**Behind chatGPT: Databronnen**

# Inhoud:

Welkom bij dit seminar over het proces van het verzamelen van gegevens voor ChatGPT. In dit seminar bespreken we de verschillende stappen die betrokken zijn bij het verzamelen van gegevens voor het trainen van het ChatGPT-taalmodel.

ChatGPT is een groot taalmodel ontwikkeld door OpenAI dat mensachtige antwoorden op tekstprompts kan genereren. Het model is getraind op een enorme hoeveelheid tekstgegevens van het internet, waaronder websites, boeken en artikelen. Het verzamelproces van gegevens voor ChatGPT omvat verschillende stappen, die we hieronder zullen bespreken.

## Stap 1: Identificatie van het corpus

De eerste stap in het verzamelproces is het identificeren van het corpus van teksten dat zal worden gebruikt om het model te trainen. Het corpus is een verzameling tekstdocumenten die worden gebruikt om ChatGPT te leren hoe mensachtige antwoorden te genereren. Het corpus moet divers en representatief zijn voor de verschillende soorten taal die ChatGPT wordt verwacht te begrijpen en te genereren.

## Stap 2: Het scrapen van de tekstgegevens

De volgende stap is het scrapen van de tekstgegevens van het internet. Dit omvat het gebruik van web scraping-tools om tekstgegevens van websites, boeken en andere online bronnen te extraheren. De gescrapte tekstgegevens moeten worden schoongemaakt en gefilterd om eventuele ongewenste tekst of tekens te verwijderen.

## Stap 3: Voorbewerking van de tekstgegevens

Zodra de tekstgegevens zijn gescrapet, moeten ze worden voorbewerkt. Dit omvat verschillende stappen, waaronder tokenisatie, zinssegmentatie en onderdelen van spraakmarkering. Tokenisatie houdt in dat de tekstgegevens in individuele woorden of tokens worden opgesplitst. Zinssegmentatie houdt in dat de tekstgegevens in zinnen worden verdeeld. Onderdelen van spraakmarkering houdt in dat het deel van de spraak voor elk woord in de tekstgegevens wordt geïdentificeerd.

## Stap 4: Het creëren van het taalmodel

Zodra de tekstgegevens zijn voorbewerkt, kunnen ze worden gebruikt om het taalmodel te creëren. Dit omvat het gebruik van machine learning-algoritmen, zoals deep neural networks, om het model op de tekstgegevens te trainen. Het taalmodel leert van de patronen en structuren in de tekstgegevens en gebruikt deze kennis om mensachtige antwoorden op tekstprompts te genereren.

## Stap 5: Het evalueren van het taalmodel

De laatste stap in het verzamelproces van gegevens is het evalueren van het taalmodel. Dit omvat het testen van het model op een set validatiegegevens om te bepalen hoe goed het presteert. De validatiegegevens moeten divers en representatief zijn voor de soorten taal die ChatGPT wordt verwacht te begrijpen en te genereren. De prestaties van het model kunnen worden gemeten met behulp van metrieken zoals perplexiteit, BLEU-score en menselijke evaluatie

# Titel

Hoe ChatGPT menselijke taal leert: De training uitgelegd.

# Taal

Nederlands

# omschrijving

# Deze workshop geeft inzicht in het proces van het verzamelen van gegevens voor het trainen van taalmodellen zoals ChatGPT, inclusief het identificeren en scrapen van tekstgegevens, voorbewerking, modellering en evaluatie. Deelnemers zullen een beter begrip krijgen van het trainingsproces van taalmodellen.

# Foto



<https://analyticsindiamag.com/behind-chatgpts-wisdom-300-bn-words-570-gb-data/>

# Resources

<https://chat.openai.com/chat>

<https://nl.wikipedia.org/wiki/ChatGPT>

<https://openai.com/blog/chatgpt>

"Language Models are Few-Shot Learners" door Tom B. Brown, et al. Deze paper beschrijft de training van het GPT-3-model van OpenAI en het gebruik van een enorme hoeveelheid aan data.

"Preparing a Corpus of Web Text for Clustering" door Mark Dredze, et al. Deze paper beschrijft de aanpak van OpenAI bij het verzamelen van een corpus van webtekst om taalmodellen zoals GPT te trainen.

De OpenAI-blog heeft verschillende artikelen gepubliceerd over de ontwikkeling en training van ChatGPT, waaronder "GPT-3: Language Models are Few-Shot Learners" en "The GPT-3 Architecture".

De GitHub-repository van OpenAI bevat ook informatie over de datasets die zijn gebruikt om ChatGPT te trainen.

# Beschrijving interactief

Voor de workshop hadden wij in gedachte om de volgende werkvormen te gebruiken?

Presentatie

Quiz

Mindmap

Discussie

# Marketing

vrienden

familie

collega’s

studie genoten

# Publiek

Homosapiens

# Extra spreker

Nee

# Verantwoordelijk omgaan met AI

Baises en data filtering