TEXT MINING

Op schattenjacht in ongelezen pdf's

Wat als we al hebben ontdekt hoe we de ziekte van Alzheimer moeten genezen, maar het niet weten omdat het antwoord begraven ligt in een pdf die niemand leest? Jammer genoeg is dat idee niet eens zo vergezocht.

Toon Verlinden

Op PubMed, waar het overgrote deel van de biomedische publicaties verschijnt, komen elke minuut twee nieuwe wetenschappelijke publicaties bij. De teller staat momenteel op meer dan 24 miljoen teksten. Hoe kunnen we dan verwachten dat onderzoekers en artsen up-to-date blijven en de link leggen tussen een symptoom uit de eerste publicatie en een werkzame stof uit de honderdduizendste?
We laten hier belangrijke kansen voor nieuwe medicijnen liggen, en dat is frustrerend. Gelukkig werken onderzoekers aan methodes die de antwoorden op onze problemen binnenkort automatisch aan ons zullen aanbieden.

Geen koffiepauze nodig

De geboorte van een nieuw medicijn start met een spelletje darts voor gevorderden. Onderzoekers proberen honderden nieuwe stofcombinaties uit die een positief effect kunnen hebben op een ziekte. Normaal gezien valt meer dan 99 procent van die combinaties af en gaat slechts een miniem percentage door naar de volgende ronde. Verlies van tijd én geld.
'Omdat onderzoekers onvoldoende tijd hebben om een overzicht te bewaren over alle nieuwe publicaties, moeten ze soms wat op hun wetenschappelijk buikgevoel afgaan', vertelt Thomas Provoost van het team Language Intelligence & Information Retrieval (LIIR) van de KU Leuven. Hij doet onderzoek naar het ontginnen van biomedische teksten. Stel je voor dat een computer de 24 miljoen PubMed-publicaties automatisch kan samenvatten tot een overzichtelijk geheel en daarbij zelf nieuwe hypotheses genereert? Geweldig. Want een computer kan blijven lezen, wordt nooit moe en heeft geen koffiepauze nodig.
Provoost: 'Door duizenden publicaties met elkaar te vergelijken, kan de computer nieuwe stofcombinaties vinden en hypotheses genereren die iedereen over het hoofd zag.' Terwijl de mens verloren loopt in een bos aan woorden en publicaties, kan een pc dag en nacht elk woord en detail met evenveel aandacht bekijken en zeldzame links ontdekken. Zo achterhaalden wetenschappers dat veel migrainelijders een magnesiumtekort hebben. Dat automatisch genereren van hypotheses laat onderzoekers toe het aantal dartspijltjes bij de start van de medicijnontwikkeling te reduceren en gerichter nieuwe stofcombinaties uit te proberen.

Compacte taalcode

De theorie klinkt mooi; de praktijk is niet vanzelfsprekend. Computers zijn kampioen in het lezen van gestructureerde databases, maar hebben het nog moeilijk met doorlopende of vrije teksten.
Walter Daelemans, hoofd van de groep Computerlinguïstiek aan de Universiteit Antwerpen, verwoordt dat als volgt: 'Taal is een erg compacte code waarbij we allerlei dingen impliceren zonder ze uit te spreken. Zeg je bijvoorbeeld: 'Jan nam de krant. Hij was op zoek naar een nieuwe baan.', dan is dat een hele andere context dan: 'Jan nam de krant. Hij was de vlieg beu.' Je interpreteert hetzelfde stukje 'Jan nam de krant' aan de hand van de tekst die errond staat. Geen sinecure voor een computer.'
Doorlopende tekst heeft geen eenduidige labels en daardoor hangt er voor een computer geen betekenis aan vast. Je kan hem dus geen cursus biochemie geven en verwachten dat hij die kennis zelf leert en toepast op andere teksten. Je moet de achtergrond en context nog steeds handmatig aanleveren. 'Een zelflerende computer is jammer genoeg nog toekomstmuziek', vertelt Provoost. 'Nu leiden we een computer nog steeds op door hem massa's voorbeelden en regels te voeren, in de hoop dat hij de patronen herkent. Dat is echter tijdrovend, weinig flexibel en erg beperkend.' De echte doorbraak in het ontginnen van teksten – of: Text Mining – komt pas op gang als een computer zichzelf dingen kan aanleren en ongesuperviseerd aan de slag kan gaan.

Computer vs mens: 1-0

Toch is er al veel mogelijk. IBM ontwikkelde enkele jaren geleden de Watsoncomputer. Door gebruik te maken van gestructureerde en ongestructureerde bronnen als Wikipedia, speelde Watson in 2011 met gemak zijn menselijke

tegenstanders naar huis in het tv-spel *Waagstuk*. Niet vanzelfsprekend, want de computer moest daarvoor ook ironie en raadsels herkennen. Iets waar computers doorgaans meer moeite mee hebben dan mensen.
Ook in het EMIF-onderzoeksproject (European Medical Information Framework) ontginnen bedrijven en universiteiten bestaande publicaties. Zij zijn onder andere op zoek naar nieuwe biomarkers die in de toekomst de ziekte van Alzheimer sneller moeten opsporen. ‘Medicatie is vaak minder efficiënt als ze laat wordt toegediend’, vertelt Bart Vannieuwenhuyse, verantwoordelijk voor Health Information Sciences bij Janssen Pharmaceutica. ‘Met nieuwe biomarkers kunnen we er sneller bij zijn.’

Onleesbaar

De grootste barrière voor de ontwikkeling van nieuwe medicijnen is echter niet het vinden van nieuwe werkzame stoffen, maar de zoektocht naar proefpersonen. Vannieuwenhuyse: ‘Vind maar eens genoeg proefpersonen als je specifiek zoekt naar mannen tussen de 40 en 45, die niet roken en al vijf tot zeven jaar een bepaalde zeldzame ziekte hebben.’
De ontwikkeling van een nieuw medicijn wordt geregeld vertraagd doordat er niet genoeg testpersonen gevonden zijn. Zo komen medicijnen niet, of later, op de markt en zijn ze minder snel beschikbaar voor de zieke mensen die ze nodig hebben.
Ook hier kan het ontginnen van ongestructureerde teksten uitkomst bieden. Een berekende gok zegt dat tachtig procent van de patiëntengegevens ongestructureerd staat neergeschreven. Dokters, verplegers of apothekers houden dossiers over hun patiënten bij of schrijven verwijs- en ontslagbrieven met massa’s tekst en medische informatie.
‘Het cliché wil dat een doktersbriefje niet te lezen is’, zegt Provoost. ‘Een getypt dossier is nog moeilijker te ontwarren. Artsen gebruiken hun eigen persoonlijke afkortingen, maken typfouten en schrijven Nederlands en Engels door elkaar. Dat maakt het erg moeilijk om een computer aan te leren hoe hij die teksten moet doorzoeken.’
De eerste tests bij het doorzoeken van gestructureerde en ongestructureerde patiëntendata waren echter positief: ondanks alle uitdagingen konden onderzoekers de gevraagde proefpersonen vinden. Maar: wat met onze privacy? Mogen computers zomaar uw en mijn medische gegevens doorsnuffelen? Daar springen artsen vanzelfsprekend erg voorzichtig mee om. Dossiers worden altijd volledig geanonimiseerd en enkel uw behandelende arts kan uw dossier aan uw naam koppelen.

Bijwerkingen

Dan verschijnt het medicijn uiteindelijk op de markt en laten de mensen hun ervaringen los op Dokter Google. ‘Man, ik heb hoofdpijn. Hoofdpijn!’ Eén persoon die op een forum over hoofdpijn klaagt na het innemen van een medicijn is geen statistiek, maar vind er duizend en je weet dat er meer aan de hand is.
Steeds meer mensen spreken het internet aan als ze zich niet goed voelen. Als ze dat doen, laten ze waardevolle ongestructureerde informatie achter in de vorm van een forumpost. De computer kan die fora ontginnen en kijken of bijvoorbeeld mensen met hoofdpijn vertellen of ze de behandeling in combinatie met een ander medicijn hebben genomen of wat ze erbij hebben gegeten. Op die manier krijgen farmaceutische bedrijven een soort medicijnreview en kunnen ze snel inspelen op problemen of vragen.
‘Mensen beseffen niet altijd wat er allemaal mogelijk is’, vertelt Sien Moens, hoofd van het LIIR aan de KU Leuven. ‘Het internet is openbaar en alles wat er staat, kunnen we wel degelijk met een computer lezen en analyseren.’ Ook fora en sociale media (zie ‘Politie en sociale media’).
Maar ook als je het internet laat voor wat het is en gewoon naar je huisarts gaat, kan Text Mining helpen. Moens: ‘Stel dat een patiënt een symptoom vertoont dat ergens in een ongelezen publicatie wordt beschreven. De computer op de dokterspraktijk koppelt het symptoom dan automatisch met de niet-gelezen publicatie en kan zo de arts én de patiënt helpen en ondersteunen.’ Een dokter raadpleegt op die manier duizend onderzoekers tegelijk, waardoor de kans op een goede diagnose stijgt.
Een bestaande toepassing is de IBM Oncology Expert Advisor, die gebaseerd is op de Watsoncomputer. Het programma vat de lijvige dossiers van kankerpatiënten samen en geeft correcte aanbevelingen voor de behandeling en opvolging. Doordat het programma onbeperkt teksten ontgint, is het advies steeds gebaseerd op de nieuwste richtlijnen en wetenschappelijke literatuur. Minpuntje: de software is momenteel slechts tachtig procent van de tijd correct en dat is nog niet genoeg om het al in gebruik te nemen op de dokterspraktijk.

Een beetje ironie

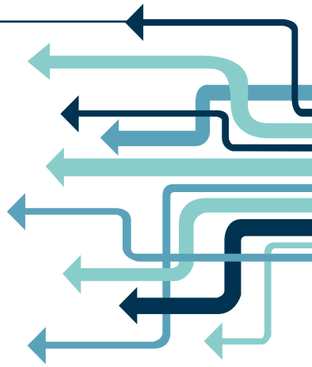
Er zijn dus al heel wat succesvolle tests gebeurd, maar de technologie om ongestructureerde teksten te doorzoeken, staat nog altijd in de kinderschoenen. Daelemans is echter optimistisch: ‘Over vijf à tien jaar zijn er echt bruikbare oplossingen.’ De heilige graal is daarbij een computer die zichzelf ongesuperviseerd dingen kan aanleren. Een computer die biochemie snapt als je hem een cursus aanreikt. ‘Pas dan zijn we vertrokken voor échte artificiële intelligentie.’ Moeten we dan bang zijn voor té slimme computers? ‘Nee hoor’, lacht Daelemans. ‘Zolang een computer niet zelfbewust wordt, is er geen probleem. De pc is en blijft een ondersteuning van de mens.’
Staat het antwoord op onze problemen dus in pdf’s die niemand leest? Waarschijnlijk wel. De Wereldbank ontdekte in mei 2014 dat maar liefst één derde van haar studies niet één keer is gedownload en dus waarschijnlijk nooit is gelezen. De studie die zegt dat haar pdf’s niet gelezen worden, hebben ze overigens verspreid als pdf. Het is het soort ironie waarvan je hoopt dat ze die zelf ook inzien. ■

En Google dan?

U denkt misschien: we kunnen toch al langer teksten doorzoeken? Ik gebruik Google elke dag. Dat is niet helemaal waar. Een zoekmachine geeft u voor uw zoekvraag inderdaad een massa relevante pagina’s en documenten, maar u moet zelf nog aan het werk om die teksten te begrijpen en er de zinvolle informatie uit te halen. Daarnaast kennen onderzoekers de meest relevante informatie al lang en leidt die maar zelden tot nieuwe inzichten.
‘Als je bijvoorbeeld zoekt naar ‘acteurs van tussen de twintig en dertig jaar oud’, geeft Google een hoop internetpagina’s waaruit je zelf de informatie moet halen’, verduidelijkt Thomas Provoost van het LIIR (KU Leuven). Text Mining combineert echter automatisch de kennis uit verschillende bronnen op één plaats en gaat op zoek naar nieuwe, zeldzame of verrassende informatie. En dat is net waar onderzoekers in geïnteresseerd zijn.

Politie en sociale media

Er is enorm veel ongestructureerde data te vinden in ons leven. Als vuistregel stelt men dat tachtig tot negentig procent van alle beschikbare data ongestructureerd is. De mogelijkheden die we aanboren door ze te ontginnen, lijken zo eindeloos dat een mens zich afvraagt waar het zal ophouden. ‘Wij zijn de uitvinders van technologie’, zegt Sien Moens, hoofd van het Language Intelligence & Information Retrieval onderzoeksteam aan de KU Leuven. ‘En zoals elke technologie kan die ten goede of ten kwade worden gebruikt.’



Op sociale media

Het is mogelijk om op basis van je berichten op Facebook automatisch je persoonlijkheid en leeftijd af te leiden. Vertel je op een forum dat je gaat wandelen in Noorwegen? Dan vertaalt een computer dat automatisch naar een nood aan warme wandelschoenen en toont hij je waar je die kan gaan kopen. Het maakt gericht adverteren makkelijker en verklaart waarom grote spelers als Google, Amazon en Facebook de techniek maar wat graag in de vingers zouden krijgen.

De politie

Een proces verbaal is een schoolvoorbeeld van een los uitgeschreven en ongestructureerde tekst. Het verslag wordt neergepend en in een archief gestopt. Als een computer echter honderdduizenden pv’s tegelijk kan doorbladeren, kan de politie links leggen tussen verschillende misdaden die voordien los van elkaar leken te staan of kan ze gevaarlijke verkeerssituaties beter inschatten aan de hand van verkeersongevallen uit het verleden.

▼ **De Watson-computer vat lijvige dossiers van kankerpatiënten samen en geeft aanbevelingen voor de behandeling.**