de Tesseract-scores leidt dit soms tot een WER of CER die ver boven de 100 ligt. De neiging van Tesseract om documenten met verschillende soorten druk (een combinatie van typemachine, handschriften en voorgedrukte tekst) toch machineleesbaar te maken, verhoogt het aantal fouten maar ÓÓk het aantal correct omgezette woorden.
- Grote verschillen tussen testset A en B. Dit heeft te maken met de samenstelling van de testsets. Set A is een random sample met onder andere handgeschreven teksten, ***hybride*** formulieren en andere voor de software lastig te verwerken materiaal. Testset B is samengesteld uit hoofdzakelijk processen-verbaal en uitspraken. Dit zijn vrijwel uitsluitend typoscripten met een hoge tekstdichtheid.

De algemene conclusie is dat Abbyy voor typoscripten de beste initiële resultaten oplevert. In geval van processen-verbaal, besluiten en ander getypt materiaal met een hoge tekstdichtheid (set B) was er sprake van een gewogen score van 15.62 ***WER***. Dit percentage is vergelijkbaar met de ***WER*** van 19.45 die in de kleine steekproef van het project Full Automatic Archival Access werd gemeten.[7] Beide foutmarges zijn laag genoeg om een full-tekst zoekfunctie te ontwikkelen en data-verrijkingen toe te passen, maar er is duidelijk nog veel ruimte voor verbetering.

Ensemble-benadering

Met enige regelmaat kwam het voor dat Abbyy iets niet perfect herkent, maar Tesseract weer wel en vice versa. In de demonstrator waarmee de 13,8 meter kon worden doorzocht via een webinterface is gewerkt met verschillende OCR-'layers'. De 'ruis' neemt weliswaar toe (d.w.z. de ***precision*** nam af), maar dit nadeel lijkt vooralsnog niet op te wegen tegen het grote voordeel van een hogere ***recall***. Bij het

[7] Final report project Full Automatic Archival Access (FAAA) (Amsterdam 2016), www.oorlogsbronnen.nl/sites/default/files/FINAL%20REPORT%20project%20Full%20Automatic%20Archival%20Access_1.pdf.

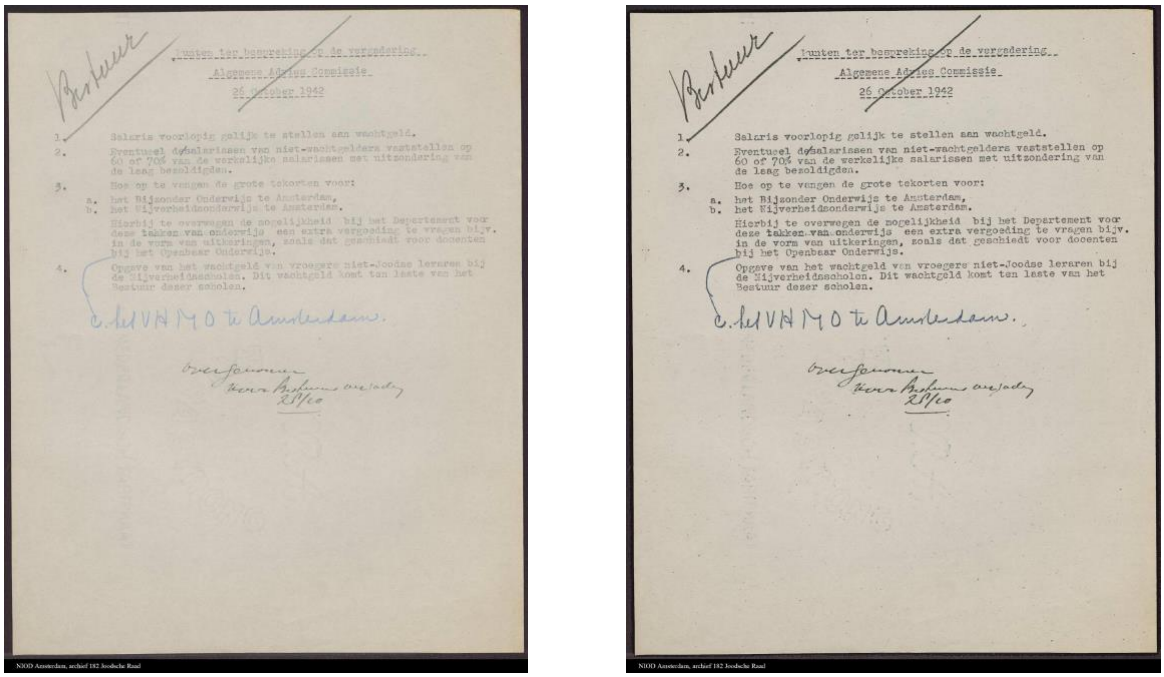
doorzoeken van het corpus heeft de gebruiker relatief weinig last van de 'ruis' in de ***OCR*** omdat het over het algemeen niet leidt tot namen maar juist tot niet-bestaande woorden.

De vindbaarheid van zoektermen kan naar verwachting fors worden verbeterd als op de achtergrond meerdere ***OCR-*** en ***Handwritten Text Recognition (HTR)***-tekstlagen worden gebruikt. In hoeverre het wenselijk en begrijpelijk is voor een eindgebruiker is iets om nader te onderzoeken.

Pre-processing images en machine learning

In TRIADO is geëxperimenteerd met een methodiek genaamd ***adaptive gaussian thresholding***. Pixels die op inkt lijken zijn gemarkeerd en daarna in het origineel zwarter gemaakt. Vervolgens is een pixelwaarde toegekend die ligt tussen de originele pixelwaarde en de pixelwaarde behorend bij zwart. De ***ground truth*** van set B in combinatie met de 'darkened' binaire images zijn vervolgens gebruikt als basis voor de ***machine learning*** met Tesseract. Set B is willekeurig verdeeld in twee delen: één voor training (80%) en één voor testen (20%).

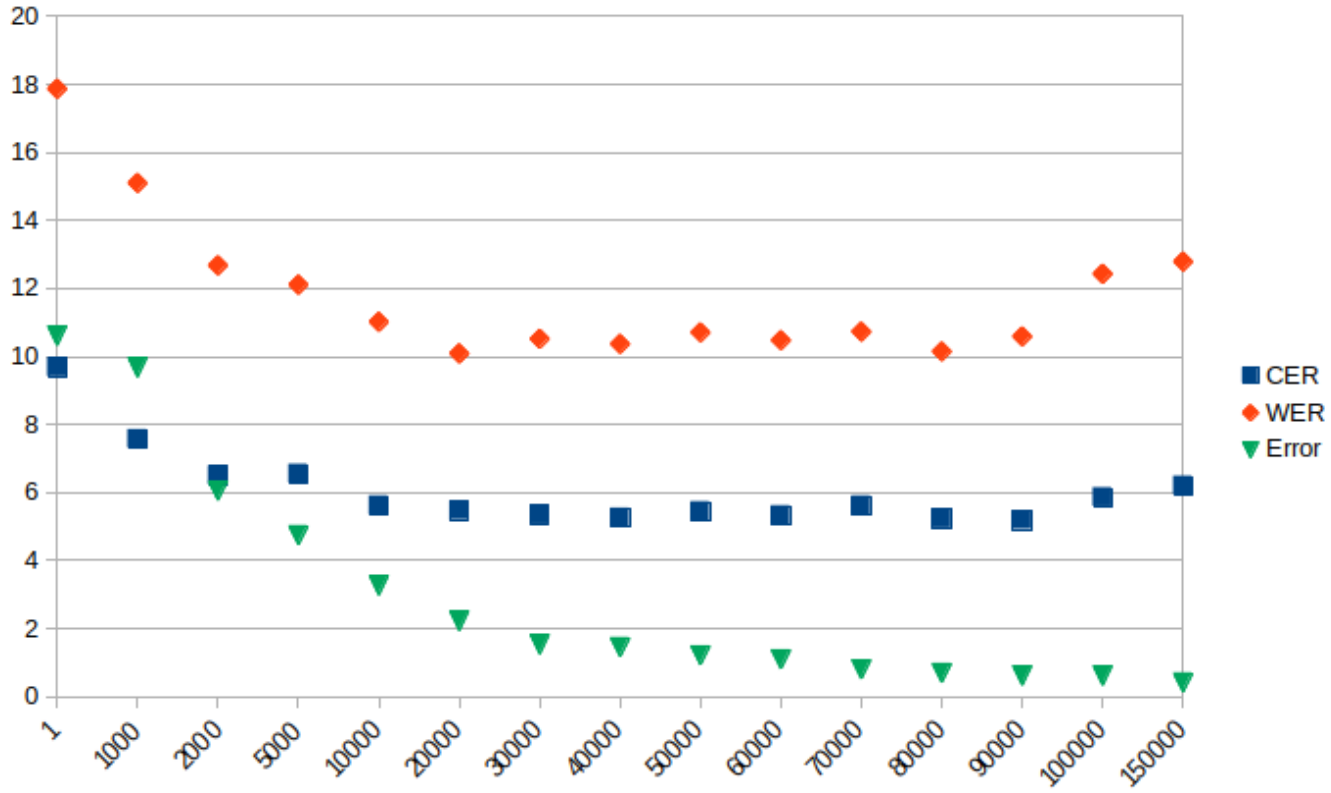
Na darkening (zie figuur 1) was de ***WER***-score van deze bewerkte images in Tesseract aanvankelijk 17.78. Dit is hoger dan de ***WER***-score van dezelfde images in onbewerkte staat. Dit is het gevolg van het feit dat Tesseract na darkening kleine vlekjes, oneffenheden e.d. voor inkt aanziet. De darkening creëert meer 'ruis', maar in geval van slecht zichtbare tekst worden uiteindelijk wel meer woorden herkend.



Figuur 1: effect van darkening (links origineel, rechts na darkening, hier op stukken uit het archief van de Joodsche Raad van het NIOD)

Bij training van Tesseract op darkened images daalde de ***WER*** naar ongeveer 10,37 (zie figuur 2). Tot 80.000-90.000 iteraties bleek er sprake van verbetering, daarna leek er sprake van overtraining.

Darkening in combinatie met machine learning blijkt dus een goede methode te zijn om de beperkingen van Tesseract tegen te gaan.



Figuur 2: aantal iteraties training Tesseract met darkening-methodiek

Aanbevelingen

Testen van OCR-resultaten in HTR *convolutional neural network* Transkribus

Tijdens deze pilot is alleen met Abbyy en Tesseract geëxperimenteerd in een streng beveiligde offline omgeving. Tests met software in de cloud waren hierdoor onmogelijk. Bij de experimenten met Tesseract blijkt dat de inzet van ***machine learning*** om betere ***OCR*** te krijgen, goede resultaten kan opleveren. In het READ-project wordt met Transkribus voor handgeschreven bronnen scores gehaald die lager liggen dan in TRIADO zijn gemeten voor getypt materiaal.[8] Zo werd recentelijk bij het transcriberen van 17e eeuwse handgeschreven notariële akten met enige handmatige training een CER gehaald van ongeveer 5.[9] Geadviseerd wordt dan ook eens met Transkribus een test te doen en de resultaten ervan te vergelijken met die uit het TRIADO-project. Niet alleen voor getypt maar ook handgeschreven of ***hybride*** materiaal lijkt Transkribus een goede optie te zijn.

[8] https://transkribus.eu.
[9] https://www.volkskrant.nl/wetenschap/nieuw-gereedschap-voor-historici-computers-die-handgeschreven-historische-papieren-omzetten-in-digitale-tekst.

Ensemble-benadering: meer ruis maar betere zoekresultaten

Voor het optimaliseren van het zoeken, blijkt een ensemble-aanpak - waarbij de resultaten van verschillende ***OCR***-output wordt gecombineerd - een goede manier om de vindbaarheid van woorden te vergroten. De 'ruis' neemt weliswaar toe, maar er kan wel meer worden teruggevonden.

Post-correctie met piccl en ticcl

Met automatische post-correctie-software (piccl en ticcl) is het mogelijk om veel voorkomende ***OCR***-fouten te verbeteren. Vanwege de technische omstandigheden in de standalone werkomgeving van TRIADO was toepassing van deze software niet mogelijk maar doorgaans leidt dit tot een eenvoudig te realiseren kwaliteitsverbetering. Veel voorkomende fouten met de letters 'l' en 'i' en 'e' en 'o' en woorden zoals als 'vrouw' en ***OCR***-equivalent 'vrouvv', kunnen met software worden verbeterd. Deze wijze van post-correctie kan worden uitgevoerd met behulp van een ***verwarringsmatrix***, ofwel een matrix die aangeeft wat de waarschijnlijkheid is dat een bepaald karakter verward wordt met een ander.

Aanpassen aan de taal in de originele documenten

De ***OCR***-software is afgesteld om enkele specifieke talen (Engels, Nederlands, Duits) te herkennen. Echter, bij een stuk waarvan we weten dat het Nederlands is, is het onzinnig en zelfs contraproductief om ook als optie Duits mee te geven. Vlekjes en kleine beschadigingen in de documenten kunnen ertoe leiden dat bijvoorbeeld een ß voorkomt in de transcriptie. Bij ingeregeld gebruik van enkel Nederlands als documenttaal worden dit soort fouten verminderd. De meest gebruikte taal in een document kan vrij betrouwbaar worden vastgesteld door het gebruik van ***n-grammen***. Deze n-grammen bepalen de waarschijnlijkheid dat een woord bij een bepaalde taal hoort.

3. Automatische classificatie

Methodiek

Het CABR is een divers archief met veel verschillende soorten documenten: formulieren, lidmaatschapskaarten, getypte correspondentie, besluiten, etc.. Er is geëxperimenteerd met ***machine learning*** om de computer te leren specifieke typen documenten automatisch te herkennen (***automatische classificatie***). In de demonstrator zijn de resultaten hiervan omgezet in filters ('type document').

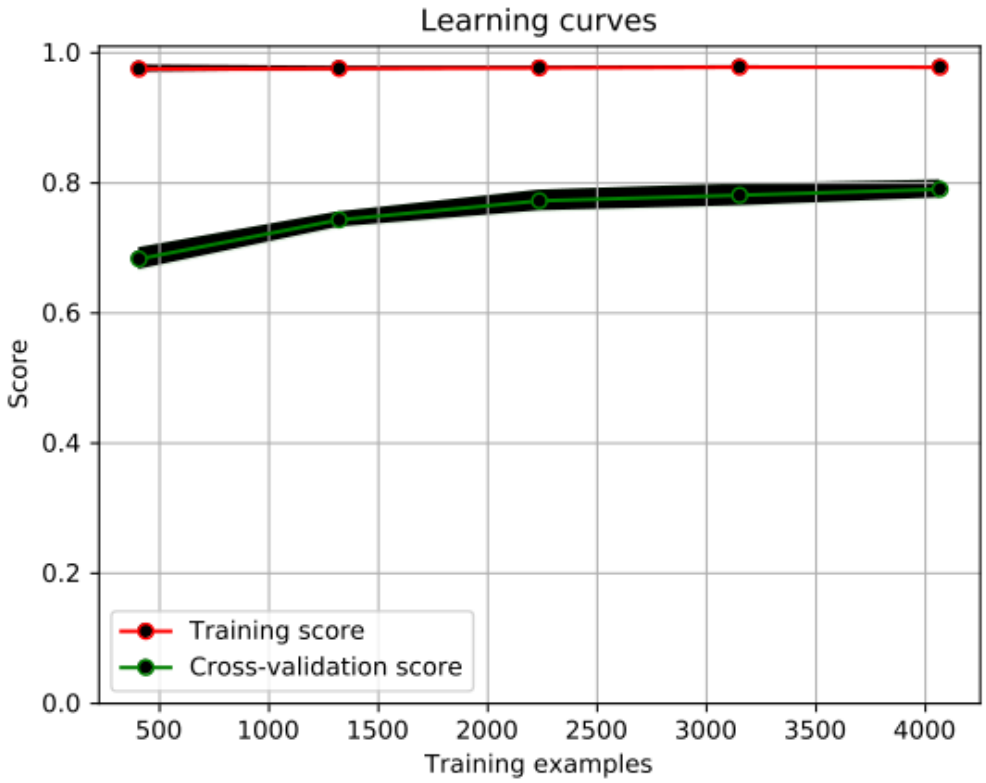
Uit de steekproef van 13,8 meter is een non-random selectie van 4768 scans handmatig in typen documenten opgedeeld. Vanuit deze ***ground truth*** set zijn 28 classes (zie figuur 4) nader uitgewerkt door twintig of meer voorbeelden te verzamelen en de computer te trainen deze documenten te herkennen. Dit waren classes met veel voorkomende documenten (bijvoorbeeld uittreksel processen-verbaal, vonnissen, getuigenverslagen). 80% van deze data is gebruikt voor training en de resterende 20% voor het testen. Bij het identificeren van typen documenten is de zogenaamde '***random forest methodiek***' toegepast op de tekstuele inhoud (Abbyy OCR) van de bestanden.[10] Aanvullend is er ook ***deep learning*** op de lay-out van de documenten uitgevoerd. De scans zijn gebinariseerd (zwart-wit

[10] P. Geurts, D. Ernst and L. Wehenkel, 2006. Extremely randomized trees. In *Machine Learning* 63(1): 3-42.

gemaakt) en verkleind naar een maximum van 150x150 pixels met behoud van de aspect ratio. Deze bewerkte images zijn vervolgens ingevoerd in een ***convolutional neural network*** dat gemodelleerd was naar het werk van Le Kang en gebruik maakte van DL4J.[11]

Scores

Aanvankelijk lag de overall ***accuracy rate*** van de automatische classificatie iets onder de 70%. Na ***machine learning*** kon dat worden verbeterd naar 80% (zie figuur 3). Het toepassen van ***machine-learning*** om classes te herkennen, blijkt dus zinnig te zijn. Daarbij was de learning curve nog stijgende aan het einde van het experiment. Met andere woorden: met meer trainingsdata had de score nog iets hoger kunnen uitvallen.



Figuur 3: effect van machine learning op accuracy rate van automatische classificatie (set B)

De zogenaamde ***confusion matrix*** (zie figuur 4) laat de verschillen per 'class' zien tussen de computerinterpretatie (predicated class) en de ***ground truth*** (actual class). Soms is een incorrecte interpretatie goed te verklaren. Zo zijn in de class processen-verbaal (class 13) 158 van de 211 pagina's herkend. De computer herkende in 23 gevallen de pagina als het type 'correspondentie_getypt' (class 3). Omdat deze documenten inderdaad erg op elkaar lijken is dat een goed te verklaren afwijking.

[11] Le Kang, Jayant Kumar, Peng Ye, Yi Li, David S. Doermann, Convolutional Neural Networks for Document Image Classification, 2014. In *ICPR '14 Proceedings of the 2014 22nd International Conference on Pattern Recognition*. Zie ook https://deeplearning4j.org.

		Predicted Class																												
		0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	Total
Actual Class	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
	1	0	0	0	3	0	1	0	0	0	0	0	0	0	3	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	9
	2	0	0	0	0	0	3	0	0	0	0	0	1	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9
	3	0	0	0	75	0	7	4	0	0	0	0	0	1	33	1	1	0	0	0	3	3	0	0	0	2	0	0	0	1	131
	4	0	0	0	8	1	2	0	0	0	0	0	0	0	11	0	2	0	0	0	0	1	0	0	0	0	0	0	0	0	25
	5	0	0	0	2	1	22	2	0	0	0	0	0	1	12	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	42
	6	0	0	0	2	0	5	17	2	0	0	0	0	0	12	0	4	0	0	0	1	0	0	0	0	0	0	0	0	0	43
	7	0	0	0	9	0	2	2	2	0	0	0	0	0	4	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	21
	8	0	0	0	3	0	2	0	0	0	0	0	0	0	2	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	8
	9	0	0	0	0	0	0	1	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4
	10	0	0	0	0	0	0	0	0	0	0	30	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	31
	11	0	0	0	0	0	1	0	0	0	0	0	0	0	2	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	5
	12	0	0	0	2	0	6	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10
	13	0	0	0	23	0	7	1	1	0	0	2	0	0	158	3	10	0	0	0	0	1	0	2	0	2	0	1	0	0	211
	14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	11
	15	0	0	0	3	0	1	0	0	0	0	0	0	0	11	0	26	0	0	0	0	1	0	0	0	0	0	0	0	0	42
	16	0	0	0	2	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4
	17	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	8
	18	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
	19	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	4
	20	0	0	0	2	0	0	1	2	0	0	0	0	0	2	1	0	0	0	0	0	7	0	0	0	0	0	0	0	1	16
	21	0	0	0	4	0	1	3	0	0	0	0	1	0	23	1	2	0	0	0	1	0	1	1	0	0	0	0	0	0	38
	22	0	0	0	10	0	0	0	0	0	0	0	0	0	2	2	0	0	0	0	1	0	0	0	0	0	0	0	0	0	15
	23	0	0	0	1	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	4
	24	0	0	0	0	0	0	0	0	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	7
	25	0	0	0	9	0	4	0	0	0	0	0	0	0	5	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	21
	26	0	0	0	2	1	0	0	0	0	0	0	0	0	4	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	9
	27	0	0	0	2	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	5
	28	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2
	Total	0	0	0	165	3	66	31	9	0	3	32	1	3	307	21	54	2	0	1	11	13	2	4	0	7	0	1	0	3	

Class no.	Description	Class no.	Description
0	beslissing_voorwaardelijke buitenvervolgstelling	15	rapport
1	besluit	16	rapport_pra
2	bevel_tot_dagvaarding	17	sententie
3	correspondentie_getypt	18	soldbuch
4	correspondentie_handgeschreven	19	staat_van_dienst_in_beweging
5	dagvaarding	20	staat_van_inlichtingen
6	gerechtelijk_schrijven	21	uitspraak
7	handgeschreven_tekst	22	uittreksel_burgerlijke_stand
8	inventaris	23	verhoor_beschuldigde
9	maandcontributie	24	verhoor_van_getuigen
10	manuscript	25	verklaring_getypt
11	openbare_terechtzitting	26	verklaring_handgeschreven
12	oproeping	27	verzoek_inlichtingen
13	proces_verbaal	28	vragenlijst

14	proces_verbaal_uittreksel		

Figuur 4: confusion matrix met scores van automatische classificatie(set B)

Aanbevelingen

Meer training, ensemble learning en input van gebruikers

Automatische classificatie-algoritmen kunnen nog worden verbeterd door tekstuele en visuele kenmerken slimmer te combineren (***ensemble learning***). De behaalde ***accuracy rate*** van circa 80% stemt hoopvol, temeer omdat de curve aan het einde van het experiment nog altijd stijgende was. Meer training zou de ***accuracy rate*** nog kunnen verbeteren. Het trainen van de computer voor bepaalde classes kost wel veel machinetijd. Het CABR bestaat uit naar schatting meer dan honderd documenttype, dus bij een vervolgproject zou een keuze moeten worden gemaakt op basis van inhoudelijk belang.

Reconstructie van de rechtsgang

Als 'classes' met een acceptabele foutmarge herkend kunnen worden, zou het ook mogelijk zijn om de specifieke stukken uit de dossiers (dagvaarding, processen-verbaal, besluit, cassatie) voor een eindgebruiker eruit te filteren.

4. Named entity recognition

Methodiek

Om named entities (personen, organisaties, locaties, producten, gebeurtenissen, overig) in de ***ground truth***-bestanden van set A en B te herkennen, is FROG-NER-software met standaardinstellingen gebruikt.[12] Vanwege de beperkingen in technische omgeving is de ***tokenizer*** vervangen door die van Apache OpenNLP.[13]

Ten behoeve van de kwaliteitsmeting zijn met de software tool BRAT handmatig alle ***named entities*** getagd conform de handleiding van Hendricks et al.[14] Vervolgens zijn voor set A en set B de volgende drie variabelen gemeten:

- De '***precision***'-waarde (true positives/(true positives + false positives). Deze indicator beantwoordt de vraag: hoeveel van wat we vinden is correct?
- De '***recall***'-waarde (true positives)/(true positives +false negatives). Deze indicator beantwoordt de vraag: hoeveel van alle named entities uit de 'ground truth' vinden we terug?

[12] https://languagemachines.github.io/frog.

[13] https://opennlp.apache.org.

[14] Voor BRAT zie http://brat.nlplab.org.Iris Hendrickx, Antal van den Bosch, Maarten van Gompel, Ko van der Sloot and Walter Daelemans, Frog. A Natural Language Processing Suite for Dutch. Centre for Language Studies and Centre for Speech and Language Technology, Radboud University, *CLST Technical Report* 16-02, https://github.com/LanguageMachines/frog/raw/master/docs/frogmanual.pdf.

- De ***F1-waarde*** (2*precision*recall)/(precision+recall). Deze indicator geeft een harmonisch gemiddelde tussen beide.

Alle waarden variëren van 0 (alles fout) tot 1 (perfect)

Scores

	Set A	**Set B**
Precision	0,271	0,281
Recall	0,253	0,387
F1	0,261	0,325

Figuur 5: resultaten named entity recognition met FROG-NER

De ***NER***-scores (zie figuur 5) waren zowel voor set A als set B laag. Set B scoorde relatief iets beter omdat dit lopende tekst is en de gebruikte ***NER***-software getraind is op krantenartikelen. Er zijn verschillende redenen voor de matige score. Moderne krantenteksten, waar de FROG-NER vooral goed bij werkt, wijken in taalgebruik af van documenten rond de periode van de Tweede Wereldoorlog. Zo is bijvoorbeeld het hoofdlettergebruik anders; namen van functietitels zoals “Wachtmeester” zijn vaak met een hoofdletter geschreven, waar dit in het hedendaags Nederlands niet meer zo geschreven wordt. Andere fouten werden veroorzaakt door Duitse zelfstandige naamwoorden die met hoofdletters gespeld waren.

Verrijking met bestaande lijsten

Het matchen van bestaande namenbestanden lijkt vooralsnog veel beter te werken. Er is in deze pilot geëxperimenteerd met de databestanden van de Nationale Database Vervolgings Slachtoffers (NDVS), de Oorlogsgravenstichting (OGS) en het Centraal Archief Bijzondere Rechtspleging (database van verdachte personen). Er is geëxperimenteerd met het matchen van namenlijsten op drie verschillende tekstlagen (op basis van Abbyy, Tesseract en ‘Tesseract met darkening en extra training’).Ondanks de beperkte testperiode bleek het al goed mogelijk in de steekproef van 13,8 meter enkele slachtoffers positief te identificeren. Het lukte eveneens om verdachten die in meerdere CABR-dossiers werden vermeld te identificeren.

De experimenten waren vooral exploratief. Zo zijn er geen exacte metingen verricht. Ook is er gewerkt met redelijk rudimentaire matchingstrategieën, die nog niet verfijnd zijn om te anticiperen op naamsvarianten, kleine spellings- en/of ***OCR***-fouten, geboorteplaats en geboortedatum in de nabijheid van de naam, etc. Er is een matcher gemaakt die op basis van ***Levenshtein-distance*** en een ***verwarringsmatrix*** een koppeling probeerde te maken tussen de namen in de ruwe ***OCR*** en die in de databestanden. Waar bij gangbare Levenshtein-functies de afstand tussen “i” en “1” één zal zijn, maakt de TRIADO-matcher gebruik van een ***verwarringsmatrix*** waarbij inruil van één letter door een specifiek

andere gekwantificeerd is in de mate van waarschijnlijkheid. Zo kunnen veel voorkomende ***OCR***-fouten grotendeels omzeild worden door ze lage 'vervangingskosten' mee te geven. De vervanging van "1" door "i" en vice versa definiëren wij bijvoorbeeld op 0.1. Daarna wordt er gezocht naar matches met alleen zeer lage aangepaste ***Levenshtein***-afstanden. Ook voor 'klinkt-als'-varianten kan deze methodiek worden toegepast.

Sommige lijsten bevatten de volledige officiële voornamen, waardoor het matchen moeilijk wordt. In de tekst kan een "Albert Klein"worden genoemd, die in een lijst "Albertus Johannes Klein" heet. Met normalisatietechnieken is het mogelijk deze te matchen, zeker wanneer ook bijvoorbeeld geboortedatum of geboorteplaats - vaak genoemd in processen-verbaal - in de waarschijnlijkheidsberekening worden meegenomen.

Aanbevelingen

Meer training van NER-software

NER is vooral bedoeld om namen 'bottom-up' uit databestanden te extraheren. Gezien de aard van het CABR-archief blijft het naar verwachting complex om hier named entities uit te halen, maar het trainen van FROG voor historische documenten uit de periode 1930-1950 zouden de resultaten wel ten goede komen.

Disambigueren van named entities

Een aanvulling op de NER is het maken van koppelingen met bestaande databestanden/lijsten van personen, organisaties, geografische locaties, etc. Met behulp van ***reguliere expressies*** kunnen named entities in 'voorspelbare' zinsconstructies automatisch worden geëxtraheerd en gekoppeld aan andere bestanden. Het herleiden van named entities tot één persoon of één locatie (disambigueren) is lastig, maar wel wenselijk. Door bijvoorbeeld regular expressions te maken die naam-geboortedatum-geboorteplaats (doorgaans unieke combinatie) uit processen-verbaal kunnen halen, kan de matching met externe bestanden nog verder verbeterd worden.

5. Datumextractie

Methodiek

De software-tool BRAT is geconfigureerd om datums te markeren. Vervolgens zijn in de ***ground truth***-documenten van set A en set B de volgende varianten getagd: Dag/maand/jaar; Maand/jaar; Dag/maand; Jaar.

Vervolgens is met een eenvoudig script in de ***ground truth*** van set A en B gezocht naar datums. Het resultaat hiervan is vergeleken met de handmatig getagde datums.

Scores

	Set A	Set B
Precision	0,688	0,571
Recall	0,284	0,707
F1	0,402	0,632

Figuur 6: resultaten datumextractie

Over het algemeen zijn de resultaten van de datumextractie (zie figuur 6) bemoedigend. De hoge mate van voorspelbaarheid maakt dat datums - in de verschillende varianten - redelijk goed door de software uit de ground truth van de beide test sets kon worden gehaald. De cijfers van set A en set B zitten bij elkaar in de buurt, alleen de ***recall*** wijkt om onbekende redenen af.

Aanbevelingen

Verdere aanscherping van extractie-script

Ter verbetering van de software kan ***machine learning*** worden toegepast. Post-correctie van ***OCR***, al dan niet in combinatie met een ***verwarringsmatrix***, kan de foutmarge doen verkleinen aangezien de voorspelbaarheid van datums heel groot is.

6. Experimenteel

Topic modelling

Methodiek

Topic modelling-software gebruikt statistische modellen om "topics" ofwel “onderwerpen” te vinden in collecties met documenten. Het is een handige manier om een soort korte samenvatting in steekwoorden te krijgen van de inhoud van een of meerdere documenten. Op basis van de ge-OCR’de documenten uit testset B zijn er twee topic modelling-methodieken toegepast: ***Non-negative matrix factorization*** (NMF, Lee en Seung 1999) en ***Latent Dirichlet Allocation*** (LDA, Blei et al. 2003).[15] Beide gaan uit van een verdeling van documenten over topics, onderwerpen die op zichzelf weer verdelingen over woorden zijn. Elk document behoort in meerdere of mindere mate tot elk van de topics, en net zo voor ieder woord dat in de documentcollectie voorkomt. Voor de test in TRIADO is om praktische redenen elke pagina als een document beschouwd. Omdat de bestanden veel ‘ruis’ bevatten, zijn vooraf niet alleen stopwoorden maar ook alle woorden met lengte twee of kleiner verwijderd. Op de resulterende documenten is **NMF** en ***LDA*** toegepast.

[15] D. D. Lee en H. S. Seung (1999). Learning the parts of objects by non-negative matrix factorization. *Nature* 401:788–791 en D. M. Blei, A. Y. Ng en M. I. Jordan (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3: 993-1022.

Resultaten

De resultaten die eruit kwamen met ***NMF*** vertonen interessante patronen. Per topic worden de dertig hoogst gewaardeerde woorden getoond. Persoonsnamen zijn vanwege privacy-overwegingen uit de topics verwijderd. Er werden uiteindelijk in totaal negentien topics geïdentificeerd. Een aantal hiervan blijkt een goede indruk te geven van de inhoud:

Topic #6: beweging lid geboortedatum godsdienst geboorteplaats onderwijs adres auto functie dienst naam genoten voornamen datum rang diploma voluit arbeidsdienst motorrijder stamboek contributie lidmaatschap wapen partij opleiding zakboekje werkzaam eereteeken front ster[16]

Topic #16: zyn volken cultuur invloed volk athene stryd zoo caesar grieken rome groote indo geest land oude tyd nieuwe ryk staat geheel griekenland eeuw ran leger macht hyk tusschen verval blykt[17]

Topic #17: groningen landwachters <naam> <naam> gearresteerd slochteren <naam> siddeburen ondergedoken woning landwacht <naam> arrestatie <naam> overgebracht onderduikers west gemeente <naam> schildwolde duitsland onderduiker huiszoeking getuige landbouwer personen boerderij wonende <naam> <naam>[18]

OCR-fouten zorgen voor een zekere ruis in de topics, maar niet zodanig dat ze onbruikbaar worden. Sterker nog: het veelvuldig voorkomen van woorden zorgt ervoor dat foutieve ***OCR***-interpretaties automatisch buiten de boot vallen.

De resultaten voor het ***LDA***-model vielen wat tegen, mogelijk veroorzaakt door de grote hoeveelheid parameters die ingesteld kunnen worden. Voor dit experiment ontbrak de tijd om de software nauwkeurig te af te stellen.

Aanbevelingen

Beperken tot documenten met hoge tekstdichtheid en verder testen van LDA
Topic modelling is nuttig bij getypte documenten met een hoge tekstdichtheid. In combinatie met ***automatische classificatie*** van documenten als besluiten of processen-verbaal, zou deze methodiek verbeterd kunnen worden. ***NMF*** leverde in de test de meest bruikbare resultaten op, maar ***LDA*** zou nog meer moeten worden getest.

Automatische clustering

Automatische clustering is een methodiek die het mogelijk maakt soortgelijke documenten op basis van tekstuele en visuele kenmerken te classificeren, zonder dat er vooraf ***ground truth*** of variabelen worden

[16] NSB-lidmaatschapsboekje.
[17] Typoscript voor een HBS-leerboek vaderlandse geschiedenis opgesteld door een overtuigd nationaalsocialist.
[18] Voorval rondom het verraad van onderduikers in de provincie Groningen.

bepaald. Omdat experimenten hiermee veel rekenkracht vergen en voor het TRIADO-project een beperkte capaciteit beschikbaar was, bleek het niet mogelijk hiermee tests te doen. Wel lijkt het CABR in potentie uitermate geschikt voor automatische clustering met behulp van bijvoorbeeld een ***DIABOLO-auto encoder*** of meer in het bijzonder ***similarity search/scale invariant feature transform matching (sift-matching)***.[19] Met ***sift-matching*** kun je bijvoorbeeld eindgebruikers op soortgelijke documenten of soortgelijke delen uit documenten (bijv. briefhoofd, familiewapen etc.) laten zoeken.

7. Glossarium

Adaptive Gaussian thresholding: een methodiek waarmee grijswaarden in een afbeelding worden omgezet in binaire waarden (zwart of wit), zodat de voor- en achtergrond van elkaar kunnen worden onderscheiden.

Automatische classificatie: een set van technieken om documenten te categoriseren in een aantal van te voren vastgestelde klassen, zonder dat er een mens aan te pas hoeft te komen. Hiervoor moet de software eerst worden getraind.

Automatische clustering: een methodiek die het mogelijk maakt soortgelijke documenten op basis van tekstuele en visuele kenmerken te groeperen zonder dat vooraf ground truth, klassen of variabelen worden bepaald. De software hoeft hiervoor niet eerst te worden getraind.

CER (Character Error Rate): het quotiënt tussen het aantal door de OCR foutief geïnterpreteerde karakters en de lengte van de tekst.

Confusion matrix: een tabel die laat zien hoeveel classificaties juist (true positives, true negatives) zijn en hoeveel onjuist (false positives, false negatives) en waar het systeem classificaties verwisseld heeft.

Convolutional neural network (ConvNet, CNN): een type deep neural network dat veel gebruikt wordt bij het analyseren en classificeren van images. ConvNets worden vaak gebruikt bij het herkennen van gezichten, objecten of verkeersborden.

DIABOLO auto-encoder: een type kunstmatig neuraal netwerk. Deze encoder probeert te leren door zijn input te verkleinen en vervolgens op basis van de verkleinde informatie zijn originele input weer te recreëren.

Ground truth: informatie vastgesteld door middel van menselijke observatie, bijvoorbeeld door mensen getranscribeerde tekst of door mensen toegekende labels. Deze worden gebruikt om software te trainen en om te bepalen hoe goed het resultaat van geautomatiseerde analyses is.

Handwritten Text Recognition (HTR): het vermogen van een computer om handgeschreven tekst te kunnen interpreteren.

Hybride documenten: documenten die deels getypt deels handgeschreven tekst bevatten.

[19] https://www.cs.ubc.ca/~lowe/papers/iccv99.pdf.

Latent Dirichlet Allocation (LDA): een statistisch model voor topic modelling, dat het mogelijk maakt om via niet geobserveerde groepen van observatiesets te verklaren waarom sommige data vergelijkbaar is. De data wordt als Dirichlet distributies gemodelleerd, ofwel een statistische modellering om bepaalde groepen documenten te onderscheiden van elkaar op basis van onderwerp (topic). Het topic kan hier het best gezien worden als een aantal kernwoorden waarmee de verschillende documenten met elkaar samenhangen.

Levenshtein-distance: een indicator die de minimale hoeveelheid bewerkingen aangeeft die nodig zijn om de ene tekenreeks in de andere te veranderen. Als je bijvoorbeeld zoekt met een Levenshtein-distance van 2, mogen twee karakters in een woord afwijken.

Named entity recognition (NER): een manier om entiteiten met een naam (zoals personen, plaatsen, organisaties, merken, etc.) uit ongestructureerde tekst te extraheren.

N-gram: een opeenvolgende sequentie van *n* items van een bepaald stukje geschreven of gesproken tekst.

Non-negative matrix factorization (NMF): een topic modelling-methodiek waarmee de inhoud van tekstdocumenten wordt verdeeld over 'topics', ofwel onderwerpen die op zichzelf weer verdelingen over woorden zijn.

Optical Character Recognition (OCR): de conversie van afbeeldingen van getypte of handgeschreven tekst naar machine-geëncodeerde tekst.

Random forest methodiek: een methodiek om met meerdere decision trees classificatie en regressie te trainen. De random forest methodiek zorgt ervoor dat de decision trees niet te specifiek op de test dataset worden getraind.

Reguliere expressies: een manier om patronen te beschrijven waardoor een computer softwarematig tekst kan identificeren.

Scale-invariant feature transform (SIFT)- matching: een wijze om via specifieke kenmerken soortgelijke afbeeldingen te vinden.

Tokenizer: software die ervoor zorgt dat karakters worden gesegmenteerd in woorden.

Topic Modelling: een methodiek waarbij door middel van een matrixvergelijking 'topics' ofwel onderwerpen uit de tekst van één of meerdere documenten worden geëxtraheerd.

Verwarringsmatrix: zie confusion matrix

Word Error Rate (WER): de quotiënt tussen het aantal door de OCR foutief geïnterpreteerde woorden en het totale aantal woorden van de tekst.

WER-independent ofwel 'bag of words': een vereenvoudigd model in natural language processing, waarbij de grammatica en woordvolgorde worden genegeerd. Alleen het aantal keer dat het woord voorkomt, is van belang. Dit model wordt over het algemeen gebruikt bij documentclassificatie.