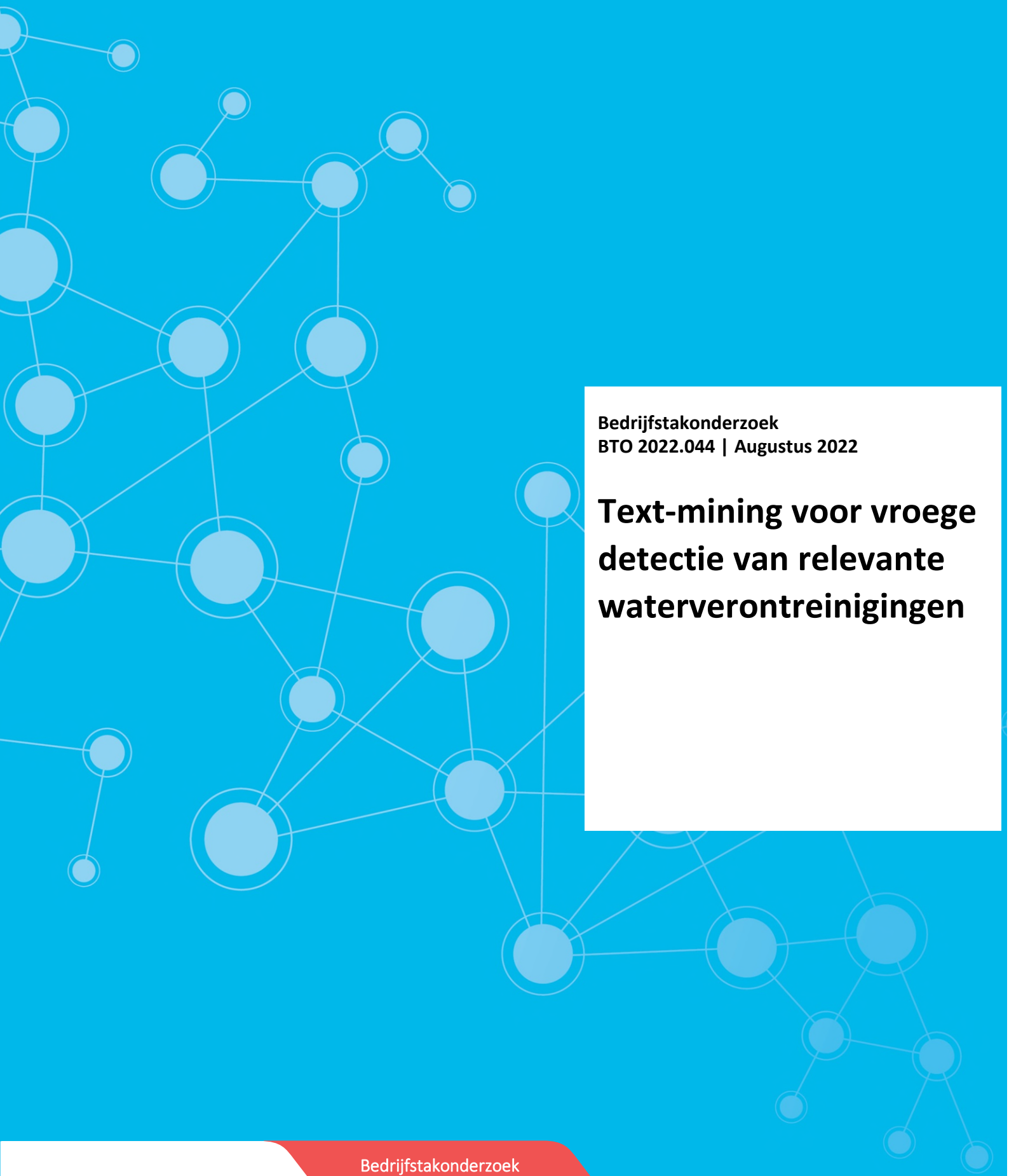


Bedrijfstakonderzoek
BTO 2022.044 | Augustus 2022

Text-mining voor vroege detectie van relevante waterverontreinigingen

Bedrijfstakonderzoek

KWR

Bridging Science to Practice

Rapport

Text-mining voor vroege detectie van relevante waterverontreinigingen

BTO 2022.044 | Augustus 2022

Dit onderzoek is onderdeel van het collectieve Bedrijfstakonderzoek van KWR, de waterbedrijven en Vewin.

Opdrachtnummer
402045-157

Projectmanager
Dr. Patrick S. Bäuerlein

Opdrachtgever
BTO - Thematisch onderzoek - Chemische veiligheid

Auteur(s)
Dr. ir. Tessa Pronk, Nienke Meekel MSc

Kwaliteitsborger(s)
Dr. Thomas ter Laak

Verzonden naar
Dit rapport is verspreid onder BTO-participanten.

Een jaar na publicatie is het openbaar.

Keywords

early warning, text-mining, waterverontreiniging, NLP, informatieverwerking

Jaar van publicatie
2022

Meer informatie
Dr. ir. Tessa Pronk
T 030-6069681
E tessa.pronk@kwrwater.nl

PO Box 1072
3430 BB Nieuwegein
The Netherlands

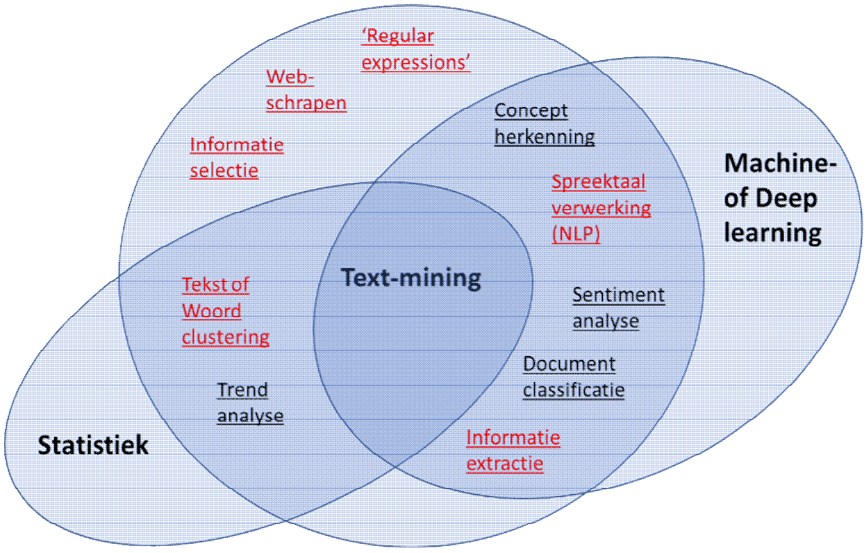
T +31 (0)30 60 69 511
E info@kwrwater.nl
I www.kwrwater.nl

Managementsamenvatting

Text-mining voor vroege detectie van relevante waterverontreinigingen

Auteur(s) Dr. ir. Tessa Pronk, Nienke Meekel MSc.

Aanwijzingen voor (toekomstige) waterverontreiniging kunnen flink verstopt zijn in diverse bronnen met digitaal beschikbare tekst, van rapporten tot social media. Het lezen en verwerken van tekst in dit soort bronnen om te zoeken naar aanwijzingen voor toekomstige waterverontreiniging 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. Er zijn vele technieken en toepassingen denkbaar. Dit rapport laat zien dat verschillende technieken kunnen worden ingezet om tekstbronnen te toetsen op het bevatten van informatie rond chemische waterverontreinigingen. Het is vervolgens aan een expert om de informatie te interpreteren en vervolgacties te nemen. Met behulp van het geleverde overzicht van technieken en bronnen, met voorbeelden van de opbrengsten en de opgeleverde R-scripts met code kunnen de technieken in de toekomst verder worden uitgewerkt.



Overzicht van beschikbare technieken in text-mining en hun relatie tot statistiek en machine (of deep-) learning. De in het rood aangegeven technieken zijn toegepast in dit rapport.

Belang: Informatie uit tekstbronnen beter ontsluiten voor vroege signalering probleemstoffen
Informatie rond nieuwe bedreigingen voor de waterketen is onder andere beschikbaar in de vorm van digitale tekst, die versnipperd over diverse bronnen beschikbaar is. Denk daarbij aan informatie in rapporten, vergunningen, memo's, nieuwsbrieven, wetenschappelijke publicaties, websites, maar ook uitingen op sociale media. Deze informatie tijdig signaleren en verzamelen draagt bij aan een vroege signalering van potentieel vervuilende stoffen. Het lezen en verwerken van tekst in dit soort bronnen, op zoek naar aanwijzingen voor toekomstige waterverontreiniging, is echter tijdrovend. Daarom is

het belangrijk een methodologie op te bouwen om met text-mining nieuwe, potentieel problematische chemicaliën te kunnen identificeren op basis van een breed scala aan informatiebronnen.

Aanpak: Een analyse van Text-mining technieken om informatie te vinden

Text-mining maakt het mogelijk om automatisch (grote) hoeveelheden tekstuele informatie te doorzoeken en de gevonden informatie op een gestructureerde manier bij elkaar te brengen. In dit onderzoek zijn hiervoor verschillende technieken verkend, met verschillende toepassingen. Zie ook bovenstaande figuur.

Resultaten: Technieken en informatiebronnen zijn in beeld

Diverse combinaties van text-mining technieken en bronnen zijn ingezet en de potentie en aard van de gevonden informatie rond nieuwe waterverontreinigingen is beschreven. Dit heeft inzicht opgeleverd over het gebruik van de technieken zelf, en de opbrengst die de technieken in combinatie met de gekozen bron opleveren. Onderzocht is ook wat de waarde van diverse tekstbronnen is voor het vinden van aanwijzingen voor nieuwe verontreinigingen. Daarnaast zijn enkele concrete bedrijfsprocessen en stoffen die mogelijk in de toekomst voor nieuwe verontreiniging kunnen zorgen, onder de aandacht gebracht. Zo heeft het onderzoek een aantal concreet toegepaste technieken voor text-mining opgeleverd, en enkele R-scripts met code om de tekst van digitale bronnen te doorzoeken. Met behulp van het geleverde overzicht van technieken en bronnen met voorbeelden van de opbrengsten en de opgeleverde R-scripts met code kunnen de technieken in de toekomst verder worden uitgewerkt.

Toepassing: Specifieke applicaties

De verzamelde kennis vormt een basis voor het verder ontwikkelen van specifieke applicaties die door drinkwaterbedrijven ingezet kunnen worden om informatie over toekomstige verontreinigingen uit digitaal beschikbare teksten te halen. Het is goed om te benadrukken dat er altijd expertkennis nodig is om de resultaten te beoordelen op relevantie, bruikbaarheid en in hoeverre het nieuwe informatie is. Door de brede inzetbaarheid van de technieken en de mogelijkheden om deze te optimaliseren per casus, is text-mining op vele vlakken toepasbaar. Het is voornamelijk van toegevoegde waarde voor onderzoeksvragen waarbij grote hoeveelheden tekst en/of online bronnen verzameld en doorzocht moeten worden.

Rapport

Dit onderzoek is beschreven in het rapport *Text-mining voor vroege detectie van relevante waterverontreinigingen* (BTO 2022.044).

Lees ook andere relevante publicaties:

- H2O: Text-mining voor de watersector (2022) https://www.h2owaternetwerk.nl/vakartikelen/text-mining-voor-de-watersector
- Verslag BTO workshop Text-mining (2022) https://www.kwrwater.nl/actueel/text-mining-for-early-detection-of-water-related-substances/
- BTO 2015.059 Signaleren van nieuwe stoffen (2014-2015) https://livelink.kwrwater.nl/livelink/livelink.exe?func=ll&objaction=overview&objid=53576812#
- PS-DRINK website https://www.rivm.nl/drinkwater/risicos-voor-drinkwater-psdrink
- BTO 2022.021 (2022) Deep Explorations: an explorative study for machine learning and deep learning applications in the water sector.
- BTO Zeer zorgwekkende stoffen (verwacht 2022)
- BTO VO integraal Voorspellen van de biologische afbraak van persistente stoffen (verwacht 2023)

Data en code

Bij dit onderzoek hoort een data- en code pakket ‘TM Data-pakket (2022)’. Deze is op verzoek benaderbaar via Sharepoint.

Meer informatie
Dr. ir. Tessa Pronk
T 030-6069681
E tessa.pronk@kwrwater.nl

PO Box 1072
3430 BB Nieuwegein
The Netherlands

Inhoud

1 Introductie

Lang voordat stoffen in regelgeving en meetprogramma's terechtkomen, kunnen er in tekstbronnen zoals rapporten, sociale media, nieuwsberichten, websites van toezichthoudende instanties of wetenschappelijke literatuur aanwijzingen zijn dat deze mogelijk naar het Nederlandse watersysteem kunnen worden geëmitteerd. Het lezen van alle mogelijk relevante informatie is niet haalbaar, zeker omdat de beschikbaarheid van digitale informatie elk jaar toeneemt en informatie versnipperd is vastgelegd. Met een 'text-mining' (TM) aanpak kunnen stoffen waarvoor een indicatie is van (toekomstige) emissie naar water semiautomatisch worden ontdekt. Deze stoffen kunnen vervolgens extra onder de loep worden genomen. Dit is vooral van belang als er ook een indicatie is voor schadelijke eigenschappen van deze stoffen. In het project PS-Drink (RIVM) is dit onderwerp in 2019 reeds aangepakt door in abstracts van wetenschappelijke artikelen gericht te zoeken naar trefwoorden die duiden op het aantreffen van nieuwe stoffen in water.[1] Na verdere selectie leverde deze aanpak ongeveer tweehonderd nieuwe stoffen op, het voorkomen van drie daarvan wordt op dit moment onderzocht in het Nederlandse oppervlaktewater. Daarnaast heeft een BTO-project dit op een vergelijkbare manier aangepakt met behulp van de TNO 'ERIS' software.[2]Dit leverde tien relevante stoffen op uit een totaal van 1733 gesignaleerde bedreigingen. Grofweg richtten beide benaderingen zich op het opsporen van woordcombinaties van stoffen of stofgroepen in teksten met relatie tot de begrippen 'nieuw' en 'water'. ERIS maakt daarnaast gebruik van een ontologie waardoor concepten en entiteiten aan elkaar gelinkt zijn en er op deze manier meer informatie toegevoegd wordt aan de tekst. De aanpak van het huidige project is breder dan de taak om nieuwe chemicaliën in wetenschappelijke publicaties achterhalen, het is ook gericht op informatie over nieuwe bronnen en emissies. Daarnaast wordt een veel breder scala aan tekstbronnen gebruikt.

Informatie over nieuwe bronnen en emissies kunnen ruwweg worden geïdentificeerd als leidend tot emissies afkomstig van industrie, landbouw en huishoudens. Voor elk van deze emissieoorzaken werden relevante en actuele tekstbronnen verzameld en beoordeeld op de toegevoegde waarde. Bijvoorbeeld de waarde als informatiebron van sociale media (bijv. Twitter), nieuwsberichten, beschikbare pdf's of teksten op relevante websites, wetenschappelijke publicaties en Wikipedia.

Verschillende technieken uit 'machine learning' (ML) en TM maken het mogelijk om van tekst tot gestructureerde data te komen. Het is van belang dat de watersector de kennis over dergelijke geavanceerde technieken verder ontwikkelt, omdat het lezen van alle beschikbare informatie onhaalbaar is. Waar mogelijk, is 'natural language processing' (NLP) toegepast om de relaties tussen woorden in zinnen te bepalen, en ook lexicons van chemische namen of andere relevante concepten worden in de aanpak opgenomen om diverse schrijfwijzen, talen, aanduidingen van stoffen en trefwoorden te kunnen herkennen en hergebruiken. Waar mogelijk werd ML toegepast om de computer te laten beslissen welke kenmerken in een informatiebron/chemische stof indicatief zijn voor een beschrijving van een proces dat een specifieke verontreiniging zal uitstoten die nieuw en relevant is voor de Nederlandse wateren.

Het onderzoeksproject heeft tot doel een methodologie op te bouwen met betrekking tot TM om nieuwe potentieel problematische chemicaliën voor de Nederlandse wateren te identificeren op basis van een breed scala aan informatiebronnen. Deze informatie gaat over mogelijke nieuwe emissies die relevant zijn voor Nederland door industrie, huishoudens, of landbouw. Dit kan worden beschouwd als onderdeel van een vroegtijdig waarschuwingssysteem om (nieuwe) processen te detecteren en te identificeren die stoffen kunnen uitstoten en zo mogelijk problemen kunnen veroorzaken voor de waterkwaliteit. Daarnaast kan het early-warning systeem gebruikt worden om meetprogramma's uit te breiden, of de aanleiding zijn om een verzoek te initiëren om nadere informatie te verkrijgen over de mogelijke emissie. De methoden zijn gecodeerd in de programmeertaal "R". Als voorbeeld voor toepassing van de TM technieken is gekozen voor het stroomgebied van de Rijn, en met name het

gebied Rheingraben-Nord in Duitsland. Na een introductie van dit gebied en de bedrijvigheid daarin, wordt per geanalyseerde informatiebron in een paragraaf zowel de methode als) het resultaat samengevat. Gedetailleerde resultaten staan in de bijlagen.

2 Case studie: Rijnstroomgebied en bedrijven

Als case studie is het gebied Rheingraben-Nord, in het Rijnstroomgebied in Duitsland geselecteerd. In dit gebied is veel bedrijvigheid die de waterkwaliteit ook in Nederland beïnvloedt. Het rapport over afvalwaterlozing van het Ministerium für Umwelt, Landwirtschaft, Natur- und Verbraucherschutz des Landes Nordrhein-Westfalen in Duitsland bevat lijsten van alle bedrijven die water lozen op de rivier de Rijn in het gebied Rheingraben-Nord.[3] De jaarlijkse lozing van afval- en koelwater wordt gerapporteerd, zowel op basis van de gerapporteerde hoeveelheden als de geschatte belasting. Deze lijst van 93 bedrijven waarvan de belasting bekend is, is handmatig geëxtraheerd en de bijbehorende websites en Twitter-gebruikersnamen van de bedrijven zijn verzameld. Bovendien is op basis van de websites een inschatting gemaakt van het producttype of type industrie. Deze lijst vormt een basis voor verdere analyses op teksten van bronnen die relevant zijn voor dit gebied.

Figuur 1. Rheingraben-Nord, de Rijn en haar kleinere toevoeren.[4]

2.1 Sociale media

Bedrijven vertegenwoordigen zichzelf vaak op sociale media, bijvoorbeeld met een Twitter-account. Twitter-gegevens zijn openbaar. Twitter-gegevens kunnen daarom legaal worden gebruikt. Er zijn wel enkele ethische beperkingen. Het is bijvoorbeeld niet toegestaan om potentieel gevoelige kenmerken (zoals religie, gezondheid) over Twitter-gebruikers te ontsluiten. Het is alleen toegestaan om Twitter-accounts te analyseren op basis van informatie waarvan een gebruiker redelijkerwijs zou kunnen verwachten dat deze gebruikt zou worden voor het doel van de analyse.[5]

Een Twitter-app is een tool die gebruikt kan worden om tweets te verzamelen van vooraf gedefinieerde gebruikers en/of tweets met specifieke woorden erin. Een Twitter-app kan, onder andere, worden aangestuurd via de open source statistische software ‘R’. Relevante tweets kunnen worden gevonden en geïmporteerd via code in een R script. Vervolgens kunnen deze worden doorzocht op specifieke woorden of patronen die duiden op nieuwe activiteit die vervolgens mogelijk kan leiden tot de aanwezigheid van aan de activiteit gelinkte stoffen in het water, bijvoorbeeld door bedrijven langs de Rijn. Hiervoor zijn steekwoorden gebruikt die te maken hebben met het

gebied gecombineerd met woorden die staan voor produceren en stoffen, de naam van het bedrijf, en het begrip 'nieuw'. De procedure waarmee de Twitterberichten zijn gefilterd is:

1. Een match met een of meer trefwoorden voor locaties in het tweet meta-dataveld 'locatie' (hiervoor is gekeken welke locaties genoemd zijn in 'locatie' en welke relevant zijn voor het casestudie gebied). Dit zijn de volgende trefwoorden: ("germany|deutschland|rhein|düsseldorf|köln|bonn|münchen|wesseling")

2. Een match met een of meer trefwoorden voor locaties in het Twitterbericht zelf:
("germany|deutschland|duitsland|noordrijn|nordrhein|rhein|rhine|rijn")

3. Een match met een of meer Duitse of Engelse trefwoorden rond bedrijvigheid in de Twitterberichten zelf:
("substance|permit|synthe|announce|increase|expand|manufact|invest")
("chemisch|erlaub|genehmigung|verkünden|steigern|erhöhen|vergröß|kläranlage")

4. De combinatie van trefwoorden voor aanwijzingen voor bedrijvigheid en het woord 'nieuw' (Duits of Engels) binnen een afstand van drie woorden in Twitterberichten zelf:
("product|plant|proces|technolog|site|locat|lab|facilit|synthe|construct|sewage")
("produkt|fabrik|prozess|stelle|standort|einrichtung|bauen|kläranlage")
De trefwoorden kunnen hier iets minder specifiek zijn dan bij 3. omdat ruis (niet-relevante tweets) wordt voorkomen door de combinatie met het woord 'nieuw'.

5. Deze zijn gecombineerd als ((1. OF 2.) EN 3.) OF (4.) waarbij het belangrijk is om te realiseren dat bij een combinatie met 'OF' de verzameling vergroot, en een combinatie met 'EN' alleen de doorsnede selecteert.

Van de meer dan 50.000 tweets* in totaal van de bedrijven die een Twitteraccount hadden, of genoemd werden op Twitter voldeden in totaal 540 aan de zoekcriteria. Na handmatige screening bleek dat 16% van deze geselecteerde tweets relevante informatie bevatte die aanleiding zouden zijn voor verder onderzoek naar de implicaties voor de waterkwaliteit. Er is daarnaast een steekproef onder niet-geselecteerde tweets gedaan die ongeveer even groot is als het totaal aan geselecteerde tweets, om vast te stellen hoeveel tweets ten onrechte niet zijn geselecteerd (vals negatieven). In de steekproef van 500 niet geselecteerde tweets bleek slechts 1% relevant. Dit betekent dat er een significante verrijking was in het aantal relevante tweets na toepassen van de zoekcriteria. Een deel van de activiteiten in de niet geselecteerde tweets was in geselecteerde tweets ook gesignaleerd, in andere berichten.

In Tabel 1 zijn een aantal voorbeelden gegeven van geselecteerde tweets met mogelijke indicatie voor nieuwe bedrijvigheid in het stroomgebied van de Rijn. Deze kunnen aanleiding geven tot nader onderzoek rond wat dit betekent voor nieuwe emissies in het Rijnstroomgebied. In het TM Data-pakket (2022) staan alle beoordeelde tweets die na het toepassen van de trefwoorden voor het filteren zijn overgebleven, alsmede een beoordeelde even grote random selectie van niet-geselecteerde tweets.

** Tijdens deze analyse was het trefwoord 'genehmigung' nog niet toegevoegd bij steekwoorden in punt 3 op deze pagina. Dit trefwoord leverde achteraf nog 18 extra twitter berichten op die voldeden aan de zoekcriteria. Een activiteit die hiermee extra werd ontdekt is sanering van de bruinkoolmijnbouw in het Rijnland (toegevoegd aan Tabel 2).*

Tabel 1. Voorbeelden van bedrijvigheid in geselecteerde Twitterberichten van of over bedrijven met emissies in het Rijnstroomgebied.

Jaar	Bericht	Korte beschrijving (handmatig toegekend)
2020	#evonik strengthens its focus on #adhesive and #sealant solutions with a new multi-purpose #silicone production facility in #geesthacht, germany:	Nieuwe siliconen productie faciliteit
2021	today is the inauguration of #evonik's largest investment to date in germany, the new #polyamide12 plant: whenever #plastics are exposed to exceptionally high stress, the high performance plastic from @evonikhp comes to the rescue:	Nieuwe polyamide12 fabriek
2022	the large-scale production of phytochol in hanau, germany, will meet an increased market demand for cholesterol	Verhoogde Phytochol productie
2022	as part of an expansion project for its us site in #theodoreal, specialty chemicals producer @evonik plans to invest $176.5 million in a new methyl mercaptan plant. the chemical is used to make metamino (dl-methionine), which is used in livestock feed.	Nieuwe methyl mercaptan fabriek
2019	many of the metals we produce are essential for megatrends such as digitalization, renewable energies, electric vehicles, and urbanization to become a reality in the first place. at the same time, these areas drive demand for our industrial metals even further.	Verhoogde metaal productie

Een overzicht van alle 32 vormen van bedrijvigheid die beschreven werd in de Twitterberichten staat in Tabel 2.

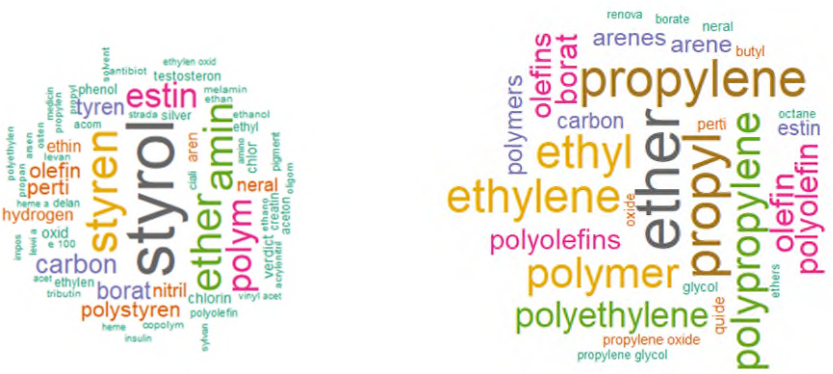
Tabel 2. Korte beschrijving van de aard van nieuwe bedrijvigheid in geselecteerde Twitterberichten.

Nieuwe siliconen productie faciliteit	Verhoogde metaal productie	Verkoop Melamine bedrijf	Meer in 3d-drug printing
Nieuwe polyamide12 fabriek	Verhoogde activiteit silica productie	Nieuwe waterstof/ elektrolyse fabriek	verhoogde koper, kobalt, nikkel behoefte voor productie elektrische auto's
Verhoogde Phytochol productie	Bouw recycling centrum	Massa productie van elektroliseerders	Chloromethaan productieverhoging
Nieuwe methyl mercaptan fabriek	Overstap van weg naar binnenvaart transport	Verhoogde productie handdesinfectiemiddelen	Nieuwe chlorine and potassium hydroxide fabriek
Investering in biomassa boiler	Nieuwe fabriek voor batterijen voor elektrische auto's	Nieuw afvalwater verwerking station	Overname BSN medical
Investering stro-pulp fabriek	Nieuwe polymeer techniek voor tabletvorm geneesmiddel	Verhoogde cholestorol productie	Bio-ptl fabriek
Investering staal-productie	Meer Methanol synthese	Meer autoproductie	Investering verbetering aluminium kwaliteit
polyactide (bioplastic) fabriek	Nieuwe aluminium alloy	Uitbreiding pvc fabriek	Sanering van de bruinkoolmijnbouw

Naast het selecteren van tweets is er ook gekeken naar mogelijkheden om de bedrijven te karakteriseren aan de hand van stofnamen in de tweets. Het is mogelijk om een chemische 'fingerprint' te extraheren uit de

Twitterberichten door per bedrijf de chemicaliën die voorkomen in de berichten te extraheren. Vervolgens kan hier een woordwolk van gemaakt worden. Dit geeft een overzicht van de stoffen waarmee het bedrijf mogelijk te maken heeft. Voor het identificeren van stoffen in de Twitterberichten is gekozen voor het herkennen van namen uit een vooraf gedefinieerde lijst. Deze lijst met namen van chemicaliën en productnamen is gedownload van https://www.ebi.ac.uk/chebi/ "chemical entities with biological interest".

Sommige productnamen hebben overlap met gewone spreektaal. Bijvoorbeeld, het product 'finish'. Dit geeft veel valse resultaten. Ook 'lead' geeft veel valse resultaten. Om die reden werken we met een 'zwarte lijst' van productnamen uit CheBi die te veel van deze valse resultaten geven. Deze namen laten we buiten beschouwing. In Figuur 2 staan twee voorbeelden van woordwolken die zijn opgebouwd uit genoemde stofnamen in Twitterberichten van/over bedrijven. Een complicatie bij het linken van deze stoffen aan het rijnstroomgebied is dat sommige bedrijven wereldwijd opereren. Indien een selectie op de relatie met het rijnstroomgebied wordt gedaan, blijven er over het algemeen te weinig stofnamen over voor een woordwolk. De woordwolken op basis van twitter-berichten kunnen dus vooral een *algemene* indruk geven van een bedrijf, maar zijn niet (altijd) relevant voor het bestudeerde watersysteem.



Figuur 2. Voorbeelden van gevonden stoffen in de tweets van gelinkt aan twee verschillende bedrijven. Deze stofnamen geven een indicatie van de activiteiten van een bedrijf. Echter, veel bedrijven hebben locaties wereldwijd waardoor deze niet specifiek aan Duitsland gekoppeld kunnen worden. Bij filteren op het relevante gebied blijven er weinig genoemde chemicaliën over.

Kader 1. Het gebruik van Regular Expressions

Bij het doorzoeken van Twitterberichten is ook gebruik gemaakt van zogenaamde 'Regular Expressions' (afgekort: 'RegEx'). Deze kunnen gebruikt worden om tekst te herkennen op basis van letterpatronen. RegEx zijn conventies om deze patronen te beschrijven. Bijvoorbeeld, '[0-9]' betekent dat de gezochte letter een cijfer is tussen 0 en 9. Met de toevoeging {4} wordt aangeven dat er in totaal vier aaneengesloten cijfers moeten worden gezocht. Met het patroon '[0-9]{4}[A-Z]{2}' worden alle patronen met vier cijfers gevolgd door twee hoofdletters gevonden. Er zijn ook conventies om aan te geven op welke positie een patroon moet staan (bijvoorbeeld aan het begin van de zin of het einde van een woord). RegEx zijn bijzonder veelzijdig en kunnen ingewikkelde letterpatronen herkennen. In het script voor het extraheren van informatie uit Twitterberichten zijn ze gebruikt om alternatieve schrijfwijzen voor bedrijfsnamen te genereren. Bedrijfsnamen die halverwege een tweede hoofdletter bevatten, worden mogelijk ook los geschreven, of zonder de tweede hoofdletter. Die opties worden met behulp van RegEx gegenereerd. Ook zijn ze ingezet om webadressen in tweets in het geheel te verwijderen en woorden binnen een bepaalde afstand van een trefwoord te selecteren.

2.2 Vergunningen

In Duitsland stroomt de Rijn door vier deelstaten (Bundesländer): Baden-Württemberg, Rheinland-Pfalz, Hessen en Nordrhein-Westfalen. Alle behalve Rheinland-Pfalz zijn verdeeld in bestuurs- of administratieve districten, zogenoemde 'Regierungsbezirk'. Beslissingen over vergunningen en vergunningverleningen worden door deze districten uitgevoerd. De aangevraagde en verleende vergunningen worden gepubliceerd in zogeheten 'Amtsblätter' die op hun websites worden gepubliceerd (Tabel 3). Aangevraagde en verleende vergunningen voor nieuwe activiteiten worden centraal gepubliceerd via https://uvp-verbund.de/startseite. De wijzigingen in de vergunningen worden elke maandag per district gepubliceerd. In plaats van deze handmatig te doorlopen om aanwijzingen te vinden die kunnen duiden op nieuwe activiteiten welke mogelijk kunnen resulteren in de emissies van (nieuwe) stoffen in de Rijn, kan dit proces versneld worden door deze publicaties automatisch te doorzoeken. Naast het automatisch doorzoeken kunnen deze publicaties ook automatisch verzameld worden via 'web-scraping'. In het Nederlands vertaald betekent deze techniek zoiets als 'website-schrapen' van informatie. Deze web-scraping herkent de relevante links voor van te voren ingegeven pagina's.

Tabel 3. Overzicht van de websites waar de Amtsblätter gepubliceerd worden door de districten die relevant zijn voor de Rijn.

Regio	Deelstaat	Webpagina
Köln	Nordrhein-Westfalen	https://www.bezreg-koeln.nrw.de/brk_internet/amtsblatt/2022/index.html
Düsseldorf	Nordrhein-Westfalen	https://www.brd.nrw.de/services/amtsblatt/amtsblaetter-2022
Münster	Nordrhein-Westfalen	http://www.bezreg-muenster.de/de/service/bekanntmachungen/amtsblaetter/index.html
Baden-Württemberg	Baden-Württemberg	https://rp.baden-wuerttemberg.de/rpk/Service/Bekanntmachung/Seiten/default.aspx
Heidelberg	Baden-Württemberg	https://www.heidelberg.de/hd,Lde/HD/Rathaus/Oeffentliche+Bekanntmachungen+Umweltrecht.html

Er werd een R-script ('Webscraping_Amtsblatter.R', zie Bijlage VII) ontwikkeld dat de Amtsblätter voor elk district bij het uitvoeren van het script downloadt en direct de tekst doorzoekt op relevante trefwoorden zoals beschreven in Bijlage I. De output van het script is een dataframe met een overzicht van de gedetecteerde trefwoorden per gescreend bestand,

Tabel 4. Dit kan aanleiding geven om het bestand daadwerkelijk te gaan lezen, omdat het mogelijk relevante informatie bevat. Het script kan ook de zinnen extraheren die het trefwoord bevatten en dit als een tekstbestand uitvoeren. Een voorbeeld is weergegeven in Kader 2, dit is de ruwe output van het script en kan dus spel- en/of interpunctiefouten bevatten welke geïntroduceerd zijn bij het inlezen en omzetten van het PDF bestand.

In het geanalyseerde Amtsblätt in Tabel 5 wordt de stof kwik opvallend vaak genoemd. Dit kan aanleiding zijn tot een nadere beschouwing van de tekst in dit document rond deze stof.

Tabel 4. Deel van het output data frame van de screening van de Amtsblätter uitgegeven tussen 1 januari 2021 en 16 maart 2021. Het document 'BW_Mannheim_201215_Genehmigungsbescheid.pdf' bevat het grootste aantal keer 'abwasser'.

	abwag[a]	abwasser[b]	abwasserabgabengesetz[c]	abwasserbehandlungsanlage[d]
BW_Mannheim_201215_Genehmigungsbescheid.pdf	1	11	1	4
BW_Karlsruhe_stadt_201208_Genehmigungsbescheid.pdf	0	8	0	0
BW_Mannheim_181218_Genehmigungsbescheid.pdf	0	5	0	0
BW_Rhein-Neckar-Kreis_180608_Genehmigungsbescheid.pdf	0	4	0	0
Düsseldorf_Amtsblatt-Nr-07-Anlage-Ziffer-42.pdf	0	4	0	0
Köln_07-2021.pdf	0	3	0	0

[a, c] AbwAG (afkorting voor Abwasserabgabengesetz) is een Duitse wet voor afvalwaterheffingen
[b] Abwasser is het Duitse woord voor afvalwater
[d] Abwasserbehandlungsanlage is het Duitse woord voor rioolwaterzuivering

Tabel 5. Resultaten van de screening van 'BW_Mannheim_201215_Genehmigungsbescheid.pdf' met de lijst met Duitse chemicaliën.

Duits trefwoord	Nederlandse vertaling	Aantal woorden
Arsen	arseen	2
Blei	lood	3
Cadmium	cadmium	3
Chrom	chromium	3
Eisen	ijzer	1
Kupfer	koper	3
Nickel	nikkel	3
Quecksilber	kwik	13
Stickstoff	stikstof	3
Thallium	thallium	2
Wasser	water	5
Zink	zink	2

Kader 2. Voorbeeld van zinnen uit een Amtsblätt: Köln_13-2021.pdf

Bundes.immissionsschutzgesetz

Rechtsgrundlag für die Fortschreibung ist § 47 Absatz 1 **Bundes-Immissionsschutzgesetz** (BImSchG) Die öffentlich-rechtlich Vereinbarung zur Bildung in Verbindung mit der Neununddreißigsten Verordnung einer Ausschreibungsgemeinschaft zur Beschaffung von zur Durchführung des **Bundes-Immissionsschutzgesetz** Lieferungen und Leistungen vom <Datum> wird auf den (39.

Sie beantragt Allgemein Vorprüfung des Einzelfal nach § 7 Abs. 1 - gemäß § 16 Abs. 4 **Bundes-Immissionsschutzgesetz** in Verbindung mit Nr. 13.3.2 der Anlage 1 und Anlage 3 (BImSchG) divers Änderungen an der Heizzentrale.
Einleitung

Dementsprechend handeln die Vertragspart- und ner im Rechtsverkehr nach außen jeweil ausschließ- dem Landschaftsverband Westfalen-Lippe, lich al Stadt oder Landschaftsverband. vertreten durch den Direktor des § 2 Verfahren Landschaftsverband Westfalen-Lippe, Landeshaus, Freiherr-vom-Stein-Platz 1, 48133 Münster (1) Vor **Einleitung** ein jeden Vergabeverfahren wird zwischen den Vertragspartnern festgelegt, welcher wird gemäß §§ 23 ff. des Gesetz über die kommunal Vertragspartn das jeweilig konkret Verfahren or- Gemeinschaftsarbeit (nachstehend GkG) vom 1.

Die Fachdienststellen der übrigen Vertragspart- ner können ein abweichend Votum formulieren. (5) Die Anwendungsvereinbarung ist, sofern ein Betei- ligungspflicht gegeben ist, vor **Einleitung** des Verga- (2) Die im Rahmen der fachtechnischen Wertung durch- beverfahren dem nach § 9 zuständigen Rechnungs- geführt Bemusterung wird gemeinsam von den Ver- prüfungsamt zuzuleiten. tragspartnern durchgeführt.

2.3 Websites van industriële bedrijven

Een veelgebruikte techniek voor het extraheren van website-updates of nieuwsberichten is het gebruik van RSS-feeds. Really Simple Syndication (RSS) is een hulpmiddel dat alle website-updates verzamelt als .xml-bestanden en gebruikers kunnen deze updates lezen met behulp van een RSS-feed reader. Dit voorkomt dat de gebruiker regelmatig alle afzonderlijke websites moet bezoeken om op de hoogte te blijven van de updates en aankondigingen die worden gedaan. Bij wijze van pilot zijn de websites van 10 bedrijven langs de Rijn gecontroleerd op de aanwezigheid van een RSS-feed. Slechts 2 daarvan hadden een RSS-feed, die leeg was. Het bleek dat RSS-feed vrij verouderd is en momenteel vooral gebruikt wordt voor bijvoorbeeld podcasts, en niet per se voor nieuwsberichten op bedrijfswebsites. Daarom werden RSS-feeds niet beschouwd als relevante bronnen voor dit onderzoek.

Webpagina's worden geschreven in HyperText Markup Language (HTML), die wordt gebruikt om de structuur van de webpagina te specificeren. Webpagina's zijn opgebouwd in een boomachtige structuur die bestaat uit elementen, omgeven door tags. Deze elementen kunnen bijvoorbeeld tekst of een afbeelding bevatten. Webpagina's kunnen worden geschraapt voor relevante informatie met behulp van R. Eerst wordt de URL (webpagina-adres) van een bedrijfswebsite gelezen en wordt de HTML-boom ervan gedefinieerd. Vervolgens wordt de structuur doorzocht op hyperlinks met behulp van het XML-pakket en deze worden gefilterd om relevante trefwoorden te bevatten zoals vermeld in Tabel 6. Het gaat hierbij om trefwoorden in links, hierbij wordt geen

rekening gehouden met hoofdletters en worden letters zoals ö, weergegeven als ‘oe’. Als resultaat wordt een lijst van relevante URL's gegenereerd en deze kunnen worden doorzocht op informatie over de producten en/of het nieuws van het bedrijf. Tot hier is het een uniforme procedure die geschikt is voor elke website. Maar het ontwerp van elke webpagina is anders en heeft een andere HTML-structuur. Over het algemeen bevat het "<body>"-element de inhoud van een webpagina. Een algemeen script dat alle beschikbare tekst van een webpagina extraheert is een uitdaging om te ontwikkelen, omdat het extraheren van de hele <body> van een webpagina resulteert in een heleboel ongewenste structuurelementen en tags. Als ‘proof of concept’ werden de websites van vijf verschillende bedrijven https://www.currenta.de/ (website 1), https://www.ilhoist.com/de_de (website 2), (https://www.venatorcorp.com (website 3), https://uniferm.de/de/ (website 4) en https://www.essity.de/ (website 5) gescreend voor informatie over hun producten en nieuws.

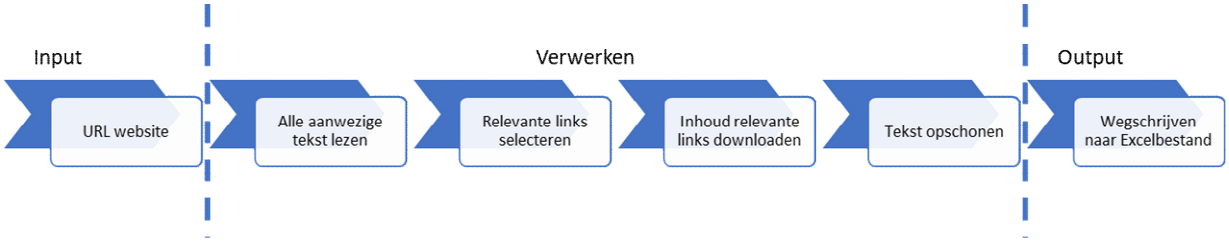
Tabel 6. Trefwoorden in hyperlinks die kunnen duiden op pagina’s gerelateerd aan productinformatie en nieuws-updates.

Trefwoorden product	Trefwoorden nieuws
product	Meldungen
products	Aktuelles
application	Presseinformationen
applications	Pressemitteilungen
produkte	Presse
solutions	Medien
loesungen	Media
market	News
produktgruppen	Neuigkeiten
geschaeft	Newsroom
produktneuheiten	Presseberichte

Deze websites zijn niet op dezelfde manier gestructureerd, waardoor de code voor elke afzonderlijke website moet worden aangepast of gewijzigd. Soms is het aanroepen van het "hoofd"-gedeelte van de website voldoende om alle informatie te extraheren. Andere websites vereisen het aanroepen van klassen, die identificatienamen zijn voor doelelementen. Voorbeelden zijn webpaginadelen die zijn onderverdeeld in een klasse met bijvoorbeeld "page-content" of "product-wrapper". Een andere manier om relevante webpaginadelen te selecteren is op basis van elementen. De meeste tekst kan worden geëxtraheerd met behulp van paragraaf-elementen "<p>", maar sommige websites bevatten relevante informatie buiten deze paragrafen om, zoals website 4.

Het script (zie ‘Webscraping_websites.R’, Bijlage VII) is ontwikkeld voor vijf websites als een case studie. Alleen het schrapen van website 4 is niet volledig gelukt vanwege de complexe structuur. Voor optimale prestaties moet handmatig worden gecontroleerd of alle gewenste informatie uit de webpagina's is gehaald, wat tijdrovend is.

De output bestaat uit een dataframe voor elke website met in de eerste kolom de URL en in de tweede kolom de geschraapte tekst van die URL. Dit dataframe wordt naar een Excel-bestand geschreven. De geschraapte tekst kan eenvoudig worden gescreend op chemische stoffen of worden onderworpen aan text-mining strategieën. Naast producten kunnen ook webpagina's worden gescand op nieuwsupdates. Dit vereist echter nog meer maatwerk per website omdat sommige websites hun nieuwsupdates op verschillende link-locaties opslaan. Het persbericht zelf is vaak verborgen en er worden slechts een paar zinnen gegeven. Dit moet voor elke website gecontroleerd worden (bijvoorbeeld voor website 5 is dit het geval).



Figuur 3. Schematisch overzicht met in- en output van het proces van web-scrapen van websites.

Een minder specifieke, alternatieve aanpak bestaat uit het schrapen van *alle* informatie die in de node met de hoofdinhoud staat, dus inclusief html tags e.d. Dit kan vervolgens doorzocht worden op informatie. Het voordeel hiervan is dat dit voor veel websites automatisch kan. Er is alleen een weblink nodig van de pagina van interesse op de website. Tevens is er een script ontwikkeld om iteraties uit te voeren, dus op de betreffende pagina te zoeken naar relevante links, de inhoud hiervan te schrapen en die pagina ook weer te doorzoeken op relevante links enz. Met dit 'springen' is het dus mogelijk om de pagina's van een website te doorzoeken. Echter dient dit ook weer te worden geoptimaliseerd per te doorzoeken website (per website zijn er bijvoorbeeld andere criteria voor relevante links). Dit is als 'proof-of-principle' uitgevoerd met websites van enkele drinkwaterbedrijven. Dit is niet hier getoond, maar gedemonstreerd in een BTO workshop Text-Mining. Vragen zoals: 'is er informatie beschikbaar over aangetroffen chemicaliën?' of 'Wordt er informatie gegeven over innamestops?' kunnen hiermee voor een (grote) groep websites beantwoord worden, bijvoorbeeld door in de pagina's te zoeken op steekwoorden.

2.4 Wat heeft dit opgeleverd?

- Een script ('Twitter_schrapen.R', zie Bijlage VII) met voorbeeld voor het doorzoeken van Twitterberichten naar industriële activiteiten, met een opsomming van te gebruiken trefwoorden en hoe deze te combineren. Daarnaast een voorbeeld van het karakteriseren van bedrijven via een woordwolk met geassocieerde chemicaliën in Twitterberichten (paragraaf 2.1).
- In totaal zijn er 32 verschillende aanwijzingen voor nieuwe industriële activiteiten gevonden in het Rijnstroomgebied in Twitterberichten tussen 2018-2022 (paragraaf 2.1).
- Een voorbeeld script ('Webscraping_amtsblatter.R', zie Bijlage VII) om documenten van specifieke websites in een keer te downloaden ('schrapen') en te selecteren voor lezen, eventueel ook specifieke zinnen, op basis van het bevatten van een lijst met kernwoorden ('informatie selectie'). Een voorbeeld van stoffen en tekst die zoal in een 'Amtsblatter' te vinden zijn (paragraaf 2.2).
- Een overzicht van de websites waar 'Amtsblatter' worden gepubliceerd in Duitsland in relevante districten voor het Rijnstroomgebied (paragraaf 2.2).
- Een voorbeeld script ('Webscraping_websites.R', zie Bijlage VII) voor het analyseren van teksten op websites via xml-formaat. Uit deze activiteit bleek dat html- structuren in websites te divers zijn om op een standaard gestructureerde manier te doorzoeken (paragraaf 2.3).
- De 'RSS' feed op websites wordt tegenwoordig niet meer gebruikt en is daardoor geen goede bron van informatie (paragraaf 2.3).
- Een voorbeeld script ('Webscraping_websites.R', zie Bijlage VII) voor het verzamelen van teksten van willekeurige websites door te 'springen' tussen pagina's van de website. Er is voor de geselecteerde bedrijven verder geen inhoudelijke analyse gedaan omdat de geselecteerde bedrijven geen/weinig informatie over specifieke chemicaliën op hun website hadden staan (paragraaf 2.3).

3 Databases van organisaties voor toelating van stoffen

Het gebruik van chemische stoffen in Europa wordt geregistreerd in de REACH-verordening (registratie en beoordeling van en de autorisatie en beperkingen ten aanzien van chemische stoffen). De beschikbare informatie over deze chemische stoffen wordt geregistreerd in een centrale databank van het Europees Agentschap voor chemische stoffen (ECHA).[6] Het ECHA coördineert ook de evaluatie van deze chemische stoffen. Op een meer lokaal niveau zijn er in Nederland verschillende organisaties die het gebruik, de verkoop en/of de toelating van diverse stoffen registreren. Voorbeelden zijn het College voor de toelating van gewasbeschermingsmiddelen en biociden (Ctgb), het College ter Beoordeling van Geneesmiddelen (CBG) en de brancheorganisatie van de diergeneeskundige farmacie (FIDIN). Het bijhouden van nieuwe goedgekeurde stoffen levert informatie op over stoffen die in de toekomst van belang kunnen worden voor de behandeling van drinkwater. Het is echter niet doenlijk om al deze lijsten handmatig door te nemen. Daarom zijn enkele scripts (zie 'Webscraping_databases.R', Bijlage VII) ontwikkeld om op eenvoudige wijze informatie op te vragen over nieuw geregistreerde stoffen, geneesmiddelen en hun ingrediënten. In Tabel 7 worden enkele stoffen gegeven die op basis van de geschraapte informatie een toename in aanwezigheid in water zouden kunnen hebben. In Bijlagen II, III, IV en V staan alle namen van stoffen die volgens de geschraapte informatie duiden op nieuw gebruik en daarmee ook een nieuwe kans op aanwezigheid in water. Echter, niet alle nieuw geregistreerde gebruiken van stoffen zullen ook daadwerkelijk in gebruik komen op manieren waarbij emissie naar de waterketen mogelijk is.

3.1 ECHA website

ECHA voert de EU-wetgeving inzake chemische stoffen uit met het oog op de bescherming van het milieu en de menselijke gezondheid. De website van het ECHA bevat informatie over de verordening inzake biociden.[7] Een biocide moet in Nederland worden toegelaten voor gebruik in een product binnen een van de 22 onderscheiden producttypen voor biociden voordat het op de markt kan worden gebracht. De potentiële stof-producttype combinaties zijn te vinden op de ECHA website. Er is hier een databank beschikbaar waarin voor stof-producttype combinaties links naar factsheets met daarin verschillende regulatoire categorieën: met werkzame stoffen in het toetsingsprogramma, niet in het toetsingsprogramma en stoffen in Bijlage I van de verordening inzake biociden. Deze databank kan gebruikt worden om nieuwe stoffen en/of nieuwe toepassingen van stoffen te signaleren. Als een stof-producttype combinatie niet in het toetsingsprogramma is opgenomen dan is het mogelijk dat dit in de toekomst op de markt gaat komen. Ook als bij een stof is aangegeven dat het een 'Annex I' stof is, kan dit het geval zijn. Ze kunnen ook al in producten terechtkomen onder 'overgangsrecht'. Zoals gezegd is in Nederland nog wel een toelating nodig voor het gebruik in een bepaald product.

Eerst werd geprobeerd alle tabelpagina's te schrapen met behulp van een algemene URL met een enkel nummer voor elke tabelpagina, maar dit is niet gelukt omdat de URL niet op die structurele manier is opgebouwd. Als alternatief kan de databank manueel worden gedownload als CSV-bestand en geïmporteerd in R. De toelatingsstatus van de stoffen kan bestaan uit 'Approved', 'Not approved', 'No longer supported', 'Expired' of 'Cancelled Application'. De factsheets van alle stoffen zonder een toelatingsstatus 'Approved', werden gescreend via de website. Hier werd de 'Regulatory Categorisation' geëxtraheerd. De 'regulatory categorization' in deze factsheet kan bestaan uit 'Review programme substance', 'Substance not in Review Programme' of 'Annex I substance', waar de laatste een manuele controle van de factsheet vereist of de toelatingsstatus nog uitgebreider is of niet.

De ECHA Biocidal Active Substances database werd gedownload op 4 Augustus 2021 en 282 van de 906 stof-producttype combinaties waren 'Approved'.[8] Voor de overgebleven 624 stof-producttype combinaties werden de factsheets gescreend voor aanvullende informatie, resulterend in 559 'Review programme substances' en 63 'Substances not in Review Programme', waarvan één stof ook een 'Annex I substance' was. Voor twee stoffen werd geen categorie aangegeven. De resultaten werden weggeschreven naar een Excel file. Omdat stoffen vaker voorkomen in de database door hun toepassingen in de diverse producttypen is een filterstap op uniek BAS (Biocidal Active Substance) nummer noodzakelijk, resulterend in 32 relevante unieke werkzame stoffen. Relevante stoffen zijn de stoffen die géén 'Approved' status hebben en gecategoriseerd zijn als 'Substance not in Review Programme'. Deze stoffen zijn nieuw aangevraagd en worden mogelijk relevant in de toekomst. Ze kunnen ook al in toepassingen terechtkomen onder 'overgangsrecht'. Ze zijn weergegeven in Bijlage II. Bij wijze van voorbeeld is voor deze stoffen ingeschat wat de waarschijnlijkheid zal zijn dat deze stoffen daadwerkelijk aangetroffen zullen worden. Met behulp van EPI Suite™ werden de dampspanning en de biodegradeerbaarheid van de stoffen voorspeld.[9] Deze parameters kunnen inzicht geven in de waarschijnlijkheid dat de biocide daadwerkelijk in het aquatische milieu aangetroffen wordt. De informatie is bijgevoegd in Bijlage II. Uit deze analyse bleek dat 12 stoffen mogelijk in het water aangetroffen kunnen worden op basis van de dampspanning < 0,003 mm Hg..

3.2 CTGB

Het college voor de toelating van gewasbeschermingsmiddelen en biociden reguleert de toelating van gewasbeschermingsmiddelen en biociden in Nederland. Ze beschikken ook over een databank van toegelaten producten. Producten die niet in deze databank zijn opgenomen, kunnen Nederland wel bereiken via gebruik van middelen die zijn geregistreerd in de databank van ECHA. Deze middelen kunnen dan elders in de Europese Unie toegelaten zijn. Deze database is niet geanalyseerd omdat er geen informatie is over datums en daarmee over toekomstige of recente toelatingen. Naast de database worden er ook periodiek collegebesluiten gepubliceerd in de Staatscourant en in een pdf bestand op de website van de CTGB. Dit zijn besluiten over toelatingen, uitbreidingen, vernieuwingen en intrekkingen. Het automatisch downloaden en scrapen van deze pdf documenten is niet eenvoudig omdat de structuur van de documenten wisselt, evenals de publicatielocatie. Om die reden is dit niet gedaan voor deze bron.

3.3 MEB/CBG

Het College ter beoordeling van geneesmiddelen (CBG, in Engels: Medicines Evaluation Board (MEB)) heeft een databank met nieuw ingeschreven geneesmiddelen, maar deze webpagina is niet geschikt voor geautomatiseerde web scraping omdat de URL niet universeel is: hij bevat bijvoorbeeld zoekdata en het aantal vermeldingen in een tabel. Deze webpagina geeft slechts de resultaten van één maand weer. Een manier om deze informatie te schrapen is dus de URL handmatig te kopiëren, ervoor te zorgen dat alle resultaten op één pagina worden getoond, en deze in een R-script te plakken dat in staat is alle geneesmiddelen en hun bestanddelen op te sommen. Een voorbeeldoutput van unieke werkzame stoffen in nieuw ingeschreven geneesmiddelen in de periode van 28 juni 2021 tot 28 juli 2021 is weergegeven in Bijlage III.

Hetzelfde geldt voor diergeneesmiddelen, CBG heeft een databank met nieuw geregistreerde diergeneesmiddelen. Opnieuw laat de webpagina alleen de resultaten van één maand zien. De unieke bestanddelen in nieuw ingeschreven diergeneesmiddelen in de periode van 11 juli 2021 tot 11 augustus 2021 zijn weergegeven in Bijlage IV. Voor beide bijlages geldt dat het om alle bestanddelen van de nieuw ingeschreven geneesmiddelen gaat, hier zitten dus ook vulmiddelen e.d. bij. Expertkennis is vereist om onderscheid te kunnen maken tussen relevante en niet-relevante stoffen.

Voor zowel geneesmiddelen als diergeneesmiddelen geldt dat nieuw ingeschreven middelen maandelijks op deze manier verwerkt zouden moeten worden voor een actueel beeld. Over meerdere maanden zou dit een trend-analyse kunnen opleveren.

STOWA heeft in 2016 een rapport gepubliceerd over diergeneesmiddelen in relatie tot waterkwaliteit en zij constateerden dat er beperkte gegevens beschikbaar zijn over het gebruik van diergeneesmiddelen.[1011] Zo registreert FIDIN (de brancheorganisatie van de diergeneesmiddelenapotheek in Nederland) de nationale markt van diergeneesmiddelen, maar deze informatie is niet openbaar beschikbaar. De relatie tussen de nieuw ingeschreven middelen en hun daadwerkelijke gebruik is dus niet makkelijk te maken.

3.4 NVWA

De Nederlandse Voedsel- en Warenautoriteit bundelt per kalenderjaar de afzetgegevens van gewasbeschermingsmiddelen per werkzame stof in Nederland. Deze gegevens zijn afkomstig van de Nederlandse Stichting voor Fytofarmacie (Nefyto) en de Rijksdienst voor Ondernemend Nederland (RVO) en worden online gepubliceerd op de website van de NVWA. De informatie kan met behulp van een R-script automatisch worden gedownload, ingelezen uit pdf en gecombineerd. Vervolgens kunnen toenames en afnames gesignaleerd worden. In Bijlage V is het verschil tussen 2018 en 2019 weergegeven, waarbij toenames zijn gemarkeerd. Hierin zijn alleen werkzame stoffen meegenomen waarvan zowel in 2018 als in 2019 gegevens beschikbaar zijn.

Ook het CBS verzamelt gegevens over gewasbeschermingsmiddelen, het gaat hier over het gebruik per teeltgroep.[12] De gegevens zijn echter alleen beschikbaar (dd. februari 2022) op de website met open data van CBS (Statline) voor 2012 en 2016, wat de data niet recent en daarmee minder relevant maakt. Om die reden is er geen analyse gedaan op deze gegevens.

3.5 Overige databases

Andere databases die mogelijk relevant kunnen zijn voor de emissies van huishoudens zijn de EU cosmetische ingrediëntendatabank en de databank van voedseladditieven. Daarnaast is de aanwezigheid van microverontreinigingen in het in- en effluent van afvalwaterzuiveringen gerapporteerd in de Watson databank.[13] Ten slotte kan de website www.emissieregistratie.nl van de Rijksoverheid ook nog inzicht geven in toename of afname van de uitstoot van verontreinigende stoffen in Nederland. Deze databases zijn voor dit rapport niet geëvalueerd als bron voor het signaleren van nieuwe verontreinigingen.

Tabel 7. Enkele opvallende stoffen die naar voren kwamen uit de analyses op web-gebaseerde databases.

Stofnaam	Database	Bijzonderheid
Metofluthrin	ECHA database biociden	Slecht afbreekbaar en niet vluchtig. Nieuwe toepassingen mogelijk in 'PT19-Repellents and attractants'
Thiacloprid	ECHA database biociden	Slecht afbreekbaar en niet vluchtig. Nieuwe toepassingen mogelijk in 'PT08-Wood preservatives'
Monochloramine generated from ammonia and a chlorine source	ECHA database biociden	Meerdere nieuwe toepassingen mogelijk: 'PT05-Drinking water' 'PT06-Preservatives for products during storage' 'PT11-Preservatives for liquid-cooling and processing systems'
Indolylboterzuur	NVWA afzetgegevens gewasbeschermingsmiddelen	Meer dan 1000 % toename in afzetgegevens van 2018 naar 2019
Pyrethrinen	NVWA afzetgegevens gewasbeschermingsmiddelen	Meer dan 1000 % toename in afzetgegevens van 2018 naar 2019

Pelargonzuur	NVWA afzetgegevens gewasbeschermingsmiddelen	Meer dan 1000 % toename in afzetgegevens van 2018 naar 2019

3.6 Wat heeft dit opgeleverd?

- Voor een aantal web-gebaseerde databases is vastgesteld wat de opties zijn om informatie te 'schrapen' (paragraaf 3.1-3.5).
- Uit de diverse databases zijn lijsten van stoffen gekomen met een toegenomen of veranderde toepassing (paragraaf 3.1-3.5) (zie Bijlagen II, III, IV en V).
- De op een rij gezette opties om informatie uit databases te 'schrapen' (paragraaf 3.1-3.5) zijn een basis voor een toekomstige analyse waarbij verschillende bronnen tegelijk gedownload en vergeleken worden. Stofeigenschappen en toxiciteitsgegevens toevoegen zal het prioriteren van deze stoffen vergemakkelijken. Het ontbreken van CAS-nummers maakt het moeilijk/tijdrovend om dit te doen, ook omdat er bij de stofnamen geen gestandaardiseerde naamgeving wordt gebruikt. Bij een specifieke gebruiks-casus kunnen CAS-nummers voor dit doel toegevoegd worden.
- Een script ('Webscraping_databases.R', zie Bijlage VII) om de ECHA database te doorzoeken op stoffen die mogelijk een nieuwe toepassing krijgen, door de database informatie met informatie uit online factsheets te combineren (paragraaf 3.1).

4 Informatie in wetenschappelijke literatuur

Als er een bepaalde activiteit of nieuw proces geïdentificeerd is uit de diverse bronnen, is het nog niet altijd direct duidelijk wat dit betekent voor emissies van stoffen. Er kan meer informatie gezocht worden in andere tekstbronnen. Uit deze tekstbronnen kunnen bepaalde associaties worden gehaald met betrekking tot de geïdentificeerde processen en stoffen.

4.1 Chemicaliën herkennen in tekst

In Figuur 2 staat een voorbeeld resultaat van het vinden van bekende namen van chemische stoffen (uit de 'ChEBI' lijst) in tekst. Dit is daar gedaan voor stoffen die samen met een bepaalde bedrijfsnaam worden genoemd in Twitterberichten. Het nadeel van deze aanpak is dat, zodra een auteur zelfs maar een spatie of komma verandert, de stofnaam door de computer niet wordt herkend. Vooral bij ingewikkelde, lange namen, zoals die vaak voorkomen bij chemicaliën, is dat een probleem. Er zijn diverse oplossingen denkbaar om dit flexibeler te maken. In R bestaan functies voor 'fuzzy matching' waarbij enkele fouten zijn toegestaan. Hiervoor wordt een 'word distance' tussen woorden uitgerekend bijvoorbeeld 1,4-dichlorobenzene en 1,4-dichloorbenzene. Deze naam zou een afstand van 2 hebben doordat er twee letters verschil is. Dat zou genoeg kunnen zijn om te zeggen dat deze naam ook een stofnaam is. Deze techniek helpt echter niet als er een geheel andere schrijfwijze voor dezelfde stof wordt gehanteerd zoals 1,4-dichlorobenzene, para-Dichlorobenzene, p-DCB en Paramoth. Om meer flexibel chemische namen te herkennen zijn er geavanceerde methoden in TM die deze zogenaamde 'entity recognition' (oftewel concept-herkenning) doen. Hierbij staat niet het herkennen van specifieke chemische namen centraal, maar de herkenning dat het concept 'chemische stof' in de tekst staat. Met behulp van 'machine learning' of 'deep learning' maken deze methoden gebruik van de context aan woorden rond chemicaliën. Het maakt in dit geval niet uit hoe de schrijfwijze van de stofnaam is, als de context van de stofnaam maar duidt op de aanwezigheid van een stofnaam. Deze robuuste aanpak wordt gebruikt in het BTO project 'Zeer zorgwekkende stoffen' (in uitvoering).

In dit project passen we een simpelere aanpak toe. We gaan woorden scoren aan de hand van *onderdelen* van chemische namen die veel voorkomen. De onderdelen van chemische woorden bepalen we met hulp van de ChEBI lijst. Het voordeel van deze aanpak is de eenvoud, de stofnaamherkenning methode bestaat namelijk in de basis uit twee lijsten met onderdelen van woorden (stofnamen en gewone woorden). Het is daardoor zeer makkelijk om toe te passen in code. Een voorbeeld van onderdelen voor een willekeurige stof, "3-hydroxy-2-methyl", staat hieronder. Deze stofnaam bestaat uit de volgende onderdelen van elk 6 tekens lang:

Part1 "3-hydr"
Part2 "-hydro"
Part3 "hydrox"
Part4 "ydroxy"
Part5 "droxy-"
Part6 "roxy-2"
Part7 "oxy-2-"
Part8 "xy-2-m"
Part9 "y-2-me"
Part10 "-2-met"
Part11 "2-meth"
Part12 "-methy"
Part13 "methyl"

Als basis voor het extraheren van woorddelen van stofnamen nemen we 'KEGG-compound' namen uit de CHeBI lijst. We verwachten dat in abstracts niet vaak zeer ingewikkelde of lange IUPAC namen te vinden zullen zijn. Is dat wel de verwachting dan kan er van namenlijst gewisseld worden. Ter illustratie, de 10 meest voorkomende onderdelen van 6 tekens lang in alle namen van de KEGG-compound lijst zijn weergegeven in Error! Reference source not found..

Als basis voor het extraheren van 'gewone woorden' nemen we 10.000 abstracts die de kernwoorden 'biology' of 'biological' bevatten, om zo tot een brede selectie van abstracts in relatie tot de life-sciences te komen.

Woord-onderdelen die zeer weinig voorkomen (<2 of 3) worden verwijderd, een kortere lijst betekent namelijk een snellere analyse. Komt het onderdeel daarna nog relatief weinig voor, dan krijgt het een iets lagere weging. De totale weging wordt vervolgens genormaliseerd door per lijst te delen door de eigen som, en te vermenigvuldigen door de gemiddelde som van de twee lijsten. Op deze manier kunnen beide lijsten (onderdelen stofnamen en onderdelen gewone woorden) in theorie precies even hoge scores toekennen.

Per woord in de tekst wordt vervolgens gekeken of deze onderdelen van stofnamen bevat en dit wordt vergeleken met de score voor onderdelen uit gewone woorden. Korte stofnamen (zoals 'cumene', dat uit 6 letters bestaat) hebben een nadeel omdat er maximaal één woorddeel in past. Om deze ook te herkennen, gebruiken we in aanvulling op deze aanpak een vaste lijst van namen voor de korte namen (7 of minder letters). Wordt een woord herkend als stofnaam via deze route, krijgt de stofnaam een score ter waarde van 10.

In Tabel 8 staat een voorbeeld van een scoring van een willekeurig zinsdeel. Is een woord geen stof dan verwachten we en hoge score voor 'woord' en een lage score voor 'stofnaam'. Alleen als de score voor 'stofnaam' hoger is dan de score voor 'woord' krijgt het woord een eindscore voor stofnaam.

Tabel 8. Voorbeeld van scoring van woorden in een willekeuring zinsdeel met de stofnaamherkenner. De drie chemicaliën in de zin worden juist geclassificeerd.

Woord	Score stofnaam	Score woord	Eindscore stofnaam
a	0	0	0
novel	0	0	0
hydrophobic	1.57	4.31	0
benzalacetone	3.36	0.94	2.42
modified	0	2.35	0
lead	10	0	10
dioxide	0.9	0	0.9
electrode	0	3.14	0

Met deze gecombineerde aanpak is de nauwkeurigheid (gebaseerd op 10 random gekozen abstracts) in het identificeren van stofnamen 75%. De nauwkeurigheid van het identificeren van niet-stofnamen is 98%. De lijst met woorden die gebruikt zijn voor de validatie en hun classificatie zijn opgenomen in het TM Data-pakket (2022). Korte 'normale' woorden worden per definitie niet geclassificeerd omdat er geen woorddeel in past maar hiervan kan aangenomen worden dat dit veelal 'normale' woorden zijn omdat ze niet voorkomen in de uitgebreide CheBI lijst van korte stofnamen. Vooral lange stofnamen worden zeer goed onderscheiden van gewone woorden. Gecombineerde woorden ('ammonium-induced') worden minder goed herkend als stofnaam. Ook zorgt in sommige gevallen de schrijfwijze van de stofnaam ervoor dat het als méér dan een woord wordt gezien (bijvoorbeeld bij een spatie) en deze onderdelen worden dan apart gescoord. Dit is een probleem van 'woordherkenning' dat we hier niet kunnen oplossen.

Gebruiken we lijsten met kortere naam-onderdelen, bijvoorbeeld 5 karakters lang, dan worden de lijsten van unieke woordonderdelen korter. Dit is logisch omdat hier minder unieke combinaties van karakters mogelijk zijn. Er passen meer woordonderdelen in een stofnaam of gewoon woord, daardoor zijn de scores hoger dan bij 6 karakters. De relatieve scores ten opzichte van elkaar blijven wel ongeveer hetzelfde.

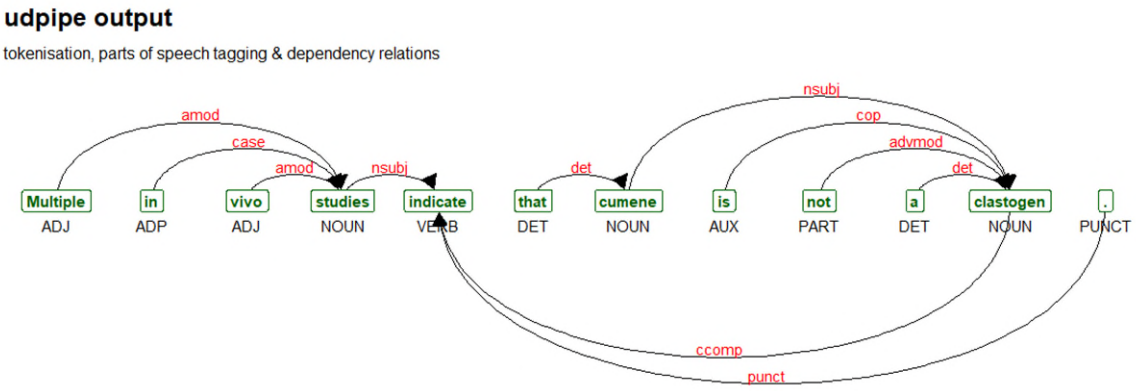
Met verdere optimalisatie kan deze methode nog verbeteren. Er kan een betere selectie uitgevoerd worden voor het corpus met de 'normale' woorden door bijvoorbeeld abstracts met trefwoorden 'chemical*', 'substance*' of 'compound*' juist niet te selecteren. Dit zal het aantal stofnamen verminderen en voorkomen dat er stukken stofnaam in de lijst van 'normale' woorden komen. Het sterretje in deze termen betekend overigens dat er een willekeurige letter(s) achter kan komen, dit is een vorm van gebruik van 'Regular expression' (zie Kader 1) door de zoekmachine van Pubmed.

Vanwege de exploratieve aard van dit onderzoek laten we het bij de twee gegenereerde lijsten met woordonderdelen van 6 karakters lang. Deze zijn te vinden in het TM Data-pakket (2022). De code om de woordonderdelen op een tijdsefficiënte manier te koppelen aan de worden in de tekst van interesse en de classificatie naar normaal woord of stofnaam staat in het 'Stofherkenner' script in het TM Data-pakket (2022).

4.2 Feiten vinden rond stoffen of processen in de vorm van 'triplets'

Teksten kunnen in hoog detail geanalyseerd worden met technieken uit de 'Natural language processing' (NLP). Met behulp van NLP kunnen bijvoorbeeld de functies van woorden (deze heten in NLP 'tokens') worden vastgesteld; is het woord, gezien de context, een werkwoord, een zelfstandig naamwoord, een bijwoord, iets anders? Dit heet 'Part-of-speech (POS) tagging'. Ook kunnen relaties tussen woorden worden afgeleid. Slaat een werkwoord op het zelfstandig naamwoord van interesse, of op een ander zelfstandig naamwoord in de zin? Dit

heet ‘Dependency parsing’. Ook kunnen stukken zin die bij elkaar horen worden geïdentificeerd. Zo kan er worden vastgesteld dat ‘de veelzijdige industriële stof cumeen’ woorden in de zin zijn die bij elkaar horen (‘Chunking’). Voor al deze taken zijn in de R-pakketten ‘OpenNLP’, ‘UDpipe’ en ‘SpacyR’ kant-en-klare functies beschikbaar die tabellen genereren waarin deze informatie per woord staat. Een visualisatie van het soort informatie staat in Figuur 5. In Bijlage VI staat een meer uitgebreide uitleg van NLP en de technieken die voor dit deel van de analyses in dit rapport zijn ingezet.



Figuur 4. Voorbeeld van de visualisatie van de zinsstructuur vastgesteld met het R-pakket ‘UDpipe’. De woorden (‘tokens’) in boxen zijn onderdeel van de geanalyseerde zin. Onder de woorden staan de labels voor woordsoort (‘Part of speech, POS’). De pijlen zijn de ‘dependency’ relaties tussen woorden, met labels voor het soort relatie. De dependency gaat altijd uit van een ‘ROOT’ werkwoord, dat is hier het werkwoord ‘indicate’. Een deel van de zin is het subject van ‘indicate’ (‘multiple in vivo studies’), en ander deel van de zin is het object van ‘indicate’ (‘cumene is not a clastrogen’). Daarom wijzen veel pijlen naar de ROOT. Voor uitleg van de labels wordt verwezen naar Bijlage VI.

Voor dit project willen we feiten extraheren in de vorm van ‘triplets’ van informatie rond een proces of stof van interesse. Het doel van ‘triplets’ is dat deze het makkelijker maken om de stof of het proces te duiden en te zien wat er zoal over deze stoffen en processen wordt geschreven. Een ‘triplet’ bestaat uit een zelfstandig naamwoord (de stof, of het proces), gevolgd door een werkwoord, gevolgd door nog een zelfstandig naamwoord. De stof of het proces kan ook het laatste zelfstandig naamwoord zijn. Voorbeelden van zulke feiten, bijvoorbeeld rond de stof cumeen, zijn (fictieve voorbeelden):

‘Cumeen’	‘is’	‘een industriële stof’
of		
‘Cumeen’	‘heeft’	‘geen mutagene werking’
of		
‘Benzeen’	‘produceert’	‘cumeen’

Met het R pakket ‘EasyPubmed’ kunnen abstracts van wetenschappelijke artikelen met R doorzocht worden via de zoekmachine Pubmed. Hieruit kunnen abstracts geselecteerd worden die een aantal steekwoorden bevatten. Deze abstracts zijn de basis voor de triplet extractie. De steekwoorden voor het selecteren van de abstracts zijn het proces of de stof (met synoniemen) gecombineerd met een aantal synoniemen voor ‘maken’ en ‘industrieel’:

“produ*[Title/Abstract] OR make[Title/Abstract] OR synthes*[Title/Abstract] OR manufact*[Title/Abstract] OR industr*[Title/Abstract] OR procedure[Title/Abstract] OR process[Title/Abstract]"

Bij het genereren van de triplets komen we een aantal moeilijkheden tegen die binnen de looptijd van dit project niet allemaal opgelost kunnen worden. De moeilijkheden hebben te maken met de manier waarop feiten worden gepresenteerd in de teksten. Met 'chunking' kunnen individuele woorden zoals 'industriële stof' en 'mutagene werking' vrij makkelijk bij elkaar gehaald worden. De gegenereerde tabel (zie Bijlage VI) geeft namelijk aan bij welk woord de Chunk begint en bij welk woord het eindigt. De ontkenning ('geen') is altijd geannoteerd met een speciale code, die kunnen we ook herkennen en bijvoegen.

Een moeilijkheid zit in de specificaties van de feiten. Het feit 'geen mutagene werking' kan bijvoorbeeld slaan op speciale omstandigheden, 'geen mutagene werking bij applicatie op de huid'. Deze specificaties kunnen we toevoegen door alle woorden met een label dat woorden aanduidt zoals 'in', 'bij', 'van', 'met', 'over' ook bij de Chunk te voegen. Dit noemen we een 'verlengde chunk'. Soms staan de specificaties echter verspreidt over verschillende zinnen en wordt het in de zin van interesse gerefereerd als 'dat' of 'deze'. In Tabel 9 staan een aantal goed interpreteerbare triplets die per enkele zin in de volgorde 'verlengde chunk zelfstandig naamwoord' 'verlengde chunk werkwoord' 'verlengde chunk zelfstandig naamwoord' zijn geëxtraheerd, en een aantal minder goede. Dit geeft een beeld van de (on-) mogelijkheden bij het genereren van de triplets. Bij het verlengen van de Chunks wordt af en toe het link-woord (zoals 'to' of 'from') bij zowel het werkwoord-Chunk als het zelfstandig naamwoord-Chunk getrokken, dat zal in vervolgonderzoek ook hersteld moeten worden.

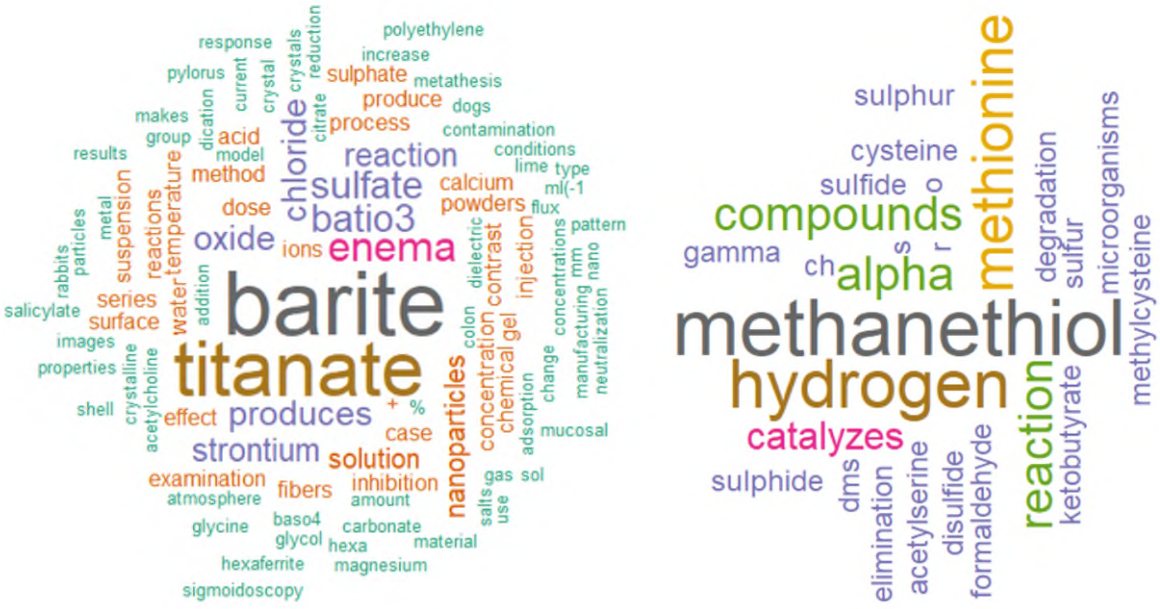
Verdere selectie op trefwoorden 'produ|make|synthes|manufact|process|fabricate|catal|generate|utili' die aanwezig zijn in het werkwoord-deel van de triplets resulteert in meer relevante triplets voor het productieproces. N.B. Als men juist selecteert op trefwoorden die met toxiciteit te maken hebben, zullen de triplets relevantie hebben voor dát onderwerp.

Tabel 9. Zes voorbeelden van uit abstracts geëxtraheerde triplets voor Bariet ('Barite'). Drie informatieve en drie niet-informatieve uit totaal 129 triplets (uit 5079 abstracts).

Zelfstandig naamwoord: verlengde Chunk	Werkwoord: verlengde Chunk	Zelfstandig naamwoord: verlengde Chunk	Informatief?
neutralization with barite hydroxide	produced	insoluble salt	Ja, informatie over een productieproces
magnetic separably barite ferrite nanomaterial (bafeo)	was synthesized via	via citrate acid	Ja, informatie over een productieproces
selective homo- and cross-desilacoupling of aryl and benzyl primary silanes	catalyzed by	by a barite complex	Ja, informatie over een productieproces
which	are controllably synthesized by tuning	the amount of barite precursor	Nee, het is onduidelijk wat het eerste zelfstandig naamwoord is.
splenic focal defect	produced by	by barite in the colon	Nee, het triplet gaat over een medische conditie
from a copolymer of polypropylene and polyethylene with barite sulfate	to make	it	Nee, het triplet is incompleet

Het R-script kan met een stof van interesse plus eventueel een of twee synoniemen triplets uit abstracts extraheren, zoals de weergegeven voorbeelden in Tabel 9. De triplets bevatten op dit moment voor een deel niet-relevante triplets, zodat deze handmatig verder moeten worden geëvalueerd voor het identificeren van nuttige informatie. Dit zal wel minder tijd kosten dan alle abstracts te moeten lezen. Op dit moment worden voor de combinatie bariet en trefwoorden rond productie 5079 abstracts teruggebracht tot 129 triplets. De huidige procedure voor triplet extractie kan dienen als een basis voor verdere ontwikkeling van triplet extractie.

De informatie in de geselecteerde triplets kan ook weer omgevormd worden tot een woordwolk, wat op een meer associatieve manier het productieproces kan weergeven. De woordwolk van bariet en die van methanethiol staan in Figuur 5. Dit zijn stoffen die geassocieerd zijn aan twee van de geïdentificeerde nieuwe activiteiten in de case studie rond het Rijnstroomgebied, in Tabel 2. Voor de woordwolken zijn alleen woorden toegelaten die relatief vaak voorkomen. Hiervoor is het 95 percentiel gebruikt. Dit lijkt een hoge selectiegrens maar de meeste worden in de triplets komen slechts éénmaal voor. Hierdoor selecteert het 95 percentiel de onderscheidende, veel voorkomende woorden. Op basis van de woordwolken kan een hypothese gemaakt worden rond welke stoffen geassocieerd zijn met nieuwe activiteiten in de case studie rond het Rijnstroomgebied. Deze stoffen zijn mogelijk ook interessant om te monitoren. Op deze manier kan rond een nieuwe activiteit een lijst met potentiële geassocieerde stoffen gemaakt worden op informatie uit abstracts uit wetenschappelijke literatuur.



Figuur 5. Woordwolk van stoffen uit geëxtraheerde triplets van zelfstandig naamwoord, werkwoord, en zelfstandig naamwoord waarbij een van de zelfstandige naamworden het woord 'barite' of 'methanethiol' bevat en het werkwoord een van de woorddelen 'produ|make|synthes|manufact|process|fabricate|catal|generate|utili' bevat. De woordwolk van bariet reflecteert ook de medische toepassing van bariet.

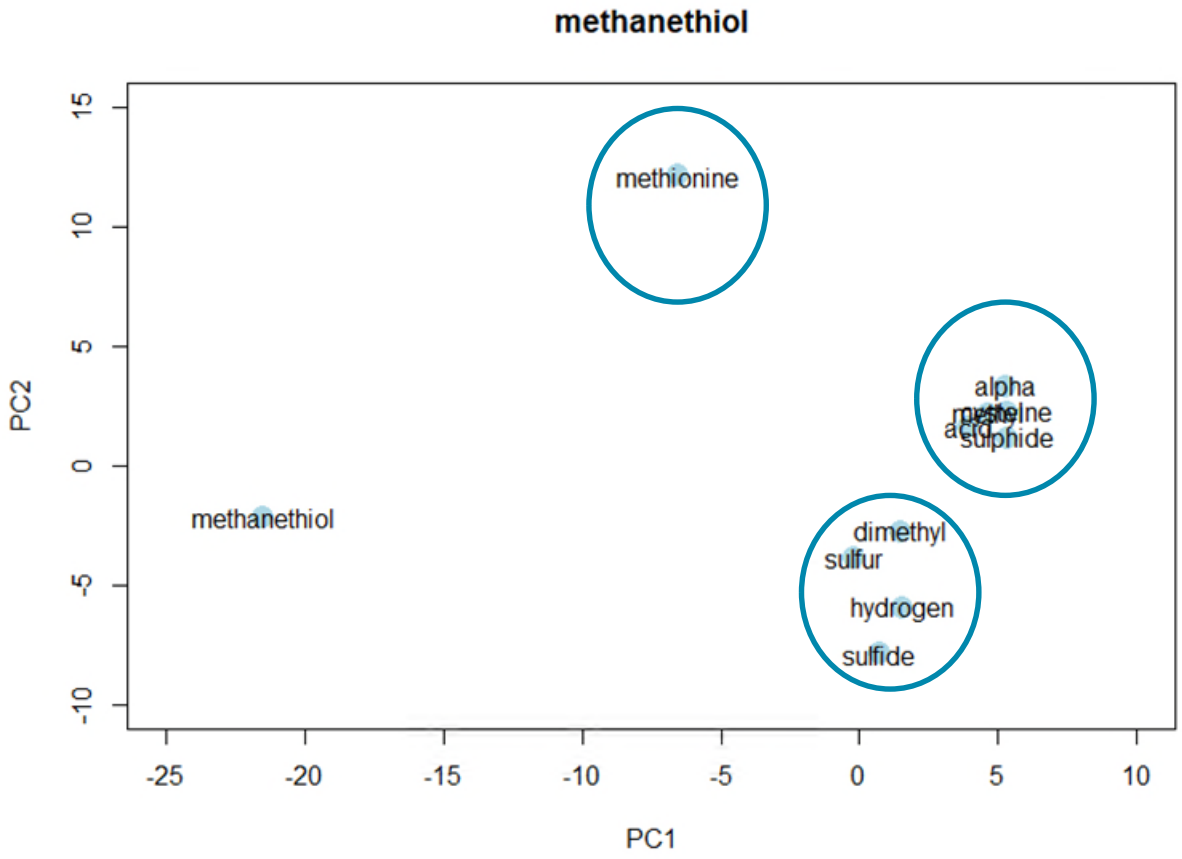
Een extra kwaliteitsslag op de triplets kan behaald worden met het includeren van de 'Dependency parsing' labels. Daarmee kan dan zeker worden gesteld dat een werkwoord echt slaat op een zelfstandig naamwoord in de triplet (als 'object' of als 'subject' van het werkwoord. Dit wordt echter bemoeilijkt door een aantal dingen. Deze labels staan soms niet direct bij het werkwoord, als het bijvoorbeeld een opsomming betreft: 'cumene, benzene, and methane'. Zo'n combinatie wordt aangeduid met het label 'compound' wat een samenstelling of groepje aangeeft. Een andere complicatie is dat het triplet niet altijd de ROOT bevat (zie ook Figuur 4). Dat betekent dat er geen duidelijk object en subject rond het werkwoord van de sub-zin zijn (zie onderschrift Figuur 4). Een optie is om in zulke gevallen alleen de geïdentificeerde sub-zin in een script opnieuw te laten annoteren door de NLP functies in het script. Dit is desgewenst voor de verdere ontwikkeling van deze techniek te implementeren.

4.3 Associaties vinden met andere stoffen: groepjes van chemicaliën

Om nog meer inzicht te krijgen rond mogelijk relevante stoffen als er een nieuw proces of stofgebruik wordt verwacht is het ook mogelijk om direct uit de tekst specifiek alleen stoffen te identificeren die samen met de stof of het proces van interesse genoemd worden. Hiervoor is het niet per se nodig om NLP te gebruiken. De tekst kan ook

als een verzameling woorden worden beschouwd. Dit wordt ook wel de 'bag of words' aanpak genoemd. Voor deze aanpak is nodig dat we namen van chemicaliën goed kunnen herkennen, en deze statistisch aan elkaar kunnen linken.

De stofnamen in abstracts die het woord of proces van interesse bevatten worden voor deze analyse herkend met de methode zoals beschreven in paragraaf 4.1. Deze worden in een 'term-document matrix' gezet. Dit is in feite een grote tabel en bevat voor alle stofnamen (rijen) in alle abstracts (kolommen) hoe vaak een stofnaam voorkomt. De stofnamen en hun frequentie kunnen wederom input zijn voor een woord-wolk. Maar op de term-document matrix kan ook een clustering worden toegepast waarmee ook eventuele subgroepen van stoffen kunnen worden gedetecteerd. Dit kan aanleiding zijn om de samenhang van de stoffen in de clusters te onderzoeken op relevantie voor emissie naar het watersysteem. In Figuur 6 staat een voorbeeld van een clustering gemaakt via een 'principal component analysis' (PCA) op de chemische woorden die genoemd zijn in abstracts waarin het woord of proces van interesse ook staat (hier: methanethiol), samen met een trefwoord dat een link met productie aangeeft 'produ|make|synthes|manufact|process|fabricate|catal|generate|utili'. Methanethiol is een van de stoffen die is geassocieerd met nieuwe activiteiten in de case studie rond het Rijnstroomgebied in Tabel 2.



Figuur 6. PCA analyse van de frequentie van stofnamen in de 'term-document' matrix. Term is hier de stofnaam, document is hier een abstract van een wetenschappelijk artikel. Clusters worden in de tekst van dit rapport toegelicht.

Een korte zoektocht op internet rond de namen in clusters van Figuur 6 levert op dat het groepje chemicaliën data bestaat uit hydrogen, sulfide, sulfur, methanethiol en dimethyl te maken kunnen hebben met het maken van wijn.[14] Methionine is een moederstof van methanethiol.[15] Het is niet helemaal duidelijk waar het groepje alpha, sulphide, acid, methyl cysteine voor staat. Mogelijk staat dit groepje woorden voor een enzymatische reactie in het productieproces. Deze techniek levert niet voor alle chemicaliën of productieprocessen diverse groepen van

chemicaliën op. Een van de oorzaken is de lage frequentie van sommige genoemde geassocieerde chemicaliën. Hierdoor worden chemicaliën die laag-frequent in verschillende context in een tekst worden genoemd toch in een en hetzelfde cluster gestopt. Ook hier is verbetering mogelijk, met name op het gebied van stofnaamherkenning. Stofnamen die uit twee delen bestaan worden bijvoorbeeld in de term-document matrix nu niet herkend als een stofnaam, maar worden opgesplitst. Met NLP chunking (zie Bijlage VI) kunnen deze delen worden herkend als 'noun-chunk'.

4.4 Wat heeft dit opgeleverd?

- Een conceptversie van een simpele classificeerder voor het herkennen van chemische namen ten opzichte van 'normale' woorden (in het script 'Stofherkenner.R', zie Bijlage VII). De classificeerder werkt op basis van lijsten met woorddelen (van chemicaliën vs. 'gewone woorden') die vergeleken worden met de woorden in tekst (paragraaf 4.1). De methode is gevoelig voor het aantal woorddelen. Een 'Deep Learning' aanpak om het concept chemische stof te herkennen is waarschijnlijk nauwkeuriger, maar ook ingewikkelder om te implementeren.
- Kennis over de toepasbaarheid en limitaties van NLP technieken (paragraaf 4.2-4.3).
- Basis-code voor het annoteren van tekst met NLP technieken via drie verschillende R-pakketten (openNLP, UDpipe, SpacyR) en een beschrijving van NLP technieken (in het script 'NLPpakketten.R', zie Bijlage VII).
- Code om 'chunks' te extraheren uit NLP-geannoteerde teksten (in het script 'NLPchunks.R', zie Bijlage VII). Dit is toegepast om 'noun-phrases' en 'verb-phrases' te extraheren uit teksten, om werkwoorden en zelfstandige naamwoorden meer context te geven bij het selecteren (paragraaf 4.2).
- Basis-code om 'triplets' te maken om zo alle informatie rond een bepaald zelfstandig naamwoord (Chunk) te extraheren. Triplets wil zeggen, een Chunk rond een zelfstandig naamwoord en een ander zelfstandig naamwoorden in dezelfde zin, gekoppeld via een werkwoord (in het script 'NLPchunks.R', zie Bijlage VII). Hierbij zit ook een optie om de geassocieerde zelfstandig naamwoorden uit de triplets in een woordwolk weer te geven.
- Een voorbeeld en code ('StofnaamClusters.R', zie Bijlage VII) om statistische associaties te vinden tussen chemicaliën in verzamelingen van teksten (bijvoorbeeld abstracts) en zo subgroepen te identificeren van chemicaliën in relatie tot de stof van interesse (paragraaf 4.3). Verdere ontwikkeling op basis van een concrete onderzoeksvraag is nodig om resultaten in meer gevallen bruikbaar te maken.

5 Samenvatting en conclusie

Voor een vroege signalering van potentiële vervuiling door stoffen zijn in dit project een aantal text-mining technieken toegepast. Dit heeft inzicht opgeleverd over het gebruik van de technieken zelf, en de opbrengst die de technieken in combinatie met de gekozen bron opleveren. Voor een aantal technieken – bron combinaties zal, om relevante waterverontreinigingen goed in beeld te krijgen, een grotere inspanning nodig zijn dan in dit project geleverd kan worden. Dit zal pas de moeite waard zijn bij concrete onderzoeksvragen. Er kan bijvoorbeeld een tool gemaakt worden voor het genereren van een lijst met stofnamen rond een industriële activiteit van interesse. Of, een onderzoeksvraag kan zijn wat de geschraapte informatie uit gecombineerde databases voor nadere inspectie-opties oplevert. Als er een concrete toepassing is bedacht voor de zeer simpele techniek van stofnaamherkenning op basis van woorddelen kan dit verder geoptimaliseerd worden. De text-mining technieken kunnen ook ingezet worden om, voor een specifieke stof van interesse, alle feiten (in het Engels 'assertions', in de vorm van 'triplets') in tekst te achterhalen of hier een tool voor te maken. Voor dit exploratieve project is gekozen voor het in kaart brengen van zulke opties.

'Web-scraping' van (pdf-)documenten, zoals vergunningen van het internet, vergemakkelijkt het verzamelen van documenten waarin aanwijzingen kunnen staan voor nieuwe activiteiten en daarbij horende stoffen in het gebied van interesse. Het zoeken op sleutelwoorden is vervolgens nuttig om te bepalen welke documenten de moeite waard zijn om te lezen. Ook kunnen zinnen met mogelijk interessante informatie uitgelicht en op een rij gezet worden, voor nog minder leeswerk. De meeste van deze documenten zullen, vermoedelijk, stoffen rapporteren die reeds gereguleerd zijn. Het kan wel nieuwe informatie opleveren m.b.t. nieuwe toepassingen of nieuwe activiteiten op nieuwe locaties. Het gebruikte voorbeeld van het scrapen van de Amtsblätter kan bijvoorbeeld gebruikt worden door RIWA-Rijn, om eventuele indicaties voor nieuwe activiteiten langs de Rijn tijdig te signaleren.

Het extraheren van relevante web-teksten bleek een uitdaging, vanwege de niet gestandaardiseerde opzet van pagina's van verschillende websites. Het is te verwachten dat veel moeite moet worden gestopt in het verwerken van deze verschillende structuren van websites, en dat dit daarvoor niet genoeg informatie zal opleveren. Het doorzoeken van RSS-feeds van bedrijven leverde ook weinig op, dit kanaal wordt niet vaak (meer) ingezet. De verwachting is ook dat een deel van de nieuwsberichten via Twitter gedeeld wordt, de mediastrategie en het gebruik van sociale media door bedrijven kan echter wel verschillen per bedrijf. Twitter is eenvoudiger gestructureerd te doorzoeken.

Het verzamelen van relevante Twitterberichten leverde een aantal interessante tweets op die kunnen wijzen op veranderde activiteit in het stroomgebied van de Rijn. Uit 50.000 tweets over en van bedrijven werden rond de 500 tweets geselecteerd, waarvan 16% de moeite waard bleek om op te volgen. In de tweets stonden ook een aantal namen van stoffen, die gebruikt kunnen worden om een idee te krijgen van de chemische activiteit van het bedrijf. Omdat het bedrijf echter ook filialen in andere gebieden kan hebben zijn deze stoffen niet altijd locatie-specifiek, en dus niet direct terug te voeren op het Rijstroomgebied, dat is geselecteerd als case studie.

Het web-scrapen van data in de vorm van o.a. tabellen leverde heel concrete lijsten van stoffen op. Deze stoffen kunnen mogelijk meer in de waterketen terecht door veranderd gebruik of veranderingen in toepassingen. Dit was mogelijk voor pesticiden, biociden, geneesmiddelen en diergeneesmiddelen. De stoffen hoeven overigens niet perse 'nieuw' te zijn. Een aanwijzing tot, in potentie, meer emissie is ook relevant. Helaas hebben veel databases geen standaard ID (unieke naam of codering) maar stofnamen met een breed scala aan schrijfwijzen en in diverse talen. Voordat een verdere analyse rond het lot van de stof in de omgeving kan worden bepaald met behulp van stofeigenschappen, zal eerst een standaard ID zoals een CAS nummer moeten worden achterhaald. Voor biociden

was er wel een CAS nummer beschikbaar. Ook hier kan een concrete onderzoeksvraag een specifieke applicatie opleveren die gebruik kan maken van deze techniek.

Het gedetailleerd doorzoeken van teksten en daar gericht relevante informatie uit halen met behulp van 'natural language processing' (NLP) heeft potentieel. Het is vrij eenvoudig om met beschikbare software pakketten in R (overigens ook in Python) de tekst te annoteren met 'part of speech' oftewel woordsoort (bijwoord, werkwoord, etc.) en 'dependency' oftewel relatie tot elkaar (bijvoeglijk naamwoord van woord x, onderwerp van werkwoord y, etc.). Het extraheren van feiten over een bepaald onderwerp in de vorm van 'triplets' van zelfstandig naamwoord – werkwoord - zelfstandig naamwoord, wordt echter bemoeilijkt door complex taalgebruik. Verdere ontwikkeling van de methodologie zal nodig zijn om het aantal niet-nuttige triplets te verkleinen. Dit zal voornamelijk zijn in de vorm van extra 'regels' die bepalen wanneer iets een goede triplet is, maar ook selectie op specifieke typen van relaties (bijvoorbeeld woorden die duiden op een productieproces). Een geheel andere aanpak is het inzetten van 'Deep-learning' om bepaalde relaties te herkennen. Met deze techniek is hier niet geëxperimenteerd.

Het herkennen van een aantal typen concepten ('entities') zoals locatie, persoon, datum in teksten kan al met de huidige gebruikte functionaliteit in de R pakketten die gebruikt zijn. Chemische stoffen als concept herkennen zit daar niet bij. Voor dit onderzoek is een simpele classificator ontworpen die gebruik maakt van lettercombinaties die vaak in namen van chemische stoffen voorkomen. Omdat in de basis de classificator bestaat uit een lijst veel voorkomende lettercombinaties in namen van chemicaliën en een lijst van lettercombinaties in woorden die geen chemicaliën aanduiden is het makkelijk deze in te passen in een ander script. Met deze methode was het mogelijk om meer chemische namen in teksten te herkennen dan via een vaste lijst met namen zoals in ChEBi. In een analyse zijn stoffen die genoemd worden in abstracts via statistiek aan elkaar verbonden tot groepjes van gerelateerde chemicaliën. Dit geeft concrete groepjes stoffen om nader onder de loep te nemen bij een aanwijzing van verandering in productieproces in een stroomgebied van interesse, bijvoorbeeld zoals geïdentificeerd via een vergunning of bericht op social media.

Samenvattend zijn er verschillende technieken verkend met verschillende toepassingen. Het is goed om te benadrukken dat er altijd expertkennis nodig is om de resultaten te beoordelen op relevantie, bruikbaarheid en in hoeverre het nieuwe informatie is. Door de uniforme eigenschappen van de technieken en de mogelijkheden om deze te optimaliseren per casus, is text-mining op vele vlakken toepasbaar. Het is voornamelijk van toegevoegde waarde voor onderzoeksvragen waarbij grote hoeveelheden tekst en/of online bronnen verzameld en doorzocht moeten worden. Concrete onderzoeksvragen zijn dan ook nodig om de technieken verder nuttig in te kunnen zetten. Hierbij kunnen ook gespecialiseerde onderzoeksbureaus of bedrijven ingezet worden.

6 Literatuurlijst

1. Hartmann, J.; Wuijts, S.; van der Hoek, J. P., Use of literature mining for early identification of emerging contaminants in freshwater resources. *Environmental Evidence* **2019,** *8* (33).
2. Sjerps, R.; Puijker, L.; Brandt, A.; ter Laak, T. *Signaleren van nieuwe stoffen (2014-2015)*; BTO 2015.059; KWR Water Research Institute: Nieuwegein, Nederland, 2015; p 51.
3. Landesamt für Natur Umwelt und Verbraucherschutz Nordrhein-Westfalen *Entwicklung und Stand der Abwasserbeseitigung in Nordrhein-Westfalen*; https://www.lanuv.nrw.de/umwelt/wasser/abwasser/lagebericht, 31 December 2020, 2020.
4. Ministerium für Umwelt Landwirtschaft Natur- und Verbraucherschutz des Landes Nordrhein-Westfalen Flussgebiete NRW. https://www.flussgebiete.nrw.de/node/311 (accessed 19 April 2022).
5. Twitter Developer terms - More about restricted uses of the Twitter APIs. https://developer.twitter.com/en/developer-terms/more-on-restricted-use-cases (accessed 19 April 2022).
6. European Commission REACH. https://ec.europa.eu/environment/chemicals/reach/reach_en.htm (accessed 26 April 2022).
7. European Chemicals Agency Information on biocides. https://echa.europa.eu/nl/information-on-chemicals/biocidal-active-substances (accessed 4 August 2021).
8. European Chemicals Agency, Active Substances. 16 July 2021 ed.; https://echa.europa.eu/nl/information-on-chemicals/biocidal-active-substances, 2021.
9. US EPA *Estimation Programs Interface Suite™ for Microsoft® Windows*, v 4.11; United States Environmental Protection Agency: Washington, DC, USA, 2021.
10. Rougoor, C. W.; Allema, A. B.; Leendertse, P. C.; van Vliet, J. *Diergeneesmiddelen en waterkwaliteit. Een verkenning van stoffen, gebruik en effecten op waterkwaliteit*; 26; Stichting Toegepast Onderzoek Waterbeheer,: 2016; p 70.
11. ter Laak, T.; Sjerps, R.; Kools, S. *Quick-scan Diergeneesmiddelen in de waterketen*; KWR 2017.037; KWR Water Research Institute: Nieuwegein, Nederland, 2017; p 49.
12. CBS Statline Gebruik gewasbeschermingsmiddelen in de landbouw; werkzame stof, toepassing. https://opendata.cbs.nl/statline/?dl=42374&ts=1602052039000#/CBS/nl/dataset/84010NED/table (accessed Februari 2022).
13. Roex, E.; van Duijnhoven, N.; van der Meiracker, R.; van Gils, J.; Derksen, A., De Watson-database brengt emissieroutes van microverontreinigingen in water beter in beeld. *Water Matters* **2020,** *Juni 2020*.
14. Viviers, M. Z.; Smith, M. E.; Wilkes, E.; Smith, P., Effects of Five Metals on the Evolution of Hydrogen Sulfide, Methanethiol and Dimethyl Sulfide during Anaerobic Storage of Chardonnay and Shiraz Wines. *Journal of Agricultural and Food Chemistry* **2013,** *61* (50), 12385-12396.
15. Philipp, T. M.; Will, A.; Richter, H.; Winterhalter, P. R.; Pohnert, G.; Steinbrenner, H.; Klotz, L. O., A coupled enzyme assay for detection of selenium-binding protein 1 (SELENBP1) methanethiol oxidase (MTO) activity in mature enterocytes. *Redox Biol* **2021,** *43*, 101972.
16. LUMITOS AG Chemikalienliste. https://www.chemie.de/lexikon/Chemikalienliste.html (accessed 16 March 2021).
17. Royal Society of Chemistry (13Z)-13-Hexadecen-11-yn-1-yl acetate. http://www.chemspider.com/Chemical-Structure.10300419.html?rid=da230127-bf03-4656-b9c3-cd86ea878d64 (accessed 15 October 2021).
18. Royal Society of Chemistry 2-Methyl-1,2-benzothiazol-3(2H)-one http://www.chemspider.com/Chemical-Structure.311863.html?rid=04ef8556-8c89-435a-8c58-a64d74befc6e (accessed 15 October 2021).
19. TM Data-pakket Sharepoint (2022). Benaderbaar voor BTO-partners. Voor andere geinteresseerden, beschikbaar op verzoek : https://teams.microsoft.com/l/team/19%3alrJ4vKFmkcx4gHW_qChl7mzZkamxi3tCFXM2mPpqgC01%40thread.tacv2/conversations?groupId=118891e7-3e7a-4936-98b7-e3e16aa4877a&tenantId=a8294dac-16db-4321-836a-c3caa9c7a8a6
20. Pronk, T.E.; Roessink, I; Smit, E. (2022). MEETSTRATEGIE BIOCIDEN, Overwegingen en criteria. Rapport KIWK 2022-07. Meetstrategie Biociden – overwegingen en criteria (KIWK) | STOWA

I Relevante woorden

Het verslag over de afvoer van afvalwater geschreven door het Ministerium für Umwelt, Landwirtschaft, Natur- und Verbraucherschutz des Landes Nordrhein-Westfalen in Duitsland[3] werd doorzocht voor Duitse woorden gerelateerd aan afvalwater, afvalwaterlozing, reguleringen rondom afvalwater en verontreinigingen. Een lijst met Duitse namen van 1314 chemicaliën werd verzameld van de Duitse lexicon uitgegeven door chemie.de.[16]

Duits	Engels	Nederlands
Direkteinleitungen	Direct discharges	Directe lozing
Indirekteinleitungen	Indirect discharges	Indirecte lozing
Abwasser	Sewage	Afvalwater
Abwasservorbehandlung	Wastewater pre-treatment	Voorbehandeling van afvalwater
Kläranlagen	Sewage treatment plants	Rioolwaterzuiveringsinstallatie
Abwasserverordnung	Wastewater ordinance	Verordening inzake afvalwater
Abwasserabgabengesetz	Wastewater tax act	Wet op de afvalwaterheffing
Abwassermenge	Amount of wastewater	Hoeveelheid afvalwater
Abwassereinleitungen	Sewage discharge	Lozing afvalwater
Einleitung	Introduction	Afvoer
Oberflächengewässer	Surface water	Oppervlaktewater
Verschmutzung	Pollution	Verontreiniging
Kanalisation	Sewerage	Riolering
Schmutzwasser	Dirty water	Afvalwater
Abwasserreinigung	Wastewater treatment	Behandeling van afvalwater
Abwasserbehandlungsanlage	Wastewater treatment plant	Afvalwaterzuiveringsinstallatie
Wasserhaushaltsgesetz	Federal Water Act	Federale Waterwet
WHG		WHG
Wasserahmenrichtlinie	Water Framework Directive	Kaderrichtlijn Water
WRRL	WFD	WFD
Umweltqualitätsnormrichtlinie	Environmental Quality Standards Directive	Richtlijn inzake milieukwaliteitsnormen
UQN-RL	EQSs	EQS richtlijn
Industrieemissionsrichtlinie	Industrial Emissions Directive	Richtlijn Industriële Emissies
IE-RL		IE-RL
Richtlinie 2013/39/EG zur Änderung der Richtlinien 2000/60/EG und 2008/105/EG		
Verordnung 166/2006/EG über die Schaffung eines Europäischen Schadstoff-freisetzungs- und verbringungs registers		
PRTR		
Kommunalabwasserrichtlinie	Urban Waste Water Directive	Richtlijn stedelijk afvalwater
Verordnung (EU) 2016/293 DER KOMMISSION zur Änderung der Verordnung (EG) Nr. 850/2004 des Europäischen Parlaments und des Rates über		

persistente organische Schadstoffe hinsichtlich des Anhangs I		
VERORDNUNG (EU) 2017/852 DES EUROPÄISCHEN PARLAMENTS UND DES RATES vom 17. Mai 2017 über Quecksilber und zur Aufhebung der Verordnung (EG) Nr. 1102/2008		
Bundes-Immissionsschutzgesetz	Federal Immission Control Act	Federale Immissie Controle Wet
BImSchG		
AbwAG		
Gesetz zur Ausführung des Protokolls über Schadstofffreisetzungs- und -verbringungsregister	Act implementing the Protocol on Pollutant Release and Transfer Registers	Akte betreffende de uitvoering van het protocol betreffende registers inzake de uitstoot en overbrenging van verontreinigende stoffen
SchadRegProtAG		
AbwV		
Oberflächengewässerverordnung	Surface Waters Ordinance	Verordening oppervlaktewateren
OGewV		
Verordnung über Anlagen zum Umgang mit wassergefährdenden Stoffen		Verordening inzake installaties voor het hanteren van stoffen die gevaarlijk zijn voor water
WasgefStAnlV		
Industriekläranlagen-Zulassungs- und Überwachungsverordnung		Verordening inzake de goedkeuring van en het toezicht op industriële rioolwaterzuiveringsinstallaties
IZÜV		
Landeswassergesetz	State water law	Staatswaterwet
LGW		
Rechtsverordnung über die Freistellung von Abwasserbehandlungsanlagen von der Genehmigungspflicht		Verordening inzake de vrijstelling van waterzuiveringsinstallaties van de vergunningsplicht
FreistVO		Vrijstellingsverordening
Emissionserklärungsverordnung		Verordening emissieverklaring
Verordnung zur Umsetzung der Kommunalabwasserrichtlinie		Verordening inzake de tenuitvoerlegging van de richtlijn stedelijk afvalwater
KomAbwV		
Verordnung über Art und Häufigkeit der Selbtstüberwachung von Abwasserbehandlungsanlagen und Abwassereinleitungen		Verordening inzake type en frequentie van zelfcontrole van installaties voor de behandeling van afvalwater en de lozing van afvalwater
SüwV-kom		
Verwaltungsvorschriften zum Vollzug der Verordnung über Anlagen zum Umgang mit wassergefährdende		Administratieve voorschriften inzake de handhaving van de verordening betreffende installaties voor het hanteren van met water gevaarlijke stoffen en betreffende gespecialiseerde bedrijven

Stoffen and über Fachbetriebe		
VV-VAwS		
Durchführungs- und Verwaltungsvorschriften		Uitvoerings- en administratieve regelgeving
Satzungen von Städten, Gemeinden und Abwasserverbanden		Statuten van steden, gemeenten en afvalwaterverenigingen
Wassergefährdenden	Water polluting	Waterverontreinigende stoffen

II ECHA stoffen met biocidale werking

Relevante stoffen uit de ECHA database, zoals beschreven in paragraaf 3.1. De biodegradeerbaarheid loopt van 5 (zeer snel biodegradeerbaar, uren) tot 1 (zeer persistent, breekt niet af). Grofweg beneden een waarde van 3 zijn stoffen vaker terug te vinden in water.[20] Alle stoffen behalve Methofluthrin zijn redelijk biodegradeerbaar, met waarden rond de 3. Stoffen met een lage dampspanning zullen niet snel vervliegen. Boven de 0.003 is er kans dat de stoffen vervliegen voor ze in het water komen.[20] Stoffen die op basis van biodegradatie of vluchtigheid in het water terug te vinden zullen zijn, zijn dikgedrukt. De getoonde stoffen zijn nieuw in combinatie met hun producttype. De stof zelf kan al wel in gebruik zijn als biocide via andere toepassingen.

Stofnaam	CAS nr.	Dampspanning (mm Hg, 25°C)[9]	Biodegradeerbaarheid (weken)	BAS nr.	Type product
(13Z)-Hexadec-13-en-11-yn-1-yl acetate	78617-58-0	1.79E-005 [17]	3.0224 [17]	1923	PT19-Repellents and attractants
1,2-benzisothiazol-3(2H)-one (BIT)	2634-33-5	0.0000257	2.8651	1218	PT10-Construction material preservatives
2-methyl-2,3-dihydro-1,2-thiazol-3-one hydrochloride	26172-54-3	0.00000303	2.8641	2043	PT06-Preservatives for products during storage
2-Propenoic acid, 2-methyl-, butyl ester, polymer with butyl 2-propenoate and methyl 2-methyl-2-propenoate (CAS nr: 25322-99-0)/ Polymeric quaternary ammonium bromide (PQ Polymer)	-	-	-	1664	PT02-Disinfectants and algaecides not intended for direct application to humans or animals
5-Chloro-2-methyl-2H-isothiazol-3-one (CIT)	26172-55-4	0.0054	2.6954	1883	PT06-Preservatives for products during storage
Active chlorine generated from chloride of ambient water by electrolysis	7782-50-5	4990	3.0425	1605	PT02-Disinfectants and algaecides not intended for direct application to humans or animals
Allyl isothiocyanate	57-06-7	3.65	2.9801	1420	PT09-Fibre, leather, rubber and polymerised materials preservatives
alpha-bromadiolone	-	-	-	1863	PT14-Rodenticides
Benzyl Alcohol	100-51-6	0.0535	3.1422	82	PT06-Preservatives for products during storage
Chloramin B	127-52-6	1.83E-11	2.7977	1263	PT02-Disinfectants and algaecides not intended for direct application to humans or animals
Chlorine dioxide generated from sodium chlorite by acidification	10049-04-4	2.74E-08	3.0501	1455	PT09-Fibre, leather, rubber and polymerised materials preservatives
Copper	7440-50-8	0	3.0587	1274	PT02-Disinfectants and algaecides not intended for direct application to humans or animals
Ethanol	64-17-5	60.9	3.2573	1303	PT06-Preservatives for products during storage
Free radicals generated in situ from ambient air or water	-	-	-	1563	PT02-Disinfectants and algaecides not intended for direct application to humans or animals
MBIT (2-Methyl-1,2-benzothiazol-3(2H)-on	2527-66-4	0.000138 [18]	2.8341 [18]	1410	PT13-Working or cutting fluid preservatives

Metofluthrin	240494-71-7	2.01E-005	0.6945 (recalcitrant)	45	PT19-Repellents and attractants
Monochloramine generated from ammonia and a chlorine source	10599-90-3	1.16E-09	3.0854	1482	PT05-Drinking water
Monochloramine generated from ammonium carbamate and a chlorine source	10599-90-3	1.16E-09	3.0854	1483	PT06-Preservatives for products during storage
Monochloramine generated from ammonium chloride and a chlorine source	10599-90-3	1.16E-09	3.0854	1583	PT11-Preservatives for liquid-cooling and processing systems
Monochloramine generated from ammonium hydroxide and a chlorine source	10599-90-3	1.16E-09	3.0854	1663	PT05-Drinking water
Monochloramine generated from sodium hypochlorite and an ammonium source	10599-90-3	1.16E-09	3.0854	1884	PT05-Drinking water
Ozone generated from oxygen	10028-15-6	49600	3.0931	1783	PT02-Disinfectants and algaecides not intended for direct application to humans or animals
Reaction mass of chloromethyl hexyl cyanocarbonodithioimidate and bromomethyl hexyl cyanocarbonodithioimidate and dihexyl cyanocarbonodithioimidate	-	-	-	1623	PT09-Fibre, leather, rubber and polymerised materials preservatives
Silicic acid, aluminium magnesium sodium salt	12040-43-6	-	-	1418	PT18-Insecticides, acaricides and products to control other arthropods
Silver chloride	7783-90-6	4.74E-18	2.8824	1380	PT04-Food and feed area
Silver nitrate	7761-88-8	4.75E-13	2.6868	1381	PT02-Disinfectants and algaecides not intended for direct application to humans or animals
Silver phosphate glass	308069-39-8	-	-	1441	PT04-Food and feed area
Sodium Azide	26628-22-8	1.78E-22	3.1041	81	PT06-Preservatives for products during storage
Sulfur dioxide released from sodium metabisulfite	7446-09-5	2600	3.0576	1419	PT09-Fibre, leather, rubber and polymerised materials preservatives
Thiacloprid	111988-49-9	0.00000164	2.2199	53	PT08-Wood preservatives
Trichoderma harzianum strain T-720	67892-31-3	-	-	2063	PT08-Wood preservatives
Willaertia subsp. magna, C2c.Maky	-	-	-	1417	PT11-Preservatives for liquid-cooling and processing systems

III CBG unieke werkzame stoffen in nieuw ingeschreven geneesmiddelen

Lijst van unieke werkzame stoffen in nieuw ingeschreven geneesmiddelen in de periode van 28 juni 2021 tot 28 juli 2021. De gegevens zijn verzameld zoals beschreven in paragraaf 3.3. Deze stoffen kunnen reeds in gebruik zijn in andere geneesmiddelen, expertkennis is vereist om te beoordelen of hier relevante stoffen tussen zitten.

Werkzame stof	
ABIRATERONACETAAT	MYCOFENOLAAT NATRIUM SAMENSTELLING overeenkomend met MYCOFENOLZUUR
TACROLIMUS 1-WATER SAMENSTELLING overeenkomend met TACROLIMUS 0-WATER	DINATRIUMCROMOGLICAAT
AMILORIDE HYDROCHLORIDE-2-WATER SAMENSTELLING overeenkomend met AMILORIDE HYDROCHLORIDE-0-WATER	NALOXONHYDROCHLORIDE 2-WATER SAMENSTELLING overeenkomend met NALOXONHYDROCHLORIDE 0-WATER
HYDROCHLOORTHIAZIDE SAMENSTELLING overeenkomend met AMILORIDE	NICOTINERESINAAT SAMENSTELLING overeenkomend met NICOTINE
BETAMETHASONDIPROPIONAAT SAMENSTELLING overeenkomend met BETAMETHASON	NITROFURANTOINE 0-WATER
CALCIPOTRIOL 1-WATER SAMENSTELLING overeenkomend met CALCIPOTRIOL	AMLODIPINEBESILAAT SAMENSTELLING overeenkomend met AMLODIPINE OLMESARTANMEDOXOMIL
BENZOYLPEROXIDE n-WATER SAMENSTELLING overeenkomend met BENZOYLPEROXIDE 0-WATER	HYDROCHLOORTHIAZIDE OLMESARTANMEDOXOMIL
CLINDAMYCINEDIWATERSTOFFOSFAAT SAMENSTELLING overeenkomend met CLINDAMYCINE	PROMETHAZINE HYDROCHLORIDE SAMENSTELLING overeenkomend met PROMETHAZINE
DUTASTERIDE TAMSULOSINEHYDROCHLORIDE SAMENSTELLING overeenkomend met TAMSULOSINE	PILOCARPINEHYDROCHLORIDE SAMENSTELLING overeenkomend met PILOCARPINE
DESOGESTREL	SILDENAFILCITRAAT SAMENSTELLING overeenkomend met SILDENAFIL
DEXAMETHASONDINATRIUMFOSFAAT SAMENSTELLING overeenkomend met DEXAMETHASONFOSFAAT	LANREOTIDEACETAAT SAMENSTELLING overeenkomend met LANREOTIDE
AZELASTINEHYDROCHLORIDE SAMENSTELLING overeenkomend met AZELASTINE FLUTICASONPROPIONAAT	SORAFENIBTOSYLAAT SAMENSTELLING overeenkomend met SORAFENIB
LEUPRORELINEACETAAT SAMENSTELLING overeenkomend met LEUPRORELINE	SUGAMMADEX NATRIUM SAMENSTELLING overeenkomend met SUGAMMADEX
EZETIMIB	SUNITINIBMALAAT SAMENSTELLING overeenkomend met SUNITINIB
FENTANYLDIWATERSTOFCITRAAT SAMENSTELLING overeenkomend met FENTANYL	SUNITINIB
FLUTICASONPROPIONAAT	TRANYLCYPROMINESULFAAT SAMENSTELLING overeenkomend met TRANYLCYPROMINE

ACETYLCYSTEÏNE	TRAZODONHYDROCHLORIDE SAMENSTELLING overeenkomend met TRAZODON
LANTHAANCARBONAAT 4-WATER SAMENSTELLING overeenkomend met LANTHAAN (LA3+)	NATRIUMALGINAAT (E 401)
BECLOMETASONDIPROPIONAAT 0-WATER	NATRIUMWATERSTOFCARBONAAT (E 500 (II))
FORMOTEROLFUMARAAT 2-WATER SAMENSTELLING overeenkomend met FORMOTEROL	PROGESTERON
FUROSEMIDE	ETOPOSIDE
CALCIUMCARBONAAT (E 170)	CARBOMEER 980
HUMAAN CHORIONGONADOTROFINE	PODOPHYLLOTOXINE
SUMATRIPTANSUCCINAAT SAMENSTELLING overeenkomend met SUMATRIPTAN	ADRENALINEWATERSTOFTARTRAAT SAMENSTELLING overeenkomend met ADRENALINE
LEVONORGESTREL	LIDOCAINEHYDROCHLORIDE 1-WATER SAMENSTELLING overeenkomend met LIDOCAINEHYDROCHLORIDE 0-WATER
LEVETIRACETAM	LEVOCETIRIZINEDIHYDROCHLORIDE SAMENSTELLING overeenkomend met LEVOCETIRIZINE
BARNIDIPINEHYDROCHLORIDE SAMENSTELLING overeenkomend met BARNIDIPINE	ONDANSETRONHYDROCHLORIDE 2-WATER SAMENSTELLING overeenkomend met ONDANSETRON
PERMETHRINE	BUPROPIONHYDROCHLORIDE SAMENSTELLING overeenkomend met BUPROPION
METHOTREXAAT DINATRIUM SAMENSTELLING overeenkomend met METHOTREXAAT	MILRINON

IV CBG unieke bestanddelen in nieuw ingeschreven diergeneesmiddelen

Lijst van unieke bestanddelen (zowel hulpstoffen als werkzame stoffen) in nieuw ingeschreven diergeneesmiddelen in de periode van 11 juli 2021 tot 11 augustus 2021. De gegevens zijn verzameld zoals beschreven in paragraaf 3.3. Deze stoffen kunnen reeds in gebruik zijn in andere diergeneesmiddelen, expertkennis is vereist om te beoordelen of hier relevante stoffen tussen zitten.

Bestanddelen
Broomhexinehydrochloride
Amoxicilline 3-water
Amoxicilline 0-water
Kaliumclavulanaat
Clavulaanzuur
Pergolidemesilaat
Pergolide
Levothyroxinenatrium 0-water
levothyroxine

V NVWA afzetgegevens gewasbeschermingsmiddelen

Lijst van toe- en afname in de afzet van gewasbeschermingsmiddelen in 2019 ten opzichte van 2018. De gegevens zijn verzameld zoals beschreven in paragraaf 3.4. De lijst is gesorteerd op het procentuele verschil t.o.v. 2018.

werkzame stof	afzet in 2018 (kg)	afzet in 2019 (kg)	verschil t.o.v. 2018 (kg)	verschil t.o.v. 2018 (%)
INDOLYLBOTERZUUR	97	29313	29216	30119.6
PYRETHRINEN	221	17955	17734	8024.4
PELARGONZUUR	2590	37283	34693	1339.5
FENPYRAZAMINE	528	5172	4644	879.5
METOBROMURON	5590	47444	41854	748.7
SULFOXAFLOR	50	422	372	744.0
SULFURYLFLUORIDE	16371	96130	79759	487.2
PIRIMIFOS-METHYL	100	580	480	480.0
(E,E)-8,10-DODECADIEN-1-OL	53	161	108	203.8
(E,Z,Z)-3,8,11-TETRADECATRIEN-1-YL ACETATE	50	145	95	190.0
(E,Z)-3,8-TETRADECADIEN-1-YL ACETATE	6	17	11	183.3
PENFLUFEN	2434	6698	4264	175.2
BROMUCONAZOOL	564	1426	862	152.8
PYRIOFENON	86	204	118	137.2
THIACLOPRID	14331	32113	17782	124.1
CHLOORMEQUAT	104178	222414	118236	113.5
AZADIRACHTINE-A	1022	2110	1088	106.5
PIRIMICARB	4214	8576	4362	103.5
METAM-NATRIUM	17340	35220	17880	103.1
AMIDOSULFURON	37	75	38	102.7
1-METHYLCYCLOPROPEEN	2	4	2	100.0
ESFENVALERAAT	4631	8525	3894	84.1
MESOTRIONE	6139	10801	4662	75.9
CARVON	10400	18087	7687	73.9
STREPTOMYCIS GRISEOVIRIDIS	48	81	33	68.8
TRIBENURON-METHYL	41	69	28	68.3
CYANTRANILIPROLE	1696	2854	1158	68.3
NAPROPAMIDE	1593	2628	1035	65.0
EMAMECTIN BENZOAAT	1444	2245	801	55.5
FLONICAMID	14852	22711	7859	52.9
2,4-DB	4280	6504	2224	52.0
HEXYTHIAZOX	192	281	89	46.4
TEFLUBENZURON	921	1335	414	45.0
ETHOPROFOS	40170	57798	17628	43.9
LENACIL	5845	8390	2545	43.5
TRICLOPYR	1212	1727	515	42.5
DESMEDIFAM	11850	16538	4688	39.6
FENPROPIMORF	18930	26345	7415	39.2
SPIROTETRAMAT	9460	12722	3262	34.5
CYROMAZINE	296	388	92	31.1

PYRAFLUFEN-ETHYL	641	839	198	30.9
IJZER(III)FOSFAAT	10041	13047	3006	29.9
AZIJNZUUR	7267	9360	2093	28.8
ZOXAMIDE	435	558	123	28.3
PYRIPROXIFEN	215	271	56	26.0
METARHIZIUM ANISOPLIAE	221	278	57	25.8
DICAMBA	6091	7642	1551	25.5
AMETOCTRADIN	10946	13704	2758	25.2
DODECAN-1-OL	8	10	2	25.0
CARFENTRAZON-ETHYL	2719	3387	668	24.6
FLUTOLANIL	10065	12502	2437	24.2
TRINEXAPAC-ETHYL	6909	8576	1667	24.1
BROMOXYNIL	5459	6753	1294	23.7
BENFLURALIN	7657	9464	1807	23.6
MALEINEHYDRAZIDE	171861	211641	39780	23.1
FLUXAPYROXAD	9100	11193	2093	23.0
ETHOFUMESAAT	42897	52215	9318	21.7
MCPA	204025	247235	43210	21.2
CYAZOFAMID	54334	65744	11410	21.0
RIMSULFURON	678	796	118	17.4
PYRIDABEN	487	571	84	17.2
MEPIQUATCHLORIDE	140	164	24	17.1
(Z)-11-TETRADECENYLACETAAT	94	110	16	17.0
FLUAZINAM	12358	14437	2079	16.8
ALUMINIUMSULFAAT	14426	16751	2325	16.1
KWARTSZAND	1692	1963	271	16.0
ETRIDIAZOOL	1781	2051	270	15.2
ISOXABEN	7160	8200	1040	14.5
PROPYZAMIDE	14823	16804	1981	13.4
FENPROPIDIN	16155	18291	2136	13.2
FORMETANATE	888	1002	114	12.8
GIBBERELLINEZUUR	96	108	12	12.5
FLUOPYRAM	19008	21259	2251	11.8
NATRIUMZILVERTHIOSULFAAT	75	83	8	10.7
CLOPYRALID	18125	19934	1809	10.0
KNOFLOOK EXTRACT	5319	5823	504	9.5
CYFLUMETOFEN	1512	1642	130	8.6
FLUFENACET	6510	7069	559	8.6
DODEMORF	15899	17246	1347	8.5
IMAZALIL	6274	6788	514	8.2
FOSETYL-ALUMINIUM	53499	57821	4322	8.1
TRIFLUSULFURON METHYL	565	610	45	8.0
PROTHIOCONAZOOL	43552	46965	3413	7.8
ACETAMIPRID	7072	7622	550	7.8
DIMETHENAMIDE-P	170494	183213	12719	7.5
CYPRODINIL	11707	12551	844	7.2
METCONAZOOL	287	307	20	7.0
METHOXYFENOZIDE	2494	2654	160	6.4
1,4 DIMETHYLNAFTALEEN	25454	26960	1506	5.9
ALUMINIUMFOSFIDE	497	526	29	5.8
DIFLUFENICAN	3323	3503	180	5.4
PENDIMETHALIN	108008	112953	4945	4.6

ETOXAZOOL	199	208	9	4.5
THIOFANAAT-METHYL	91970	95400	3430	3.7
PROPICONAZOOL	4061	4186	125	3.1
TRITOSULFURON	853	876	23	2.7
LUFENURON	244	250	6	2.5
BIXAFEN	3888	3967	79	2.0
TOLCLOFOS-METHYL	35340	36000	660	1.9
QUINOCLAMIN	1412	1438	26	1.8
2,4-D	75843	77235	1392	1.8
IJZER(II)SULFAAT	16243	16504	261	1.6
FLUROXYPYR	62859	63455	596	0.9
BUPIRIMAAT	4730	4772	42	0.9
PENCONAZOOL	699	703	4	0.6
FLUOXASTROBIN	6266	6293	27	0.4
ABAMECTINE	1059	1059	0	0.0
TERTRADECANOL	2	2	0	0.0
COS-OGA	19	19	0	0.0
(Z)-TETRADEC-9ENYLACETAAT	11	11	0	0.0
FLUDIOXONIL	12449	12379	-70	-0.6
FOLPET	86315	85524	-791	-0.9
CONIOTHYRUM MINITANS	610	601	-9	-1.5
FENMEDIFAM	51768	50721	-1047	-2.0
LAMBDA-CYHALOTHRIN	2374	2319	-55	-2.3
INDOXACARB	717	700	-17	-2.4
PYRIDALYL	5803	5660	-143	-2.5
OXAMYL	55878	54468	-1410	-2.5
KRESOXIM-METHYL	5186	5055	-131	-2.5
MILBEMECTINE	35	34	-1	-2.9
FLORASULAM	1345	1304	-41	-3.0
SPIRODICLOFEN	366	354	-12	-3.3
METAMITRON	242618	234649	-7969	-3.3
MANCOZEB	2218646	2132832	-85814	-3.9
CYFLUFENAMID	374	357	-17	-4.5
FENHEXAMIDE	2121	2013	-108	-5.1
PYROXSULAM	672	634	-38	-5.7
CAPTAN	434546	409917	-24629	-5.7
BIFENAZAAT	804	753	-51	-6.3
AMISULBROM	20343	19009	-1334	-6.6
IMAZAMOX	144	134	-10	-6.9
SEDAXAAN	930	865	-65	-7.0
BIFENOX	4483	4169	-314	-7.0
PROCHLORAZ	33289	30869	-2420	-7.3
METAZACHLOOR	9396	8689	-707	-7.5
TEBUCONAZOOL	21658	20000	-1658	-7.7
MEPANIPYRIM	6855	6305	-550	-8.0
PROPAQUIZAFOP	174	160	-14	-8.0
TEMBOTRIONE	8831	8072	-759	-8.6
CHLOORPROFAM	100122	91462	-8660	-8.6
ETHEFON	4447	4057	-390	-8.8
PROHEXADION-CALCIUM	1721	1567	-154	-8.9
CLETHODIM	7821	7120	-701	-9.0
PINOXADEN	152	138	-14	-9.2

METALAXYL-M	5436	4934	-502	-9.2
EPOXYCONAZOOL	17333	15723	-1610	-9.3
DIMETHOMORF	33504	30354	-3150	-9.4
CYMOXANIL	41888	37774	-4114	-9.8
PYRACLOSTROBIN	36708	32897	-3811	-10.4
BOSCALID	19601	17456	-2145	-10.9
MESOSULFURON-METHYL	6329	5574	-755	-11.9
KALIUM WATERSTOFCARBONAAT	53040	46627	-6413	-12.1
FLUAZIFOP-P-BUTYL	3527	3100	-427	-12.1
FLUPYRADIFURON	3175	2783	-392	-12.3
MECOPROP-P	8069	7063	-1006	-12.5
METSULFURON-METHYL	88	77	-11	-12.5
MANDIPROPAMID	54641	47280	-7361	-13.5
CHLORIDAZON	65939	56865	-9074	-13.8
ASULAM	37432	32140	-5292	-14.1
SPINOSAD	10726	9121	-1605	-15.0
PROSULFOCARB	293197	249168	-44029	-15.0
BENTAZON	19459	16462	-2997	-15.4
AZOXYSTROBINE	55315	46727	-8588	-15.5
THIENCARBAZON-METHYL	3641	3062	-579	-15.9
DAMINOZIDE	37362	31252	-6110	-16.4
GLYFOSAAT	772234	639050	-133184	-17.2
BUPROFEZIN	438	360	-78	-17.8
TERBUTYLAZIN	69505	56932	-12573	-18.1
FORAMSULFURON	417	340	-77	-18.5
CHLORANTRANILIPROLE	2383	1938	-445	-18.7
KALIUMFOSFONATEN	16146	13015	-3131	-19.4
METALDEHYDE	765	603	-162	-21.2
6-BENZYLADENINE	942	740	-202	-21.4
METHIOCARB	3250	2525	-725	-22.3
METRIBUZIN	40627	31427	-9200	-22.6
NICOSULFURON	4264	3292	-972	-22.8
DELTAMETHRIN	2106	1617	-489	-23.2
SPIROMESIFEN	556	426	-130	-23.4
IODOSULFURON-METHYL-NATRIUM	2547	1946	-601	-23.6
TRIFLUMIZOOL	1467	1105	-362	-24.7
DIQUATDIBROMIDE	149509	112606	-36903	-24.7
FOSTHIAZATE	61483	46263	-15220	-24.8
ACIBENZOLAR-S-METHYL	8	6	-2	-25.0
BENTHIAVALICARB-ISOPROPYL	11494	8526	-2968	-25.8
QUIZALOFOP-P-ETHYL	544	391	-153	-28.1
DIFENOCONAZOOL	101577	72680	-28897	-28.4
FENAMIDONE	3744	2664	-1080	-28.8
HALOXYFOP-P-METHYL	2700	1915	-785	-29.1
TRIFLOXYSTROBIN	14260	10033	-4227	-29.6
MALTODEXTRINE	13036	9125	-3911	-30.0
ACEQUINOCYL	1615	1128	-487	-30.2
CLOMAZONE	4925	3399	-1526	-31.0
METRAFENONE	639	438	-201	-31.5
CHLOORTHALONIL	63900	43631	-20269	-31.7
ACLONIFEN	82353	55164	-27189	-33.0
GIBBERELLINEN	170	106	-64	-37.6

QUINMERAC	1678	1044	-634	-37.8
S-METOLACHLOOR	89369	55079	-34290	-38.4
PROPAMOCARB HYDROCHLORIDE	210890	128174	-82716	-39.2
CYCLOXYDIM	4804	2810	-1994	-41.5
FENOXAPROP-P-ETHYL	327	191	-136	-41.6
CHLOFENTEZIN	840	485	-355	-42.3
THIFENSULFURON-METHYL	441	252	-189	-42.9
PROSULFURON	187	98	-89	-47.6
DITHIANON	22884	11245	-11639	-50.9
CYPROCONAZOOL	3337	1628	-1709	-51.2
FLUOPICOLIDEN	21178	10266	-10912	-51.5
DODINE	10800	5208	-5592	-51.8
ZWAVEL	78455	36807	-41648	-53.1
ISOPYRAZAM	2928	1363	-1565	-53.4
1-NAFTYLAZIJNZUUR	188	81	-107	-56.9
(Z)-8-DODECEN-1-YL ACETAAT	7	3	-4	-57.1
PYMETROZINE	7645	3145	-4500	-58.9
FLUMIOXAZIN	4586	1831	-2755	-60.1
PYRIMETHANIL	18295	6753	-11542	-63.1
PACLOBUTRAZOL	170	60	-110	-64.7
THIAMETHOXAM	2318	707	-1611	-69.5
BACILLUS THURINGIENSIS	34120	9725	-24395	-71.5
KOOLZAADOLIE	22292	6174	-16118	-72.3
THIRAM	60340	14472	-45868	-76.0
PENCYCURON	24290	5597	-18693	-77.0
SILTHIOFAM	168	35	-133	-79.2
IMIDACLOPRID	660	135	-525	-79.5
LAMINARIN	372	63	-309	-83.1
ETHYLEEN	13290	1696	-11594	-87.2
ISOXAFLUTOOL	130	9	-121	-93.1
GROENEMUNTOLIE	17024	570	-16454	-96.7
CARBEETAMIDE	2652	34	-2618	-98.7
GLUFOSINAAT-AMMONIUM	43040	492	-42548	-98.9
BENZOEZUUR	4887	7	-4880	-99.9
FIPRONIL	3952	-800	-4752	-120.2

VI Informatie extraheren met 'Natural Language Processing' (NLP)

Met NLP is het mogelijk spreek- of schrijftaal te interpreteren. In deze bijlage zetten we uiteen wat de mogelijkheden zijn om informatie rond stoffen of processen te achterhalen uit teksten. De technieken zijn universeel en kunnen ook voor andere informatie extractie doeleinden worden ingezet. De analyses zijn voor deze bijlage uitgevoerd met functies die beschikbaar zijn in diverse te installeren pakketten voor de statistische software 'R'. In de software 'Python' zijn soortgelijke functies ook beschikbaar. Als tekstbron voor informatie hebben we samenvattingen van wetenschappelijke artikelen welke bepaalde steekwoorden bevatten gedownload uit de database 'PubMed'. Dit is mogelijk binnen 'R' met functies uit het pakket 'EasyPubmed'.

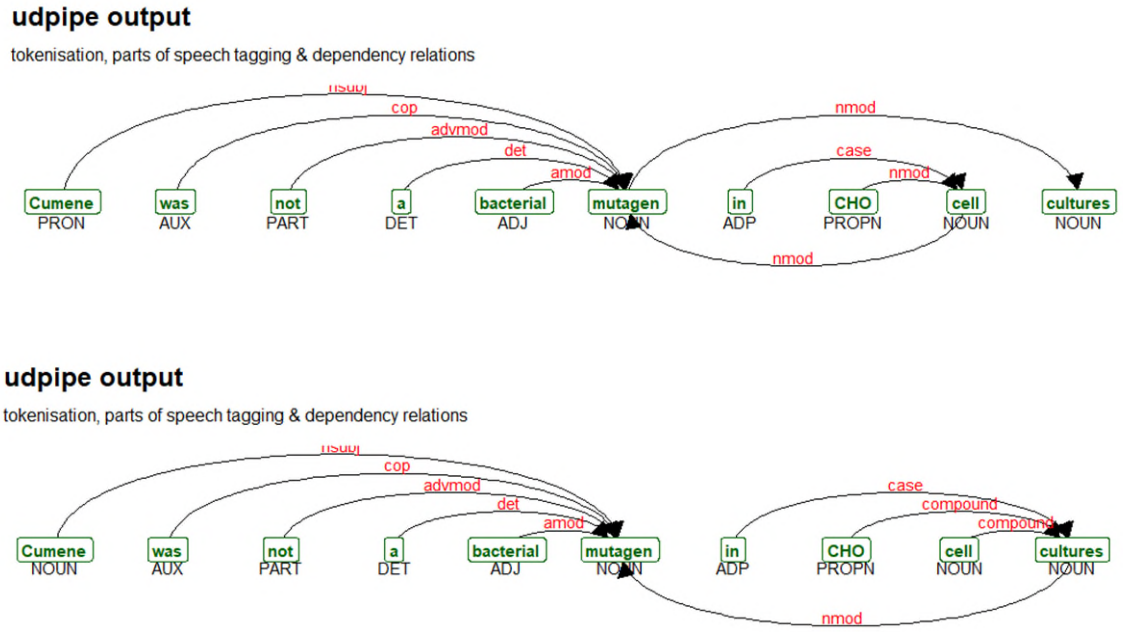
Zinsopbouw afleiden door labels toe te voegen: Parsers

Voor de automatische interpretatie van een tekst is het handig om te weten welke van de woorden werkwoorden zijn, welke zelfstandig naamwoorden, wat de bijwoorden zijn en op welk woord ze slaan, enzovoort. Van een tekst kan met 'parsers' bepaald worden wat voor labels ('tags' of 'annotations') er kunnen worden gehangen aan woorden of stukken uit de tekst, of relaties daartussen. De woorden of stukken uit de tekst heten in NLP 'tokens' en worden vóór het parsen geïdentificeerd door een 'tokenizer'.

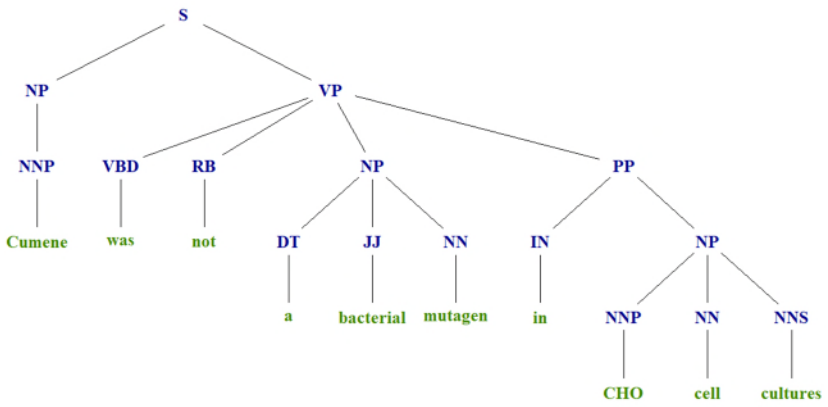
Parsers analyseren een tekst op basis van onderliggende grammatica. Parsing-benaderingen kunnen statistisch, probabilistisch zijn of op machine leren gebaseerd. Enkele voorbeelden van parsers zijn de Stanford en de OpenNLP parser. Deze kunnen ingezet worden via beschikbare functies in pakketten in 'R'. De parser-benaderingen zijn getraind op bepaalde teksten. Bijvoorbeeld nieuwsberichten, wetenschappelijke literatuur, of Wikipedia. Afhankelijk van de tekst waarop ze getraind zijn, zijn ze beter of slechter in staat om de tekst goed te labelen. Diverse labels kunnen worden toegekend. Hieronder een overzicht met enkele voorbeelden.

- Wat in de tekst is een zin ('sentence annotator'). Een goede parser stopt bijvoorbeeld niet bij elke punt.
- Wat in de tekst is één woord ('word annotator'). Een goede parser herkent ook samengestelde woorden of woorden met nummers erin zoals chemische verbindingen.
- Wat is de functie van het woord ('part of speech', POS tag) zoals zelfstandig naamwoord, werkwoord, bijvoeglijk voornaamwoord, punctuatie, etc.)
- Afhankelijkheid van woorden ten opzichte van elkaar ('dependency'), slaat het bijwoord 'bacterieel' op 'mutageen' of 'cell-cultures'?
- Deelzinnen, bijvoorbeeld rond een zelfstandig naamwoord- of werkwoord ('noun phrase' of 'verb phrase').

De diverse functies uit verschillende pakketten in 'R' doen dit labelen automatisch, dat maakt het eenvoudig in de uitvoering. In Figuur 1 en Figuur 2 staan visualisaties van zulke labeling. In Tabel 1 staat een voorbeeld van een tabel waarin deze informatie staat. Voor dit project zijn drie pakketten uitgeprobeerd: OpenNLP, UDpipe, en SpacyR.



Figuur 1. Tokenisatie, afhankelijkheidsbepaling ('Dependency parsing') en POS-tagging met het pakket 'UDPIPE'. Twee ingeladen modellen (boven en onder) geven nét iets andere labels. Voor de betekenis van de dependency labels zie https://universaldependencies.org/u/dep/index.html.



Figuur 2. Tokenisatie, deelzin ('chunk') tagging en POS-tagging met het pakket 'OpenNLP'. Deze verwerker kent 'Penn Treebank' (https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html) labels toe aan de hand van de 'Apache OpenNLP chunking parser' voor Engels.

Tabel 1. Voorbeeld van een tabel met labels per 'token', gegenereerd met functies uit het pakket 'SpacyR'.

tekst_id	zin_id	token_id	token	lemma	pos	head_ token_id	dep_rel	entity nounphrase
text1	1	2	Decontamination	Decontamination	PROPN	3	compound	beg
text1	1	3	procedures	procedure	NOUN	3	ROOT	end_root
text1	1	4	for	for	ADP	3	prep	
text1	1	5	skin	skin	NOUN	4	pobj	beg_root
text1	1	6	exposed	expose	VERB	5	acl	
text1	1	7	to	to	ADP	6	prep	
text1	1	8	phenolic	phenolic	ADJ	9	amod	beg
text1	1	9	substances	substance	NOUN	7	pobj	end_root
text1	1	10	.	.	PUNCT	3	punct	

Al deze informatie kan gebruikt worden voor het extraheren van informatie. Het gegeven dat een woord gelabeld is als POS kan bijvoorbeeld al onderscheid maken of het woord 'finish' in een zin een Biocide kan zijn (dat zal als een zelfstandig naamwoord zijn gelabeld) of het als een werkwoord is bedoelt. Ook kan POS gebruikt worden voor het extraheren van sleutelwoorden uit de tekst voor bijvoorbeeld weergave in een woordwolk; de zelfstandig naamwoorden zullen hier de beste informatie geven.

Het identificeren van deelzinnen geeft de onderwerpen in de tekst duidelijk aan. In Tabel 1 is bijvoorbeeld te zien dat de woorden 'decontamination procedure' bij elkaar horen. De kolom 'entity nounphrase' geeft namelijk aan dat de noun phrase begint bij 'decontamination' en eindigt bij 'procedure'. Ook de woorden 'phenolic substances' horen op deze manier bij elkaar.

Als we van alles willen weten over het voorbeeld 'cumene' kunnen we opzoeken welke verb phrases bij noun phrases horen die het woord 'cumene' bevatten, en welke noun phrases weer subjects zijn daarvan. Vanaf hier wordt het ingewikkeld, omdat de 'R' pakketten geen kant en klare functies hebben voor zulke specifieke informatie rond een woord van interesse. Hier moet dus code geschreven worden dat met de beschikbare informatie begrijpelijke feiten kan extraheren. Ingewikkelde zinsopbouw is daarbij een complicerende factor met voegwoorden zoals 'and', het bestaan van bijzinnen, en ontkenningen zoals 'not'. Informatie staat soms ook verspreid over meerdere zinnen. In de code zullen veel verschillende labels (POS) moeten worden geadresseerd in de code, met hun specifieke relatie. Als de regels niet alle situaties adresseren, zullen er fouten in de resultaten voorkomen. Op basis daarvan kan dan een nieuwe regel geschreven worden.

Specifieke concepten herkennen in tekst: Named Entity Recognition (NER)

Voor het bepalen van potentieel nieuwe chemische bedreigingen voor de waterketen is het nuttig om te kunnen herkennen wanneer het in een tekst gaat over chemicaliën. Een optie is om een lijst met chemische namen, een 'lexicon', te gebruiken. Voor het herkennen van chemicaliën in Twitterberichten hebben we met deze techniek gewerkt. Het nadeel hiervan is dat chemicaliën in de praktijk vele synoniemen hebben en deze staan niet allemaal in de lijst. Een ander nadeel is dat de namen van chemicaliën in de lijst soms dubbele betekenis hebben. Om niet te veel resultaten te krijgen die geen chemicaliën zijn, moet er daardoor met 'blacklists' gewerkt worden waarbij een woord uit de lijst uitgesloten wordt vanwege de vele valse resultaten die het oplevert.

Voor een deel kunnen chemicaliën opgespoord worden door hun unieke naamgeving. Een CAS-nummer bestaat bijvoorbeeld altijd uit nummers met een streepje ertussen, en het laatste cijfer is een controlecijfer dat afgeleid kan worden van de voorgaande cijfers. 'Regular expressions' (Regex) kunnen dit opsporen. Regex werken met

conventies voor het herkennen van letterpatronen (zie ook Kader 1 in het rapport). Een te herkennen patroon kan zijn '1 of meer cijfers gevolgd door een '-', gevolgd door 1 of meer letters'. Dit wordt geschreven in code als "[0-9]+ - [a-z]+" en alle tekst die hieraan voldoet, wordt herkend en teruggegeven. De '+' is de code voor 'een of meer'. Een ander karakteristiek voor chemicaliën is dat de naam vaak wordt geschreven als een combinatie van cijfers en letters, zoals 1,2-dichloor. Er blijven echter genoeg namen over die hier niet aan voldoen, dit zal dus maar een deel van de chemicaliën kunnen achterhalen.

Een alternatief voor het herkennen van concepten is het gebruiken van 'Machine Learning' (ML). Hierbij worden teksten ('corpus') gebruikt waarbij de chemicaliën al van te voren zijn geïdentificeerd door een expert. Aan de hand van de plek van de chemicaliën binnen de zinsstructuur en de specifieke gebruikte woorden kan de computer 'leren' wanneer het over een chemische stof gaat in de tekst. Kant en klare NER annotaties die zijn gebaseerd op ML bestaan in de 'R' pakketten voor NLP voor personen, plaatsen, datums, geld, organisatie, percentages. Voor andere concepten moet een ML model gebouwd worden of een commercieel pakket aangeschaft worden.

Statistische verbanden zoeken met een 'bag of words' aanpak

Er bestaat ook een andere manier van het relateren van woorden aan elkaar. Deze aanpak is veel simpeler, en bestaat uit het simpelweg tellen van woorden en kijken of ze vaak bij elkaar voorkomen in teksten. In combinatie met NER kunnen zo bijvoorbeeld clusters van chemicaliën worden geïdentificeerd. Een combinatie van 'bag of words' en NLP is ook mogelijk, het gaat dan om het tellen van zelfstandig naamwoorden of werkwoorden die voorkomen bij een bepaald woord van interesse. Bij de 'bag of words' aanpak wordt vaak gebruik gemaakt van het afkorten van woorden tot de woordstam, 'stemming'. Hierdoor is het tellen van woorden niet afhankelijk van vervoegingen of het gebruik van enkelvoud / meervoud.

Tabel 2. Overzicht veel voorkomende 'Dependency' relatie labels.

det	determiner
compound	compound
nsubj	nominal subject
amod	adjectival modifier
cc	coordinating conjunction
conj	conjunct
dobj	direct object
relcl	relative clause modifier
punct	punctuation
ccomp	clausal complement
prep	prepositional modifier
pobj	object of preposition
pcomp	complement of preposition
advmod	adverbial modifier
aux	auxiliary
mark	marker
nsubjpass	nominal subject (passive)
auxpass	auxiliary (passive)
advcl	adverbial clause modifier
appos	appositional modifier
dative	dative
acomp	adjectival complement
nummod	numeric modifier
attr	attribute
dep	unclassified dependent
agent	agent
expl	expletive
acl	clausal modifier of noun (adjectival clause)
npadvmod	noun phrase as an adverbial modifier
neg	negation modifier
xcomp	open clausal complement
intj	interjection

VII Beschrijving van de scripts met R code

In deze bijlage staat een opsomming van de scripts die gemaakt zijn tijdens dit project. Elk script voert een text-mining techniek uit, al dan niet als 'proof-of- concept'. De scripts zijn beschikbaar via het TM Data-pakket (2022).

Nr	Naam ('Script_402045_157' gevolgd door…)	Functie
1	Stofherkenner.R	Bevat de code om 'woorddelen' te maken van stofnamen en woorden uit abstracts. Deze twee lijsten woorddelen worden vervolgens in de rest van de code gebruikt om tekstwoorden te classificeren als stofnaam of niet-stofnaam.
2	NLPpakketten.R	Bevat voorbeelden van het verwerken van een abstract met de functies in de R-pakketten 'SpacyR', 'UDpipe', 'OpenNLP'.
3	Twitter_schrapen.R	Bevat code voor het binnenhalen van twitterberichten op basis van trefwoorden (hier namen van bedrijven in het Rijnstroomgebied) en deze te doorzoeken op aanwijzingen voor nieuwe bedrijvigheid met behulp van trefwoorden en een CHeBi-lijst met chemische namen.
4	NLPchunks.R	Bevat code voor het maken van 'lange chunks' van woorden die bij een zelfstandig naamwoord of werkwoord horen. Deze chunks worden vervolgens bij elkaar gezet als combinatie van zelfstandig naamwoord, werkwoord, zelfstandig naamwoord om zo 'triplets' van feiten (ook wel 'assertie') te maken.
5	StofnaamClusters.R	Bevat code voor het genereren van een 'term-document-matrix' met stofnamen, die via een clustering tot groepen worden gebracht. Input is een stofnaam of proces van interesse. Deze wordt als trefwoord gebruikt om relevante Pubmed abstracts te downloaden.
6	Webscraping_amtsblatter.R	Bevat code om Amtsblätter te downloaden en deze te doorzoeken op relevante woorden.
7	Webscraping_websites.R	Bevat code voor het scrapen van websites en om de links op webpagina's te doorzoeken op nieuws of productinformatie.
8	Webscraping_databases.R	Bevat code voor het binnenhalen van ECHA factsheets, en het downloaden van informatie over nieuw ingeschreven (dier)geneesmiddelen en afzetgegevens van gewasbeschermingsmiddelen. De gegevens worden verzameld en in overzichtelijke tabellen gezet.