

Exercise sheet 12

Natural Language Processing

Hand-in (voluntarily): 01/26/2024 until 10:00 a.m. via Moodle

In moodle, you will find the file `Language_Detection.csv`. In it, you will find a number of texts and the respective language they stem from. Your task is to determine the language of the texts as efficiently as possible in an unsupervised analysis.

To do this, please visit Huggingface, a website on which open source transformer models are available for the public to download. Look for appropriate models that are able to solve the given task without additional training and compare their performances. Some models on Huggingface also provide the option for an inference API directly on the website, enabling you to test out even large language models that usually would not run on local hardware.

Please keep in mind, that you are not meant to train the model yourselves, as transformer models take a lot of time to train on weak GPUs or even CPUs (like most Laptops have).

Recommended packages & functions

R: `huggingfaceR`

Python: `transformers`