CLARIAH-project Digital Film Listings

Reconstructie van oude filmladders

Vroeger stond in de krant waar en wanneer bioscoopfilms draaiden. Onlangs is de informatie uit deze zogenoemde Filmladders uit kranten van 1948 tot 1994 digitaal gereconstrueerd.
Kaspar Beelen

Het CLARIAH-project Digital Film Listings (DIGIFIL) spitst zich toe op het automatisch extraheren, digitaliseren en publiceren van de informatie uit Filmladders. Het project demonstreert hoe machinegetranscribeerde tekst, geproduceerd met behulp van *Optical Character Recognition* (OCR), op een automatische manier kan worden geconverteerd naar verrijkte en gestructureerde data, voorzien van semantische annotaties.

Verrijkt en gelinkt

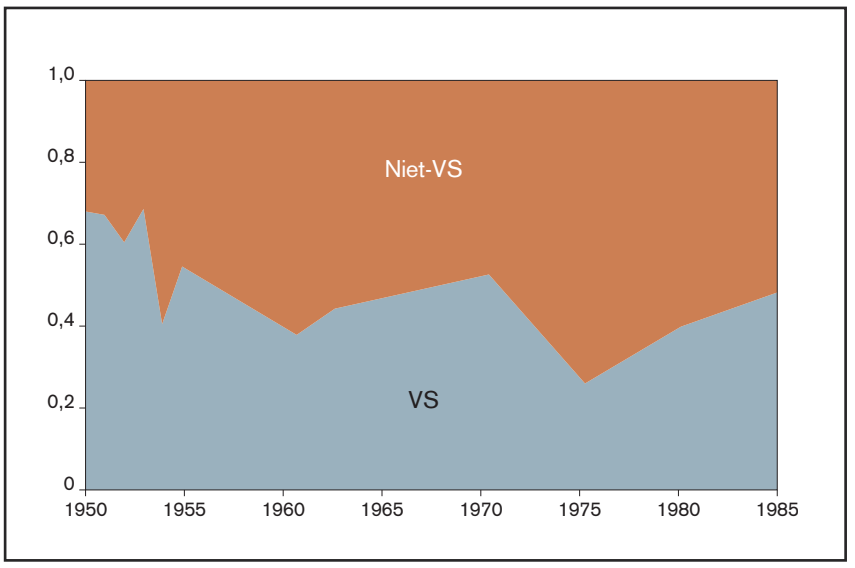
De eerste fase bestond uit een klassiek 'naald in een hooiberg'-probleem: de Filmladders vissen uit de gigantische stapel artikelen in de Delpher-krantendatabank. Vervolgens maakte *machine learning* de impliciete structuur van de ladders expliciet: elk woord in de ladders werd op basis van de context voorzien van een label ('titel', 'tijdstip', 'bioscoop'). De geïdentificeerde filmtitels werden daarna gekoppeld aan externe databanken, zoals de Internet Movie Database (IMDb). Met de verrijkte en gelinkte DIGIFIL-data is het mogelijk om ontwikkelingen in het naoorlogse filmlandschap kwantitatief te onderzoeken, en de patronen te vergelijken met kwalitatief onderzoek (zie afbeelding).

Grotere ambities

DIGIFIL richt zich op de extractie van één type micro-evenementen, namelijk filmvertoningen, maar de onderzoekers koesteren grotere ambities. De tools kunnen misschien ook worden toegepast voor de extractie van andere soorten informatie, zoals theateragenda's of scheepsberichten. In die zin poogt DIGIFIL om alledaagse geschiedenis en digitale geesteswetenschappen met elkaar te verbinden.

De onderzoekers binnen het DIGIFIL-project zijn: Kaspar Beelen, Ivan Kisjes, Thunnis van Oort, Kathleen Lotze en Julia Noordegraaf.

Op Gitlab staan diverse scripts: ***gitlab.com/uvacreate/digifil/***



De afbeelding toont het aandeel Hollywoodfilms (blauw) versus niet-in-Amerika-geproduceerde films (bruin) in de periode 1948 tot 1994 in Nederlandse bioscopen. De data laten een geleidelijke afname zien van het marktaandeel van Amerikaanse films, gedefinieerd in termen van het aantal vertoningen. Dit is in strijd met de bevindingen van Bart Hofstede, die in zijn proefschrift over de mondiale positie van de Nederlandse film uit 2000 een andere dynamiek waarnam: een relatieve daling van het Amerikaanse marktaandeel in de jaren 1960 en 1970, gevolgd door een sterke stijging in de jaren 1980 tot een aandeel van meer dan 80%. Of er werkelijk sprake is van een nieuw inzicht zal vervolgonderzoek moeten uitwijzen.

UITGAAN IN ROTTERDAM

BIOSCOPEN. Arena: Lady Godiva ...Pin-up girl, 18 j.; Capitol: Messalina, de keizerin der zonde, 18 j.; Cineac (dag.) Actualiteiten; Cineac (avond) en Victoria: Tramlijn „Begeerte", 18 j.; Colosseum: Ivoorjagers, 14 j.; Harmonie: Lach en vergeet, a.l.; Lutusca: Teresa, 14 j.; Luxor: Grens incident, 18 j.; Passage: Vr. t/m Zond.: Vrouwen en opium; Ma. t/m Do.: Hoogverraad, 18 j.; Prinses: De Jantjes. 14 j.; Rex: Om het goud van Nevada, 14 j.; 't Venster: Les Enfants du paradis. 18 j.

Schouwburg: Rott. Philh. Orkest, 12de concert, serie A, he. 8 u.; Bonaventura. mo. en Zo. 8 uur. Aula Museum Boymans: Solistenconcert, m.m.v. Gerard de Beer. José Stordiau, mo. 8 u.

Voorbeeld van een filmladder, Algemeen Handelsblad, 14-03-1952
Credits: Delpher

GELEZEN

LCRDM positioning Paper 2019 en verder
LCRDM

De adviesgroep van het het Landelijke Coördinatiepunt Research Data Management (LCRDM) heeft een Positioning Paper 2019 en verder opgesteld. In dit paper wordt - aan de hand van een thematische prioritering en drie bredere (beleidsmatige) werkgebieden - beschreven wat wel en niet binnen de scope van het LCRDM valt.
Belangrijk is dat sinds 2018 het LCRDM werkt met een pool van experts. Deze pool is inmiddels uitgegroeid tot ruim 190 deelnemers uit 60 Nederlandse onderzoeksinstellingen.
De experts werken in taakgroepen aan diverse aspecten van Research Data Management uit de praktijk van Nederlandse onderzoeksinstellingen. De samenwerking in de taakgroepen plaatst de activiteiten binnen de instellingen in een breder landelijk perspectief. Dit zorgt voor samenhang, herkenbaarheid en onderbouwing.
De paper en resultaten van werkgroepen staan op de LCRDM-website.

LCRDM.nl

COLUMN

Micro-data voor macro-onderzoek

In 2011 verhuisde ik na een aantal jaar werken in het Verenigd Koninkrijk (VK) terug naar Nederland. Voor mijn onderzoek in het VK maakte ik veel gebruik van de fantastische Britse longitudinale survey data en de longitudinale data van de volkstellingen. In het VK is er nog iedere tien jaar een echte volkstelling en voor een kleine steekproef worden de individuele bestanden aan elkaar gekoppeld. Daardoor is het mogelijk om mensen door de tijd te volgen met iedere tien jaar een meetopname van het leven van mensen; niet ideaal, maar als je vier volkstellingen aan elkaar koppelt, kun je mensen toch al dertig jaar volgen.

Eenmaal terug in Nederland stuitte ik op de microdata van het Centraal Bureau voor de Statistiek. De Britse data was al mooi, maar de Nederlandse microdata is het ware paradijs voor sociaalwetenschappers: individuele longitudinale data van de hele Nederlandse bevolking vanaf eind jaren '90; geen steekproef, maar gewoon iedereen. De Nederlandse microdata is zeer vergelijkbaar met wat er beschikbaar is voor onderzoek in Zweden en Finland waardoor (in theorie) vergelijkbaar onderzoek mogelijk is.

Microdata is registerdata op basis van de Basisregistratie Personen. Een uniek kenmerk van deze data is de mogelijkheid om surveydata te koppelen aan de CBS registerdata. Bijvoorbeeld: informatie uit een survey over zoekgedrag naar werk kan worden gekoppeld aan de microdata met informatie of mensen werk hebben gevonden, ook jaren na het afnemen van de survey. Het slim linken van survey data aan registerdata heeft enorme potentie voor onderzoek.
Met de enorme schat aan gegevens op individueel niveau kunnen onderzoekers onder zeer strikte voorwaarden zelf onderzoek doen. Hiermee ontstaan ook nieuwe uitdagingen, zoals het waarborgen van privacy versus de enorme maatschappelijke baten van onderzoek.
De schat aan data stelt ook nieuwe eisen aan de hardware en software die sociaalwetenschappers nodig hebben om hun onderzoek te kunnen doen. Waar het gebruik van een supercomputer het domein was van klimaatonderzoekers en de kwantummechanica, ook sociaalwetenschappers hebben in toenemende mate grote rekenkracht nodig.

Eerder dit jaar had ik de eer om samen met een van mijn promovendi binnen een pilot van de ODISSEI Data Facility te werken met de high-performance computer van SURFsara. Deze supercomputer heeft enorme potentie voor de sociale wetenschappen. Ik hoop dan ook zeer dat er financiering beschikbaar komt voor het verder ontwikkelen en opschalen van ODISSEI (Open Data Infrastructure for Social Science and Economic Innovations) waardoor Nederland zich kan ontwikkelen tot de absolute wereldtop qua data infrastructuur voor de sociale wetenschappen.

Maarten van Ham
Maarten geeft de volgende column graag aan Sander Steijn van het Sociaal en Cultureel Planbureau.

Maarten van Ham is hoogleraar stadsgeografie aan de Faculteit Bouwkunde van de Technische Universiteit Delft. Hij promoveerde als geograaf in Utrecht en werd in 2011 hoogleraar aan de universiteit van St Andrews in het Verenigd Koninkrijk. In 2014 ontving hij een European Research Council (ERC) subsidie voor onderzoek naar de oorzaken en gevolgen van ruimtelijke ongelijkheid in Nederland, Zweden, Estland en het Verenigd Koninkrijk.