Exercise Sheet #9

Deadline: 02.02.2026, 12:00h

Problem 1 (*Tokenization in ChatGPT*) (10 points)

Most LLMs, such as ChatGPT, tokenize texts based on subwords. The tokenizer used in ChatGPT models is called *tiktoken*, an open-source project by OpenAI. You can find the repository here.

(a) Play around with tiktoken. Install toktoken and encode and decode a few strings.

(b) A web app is available for tiktoken which visualizes the inner workings of the tokenizer.

`https://tiktokenizer.vercel.app/`

Play around with the web app. Do you recognize patters where a single token might be a full word and where it might only be a subword? Experiment with different input texts. Does the token representation depend on capitalization? Write a small report (a couple sentences).

Problem 2 (*NotebookLM*) (10 points)

NotebookLM[1] is an AI research assistant by Google that is helpful when interacting with sources. By uploading PDFs, notes, or links, you create a personalized expert that summarizes content, answers questions, and generates insights based on your provided sources.

(a) Create a notebook in NotebookLM and add the lecture notes as sources.

(b) Test out all the ways NotebookLM can help you interact with these sources (e.g. create a mind map or flash cards).

Google Gemini introduces a feature called *Gems*. Gems are custom versions of Gemini that you can tailor with specific instructions and knowledge to act in a certain way and also allow including NotebookLM sources.

(c) Let's replace our professor with AI![2] Create a Gem that teaches you the lecture notes, including knowledge from the NotebookLM from part (a). Customize the Gem to be a useful teacher to you.

[1] You need a Google account to do this problem.
[2] Please still attend the lecture and tutorials :)