Taaltechnologie doorzoekt miljoenen online krantenartikelen

Samen dieren vinden in Delpher

In het SERPENS-project werken lexicologen, taaltechnologen en historisch ecologen samen om krantenartikelen over flora en fauna beter vindbaar te maken.

Marieke van Erp

Historisch ecologen bestuderen de impact van de natuur op het menselijk handelen door de tijd. Kranten zijn hiervoor een uitstekend medium omdat ze alledaagse gebeurtenissen beschrijven, frequent gepubliceerd worden en met dank aan de Koninklijke Bibliotheek gemakkelijk digitaal te doorzoeken zijn via Delpher.nl, de historische database van Nederland.

Zoekresultaten op 'wolf'

Relevante artikelen over flora en fauna vinden in miljoenen online artikelen is echter geen kwestie van een zoekterm invullen en dan de resultaten bekijken. Je moet rekening houden met de historische taalontwikkelingen in spelling en woordenschat, en er is een efficiënte manier nodig om de gewenste data uit Delpher te halen en voor onderzoek geschikt te maken. Bovendien kom je bij het zoeken naar bijvoorbeeld een 'wolf' deze niet alleen als lid van de soort Canis lupus tegen, maar ook als achternaam (John de Wolf), als locatie (Wolf Rock, 15 kilometer van Kaap Land's End) of bijvoorbeeld in de naam van een filmscript 'The Sea Wolf'. Alleen al tussen 1930 en 1939 levert het trefwoord 'wolf', zonder synoniemen, 66.809 resultaten op. Het is ondoenlijk om deze resultaten handmatig te analyseren.

Supervised machine learning

Uit het DiaMaNT-lexicon van het Instituut voor de Nederlandse Taal werden historische spellingvarianten en synoniemen van onze zoekterm gehaald. Vervolgens werden de data via de API van Delpher binnengehaald en in een databaseomgeving gezet. Een historisch lexicon maakte een inschatting van de OCR-kwaliteit. Om de relevante artikelen te selecteren, is een automatisch classificatiesysteem ingezet om het onderscheid dier versus geen dier te kunnen maken. Ook gaf het aanvullende informatie: bevat het artikel bijvoorbeeld een beschrijving van een jachtgebeurtenis of een verslag over materiële schade door een dier. Het systeem gebruikte een 'supervised machine learning'-algoritme: het leerde van een door een domeinexpert geannoteerde trainingsdataset door in artikelen karakteristieken te herkennen die tot een bepaalde classificatie leiden. Getest op nieuwe artikelen, wist het systeem in 92% van de gevallen correct te identificeren of een artikel al dan niet over een dier gaat. Uiteraard willen historisch ecologen nog veel meer weten. Door de computer de gigantische bak resultaten te laten filteren, kunnen zij zich richten op de diepere analyse van relevante artikelen.

SERPENS is een samenwerking tussen de Radboud Universiteit Nijmegen, het Instituut voor de Nederlandse Taal en het KNAW Humanities Cluster. SERPENS is mogelijk gemaakt door het NWO-project CLARIAH-CORE.

clariah.nl/projecten/research-pilots/serpens

Het woord 'wolf' levert vele zoekresultaten op. Hiernaast drie voorbeelden:

- ***Zoekplaatje Rood Kapje en de wolf***
  credits Provinciale Geldersche en Nijmeegsche courant. Nijmegen, 21-09-1935. https://resolver.kb.nl/resolve?urn=MMRANM02:000035671:mpeg21:a0105,
- ***Het wondere wolfje***
  credits Limburgsch dagblad. Heerlen, 03-12-1990. https://resolver.kb.nl/resolve?urn=ddd:010624151:mpeg21:a0083).
- ***De wolf in den Haagschen Dierentuin***
  credits Bredasche courant. Breda, 16-06-1932. https://resolver.kb.nl/resolve?urn=MMSAB03:000065143:mpeg21:a0057

Lowlandsgangers testen eerste hiphopgenerator

Machine learning: wat is echt en wat is nep?

Ruim 700 bezoekers testten afgelopen zomer de rapbot die ontwikkeld werd binnen het project Deepflow. Taaltechnoloog Folgert Karsdorp praat ons bij over de resultaten. *Mathilde Jansen*

Karsdorp, postdoc aan het Meertens Instituut, ontwikkelde samen met onderzoekers van de Universiteit Antwerpen een hiphopgenerator, de rapbot. Deze genereert zelf hiphopteksten op basis van 64.000 (vooral Amerikaanse) originele raps. Tijdens Lowlands, een drie dagen durend festival boordevol muziek, theater, film, comedy, literatuur en wetenschap, lieten ze de rapbot testen in een zogenaamde MC Turing-test. Bezoekers aan Lowlands Science zagen een tekst op een beeldscherm. Van elke tekst moesten ze beoordelen of de tekst een echte rap was of nep. Karsdorp wilde daarbij vooral weten: waar letten mensen op? En: werkt het bij iedereen hetzelfde?

Het antwoord op de laatste vraag is 'nee'. Uit de analyses blijkt dat er grote verschillen zijn tussen de antwoorden van zogenaamde amateurs en experts. "De eerste groep gokt maar wat", zegt Karsdorp, "terwijl de tweede groep let op talige eigenschappen van de tekst, zoals rijm, alliteratie en flow."

Letter- of woordniveau

In totaal werden 6 taalmodellen getest. Het eerste model werkt op letterniveau: het computermodel bepaalt per letter wat de meest waarschijnlijke letter is die volgt. Het tweede model werkt op woordniveau, het derde combineert beide modellen. "Een voordeel van het lettermodel is dat je te maken hebt met een klein vocabulaire", legt Karsdorp uit, "26 letters en nog allerlei interpunctie en diakritische tekens, zeg een stuk of 100 symbolen. Een nadeel van dit letter voor letter genereren, is dat je na 100 letters maar een paar woorden verder bent. Dan kun je beter overgaan op het woordniveau, maar daar is het probleem juist dat je vocabulaire heel groot wordt. In ons corpus hadden we bijvoorbeeld te maken met 380.000 unieke woorden. Naarmate je vocabulaire groeit, heeft de computer meer trainingsmateriaal nodig om de juiste voorspellingen te doen. Daarom hebben we een hybride model toegevoegd." Vervolgens werden alle modellen getest met en zonder bepaalde talige condities zoals de flow van de regels, rijmelementen en ritmische aspecten.

JONG TALENT

'Rapbot genereert zelf hiphopteksten'

Sterkste effect

Karsdorp: "Bij het lettermodel pikten de deelnemers de gegenereerde fragmenten er het makkelijkst uit. Bij het hybridemodel zat het slagingspercentage echter ongeveer op kansniveau. Als er talige condities aan het model waren toegevoegd, presteerden alle modellen beter – en werd het dus moeilijker voor de deelnemers. Maar het sterkste effect hadden ze op het woordmodel. Ritmische aspecten zitten op het lettergreepniveau, en vallen daardoor het beste samen met het woordmodel."

Karsdorp gebruikt de resultaten van dit onderzoek om nieuwe modellen voor culturele verandering te ontwikkelen. Niet alleen in taal, maar ook bijvoorbeeld in melodie. Voor zijn onderzoek put hij onder andere uit de liederenbank en volksverhalenbank van het Meertens Instituut.

deep-flow.nl

Folgert Karsdorp studeerde Nederlands en Taalkunde in Leiden en promoveerde in 2016 cum laude aan de Radboud Universiteit. Voor zijn proefschrift ontwikkelde hij een computermodel om de hervertellingen van verhalen te simuleren. *foto Evy van Schelt*