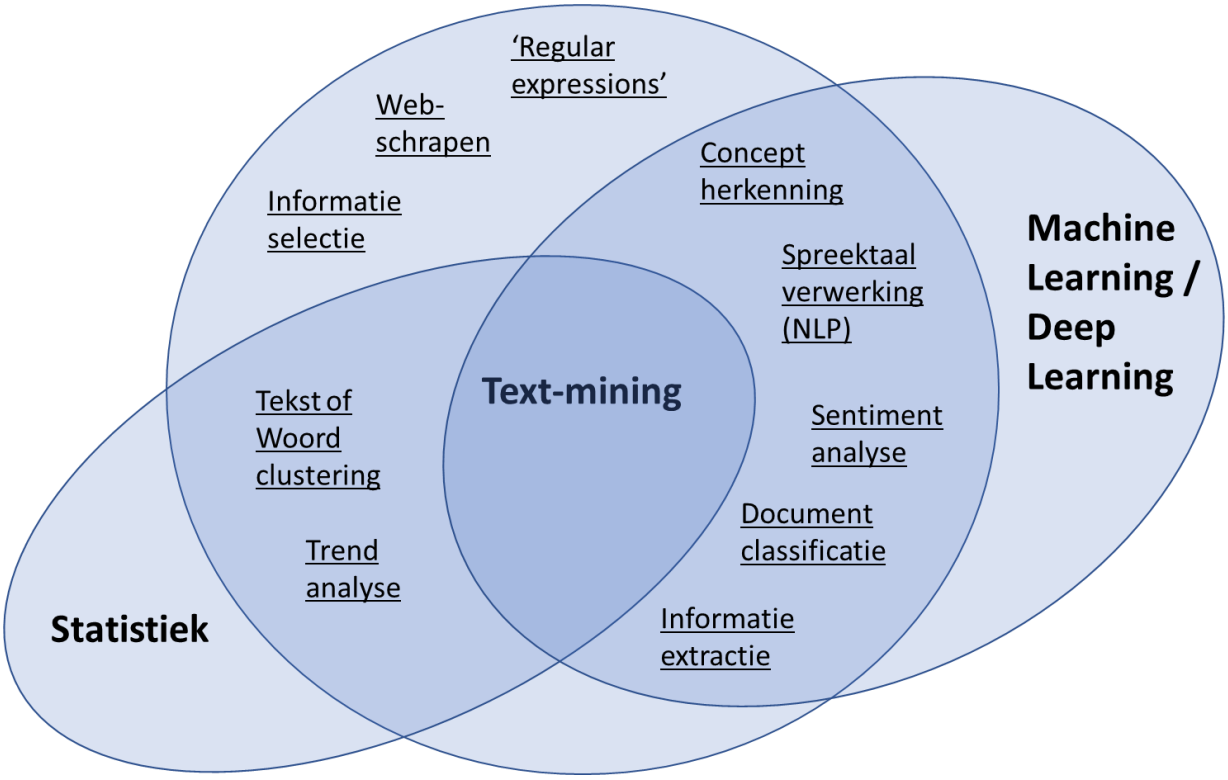
Text-mining voor de watersector

Het automatisch doorzoeken van (grote) hoeveelheden tekstuele informatie en de gevonden informatie op een gestructureerde manier bij elkaar brengen is mogelijk met text-mining. Er zijn vele technieken en toepassingen denkbaar. Dit artikel schetst technieken en toepassingen waar de watersector van kan profiteren.

Tessa Pronk, Nienke Meekel, Xin Tian, Sotirios Paraskevopoulos (KWR water research institute)

Net als in elk ander vakgebied wordt in de watersector sommige informatie vastgelegd in de vorm van tekst. Denk daarbij aan rapporten, vergunningen, memo's, nieuwsbrieven, e-mails, wetenschappelijke publicaties, websites, maar ook uitingen op sociale media en schriftelijke klant-interacties. Het lezen en verwerken van dit soort teksten is tijdrovend. De techniek text-mining wordt gebruikt voor het automatisch doorzoeken en verwerken van teksten. In potentie kan deze techniek gewenste informatie uit teksten halen en netjes op een rij zetten, zonder dat de (gehele) tekst door een persoon gelezen hoeft te worden. Dit kan veel tijdwinst opleveren. Een ander voordeel is dat zeer grote verzamelingen teksten geanalyseerd kunnen worden op informatie en trends, ook bijvoorbeeld in een historisch perspectief. Omdat de informatie volgens een gestandaardiseerde manier door de computer wordt geanalyseerd zijn de resultaten ook reproduceerbaar en objectief. In 2021 is deze techniek toegepast in het bedrijfstakonderzoek (BTO) van de gezamenlijke Nederlandse drinkwaterbedrijven en heeft KWR Water een aantal voorbeelden uitgewerkt. Deze voorbeelden lopen via twee sporen: communicatie met klanten en het vergaren van informatie voor bijvoorbeeld operationele, monitoring- of onderzoeksdoeleinden.



Afbeelding 1. Overlap van verschillende technieken met 'text-mining'. Dikgedrukt zijn algemene technieken. Onderstreept zijn diverse specifieke text-mining technieken, of taken

Technieken binnen text-mining

De mogelijkheden om text-mining toe te passen zijn divers. In afbeelding 1 is een aantal technieken of taken binnen het vakgebied van text-mining gegeven. Deze maken op hun beurt weer gebruik van statistiek, 'machine learning' en, meer recent, 'deep learning'. Hieronder worden deze technieken of taken nader toegelicht.

'Web-schrapen'

Een eerste toepassing is het automatisch binnenhalen van informatie van websites ('web-scraping'). Veel websites hebben weliswaar een zoekfunctie waarmee de gebruiker naar de goede pagina geleid kan worden, maar er is meer mogelijk. Door de teksten van pagina's ook te downloaden ('schrapen') kunnen er diverse analyses op de teksten worden uitgevoerd, inclusief op combinaties van teksten. Hierdoor wordt het mogelijk overkoepelende analyses uit te voeren, zoals: welke informatie komt voor op websites van chemische bedrijven? Wat is de informatie over verwijderingsrendementen op de webpagina's van drinkwaterbedrijven? Ook documenten of datasets op websites kunnen automatisch geïdentificeerd worden en met een enkel commando gedownload. Dat scheelt in sommige gevallen veelvuldig handmatig klikken en downloaden. In het BTO is gekeken of het mogelijk is door websites te 'springen' van pagina naar pagina en of op deze manier de tekst van de website in beeld kan komen. Dit bleek eenvoudig te realiseren, al is dit een tijdrovend proces omdat het aantal op te volgen links exponentieel stijgt bij het 'springen' door pagina's.

Een concrete uitgewerkte casus binnen het BTO is het combineren van stofnamen in een tabel met toegestane stoffen op de website van het Europese Chemicaliënagentschap (ECHA) met gegevens over toelating op de website, in losse factsheets per stof. De informatie in de factsheets is nodig om te zien of een stof voor een nieuwe toepassing ingezet mag worden. Door de tabel te combineren met de factsheets op de website, kon worden achterhaald welke mogelijke nieuwe soort emissie van deze stof te verwachten is bij deze toepassing.

Selectie van relevante documenten of tekst

Het kan voorkomen dat slechts enkele documenten van een verzameling documenten relevante informatie bevatten. Deze documenten kunnen ofwel op een computer staan of op een website (zie 'web-schrapen'). Om te voorkomen dat de documenten allemaal gelezen moeten worden kan er geselecteerd worden op kernwoorden die aanwezig zijn in de tekst van het document. Dit doel heet in text-mining 'information retrieval', oftewel 'informatieselectie', en staat voor het verzamelen van letterlijke (stukken) informatie. Nog verder inzoomen kan door in geselecteerde documenten alleen de alinea's met de kernwoorden te selecteren en samen te brengen voor het doorlezen. Hiermee kan de hoeveelheid te lezen tekst gereduceerd worden.

Met 'regular expressions', afgekort 'regex', kunnen specifieke tekstpatronen teruggevonden worden. Bijvoorbeeld postcodes (vier cijfers, twee letters), mailadressen (letters/cijfers, gevolgd door een '@', letters, een '.' en dan weer letters). Met regex kunnen kernwoorden dus in plaats van letterlijke woorden ook uit patronen bestaan, zoals een identificatienummer met een vast format.

Een voorbeeld van toepassing van information retrieval zijn vergunningen. Deze kunnen centraal beschikbaar zijn en informatie bevatten over (nieuwe) emissies of aanpassingen aan bijvoorbeeld

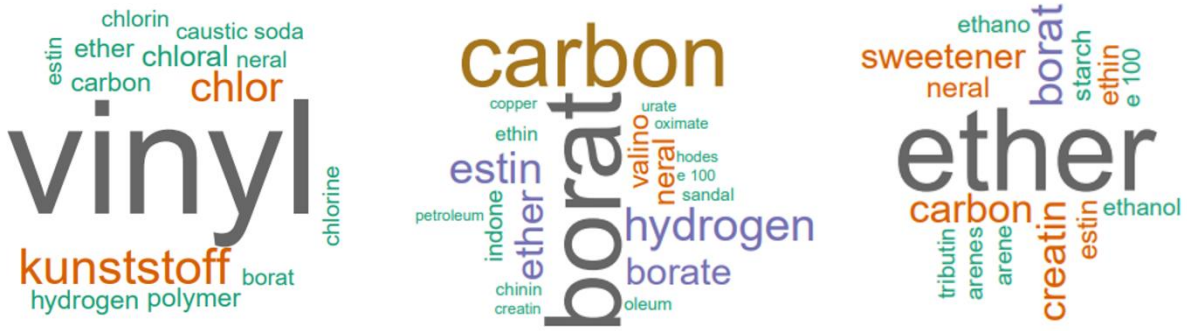
afvalwaterzuiveringen. In het BTO kon de hoeveelheid te lezen informatie over vergunningen die beschikbaar waren via het zoeken met relevante woorden teruggebracht worden naar een aantal zinnen per document. In het project 'PS-drink' zijn met deze techniek in wetenschappelijke artikelen 200 stoffen achterhaald die een nieuwe bedreiging voor de waterkwaliteit kunnen zijn. Het voorkomen van drie daarvan in oppervlaktewater wordt nu onderzocht [1].

Conceptherkenning

Voor sommige toepassingen is het nodig om 'concepten' in teksten te herkennen. Concepten zijn containerbegrippen, waarbinnen weer specifieke begrippen vallen. Met conceptherkenning kan de computer bijvoorbeeld herkennen dat een woord in een tekst een bacterie betreft of een chemische stof, ook al is de naam iedere keer anders. Ook kan conceptherkenning bijvoorbeeld achterhalen dat een tekst over waterzuivering gaat, ook al wordt dat woord zelf niet genoemd. Dit heet in dit geval 'documentclassificatie'. Zulke documentclassificatie kan ook gaan over beslissingen over de bruikbaarheid van het document. Een voorbeeld van het gebruik van conceptherkenning uit het BTO is het verzamelen van informatie over chemicaliën uit teksten, zoals de afbraaksnelheid of welke afbraakproducten er gevormd kunnen worden. Een ander voorbeeld is het verwerken van klantenberichten. Conceptherkenning kan worden gebruikt om gemelde problemen (bijvoorbeeld lekkage of een probleem met de waterkwaliteit) te detecteren [2]. Dit kan ervoor zorgen dat een bericht automatisch en snel bij de juiste persoon terecht komt.

Clustering van concepten of woorden

In een tekst staan veel woorden in samenhang bij elkaar. Het is mogelijk om concepten aan elkaar te verbinden door te kijken of ze in verschillende teksten vaak bij elkaar staan. Zo kunnen groepjes van gerelateerde concepten gevonden worden, of groepjes van woorden binnen concepten, zoals groepjes bacteriën. Het is op deze manier bijvoorbeeld mogelijk om kandidaatstoffen te vinden die mogelijk samen met een stof van interesse uitgestoten zouden kunnen worden. Tekstbronnen om associaties uit te extraheren kunnen een diverse oorsprong hebben, van wetenschappelijke artikelen tot rapporten, Wikipedia-teksten (zie 'web-schrapen') of twitterberichten, zoals in afbeelding 2.



Afbeelding 2. Drie voorbeelden van chemicaliën-signaturen uit Duits- en Engelstalige twitterberichten van/over bedrijven met afvalwateremissies in het Rijnstroomgebied

Spreektaalverwerking

Natural Language Processing (NLP) maakt gebruik van technieken die op hun beurt gebruik maken van het interpreteren van tekst in spreektaal. Technieken binnen NLP zijn in staat om functies aan

worden toe te kennen, zoals ‘werkwoord’, ‘bijvoeglijk naamwoord’, ‘voorzetsel’, et cetera. Mocht een woord zowel een werkwoord als een zelfstandig naamwoord kunnen zijn, zoals bij het woord ‘zagen’, dan kan NLP het onderscheid maken op basis van de context. Ook kan NLP afhankelijkheden tussen woorden toekennen. Zo is het mogelijk vast te stellen welk zelfstandig naamwoord het onderwerp is van een werkwoord, of bij welk zelfstandig naamwoord een bijvoeglijk naamwoord hoort. Hierdoor is het mogelijk gerelateerde tekst uit een zin te selecteren, zoals een combinatie van woorden met labels onderwerp-werkwoord-lijdend voorwerp (een ‘triplet’). Voorbeelden: ‘bacterie x’ ‘degradeert’ ‘stof z’; ‘zuiveringstechniek x’ ‘verwijdert’ ‘stof y’.
Recentelijk zijn ‘Deep Learning’-technieken in opkomst in het NLP-werkveld. Dit zijn krachtige modellen die woordverbanden nog beter kunnen detecteren dan de traditionele NLP. Gebaseerd op een collectie gelabelde voorbeelden, kan Deep Learning zulke onderwerp-werkwoord-lijdend voorwerp triplets goed detecteren. Een prominent voorbeeld van een Deep Learning-techniek is het ‘Bidirectional Encoder Representations from Transformers’ (BERT)-model. Het BERT-model maakt effectief gebruik van context: woorden die aan een woordtype, woordfunctie of woordrelatie van interesse voorafgaan en erop volgen. Tegenwoordig behoren modellen die vergelijkbaar zijn met BERT tot de meest bruikbare tools die worden gebruikt om NLP-taken uit te voeren [3].

Extractie van informatie

Bij informatie-extractie wordt specifieke, gestructureerde informatie, inclusief relaties tussen woorden, uit grote aantallen tekstdocumenten geëxtraheerd. Dit gebeurt vaak met behulp van NLP-technieken. Dit in tegenstelling tot ‘Information Retrieval’ (zie ‘Selectie van relevante documenten’,), dat gaat over het teruggeven van letterlijke tekst die relevant is voor een specifieke zoekopdracht [4]. Informatie-extractie wordt veel gebruikt in de gezondheidszorg, waar experts proberen belangrijke informatie te extraheren uit vrije tekst, klinische notities en ongestructureerde voortgang-/
behandelingsnotities. Met informatie-extractie wordt direct het gezochte feit of verband achterhaald, getoetst op relevantie en in een bruikbare vorm weergegeven.

Voorbeelden van toepassingen

Sociale media doorzoeken op nieuws rond waterkwaliteit

Nieuws kan informatie bevatten over nieuwe activiteiten van bedrijven, die mogelijk de waterkwaliteit beïnvloeden. Twitter kan met functies in de statistische programmeertalen R of Python eenvoudig doorzocht worden op tweets met bepaalde trefwoorden, welke vervolgens gefilterd kunnen worden op locatie of de aanwezigheid van andere trefwoorden. In een van de projecten binnen het BTO zijn met behulp van een lange lijst van bedrijven met emissie in het Rijnstroomgebied, geassocieerde twitterberichten met ‘information retrieval’-technieken doorzocht op aanwijzingen van wijzigingen in productie(processen) voor deze bedrijven. Het is de moeite waard deze te achterhalen, omdat er mogelijk emissies mee gepaard gaan die eerder niet voorkwamen. Zo kwam aan het licht dat er een nieuwe cumeen (ook wel isopropylbenzeen genoemd)-fabriek gepland is in het Rijnstroomgebied en dat er investeringen ophanden zijn voor de productie van batterijen. In totaal werden er dertien nieuwe activiteiten aangekondigd in het Rijnstroomgebied tussen 2018 en 2020. Een andere toepassing voor sociale media is ‘sentiment analysis’. Een bericht geassocieerd met drinkwater op sociale media kan bijvoorbeeld geclassificeerd worden op positieve of negatieve

betekenis. Hiermee kan een waterbedrijf zien of er bij gebruikers problemen voordoen die mogelijk niet gemeld worden bij de klantenservice.

Chatrobot bij het verwerken van klantcontact

De meeste drinkwaterbedrijven moeten elk jaar een aanzienlijk aantal klachten van klanten afhandelen. Traditioneel worden klachten behandeld door bekwaam personeel dat weet hoe ze primaire problemen kunnen identificeren, klachten classificeren, oplossingen vinden en met klanten communiceren. De inspanning die gepaard gaat met de afhandeling van klachten is vaak groot, afhankelijk van het aantal klanten dat een bedrijf bedient. De opkomst van NLP, mogelijk gemaakt door Deep Learning, heeft echter nieuwe mogelijkheden gecreëerd voor het begrijpen en interpreteren van tekstklachten. Als zodanig hebben KWR en Waterbedrijf Groningen de waarde onderzocht van het gebruik van NLP voor het verwerken van klantenklachten. NLP kan taalstructuren ontleden en intenties en gevoelens extraheren uit klachten van klanten, speciaal gefocust op relevante onderwerpen voor de drinkwaterbedrijven. Deze studie vormt een stap in de richting van de volledige automatisering van de verwerking van consumentenklachten voor Waterbedrijven [2].

Verkrijgen van actuele informatie over een watergedragen ziekteverwekker

Als onderdeel van het Horizon 2020 PathoCERT-project heeft KWR water informatie-extractie ingezet in het domein van de omgevingsmicrobiologie. In dit geval ging het erom kenmerken van een pathogeen te vinden in wetenschappelijke publicaties. De geëxtraheerde informatie helpt experts om veel sneller belangrijke en actuele informatie over pathogeeneigenschappen te verkrijgen. De verzamelde informatie kan worden ingevuld in vooraf gedefinieerde sjablonen (‘template-filling’) en vervolgens worden gebruikt voor analyse- of beslissingsdoeleinden. Onderstaande tabel is een voorbeeld van een dergelijk proces waarbij informatie over de watergedragen ziekteverwekker legionella automatisch uit een wetenschappelijke publicatie wordt gehaald en in het sjabloon wordt ingevuld.

Tabel 1. Voorbeeld (vertaald uit het Engels) van ‘template filling’ gebruikmakend van informatie-extractie uit wetenschappelijke literatuur rond de watergedragen ziekteverwekker legionella

Informatie trefwoord	**Resultaat**
Soort	*Legionella pneumophila*
Incubatietijd (dagen)	2-14
Symptomen	Hoofdpijn, spierpijn, algehele zwakheid, gebrek aan eetlust, koorts, hoesten, rillingen, kortademigheid, gewrichtspijn
Transmissieroute	Inademing, microaspiratie, direct contact met operatiewonden
Milieuhabitat	Aquatische habitats, waterleidingsystemen
Klinische manifestatie	Veteranenziekte, atypische longontsteking, Pontiackoorts
Bron van blootstelling	Watervoorziening, besmettelijke aerosolen, koeltorens, bubbelbaden, potgrond

Horden voor het toepassen van text-mining

In de programmeertalen Python en R zijn functies te vinden die text-mining-taken kunnen uitvoeren. Hiervoor moet kennis aanwezig zijn van deze talen. Programmeervaardigheden kunnen daarom limiterend zijn voor het uitvoeren van text-mining-taken. Zowel voor Python als voor R bestaan gebruikersinterfaces die het makkelijker kunnen maken. Er bestaan ook kant-en-klare (commerciële) applicaties die een bepaalde taak kunnen uitvoeren.

Teksten vallen veelal onder copyright, of dit nu wel of niet expliciet is aangegeven. Feiten daarentegen vallen niet onder copyright. Het is dus toegestaan om statistiek uit te voeren op de feiten in teksten: *"the right to read, is the right to mine"* [5]. Maar het lokaal opslaan van teksten is weer niet toegestaan omdat dit onder het verwerken van teksten valt. Het risico op gebruik voor doeleinden die niet toegestaan zijn (zoals verspreiding of plagiaat) is dan aanwezig. Het rechtstreeks verwerken van teksten in een applicatie via weblinks, zonder de teksten op te slaan zou hiervoor een oplossing kunnen zijn. Uitgevers van wetenschappelijke tijdschriften staan het massaal automatisch downloaden van teksten voor text-mining alleen toe (soms is expliciete toestemming nodig) als het via hun eigen 'applicatie-interface' (API) gaat. Teksten met de juiste licentie die aangeeft dat er geen copyright is (bijvoorbeeld een 'creative commons'-licentie, zoals CCBY, of CC0) kunnen wel altijd lokaal verwerkt worden. Het is daarom aan te bevelen om tekstuele informatie onder zo'n licentie aan te bieden of te publiceren. Dit wordt ook wel 'Open Access' publiceren genoemd.

Informatie die terug te leiden is tot een persoon valt onder privacygevoelige informatie. Het is wettelijk niet toegestaan om deze informatie te gebruiken voor andere doeleinden dan waarvoor de informatie oorspronkelijk bedoeld was. Ook het tijdelijk opslaan van privacygevoelige informatie valt onder de verwerking van deze informatie en mag niet zonder medeweten van de personen zelf. Het is dus belangrijk om van te voren vast te stellen of de toegepaste text-mining binnen het te verwachten gebruik van de informatie valt.

Niet alle teksten zijn toegankelijk. Alleen informatie in het publieke deel van het internet kan gebruikt worden. Alle tekst waarvoor eerst moet worden ingelogd, is niet bereikbaar. De structuur van websites (bijvoorbeeld veel pagina's binnen pagina's) kan het moeilijk maken om informatie te vinden.

Taalgebruik kan ook een horde vormen voor het extraheren van de juiste informatie. Een techniek als spreektaalverwerking (NLP) wordt bemoeilijkt door het gebruik van ingewikkelde zinsopbouw. Er kunnen opsommingen plaatsvinden, ontkenningen ('Stof x wordt *niet* afgebroken tot Stof y'), er kunnen meerdere bijzinnen worden gebruikt of informatie kan in meerdere zinnen worden gepresenteerd. Een concept of woord kan geïntroduceerd worden, waarna ernaar gerefereerd wordt met woorden als 'deze' of 'die'. Daardoor is er nooit 100% nauwkeurigheid bij het extraheren van informatie uit tekst. Voor het uitsluiten van vals-positieven bij extractie of selectie van tekst kan het nodig zijn de resultaten toch nog door een persoon te laten controleren.

De voorbereiding en training van een nieuw NLP-model (bijvoorbeeld het kunnen herkennen van omstandigheden waarin zuiveringsefficiënties zijn gerapporteerd) is niet altijd een gemakkelijke taak. Om een robuust en efficiënt model te creëren, moet een deskundige van te voren een voldoende aantal gegevens van hoge kwaliteit verzamelen. Dit kan een tijdrovend en soms intensief proces zijn. Zodra deze stap is voltooid, bestaat de volgende stap uit het voorbereiden van de gegevens (door handmatige labeling van de gegevens) en het model te trainen zodat de computer de tekst kan 'lezen' en 'begrijpen' en de juiste informatie kan extraheren. Ten slotte moet het model getest

worden met behulp van meerdere 'testgegevens' en is een evaluatie van de prestaties nodig. Dit proces (voorbereiding, trainen en testen van gegevens) vereist een deskundige en neemt gewoonlijk heel wat uren in beslag. Het kan zijn dat de verbeterde prestatie ten opzichte van een simpele information retrieval-aanpak (met wat leeswerk) niet opweegt tegen de moeite.

Conclusies en vooruitblik

Grote hoeveelheden informatie zijn beschikbaar in tekstvorm. Text-mining kan een middel zijn om de rijkdom aan informatie met betrekking tot de watersector op efficiënte wijze te ontsluiten. Het is duidelijk dat verschillende technieken in text-mining ingezet kunnen worden voor veel verschillende taken. Voor een succesvolle ontsluiting van informatie zijn concrete doelen en gebruikerscasussen, toegankelijke tekstbronnen en meer kennis van deze technieken in de watersector nodig. Het ontwikkelen van bruikbare, specifieke applicaties voor de watersector zou het ontsluiten van informatie uit teksten ook vergemakkelijken.

Referenties

1. Hartmann, J. et al (2019). 'Use of literature mining for early identification of emerging contaminants in freshwater resources'. *Environ Evid* 8, 33. https://doi.org/10.1186/s13750-019-0177-z

2. Tian, X., Vertommen, I., Tsiami, L., Thienen, P. van, Paraskevopoulos, S. (2022). 'Automated Customer Complaint Processing for Water Utilities Based on Natural Language Processing—Case Study of a Dutch Water Utility'. *Water* 14, 674. https://doi.org/10.3390/w1404067

3. Zhu, R., Tu, X., Xiangji Huang, J. (2021). 'Utilizing BERT for biomedical and clinical text mining'. In: Lee, K.C., Roy, S.S., Samui, P., Kumar, V. (eds.), *Data Analytics in Biomedical Engineering and Healthcare*. Academic Press, 73-103. ISBN 9780128193143, https://doi.org/10.1016/B978-0-12-819314-3.00005-7.

4. Mulins, M. (2008). 'Information extraction in text mining'. *Computer Science Graduate and Undergraduate Student Scholarship*. 4. https://cedar.wwu.edu/computerscience_stupubs/4

5. Joseph, H. (2015). Citaat *"the right to read is the right to mine"* https://sparcopen.org/news/2015/the-right-to-read-is-the-right-to-mine/