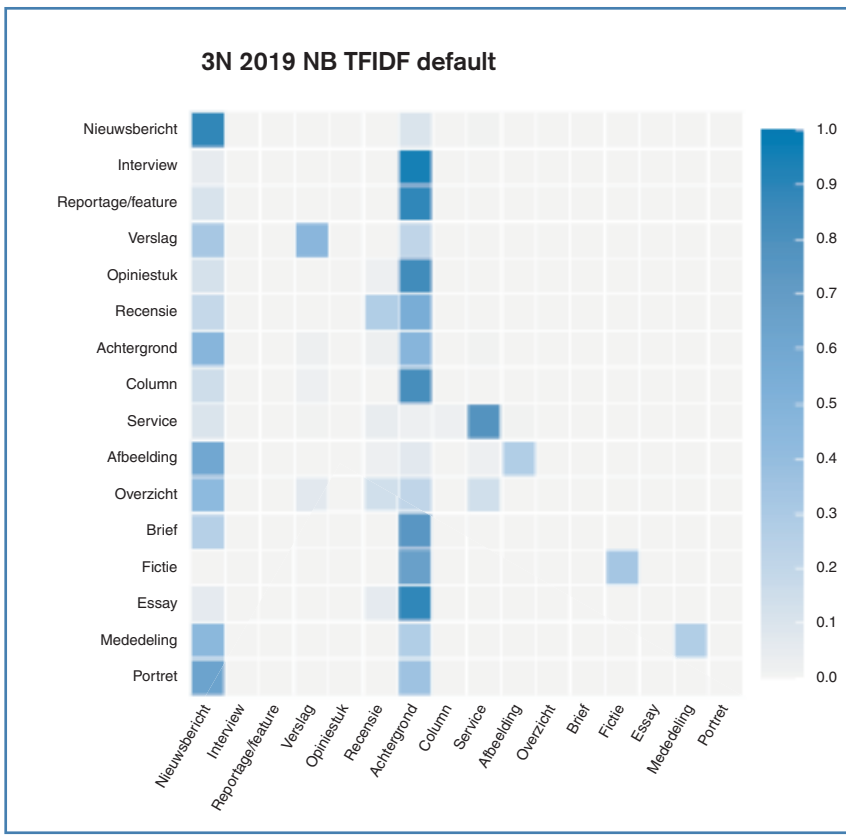
Platform voor transparantie in machine learning

NEWSGAC biedt een kijkje onder de motorkap

Met digitale methoden kan op grote schaal onderzocht worden hoe journalistiek zich in de twintigste eeuw heeft ontwikkeld. Het beste algoritme daarvoor kiezen, is echter een uitdaging.

Kim Smeenk



Deze confusion matrix heeft op de verticale as het daadwerkelijke genre en op de horizontale as het voorspelde genre. De visualisatie laat zien hoe goed het algoritme het genre van een groep artikelen heeft voorspeld en tussen welke genres verwarring is opgetreden. *Credits: NEWSGACS*

Digitale krantenarchieven maken het tegenwoordig mogelijk om op grote schaal onderzoek te doen naar ontwikkelingen in verslaggeving. Het NEWSGAC-project brengt journalistieke ontwikkelingen in beeld door de genres van krantenartikelen te classificeren. Een genre is bijvoorbeeld 'reportage', waarin verslaglegging op locatie belangrijk is. Een ander voorbeeld is 'opiniërende essay', een genre met de nadruk op de mening van de schrijver. Het classificeren gebeurt met machine learning. Onderzocht wordt of de journalistiek in de twintigste eeuw inderdaad steeds minder politiek en opiniërend werd, zoals vaak wordt aangenomen.

NEWSGAC wil niet alleen deze inhoudelijke vraag beantwoorden, maar ook een platform voor transparantie in machine learning bieden. Automatisch genres toekennen aan krantenartikelen is echter geen vanzelfsprekende klus. Tijdens de eerste experimenten bleken twee zaken van essentieel belang: algoritmes moeten met elkaar vergeleken kunnen worden en de onderliggende opties en aannames moeten inzichtelijk zijn. NEWSGAC maakt via visualisaties inzichtelijk hoe algoritmes kiezen, welke 'features' belangrijk zijn bij de classificatie van een artikel.

Onderwerp of genre?

Deze aanpak blijkt succesvol. Niet alleen kan de nauwkeurigheid van algoritmes vergeleken worden, ook de vraag of ze hun keuzes maken op de juiste gronden kan worden beantwoord. De meeste algoritmes blijken bijvoorbeeld vooral goed in het onderscheiden van onderwerpen en minder in het onderscheiden van genres, terwijl juist de levensloop van genres centraal staat in het project. Doordat NEWSGAC dit proces inzichtelijk maakt, kon het beste algoritme gekozen worden. Dat algoritme classificeert nu ook miljoenen artikelen.

ADAH-project

NEWSGAC is een van de vier ADAH-projecten (Accelerating Scientific Discovery in the Arts and Humanities) van CLARIAH en het eScience Center. Andere partners zijn: Rijksuniversiteit Groningen (RUG), CWI, KB en Beeld & Geluid. Hoofdonderzoeker is prof. Marcel Broersma (RUG). Binnenkort wordt NEWSGAC geïntegreerd in de CLARIAH-infrastructuur zodat ook andere onderzoekers met andere vragen en datasets er gebruik van kunnen maken.

esciencecenter.nl/project/newsgac

GELEZEN

Open a GLAM Lab
Digital Cultural Heritage Innovation Labs,
Book Sprint, Doha, Qatar,
23-27 September, 2019

Lotte Wilms, coördinator van het KB Lab van de Koninklijke Bibliotheek, deed eind vorig jaar in Qatar mee met een 'Book sprint'. In vijf dagen schreven 15 experts gezamenlijk een boek over het opzetten van een GLAM Lab. Een GLAM Lab is een innovation lab voor Galleries, Libraries, Archives and Museums. Naast de voordelen van een GLAM Lab behandelt het boek ook tips en manieren om een eigen GLAM Lab op te zetten met behulp van key points en case studies. Ook het KB Lab krijgt aandacht in het boek. In het KB Lab (lab.kn.nl) staan experimentele digitale tools en data gemaakt van, voor en met de digitale collecties van de Koninklijke Bibliotheek.
Het boek 'Open a GLAM Lab' is gratis te downloaden in .pdf en .epub bestand, via glam-labs.io.
hdl.handle.net/10576/12115

COLUMN

Kijkje in de keuken

Soms kan "weten hoe de worst gemaakt wordt" alle eetlust bederven. Zo heeft één middagje bijverdienen in de Pizza Hut op Amsterdam Centraal 15 jaar geleden me voorgoed genezen van elke behoefte om een pizza-slice te bestellen bij een fastfoodrestaurant. Geldt dat eigenlijk ook voor surveys? Zouden de bevindingen van het Sociaal en Cultureel Planbureau met nog meer smaak door het publiek worden verorberd als men een beter zicht had op wat er allemaal komt kijken bij het uitvoeren van een survey? Die vraag kwam afgelopen zomer bij mij op tijdens het keynotespektakel van mijn voormalige baas Ineke Stoop op het methodologiecongres ESRA. In vliegende vaart schetste Ineke de best practices in survey-onderzoek. Wie dacht na deze spoedcursus survey-onderzoek in 2 minuten – overigens beloond met een ovationeel applaus – weer naar huis te kunnen, verkeek zich. Want kloppen die best practices wel? En als ze kloppen, waarom worden ze dan niet altijd toegepast? Als we niet weten of ze kloppen, hoe komen we daar dan achter? Als ze soms kloppen, wanneer dan en waarom juist dan? En als ze niet kloppen, wat kunnen we dan wel doen?

Een grondige weging van de best practices onthulde dat survey-onderzoekers zich met talloze uitruilen en dilemma's geconfronteerd zien. Zo kunnen goede face-to-face interviewers mensen overtuigen om mee te doen aan onderzoek, aanmoedigen om (privé-)informatie te delen en respondenten helpen met complexe vragenlijsten. Maar ze kunnen ook in de verleiding komen respondenten door het interview heen te jagen, vragen ongewenst te versimpelen of -in het ergste geval- data zelf te verzinnen. Valt dergelijk ongewenst interviewergedrag te voorkomen door met ervaren interviewers te werken? Misschien, maar staan die ervaren rotten ook open voor specifieke instructies of nieuwe inzichten? Ook zij kunnen en moeten bijleren. En hiermee heb ik slechts twee van de negentien best practices besproken.

Ineke's synthese van decennia aan theoretische kennis over- en praktische ervaring met survey-onderzoek bevatte gouden raad voor iedereen die ooit overweegt een survey uit te voeren, én vormde een belangrijke onderzoeksagenda voor iedereen die eraan wil bijdragen dat surveys in de toekomst nog beter worden uitgevoerd.

Ongetwijfeld zal de rol van het survey binnen de sociale wetenschap door de toenemende beschikbaarheid van andere databronnen (big en organic data, maar ook kwalitatief onderzoek) in de komende jaren blijven veranderen. Maar surveys blijven nodig om te begrijpen hoe burgers de wereld om hen heen en hun eigen rol in die wereld ervaren en waarderen. Zolang we de keuzes en overwegingen bij het ontwerpen en uitvoeren van die surveys net zo consciëntieus en weloverwogen blijven maken als Ineke deed, durf ik iedereen een kijkje in de keuken te gunnen en de resulterende datamaaltijd met trots te serveren.

Sander Steijn

Sander Steijn is methodoloog bij het Sociaal en Cultureel Planbureau (SCP). Op het SCP adviseert hij in alle fases -dataverzameling, analyse en rapportage- van het (kwantitatieve) onderzoeksproces en werkt aan de langdurige (survey-)data infrastructuur van het SCP. Daarnaast is hij lid van het Core Scientific Team van het European Social Survey en namens de planbureaus en onderzoeksinstellingen lid van het Supervisory Board van ODISSEI.

Sander geeft de pen door aan Marieke Houben-van Herten, statistisch onderzoeker en projectleider bij het CBS.