Topic 10: Named Entity Recognition for German official documents

Instructor Gabor Recski

Overview Named Entity Recognition (NER) is a key bottleneck in many NLP applications, including relation extraction, question answering, or anonymization. Your goal in this project is to select 1-2 corpora of German official documents and define 2-3 entity types (e.g. person, organization, address, phone number) for which you develop NER solutions

Resources

• FinCorpus-DE10k: A Corpus for the German Financial Domain (Hamotskyi, Kozaeva, and Hänig, 2024)

• German Parliamentary Corpus (GerParCor) (Abrami, Bagci, and Mehler, 2024)

• Open Legal Data dataset of 100,000 German court decisions