LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN
CENTRUM FÜR INFORMATIONS- UND SPRACHVERARBEITUNG
STUDIENGANG COMPUTERLINGUISTIK

Thesis proposal

Topic: **Exploring Archival of Queer Language across Time and Communities**

Supervisor: Leonor Veloso

Examiner: Prof. Hinrich Schütze

Level: BSc/MSc

Summary: Online dictionaries, ontologies, and linked vocabularies of queer terminology are widely used in the fields of digital archival and NLP. These sources often offer metadata regarding (i) original/first sources that document and define a term, and (ii) subcommunities of the queer/LGBTQ+ community that the term is most closely associated with. The objective of this project is to analyze how terminology related to queer subcommunities evolved over time – e.g., coinage of new non-binary identity labels in the 2010s (Filardo-Llamas and Roldán-García 2025). This can be achieved ideally through the creation of a structured dataset. The exact structure and format of the final data artifact can be modified, but should contain the following:

- A normalized list of vocabulary `terms`, each associated with a normalized `community name` and one or more `documentation entries`.
- Each `documentation entry` should have a `date`, `definition`, `source name`, and a `url` associated with the source. If present in the original source metadata, each `documentation entry` should also have a `community` that the term is associated with.

A potential work pipeline is:

- Identifying queer terminology sources, especially those that note the origin of terms and/or subcommunities where that term originates from (some leads are GSSO[1], Chew Glossary[2], Green's Dictionary of Slang[3], Homosaurus[4]). This is not trivial, since common terms can appear in multiple sources with different community-based categorizations, or be classified as variants/synonyms of other terms. A large component of this work is normalization of entities;
- Scraping desired sources and creating a data artifact as previously described;
- Analyzing and visualizing the evolving documentation of subcommunity-specific terminology across time.

The expected workload of this project should be suited for a BSc, but can be extended to a MSc thesis by, for example, using the final data artifact to survey representation of historical and subcommunity-specific queer terminology within NLP fairness literature. This project is largely inspired by Wang and Adamidou 2024.

Requirements:

- Enthusiasm!
- Knowledge of Python and data handling-related libraries.

References:

[1] `https://gsso.research.cchmc.org/#!/home`
[2] `https://itg.nls.uk/wiki/`
[3] `https://greensdictofslang.com/`
[4] `https://homosaurus.org/`

- Shuai Wang and Maria Adamidou (2024). "Examining LGBTQ+-Related Concepts in the Semantic Web: Link Discovery, Concept Drift, Ambiguity, and Multilingual Information Reuse". In: *International Conference on Knowledge Engineering and Knowledge Management*. Springer, pp. 1–17