Course code: MLC_VJMGT

This course is designed for anyone who is fascinated by the capabilities of large language models (LLMs) and generative artificial intelligence, and wants to delve into the subject beyond the basic user level. We will learn about transformers, the basic building blocks of modern language models, introduce the most well-known architectures, and show how large language models can be used for a variety of applications. No paid third-party account is required for the hands-on exercises. We'll use open-source models that, when used properly, are just as good as the biggest commercial models.

Who is the course for

This course is designed for anyone who is fascinated by the capabilities of large language models (LLMs) and generative artificial intelligence, and wants to delve into the subject beyond the basic user level.

What we teach you

We will learn about transformers, the basic building blocks of modern language models, introduce the most well-known architectures, and show how large language models can be used for a variety of applications.

Required skills

- Basic knowledge of programing in Python
- Basics of machine learning on the level of our course Introduction to machine Learning

Course outline

- Overview of generative AI (text, images)
- Evolution of language modeling
- Transformers
- Types of transformer-based LLMs (encoder, decoder, encoder-decoder)
- Reinforcement learning with human feedback
- Overview of the most popular transformer-based LLMs (BERT, GPT, LLAMA, T5, BART…)
- Transformer-based classification example with HuggingFace in Google Colab
- Prompt engineering: in-context learning, zero shot, one shot and few shot prompting, configuration parameters of the generative process
- Practical examples of in-context learning with HuggingFace in Google Colab
- Full fine-tuning of large language models, parameter-efficient fine-tuning (LoRA)
- Text generative AI evaluation (ROUGE, BLEU)
- Practical example of parameter efficient fine-tuning with HuggingFace in Google Colab
- Retrieval Augmented Generation (RAG)