Enorme stap om archieven voorwaarts toegankelijk te maken

Het Centraal Archief Bijzondere Rechtspleging telt ruim 300.000 dossiers van personen die na de Tweede Wereldoorlog zijn onderzocht omdat ze werden verdacht van ‘collaboratie’. In het project Tribunaal-archieven als Digitale Onderzoeksfaciliteit (TRIADO) worden de mogelijkheden om dit archief digitaal te ontsluiten onderzocht.


In 2016 startten het Nationaal Archief, Huygens ING, NIOD en Netwerk Oorlogsbrommen (NOB) het project Tribu- naarchieven als Digitale Onderzoeksfaciliteit (TRIADO). De bedoeling was om op basis van een steekproef uit het Centraal Archief Bijzondere Rechtspleging (CABR) de mogelijkheden van digitale ontsluiting te verkennen door te experimenteren met ‘proven technology’, oftewel bestaande software, zowel open source als commercieel.

Twee pakketten getest

Een scan van een tekstdocument is voor een computer niet veel anders dan een vakantiefoto. Pas wanneer tekst-herkenningsoftware de pixels transformeert naar letters en woorden, kan de tekst digitaal doorzoekbaar worden gemaakt. In TRIADO zijn tests gedaan met twee bekende OCR (Optical Character Recognition) pakketten: Abby Finereader en Tesseract. Beide hadden de beste resultaten, maar volledig geptyte documenten zoals procesverslagen en besluiten, Abby Finereader scoorde het best voor dit soort documenten: er werd een ‘word error rate’ (WER) ofwel het percentage foutief omge- zette woorden) gemeten van 15.62. Ter vergelijking: voor een kleine steekproef van historische kranten werd door de Koninklijke Bibliotheek een WER van 9 gemeten. Het minder resulteert voor archiefmateriaal versus drukwerk valt grotendeels te verklaren door de conditie van het papier (doorgangen), breukkeuze typemachine en vervuilde inkt.

Voor een archief zoals het CABR, met een hoge informatiedichtheid en veel feitelijke vermeldingen van namen, plaatsen, datums en gebeurtenissen, is een WER van 15 hoog maar goed genoeg om een zoekfunctie ontwikkelen te bouwen en nadere verfijningen aan te brengen. In TRIADO is vanwege de privacygevoelige informatie in het CABR binnen een streng beveiligde omgeving gewerkt. Het was niet mogelijk om deze te experimenteren met software die wordt ingezet voor handschriftsherkenning. In projecten zoals READ (read.transkrivaat.eu) worden hiermee inmiddels goede resultaten behaald.

Verschillen

Tests in TRIADO waren uit dat voor- bewerking van de scans door het zware maken van pixels die op intk lijken (‘darkening’), in combinatie met machine learning de kwaliteit van de machineleesbare tekst nog verder kon verbeteren. Met enkele handmatige tra- ing leerde de computer zichzelf om te leren. Er werden meer fouten gemaken, maar tegelijkertijd herkende de software ook meer woorden. Interessant was het verschil tussen Tesseract en Abby. Tesseract heeft de rein- ging om van alles op een pagina letters en woorden te maken. Abby daarentegen laat twijfelốgen buiten be- schouwing en is erg goed in het omschat- ten van relevante getypete tekst. Door dat Tesseract er heel veel extra letters en woorden bij verleent, scoort de lager in de WER-score. Uiteindelijk herkent het soms wel woorden die Abby negeert. In technische termen: de ‘precision’ (percentage van de gevonden woorden dat correct is) van Abby is hoger, maar de ‘recall’ (percentage woorden dat kan worden teruggeve- den) is lager. In TRIADO is ervoor gekozen met verschillende OCR-lagen (Abby, Tesseract, Tesseract, Tesseract) te werken om het beste van alle drie methodieken te benutten.

Zoekfilter

Slotconferentie over TRIADO
Op 13 september organiseert het Netwerk Oorlogsbronnen een slotconferentie over TRIADO. Meld u op oorlogsbronnen.nl aan voor de nieuwsbrief en blijf zo op de hoogte.

Inschrijving
VOOR WISSELSSTEENWARK BIE HET WINTERHUIS

1. Vraag: Nummer?

2. Voornaam:

3. Achternaam:

4. Adres:

5. Postcode:

6. Stad:

7. Land:

8. Email adres:

9. Telefoonnummer:

10. Vraag: Waarom?

UITVOERING VAN HUIZEN – OTTOFLIEP – KADE 6 – 1324 ZJ LEIDEN – PH: 017-612 52 52

Antwoord: We zullen de inschrijving meteen voorbereiden. 

Dit is een experimentele inschrijving op de website van de TRIADO-project. Indien u aangemeld bent, zult u alle relevante informatie ontvangen.

Autoclassificatie
Een van de belangrijkste onderdelen van TRIADO was de test met het automatisch herkennen van soorten documenten (autoclassificatie). Voor 28 type documenten zijn handmatig voorbeelden verzameld. Vervolgens is de computer getraind om op basis van deze voorbeelden soortgelijke documenten in de rest van de dertien meter te herkennen.

De software heeft zichzelf verder getraind met de gevonden resultaten (deep learning). In eerste instantie werd 70% procent van de documenten correct herkend. Machine learning leverde een verbetering op van circa tien procent. De computer voortouwde opvallend menselijke trekjes: processen-verbaal en getypeerde correspondentie werden vaak door elkaar gehaald, gedrukte lidmaatschapsschakelen tonen daarentegen goed worden onderscheiden. Conclusie: autoclassificatie heeft grote potentie om omvangrijke archieven digitaal nader toegankelijk te maken.

Topic modelling
Topic modelling-software analyseert machineleesbare tekst uit documenten en vat het samen in enkele veelvoorkomende steekwoorden. Een van de bijkomstige voordelen van topic modelling is dat door de OCR foutief of verzette woorden er eenvoudig kunnen worden uitgefilterd omdat ze vaak niet meer dan eenmaal voorkomen. Wat overblijft is bijvoorbeeld het volgende: Topic #17: Groningen landwachters en naam gedeburend ondergedoken woning landwacht naam arrestatie naam overgebracht onderduikers west gemeente naam schildwolde duitsland onderduiker huiszoeking getuige landbouwer personen boorden wonende naam naam naam

Op basis van deze staccato karakteriserende (de namen zijn weggelaten) kan je als gebruiker al een goede indruk krijgen waar de hieraan gekoppelde documenten over gaan. In dit geval een proces-verbaal over een voorval waarbij enkele Groningse Landwachters - hulpbestuursmensen gerekruiterd vanuit de Nationaal Sociaal-Wetenshappelijke Beweging - verdacht werden van hulp bij het oppakken van onderduikers.

Digitaal ontsluit van archieven
Na afloop van alle experimenten is er een demonstrator gebouwd. Wie nu zoekt op ‘Oostfront’, krijgt in een paar seconden 619 treffers met verwijzingen naar documenten waar dit woord gebruikt wordt. Dit is een wereld van verschil met de huidige status waarin het CABR alleen op dosierniveau is ontsloten op naam van verdachte.

Het TRIADO-project is nog niet afge- lopen. Tot juli zal NIOD-onderzoeker Ismee Tames verkennen welke toegevoegde waarde het digitaal ontsluiten van het CABR voor de wetenschap heeft. Nu al, met alle onvolkomenheden, kan geconcludeerd worden dat de inzet van digitale technologie een immense stap voorwaarts betekent bij het toegankelijk maken van archieven. Niet alleen ligt voor het eerst toegang tot op pagina niveau binnen bereik, ook biedt machineleesbare tekst uitkomende aanknopingspunten om de informatie in verschillende archieven aan elkaar te koppelen. Zover zijn we nog niet. Eerst moet er meer gedigitaliseerd worden en vooral machineleesbaar gemaakt.

Lees verder
> Onderzoeksrapport TRIADO, zie www.oorlogsbronnen.nl/rapport-triado-verrijkingsfase
> Projectpagina: tinyurl.com/yycqvp8m