Text Technology Project

INITIALIZING PHASE

Participants:

Name: Sushma Kumari
Matrikelnummer: 3587711
Studium: M.Sc. Computational Linguistics

Name: Rishi Anil Rajani
Matrikelnummer: 3746976
Studium: M.Sc. Computer Science

Left-Brained Written Word

LLMs like GPT-4 are becoming increasingly creative in generating unseen texts, given the vast libraries on which they are trained. On the other hand, human-made pieces of texts like passages, poetry and theatre have become dull and uninspired; they no longer possess the charm or panache that could be found in the Shakespearean era. The proposal is to create a corpus of tokens, words, constructed sentences and complete texts written in Early Modern English, found in the Helsinki Corpus of English Texts, that can act as the groundwork for building a library which shall be used by an LLM to generate creative pieces of art. Following would be the steps to build this corpus:

- Collect:
  Extract Early Modern English texts from the XML-based Helsinki Corpus files
- Prepare:
  Perform cleaning of the texts (if needed) and store the texts as tokens, words, and sentences in PostgreSQL
- Access:
  Query the database using XQuery