

Instructions

Go to

www.menti.com

Enter the code



Or use QR code

Welkom bij de workshop:

De training van een taalmodel

Wie zijn wij?

- **Miranda Jorna, Wouter van Willegen en Jesper Klop**
- **Studenten Finance & Control + Technische bedrijfskunde**
- **Minor Datadriven Solutions**



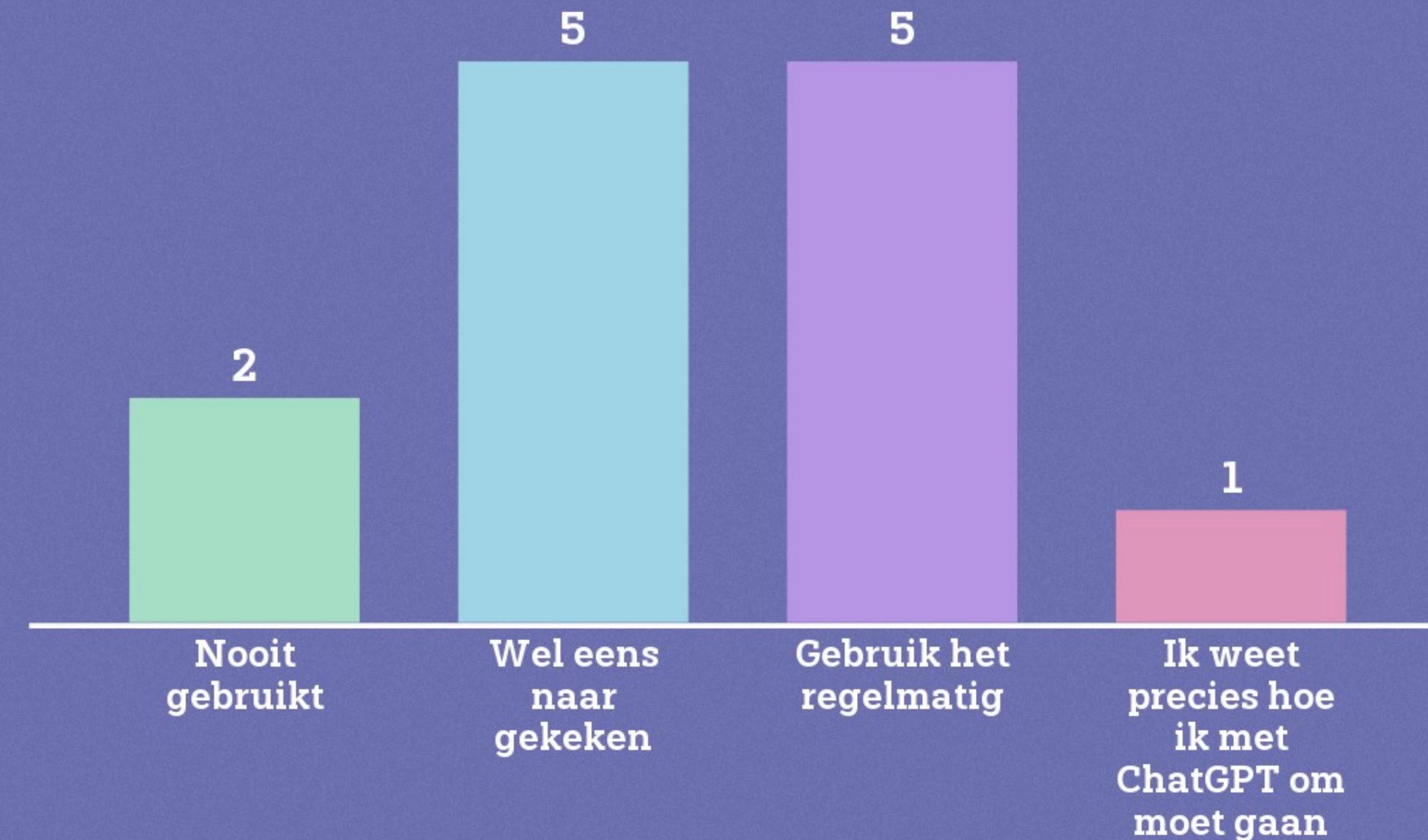
PLAN:



Inhoud workshop

- Het verzamelen van gegevens
- Het identificeren en scrapen van tekstgegevens
- Voorbewerking van de tekstgegevens
- Het creëren van een taalmodel
- Het evalueren van een taalmodel

Hoeveel ervaring heb je met ChatGPT?



1. Het verzamelen van gegevens

- Kwaliteit en kwantiteit
- Enorme hoeveelheid kennis
- Machine learning
- Bias in ChatGPT

Waar denk je aan bij bias dat in ChatGPT zou kunnen zitten?

culturele 'aannames'

verkeerde bronnen
vooroordelen

negatief over vrouwen

eenzijdige info ethiek
racisme

negatief over gekleurde m

Het identificeren en scrapen van tekstgegevens

- Niveaus tokenization
- Lowercasing
- Stemming and Lemmatization



Tokens	Characters
302	782

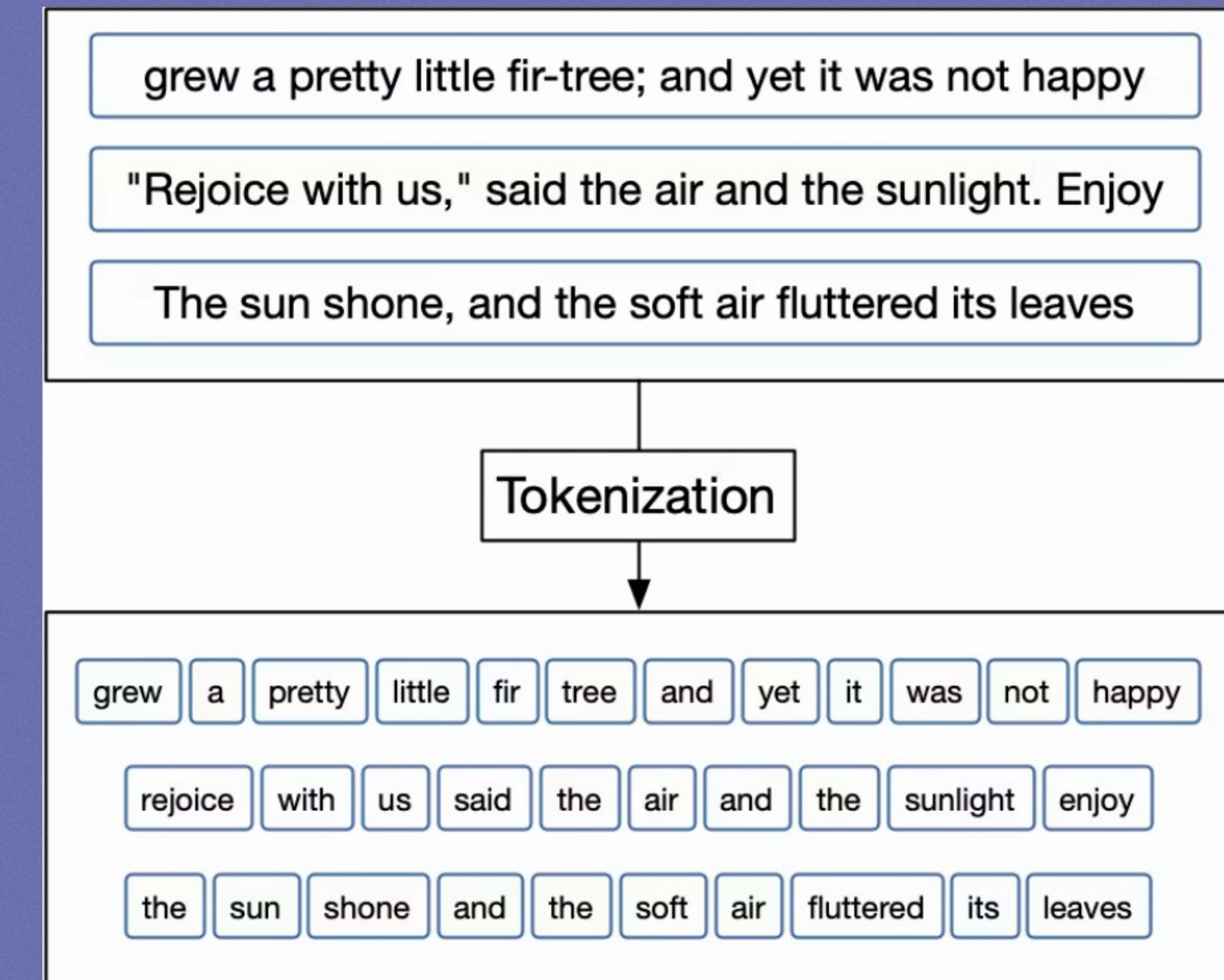
Een gazelle is een van de vele antilopensoorten in het geslacht *Gazella* / ♀ ♂ ' z ♀ l ♀ / . [2] Dit artikel behandelt ook de zeven soorten die deel uitmaken van twee andere geslachten, *Eudorcas* en *Nanger*, die vroeger werden beschouwd als ondergeslachten van *Gazella*. Een derde voormalige onderklasse, *Procapra*, omvat drie levende soorten Aziatische gazzellen.

Gazellen staan bekend als snelle dieren. Sommigen kunnen rennen met bursts tot 100 km / u (60 mph) of rennen met een aanhoudende snelheid van 50 km / u (30 mph). [3] Gazellen worden meestal gevonden in de woestijnen, graslanden en savannes van Afrika; maar ze komen ook voor in Zuidwest- en Centraal-Azië en het Indiase subcontinent. Ze leven meestal in kuddes en eten fijne, licht verterbare planten en bladeren.

TEXT

TOKEN IDS

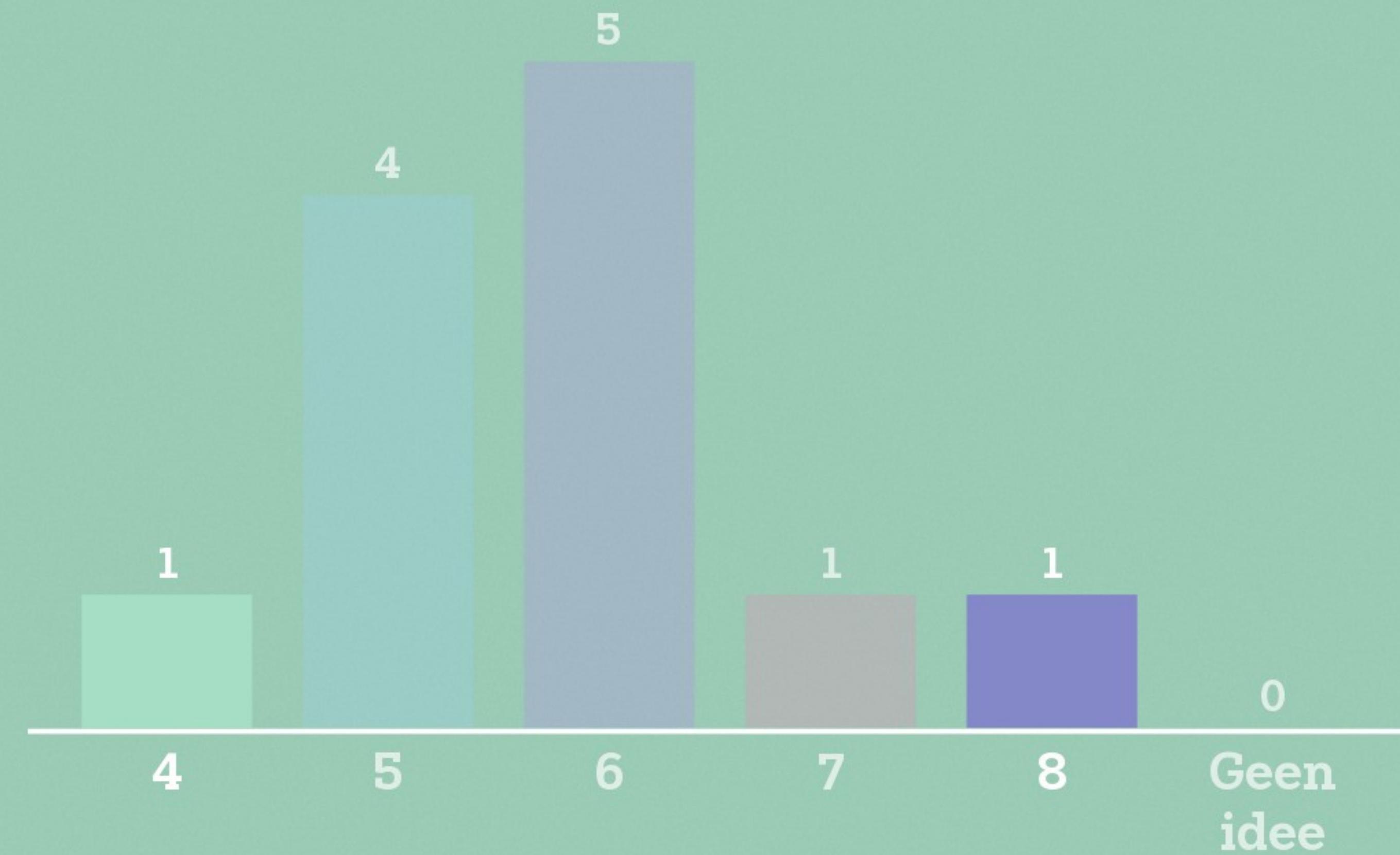
[HTTPS://PLATFORM.OPENAI.COM/TOKENIZER](https://platform.openai.com/tokenizer)**ChatGPT**



[HTTPS://SMLTAR.COM/TOKENIZATION.HTML](https://SMLTAR.COM/TOKENIZATION.HTML)

De basis

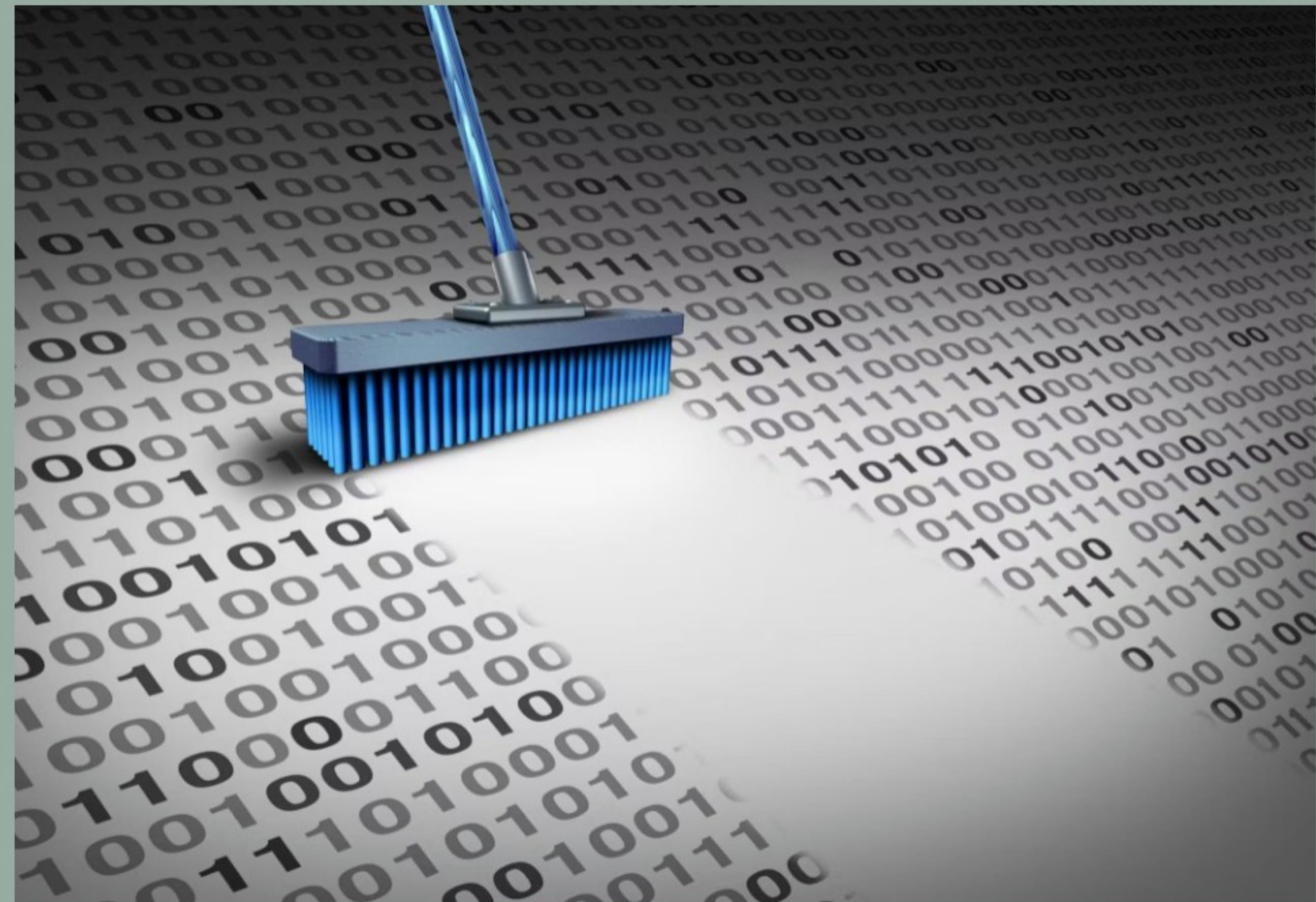
Hoeveel tokens heeft de volgende zin: "Mijn naam is Wouter!"



Stemming and Lemmatization

Het verkleinen van de dataset en het vergroten van de identificeerbaarheid van woorden. door het reduceren van woorden naar de basis vorm. Bijvoorbeeld: vliegen en gevlogen worden gereduceerd naar vlieg.

Voorbewerking van de tekstgegevens

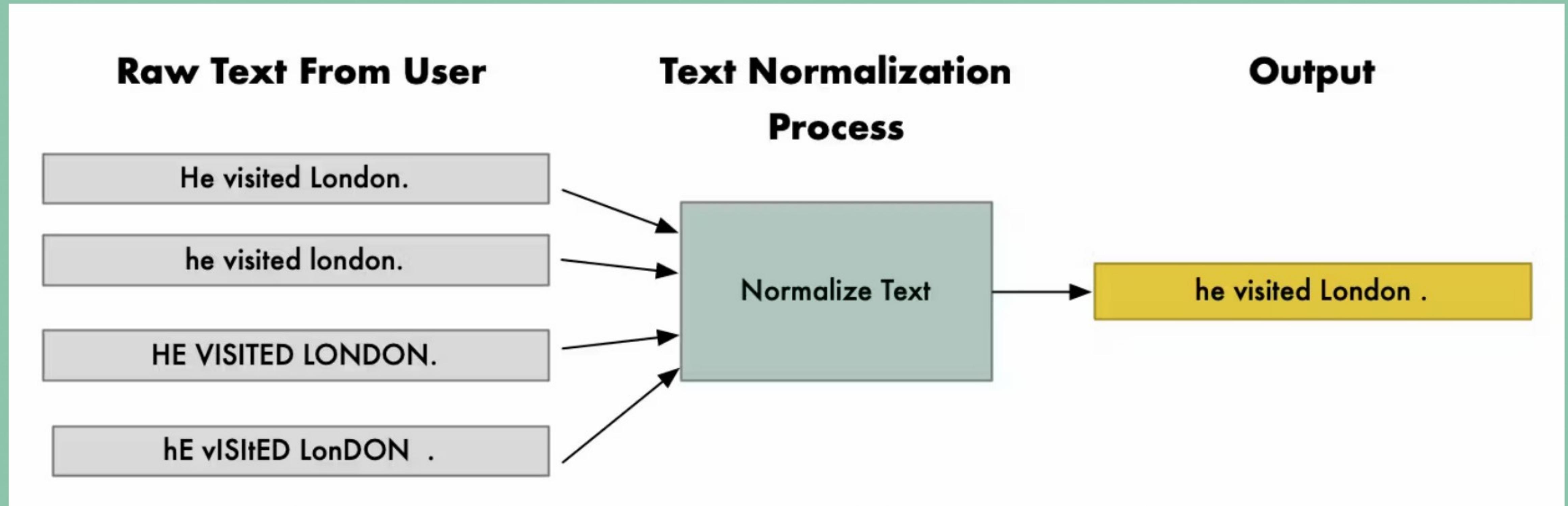


Tekstreiniging

Voorbewerking van de tekstgegevens

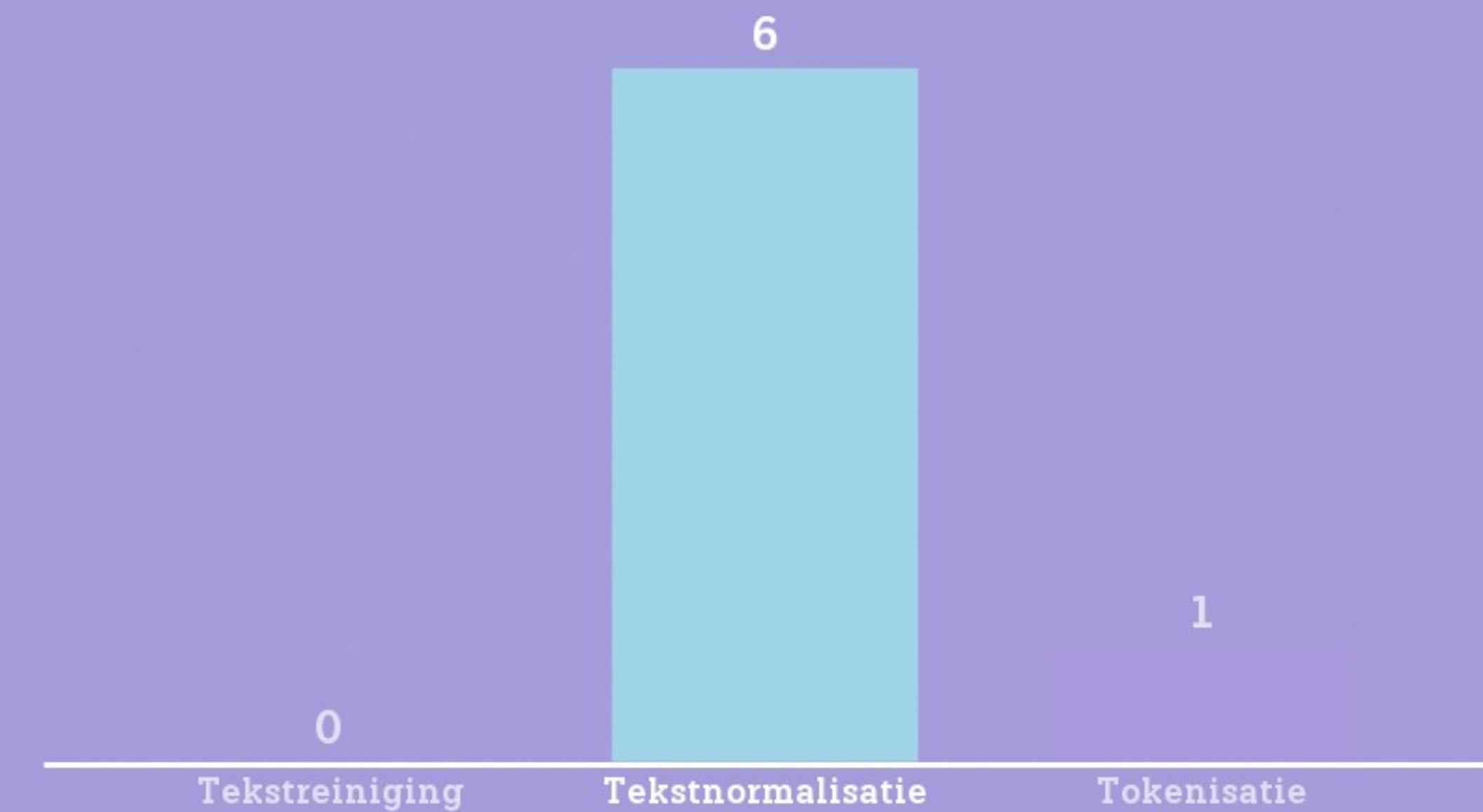
- Woordembeddings
- Gegevensaugmentatie
- Datasetbalancing
- Verwijderen van ruis





Tekstnormalisatie

Welke stap in het proces van voorbewerking van tekstgegevens is gericht op het omzetten van tekst naar een gestandaardiseerd formaat?

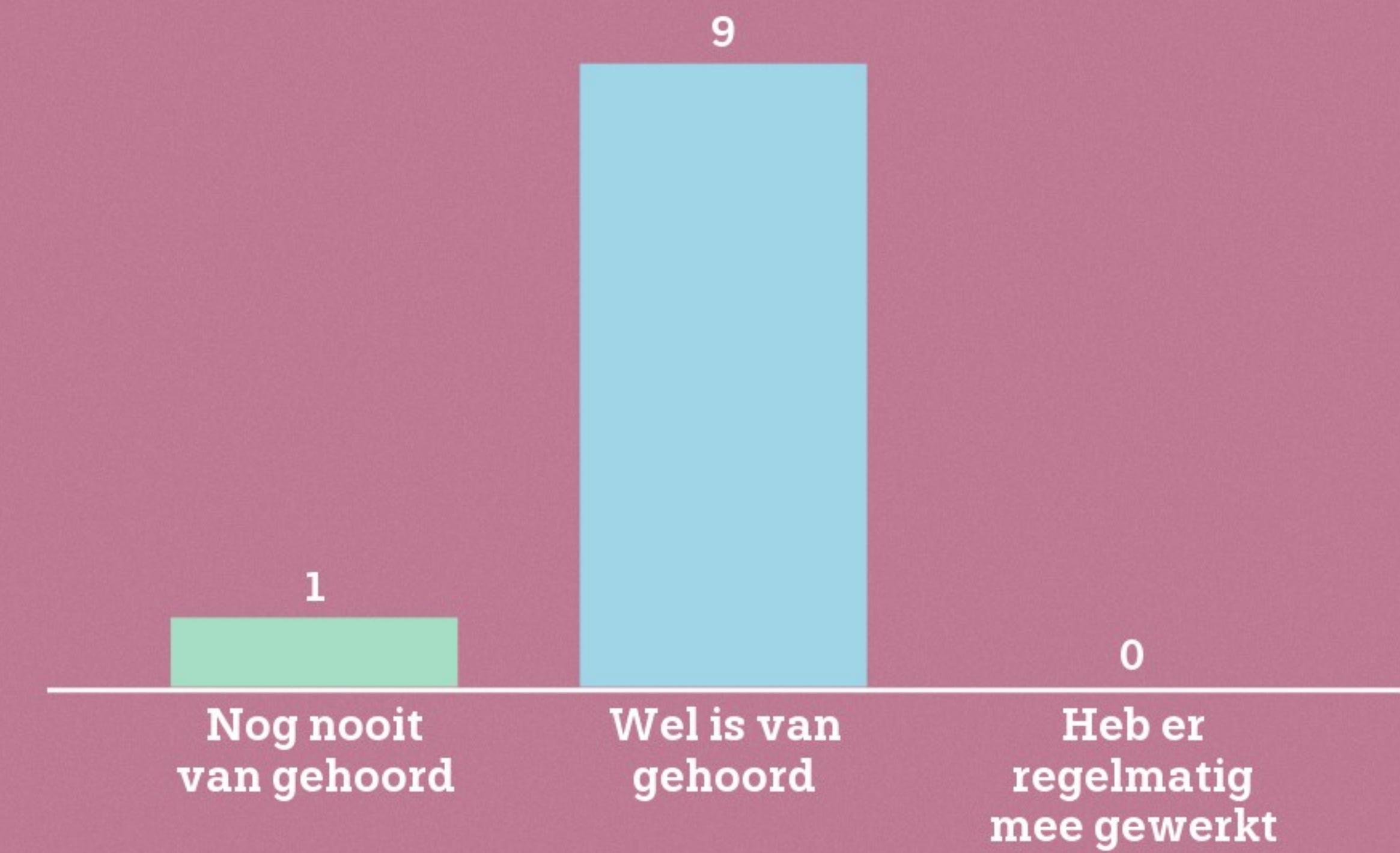


Het creëren van een taalmodel

→ Machine learning

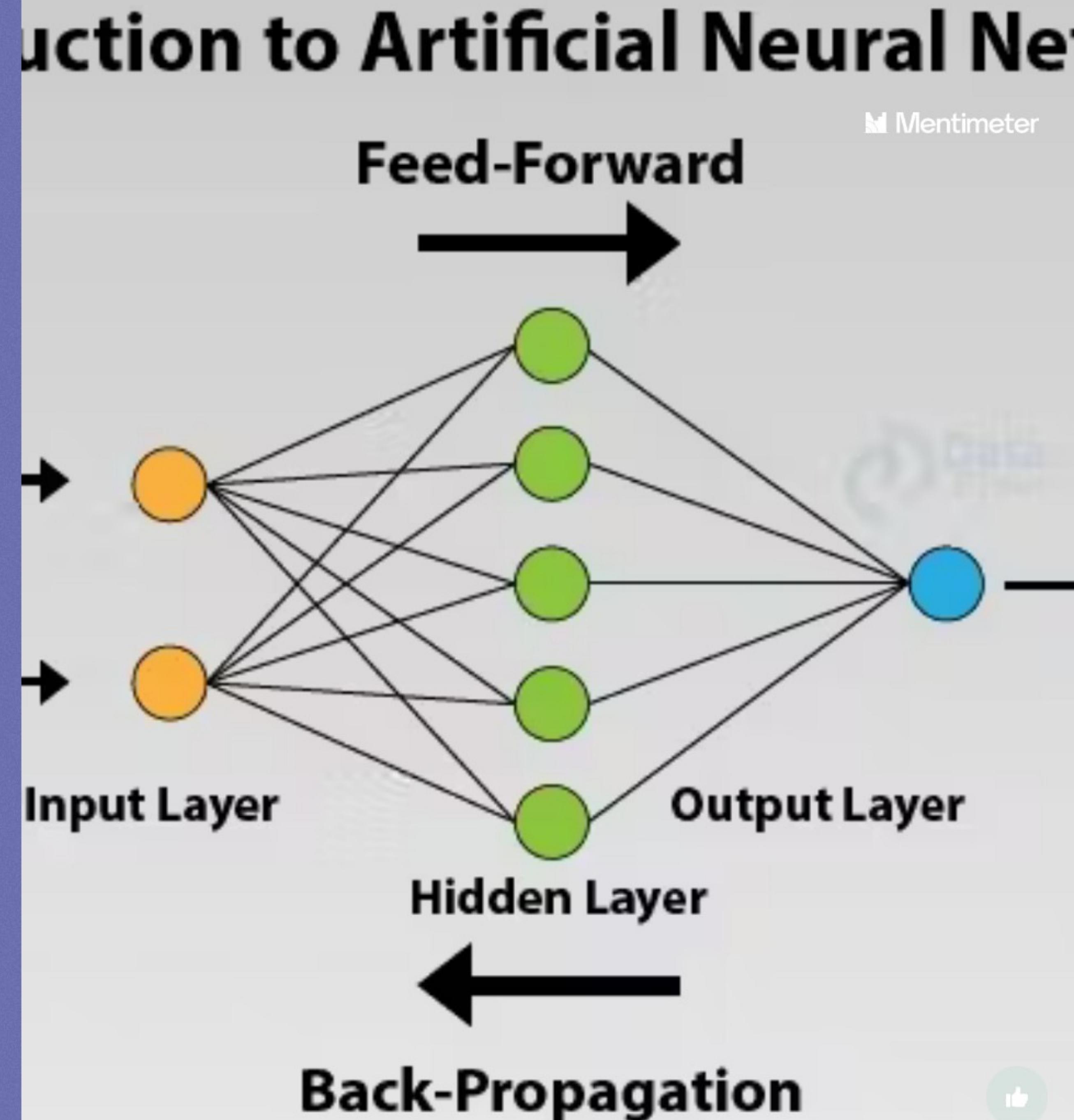


Weet je wat machine learning is en wat dit inhoudt

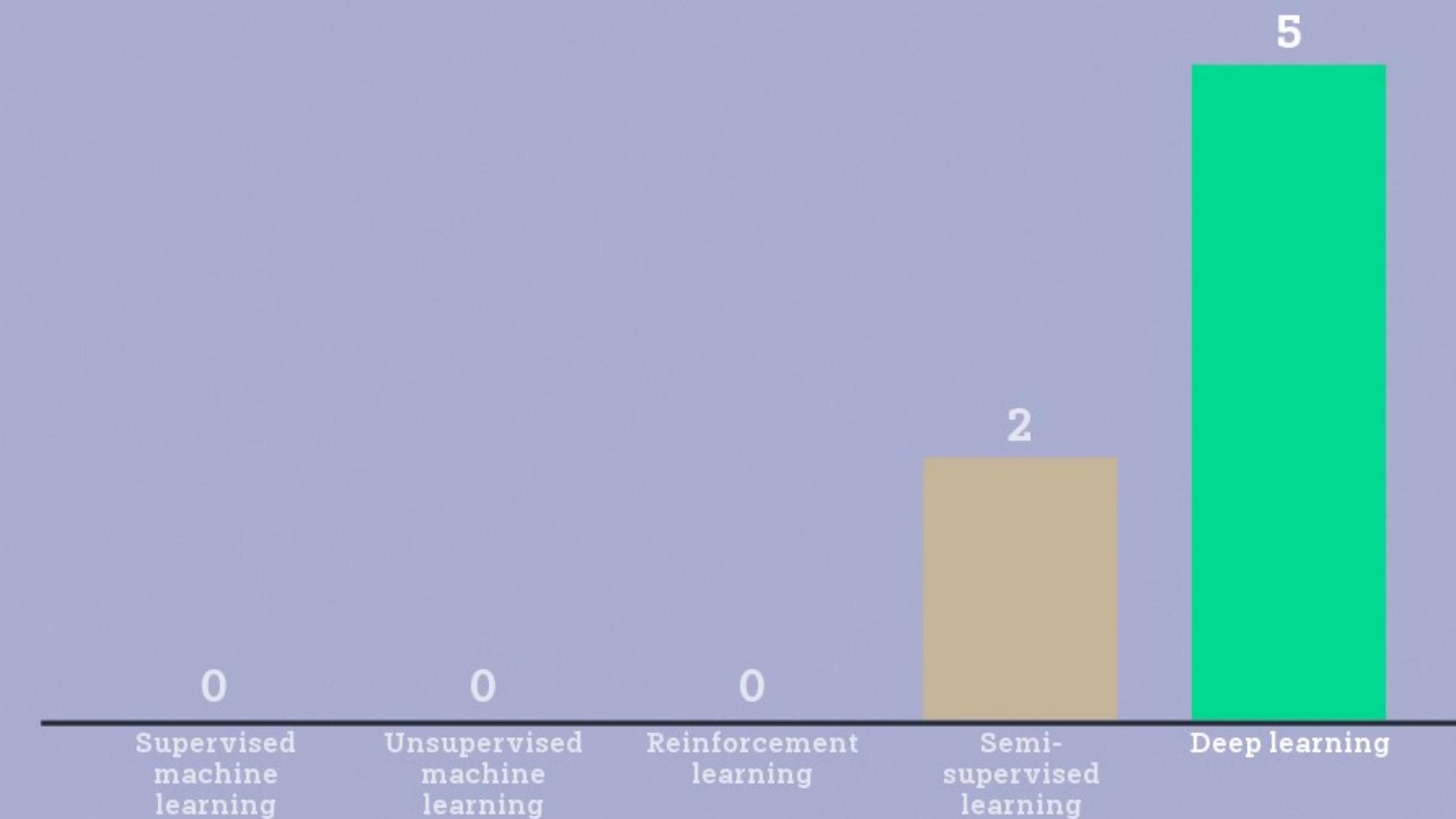


Verschillende vormen van machine learning

- Supervised machine learning
- Unsupervised machine learning
- Reinforcement learning
- Semi-supervised learning
- Deep learning



Welke vorm van machine learning wordt door ChatGPT gebruikt



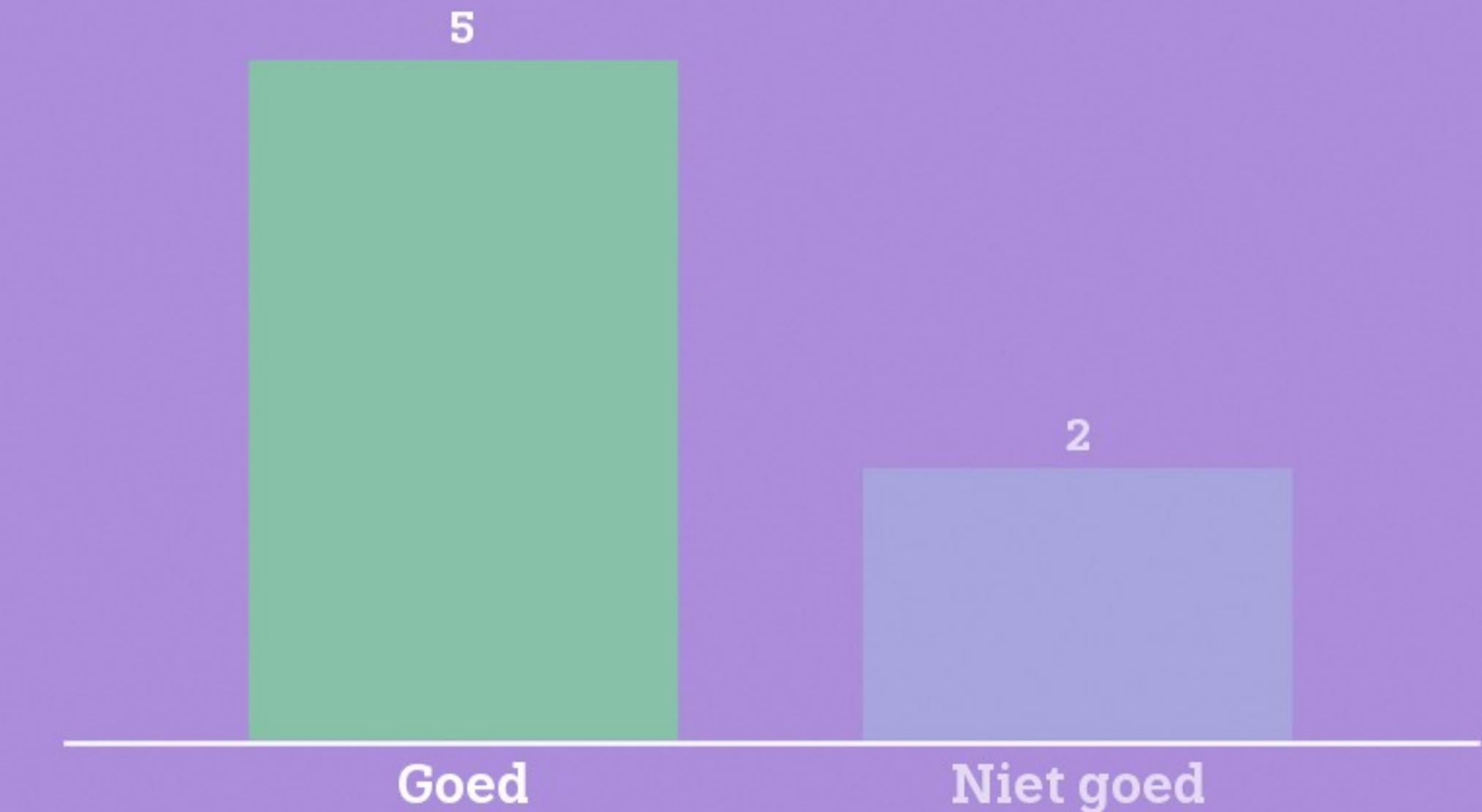
EVALUATIE



Het evalueren van een taalmodel

- De perplexiteit
- De F1-score
- Kwalitatieve evaluatiemethode

Ik vertelde net dat ChatGPT een lage perplexiteit heeft van 20. Is dit juist goed of niet goed?



Dit is het einde van onze seminar. Zijn er vragen?



Enquête ChatGPT seminar

