Mixture of Experts for Topic Annotation

This repository implements a full pipeline for extracting themes from the Bullinger correspondance using GPT-4o and DeepSeek, and combining their outputs through a Mixture of Experts (MoE) strategy. It supports converting raw XML files, running model annotations, converting outputs, and performing comparative analysis.

Workflow Overview

1. `xml2txt.py` converts XML to plain text

2. Set your API keys

- `export OPENAI_API_KEY="XXX-your-openai-key"`
- `export DEEPSEEK_API_KEY="XXX-your-deepseek-key"`

3. `query_LLMs.py` can be used to:

```
- annotate with DeepSeek
- test annotation on a small sample with both models, for example on
`test_files/eng/`
```

4. Annotate with GPT-4 (Batch Processing)

- `batches_create.py`
- `batches_process.py` sends batches to OpenAI and downloads processed results
- `batches_concat.py` concatenates the output of all batches into one file

5. Convert Outputs to CSV

- `convert_gpt4_batch_annos_to_csv.py`
- `convert_deepseek_annos_to_csv.py`

6. `MixtureOfExperts.py` performs Mixture of Experts (MoE)

Combines annotations from both models using topic-specific logic:

Topic	Strategy	Threshold(s)
Requests and Petitions & Military and Political Affairs	Intersection	Both models ≥ 90
All other topics	Soft Union	One ≥ 60, other ≥ 20

Outputs:

- `topic_mapping.csv`: English ↔ German ↔ ID topic mapping

- A csv file with 2 columns: File ID (int) and Topics (str), which is a list of topic IDs mapped in the topic_mapping.csv
- (optional) `gestapelte_zuweisungen_modelle.png`: Stacked bar plot of topic assignments by both models

7. (Optional) Merge new and old annotations

If new annotations are produced, merge the two csv files together to create the final file

- `combine_merged_topics_ids_files.py` file1 file2
- output: `merged_topics_ids.csv`