Sentence embeddings, je nieuwe beste vriend



(of: waarom eigen taalmodellen nog altijd nuttig zijn)

dr. ir. Thomas Winters

Postdoctoraal onderzoeker @ KU Leuven

thomaswinters.be

Slides: thomaswinters.be

Wie kent sentence embeddings?



Wie heeft sentence embeddings al zelf gebruikt?



Thomas Winters

Onderzoeker in creatieve AI & taalmodellen aan KU Leuven sinds 2018

RobBERT taalmodel met Pieter Delobelle (2020)
(state-of-the-art Nederlandstalige BERT taalmodel)

Werk met GPT-modellen sinds 2019
(o.a. voor humor, Improbotics...)

Creative AI workshops & advies

KU LEUVEN LEUVEN.AI DTAI




Wat is een sentence embedding?

Probleem

Welke twee van deze drie zinnen zijn meest gelijkaardig qua betekenis?

1: *“The train is quite noisy”*

2: *“Steam locomotives are loud”*

3: *“I train the quite pretty dog”*

Hoe weet je dat? Hoe kan je dat kwantificeren?

→ **Sentence embedding:**

bereken “coördinaat” op basis van betekenis

Woord embeddings

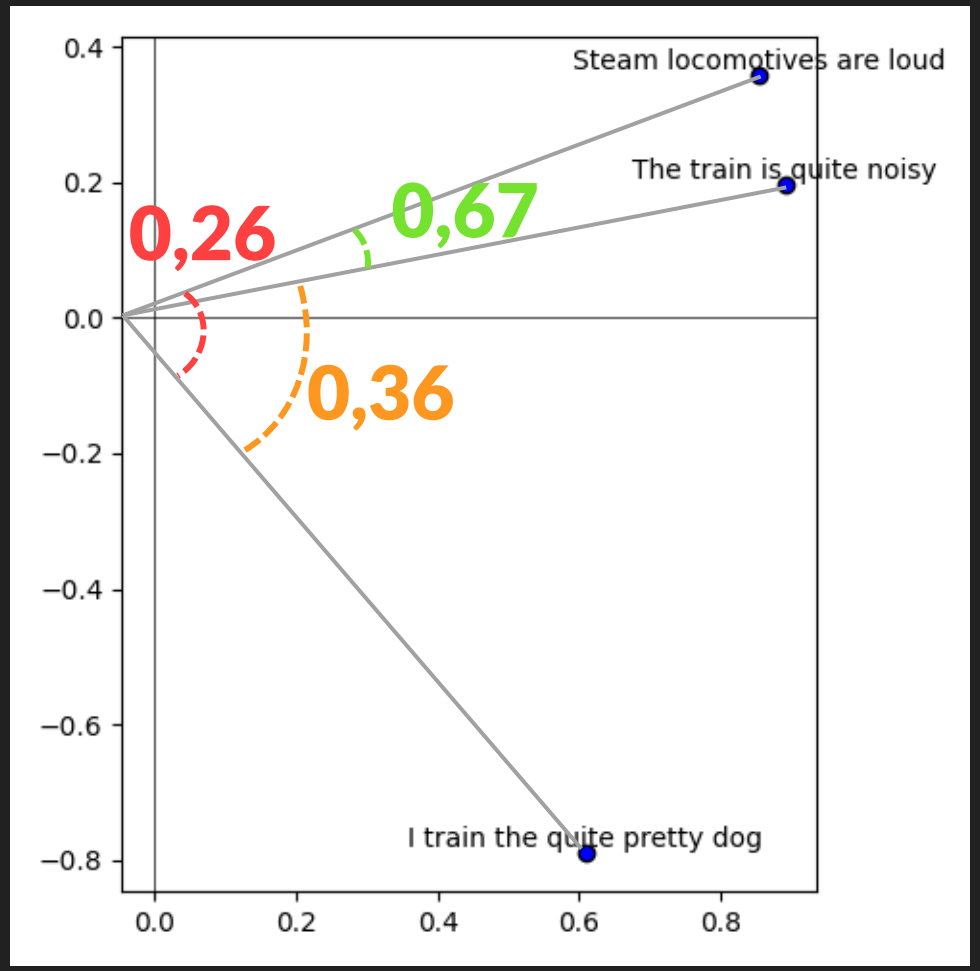


Leer welke woorden in zelfde context voorkomen

Elke as = een woord
Daarna: reductie van assen (PCA)

Niet vogelvlucht afstand, maar hoek

Grote taalmodellen: Sentence embeddings



Cosine similarity:

$$\frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \cdot \sqrt{\sum_{i=1}^{n} B_i^2}}$$

→ Sentence embeddings vinden wel degelijk dat de eerste twee zinnen het meest gelijkaardig zijn

Calculating Embeddings

Through APIs

(e.g. OpenAI)

```python
from openai import OpenAI
client = OpenAI()

response = client.embeddings.create(
    input="I train the quite pretty dog",
    model="text-embedding-3-small"
)
```

https://platform.openai.com/docs/guides/embeddings

Locally

(sentence-transformers)

```python
from sentence_transformers import SentenceTransformer

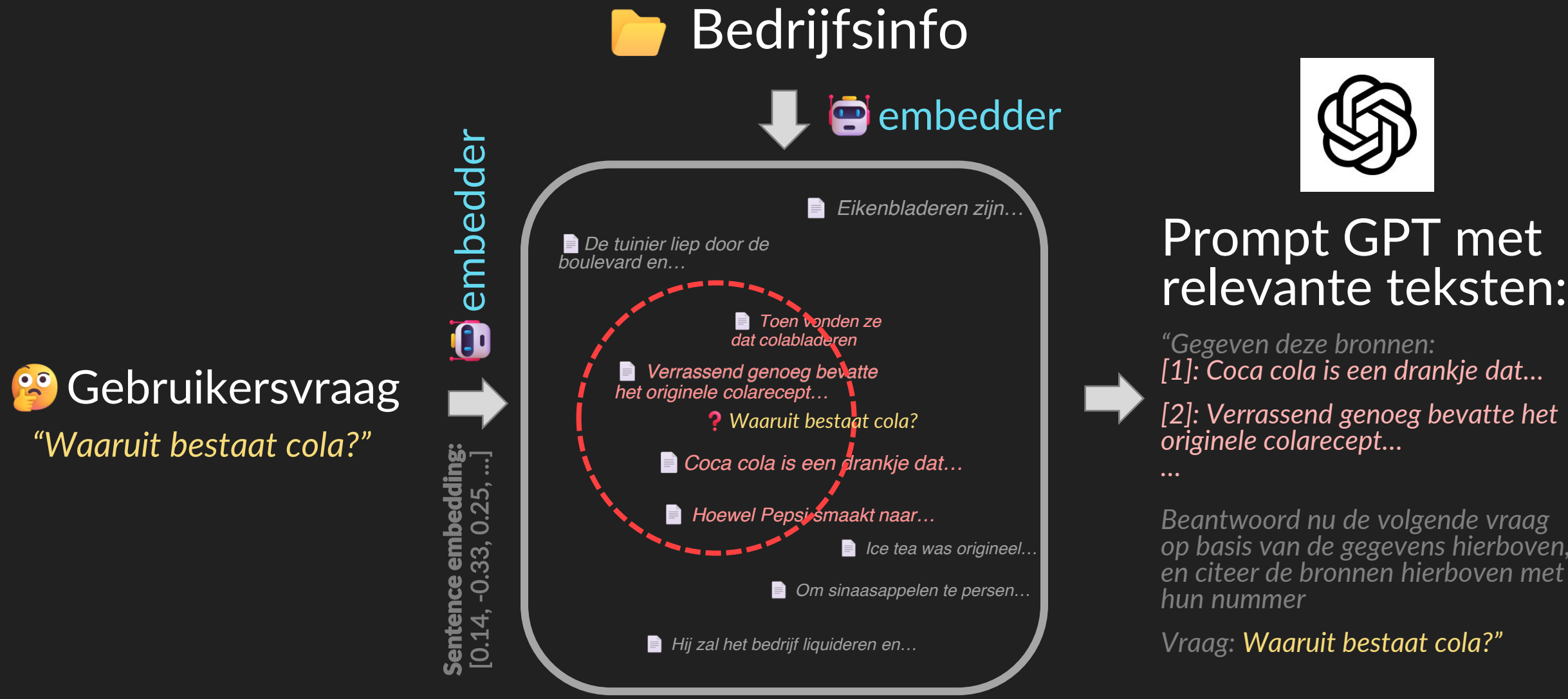
model = SentenceTransformer("all-MiniLM-L6-v2")
embeddings = model.encode(["I train the quite pretty dog"])
```

https://sbert.net/

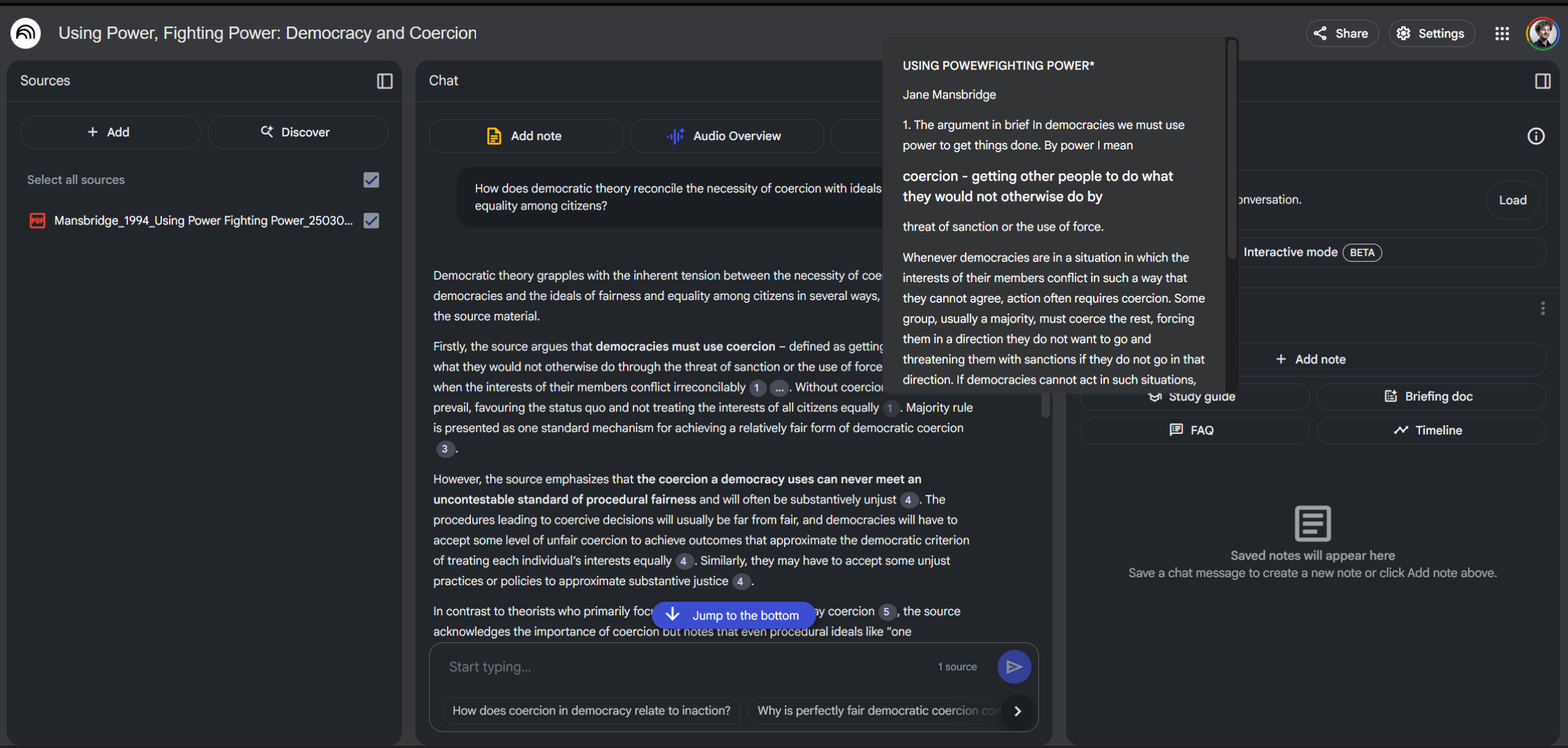

Toepassingen met embeddings


Probleem 1: GPT kent je bedrijfsdata niet & hallucineert

Oplossing: Retrieval-Augmented Generation (RAG)



NotebookLM: Google's RAG interface

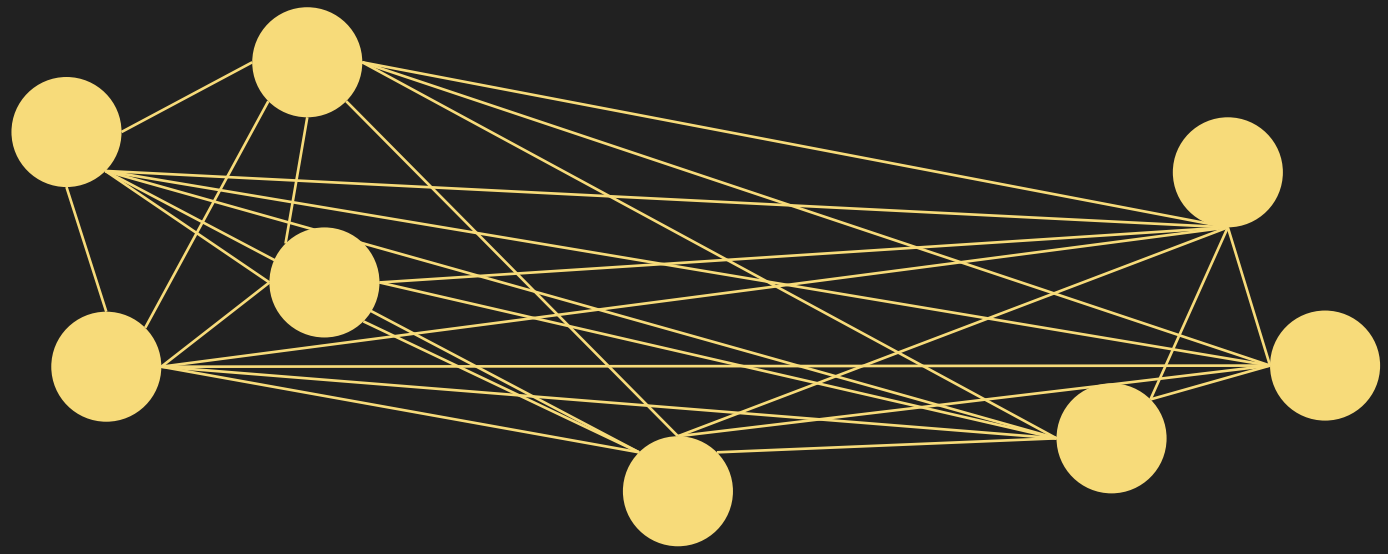




Probleem 2: Ongestructureerde tekstuele data

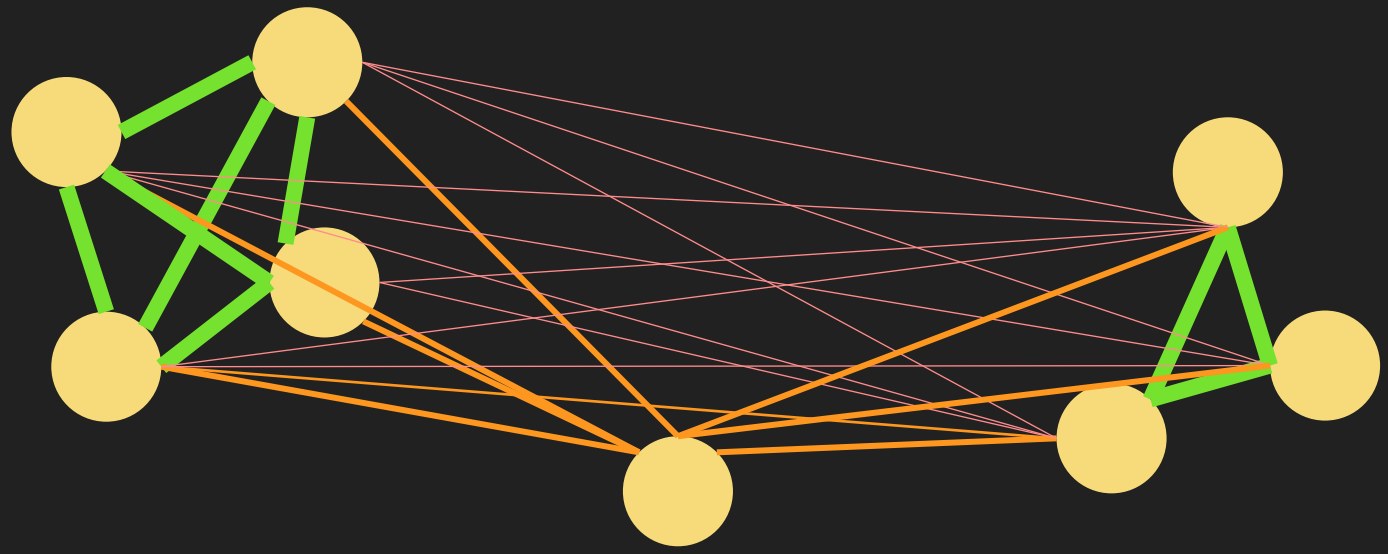
Vindt betekenisvolle categorieën voor teksten

1. Zoek alle paren



Vindt betekenisvolle categorieën voor teksten

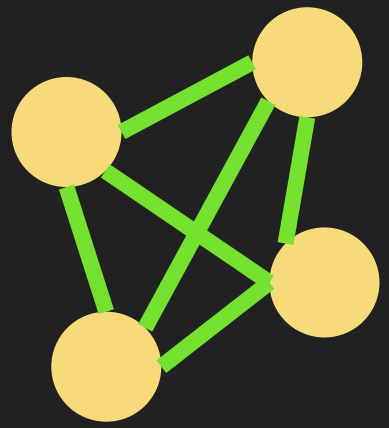
2. Bereken de similarity voor al die paren



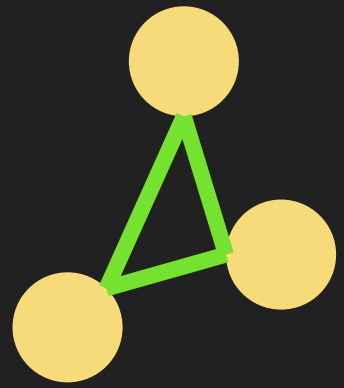
Vindt betekenisvolle categorieën voor teksten

3. Vind klieken die onderling hoge similarity hebben

Community 1



Community 2





Voorbeeld: Rik Torfs tweets communities

Community 128: Kort cynisme

- Spaar de cynicus, want cynisme is onuitgesproken verdriet.

- Er zijn veel cynici omdat achter cynisme verdriet schuilgaat.

- Cynisme is weemoedige humor van de verliezer.

- Cynisme is een vergissing. Maar cynisch willen zijn, en er het talent voor missen... Dat is pure tragiek.

- Word nooit cynisch, zelfs al kan u soms niet anders dan cynisme als wapen te hanteren.

- Cynisme vloeit voort uit het onvermogen tot verdriet.

- Een beetje cynisme hoeft geen ramp te zijn. Zolang je er niet trots op bent.

- Geloof is de beste remedie tegen cynisme. Daarom doen sommigen er cynisch over.

- De cynicus beweent zijn eigen mislukking en acht de samenleving daarvoor schuldig. Cynisme is een vorm van zelfoverschatting.

- Troost de cynici. Verzacht hun verdriet.

Community 128: Centrum van de wereld

- Als je niet het centrum van de wereld bent, ziet de wereld er plotseling beter uit.

- Als je er zelf niet het centrum van bent, wordt de wereld groter.

- Als je jezelf niet het centrum van de wereld voelt, ziet hij er opeens veel beter uit.

- Zolang we niet denken dat we zelf het centrum van de wereld zijn, heeft de wereld toekomst.

- Als je zelf niet het centrum van de wereld bent, wordt hij meteen veel groter.

- Wie zichzelf het centrum van de wereld acht, kent de wereld niet goed.

- Als je zelf niet langer het centrum van de wereld bent, ziet hij er plotseling veel mooier uit.

- Als je zelf niet het centrum van de wereld bent, wordt hij plotseling veel mooier.

- Als je niet het centrum van de wereld bent maar een doodgewone voorbijganger, kun je veel beter kijken.

- Als je zelf niet het centrum van de wereld bent, kun je hem beter zien met de nodige afstand.

Community 131: Nu zijn we slimmer

- Waarom denken wij slimmer te zijn dan de vorige generaties, terwijl zij exact hetzelfde dachten?

- We denken altijd dat we nu slimmer zijn dan vroeger. Daarin leren we niet bij.

- Onze generatie is slimmer dan de vorige. Dat vond de vorige ook. Niets verandert zolang alles beter wordt.

- Te gemakkelijk denken we intelligenter te zijn dan wie anders denken. Terwijl ze misschien juist anders denken omdat ze intelligenter zijn.

- Geef mensen altijd de indruk dat ze slimmer zijn dan jezelf. Dat hebben ze graag. En vaak is het nog waar ook.

- Wie zich slimmer voelt dan de rest van de mensheid heeft wellicht te weinig nagedacht.

- We denken graag dat we alles beter weten dan de vorige generatie. Tot we zelf de vorige generatie zijn.

- Waarom denken we toch dat we nu wijzer zijn dan vroeger? Terwijl het misschien juist andersom is.

- Waarom denken wij dat wij met de jaren wijzer worden, en nooit het tegendeel?

- Geloof is iets van vroeger toen we nog maar weinig wisten: Thomas van Aquino, Erasmus, Pascal. Gelukkig zijn we veel slimmer geworden.

Community 132: Wat is humor?

- Humor is wat overblijft als de lust tot lachen iedereen is vergaan.

- Humor is de pijnlijke stilte die valt na een poging tot grap.

- Humor is nooit kwetsend als er werkelijk niemand is die lacht.

- Humor is niet meer van deze tijd. Want het gevaar bestaat dat er gelachen wordt.

- Humor is vooral belangrijk als er niet veel te lachen valt.

- Humor is wat overblijft wanneer redelijkheid niet helpt.

- Humor is liefde. Als je iemand niet aardig vindt, veins dan ernst, prijs het verstand van je gesprekspartner en verlies zijn visitekaartje.

- Humor is wat overblijft als we het over alles oneens zijn en elkaar als persoon toch respecteren.

- Humor mag nog altijd. Zolang geen mens zich gekwetst voelt, de samenleving er rechtvaardiger door wordt en niemand lacht.

- Humor is populisme. Tenzij niemand ermee lacht, dan is het kwaliteit.

Package vol bestaande tools

sentence_transformers.util.community_detection(*embeddings: Tensor | ndarray, threshold: float = 0.75, min_community_size: int = 10, batch_size: int = 1024, show_progress_bar: bool = False*) → list[list[int]] [source]

Function for Fast Community Detection.

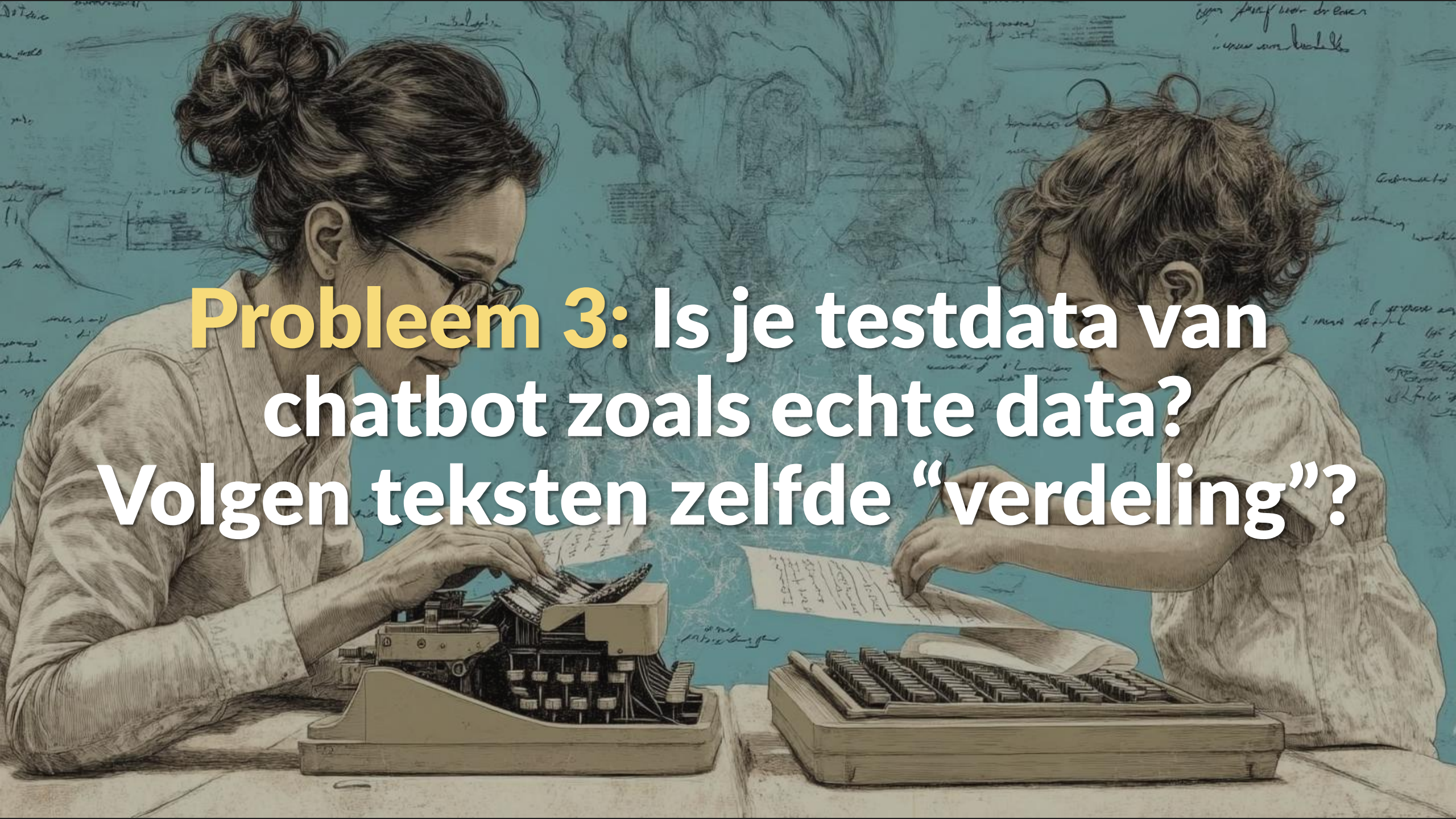
Finds in the embeddings all communities, i.e. embeddings that are close (closer than threshold). Returns only communities that are larger than min_community_size. The communities are returned in decreasing order. The first element in each list is the central point in the community.

Parameters:
- **embeddings** (*torch.Tensor or numpy.ndarray*) – The input embeddings.
- **threshold** (*float*) – The threshold for determining if two embeddings are close. Defaults to 0.75.
- **min_community_size** (*int*) – The minimum size of a community to be considered. Defaults to 10.
- **batch_size** (*int*) – The batch size for computing cosine similarity scores. Defaults to 1024.
- **show_progress_bar** (*bool*) – Whether to show a progress bar during computation. Defaults to False.

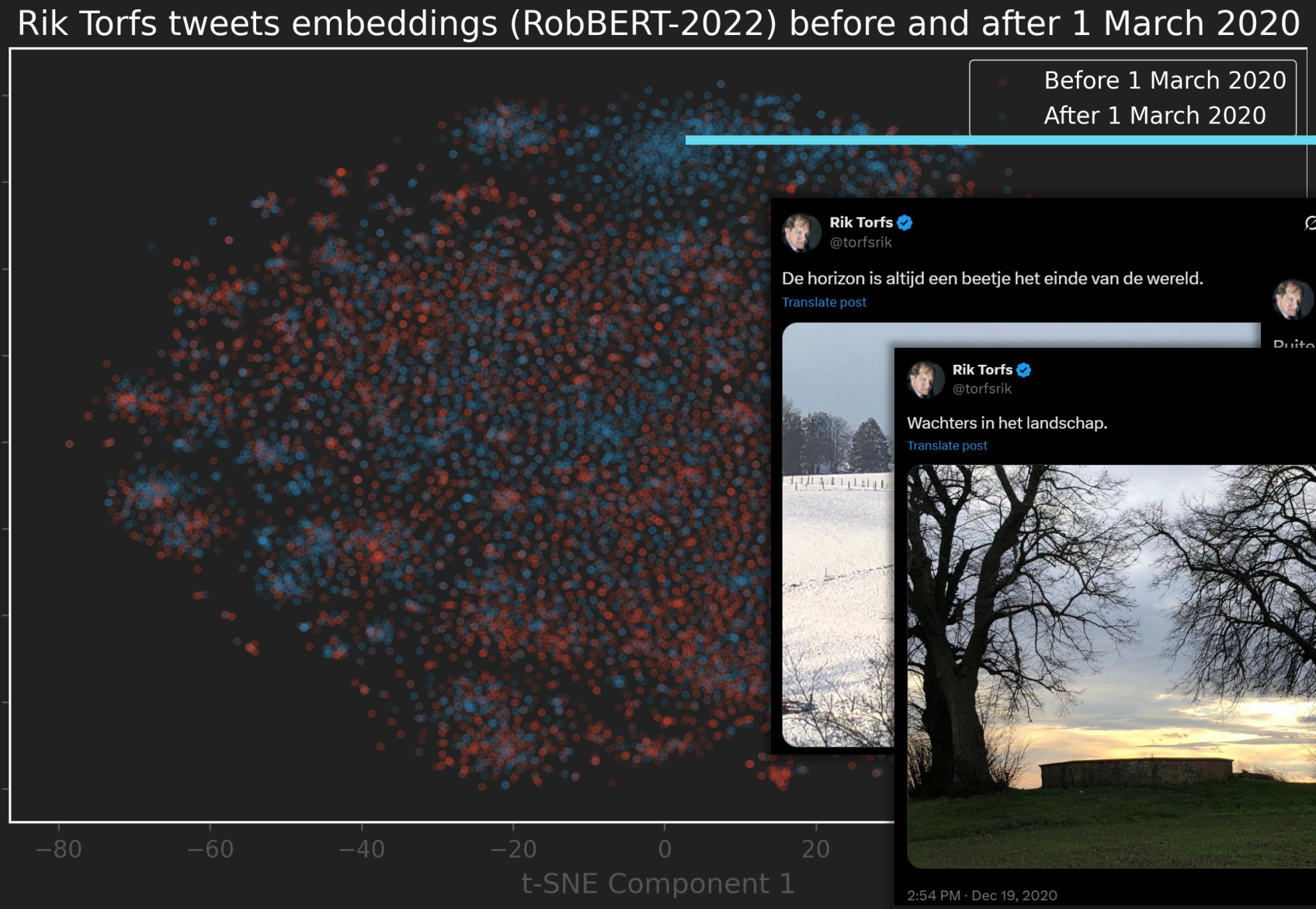
Returns: A list of communities, where each community is represented as a list of indices.

Return type: List[List[int]]

https://sbert.net/docs/package_reference/util.html#sentence_transformers.util.community_detection


Probleem 3: Is je testdata van chatbot zoals echte data? Volgen teksten zelfde “verdeling”?

Check of gelijkaardige “verdeling” van teksten

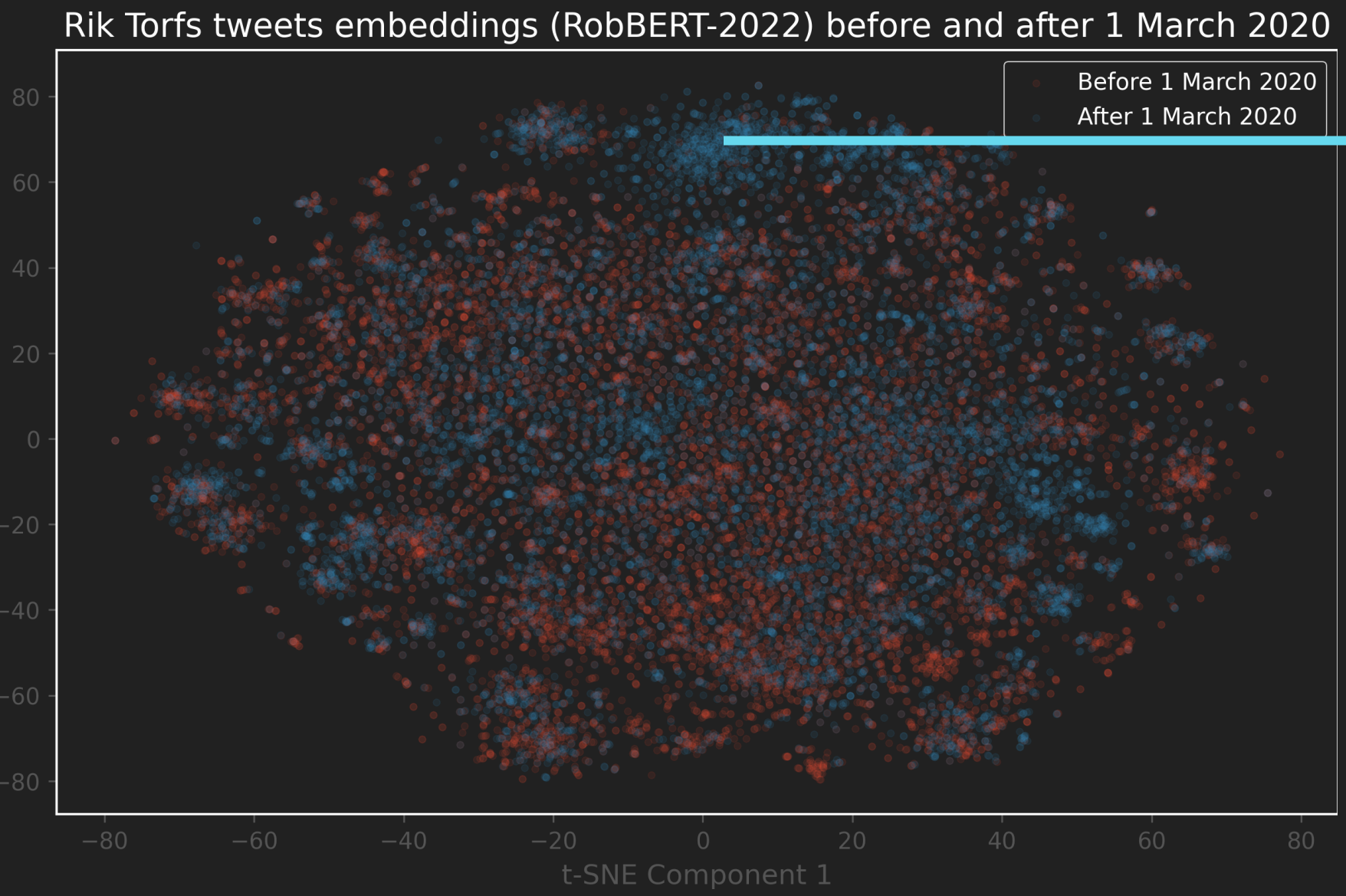


Kmeans clustering centroid afstand verschil

→ vindt zelf nieuwe categorie: Rik Torfs wandeltweets



Check of gelijkaardige “verdeling” van teksten



Kmeans clustering centroid afstand verschil

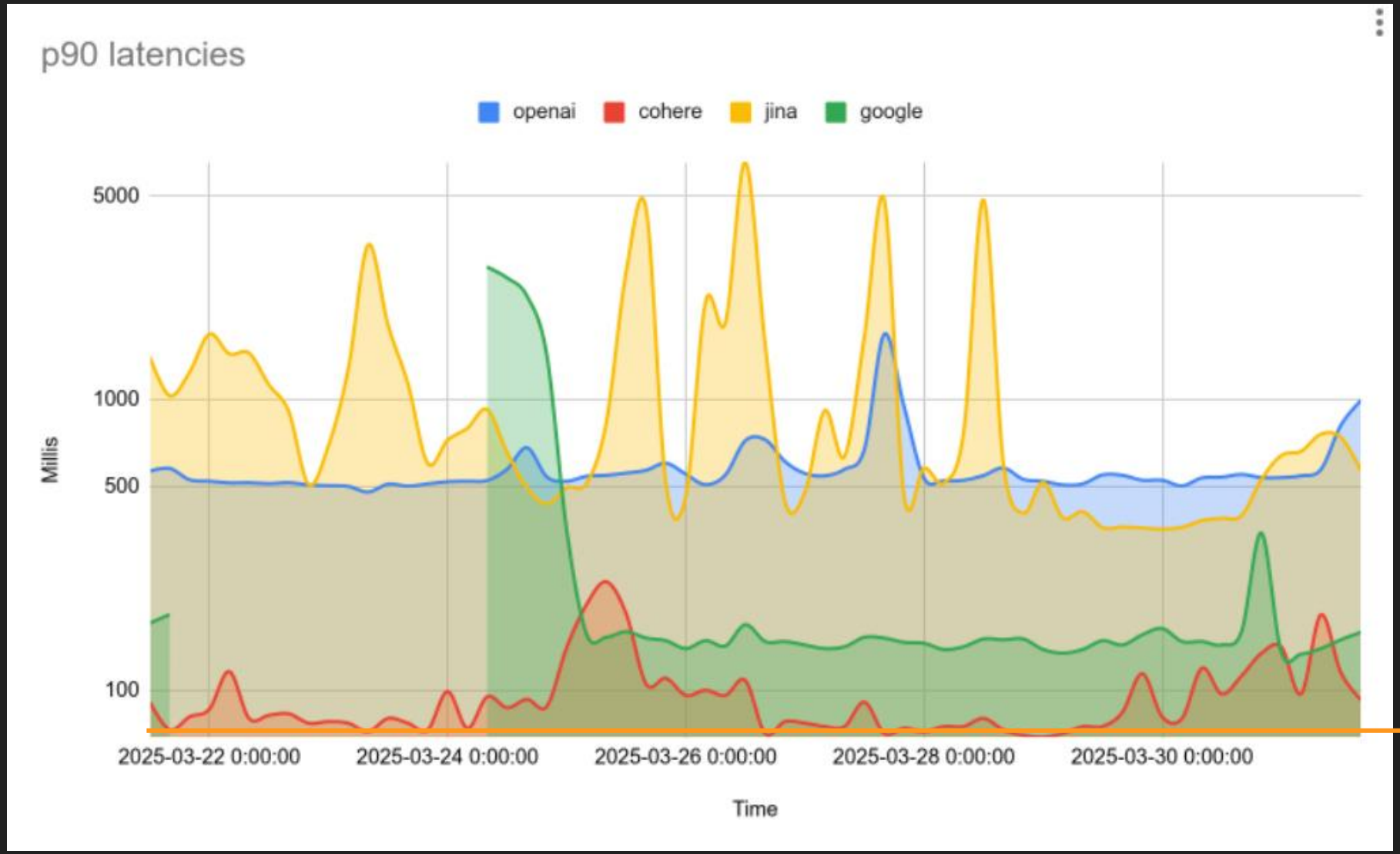
→ vindt zelf nieuwe categorie: Rik Torfs wandeltweets

Nuttig voor:

- Volgt mijn chatbot test data dezelfde distributie als de echte data van klanten?
- Identificeren wanneer gebruikers veel jailbreaks proberen
- Anomalieën vinden in de output van je chatbot

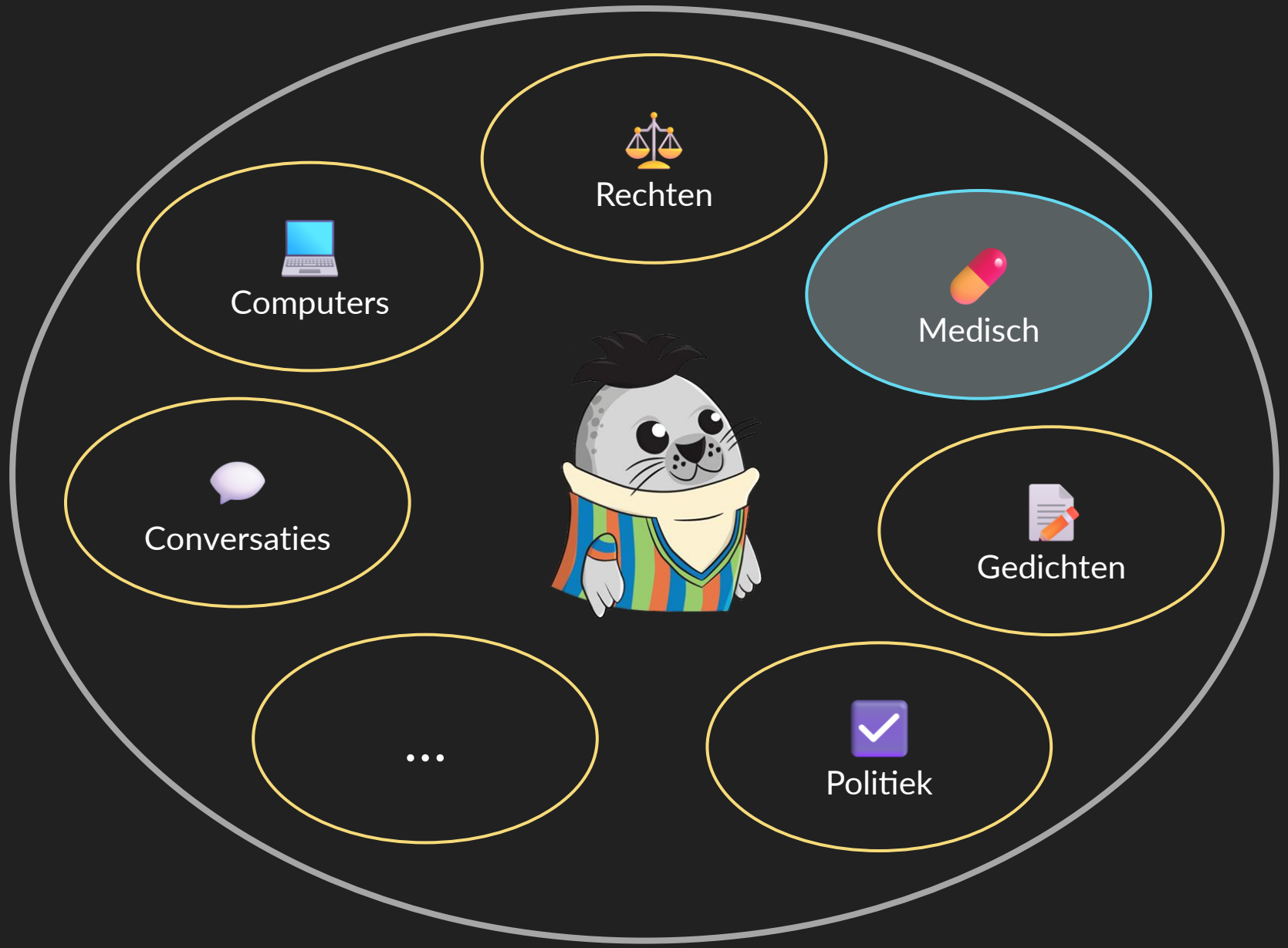

Waarom eigen & lokaal model?

Snelheid van APIs is unreliable: batching

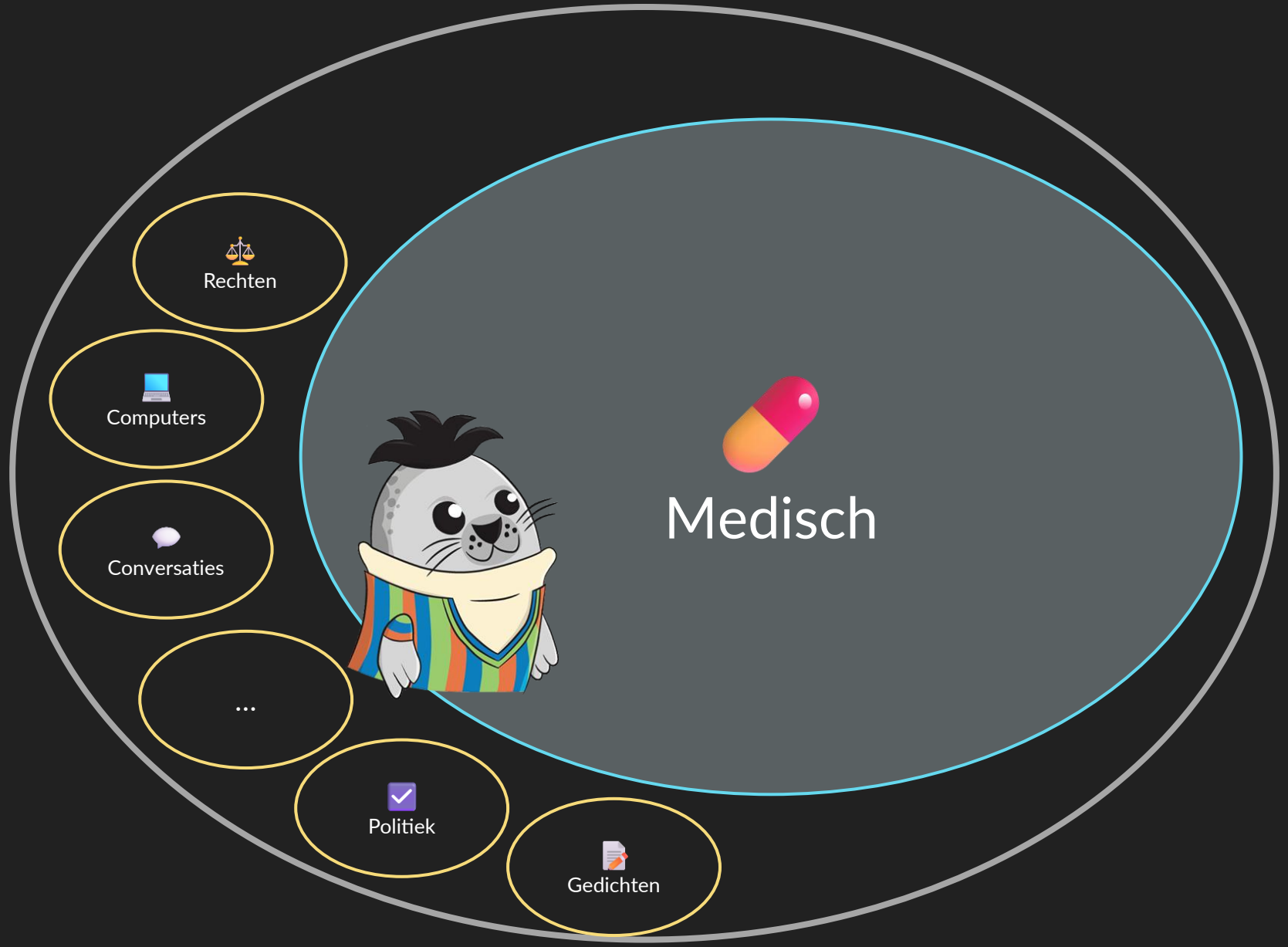


https://nixiesearch.substack.com/p/benchmarking-api-latency-of-embedding

Meer fijnkorrelig voor jouw domein



Finetune op medische data → Preciezere embeddings!



Sentence embeddings, je nieuwe beste vriend



(of: waarom eigen taalmodellen nog altijd nuttig zijn)

dr. ir. Thomas Winters

Postdoctoraal onderzoeker @ KU Leuven

thomaswinters.be

Slides: thomaswinters.be