**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

Wat zijn de gevolgen van dat er gebruik gemaakt wordt van deze databronnen

# Werkvorm:

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

* Presentatie
* Quiz
* Mindmap
* Discussie

# Communicatie:

"Word een meester in taal: ontdek hoe ChatGPT getraind wordt om mensachtige antwoorden te genereren!"

"Ontdek de geheimen achter de training van ChatGPT, het taalmodel dat mensachtige antwoorden genereert"

Hoe ChatGPT leert praten: een kijkje in het verzamelen van gegevens voor een AI-taalmodel.

Hoe we ChatGPT hebben getraind: gegevens verzamelen en verwerken

"Hoe ChatGPT wordt getraind om mensachtige antwoorden te genereren: een kijkje in het

verzamelproces van gegevens".

"Van Tekst tot Taalmodel: Het Verzamelproces van Gegevens voor ChatGPT"

"Ontdek de geheime formule achter de meest geavanceerde AI-technologie ooit gecreëerd - ChatGPT: Het revolutionaire taalmodel dat de manier waarop we communiceren zal veranderen!"