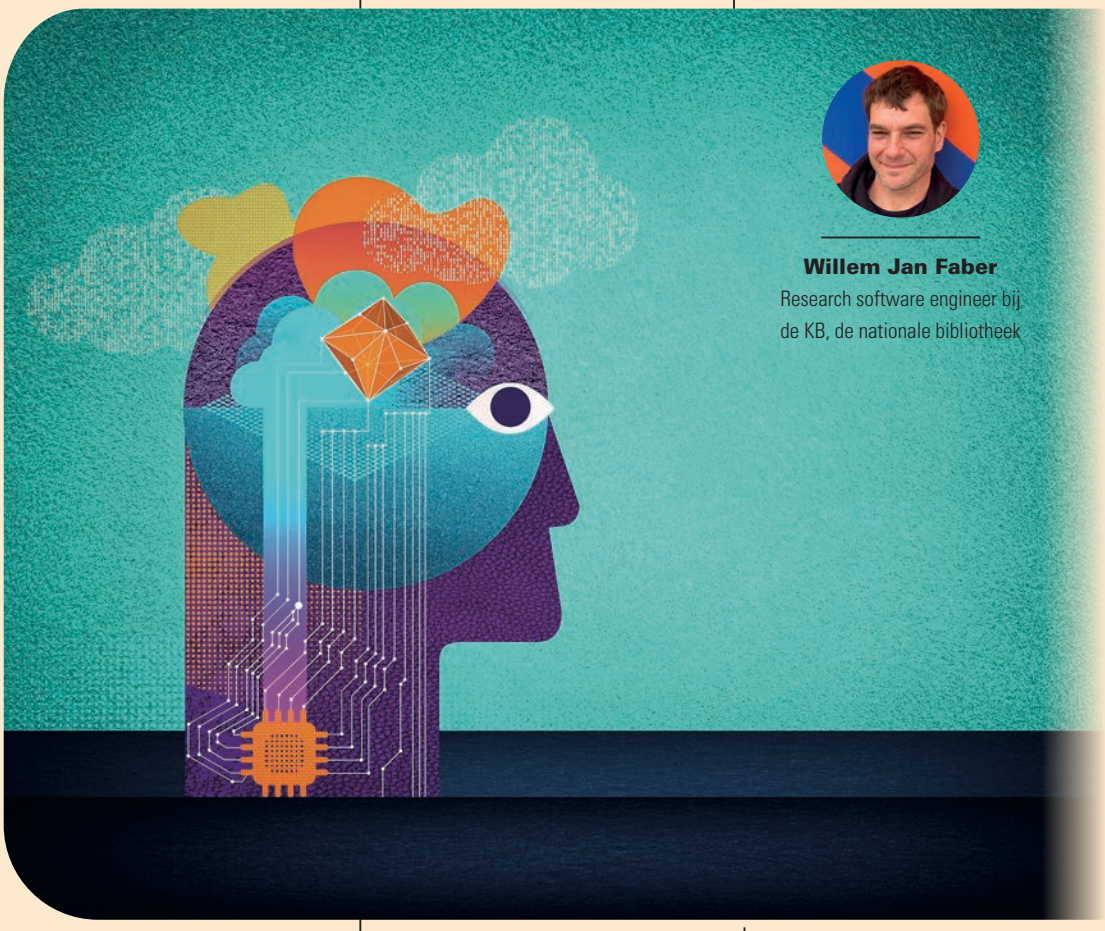
WAAR TAALMODEL EN GAMINGHARDWARE ELKAAR ONTMOETEN

Fijnmazig zoeken in de collecties mogelijk maken en al deze data aan elkaar en aan andere bronnen verbinden, zo luidt een van de doelen van de KB. De huidige ontwikkelingen in NLP en AI, en 'wat tweak-werk', helpen de nationale bibliotheek daarbij.

In de rubriek 'KB Onderzoekskroniek' beschrijven medewerkers van de afdeling Onderzoek van de Koninklijke Bibliotheek hun resultaten, trends en vondsten.

Geen probleem, hoe simpel het ook lijkt, is daadwerkelijk simpel, als je maar lang genoeg inzoomt. Zo is het voor de lezers van dit stuk nog vrij eenvoudig om in de volgende zin te kunnen aangeven wat een zogenoemde entiteit – een persoon, locatie of organisatie – is en wat niet: *Ook hoogleraar staatsrecht Wim Voermans wijst erop dat het verwijderen van de sms'jes door Rutte niet mag vanuit zijn kantoor te Den Haag*. Het potlood mag weer weg, goed gedaan! De personen Rutte en Wim Voermans en de plaats Den Haag moeten inderdaad worden aangestreept. Het wordt echter al snel ingewikkeld. Kijk bijvoorbeeld naar deze zin: *Best is volgens mevrouw Van den Haag de leukste plek op de wereld*. Weg valt alle wetmatigheid waar je je in de vorige zin aan vasthield. Best is hier een plaats en mevrouw Van den Haag een persoonsnaam. En dit is nog een relatief simpel voorbeeld, want in historische documenten is het aantal uitzonderingen niet op te sommen.

'Gamingvideokaarten zijn in staat een taalmodel te genereren binnen dagen, terwijl dit met een standaard processor weken tot maanden duurt'



Willem Jan Faber
Research software engineer bij de KB, de nationale bibliotheek

NLP en NER

De taak van het vaststellen van entiteiten wordt binnen het vakgebied natural language processing (NLP) ook wel named entity recognition (NER) genoemd. Er zijn vele pogingen gedaan om de wetmatigheid en de structuur van zinnen te ontrafelen met code. Dit maakt het mogelijk om geautomatiseerd grote hoeveelheden tekst te ontsluiten met zeer fijnmazige zoekmogelijkheden. Bovendien stelt het je in staat om op basis van de gevonden entiteiten bronnen aan elkaar te verbinden.

Reusachtige rekenkracht

De laatste jaren zijn er grote doorbraken geweest binnen het vakgebied NLP. Dat komt doordat kunstmatige intelligentie in hoog tempo volwassen is geworden. Bedrijven als NVIDIA, Google/DeepMind, OpenAI, Microsoft en IBM tillen deze ontwikkelingen razendsnel naar grote hoogten. Zelfs de kleintjes doen mee. Zo heeft webshop Zalando een eigen NLP-engine.

De grote bedrijven storten zich met reusachtige rekenkracht op gigantische datasets. Neem het BERT-taalmodel van Google, dat onder andere wordt gebruikt in Google Translate, Alexa en Siri. Het wordt getraind met Wikipedia (zo'n 2,5 miljard woorden) en het Google Books-corpus (zo'n 80 miljard woorden) in meer dan tien talen. Om dit te kunnen verwerken heeft Google

eigen specifieke AI-hardware ontwikkeld en 64 zware systemen vier dagen laten draaien.

Snellere videokaart

De speciale hardware van Google is gemaakt om gigantisch veel matrix-/vectorberekeningen te doen, vergelijkbaar met een klassieke videokaart. In een 3D gamewereld is de videokaart verantwoordelijk voor het genereren van zestig beelden per seconde. Het toeval wil dat de berekeningen die worden uitgevoerd tijdens het trainen van een taalmodel lijken op die van het genereren van een 3D gamewereld. De videokaart is efficiënter in een specifiek soort wiskunde dan de standaard processor. Dit verschil is goed te merken als je met grote taalmodellen gaat werken; de duurdere gamingvideokaarten zijn in staat een taalmodel te genereren binnen dagen, terwijl dit met een standaard processor weken tot maanden duurt.

Indonesisch taalmodel

Binnen de KB profiteren we enorm van deze grootschalige NLP-onderzoeken. Veel van de resultaten worden vrijgegeven en zijn met wat tweak-werk ook in te zetten voor onze doelen: fijnmazig zoeken in onze collecties mogelijk maken, en deze collecties zo goed mogelijk aan elkaar en aan andere bronnen verbinden.

Om te kunnen profiteren van de recente ontwikkelingen gebruiken we bij de KB regelmatig het rekencentrum van SURF (surf.nl). Binnen de SURF Research Cloud-omgeving is het mogelijk een beroep te doen op zware videokaarten die nodig zijn om de bestaande taalmodellen te hergebruiken. Zo hebben we voor onze researcher in residence Simon Kemper een Indonesisch taalmodel ontwikkeld op basis van Googles BERT. De resultaten die we hebben behaald, zijn ronduit verbluffend. Wil je precies weten hoe goed de aanpak werkt, check dan Simons blog: tinyurl.com/KBbert. <