Informatica bij de Nationale Politie

Hoi! Wat leuk dat je vandaag op de open dag komt kijken. Ik ben Daphne Odekerken, en in 2010 besloot ik om informatica te gaan studeren in Utrecht. Ik had op de middelbare school een beetje leren programmeren en koos voor informatica omdat ik graag technisch maar ook creatief bezig ben en informatica geeft deze mogelijkheid. Informatica in Utrecht is erg veelzijdig! Tijdens mijn studie werkte ik bijvoorbeeld aan een programma dat logische puzzels genereert en/of oplost, maakte ik een systeem voor automatische analyse van Facebookdata en ontwierp en implementeerde ik een systeem voor het automatisch herkennen van akkoorden in muziek. Inmiddels ben heb ik mijn bacheloropleiding Informatica en master Computing Science afgerond, maar ik ben nog steeds vaak op de universiteit te vinden! Ik doe namelijk een parttime PhD en daarnaast werk ik parttime bij de Nationale Politie. Op beide plekken werk ik aan hetzelfde onderwerp: dialoogsystemen. Kort gezegd zijn dat computerprogramma's waarmee je kunt communiceren; denk bijvoorbeeld aan Apple's Siri of Google Assistant.

Een onderdeel van mijn onderzoek gaat over automatische tekstclassificatie. Dat houdt in dat je een computerprogramma automatisch laat bepalen of een tekst in een vooraf bepaalde categorie hoort.

Er zijn allerlei methoden om automatisch teksten te classificeren. Een relatief eenvoudige methode is het gebruiken van een linear model op de "bag-of-words" representatie van de tekst. Laten we eens kijken naar een voorbeeld vanuit mijn werk bij de politie: hier werk ik o.a. aan een classifier die herkent of er in een tekst staat dat een product geleverd is. Dat is belangrijke informatie bij het afhandelen van aangiftes van mensen die opgelicht zijn via websites zoals Marktplaats of eBay.
De bag-of-words representatie van de tekst is simpelweg een lijstje waarin je het aantal voorkomens van elk woord telt. Bijvoorbeeld: de bag-of-words van de zin "Ik heb een playstation besteld maar heb die nooit gekregen." is:

ik (1)	heb (2)	een (1)	playstation (1)	besteld (1)	maar (1)	die (1)	nooit (1)	gekregen (1)

Wat is de bag-of-words van de volgende zinnen:
1) "Ik heb de computer gisteren ontvangen en ik heb hem meteen getest."
2) "Ik heb antwoord op mijn vraag gekregen."

Een lineair model is een soort wiskundige formule die aan elk relevant woord een waarde toekent en combineert in één einduitkomst. In ons voorbeeld voorspelt de classifier de categorie "product geleverd" als die waarde een bepaalde drempel overschrijdt. Een mogelijke classifier zou bijvoorbeeld de categorie "product geleverd" toekennen als de uitkomst van volgende formule hoger is dan 0.5:

$$0.6 * F_{ontvangen} + 0.6 * F_{gekregen} + 0.6 * F_{geleverd} - 0.2 * F_{niet} - 0.2 * F_{nooit}$$

In deze formule staat F_{woord} voor het aantal keren dat dit $woord$ in de tekst voorkomt. Dat haal je uit de bag-of-words. Onze classifier zou voor de eerste voorbeeldzin (over de playstation) de waarde $0.6 * 0 + 0.6 * 1 + 0.6 * 0 - 0.2 * 0 - 0.2 * 1 = 0.4$ berekenen. Dat is minder dan 0.5, dus de classifier wijst de categorie "product geleverd" **niet** toe.

Welke categorie voorspelt de classifier voor de andere twee voorbeeldzinnen?
Wat denk je, heeft de classifier beide keren gelijk? Waarom wel of niet?

Antwoorden

1) *"Ik heb de computer gisteren ontvangen en ik heb hem meteen getest."*

ik (2)	heb (2)	de (1)	computer(1)	gisteren (1)	ontvangen (1)	hem (1)	meteen (1)	getest (1)

2) *"Ik heb antwoord op mijn vraag gekregen."*

ik (1)	heb (1)	antwoord (1)	op (1)	mijn (1)	vraag (1)	gekregen (1)

Voorspelling van classifier
1) "Ik heb de computer gisteren ontvangen en ik heb hem meteen getest."
$0.6 * 1 + 0.6 * 0 + 0.6 * 0 - 0.2 * 0 - 0.2 * 0 = 0.6$. Dat is meer dan 0.5, dus de classifier kent de categorie "product geleverd" **wel** toe.
2) "Ik heb antwoord op mijn vraag gekregen."
$0.6 * 0 + 0.6 * 1 + 0.6 * 0 - 0.2 * 0 - 0.2 * 0 = 0.6$. Dat is meer dan 0.5, dus de classifier kent de categorie "product geleverd" **wel** toe.

Evaluatie voorspelling van classifier
In de eerste zin heeft de classifier gelijk, want er staat inderdaad dat er een product geleverd is. Bij de tweede zin gaat het niet goed: de classifier denkt dat er een product geleverd is maar eigenlijk is dat niet zo: het lijdend voorwerp horend bij "gekregen" is "antwoord", maar dat is geen product. Hier zie je dat een lineair model op de bag-of-words representatie niet altijd het juiste antwoord geeft. In de praktijk worden dergelijke problemen opgelost door bijvoorbeeld eerst een (automatische) grammaticale analyse op de tekst te doen.