Pop Culture & Meme Analysis

Created by Marcell Bérces

Original Project Descrption

Scrape memes, local popcultural content from the internet, gather background information and create text only instructions on the events, ideas, entities from the memes/popcultural references.

Data Sources

Due to practical constraints with scraping, the project uses the MemeCap dataset, a Reddit-based collection of images, image captions, meme captions as the primary source of meme text. The images are put through OCR to extract any available text and then discarded. To provide cultural and factual background, relevant Wikipedia articles are scraped and used as external knowledge sources.

Method

The proposed solution follows a multi-stage NLP pipeline:

- Data selection: Meme captions are sampled from the MemeCap dataset (title, image caption, meme caption). Relevant Wikipedia pages are retrieved as background context.
- OCR of images: Meme images are run through an OCR to extract any text which are later filtered.
- Reference grounding: Relevant information is extracted from the MemeCap dataset data and image OCR outputs.
- Chunking: Wikipedia documents are chunked into smaller text segments for efficient processing.
- Indexing layer: Chunks are indexed and combined into a database.
- LLM explanation generation: The university LLM endpoint is used to generate explanations for the memes. If the response is unsatisfactory, a fixing step is implemented.
- Dataset assembly: Final LLM outputs are converted into instruction–completion examples.

Output

The final result is a 3000-entry supervised fine-tuning (SFT) dataset formatted according to Hugging Face standards. Each entry consists of an instruction prompt and an explanatory completion describing the cultural references present in meme-related text.

The dataset is intended for research and educational use.

Limitations

The dataset is based on the MemeCap dataset, therefore the memes are only in English and the diversity is not high.

Conclusion

This project demonstrates how meme text can be transformed into structured cultural knowledge and instruction-style training data using LLMs. The resulting dataset provides a foundation for further research into improving LLM understanding of pop culture and internet-native communication.