Ministerie van Binnenlandse Zaken en Koninkrijksrelaties

Pilot ‘Zoeken en opruimen’

Eindrapport

April 2023 - Versie 1.0

Rijksprogramma Duurzaam Digitale **Informatiehuishouding**

Inhoud

1 Inleiding en opdracht

Voor het ministerie van Financiën is in de maanden november, december 2020, januari en februari 2021 een Pilot uitgevoerd dat als doel heeft om duplicaten en bijna-duplicaten te identificeren met behulp van Zoek en Vind. Deze PoC heeft geleid tot een eindrapport (zie referentie 1) waarin een aantal aanbevelingen worden gegeven om de beschreven methodes in een productie-omgeving te kunnen gebruiken.
De opdracht van dit project is het toepassen van de functionaliteit om duplicates en bijna-duplicaten te detecteren in drie pilots. Zoals in de offerte al is benoemd zal dit in drie fases wordt uitgevoerd:

1. Het overzetten van de functionaliteit uit de PoC naar de productie-omgeving.
   Deze fase wordt opgedeeld in drie activiteiten:
   a. Het productierijp maken van de detectie van bijna-duplicaten
   b. Het overzetten van duplicate detectie naar de productie-omgeving
   c. Het overzetten van de detectie van bijna-duplicaten naar de productie-omgeving
2. Uitwerken van processen en de rol van Zoek en Vind in deze processen
3. Het uitvoeren van de drie pilots

In dit eindrapport zijn de resultaten van fase 1 beschreven.
RDDI en het Ministerie van Financiën hebben fase 2 en 3 uitgevoerd en als eindresultaat een video opgenomen.

2 Fase 1

In Fase 1 wordt de functionaliteit van de eerste PoC overgezet naar de productieomgeving van Zoek en Vind van het ministerie van Financiën.

2.1 Fase 1.1 – Functionaliteit m.b.t. bijna-duplicaten *productierijp* maken

In deze fase wordt de functionaliteit om bijna-duplicaten te detecteren productierijp gemaakt. Dit houdt in dat:

- De verschillen tussen (bijna) dezelfde documenten in verschillende formaten verder wordt geanalyseerd. De resultaten van deze analyse kunnen tot een uitbreiding van de zogenaamde '*wasstraat*' leiden.
- De aanroep van de functie om voor elk document fingerprints te bepalen wordt geoptimaliseerd. Doel is om een instelling te vinden waarbij (bijna) duplicaten goed worden gedetecteerd. Snelheid van indexering en zoeken wordt ook meegenomen in het bepalen van de instellingen.

2.1.1 Verbeteren van de 'wasstraat'

Tijdens het analyseren van de teksten die IDOL uit de documenten haalt (waarbij de wasstraat uit de eerste PoC is gebruikt) zien we dat bij het gebruik van bepaalde opmaak de verschillen van dezelfde documenten in verschillende formaten groot is. Daardoor zal het bepalen van bijna-duplicaten op basis van fingerprints minder goede resultaten opleveren. Hieronder volgt een aantal elementen die worden gebruikt en is beschreven welke verschillen in verschillende formaten dit in IDOL oplevert. Waar mogelijk is aangegeven welke mogelijkheid er is om de tekst in de wasstraat op te schonen

Afbeeldingen – In sommige gevallen wordt in MS Word een zogenaamde alt-tekst als eigenschap van een afbeelding opgenomen. Deze alt-tekst is bij MS Word-documenten aanwezig in de zoekindex. Bij de PDF-documenten is deze alt-tekst niet aanwezig. Omdat de alt-tekst niet als zodanig is te herkennen, kan deze niet worden opgeschoond.

Tabellen – In de wasstraat van de eerste PoC werden waarden van een rij van een tabel aan elkaar geplakt in PDF-documenten. In de eerste wasstraat is een aanpassing gemaakt zodat tabellen in PDF en MS Word-documenten meer overeenkomen.

Headers en footers – De manier waarop headers en footers in PDF- en MS Word-tekst staan, verschilt zeer. Omdat de header en footers niet in de tekst herkenbaar zijn, is het niet mogelijk deze te op te schonen.

Hyperlinks in de tekst – Onder tekst kan een hyperlink worden opgenomen. In de geïndexeerde tekst van een Word-document is deze hyperlink niet aanwezig. Bij PDF-documenten is deze herkenbaar door {HYPERLINK <URL>}. In de wasstraat wordt deze tekst verwijderd.

Lijsten – Het gebruik van spaties en tabs is bij lijsten (zowel genummerd als niet genummerd) voor MS Word- en PDF-documenten verschillend. Ook de bullet zelf wordt in PDF wel getoond en bij MS Word niet. De wasstraat kan worden verbeterd door bij lijsten

- De bullet te verwijderen. De letter 'o' kan in een hoger niveau van de lijst worden gebruikt.
- Enkele letters te verwijderen. Het is mogelijk lijsten te maken die enkele letters gebruiken ('a', 'b', 'c' etc.). Hiermee wordt ook de letter 'o', die soms als hoger niveau bullet wordt gebruikt, meegenomen

Genummerde lijst 1., 1.1 e.d. worden in de bestaande wasstraat (1.0) al verwijderd.
In onderstaande tabel staan de resultaten van de verbeterde wasstraat. Als instelling van de fingerprint-functie is de instelling genomen die ook tijdens de eerste PoC is gebruikt. Daarmee is het makkelijker om de resultaten te vergelijken met de wasstraat uit de eerste PoC.

Testdocumenten	Scenario's	Scenario's	Scenario's
Wasstraat	**Geen wasstraat**	**Wasstraat 1.0**	**Wasstraat 2.0**
Document van 26 pagina's	#fp: 119 Match: 2 (1%)	#fp: 93 Match: 38 (42%)	#fp: 90 Match: 61 (68%)
Document van 8 pagina's MS Word t.o.v. PDF	#fp: 29 Match: 1 (3%)	#fp: 30 Match: 14 (47%)	#fp: 28 Match: 19 (67%)

2.1.2 Fingerprint-functie

In het eindrapport van de eerste PoC is kort weergegeven welke parameters de functie voor het bepalen van de fingerprints van een document gebruikt.
In de volgende tabel staat de officiële documentatie van deze functie:

string ... **fingerprint_string**(*string* data, *unsigned int* min_chars = 10, *unsigned int* mask_size = 8, *unsigned int* num_bytes = 48)

Arguments

Data	String	The string to fingerprint
min_chars	*unsigned int*	Absolute minimum chunk size in UTF8 characters.
mask_size	*unsigned int*	Size of bitmask when calculating string chunk boundaries. Valid range 1-31. Smaller values increase number of fingerprints.
num_bytes	*unsigned int*	Length of substring to check when calculating string chunk boundaries.

Vrij vertaald betekenen de parameters het volgende:
min_chars – De absolute minimale grootte van het tekstblok in UTF-8 karakters
mask_size – De bitmask-grootte voor het berekenen van de grenzen van een tekstblok. De mask_size moet tussen 1 en 31 liggen. Kleinere waardes van de mask_size leidt tot meer fingerprints.
num_bytes – Lengte van de substring die gecontroleerd wordt als de grenzen van een tekstblok worden berekend.

In de volgende paragrafen wordt beschreven wat we gedaan hebben om waardes van deze parameters te vinden waarmee goede resultaten kunnen worden gehaald zonder de indexering- en query-snelheid van Zoek en Vind te veel negatief te beïnvloeden.

2.1.3 Testen parameters van de fingerprint-functie

Bij het testen van de optimale instelling voor het bepalen van fingerprints is een aantal verschillende documenten gebruikt om een zo groot mogelijke variëteit van documenten te krijgen: grote en kleine documenten, met veel en zonder afbeeldingen, met en zonder tabellen etc.

De documenten die die zijn gebruikt staan in onderstaande tabel met daarbij het aantal pagina's en het formaat:

Bestandsnaam	P	Formaat
2_antwoorden-op-kamervragen-over-nijpende-tekorten-aan-forensische-artsen	5	MS Word, PDF
7_besluit-wob-verzoek-over-staatssteun-klm(search=klm)	3	MS Word, PDF
9_geannoteerde-agenda-voor-de-landbouw-en-visserijraad-van-21-februari-2022	6	MS Word, PDF
1_5-a-day_portion-sizes	4	MS Word, PDF
4_beantwoording-kamervragen-over-effecten-van-de-europese-green-deal-op-de-landbouwopbrengst.	5	MS Word, PDF
9a_geannoteerde-agenda-voor-de-landbouw-en-visserijraad-van-21-februari-2022	6	PDF
beheer afspraken BDAP v 0.3	6	MS Word
beheer afspraken BDAP v 0.2	6	MS Word
beheer afspraken BDAP v 0.1	5	MS Word
17_refman-5.0	75	MS Word, PDF
7_apache_nifi_tutorial.pdf	67	MS Word, PDF
8_besluit-en-openbaar-gemaakte-documenten-wob-verzoek-repatriering-nederlanders-uit-marokko	103	MS Word, PDF
11_Gemeenteraadsverkiezingen+2022_Kant+en+klare+boodschappen+voor+op+sociale+media	22	MS Word, PDF
10_factsheet-11-medicatie-polyfarmacie-psychofarmaca	24	MS Word, PDF

De fingerprint-functie heeft drie parameters waarmee geëxperimenteerd kan worden:

1. Het minimale aantal karakters om een fingerprint te bepalen. Deze staat default op 10. Hier is tijdens de tests niet mee gevarieerd omdat in de eerste PoC bleek dat dit weinig tot geen effect had.
2. De mask_size. Hiermee wordt de grootte van de tekst bepaald waar een fingerprint voor wordt berekend. Deze parameter heeft de meeste invloed op het aantal fingerprints dat voor een document wordt berekend. De default waarde is 10.
3. Met de laatste parameter, num bytes, wordt de lengte van de substring gecontroleerd. Ook met deze parameter zou kunnen worden 'gespeeld'. Default waarde is 48.

De tests zijn in twee stappen uitgevoerd. Eerst is getest met verschillende waardes van de *mask_size. Dit is met drie scenario's uitgevoerd waarbij van de testdocumenten is gekeken naar het aantal fingerprints dat match met een (bijna) duplicaat.*
In de tweede test is gekeken naar de optimale waarde van de num bytes-parameter.

Voor de tests met de mask_size is met drie scenario's gewerkt:

1. Een *mask_size van 10 (de default)*
2. *Een mask_size van 7*
3. *Een mask_size van 5*

Deze drie scenario's zijn uitgevoerd met de default waarde van de *num_bytes-parameter en de wasstraat uit de eerste PoC.*

Het testen van een *mask_size* groter dan 10 heeft geen zin omdat voor een document dusdanig weinig fingerprints worden berekend dat bijna-duplicaten niet kunnen worden gevonden.

Naast de drie genoemde scenario's is ook gekeken naar scenario 2 i.c.m. de uitgebreide wasstraat (wasstraat 2.0). Dit hebben we scenario 2.1 genoemd.

In de volgende tabellen worden de resultaten van de tests weergegeven.

Kleine documenten – Verschillende formaten

Bestand	**Scenario 1** *mask_size 10 wasstraat 1.0*	**Scenario 2** *mask_size 7 wasstraat 1.0*	**Scenario 3** *mask_size 5 wasstraat 1.0*	**Scenario 2.1** *mask_size 7 wasstraat 2.0*
2_antwoorden-op-kamervragen-over-nijpende-tekorten-aan-forensische-artsen	#fp: 11 Match: 7 (63%)	#fp: 86 Match: 77 (88%)	#fp: 241 Match: 225 (93%)	#fp: 84 Match: 75 (89%)
7_besluit-wob-verzoek-over-staatssteun-klm	#fp: 4 Match: 1 (33%)	#fp: 10 Match: 4 (40%)	#fp: 54 Match: 39 (72%)	#fp: 11 Match: 6 (55%)
9_geannoteerde-agenda-voor-de-landbouw-en-visserijraad-van-21-februari-2022	#fp: 18 Match: 11 (61%)	#fp: 135 Match: 118 (87%)	#fp: 389 Match: 369 (95%)	#fp: 132 Match: 115 (87%)
1_5-a-day_portion-sizes	#fp: 32 Match: 0 (0%)	#fp: 25 Match: 0 (0%)	#fp: 36 Match: 0 (0%)	#fp: 31 Match: 22 (71%)
4_beantwoording-kamervragen-over-effecten-van-de-europese-green-deal-op-de-landbouwopbrengst	#fp: 9 Match: 4 (44%)	#fp: 88 Match: 76 (86%)	<url te lang>	#fp: 86 Match: 78 (88%)

Kleine documenten – Zelfde formaat – Versies

Bestand	**Scenario 1** *mask_size 10 wasstraat 1.0*	**Scenario 2** *mask_size 7 wasstraat 1.0*	**Scenario 3** *mask_size 5 wasstraat 1.0*	**Scenario 2.1** *mask_size 7 wasstraat 2.0*
9_geannoteerde-agenda-voor-de-landbouw-en-visserijraad-van-21-februari-2022 (t.o.v. 9a-versie van het document)	#fp: 18 Match: 17 (94%)	#fp: 128 Match: 127 (99%)	<url te lang>	#fp: 124 Match: 123 (99%)
beheer afspraken BDAP v 0.3	#fp: 10 Match: 0,2: 9 (90%) 0.1: 4 (40%)	#fp: 73 Match: 0.2: 71 (97%) 0.1: 41 (56%)	#fp: 215 Match: 0.2: 210 (99%) 0.1: 127 (60%)	#fp: 69 Match: 0.2: 68 (98,5%) 0.1: 39 (56.5%)

Grote documenten – Verschillende formaten

Bestand	Scenario 1 *mask_size 10* *wasstraat 1.0*	Scenario 2 *mask_size 7* *wasstraat 1.0*	Scenario 3 *mask_size 5* *wasstraat 1.0*	Scenario 2.1 *mask_size 7* *wasstraat 2.0*
17_refman-5.0	#fp: 134 Match: 56 (42%)	#fp: 1066 Match: 834 (78%)	#fp: 3320 Match: 2789 (84%)	#fp: 1044: Match: 1024 (98%)
7_apache_nifi_tutorial.pdf	#fp: 54 Match: 15 (27%)	#fp: 414 Match: 247 (59%)	#fp: 1385 Match: 847 (68%)	#fp: 446 Match: 273 (62%)
8_besluit-en-openbaar-gemaakte-documenten-wob-verzoek-repatriering-nederlanders-uit-marokko	#fp: 34 Match: 19 (59%)	#fp: 191 Match: 157 (81%)	<url te lang>	#fp: 179 Match: 149 (83%)
11_Gemeenteraadsverkiezingen 2022_Kant en klare boodschappen voor op sociale media	#fp: 43 Match: 1 (3%)	#fp: 197 Match: 53 (30%)	<url te lang>	#fp: 186 Match: 97 (52%)
10_factsheet-11-medicatie-polyfarmacie-psychofarmaca	#fp: 17 Match: 0 (0%)	#fp: 148 Match: 47 (28%)	<url te lang>	#fp: 144 Match: 59 (40%)

In de tabel met grote bestanden is niet voor alle testdocumenten het resultaat van scenario 3 bepaald. Vanwege de grote aantallen fingerprints per document was het niet mogelijk om dit met de frontend van Zoek en Vind te bepalen omdat de URL te lang werd (hiervoor is ondertussen een oplossing in de back-end van Zoek en Vind gerealiseerd). Van de twee documenten *(17_refman-5.0 en 7_apache_nifi_tutorial)* waarbij wel de resultaten zijn weergegeven, is dit met de hand bepaald.

Zoals in de tabellen is af te lezen groeit het aantal fingerprints sterk als de parameter *mask_size* kleiner wordt. In onderstaande tabel is dit weergegeven voor een document van 5, 26 en 67 pagina's

#fp met mask_size	1	2	3	4	5	6	7	8	9	10
5 pagina's	826	-	542	-	215	115	73	37	-	10
26 pagina's	-	-	-	-	560	312	167	-	-	-
67 pagina's	-	-	-	-	3312	1942	1044	-	-	-

Uit deze tabel valt af te lezen dat het aantal fingerprint erg groot wordt naar mate de *mask_size* kleiner wordt. Bij een *mask_size* van 3 wordt per pagina meer dan honderd fingerprints berekend.

Uit de resultaten van de drie scenario's zijn een aantal conclusies te trekken:

- Het aantal matches van fingerprints neemt significant toe door een *mask_size* van 7 in plaats van 10 te gebruiken
- De verbetering in matches met een *mask_size* van 7 naar 5 is beperkt maar het aantal fingerprints groeit met deze stap zo veel dat we deze stap niet aanbevelen
- De verbeterde wasstraat heeft ook tot een verbetering in de matches geleid zonder impact op de indexering.

De tweede test die is uitgevoerd heeft betrekking op de *num bytes*-parameter. Hier is ook met drie scenario's gewerkt:

1. *Num bytes* is 40
2. *Num bytes* is 48 (de default)
3. *Num bytes* is 60

In deze scenario's is gewerkt met een *mask_size* van 7 en de wasstraat van de eerste PoC.

Kleine documenten – Verschillende formaten

Bestand	Scenario 1 *Num bytes 40*	Scenario 2 *Num bytes 48*	Scenario 3 *Num bytes 60*
2_antwoorden-op-kamervragen-over-nijpende-tekorten-aan-forensische-artsen	#fp: 73 Match: 66 (90%)	#fp: 86 Match: 77 (88%)	#fp: 70 Match: 60 (85%)
7_besluit-wob-verzoek-over-staatssteun-klm(search=klm)	#fp: 11 Match: 5 (45%)	#fp: 11 Match: 4 (40%)	#fp: 19 Match: 12 (63%)
9_geannoteerde-agenda-voor-de-landbouw-en-visserijraad-van-21-februari-2022	#fp: 148 Match: 136 (93%)	#fp: 132 Match: 120 (93%)	#fp: 119 Match: 107 (92%)
1_5-a-day_portion-sizes	#fp: 73 Match: 66 (90%)	#fp: 86 Match: 77 (89%)	#fp: 70 Match: 60 (85%)
4_beantwoording-kamervragen-over-effecten-van-de-europese-green-deal-op-de-landbouwopbrengst.	#fp: 11 Match: 5 (45%)	#fp: 11 Match: 4 (36%)	#fp: 19 Match: 12 (63%)

Kleine documenten – Zelfde formaat – Versies

Bestand	Scenario 1 *Num bytes 40*	Scenario 2 *Num bytes 48*	Scenario 3 *Num bytes 60*
9_geannoteerde-agenda-voor-de-landbouw-en-visserijraad-van-21-februari-2022 (t.o.v. 9a-versie van het document)	#fp: 145 Match: 144 (98.6%)	#fp: 129 Match: 127 (99%)	#fp: 118 Match: 114 (98%)
beheer afspraken BDAP v 0.3	#fp: 68 Match: 66 (98.5%)	#fp: 73 Match: 71 (97%)	#fp: 85 Match: 82 (98.8%)

Grote documenten – Verschillende formaten

Bestand	Scenario 1 *Num bytes 40*	Scenario 2 *Num bytes 48*	Scenario 3 *Num bytes 60*
17_refman-5.0	<url te lang>	<url te lang>	<url te lang>
7_apache_nifi_tutorial.pdf	#fp: 36 Match: 27 (79%)	#fp: 32 Match: 28 (77%)	#fp: 37 Match: 28 (77%)
8_besluit-en-openbaar-gemaakte-documenten-wob-verzoek-repatriering-nederlanders-uit-marokko	#fp: 205 Match: 169 (82%)	#fp: 186 Match 152 (82%)	#fp: 170 Match: 129 (76%)
11_Gemeenteraadsverkiezingen 2022_ Kant en klare boodschappen voor op sociale media	#fp: 186 Match: 58 (31%)	#fp: 178 Match: 49 (28%)	#fp: 189 Match: 52 (28%)
10_factsheet-11-medicatie-polyfarmacie-psychofarmaca.pdf(search= psychofarmaca)	#fp: 167 Match: 58 (35%)	#fp: 148 Match: 48 (32%)	#fp: 118 Match: 27 (23%)

Uit de resultaten valt af te leiden dat afwijken van de default-waarde van de *num bytes*-parameter geen significante verbeteringen geeft in het aantal matches.

2.1.4 Zoeken met fingerprints

In de eerste PoC kon in de frontend worden gezocht met de fingerprints van een document. Bij grote documenten die veel fingerprints (meer dan 250) hebben, werd de URL van de zoekvraag te lang en werd een foutmelding getoond.
De Business Logic Layer van Zoek en Vind is nu zo aangepast dat ook voor grote documenten kan worden gezocht met fingerprints.

2.1.5 Conclusie fase 1.1

Bij de testgevallen met verschillende waardes voor de *mask_size*, één van de parameters van de functie die de fingerprints bepaalt, zien we dat de grootste verbetering te zien is in de stap van *mask_size* 10, de default waarde, naar een *mask_size* van 7.
In onderstaande tabel zijn de resultaten van de dertien testdocumenten (waarvoor in de testset één bijna-duplicaat document aanwezig is) weergegeven verdeeld over een bereik van matchings-percentage van de bijna-duplicaten.

Overzicht aantal matching documenten per scenario

Bereik matchings-percentage	*Scenario 1* *Mask_size 10* *Wasstraat 1.0*	*Scenario 2* *Mask_size 7* *Wasstraat 1.0*	*Scenario 3* *Mask_size 5* *Wasstraat 1.0*	*Scenario 2.1* *Mask_size 7* *Wasstraat 2.0*
0 – 20%	3 (23%)	1 (8%)	1 (1%)	0 (0%)
21 – 40%	3 (23%)	3 (23%)	0 (0%)	1 (8%)
41 – 60%	3 (23%)	2 (15%)	1 (8%)	3 (23%)
61 – 80%	2 (15%)	1 (8%)	1 (8%)	2 (15%)
81 – 100%	2 (15%)	6 (46%)	6 (46%)	7 (54%)
URL te lang	0 (0%)	0 (0%)	4 (31%)	0 (0%)

De percentages opgeteld per kolom kunnen iets meer of minder dan 100% zijn in verband met afronding.

Hierin zie je dat in scenario 1 (*mask_size* is 10) 30 procent van de documenten (4 documenten) tussen de 61 en 100% van de fingerprints matchen met het bijna-duplicaat document in de testset.
In scenario 2 (*mask_size* is 7) is dit percentage gestegen naar 54 procent van de documenten. In combinatie met de wasstraat 2.0 (scenario 2.1) loopt dit op tot 69% (negen van de dertien documenten).

Uit fase 3 van deze pilot, waar de drie pilot scenario's zullen worden uitgevoerd, zal blijken welke matchingspercentages als best practice gebruikt gaan worden om bijna-duplicaten te vinden. Ook zal dan gekeken worden of er *false positives* als bijna-duplicaat gevonden worden en welke maatregelen hiertegen genomen kunnen worden.

De stap naar een *mask_size* van 5 (scenario 3) levert wel een verbetering op in de opsporing van bijna-duplicaten, maar deze verbetering is niet heel groot. Voor het bereik van de het matchingspercentage tussen 61 en 100% worden 69% bijna-duplicate documenten gevonden. Dat zijn negen van de dertien documenten.
Het nadeel van een *mask_size* van 5 is dat het aantal fingerprint drie keer zo groot is dan bij een *mask_size* van 7.

De tests met de varianties in de parameter num_bytes levert geen significatie verbetering (en verslechtering) van het matchings-percentage.

2.2 Fase 1.2 – Functionaliteit m.b.t. duplicaten naar productie overzetten

2.2.1 Uitgevoerde acties

Bij de uitrol van de functionaliteit voor het opsporen van duplicaten zijn de volgende acties uitgevoerd

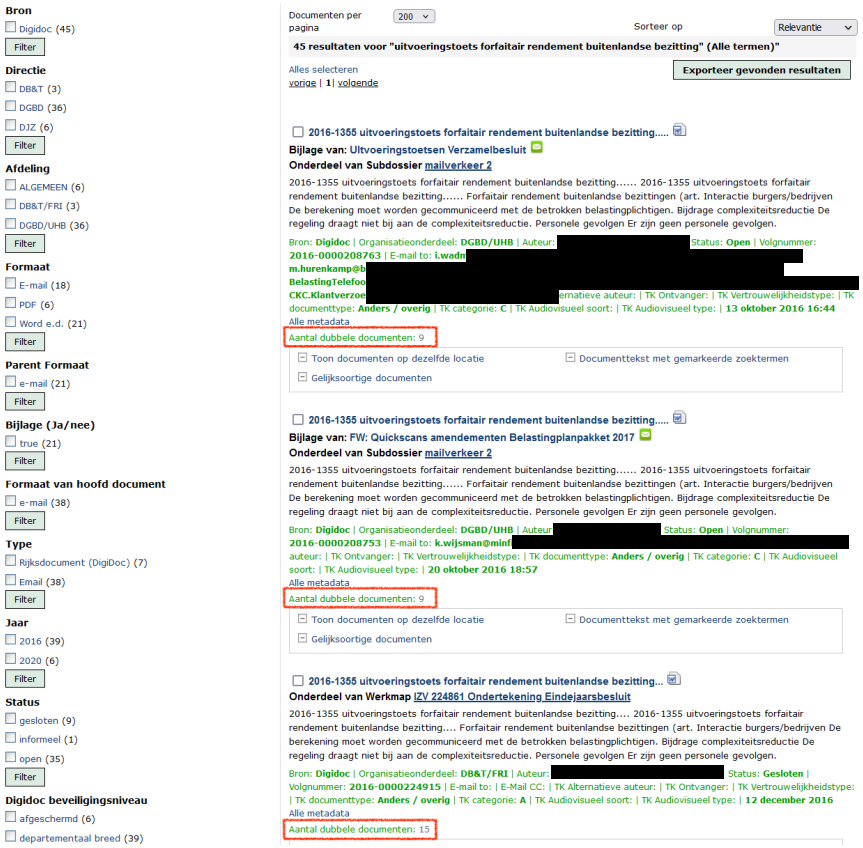
- Er is in de bronnen Docman, DigiDoc en alle Orgdata-netwerkschijven aan alle documenten een hash-waarde toegekend. Hash-waardes worden berekend op basis van het hele bestand. Identieke bestanden krijgen eenzelfde hash-waarde.
- Uitrol van een nieuwe versie van Zoek en Vind waarin het overzicht van duplicaten is verbeterd.
- De software die duplicaten opspoort en metadata-velden vult is verder geoptimaliseerd. Incrementele runs zijn daardoor veel sneller afgerond.
- De indexeerstraat is zodanig aangepast dat documenten verrijkt kunnen worden met metadata zonder dat het document volledig geïndexeerd hoeft te worden. Het volledig indexeren van een document is een veel zwaarder proces dan het verrijken. Het vullen van metadata-velden ten behoeve van duplicaten-detectie zal daardoor veel sneller kunnen worden uitgevoerd.

2.2.2 Resultaat

Zoek op **uitvoeringstoets forfaitair rendement buitenlandse bezitting** en controleer in de resultaatlijst bij elk document of er staat: dubbele documenten.

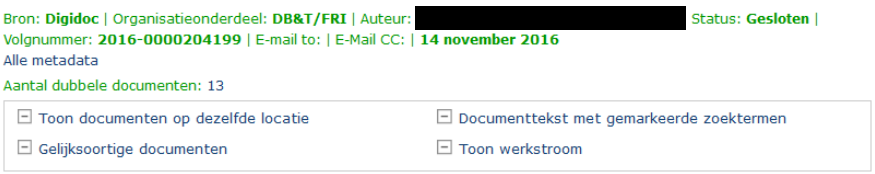
De documenten van Digidoc, Docman en alle Orgdata-netwerkschijven zijn voorzien van extra metadata die gebruikt kunnen worden als filter bij het opsporen van duplicaten. Hierbij zijn trouwens plaatjes (bestanden met de extensie PNG, IMG, GIF, JPEG, etc.) uitgesloten.
Het gaat om de volgende velden die als filter te gebruiken zijn in de productie-omgeving van Zoek en Vind Financiën (https://zoeken.rijksweb.nl).

- **Duplicaat aanwezig**: dit geeft aan dat er duplicaten zijn
- **Aantal duplicaten**: het aantal duplicaten voor een document (incl. het document zelf)

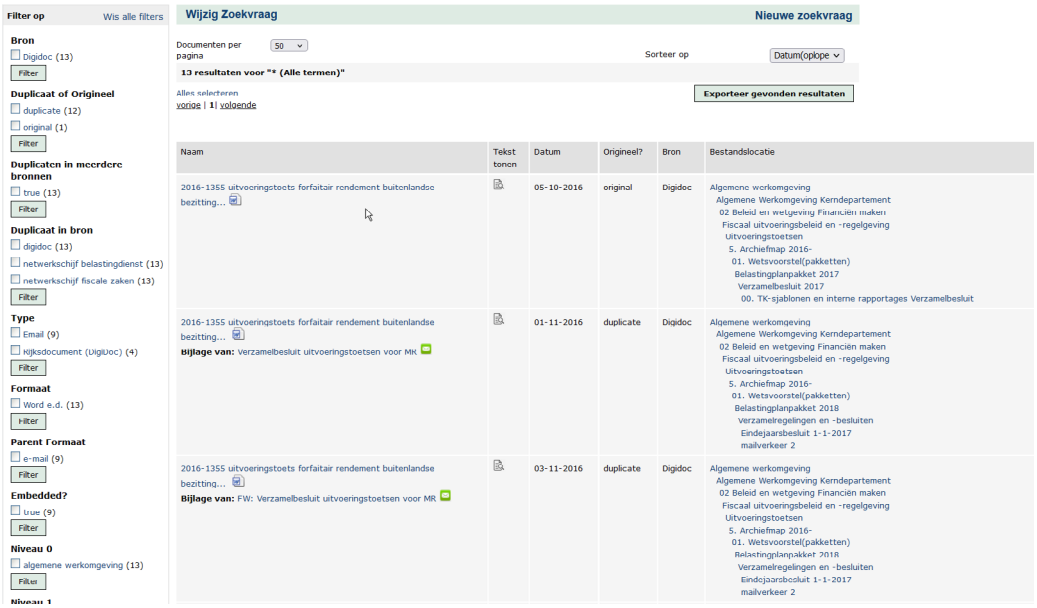


- **Duplicaat of Origineel**: het document met oudste datum wordt als *origineel* gekenmerkt, en de andere duplicaten als *duplicaat*
- **Duplicaat in Bron**: geeft het aantal duplicaten in de verschillende bronnen
- **Duplicaten in meerdere bronnen**: geeft aan dat een document duplicaten heeft in verschillende bronnen (nu dus Digidoc, Docman en alle Orgdata-netwerkschijven)

Met deze filters kun je snel duplicaten opsporen, en kun je vanuit de resultaatlijst doorklikken naar een lijst van duplicaten door op het getal te klikken achter *Aantal dubbele documenten.*



Je ziet dan de lijst met dubbele documenten:



In dit overzicht is een aantal filters beschikbaar die gebruikt kunnen worden om een subset te creëren.
Het overzicht van duplicaten heeft 6 kolommen:

- **Naam** – Deze kolom toont de bestandsnaam gevolgd door het icoontje dat het type van het document weergeeft. Als het document onderdeel is van een zipbestand of een bijlage is van een e-mailbericht, wordt dat onder de naam weergegeven.
- **Tekst tonen** – Door op het teksticoontje in deze kolom te klikken verschijnt de tekst van het document in een nieuw tabblad.
- **Datum** – Deze kolom geeft de documentdatum weer
- **Origineel?** – In deze kolom staat of het document het (waarschijnlijke) origineel is of een duplicaat.
- **Bron** – Deze kolom toont de bron van het document.
- **Bestandlocatie** – In deze kolom wordt de locatie in de bron van het bestand weergegeven. Door op de locatie te klikken wordt een overzicht van alle bestanden op deze locatie getoond.

2.3 Fase 1.3 – Functionaliteit m.b.t. bijna-duplicaten naar productie overzetten

2.3.1 Uitgevoerde acties

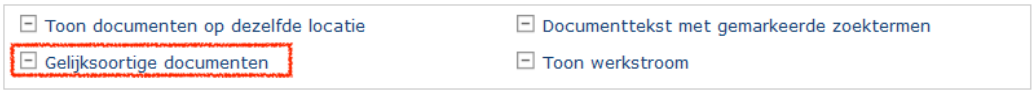
Bij de uitrol van de functionaliteit voor het opsporen van gelijksoortige documenten zijn de volgende acties uitgevoerd:

- De indexeerstraat waarin de wasstraat voor het bepalen van fingerprints is in productie genomen.
- Uitrol van een nieuwe versie van Zoek en Vind waarin de pagina voor het vergelijken van tekst van gelijksoortige documenten is opgenomen.
- In de nieuwe versie van Zoek en Vind worden de fingerprints waarmee wordt gezocht, niet meer via de URL verstuurd. Tijdens de PoC leverde dit fouten op bij documenten met veel fingerprints.

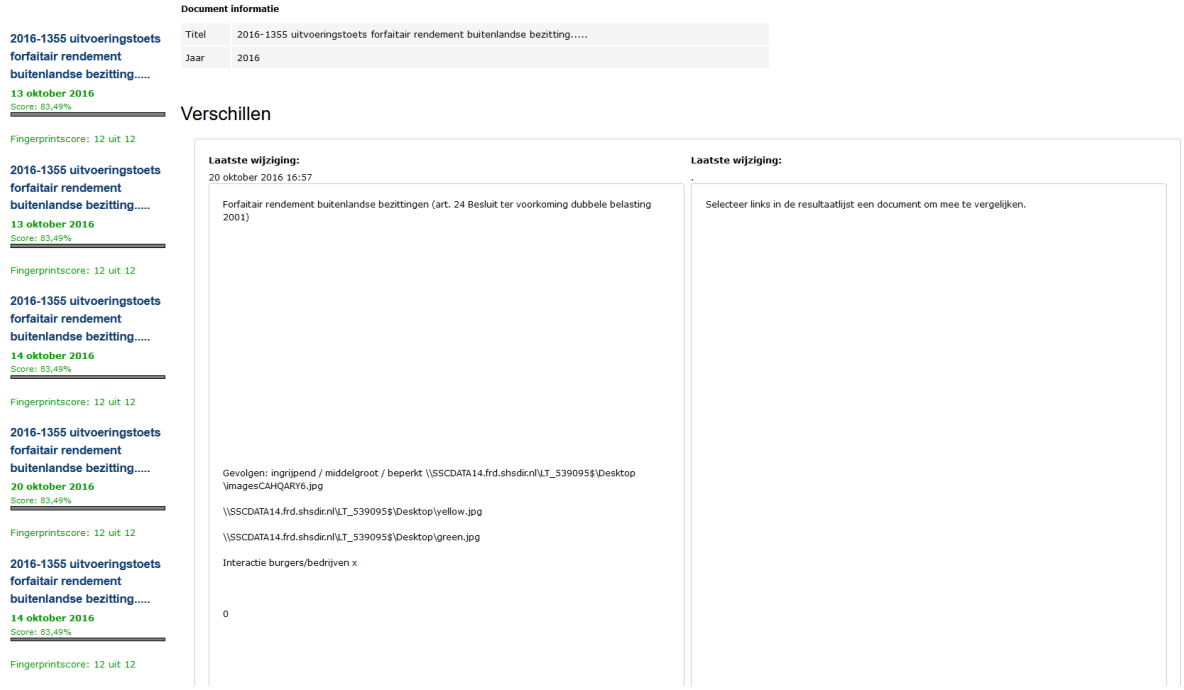
Deze opties kunnen d.m.v. configuratie aan/uit gezet worden voor groepen gebruikers.

2.3.2 Resultaat

Bij elk document wordt de optie gelijksoortige documenten getoond, waarmee o.b.v. fingerprints gelijksoortige documenten gezocht worden.

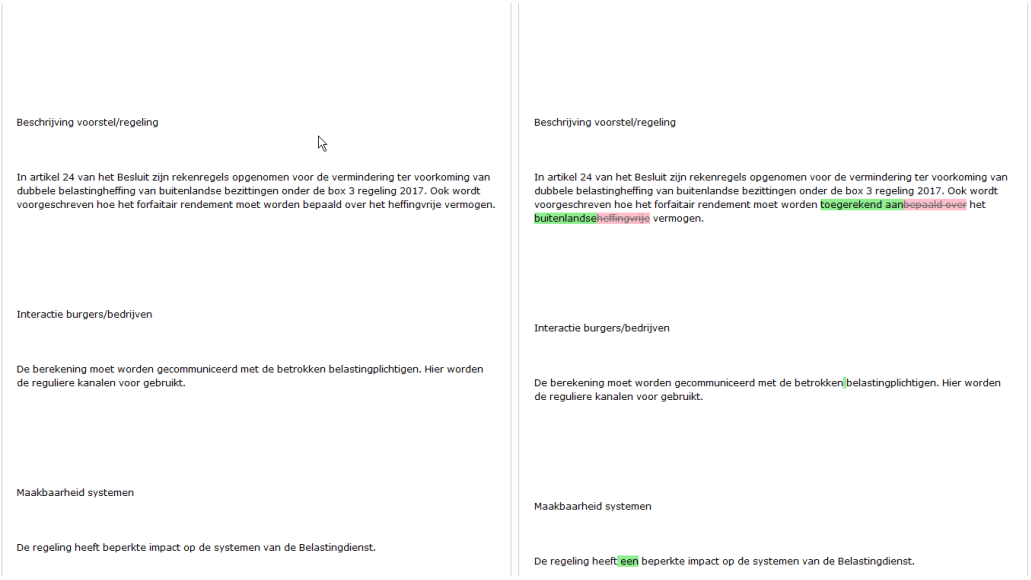


Als op deze link wordt geklikt opent het volgende scherm:



Aan de linker kant van het scherm wordt een lijst met gelijksoortige documenten getoond. Per *gelijksoortig* document wordt de titel en datum getoond en hoeveel fingerprints overeenkomen.
Een *fingerprintscore* van 10 uit 12 betekent dat dit document 10 van de 12 dezelfde fingerprints heeft als het document waarmee is gezocht.

Als je een van de documenten opent door op de titel te klikken wordt deze in de rechter kolom getoond en worden verschillen gemarkeerd tussen het document waarmee gezocht is (linker kolom) en het gevonden document.



3 Conclusie en aanbeveling

3.1 Conclusies fase 1

Het verbeteren van de 'wasstraat' voor het opsporen van bijna-duplicaten heeft voor bepaalde documenten geleid tot een veel betere match. Met name bij documenten met tabellen (zoals het document *1_5-a-day_portion-sizes*) en documenten met opsommingen en lijsten (zoals document *17_refman-5.0*) worden aanzienlijke verbeteringen in het percentage matchende fingerprints bereikt.
Bij andere documenten zijn de resultaten iets verbeterd of gelijk. Nergens leidt deze wasstraat 2.0 tot een lager matchings-percentage.

Onze conclusie is dat voor bepalen van de fingerprints een *mask_size* van 7 en de default *num_bytes*-waarde van 48 goed werkbaar is bij het opsporen van bijna-duplicaten.

3.2 Aanbevelingen naar aanleiding van fase 2 en 3

Tijdens het uitvoeren van fase 2 en 3 is een aantal bevindingen gedaan.

Verwijderde documenten
Als documenten uit een bron worden verwijderd, detecteert de software die duplicaten opspoort dit momenteel nog niet tijdens incrementele runs. Dit kan betekenen dat het filter *aantal duplicaten* aantallen toont die niet correct hoeven te zijn.
De software moete worden aangepast zodat telling van duplicaten goed wordt bijgewerkt, ook als er documenten zijn verdwenen.

Aantal fingerprint kan erg veel zijn
Bij (erg) grote documenten kan het aantal fingerprints dat wordt berekend erg groot zijn. Het zoeken van vergelijkbare documenten kan daardoor traag zijn er wordt nl. gezocht met alle documenten.
Aanbevolen wordt om per document een maximaal aantal van bijvoorbeeld 4000 fingerprints in de zoekindex op te slaan. Voor een pdf-document betekent dit dat fingerprints voor de ca. eerste 300 pagina's worden bepaald. Bij MS Word zijn dit ca. 250 pagina's. Dat betekent dat bij grote documenten dat alleen het begin gebruikt zal worden voor opsporen van vergelijkbare documenten.

Aantallen van de facetwaardes
Tijdens het testen is af en toe gebleken dat de waardes achter filterwaardes niet correct is. Het blijkt dat af en toe een deel van de zoekindex niet correct is. De oorzaak hiervan is niet bekend en moet worden geanalyseerd.
Om deze situatie vroegtijdig op te sporen is er speciale monitoring ingericht. Als er een discrepantie tussen de telling van filterwaardes en het daadwerkelijke aantal wordt geconstateerd, kan dat worden hersteld.

Vergelijken van vergelijkbare documenten
Op de pagina waar vergelijkbare documenten worden getoond, zijn een aantal verbeteringen mogelijk.
- In de linker kolom is nu niet duidelijk welk document geselecteerd is voor de vergelijking.
- Bij grote documenten duurt het vergelijken van de teksten lang zonder dat de gebruiker een melding ziet.

Rechten

Het bepalen van exacte duplicaten gebeurt voor alle documenten in de informatiebronnen. Als een gebruiker alle documenten in alle bronnen dan zullen de aantallen die getoond worden *kloppen*, d.w.z. als in de filters aangegeven wordt dat een document 22 duplicaten heeft dan zal de gebruiker die 22 documenten ook kunnen zien in de resultaatlijst. Als een gebruiker niet de rechten heeft op alle 22 documenten dan zal in de lijst van dubbele documenten ze ook niet alle 22 getoond worden.
Ons advies is dan ook om het opruimen van duplicaten te laten uitvoeren door medewerkers met uitgebreide rechten.

Zoekresultaat highlighten

Het zou handig zijn wanneer het zoekresultaat wordt ge-highlight als we de inhoud van de map bekijken, zeker als er veel bestanden in zitten. Wanneer je nu de locatie opent, dan zie je niet direct het bestand dat als duplicaat is aangemerkt.

Naam	Tekst tonen	Datum	Origineel?	Bron	Bestandslocatie
Deel 3: Mandaatregisters regio Belastingkantoren en - onderdelen.		08-01-2016	original	Docman	Uitvoeren van de bedrijfsvoering m.b.t. personeel over 2013. MACHTIGINGSBESLUITEN / MANDAATBESLUITEN
Deel 3 Mandaatregisters regio Belastingkantoren en - onderdelen. **Bijlage van:** Archief - 2019-09-04		04-09-2019		Digidoc	Algemene werkomgeving Algemene Werkomgeving Kerndepartement 02 Beleid en wetgeving Financiën maken Fiscaal uitvoeringsbeleid en -regelgeving Advisering Belastingdienst Procesondersteuning DGBD 2017-2021 05 PROCES PRODUCTINNOVATIE 06 ACCOUNTHOUDERSCHAP 0602 TOESLAGEN 02 AUT 01 Aangeleverde stukken voor commissie
Deel 3 Mandaatregisters regio Belastingkantoren en - onderdelen. **Bijlage van:** Archief - 2019-09-04		16-09-2019	duplicate	Digidoc	Algemene werkomgeving Algemene Werkomgeving Kerndepartement 02 Beleid en wetgeving Financiën maken Fiscaal uitvoeringsbeleid en -regelgeving Advisering Belastingdienst Procesondersteuning DGBD 2017-2021 05 PROCES PRODUCTINNOVATIE 06 ACCOUNTHOUDERSCHAP 0602 TOESLAGEN 02 AUT 01 Aangeleverde stukken voor commissie Aangeleverde stukken voor de commissie deel 2

Dit punt staat op de backlog van Zoek en Vind.

Dit is een uitgave van:

Rijksprogramma Duurzaam Digitale
Informatiehuishouding (RDDI)

Rijnstraat 50
Postbus 16375
2500 BJ Den Haag

dashboardcoachihh@minocw.nl
www.informatiehuishouding.nl

April 2023 - Versie 1.0