MAKING A CHATBOT

Data Science Project

Tuule Tars
Siim Tänavots
Mark Tarnovski

Main Goal

1. Training a simple chatbot
2. Comparing the models fine-tuned on human- and AI-generated datasets

DATA USED

1. **Kaggle dataset** with simple Q&A pairs
2. Similar dataset generated via **ChatGPT**
3. **SQuAD** – Stanford Question Answering Dataset

HARDSHIPS

Working with the computational limitations on Google Colab; access is volatile
Generating (semi)large amounts of data via ChatGPT
Formatting said data
Fine-tuning with the Azure OpenAI (impossible)
Training an adequate bot **with the Kaggle dataset**

APPROACH

1. **Fine-tuning ChatGPT4o mini** with the Kaggle and generated datasets
2. **Training several basic chatbots/QA bots** from scratch using different approches: Bag-of-words, Word2Vec, Seq2Seq, ...
3. **Fine-tuning a DistilBERT model for QA** (question answering) on the SQuAD dataset
4. **Fine-tuning DialoGPT** using the Kaggle dataset

RESULTS AND CONCLUSIONS

The **basic chatbots** trained from scratch **showed insufficient results**, even when increasing the amount of training data. For example, the **Bag-of-words models only replied with the same answer**, no matter the question.
There were also **no noticeable differences** when comparing the results of training **on the Kaggle dataset and the ChatGPT generated dataset**. This suggests that a more complex structure is necessary for the model to achieve good results.

The **fine-tuning of DistilBERT for QA purposes worked remarkably well**, even for small amounts of training data. For example, fine-tuning on 5000 lines of data already yielded reasonably good results when compared to the non-fine-tuned version. These results are likely **thanks to the high quality of the SQuAD** dataset, which is human-generated, in terms of question formulation as well as answer selection. **Fine-tuning DialoGPT on the Kaggle dataset did not show much improvement** compared to the default version, however.

A conversation with a BoW chatbot

```
You: how are you doing
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 24ms/step
Bot: what do you mean?
You: what did you have for breakfast?
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 21ms/step
Bot: what do you mean?
```

A conversation with DialoGPT

```
>> User:hi, how are you doing?
DialoGPT: i am doing well.
>> User:what do you mean by that?
DialoGPT: !!!!!!!!!!!!!!!!!!!!!!!!!!!!
>> User:are you okay?
```