Beeldbanken verrijken

Erik van den Hooven
0323268

Bachelor thesis
Credits: 9EC

Bachelor Opleiding Kunstmatige Intelligentie

Faculteit der Natuurwetenschappen, Wiskunde en Informatica
Universiteit van Amsterdam
Science Park 904
1098 XG Amsterdam

Begeleider
dr. M.J. Marx

Faculteit der Natuurwetenschappen, Wiskunde en Informatica
Universiteit van Amsterdam
Science Park 107
1098 XG Amsterdam

25 juni 2010

Inhoudsopgave

Abstract

Het probleem met het huidige aanbod van beeldbanken is dat ze verdeeld en ongestandaardiseerd zijn. Door de databases van deze lokale beeldbanken naar een globaal schema te mappen en zodoende een datawarehouse te creëeren en door normalisatie wordt deze data centraal toegankelijk. Door het toepassen van tekstanalyse technieken op deze gegevens wordt dit datawarehous verrijkt met gegevens die voorheen enkel impliciet beschikbaar waren.

1 Introductie en context

Sinds jaar en dag voorziet het internet naast tal van andere informatie ook in digitaal beeldmateriaal. Niet een enkele afbeelding in een webpagina, maar databanken die zich richten op grote collecties beeldmateriaal. Deze banken bevatten prenten, ansichtkaarten, landkaarten en fotomateriaal van verleden tot heden. Maar over het algemeen toch wat ouder beeldmateriaal (Tabel 2). Bij dit fotomateriaal horen ook metadata waaronder een beschrijving, een datum en een locatie. Welke metadata er beschikbaar zijn verschilt per beeldbank.

Afbeelding 1: Voorbeeld van een beeldbank. In dit geval Beeldbank Groningen.Links ziet u een foto, rechts ziet u desbetreffende metadata.

Met dit onderzoek wordt aangetoond hoe het mogelijk is om met technieken uit de information retrieval gegevens met zichzelf te verrijken. Dat wil zeggen dat er informatie die al in de data gelegen is wordt geannoteerd zodat een computersysteem rijkere queries hierop kan toepassen.
De onderzoeksvraag luidt dan ook:

Is het mogelijk om…

1. Door samen voegen van data van verschillende beeldbanken;
2. Door het standaardiseren en normaliseren van waardes;
3. Door het verrijken van deze metadata d.m.v. tekstanalyse technieken, waarbij implicite informatie expliciet wordt gemaakt;

…een webportal te bouwen waarmee beter gezocht kan worden naar beeldmateriaal dan in het huidige aanbod van niet genormaliseerde en tevens gedistribueerde beeldbanken.

2 Aanpak en methoden

2.1 Data verzamelen

De woorden ‘gegevens’ en ‘data’ worden door elkaar gebruikt maar betekenen hetzelfde. Om verwarring met het meervoud van datum (data) te voorkomen wordt van de voorkeursspelling afgeweken en wordt het meervoud datums gebruikt.
De pagina <http://beeldbanken.startpagina.nl> is als begin genomen voor de zoektocht naar goede beeldbanken. Het aanbod van beeldbanken blijkt in kwalilteit uiteenlopend. Door handmatige selectie is het bronmateriaal tot stand gekomen (Tabel 1).

Bronnen van beeldbanken

	Naam	Bron
1	Amsterdam	http://beeldbank.amsterdam.nl
2	BHIC[1]	http://www.bhic.nl
3	Groningen	http://www.beeldbankgroningen.nl
4	Leeuwarden	http://hcl.pictura-dp.nl
5	Nationaal Archief	http://beeldbank.nationaalarchief.nl
6	NIOD[2]	http://www.beeldbankwo2.nl
7	Utrechts Archief	http://www.hetutrechtsarchief.nl
8	Zeeuwse B.	http://beeldbank.zeeuwsebibliotheek.nl

Tabel 1: Onder ‘Bron’ vindt u de webpagina’s die als bron zijn gebruikt voor het verzamelen van de data.

Overzicht bronnen

	Naam	Aantal	Min.	Max.	Gem.
1	Amsterdam	187.191	1100	2020	1881
2	BHIC	106.760	1102	2014	1977
3	Groningen	108.007	1100	2016	1923
4	Leeuwarden	61.632	1160	2011	1830
5	Nationaal Archief	484.422	1100	2012	1967
6	NIOD	118.960	1100	2020	1940
7	Utrechts Archief	111.100	1100	2018	1935
8	Zeeuwse B.	188.616	1100	2011	1967
	Totaal	1.366.688	1100	2020	1927

Tabel 2: Onder ‘Aantal’vindt u het aantal afbeeldingen of records dat de beeldbank bevat. Onder ‘Min.’, ‘Max.’ en ‘Gem.’ vind u respectievelijk de minimale-, de maximale- en de gemiddelde datum van de betreffende beeldbank.
Wat direct opvalt is dat bijna alle beeldbanken 12e eeuw als vroegste afbeelding aangeven. Dat is de vroegste datum die de tijdklassificeerder herkent. Wat ook opvalt is dat de maximale jaartallenvoorbij 2010 komen.Het kan zijn dat deze jaartallen voorkomen omdat er wellicht over de toekomst wordt gesproken in de foto-omschrijvingen maar dat is verder niet onderzocht.

Op deze acht bronnen kan het onderzoek gebaseerd worden. Het gaat hier in alle gevallen om gemeentelijke, provinciale of nationale instanties. Deze instanties hebben hun digitale bron in sommige gevallen kunnen baseren op georganiseerd archiefmateriaal (Zie correspondenties in appendix). D.w.z. archiefmateriaal waarvan de metadata al in enige vorm aanwezig was.

[1] *BHIC – Brabants Historisch Informatiecentrum*
[2] *NIOD – Nederlands Instituut voor Oorlogsdocumentatie*

Een blik op Tabel 2 leert ons direct dat het Nationaal Archief in zijn eentje al één derde van de beeldbank omvat. Verder is het eigenaardig dat datums in de beeldbanken zitten die later zijn dan nu (2010). Ook valt op dat de datums niet lager dan het jaar 1100 komen. Hier wordt later onder Data mutatie → Tijd in gegaan. Wat ook aardig is om te zien is dat het jaar 1940 de gemiddelde datum van het Nederlands Instituut voor Oorlogsdocumentatie is. Dat klinkt in ieder geval logisch. Blijkbaar zit er ook veel fotomateriaal van de aanloop naar WOII in.

2.1.1 Techniek

De opgehaalde gegevens zijn publiekelijk toegankelijk via bovenstaande websites. Het ophalen van deze gegevens is niet triviaal. Al deze websites maken gebruik van een query interface om toegang te krijgen tot de informatie in de vorm van zoekresultaten. Men kan dus enkel alle informatie kopieren door handig op het desbetreffende querysysteem in te spelen.

Voor iedere bron is een programma op maat geschreven om er voor te zorgen dat de gegevens konden worden opgehaald. Zo'n programma heet een scraper (Zie Pseudocode 1). In de meeste gevallen was het zo dat een lege zoekopdracht in alle beschikbare foto's resulteerde, waarna makkelijk over alle overzichtspagina's geïtereerd kon worden om de gewenste informatie op te halen.

Een belangrijke subtaak van de scrapers is het correct parseren van de opgehaalde pagina's zodat het niet steigert op een kleine html syntax error. Hier is de python bibliotheek BeautifulSoup voor gebruikt. Tot groot genoegen lost deze bibliotheek ook moeilijkheden rond karakter encodering op.

BeautifulSoup retouneert een boom van HTML tags, het DOM[3]. Het is nu mogelijk om per pagina door deze boom te navigeren om zodoende de gewenste informatie te abstraheren. Dit proces gaat als volgt. Zoek uit waar de gewenste informatie in de pagina te vinden is. Schrijf code die naar de gewenste plaats navigeert, en haal de informatie op. Om het DOM van een webpagina te inspecteren kan gebruik worden gemaakt van de Firefox[4] plugin Firebug[5].

2.1.2 Problemen

Tijdrovende bezigheid

Het ophalen van gegevens kan wel enkele dagen per beeldbank in beslag nemen. Het kopieren van grote hoeveelheden gegevens via het internet is op zichzelf een tijdrovende bezigheid. Hiernaast is ook belang gehecht aan het Robot Exclusion Protocol. Dit protocol geeft richtlijnen omtrent beschaafd gebruik van publiekelijk toegankelijke databronnen. Hieronder valt ook de crawl-delay directive. Welke voorschrijft het externe robots.txt op te halen en zicht te houden aan de al dan niet gespecificeerde Crawl-delay parameter. Dit resulteert over het algemeen in enkele seconden pauze tussen iedere connectie.

Ontbrekende informatie op overzichtspagina

Normaal vindt men de metadata al op de overzichtspagina[6]. Bij een aantal databanken is per foto een zoekopdracht op foto ID noodzakelijk gebleken. Hierdoor duurde het binnenhalen zeker nog eens tien tot twintig keer zo lang. Aangezien nu voor iedere foto afzonderlijk de crawl-delay in acht moet worden genomen.

Directe links

Voor het webportal is het belangrijk referenties naar het bronmateriaal op te slaan. Hiervoor is het wel noodzakelijk dat de oorspronkelijke beeldbank in permalinks voorziet. Het idee achter een permalink is dat het een url is die altijd naar hetzelfde materiaal zal blijven verwijzen. Sommige van de beeldbanken maken geen gebruik van permalinks. Het wordt zo onmogelijk om in het portal nog terug te verwijzen.

Karakter encodering

Het Nationaal archief blijkt in zijn website problemen met karakter encodering te hebben. Hierdoor is de data vervuild met onleesbare tekens. Een deel van deze problemen zijn handmatig verholpen.

Beperkt aantal zoekresultaten

Sommige sites beperken het aantal zoekresultaten tot bijvoorbeeld duizend hits. Een lege zoekopdracht versturen zodat alle foto's als resultaat worden gegeven werkt hierbij niet. Wanneer men op de site de mogelijkheid biedt per tijdsperiode te zoeken wordt het mogelijk de zoekresultaten onder de duizend te houden.

2.2 Globaal Databankschema

In Data Exchange gaat het om het verplaatsen van data tussen twee relationele gegevensbanken. Er is dus echt sprake van een mapping tussen twee schemas, het lokale schema en het nieuwe globale schema (Zie appendix → globaal schema). In ons geval ligt dat gecompliceerder omdat we geen rechtstreekse toegang hebben tot de bron databanken. We weten niet eens of ze relationeel zijn. We weten zelfs niet eens of er wel gebruik gemaakt wordt van een databank. Afgezien daarvan kunnen we wel de techniek van schema mappings gebruiken. De beeldbanken bieden eigenlijk een 'view' (1) op de database, waarin voor elke foto al zijn publiek toegankelijke attributen met hun waardes worden opgesomd[7].

Het globale schema is niet relationeel maar het is een geneste structuur, vergelijkbaar met XML[8] of JSON[9], in de vorm van een python dictionary[10]. Omdat de gegevens in een dictionary zitten praten we telkens over attributen en attribuutwaarden.

---

[3] *Document Object Model <http://www.w3.org/DOM/>*

[4] *Firefox <http://www.mozilla-europe.org/nl/firefox/>*

[5] *Firebug <https://addons.mozilla.org/en-US/firefox/addon/1843/>*

[6] *Als een zoekopdracht naar een beeldbank wordt verstuurd resulteert dit in een eerste overzichtspagina uit vaak vele duizende overzichts pagina's. Zo'n pagina bestaat uit een verzamling verkleinde foto's veelal gerangschikt in een tabel of een lijst van tussen de 6 tot 24 foto's. Vanaf deze pagina navigeert men naar één van de andere overzichtspagina's of men kiest er voor een foto uit te vergroten.*

[7] *Er kunnen natuurlijk meer attributen aanwezig zijn voor een foto, maar daar kunnen we dan niet bij. Een voorbeeld is de permanente URL van een foto en zijn beschrijving. Die is bij beeldbank X niet publiek beschikbaar, maar prive' natuurlijk wel.*

[8] *XML <http://www.w3.org/TR/REC-xml/>*

[9] *JSON <http://tools.ietf.org/html/rfc4627>*

[10] *Python dict <http://docs.python.org/library/stdtypes.html#dict>*

De voor dit onderzoek belangrijkste aspecten van foto's zijn tijdsdatering, locatie en entiteiten met een naam (named entities). Wanneer aanwezig zijn deze gegevens in de website geselecteerd en op opgeslagen. De conversie van lokaal naar globaal heeft dus plaats gevonden tijdens het ophalen van de gegevens. Er is dan ook geen kopie van de lokale schema's bewaard.
Het globale schema is ontworpen om deze drie aspecten zo precies mogelijk te kunnen weergeven. We behandelen deze aspecten afzonderlijk en daar zullen we de exacte vorm van het globale schema steeds geven. Naast deze drie aspecten is ook telkens de foto-omschrijving opgeslagen. Hierin hopen we de drie aspecten weer te vinden. Wat kans biedt op een verrijking van de data.

2.3 Data mutatie

De waarden van de tijd- en lokatieattributen in het globale schema zijn geformateerd zoals ze dat lokaal bepaald hebben. U kunt zich voorstellen dat er afwijkingen optreden in de manier waarop een beeldbank lokaal zijn attribuutwaarde heeft gerepresenteerd t.o.v. een andere beeldbanken. Teneinde dit doel te bereiken is een klassificeerder geschreven. Deze klassificeert de ruwe waarden en zo kunnen de ruwe waarden genormaliseerd opgeslagen worden.

2.3.1 Tijd

Onder het kopje Tijd → Implementatie tijdklassificeerder leest u hoe de klassificeerder datums probeert te vinden in de ruwe datum waarden. Verder wordt er naar datums gezocht in de foto-omschrijvingen. Hier kan zich extra tijdsinformatie in bevinden. Deze geabstraheerrde informatie wordt in het globale schema opgenomen.

```
'location':{
      'description':[ (start_date, end_date, type,
gran), …]
      'normalized':
            (start_date, end_date, type, gran)
      'source':""
}
```

code fragment 1: Hier ziet u hoe het globale schema onder verdeeld is voor het attribuut tijd. Onder description staan in een lijst alle genormaliseerde datums die in de foto-omschrijving gevonden zijn. Onder normalized staat de genormaliseerde datum die uit het datumattribuut van de lokale vieuw is opgehaald. Onder source staat de niet genormaliseerde versie hiervan.Veder ziet u dat de genormaliseerde datums bestaan uit enkele onderdelen. ***start_date*** *en* ***end_date*** *staan voor een ISO representatie van een datum,* ***type*** *wijst uit of het hier een periode of een tijdspunt betreft.* ***gran*** *staat voor granulariteit en dit geeft informatie over in welke mate de datum gespecificeerd is, waardes hiervoor kunnen zijn 'day', 'month' of 'year'. Wanneer de granulariteit op 'year' staat betekent dit biijvoorbeeld een periode zoals: 1920 - 1930. Een granulariteit met waarde 'day' zou kunnen zijn: 1920-05-12 -1930-07-28*

Het is een mooi gegeven dat er extra tijdsinformatie aan de foto's gerelateerd kan worden. De recall zal hierdoor stijgen, maar er is geen garantie dat de foto alleen gevonden wordt wanneer het ook relevant is voor de gebruiker. Wat zeggen die toegevoegde datums nog over de foto? Om hierop te kunnen antwoorden is een kwalitatieve analyse gemaakt van de toegevoegde datums. Hieruit blijkt dat, van alle datums, zij in de meeste gevallen wel op directe of indirecte manier aan de afbeelding gekoppeld zijn. In andere gevallen is het onduidelijk waar de datum precies voor staat. De categorieën in Tabel 3 zijn handmatig vastgesteld.

Overzicht handmatig vastgestelde semantische tijd-klassen

nr	Categorie	Percentage
1	leefperiode	12%
2	werkperiode	5%
3	creatie/destructie object (gebouw, foto, schilderij)	37%
4	evenement (koninklijk bezoek, manifestatie)	13%
5	ramp (relletjes, overstroming, brand)	5%
6	overig (demping, signering, uitgave)	28%

Tabel 3: Onder het kopje 'Categorie' vind u de verschillende semantische klassen van de gevonden datums in de foto-omschrijvingen. Onder het kopje 'Percentage' ziet u hoe vaak deze klasse voor kwam.
Deze tabel geeft de resultaten weer het kwalitatief onderzoek naar de semantische klassen waarin datums vallen door de gehele datawarehous. Het onderzoek is gebaseerd op 60 handmatig geklassificeerde datums.

Hoewel deze categorieën misschien niet altijd direct met de periode waarin de foto gemaakt is te maken hebben, is de informatie toch relevant want het heeft altijd wel een relatie met wat is afgebeeld. Uiteraard zou het beter zijn als de datums met dit soort categorieën geannoteerd zouden zijn. Dit zou de mogelijkheid bieden om bijvoorbeeld een evenement in mei 1980 te zoeken.

Implementatie tijdklassificeerder

Dit onderzoek baseert zich op de ISO[11] standaard. De ISO datum-tijd richtlijnen schrijven een datum voor die van links naar rechts en van meest naar minst significant attribuut opgeschreven wordt. Het mag ook alleen cijfers bevatten. D.w.z. YYYY-MM-DD:hh:mm:ss. De datums zijn in het globale schema beperkt tot een YYYY-MM-DD representatie. Om de 'scherpte' van de datum aan te duiden wordt telkens de term granulariteit gehanteerd. De fijnste granulariteit in het globale formaat is de dag, daarna de maand en als laatste het jaar.
Binnen het tijdsattribuut van de beeldbanken zijn zes klassen onderscheiden. In Tabel 5 vindt u de zes vastgestelde klassen. Van deze zes klassen zijn alleen de eerste vier in het globale schema opgenomen. Alle klassen worden puur bepaald op basis van syntaxis. Onder de periode-types vallen alle datum types die als eigenschap hebben dat ze een periode beslaan. D.w.z. dat er twee datums gevonden zijn die aan elkaar gerelateerd zijn bijvoorbeeld m.b.v. een verbindingsstreepje. Het kan ook zijn dat er een datumpunt met een verminderde granulariteit wordt gevonden, deze datumpunten zijn geklassificeerd als datumperiode waar naast ook de mate van granulariteit is opgeslagen. Verder is er nog het tijdspunt in de vorm van een enkele volledige datum. De klassificeerder zoekt deze klassen d.m.v. reguliere expressies waarbij de klassen één t/m zes middels een cascade effect worden gezocht. Hierdoor kunnen de categorieën geen overlap vertonen. Het zijn dus werkelijk klassen in de technische zin des woords.

[11] *<http://www.iso.org/iso/support/faqs/faqs_widely_used_standards/widely_used_standards_other/date_and_time_format.htm #what-iso-8601-covers>*

Overzicht variabelen die reguliere expressie vertegenwoordigen

nr	Label	Reguliere expressie
1	dash	'(\W+|\W+en\W+|\W+tot\W+)'
2	dag	'([0-3]|)[0-9]'
3	maand*	'((januari|…|december|jan|…|dec)|((0[0-9])|(1[0-2])|[0-9]))'
4	iso_maand	'(0[0-9])|(1[0-2])|[0-9])'
5	jaar	'(((11|12|13|14|15|16|17|18|19)[0-9][0-9])|(20(0|1)[0-9]))'
6	datum	dag + '\W+' + maand + '\W+' + jaar
7	iso	jaar + '\W+' + iso_maand + '\W+' + dag
8	mdjr	maand + '\W+' + jaar
9	jrmd	jaar + '\W+' + maand
10	context	'(^|\W+)' + reguliere expressie + '($|\W+)'

Tabel 4: De labels onder 'Label' moeten beschouwd worden als variabelen die in Tabel 5 gebruikt worden. Onder 'Reguliere Expressie' ziet u dat de variabelen voor een reguliere expressie staan.
De tijdsklassificeerder bepaalt zijn klassen door een samenstelling van bovenstaande reguliere expressies. Iedere datumklasse is opgebouwd uit één of meer van bovenstaande expressies. Alle reguliere expressies worden ingebed in de reguliere expressie achter het context label. Het + teken, zolang niet binnen de quotes, stelt een concatenatie van twee string-representaties van een reguliere expressie voor. De quotes zijn geen onderdeel van de reguliere expressie.
** Vul voor de drie puntjes (…) de tussenliggende maanden in.*

Overizicht syntactische tijd-klassen met hun reguliere expressies

nr	Klasse	Type	Reguliere expressie	Gran.
1	ISO	point	jaar + '\W+' + iso_maand + '\W+' + dag	dag
2	ISO	periode	iso + dash + iso	dag
3	jrmd*	periode	jrmd + dash + jrmd	maand
4	jaar	periode	jaar + dash + jaar	jaar
5	Overig**			
6	leeg		''	

Tabel 5: Onder 'Klasse'vind u de namen van de verschillende klassen die samen met hun type een unieke klasse voorstellen. Onder 'Type' vindt u deze types. Onder 'Reguliere Expressie'vindt u de reguliere expressies die bepalen tot welke klasse een string behoort. Onder 'Gran.'vindt u de granulariteit van de klasse.
De klassificeerder sorteert alle datums in de bovenstaande zes klassen. Om de reguliere expressies leesbaar te houden is gebruik gemaakt van de labels uit Tabel 4.
** jrmd staat voor jaar-maand*
*** In deze categorie valt alles wat geen lege string is en ook geen van de bovenstaande categorieën. Voorbeelden hiervan zijn 'onbekend', ' ', ' ', '12-apr-1012'*

Evaluatie tijdklassificeerder

Door de klassificeerder is het mogelijk statistische gegevens over de verkregen en samengevoegde data te verzamelen. Om iets te kunnen zeggen over de kwaliteit van deze uitkomsten is het zaak ook de klassificeerder opzich te evalueren. Teneinde dit doel te bereiken is per beeldbank telkens een steekproef van honderd datums uit het datum attribuut genomen en handmatig geklassificeerd. Dit is ook gebeurt met telkens twintig foto-omschrijvingen per beeldbank.
Uit de evaluatie van het datum attribuut is gebleken dat de klassificeerder 100% correct klassificeert. Hieruit blijkt dat men lokaal per beeldbank ook standaardisatie heeft toegepast, bijvoorbeeld in de vorm van de gebruikers interface. Verder zijn er geen context verschijnselen waarop de reguliere expressies zouden kunnen falen, zoals men dat wel heeft bij de foto-omschrijvingen. Desalniettemin gaat de klassificatie ook bij de foto-omschrijvingen slechts in enkele gevallen mis. Het betreft hier enkele huisnummers die boven de elfhonderd[12] komen, enkele identificatienummers die in de omschrijving terecht zijn gekomen en tenslotte nog wat andere eenheden die boven de elfhonderd komen zoals meters en kilogrammen. Voor een exacte tellingen bekijkt u de resultaten onder Resultaten.

Problemen tijdklassificeerder

De datums van het Utrechts Archief hebben de klassificeerder echt op de proef gesteld. Strikt genomen zijn de klassificaties correct maar er lijkt iets anders aan de hand.
Wanneer u de tabel met reguliere expressies bekijkt ziet u dat enkele van de volgende datums tot de verkeerde klasse worden gerekend.
'22/01/1976(22 januari 1976, 20e eeuw)' wordt een ISO van het type periode. Het haakje matcht met de '\W+' uit dash. Hierom wordt deze datum geklassificeerd als een periode. Om dit probleem te verhelpen is de klassificeerder uitgebreid met een functie die stelt dat een periode niet uit twee dezelfde datums mag bestaan.
01/01/1900 – 31/12/1957 wordt als een ISO van het type periode geklassificeerd met een granulariteit van het type dag. Ook deze klassificatie lijkt op het eerste gezicht correct. Maar het is een periode met een ongewenste granulariteit. Deze periode drukt feitelijk een periode jaar uit. Namelijk 1900 – 1957. De klassificeerder is uitgebreid met een functie die dit verhelpt..

2.3.2 Locatie

Tijd en plaats zijn qua aanpak vrijwel hetzelfde gebleken. Het oorspronkelijke locatieattribuut wordt geklassificeerd en genormaliseerd. De data worden verrijkt door naast het klassificeren en normaliseren van het locatieattribuut ook het descriptieattribuut te doorzoeken. Behalve deze verrijking is de data ook aangevuld met geocodes in de vorm van een lengte en een breedte graad.
Veel van de foto's zijn oorspronkelijk al voorzien van een locatieattribuut. Sommige waarden hiervan bestaan enkel uit een straatnaam en een nummer, weer andere hebben hierbij ook een plaatsnaam. Bij de tijden is onderscheid gemaakt naar granulariteit. Bij de plaatsen is iets dergelijks gebeurd. Wanneer er een plaats, een straat en een huisnummer bekend is kan de locatie nauwkeuriger bepaald worden dan wanneer er alleen een plaats en een straat is.
Op basis van dit gegeven is een klassficeerder gebouwd. Deze heeft per record klassen gezocht. In Tabel 6 ziet u welke klassen er bepaald zijn en welke reguliere expressies daar bij horen.
Het is logisch dat er records bestaan waarbij meerdere locaties in het descriptie attribuut worden gevonden. Een verschil met de klassificatie van de tijden is dat er deze keer ook in het locatieattribuut meerder locaties zijn gevonden.

[12] Zoals in Tabel 4 is te zien wordt een cijfer pas vanaf elfhonderd als datum geinterpreteerd.

Implementatie locatieklassificeerder

Om de gegevens te klassificeren is er een main-loop die over de records van de database loopt. Binnen deze loop wordt de locatieklassificeerder telkens aangeroepen. De klassificatie vindt plaats op het locatieattribuut én op het descriptie attribuut.

De klassificeerder probeert in de tekst eerst klasse 1 te vinden (zie Tabel 6). Als de reguliere expressie matcht zal hij de de gematchte tekst verwijderen, en nogmaals dezelfde reguliere expressie op de aangepaste tekst toepassen. Wanneer er geen match meer is, stapt hij naar de 2e klasse. Ook hier probeert hij uitputtend te matchen. Dit gebeurt tot en met de 5e klasse. Wanneer alle klassen uit de tekst gefilterd zijn worden deze aan het database record toegevoegd. In het globale schema is ruimte gemaakt om de gevonden informatie gestandaardiseerd op te slaan.

```
'location':{
      'description':{..}
      'normalized':{
            'city':{
                  'Amsterdam':{
                        'lng':52.37,
                        'lon': 4.89
                  }
                  '…':{…}
             },
            'city street':{…},
            'city street number':{…},
            'street':{…},
            'street number':{…},
      }
      'source':""
}
```

code fragment 2: Weergave van de structuur van het globale schema voor het aspect plaats.Location wordt verder onderverdeeld in drie labels description, normalized en source. Onder descriptionvind men exact dezelfde onderverdeling als onder normalized. Het verschil zit hem erin dat hier de gegevens staan die de klassificeerder in de foto-omschrijving heeft gevonden. Onder normalized staan de klassen en hun instanties die de kalssificeerder in het locatieattribuut van de lokale view heeft gevonden. onder source staat wat het oorspronkelijke waardce van locatieattribuut in de view van de lokale schema was.

Voor klasse 1 maakt het niet uit of plaats voor, of na de straat en het huisnummer komt. Beiden schrijfwijzen komen voor in het Nederlands. Dat geld ook voor de 3e klasse 'plaats straat'. Een straat-huisnummer combinatie hebben we als ondeelbaar beschouwd.

Overzicht syntactische locatie-klassen

nr	Klasse	Reguliere expressie
1	plaats straat nummer	'(^|\W+)(' + plaatsnaam + '[A-Z][a-z]+' + suffix + '(\W+\d+)+)($\|\W+)' of '(^\|\W+)([A-Z][a-z]+' + suffix + '(\W+\d+)+' + plaatsnaam + ')($|\W+)'
2	straat nummer	'(^|\W+)([A-Z][a-z]+' + suffix + '(\W+\d+)+)($|\W+)'
3	plaats straat	'(^|\W+)(' + plaatsnaam + '[A-Z][a-z]+' + suffix + ')($\|\W+)' of '(^\|\W+)([A-Z][a-z]+' + suffix + '\W+' + plaatsnaam + ')($|\W+)'
4	plaats	'(^|\W+)(' + plaatsnaam + ')($|\W+)'
5	straat	'(^|\W+)([A-Z][a-z]+' + suffix + ')($|\W+)'
6	leeg	'' of '(^\W+$)'

Tabel 6: Onder 'Klasse' vind uwelke klassen er zijn onderscheiden voor locaties. Onder 'Reguliere Expressie' ziet u welke expressies gebruikt zijn om in de tekst deze klassen te onderscheiden.Binnen de reguliere expressies bevinden zich de variabelen 'plaatsnaam' en 'suffix'. Terwijl de reguliere expressies in een loop tegen de tekst worden gematcht worden 'plaatsnaam' en 'suffix' telkens afgewisseld met respectievelijk één van de verzameling straatnamen en één van de verzameling suffixen.
Voor 'plaats straat nummer' en voor 'plaats straat' is bepaald dat er twee mogelijkheden zijn. De plaats voor de straat of juist erachter. Dit voorkomt mede dat er niet onterecht plaatsen en straten in de klassen 'plaats' en 'straat' terecht komen.
Het + teken, zolang niet binnen de quotes, stelt een concatenatie van twee string-representaties voor. De quotes zijn geen onderdeel van de reguliere expressie.

De klassen zijn gebaseerd op reguliere expressies (Tabel 6). Hierin komt u telkens de variabelen 'plaatsnaam' en 'suffix' tegen. 'plaatsnaam' wordt gesubstitueerd door een plaatsnaam uit een lijst van plaatsnamen. 'suffix' wordt gesubstitueerd door een lijst van suffixen (code fragment 3).

```
suffixen = ['boulevard', 'brug', 'gracht', 'dijk'
            'kade', 'laan', 'markt', 'pad',
            'park', 'passage', 'plantsoen',
            'plein', 'singel', 'steeg', 'straat',
            'wal', 'weg']
```

code fragment 3: De lijst met suffixen voor straatnamen. Deze worden gebruikt om straatnamen te herkennen in de reguliere expressies uit Tabel 6 .

Straatnamen zijn redelijk goed te vinden op basis van suffix. Voor plaatsnamen zijn er geen veel vookomende suffixen. Het is daarom noodzakelijk gebleken een lijst van Nederlandse plaatsnamen tegen alle database records te matchen. Zowel voor het locatie-attribuut als voor het decsriptie-attribuut. In eerste instantie is dat geprobeerd met een plaatsnamenbestand van 5.400 Nederlandse plaatsnamen. Om deze plaatsnamen binnen een redelijke tijd te kunnen matchen is het de slimste keuze om alle plaatsnamen als één grote disjunctie in de reguliere expressie te verwerken. Helaas kan de reguliere expressie machine van Python dit niet aan.

Een andere aanpak is om door de lijst met plaatsnamen te loopen en de plaatsnaam telkens te substitueren in de reguliere expressie. Dit betekent 5.400 x 2 x 1.366.688 zoekoperaties. Een operatie die in weken berekeningstijd zou lopen.

Door het plaatsnamenbestand heftig in te korten wordt de berekeninstijd verkort. Omdat bekend is waar de beeldbanken gelegen zijn bleek het handig om een stedenlijst samen te stellen van belangrijke steden in de buurt van de beeldbank, om de kans van herkenning te vergroten. Bij het samenstellen van deze lijsten is uit gegaan van gemeentenamen. Waarachter de opvatting schuil gaat dat wanneer een stad tot gemeente wordt uitgeroepen, deze stad belangrijk moet zijn voor de provincie.

Evaluatie locatieklassificeerder

Het evalueren van een lokatieklassificeerder is lastiger dan het controleren van een tijdklassificeerder. Aan een datum is goed te zien of deze als juist is geklassificeerd. Aan een plaatsnaam is dat niet altijd te zien. Denkt u aan een plaatsnaam als 'Koewacht". Niemand die daar niet uit de buurt komt zal weten dat het een plaats in Zeeland betreft. Een locatieklassificeerder zal dus geëvalueerd moeten worden door een persoon met goede kennis van plaatsnamen. Gezien mijn Zeeuwse afkomst is het

verstandig dan ook de klassificaties op de Zeeuwse beeldbank te evalueren.

Percentages bij evaluatie

	Pos	Neg
True	49%	18%
False	0%	33%

Tabel 7: Op de kruising van 'True' en 'Pos' vindt u het percentage locatieattributen dat terecth als 'stad' geklassificeerd is. Op de kruising van 'True' en 'Neg' vindt u het percentage dat terecht niet als 'stad' geklassificeerd is. Op de kruising van 'Neg' en 'False' vindt u het percentage dat wel een 'stad' is, maar niet als zodanig herkend is. Op de kruising van 'Pos' en 'False' vindt u het percentage dat onterecht als 'stad' geklassificieert is.

In Tabel 7 ziet u hoe deze percentages zijn uitgepakt. Het betreft hier een trekking van 100 locatieattributen. Onder de True Negatives bevinden zich landstreken en riviernamen en ook in enkele gevallen de naam van de provincie, Zeeland dus.

Problemen locatieklassificeerder

Beperking reguliere expressie machine van Python
De reguliere expressie machine van python kan maar een beperkt aantal karakters aan. Wanneer dit aantal wordt overschreden resulteert dit in een runtime error met de mededeling 'maximum recursion limit'[13].

Google Query limiet
Google heeft een limiet op de Google Maps Geocoding API[14] van 2500 requests per dag per IP adres. Het is niet gelukt om voor de deadline alle locaties tot lengte en breedte graden om te zetten. Het opvragen van coordinaten voor alle locaties zou bijna vijf weken in beslag nemen.

Meerdere locaties per foto
Anders als bij het datumattribuut staan er soms meerdere plaatsen in het locatieattribuut. Hierdoor is niet meer te stellen dat een foto in één locatie-klasse valt. Daarnaast kan het zijn dat één record meerdere van dezelfde lokatieklassen bevat. Er ontstaan hierdoor anti-intuïtive waarden zoals de 119% gevonden klassen bij Beeldbank Groningen (Tabel 14). Maar er zijn nog andere problemen.
Stel dat er feitelijk drie locaties in het locatieattribuut staan, maar dat de klassificeerder er daar maar één als klasse herkent. In principe zouden er dan twee datums aan de categorie 'Overig' moeten worden toegevoegd.

2.3.3 Naam dragende entiteiten

Het laatste aspect van dit onderzoek betreft naam dragende enteiten, of in de meer gangbare engelse termen 'named entities'. Voorbeelden hiervan zijn namen van gebouwen, bruggen, mensen en locaties.

NER klassificeerder

Wat de NER Fietstas doet is meer dan alleen klassificeren. Het programma bestaat uit drie subprogramma's:

- Named entity Recognition
- Named entity Classification
- Named entity Normalisation

Het recognition gedeelte zorgt er voor dat named entities worden herkend. Een simpele reguliere expressie hiervoor zou kunnen zijn '([A-Z][a-z]+\s+)+'. Deze expressie kom neer op het herkennen van woorden die beginnen met een hoofdletter.
De klassificeerder maakt onderscheid in vier klassen. Locatie (LOC), organisatie (ORG), persoon (PER) of overig (MISC).
Het normaliseren van named entities gebeurt met behulp van wikipedia pagina's. Zoals men wellicht weet maakt wikipedia gebruik van redirects. Redirects op wikipedia worden in eerste plaats ingezet om synonymen naar dezelfde inhoud te verwijzen. Maar wanneer het named entities betreft bestaan er vaak ook vele goede schrijfwijzen die naar dezelfde entiteit verwijzen. Fietstas kent deze verwijzingen en kan op basis hiervan het wikipedia adres van die named entity als normalisatienaam gebruiken.
In het globale schema bestaat nu per databse record een label 'entity'. Dit label is verder onderverdeeld in de brontekst, de normalisatie en de categorie.
Stel we hebben de volgende foto-omschrijving:

> *"Koningin Wilhelmina tijdens het uistpreken van de troonrede bij de opening van de Staten-Generaal"*

In de database zal dit er zo uitzien:

```
'entity':{
    'http://nl.wikipedia.org/wiki/Wilhelmina_de
        r_Ned erlanden':{
          'class':'PER'
          'source':'Koningin Wilhelmina'
     }
     'http://wiki.../Some_Entity':{
          …
     }
     …
}
```

code fragment 4: Weergave de structuur van het globale schema voor het aspect named entity en tevens een mogelijke instantiatie hiervan.

Met de lijsten van deze verschillende schrijfwijzen kunnen we de gebruiker, ook bij een alternatieve spelling, naar het juiste document sturen. Dit zal de recall verhogen.

Evaluatie NER Klassificeerder

ook bij de NER is een evaluatiestap toegepast, waarvan u in de Tabel 18 resultaten ziet. Deze evaluatie is gebaseerd op een handmatige klassificatie. Er is een twintigtal foto-omschrijvingen automatisch en handmatig geklassificeerd. De klassificeerder vond hierbij honderdzes named entities. Handmatig werden er maar negenenzestig entiteiten gevonden. Alleen dat gegeven wijst al op een matige precisie.

[13] < http://docs.python.org/library/re.html#avoiding-recursion>
[14] < http://code.google.com/intl/nl/apis/maps/documentation/geocoding/#Limits>

Percentages bij evaluatie

	LOC		MISC		ORG		PER	
	Pos	Neg	Pos	Neg	Pos	Neg	Pos	Neg
True	30%	48%	0%	89%	2%	87%	9%	76%
False	8%	14%	11%	0%	4%	7%	15%	0%

Tabel 8: In de bovenste rij van de header vindt u de klassen die de NER herkent. Voor deze klasse is telkens een subkolom 'Pos' en 'Neg' aangemaakt. Verder zijn er de rij-labels 'True' en 'False'.
Op de kruising van 'True' en 'Pos' vindt u het percentage van de desbetreffende klasse dat terecth als die klasse geklassificeerd is. Op de kruising van 'True' en 'Neg' vindt u het percentagevan de desbetreffende klasse dat terecht niet als die klasse geklassificeerd is. Op de kruising van 'Neg' en 'False' vindt u het percentage van de desbetreffende klasse die als zodanig geklassificeert had moeten worden, maar die niet als zodanig geklassificeerd zijn. Op de kruising van 'Pos' en 'False' vindt u het percentage dat onterecht als de betreffende klasse geklassificieert is.

De NER kan ook lokaties vinden in de foto-omschrijvingen. Uit de vergelijking van de lokatieklassificeerder met de lokatie klasse uit de NER blijkt dat de NER dat wel iets beter doet. Dat mag ook wel, want de lokatieklassificeerder biedt nog heel veel ruimte voor verbetering. Daarnaast is in de lokatieklassificeerder een lichte mate van domeinkennis ingebouwd in de vorm van plaatsnaamlijsten. Terwijl de NER juist geschikt moet zijn voor vele domeinen. Hierom is hij in het nadeel.
Hoewel het misschien te begrijpen is hoe het komt dat de NER klassificeerder het niet veel beter doet dan de lokatieklassificeerder speelt er nog wel een ander probleem. Het is zo dat deze klassificeerder, in tegenstelling tot de andere twee, meer false positives maakt. Dat is bij de lokcatieklassificeerder, ten koste van de recall, veel minder het geval (afgerond 0% zelfs).
Wanneer de resultaten van de NER klassificeerder in de data zijn verwerkt zorgt dit dus voor een hogere recall, maar de precisie gaat achteruit.

Problemen NER klassificeerder

Karakter encodering

Bij het analyseren van de data die terug is gekomen van de NER klassificeerder is na het analyseren van de resultaten vast gesteld dat de data niet langer alle beeldbanken bevatte. Beeldbanken Leeuwarden en Nationaal Archief zijn hierom niet in de resultaten verwerkt. Door problemen met karakter encodering heeft het NER klassificatie proces langer geduurd dan gehoopt. Ten tijde van constatering van deze onvolledige resultaten was geen tijd meer om dit probleem terug te koppelen en op te lossen.

3 Resultaten

3.1 Tijd

Uit de resultaten van de Klassifceerder blijkt dat Beeldbank Leeuwarden en Beeldbank Zeeland de grootste uniformiteit te vertonen in hun tijdsattribuut. Voor beiden beeldbanken geldt dat ze of een tijdspunt of helemaal niks hebben. Bij de andere beelbanken maakt men juist meer gebruik van periodes.
Wanneer we de periodes verder uitsplitsen op basis van granulariteit zien we dat ook hier Beeldbank Leeuwarden erg consequent is voor de acht procent periode's want deze hebben allen dezelfde granulariteit. Sowieso komen periodes, met een granulariteit van het type jaar, veel voor, zowel in de attributen als in de foto-omschrijvingen.

Verdeling klassen tijdsattribuut

	Beeldbank	ISO	Periode	Overig	Leeg
1	Amsterdam	37%	63%	0%	0%
2	BHIC	14%	70%	2%	14%
3	Groningen	14%	77%	0%	8%
4	Leeuwarden	75%	8%	0%	16%
5	Nationaal Archief	22%	78%	0%	0%
6	NIOD	28%	70%	0%	1%
7	Utrechts Archief	45%	40%	0%	15%
8	Zeeuwse B.	31%	0%	0%	69%
	Totaal	33%	51%	0%	15%

Tabel 9: Onder 'Beeldbank' vindt u op welke beeldbank de percentages van toepassing zijn. Onder 'ISO' staat het percentage records dat een ISO datum van het type 'point' bevat. Onder 'Periode' staat het percentage records dat een periode bevat. Onder 'Leeg' staat het percentage records dat niks bevatte.
Deze tabel geeft voor alle beeldbanken weer welke datum types gevonden zijn en hoe deze verdeeld zijn in de beeldbank. Het betreft hier datums uit het datum attribuut, en niet uit de foto-omschrijvingen. Ongeveer de helft van de foto's is voorzien van een tijdsattribuut. Het Nationaal archief en het NIOD blinken hier uit in uniformiteit.

Verdere onderverdeling 'Periode kolom

	Beeldbank	dag	maand	jaar
1	Amsterdam	2%	19%	42%
2	BHIC	1%	4%	65%
3	Groningen	1%	11%	66%
4	Leeuwarden	0%	0%	8%
5	Nationaal Archief	3%	17%	58%
6	NIOD	0%	9%	61%
7	Utrechts Archief	0%	5%	34%
8	Zeeuwse B.	0%	0%	0%
	Totaal	1%	8%	42%

Tabel 10: Onder 'Beeldbank' staat op welke beeldbank de percentages van toepassing zijn.Onder 'ISO' staat het percentage records dat een datum van de klasse 'ISO'van het type 'period' bevat. Onder 'jrmd'staat het percentage records dat een datum van de klasse jaar-maand van het type 'period' bevat. Onder 'Jaar' staat het percentage records dat een datum van de klasse Jaar' van het type 'period'bevat.
Deze tabel geeft de verder opsplitsing van de periode klassen uit Tabel 9 weer. De percentages zijn allemaal relatief tot de percentages onder periode in Tabel 9.

Het aantal datums dat gemiddeld per record gevonden wordt in de omschrijvingen

	Beeldbank	0	1	2	3	4	Ov.
1	Amsterdam	91%	6%	1%	0%	0%	0%
2	BHIC	75%	21%	2%	0%	0%	0%
3	Groningen	90%	7%	1%	0%	0%	0%
4	Leeuwarden	80%	13%	3%	1%	0%	0%
5	Nationaal Archief	96%	3%	0%	0%	0%	0%
6	NIOD	75%	18%	3%	1%	0%	0%
7	Utrechts Archief	86%	10%	2%	0%	0%	0%
8	Zeeuwse B.	96%	2%	0%	0%	0%	0%
	Totaal	90%	7%	1%	0%	0%	0%

Tabel 11: Onder het kopje 'Beeldbank' staan de beeldbanken waarop de percentages van toepassing zjin. Onder'1' t/m '4' staan de percentages van het aantal records dat gevonden is in de foto-omschrijvingen en dat respectievelijk dat aantal datums bevat. Onder 'Ov.' vindt u de percentages van het aantal records dat meer dan vier datums bevatte.
Sommige van de verwaarloosbaar lage percentages zijn naar nul afgerond.

Verdeling tijd-klassen in omschrijvingen

	Beeldbank	ISO	per. dag	per. maand	per. jaar
1	Amsterdam	13%	3%	5%	79%
2	BHIC	10%	0%	2%	87%
3	Groningen	22%	1%	3%	74%
4	Leeuwarden	7%	0%	2%	91%
5	Nationaal Archief	11%	0%	14%	75%
6	NIOD	38%	1%	14%	46%
7	Utrechts Archief	48%	0%	1%	52%
8	Zeeuwse B.	19%	0%	4%	76%
	Totaal	22%	0%	7%	70%

Tabel 12: Onder het kopje Beeldbank staan de beeldbanken waarop de percentages van toepassing zjin. De rest van de kopjes zijn de klassen die het tijdattribuut kent.
Deze tabel geeft aan wat de verdeling is van de verschillende tijdsklassen in alle descriptions. Deze percentages zijn relatief aan het aantal gevonden datums, en dus niet aan het aantal records in de databse.
Sommige van de verwaarloosbaar lage percentages zijn naar nul afgerond.

Percentage datums wanneer descripties of tijd attribuut al dan niet leeg is.

	A = Leeg	A = Vol	
D = Leeg	12.95%		77.27%
D = Vol	1.91%	dubbel*	1.34%
		uniek**	6.52%
		totaal	7.86%

Tabel 13: 'A' staat voor locatieattribuut, 'D' staat voor description. 'D=Leeg' of 'A=Leeg' betekent dat er geen datum is gevonden in respectievelijk de description of de attribuutwaarde.Bij 'D=Vol' of 'A=Vol'is er juist wel een datum gevonden.
Deze tabel geeft het percentage van records aan dat verrijkt is. Wanneer er geen datum in het oorspronkelijke tijdsattribuut is gevonden, maar wel in de foto-omschrijvingen dan hebben we hiermee de data verrijkt. Wanneer zowel het tijdsattribuut als de foto-omschrijving een datum blijkt te bevatten is middels deze tabel aangegeven welk percentage van de foto-omschrijvingen dezelfde datum bevat als het tijdsattribuut. Wanneer dit het geval is betekent het dat deze datum dubbel is opgeslagen, dit hoeft echter niet te betekenen dat er niet nog meer datums zijn gevonden in de omschrijving. Wanneer dat het geval is kan toch nog een verrijking hebben plaatsgevonden.
** De datum in het attribuut is ook gevonden in de lijst van gevonden datums in de foto-omschrijving*
*** De datum in het attribuut is niet gevonden in de lijst van gevonden datums in de foto-omschrijving*

3.2 Plaats

Wanneer we de resultaten van de lokatieklassificeerder bekijken zien we direct dat het merendeel uit enkel plaatsen bestaat. En dat ook een groot deel niet geklassificeerd wordt. Hier is al uitgebreid op ingegaan tijdense het bespreken van de klassificeerder.
Beeldbank Utrecht doet het heel slecht. De rede hiervoor is dat ze geen locatieattribuut in hun view hebben. Verder blijkt uit deze resultaten dat er ook in de omschrijvingen geen locaties zijn gevonden.
Beeldbank Groningen daarentegen doet het eigenaardig goed. Het percentage kan boven de 100% stijgen omdat er meerdere locaties in het locatieattribuut kunnen zitten.

Verdeling locatie klassen in locatieattribuut

	Bron	Klasse	Overig	Leeg
1	Amsterdam	78%	22%	16%
2	BHIC	72%	28%	1%
3	Groningen	119%	3%	9%
4	Leeuwarden	90%	10%	17%
5	Nationaal Archief	16%	49%	35%
6	NIOD	6%	94%	0%
7	Utrechts Archief	0%	0%	100%
8	Zeeuwse B.	55%	45%	0%
	Totaal	44%	38%	24%

Tabel 14: Onder 'Klasse'vind u het percentage gevonden klassen t.o.v. het aantal records in de database. Omdat er meerdere datums per record gevonden kunnen worden, bestaat de mogelijkheid, zoals bij Groningen, dat het percentage boven de 100% uit stijgt. Onder 'Overig' vind u het percentage locatieattributen dat niet leeg was, maar waar de klassificeerder ook niet instaat is gebleken een klasse toe te wijzen.Ook hier is het percentage moeilijk te begrijpen, voor meer informatie leest u sectie "2.3.2 →Problemen →Meerdere locaties per foto". Onder 'Leeg' vind u het percentage records dat leeg was, en waar dus ook geen klasse is gevonden.

Verdere onderverdeling 'Klasse' kolom

	Plaats Straat Nummer	Straat Nummer	Plaats Straat	Plaats	Straat
1	0%	0%	56%	1%	43%
2	10%	4%	13%	67%	6%
3	0%	5%	0%	68%	27%
4	0%	31%	0%	46%	23%
5	0%	0%	0%	99%	1%
6	0%	0%	0%	99%	1%
7	0%	0%	0%	0%	0%
8	0%	0%	0%	100%	0%
	1%	5%	15%	60%	19%

Tabel 15: Als kop ziet u de telkens de klassen die het locatieattribuut kent. Dit is een onderverdeling van de kolom 'Klasse' in Tabel 14 .Op de onderste rij ziet u de totale percentages. De nummers 1 t/m 8 stellen de beeldbanken voor in dezelfde volgorde als in Tabel 14. De percentages zijn relatief aan het desbetreffende percentagein de 'Klasse'kolom in Tabel 14.De cijfers in de meest linkse kolom corresponderen met de cijfers en de beeldbanken uit Tabel 14.

Verdeling locatie klassen in descriptions

	Plaats Straat Nummer	Straat Nummer	Plaats Straat	Plaats	Straat
1	0%	38%	0%	7%	56%
2	2%	4%	10%	32%	27%
3	0%	1%	0%	6%	33%
4	0%	0%	0%	13%	3%
5	1%	34%	0%	84%	51%
6	0%	1%	0%	6%	5%
7	0%	5%	0%	12%	16%
8	0%	1%	0%	4%	7%
	0%	11%	1%	20%	25%

Tabel 16: Als kop ziet u de telkens de klassen die het locatieattribuut kent. De percentages zijn relatief genomen aan het aantal database records en ze slaan op het aantal locaties die gevonden zijn in de omschrijvingen.

	A = Leeg	A = Vol
D = Leeg	39%	23%
D = Vol	15%	23%

Tabel 17: : 'A' staat voor locatieattribuut, 'D' staat voor description. 'D=Leeg' of 'A=Leeg' betekent dat er geen locatie is gevonden in respectievelijk de description of de attribuutwaarde. Bij 'D=Vol' of 'A=Vol' is er juist wel een lokatie gevonden.
Deze tabel geeft het percentage van records aan dat verrijkt is. Wanneer er geen datum in het oorspronkelijke tijdsattribuut is gevonden, maar wel in de foto-omschrijvingen dan hebben we hiermee de data verrijkt.
Een percentage van 15% aangevuld lijkt heel groot vergelijken bij de 1.91% van het tijdattribuur. Maar zoals u hier onder inTabel 18 kunt zienbestaat hiervan een groot deel uit een 'straat nummer' of een 'straat'. De twee klassen waar tot op heden nog geen geocodes voor zijn gevonden.

Unieke locaties in descsriptie- én locatieattribuut

Klasse	Geklassificeerd	Waarvan geocodes
plaats straat nummer	7.154	7.122
straat nummer	70.100	
plaats straat	3.081	3.044
plaats	233	232
straat	13.215	
totaal	93.783	10.398

Tabel 18: Onder 'Klasse' vindt u de bekende klassen die onderscheiden zijn voor de locatie. Onder geklassificeerd vind u het aantal unieke locaties die gevonden zijn door de klassificeerder, door alle beeldbanken heen voor zowel het locatieattribuut als de descriptiie. Onder 'Waarvan geocode' vindt u welk van deze unieke locaties door google omgezet zijn in een lengte- en een breedtegraad.
'straat nummer' en 'straat' zijn niet omgezet naar geocodes. Zie '2.3.2 → problemen → Google query limiet' voor meer informatie. Verder zijn deze klassen zo onprecies dat google met meerdere mogelijkheden zal terugkomen.

Afbeelding 2: Wat u hier aan witte puntjes over Nederland verspreid ziet zijn de 10.398 unieke geocodes uit Tabel 18 .Ieder wit puntje stelt één of meer afbeeldingen voor.Aan de verdeling is de herkomst van de beeldbanken vrij goed terug te zien. U ziet duidelijke concentraties rond Walcheren, heel Brabant,Amsterdam en Utrecht, Leeuwarden en Groningen.

3.3 Entiteit

Zoals meteen zichtbaar wordt in onderstaande tabel zijn de beeldbanken Nationaal Archief en Leeuwarden niet meegenomen in de resultaten. Deze problematiek reeds besproken onder Evaluatie NER klassificeerder.
Wat wel eigenaardig is, is dat de NER de meeste locaties in de beeldbank van het Utrechts Archief vindt, terwijl daar nu juist geen locaties zijn gevonden door de lokatieklassificeerder. Dit is onlosmakelijk verbonden met de problematiek rond de lokatieklassificeerder.

Verdeling klassen NER

	Beeldbank	LOC	MISC	ORG	PER
1	Amsterdam	43%	13%	16%	28%
2	BHIC	29%	8%	11%	52%
3	Groningen	59%	7%	8%	27%
4	Leeuwarden				
5	Nationaal Archief				
6	NIOD	31%	20%	16%	34%
7	Utrechts Archief	60%	5%	9%	25%
8	Zeeuwse B.	36%	10%	12%	42%
	Totaal	43%	10%	12%	35%

Tabel 19: Onder 'Beeldbank' vindt u op welke beeldbank de percentages van toepassing zijn.Verder ziet u In de header van de tabel de klassen die de NER herkent. Daaronder ziet u de percentages van iedere klasse die er in de foto-omschrijvingen te vinden zijn per beeldbank en in totaal.
De beeldbank Leeuwarden en het Nationaal Archief zijn leeg omdat deze tentijden van dit schrijven nog niet berekend waren.

4 Conclusie

In zijn algemeenheid kan geconcludeerd dat er verrijking heeft plaatsgevonden via de drie aspecten tijd, plaats en enteit. Er zijn afbeeldingen beschikbaar gemaakt die voorheen niet beschikbaar waren op basis van die zoekcriteria. De percentages lijken soms wat tegen te vallen

maar omgerekend naar een absoluut aantal afbeeldingen gaat het om ongeveer 25.000 afbeeldingen die eerst niet en nu wel op datum te vinden zijn. En om ongeveer 210.000 afbeeldingen die eerst niet en nu wel op locatie te vinden zijn. Daarnaast zijn ook records aangevuld met extra datums en locaties. Van named entities was in geen enkele beeldbank sprake, dus dat aspect is met 100% verrijkt.
Een vraag waar wel over nagedacht kan worden is wat deze verrijkingen precies inhouden. Wat zeggen die lokaties en datums nog over de foto? Het wordt uit Tabel 3 duidelijk dat van datums niet gesteld kan worden dat het het moment van de foto-opname is. Ook van de lokaties kan niet gesteld worden dat deze altijd de fotolokatie zijn. Dat blijkt wel uit dit sample:

> *"Afbeelding van enkele tafels afkomstig uit het 17e eeuwse poppenhuis, onderdeel van de collectie van het Stedelijke Museum van Oudheden te Utrecht, gevestigd in het huis Het Hogeland (Museumlaan 2) te Utrecht. Het museum was van 1890 tot 1921 op dit adres gevestigd . Daarna werd de collectie integraal opgenomen in het Centraal Museum (Agnietenstraat 1)"*

Als laatste wordt bijvoorbeeld de lokatie van het Centraal Museum aangeduid, en niet de lokatie van de foto. Toch hoeft dat niet uit te maken. Het gaat hier in de omschrijving om waar de collectie allemaal heeft gestaan. En het heeft dus wel te maken met wat er op de foto wordt afgebeeld. Er kan zelfs algemeen gesteld worden: Alles wat in de omschrijving staat heeft met de foto te maken anders waren zij niet als zodanig beschreven. Dus alles wat aan impliciete informatie expliciet wordt gemaakt zal met de foto te maken moeten hebben. Daarnaast, als dat niet het geval zou zijn is dat eerder een kritiek op het bronmateriaal dan op de klassificeerders.

5 Discussie en toekomstig werk

In de conclusie wordt gesteld dat alles wat aan impliciete informatie uit de foto-omschrijving expliciet wordt gemaakt altijd met de foto te maken zal hebben. Hoewel dat waar is zal ons mensen toch opvallen dat hier in de meeste gevallen wel een kwalitatieve rangschikking mogelijk is. Vooralsnog is het met name de recall die omhoog zal gaan, maar er zal vervuiling op treden als datums die indirect met een foto te maken hebben even zwaar geranked zullen worden als datums die direct met het afgebeelde te maken hebben. Ik neem nu telkens het tijdattribuut als voorbeeld, maar dit geldt ook voor lokatie en entiteit. Toekomstig werk zou zich kunnen richten op een maat voor relevantie van deze drie aspecten.
Momenteel is het alleen nog mogelijke de database te raadplegen d.m.v. een programmeertaal. In de toekomst moet dit via een query interface gaan.

Tijd
Wanneer er enkel een jaartal is genoteerd is de tijdklassificeerder op dit moment niet in staat dit te herkennen wanneer datums beneden het jaar 1100 komen. Sommige van de landkaarten stammen uit deze tijd en uit vroegere tijden. Het nadeel van lagere jaartallen is dat ze makkelijker verward worden met andere nummers die in de omschrijvingen voor komen, zoals huisnummers en identificatienummers. Door het gebruik van heftigere tekstanalysetechnieken zal dit mogelijk zijn. Het gebruik van n-grammen zal hier mogelijk uitkomst kunnen bieden.

Plaats
De plaatsklassificeerder biedt nog veel ruimte voor verbetering. Sommige van de ontwerpbeslissingen zijn in een haast voor het genereren van resultaten niet voldoende doordacht. Hierdoor is de recall omlaag gegaan.
Ik heb het dan over het integreren van meer kennis in de klassifiseerder zonder daarbij al te veel computatietijd te verliezen.
Naast dat de klassificeerder kan worden aangevuld met een groter plaatsnamen bestand ten behoeve van de recall en precisie, zal ook een straatnamenlijst van Nederlandse straatnamen de klassificeerder verbeteren. Hier aan zou kunnen worden toegevoegd een lijst van provincienamen en landstreken. Tijdens de evaluatie van de klassificeerder ben ik enkele malen tegen lokatieomschrijvinen aangelopen waar ik de landstreken 'Walcheren' en 'Noord-Beveland' tegenkwam. Google Maps kan ook goed met dit soort landstreken overweg.
Verder denk ik dat het wel verstandig is om rekening met de oorsprong van een beeldbank te houden om de zoekboom te verkleinen, waarmee ik niks meer wil zeggen dat het niet slim is om alle straatnamen van Amsterdam te evalueren wanneer men de records uit de beeldbank van Leeuwarden wilt verrijken.

Entiteit
De standaardisatie van de named entities heeft nog niet zo'n effect als bijvoorbeeld het normaliseren van de datums en de lokaties. De normalisatie van de datums en de lokaties levert ons direct nieuwe zoekmogelijkheden. Dankzij de geocodes kan direct de geografische afstand tussen foto's bepaald worden. Ook de temporele afstand tussen foto's is nu te bepalen. Voor entiteiten bestaat er niet zo'n maat. Er bestaat niet zoiets als een afstand tussen namen in de semantische zin des woords[15]. De categorie entiteiten biedt een ander doel. Zoals al eerder besproken wordt het door standaardisatie van named entities mogelijk dezelfde foto's te vinden terwijl verschillende veel voorkomende schrijfwijzen van deze entiteit worden gebruikt. Hiervoor moet wel een systeem zijn wat deze verschillende schrijfwijzen mapt naar de gestandaardiseerde naam. Dit is nog niet geïmplementeerd.

Computer vision
In het computer vision domein is het mogelijk impliciete informatie op het niveau van beelden expliciet te maken. Dit zou een mooie aanvulling kunnen zijn op de data. Verder zal ook deze informatie mogelijk een rol kunnen vervullen bij het verrijken.

[15] *Ik doel hier op het levensthein algoritme.*

6 Literatuurlijst

1. *Data Integration: The Teenage Years.* **Halevy, Alon, Rajaraman, Anand en Ordille, Joann.** 2006.
2. *Answering Queries Using Views: A Survey.* **Levy, Alon.**
3. *Named Entity Normalization in User Generated Content.* **Marx, Maarten, et al.** 2008.

7 Apendix

7.1 Codefragmenten

binnen de code fragmenten is gebruik gemaakt van syntax highlighting. gereserveerde woorden zijn rood, strings zijn blauw en commentaar is groen.

7.1.1 Scraper

```
database = {}
for p=1; p <= aantal_paginas; p++:
        url  = 'http://server.domein.nl/?zoekopdracht=alles&pagina_index=' + p
        html = laadPagina(url)
        dom  = parseerHtml(html)
        meta_data_tabellen = dom.findAll('html_tag', 'html_attribuut'='meta_data')        foreach
metadata_tabel in meta_data_tabellen:
                # record volgens globaal schema
                nieuw_db_record = {
                        eigenschap_1:'',
                        eigenschap_2:'',

                        ...
                        eigenschap_n:'',
                }
                foreach tabel_rij in metadata_tabel:
                        if tabel_rij.label == 'eigenschap_1':
                                nieuw_db_record.eigenschap_1 = tabel_rij.waarde
                        if tabel_rij.label == 'eigenschap_2':
                                nieuw_db_record.eigenschap_2 = tabel_rij.waarde

                        ...
                        if tabel_rij.label == 'eigenschap_n':
                                nieuw_db_record.eigenschap_n = tabel_rij.waarde
                database.append(nieuw_db_record)
        wachtEenSeconde()
schrijfNaarBestand(database, '/map/bestandsnaam.extensie')
```

Pseudocode 1: Dit is het skelet van de scrapers. In principe zijn alle data op deze wijze opgehaald en naar een standaard formaat omgezet. Uiteraard waren de data per website niet altijd zo mooi gestructureerd in een label-waarde-paar.

7.1.2 Globaal schema

```
'http://www.locatievanfoto.nl/?id=1:{
        'time':{…}
        'location':{…}
        'entity':{…}
        'description':'…'
        'thumb':'http://www.locatievanfoto.nl/?id=1&thumb=t'
}
'http://www.locatievanfoto.nl/?id=2:{…}
…
'http://www.locatievanfoto.nl/?id=n':{…}
```

code fragment 5: In dit fragment ziet u hoe de database gestructureerd is. Het heeft als labels telkens de id's van de foto's welk ook tevens het webadres zijn van waar de foto is opgehaald. Hier is een voorbeeld van zo'n adres gegeven maar dit adres kan alle vormen aannemen die men op het web ook tegen komt. Verder is er een opsplitsing gemaakt voor de drie aspecten tijd plaats en entiet, die hier in het engels zijn. Ook is er een label waaronder de originele foto-omschrijving wordt opgeslagen, dat is het label 'description'. Als laatste is daar de referentie naar de thumbnail, een klein voorbeeldje van de afbeelding.

7.2 Correspondenties

Hieronder vindt u de correspondenties met enkele beeldbanken. Als reacite op onderstaande mail.

Geachte meneer/mevrouw,
Ik doe onderzoek naar de kwaliteit van Nederlandse beeldbanken op internet. Ik wil graag weten waarop u uw meta-gegevens heeft gebaseerd. Zijn de beschrijvingen en andere data vlak voor het stichten van de beeldbank "verzonnen", of zijn ze bijvoorbeeld gebaseerd op archiefmateriaal? Met andere woorden, wat is de kwaliteit van de gegevens?
Bij voorbaat dank voor uw reactie.

Zeeland

De metadata bij de afbeeldingen in de Beeldbank Zeeland is gebaseerd op een drietal pijlers. Hieronder volgen ze in de volgorde zoals wij ermee omgaan:

1. De eventuele informatie / gegevens die bekend zijn bij schenking / aankoop. Dikwijls staat er op de achterzijde van een foto wat informatie. Wanneer we een schenking aannemen willen we van de schenkers altijd weten wat er op de afbeeldingen te zien is. Zonder metadatering kunnen we vrij weining beginnen.

2. De kennis voor zover deze aanwezig is binnen het Zeeuws Documentatiecentrum. Degenen die de afbeeldingen invoeren en registreren, komen allemaal hier uit de regio, dus er is altijd een basiskennis aanwezig. Verder kan informatie worden ingewonnen bij vakspecialisten, literatuur en internet.

3. Bij veel beeldbanken is het mogelijk om reacties op foto's te kunnen geven. Soms dat mensen van buitenaf het nodige weten van hetgeen wat op de afbeelding zichtbaar is. Ze geven die informatie door middels de reacties. In ons geval zijn het veelal mensen op leeftijd die geïnteresseerd zijn in de lokale geschiedenis en thuis een heel documentatiecentrum hebben. Soms krijgen we zeer gedetailleerde informatie binnen over b.v. bewoners van een pand, functie, etc. Want niet alle informatie is exact op te zoeken. Dan is deze informatie zeer kostbaar.

Utrechts Archief

De beschrijvingen bij onze afbeeldingen zijn bij een gedeelte van de collectie gebaseerd op oude beschrijvingen die nog uit een papieren catalogus stammen (en die stuk voor stuk een redactionele slag ondergaan en waarnodig gecorrigeerd/aangevuld worden) en voor een gedeelte betreft het “nieuwe” foto’s, waarbij beschrijvingen gemaakt worden door specialisten die zich baseren op gegevens die op de foto’s te vinden zijn, gegevens die uit het beeld zijn af te leiden of gegevens die gebaseerd zijn op archief- en literatuuronderzoek.
Daarnaast zijn we voortdurend bezig om oude beschrijvingen in de beeldbank aan te vullen of te corrigeren
Als u nog meer vragen hebt, dan bent u uiteraard welkom deze te stellen.

Groningen

Het overgrote deel van de foto’s (althans, van de foto’s uit de collectie van de Groninger Archieven) was al beschreven lang voor dat het begrip ‘beeldbank’ was uitgevonden. Zelfs het begrip metadata zal de vroegere beschrijvers onbekend zijn geweest.
De beeldbank die wij gebruiken (ontwikkeld door Pictura) bevat standaard een aantal metadata die gebaseerd zijn op de Dublin Core. Deze komen min of meer overeen met de metadata van de ISBD. De ‘oude’ beschrijvingen waren gemaakt op basis van de ISBD.
De inhoud van de beschrijvingen is in het algemeen gesproken gebaseerd op de gegevens die bij de foto’s zijn meegekomen (mondeling of schriftelijk of aantekeningen op de achterkant van de foto) en als die ontbraken, zijn ze gebaseerd op de kennis van de beschrijver.