



<http://www.diva-portal.org>

Postprint

This is the accepted version of a paper presented at *Huminfra Conference (HiC 2024, 10 jan 2024 - 11 jan 2024, Gothenburg)*.

Citation for the original published paper:

Alfter, D., Falk, O., Ihrmark, D., Golub, K., Humlesjö, S. (2024)

Automatic subject indexing of Swedish LGBTQ+ fiction

In: *Presented at Huminfra Conference (HiC), Gothenburg, 10 jan 2024 - 11 jan 2024*

N.B. When citing this work, cite the original published paper.

Permanent link to this version:

<http://urn.kb.se/resolve?urn=urn:nbn:se:lnu:diva-128148>

Automatic subject indexing of Swedish LGBTQ+ fiction

David Alfter (GU), Olof Falk (UB), Daniel Ocic Ihrmark (LNU),
Koraljka Golub (LNU) and Siska Humlesjö (GU)

Huminfra Conference (HiC 2024), 10 jan 2024 - 11 jan 2024, Gothenburg





Background

- With expanding and unstructured text collections, automatic and computationally assisted indexing can potentially alleviate organizing workloads (Golub, 2006; Golub et al., 2016; Moulaison-Sandy et al., 2021; Short, 2019).
- A few studies on automatic and computationally assisted fiction indexing exist, such as Short (2019) and Moulaison-Sandy et al. (2021), but overall, the subject has been explored to a relatively small extent.
- Our goal was to begin investigations on whether and how computational methods can effectively be used to support indexing of LGBTQI+ themes in fiction.
- To do this, we departed from the *Queerlit* database and its *QLIT* subject headings (see Golub et al., 2022, 2023). As a secondary objective, we also considered more general themes from SAO* in the analysis.

*Svenska Ämnesord, see: <https://id.kb.se/>.



An early experiment: LDA topic modeling as a tool for computer-assisted fiction indexing*

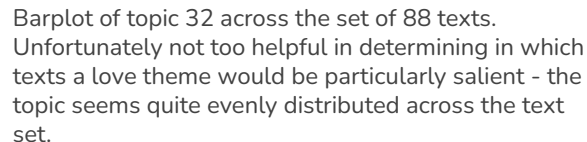
- Data: 88 full texts indexed in Queerlit, and collected from Litteraturbanken (<https://litteraturbanken.se/>). The dataset included novels, poetry, short stories, and essays (proofread e-texts were prioritized).
- Text pre-processing: Metadata removal, stop word removal (according to a standard Swedish stop word list), and slicing of full texts into 500-word chunks (upper limit). Topical probabilities was then calculated for each of the full texts.
- The number of subject headings applied in Queerlit amounted to 97. Thus, as an experiment, the LDA algorithm (through MALLET) was instructed to organize the words in the text set into 97 computationally derived topics.
- These 97 topics were then qualitatively assessed to determine whether any of them corresponded to the applied subject headings in Queerlit (and SAO).

*See Blei (2012) for an overview of the LDA algorithm and its implications, and see Golub (2006) for details on computer-assisted fiction indexing. These experiments closely followed methods explained by Jockers (2013, 2014) and made use of the R wrapper for MALLET (Magnusson & Mimno, 2022). The stop word list was provided by Dahlgren (2022).

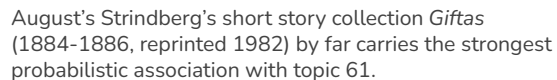


Tentative observations from a 97-topic model

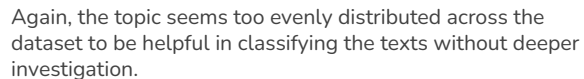
- 63 out of 97 topics were deemed too vague, possibly due to too much variation present in the dataset (such as different genres, different stages of OCR proofreading, older vs. modern Swedish, and considerable amounts of proper nouns). The method thus needs refinement.
- 34 topics were deemed sufficiently interpretable and coherent, although the algorithm (somewhat unsurprisingly) did not appear to pick up on the subtextual and peripheral themes manually indexed by Queerlit very well.
- Through a qualitative comparison between topics and QLIT subject headings, three topics - 32 (Love), 61 (Marriage), and 70 (Death and dying) were found to resemble QLIT headings (although not clearly connected to LGBTQI+ perspectives), and were tentatively labelled accordingly.



Looking at the wordcloud, the topic also seems quite general, and arguably not clearly indicative of LGBTQI love specifically.



Giftas is indexed under “Marriage” (among other terms) in both SAO and Queerlit. However, the topic does not seem to carry an immediate LGBTQI+ connection.



The 3rd most prominent text for this topic - Edith Södergran's *Landet som icke är* (1925) - is labelled with the SAO term Döden (English: Death); however, not the QLIT term Döden och döende (HBTQI).

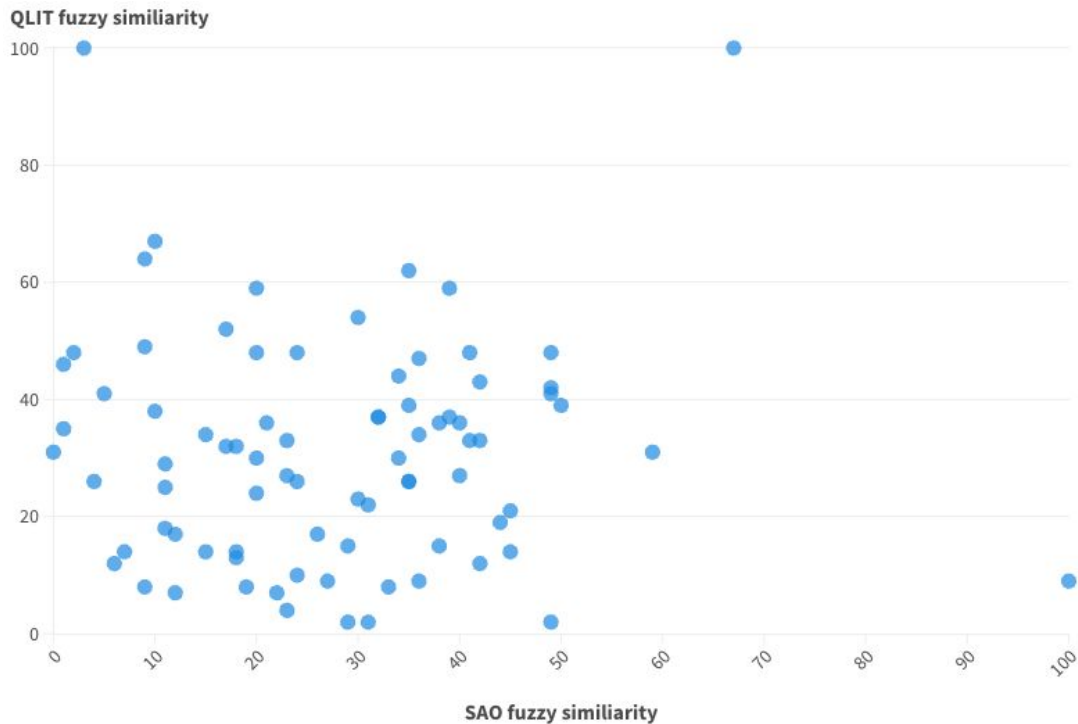


Zero-shot classification for automated fiction indexing

- Data: 82 short descriptions by indexers
- Methodology:
 - Zero-shot classification pipeline with short text and SAO/QLIT labels
 - Retain only labels with a probability of over 0.9
- Evaluation: Manual inspection of labels in subset of data
 - 20-40% accuracy
- Evaluation: String matching
 - “Fuzzy” string matching
 - Token Sort Ratio



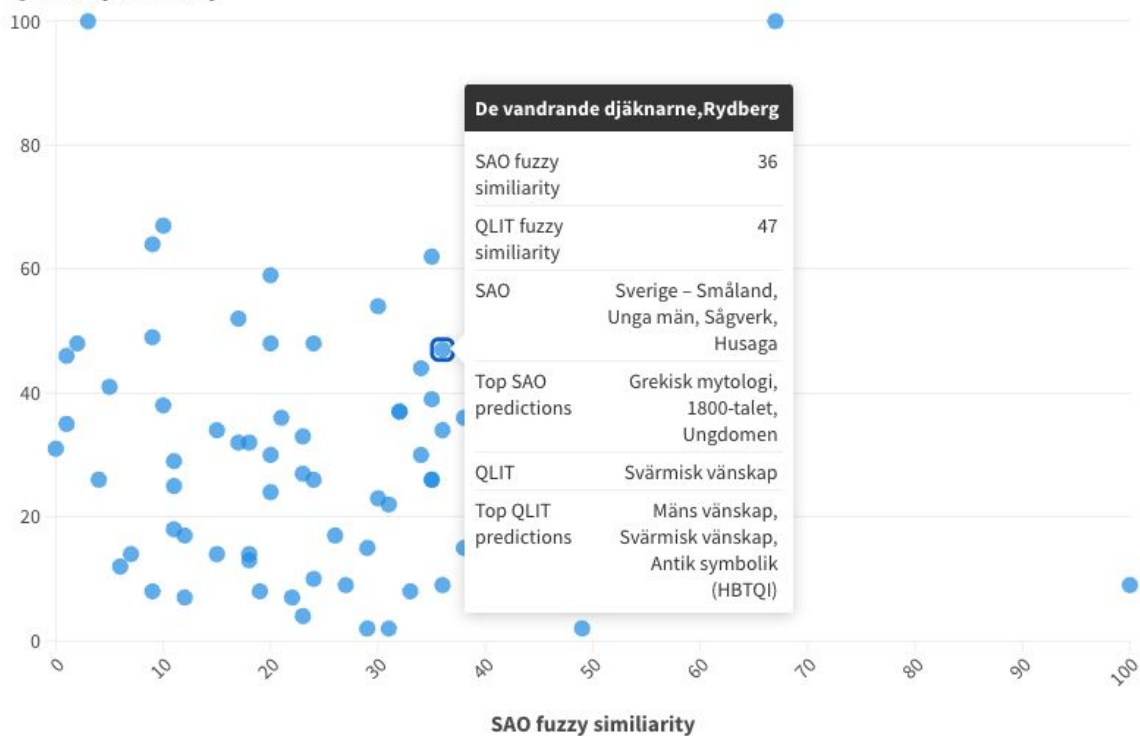
QLIT and SAO Labeling using Zero-Shot: Results of Fuzzy String Matching



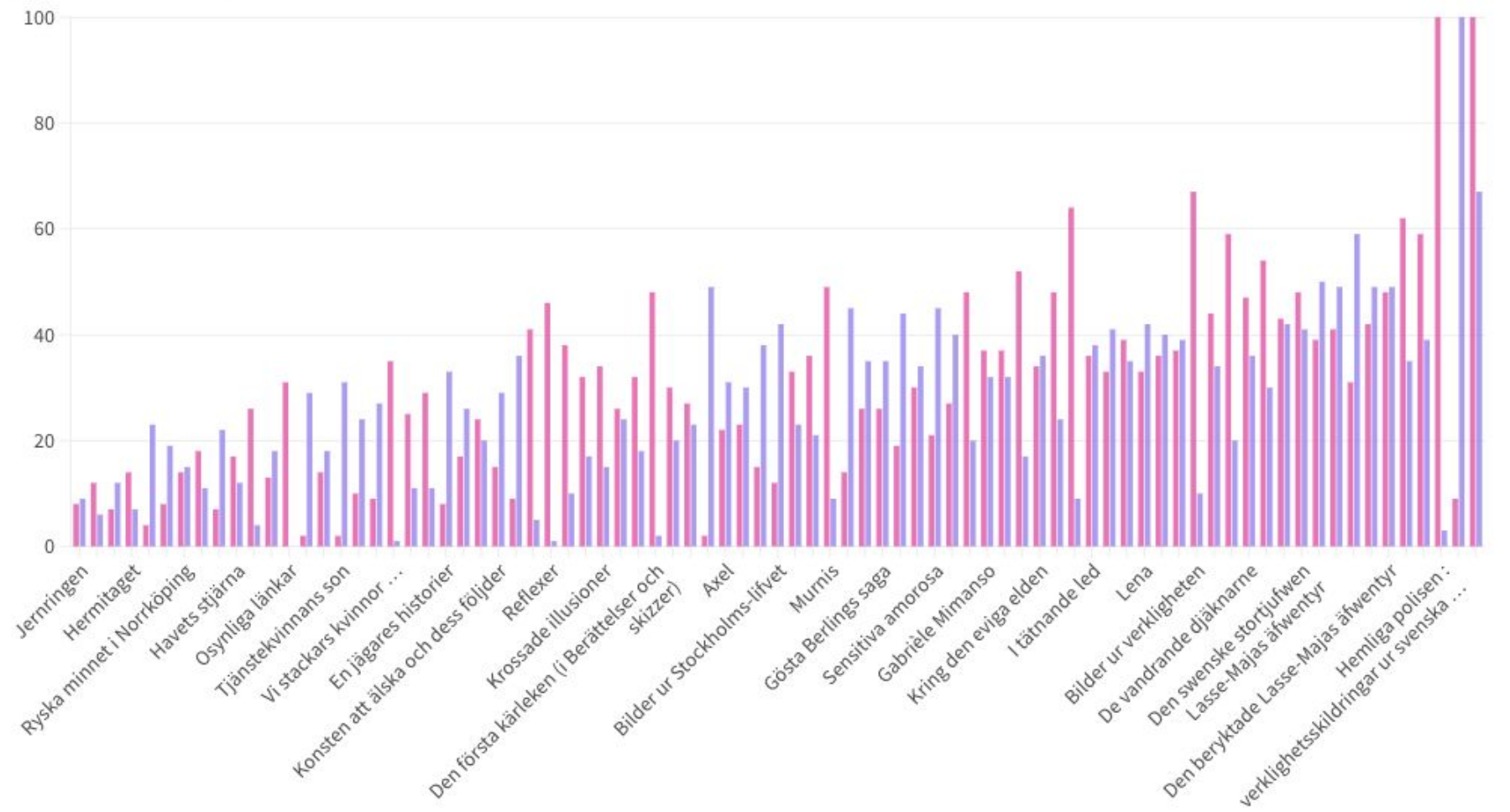


In Practice

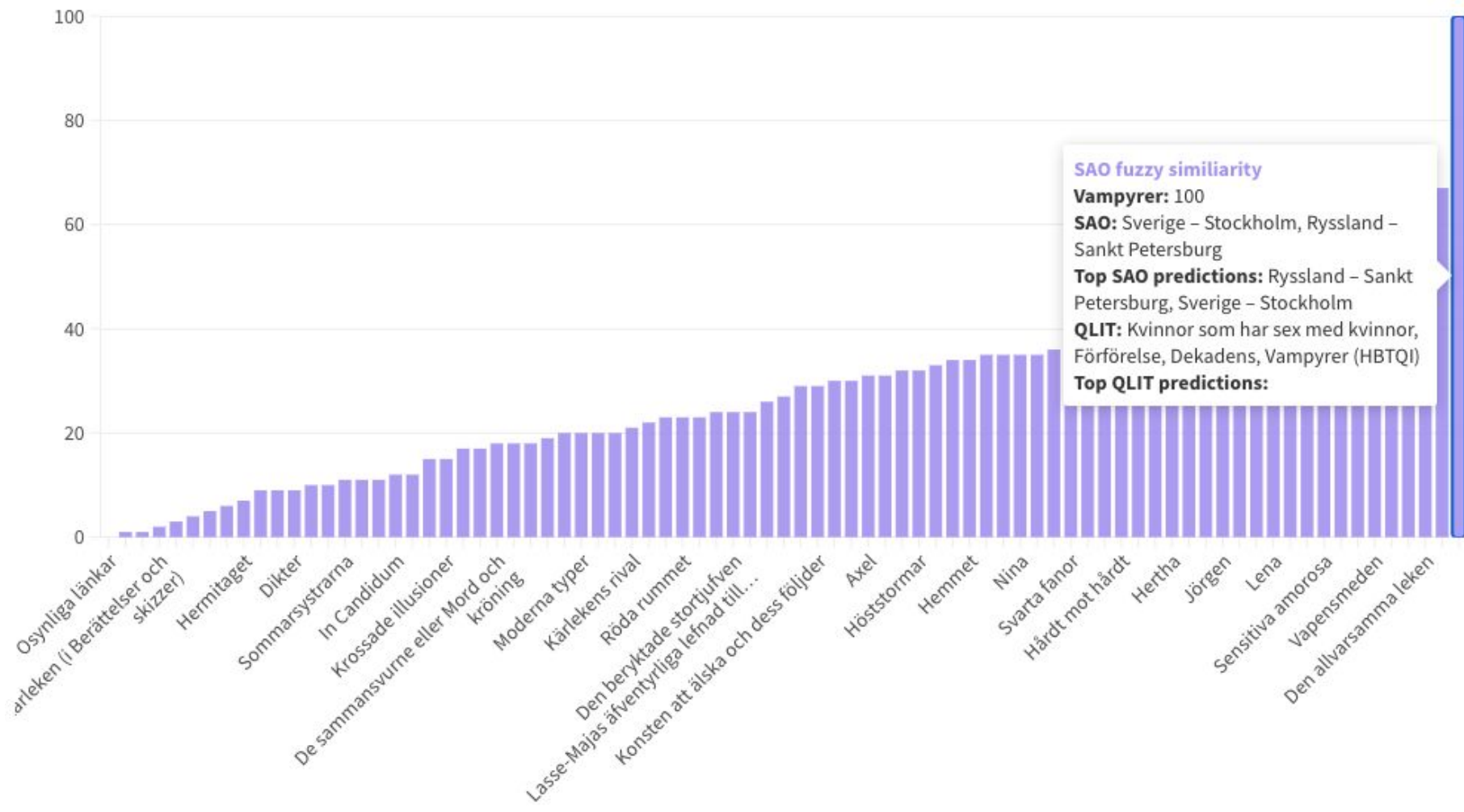
QLIT fuzzy similarity



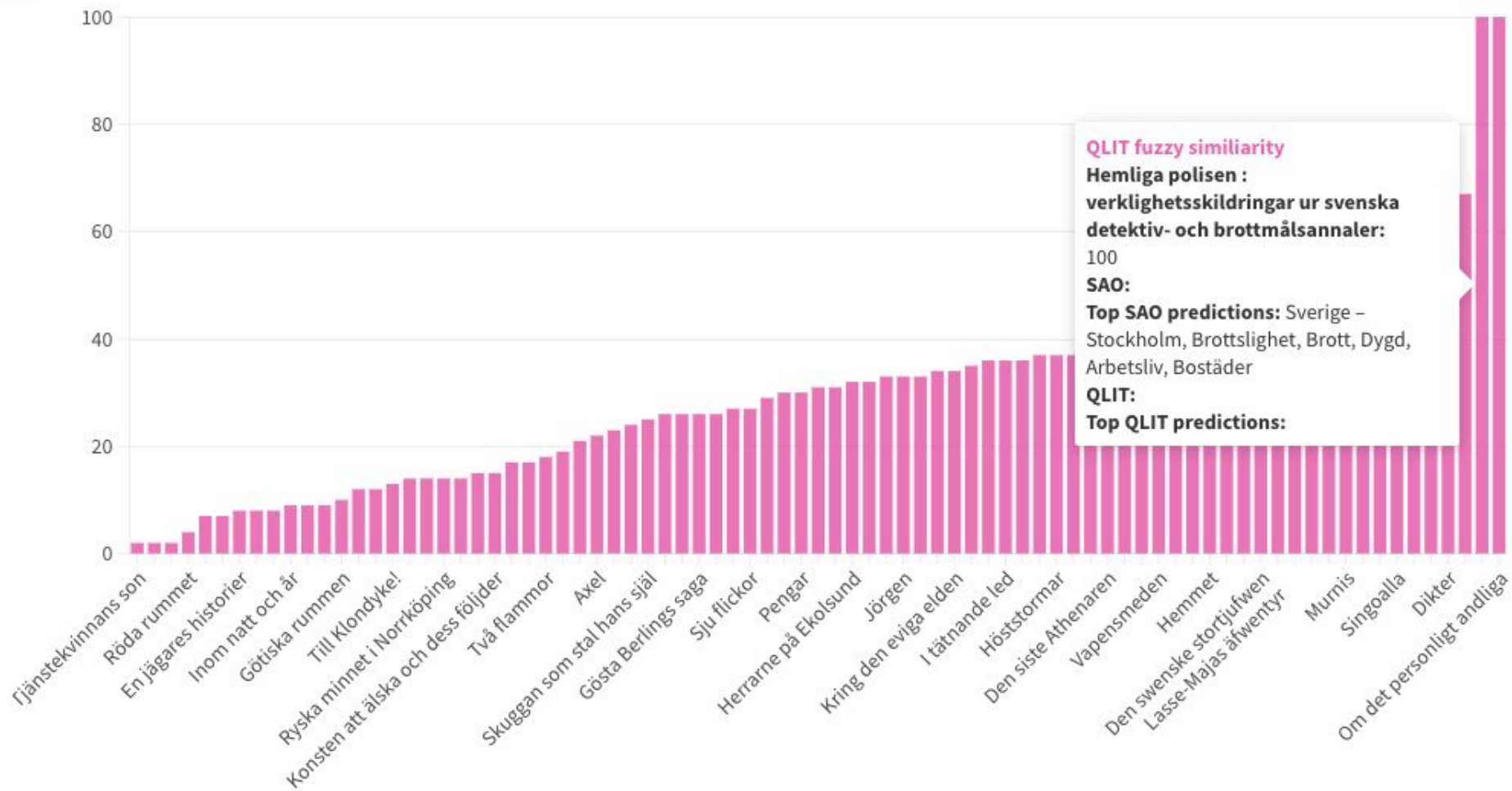
QLIT fuzzy similarity SAO fuzzy similarity



SAO fuzzy similiarity



QLIT fuzzy similiarity





Tentative conclusions & further research suggestions

- LDA topic modeling seems an interesting way for initial exploration of fiction collections for topical content. However, it does not seem to present a ready way of dealing with subject classification for unstructured fiction collections.
- Overall, LDA seems more promising in picking up more general themes, suggesting that manual indexing may still be required to pick up peripheral or subtextual themes. (The method could likely be improved further).
- Zero-shot classification initially seems promising, but requires manual intervention for practical uses.
- Future work: zero-shot classification on full text.



References & further reading

Blei, D. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4), 77-84.

Dahlgren, P. M. (2018). Svensk text. *Svensk nationell datatjänst*.
<https://snd.gu.se/sv/catalogue/study/ext0278>

Golub, K. (2006). Automated subject classification of textual Web pages, based on a controlled vocabulary: challenges and recommendations. *New Review of Hypermedia and Multimedia*, 12(1), 11–27. <https://doi.org/10.1080/13614560600774313>

Golub, K., Dagobert, S., Buchanan, G., Tudhope, D., Lykke, M., & Hiom, D. (2016). A framework for evaluating automatic indexing or classification in the context of retrieval. *Journal of the Association for Information Science and Technology*, 67(1), 3–16.
<https://doi.org/10.1002/asi.23600>

Golub, K., Bergenmar, J., & Humelsjö, S. (2022). Searching for Swedish LGBTQI fiction: challenges and solutions. *Journal of Documentation*, 78(7), 464–484.
<https://doi.org/10.1108/JD-06-2022-0138>

Golub, K., Bergenmar, J., & Humelsjö, S. (2023). Searching for Swedish LGBTQI fiction: the librarians' perspective. *Journal of Documentation*, 79(7), 261–279.
<https://doi.org/10.1108/JD-05-2023-0080>

Jockers, M. L. (2013). *Macroanalysis: Digital Methods and Literary History* (1st ed.). University of Illinois Press.

Jockers, M. L. (2014). *Text Analysis with R for Students of Literature* (2014th ed.). Springer Nature.
<https://doi.org/10.1007/978-3-319-03164-4>

Magnusson, M., & Mimno, D. (2022). *Mallet: An R Wrapper for the Java Mallet Topic Modeling Toolkit*. Retrieved 2024-01-03, from: <https://cran.r-project.org/web/packages/mallet/index.html>

Moulaison-Sandy, H., Adkins, D., Bossaller, J., & Cho, H. (2021). An Automated Approach to Describing Fiction: A Methodology to Use Book Reviews to Identify Affect. *Cataloging & Classification Quarterly*, 59(8), 794–814. <https://doi.org/10.1080/01639374.2021.1992694>

Short, M. (2019). Text Mining and Subject Analysis for Fiction; or, Using Machine Learning and Information Extraction to Assign Subject Headings to Dime Novels. *Cataloging & Classification Quarterly*, 57(5), 315–336. <https://doi.org/10.1080/01639374.2019.1653413>

Thanks for your time and attention!