Few-Shot Prompting for Subject Indexing of German Medical Book Titles

Lisa Kluge
Deutsche Nationalbibliothek
Frankfurt am Main, Germany
l.kluge@dnb.de

Maximilian Kähler
Deutsche Nationalbibliothek
Leipzig, Germany
m.kaehler@dnb.de

Abstract

With the rise of large language models (LLMs), many tasks of natural language processing have reached unprecedented performance levels. One task LLMs have not yet been evaluated on is subject indexing with a large controlled target vocabulary. In this work, an LLM is applied to the task of subject indexing a dataset of German medical book titles, compiled at the German National Library. The results are compared to two common baseline methods already in productive use at this institution. One critical parameter in a few-shot prompting approach is the composition of examples given to the LLM for instruction. In order to select examples, two similarity measures between book title and gold-standard labels are applied. We hypothesise that these notions of similarity can serve as a measure of task difficulty. Our findings indicate that the LLM does not outperform the baselines. Still, (off-the-shelf) LLMs can be a valuable addition in an ensemble of methods for subject indexing as they do not depend on training data.

1 Introduction

At the German National Library (Deutsche Nationalbibliothek, DNB), incoming publications undergo subject indexing not only in an intellectual fashion. Digital publications can be indexed in an automated way. In both cases, each medium is annotated with fitting entities from the Integrated Authority File[1] (GND) in order to to make them accessible to users. In the present study, a large language model (LLM) is compared to two baseline approaches for automated subject indexing in productive use at the DNB. These are available via the Annif framework (Suominen, 2019) developed by the Finnish National Library.

The focus of this work is on improving the selection and composition of examples used in an LLM few-shot prompting approach to make further progress towards solving the GND-annotation problem.

Our work makes the following contributions:

- To our knowledge, this is the first application of LLMs to subject indexing of German scientific publications.
- We provide a comparison between an LLM-based approach and widespread methods for subject indexing at libraries.
- We investigate the influence of purposeful prompt variation on the model's performance.
- Two measures of similarity, one accounting for lexical and one for semantic similarity, are used for two purposes. First, as a guide for our selection of samples for the prompts and, second, as a heuristic for predicted indexing difficulty.

2 Related Work

2.1 Subject Indexing

Automated subject indexing (e.g. see Golub (2021)) can be approached as either a multi-label classification (MLC) task, a keyword extraction/generation problem, or a combination of both (Erbs et al., 2013). To exploit their individual strengths and improve performance, the results from different methods can be combined into a fusion or ensemble (Toepfer and Seifert, 2020).

2.2 Annif

Annif[2] (Suominen, 2019) is a toolkit for automated subject indexing. Two of its implemented methods serve as baseline for our experiments. The first is a Rust implementation of the partitioned label tree approach (cf. Parabel (Prabhu et al., 2018) and

[1] https://gnd.network/Webs/gnd/EN/Home/home_node.html

[2] https://github.com/NatLibFi/Annif/

Bonsai (Khandagale et al., 2020)), called Omikuji[3]. The second baseline is a lexical method based on Maui (Medelyan, 2009), called Maui-Like-Lexical-Matching (MLLM)[4].

2.3 LLMs

LLMs have been applied to a range of tasks (Zhao et al., 2023), including multi-label classification (Peskine et al., 2023; D'Oosterlinck et al., 2024; Zhu and Zamani, 2024), as well as keyword extraction (Maragheh et al., 2023) and keyword generation (Maragheh et al., 2023; Lee et al., 2023). With a prompting procedure analogical to ours, Lee et al. (2023) applied few-shot prompting to generating keywords from abstracts in order to provide an alternative for missing author-defined keywords. D'Oosterlinck et al. (2024) proposed a method utilising interactions of multiple LLMs to infer, retrieve and rank keywords, and thereby bootstrapping prompts in an automated fashion from a set of given few-shot examples.

3 Method

3.1 Model

In our experiments, we opted for a family of LLMs called Luminous[5], developed by the German Company Aleph Alpha[6]. The majority of experiments was done with the Luminous-base model (13B parameters[7]). Fewer experiments were also done with the bigger models, Luminous-extended (30B[7]) and Luminous-supreme (70B[7]), as they have an increased price compared to the base model. For simplicity, we only included findings here that are related to the alteration of prompts.

3.2 Data

All methods were compared on a test set of 486 German scientific book publications. The data was randomly sampled from the catalogue of the DNB. It was filtered for these criteria: German language, publication year 2017 to 2023, and publisher from a list of scientific publishers. To reduce the cost of the experiments, we only included publications from the medicine subject category. Omikuji was trained on a larger dataset (approx. 950.000 training items), disjoint from the above test set and also including other subject categories. This reflects our production settings at the DNB, where all subject categories are indexed by a unified model.

As textual input for the automatic indexing only plain book titles were used. Whereas the full texts of the publications are available, they would need to be cut-off or separated into smaller chunks to process them with the chosen LLM. The heterogeneous structures of these texts also make it difficult to automatically scrape summaries or abstracts from them. Due to our limited resources, we decided not to investigate this additional step and to first experiment on titles before moving on to more costly experiments on longer texts. To be noted, experiments with (shortened) full texts have already been done with the baseline methods and are planned for the LLM-based method, too.

All of the selected publications have previously been intellectually subject indexed with GND entities by professionals with profound expertise in the respective field and the taxonomy. These annotations, further referred to as *labels*, are the gold-standard of our data. The labels all have a unique identifier and one or more short textual description(s). Labels fall under the rough categories of subject headings representing concepts of the various scientific (sub-)disciplines and named entities (personal names, corporate bodies, geographic entities, etc.), the latter constituting the majority of concepts represented in the GND.

3.3 Procedure

Our approach consisted of two steps which have previously been utilised for keyword generation and MLC respectively.

First, as done by Lee et al. (2023), keywords were generated via few-shot prompting. A prompt comprises an instruction ("Extract keywords from book titles.") and a set of examples, illustrating the desired output format of the keywords. See Appendix A for the structure of the prompts.

Next, the generated keywords were mapped to the GND vocabulary, similar to the mapping Zhu and Zamani (2024) conducted in their MLC approach. Here, we used Aleph Alpha's symmetric semantic embeddings[8]. Before vectorisation, the label texts in the target vocabulary as well as the model-produced keywords underwent a simple step

[3] https://github.com/tomtung/omikuji

[4] https://github.com/NatLibFi/Annif/wiki/Backend:-MLLM

[5] https://docs.aleph-alpha.com/docs/introduction/luminous/

[6] https://aleph-alpha.com/

[7] https://docs.aleph-alpha.com/docs/introduction/model-card

[8] https://docs.aleph-alpha.com/docs/tasks/semantic_embed/

of preprocessing by being integrated into a sentence ("A good keyword for this document is *label text / keyword*."). These sentences were vectorised. Via cosine similarity, the most similar label was retrieved for each generated keyword.

3.4 Similarity Measures

Inferring GND entities from book titles alone is a task that can be impossible even for humans, depending on the amount of information or degree of specificity in the particular title. To illustrate, a title like "Report" gives no hints as to what the report is about. To address this problem, we estimated the difficulty of indexing a particular book title by considering two simple notions of similarity between book title and the set of its annotated gold-standard labels. The two similarity measures were used both in the prompt design to select examples and in the evaluation as hypothesised indicator of difficulty.

3.4.1 Lemma Overlap

The first measure aims to capture lexical similarity and is referred to as Lemma Overlap, abbreviated LO (cf. Equation 1). The size of the intersection of lemmas (λ_l) of each label l and lemmas (λ_t) of title t is divided by the number of lemmas in the label[9]. Per book title t, the final score is obtained by averaging over the entire set of annotated gold-standard labels (L_t).

$$\text{Lemma Overlap}(t, L_t) = \frac{1}{|L_t|} \times \sum_{l \in L_t} \frac{|\lambda_l \cap \lambda_t|}{|\lambda_l|} \quad (1)$$

3.4.2 MeanSBERT

To be able to also capture similarity beyond textual overlap, we defined a second measure using Sentence-BERT (Reimers and Gurevych, 2019), called MeanSBERT (cf. Equation 2). The cosine similarity (S_c) between embeddings of title ($\overrightarrow{t}$) and all label texts ($\overrightarrow{l}$) was computed and averaged over all labels[10].

$$\text{MeanSBERT}(t, L_t) = \frac{1}{|L_t|} \times \sum_{l \in L_t} S_c(\overrightarrow{l}, \overrightarrow{t}) \quad (2)$$

3.4.3 Splitting the Dataset

Based on LO and MeanSBERT, the entire test set was split into roughly similarly-sized groups of documents with low, medium and high title-label-similarity. Constructing a cross-table from these groups over both measures lead to a division of the test set into nine separate groups (find details in Appendix C). To exemplify, if the fictitious book title *"Natural language processing"* had *"Computational linguistics"* as its only label, this title would be low LO and high MeanSBERT.

3.4.4 Prompt Design

Analysing similarity between title and labels can also be beneficial for prompt design. If the model is only instructed with examples with high similarity, the labels produced might turn out be closely related to the test title, too, and vice versa. By considering title-label-similarity when constructing the prompt, different behaviour was elicited in the LLM.

4 Experiments

Factoring out base-model selection and other hyperparameters, our experiments were directed at trying out different few-shot sample combinations for the prompts. Table 1 gives an overview of the idea behind them. All of the individual examples in the prompts adhere to the same criteria as the medical test set, but are not part of it. Some prompts only contain samples falling into specific similarity categories of LO and MeanSBERT (`low_low`, `high_low`, `high_high`), while another one includes heterogeneous similarities (`mixed_sim`). Additionally, three more prompts were constructed unrelated to the similarity measures (`deducible`, `combination`, `many_labels`). More details concerning the prompts and the examples used in prompt `high_low` can be found in Appendix B. The previously described procedure was applied to our dataset with all of the seven prompts.

5 Evaluation

5.1 Prompt Variation

Table 2 shows the results of 7 different prompt specifications in comparison (see Appendix D for result set sizes). The prompt with low-similarity examples has the worst F1-performance, suggesting unrelated examples don't guide the LLM well enough. The two prompts with high LO (`high_low`, `high_high`) achieve the two best precision scores, which may, in addition to the high similarity, also be related to the fact that these prompts both contain and generate

[9] Lemmatization was done using spaCy (https://spacy.io/).

[10] We used the Python sentence-transformers library (https://www.sbert.net/) with model "distiluse-base-multilingual-cased-v1".

Prompt name	Comment
low_low	low LO, low MeanSBERT
high_low	high LO, low MeanSBERT
high_high	high LO, high MeanSBERT
mixed_sim	different similarities
deducible	only deducible labels
combination	combination of samples
many_labels	more labels per title

Table 1: Prompt specifications and short explanation.

Prompt	Prec	Rec	F1
low_low	0.231	0.244	0.237
high_low	0.459	0.223	**0.300**
high_high	**0.516**	0.210	0.298
mixed_sim	0.278	0.303	0.290
deducible	0.307	0.280	0.293
combination	0.237	**0.326**	0.274
many_labels	0.207	0.295	0.243

Table 2: Micro-averaged performance of seven prompt combinations.

the smallest number of labels. The best recall scores are attained by prompts with examples from different similarity categories (combination, mixed_similarity). Perhaps this diversity allows the LLM to pick up on a variety of relationships between title and labels and, thus, it can find more correct labels. The best trade-off in terms of F1-score is produced by the prompt high_low.

5.2 Prompt Ensemble

In addition to the individual results, we investigated if the performance would improve when the suggestions of multiple prompt experiments are combined. We used the results of the four prompts high_low, mixed_similarity, deducible and combination. These were selected because they performed well in at least one of the metrics recall, precision or F1-measure. Table 3 shows the outcomes of the combination. The number of experiments i a label was suggested by can serve as a measure of confidence that a label is relevant to a particular title. Keeping all suggestions generated using at least one of the prompts ($i \geq 1$) leads to a high recall strategy. In contrast, considering only those suggestions that all prompts produce ($i \geq 4$) gives a high precision strategy. The best trade-off in terms of F1-score is found in the $i \geq 2$ scenario (a keyword is generated using at least two prompts).

$i \geq$	Prec	Rec	F1
1	0.203	**0.394**	0.268
2	0.322	0.326	**0.324**
3	0.416	0.260	0.320
4	**0.576**	0.166	0.257

Table 3: Micro-averaged results of the prompt ensemble (4 prompts). Parameter i indicates by at least how many prompts a suggestion was made.

5.3 Baselines

Previously introduced baselines, MLLM and Omikuji, are currently well-performing methods in our productive environment. As ranked retrieval methods, they both return a long ranked list of labels, which we truncated at the 5th position. Thus, scores reported are precision@5, recall@5 and F1@5. For the ensemble of prompts, the frequency-of-suggestion i was converted into a score to allow a ranking, too, making results comparable to MLLM and Omikuji. As this ranking is discrete, it is possible for ties between suggestions to appear, so we decided not to include, e.g., precision@1 or precision@2, which could be impacted more severely by this impreciseness.

Table 4 shows the outcomes with 95% confidence intervals. All confidence intervals presented in our evaluation are obtained by bootstrapping the test set, i.e. randomly resampling the documents of the test set. This expresses the uncertainty of results with respect to the variability of the underlying data, but does not include an estimation of model uncertainty. Regarding F1-measures, our LLM method is outperformed by Omikuji and MLLM. Yet, it has better recall than MLLM.

5.4 Similarity Measures

A more detailed comparison between the methods can be found in Figure 1, showing performance stratified by similarity measures with 95% confidence intervals. Generally, F1-scores increase with higher LO. In particular, MLLM, being a lexical method, performs best of all methods in the high LO strata. With MeanSBERT, we do not observe a strong correlation of similarity and F1-score, especially not for the LLM-prompt-ensemble and Omikuji. However, one may observe that the LLM-prompt-ensemble has a slight advantage over MLLM in the low LO strata, indicating that the LLM is able to leverage some sort of world knowledge in order to suggest labels that are not directly

Method	Prec@5	Rec@5	F1@5
Omikuji	0.274 [0.260, 0.292]	**0.462** [0.433, 0.486]	**0.344** [0.326, 0.362]
MLLM	**0.275** [0.262, 0.292]	0.297 [0.281, 0.316]	0.286 [0.271, 0.303]
LLM-prompt-ensemble	0.207 [0.196, 0.218]	0.393 [0.370, 0.413]	0.271 [0.258, 0.285]

Table 4: Micro-averaged results of LLM-prompt-ensemble and baselines. Values in brackets indicate 95% confidence intervals.

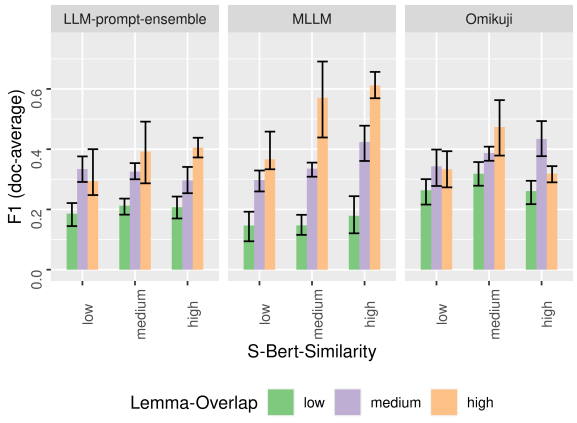

Figure 1: Performance by similarity categories.

derived from the title itself. Still, in this domain of low LO, Omikuji, the trained method, outperforms the other approaches. To conclude, we found tentative support for our assumption that the similarity between title and labels reflects the difficulty of subject indexing a particular book title. We acknowledge that the significance of this graph should not be overestimated, as the number of documents varies between the nine groups (see Appendix C for details).

6 Discussion

Assigning labels to book titles is a difficult task. In a small feasibility study we conducted on 250 titles, almost half of the not-found labels were not deducible for the human annotator by means of the title alone. Even professionals usually need more context. The Luminous models had to perform this task with only a few examples provided. In contrast, Omikuji, as a learning method, has the advantage of observing a multitude of label assignments during training. However, both MLLM and our LLM method can handle labels not observed in training, whereas Omikuji can't.

Our experiments revealed that the combination of prompt examples can impact performance in terms of quality and quantity of the results. The variation in F1-Score between prompts was small, though, with no prompt clearly exceeding all others. Using different or enhanced sets of examples could further improve performance.

7 Conclusion and Future Works

While we didn't find our LLM-based method to outperform the baselines at hand, our experiments on subject indexing German medical book titles revealed insights on factors for successful prompt combination. With the few examples fitting into a prompt, one can tweak results in specific directions, e.g., to optimise precision. In our case, the similarity measures were the main criterion for the selection.

In the future, our goal is to provide a benchmark study on the task of subject indexing, in order to support other libraries and institutions. A new perspective for the evaluation of this task has been introduced by the similarity measures. We plan to include results from a larger dataset of more diverse titles as well as a dataset with the complete texts of the scientific publications. We also want to evaluate our LLM-based approach on these. Furthermore, we will look into automated procedures for prompting, as done in D'Oosterlinck et al. (2024).

Acknowledgements

This work reports on findings from an ongoing project at the DNB[11], which is funded by the German Minister of State for Culture and the Media. With the AI strategy, the German federal government is supporting the research, development and application of innovative technologies.

We kindly acknowledge the support of Aleph Alpha for setting up our experiments with their API and inspiring the few-shot prompting approach as a potential solution for our automated subject indexing problem.

[11] https://www.dnb.de/EN/Professionell/ProjekteKooperationen/Projekte/KI/ki_node.html

Ethical Considerations

Bommasani et al. (2023) compared what they refer to as *Foundation Models* with respect to their current compliance with the upcoming EU AI Act. Aleph Alpha's Luminous models were among the examined models. Regarding different factors, including, for example, data transparency and energy consumption, the Luminous models (and other LLMs) didn't fulfill (all) the defined compliance criteria. This is a reminder that LLMs have to be utilised under great care and responsibility and that it is important to acknowledge their shortcomings in terms of transparency and reproducibility.

Stereotypes and other discriminatory artifacts in the LLM, which could have been present in the model's training data, might impact which entities are assigned to an incoming publication, either in the generation or the mapping step. Users visiting the DNB use subject headings and other GND descriptors (automatically or intellectually assigned) to research literature. Misleading terms, no matter if they result from stereotypes in the data, lack of model-performance or human mistakes, can negatively impact the results of this search.

Limitations

All our findings only relate to one (family of) LLM(s). The performance of other language models may differ.

Furthermore, the present study was done on a small restricted dataset. Thus, findings cannot be transferred or generalised to different datasets and other tasks.

Also, our experiments of interchanging few-shot examples are not exhaustive. Better prompt combinations, prompt structures and prompt instructions may exist. Samples for the prompts were partially chosen from a specific data subset (e.g. with specific similarities) and in other cases from the entire dataset, but always by subjective perception and not in a randomised way. This may have introduced unintentional bias in the composition of the examples.

Finally, the experiments presented in this study originate from a project with limited resources. Inevitably, this has affected our choices in our experiments, which always have the primary objective of improving our production settings.

References

Rishi Bommasani, Kevin Klyman, Daniel Zhang, and Percy Liang. 2023. Do Foundation Model Providers Comply with the EU AI Act? https://crfm.stanford.edu/2023/06/15/eu-ai-act.html.

Karel D'Oosterlinck, Omar Khattab, François Remy, Thomas Demeester, Chris Develder, and Christopher Potts. 2024. In-context learning for extreme multi-label classification. *arXiv preprint arXiv:2401.12178*.

Nicolai Erbs, Iryna Gurevych, and Marc Rittberger. 2013. Bringing order to digital libraries: From keyphrase extraction to index term assignment. *D-Lib Magazine*, 19(9/10):1–16.

Koraljka Golub. 2021. Automated subject indexing: An overview. *Cataloging & Classification Quarterly*, 59(8):702–719.

Sujay Khandagale, Han Xiao, and Rohit Babbar. 2020. Bonsai: diverse and shallow trees for extreme multi-label classification. *Machine Learning*, 109(11):2099–2119.

Wanhae Lee, Minki Chun, Hyeonhak Jeong, and Hyunggu Jung. 2023. Toward keyword generation through large language models. In *Companion Proceedings of the 28th International Conference on Intelligent User Interfaces*, IUI '23 Companion, page 37–40, New York, NY, USA. Association for Computing Machinery.

Reza Yousefi Maragheh, Chenhao Fang, Charan Chand Irugu, Parth Parikh, Jason Cho, Jianpeng Xu, Saranyan Sukumar, Malay Patel, Evren Korpeoglu, Sushant Kumar, and Kannan Achan. 2023. LLM-TAKE: Theme-aware keyword extraction using large language models. In *2023 IEEE International Conference on Big Data (BigData)*, pages 4318–4324.

Olena Medelyan. 2009. *Human-competitive automatic topic indexing*. Ph.D. thesis, The University of Waikato, New Zealand.

Youri Peskine, Damir Korenčić, Ivan Grubisic, Paolo Papotti, Raphael Troncy, and Paolo Rosso. 2023. Definitions matter: Guiding GPT for multi-label classification. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4054–4063, Singapore. Association for Computational Linguistics.

Yashoteja Prabhu, Anil Kag, Shrutendra Harsola, Rahul Agrawal, and Manik Varma. 2018. Parabel: Partitioned label trees for extreme classification with application to dynamic search advertising. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, page 993–1002, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Osma Suominen. 2019. Annif: DIY automated subject indexing using multiple algorithms. *LIBER Quarterly: The Journal of the Association of European Research Libraries*, 29(1):1–25.

Martin Toepfer and Christin Seifert. 2020. Fusion architectures for automatic subject indexing under concept drift: Analysis and empirical results on short texts. *International Journal on Digital Libraries*, 21(2):169–189.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models. *arXiv preprint arXiv: 2303.18223*.

Yaxin Zhu and Hamed Zamani. 2024. ICXML: An in-context learning framework for zero-shot extreme multi-label classification. *arXiv preprint arXiv:2311.09649*.

A Prompt Structure

The following table shows the prompt structure. *[...]* indicates positions to fill with example titles and keywords. Keywords are comma-separated. For an incoming test title, the *Schlagwörter*-field remains empty.

Original	Translation
Extrahiere Schlagwörter aus Titeln. Text: *[...]* Schlagwörter: *[...]* ### Text: *[...]* Schlagwörter: *[...]* ### ... ### Text: *[...]* Schlagwörter:	Extract keywords from titles. Text: *[...]* Keywords: *[...]* ### Text: *[...]* Keywords: *[...]* ### ... ### Text: *[...]* Keywords:

B Prompt Details

B.1 Prompts Unrelated to Similarity

The prompt `deducible` contains examples where all assigned labels are deducible from the title. As the defined similarity measures are each averaged over the entire set of labels of a title, even titles in both high-similarity categories may have labels not inferable from the title. The prompt `combination` contains examples used in other prompts, but designed without a focus on a given similarity category. Just as the `mixed` prompt, it was meant be more diverse in the nature of its included examples than the prompts with only samples from a single similarity group. The prompt `manylabels` contains more labels per title than any of the other prompts. As such, it is like a counterpart to prompts `high_low` and `high_high` with only few labels per title.

B.2 Prompt Characteristics

The table below shows the number of examples and average number of labels in the prompts.

Prompt	Examples	Avg. Labels
`low_low`	8	2,75
`high_low`	8	1,25
`high_high`	8	1,38
`mixed_sim`	8	3,38
`deducible`	8	2,63
`combination`	8	4,88
`many_labels`	6	9

B.3 Example Prompt Combination

In the following, the examples in the prompt `high_low` are listed, along with translation and a reference to the title in the catalogue of the German National Library. Other sample combinations are available on request.

- Stottern Erkenntnisse, Theorien, Behandlungsmethoden (*Stammering Findings, Theories, Methods of Treatment*); **Labels**: Stottern (*stammering*) [https://d-nb.info/1003711952]
- Last minute - Gynäkologie und Geburtshilfe [fit fürs Examen in 2 Tagen!] (*Last minute - gynecology and obstetrics [prepared for the exam in 2 days!]*); **Labels**: Gynäkologie, Geburtshilfe (*gynecology, obstetrics*) [https://d-nb.info/1010285904]

- Hilferuf Essstörung Rat und Hilfe für Betroffene, Angehörige und Therapeuten (*Cry for help Eating disorder advice and help for persons concerned, relatives and therapists*); **Labels**: Essstörung (*Eating disorder*)[https://d-nb.info/1017606552]
- Rückenschule für Kinder mit Spiel und Spaß Schmerzen lindern und Haltungsschäden vorbeugen (*Back therapy training for kids Relieve pain with fun and games and prevent postural defects*); **Labels:** Kind, Rückenschule (*Child, back therapy training*) [https://d-nb.info/1102547840]
- Organsysteme verstehen - Niere integrative Grundlagen und Fälle (*Understanding organ systems - Kidney Integrative foundations and cases*); **Labels:** Niere (*Kidney*) [https://d-nb.info/113137469X]
- Schlafstörungen wieder tief und gesund schlafen; New-Age-Musik (*Sleep disorders Sleep soundly and healthily again; New age music*); **Labels:** Schlafstörung (*Sleep disorder*) [https://d-nb.info/1201018668]
- Wenn Töne Farben haben Synästhesie in Wissenschaft und Kunst (*When sounds have colours Synesthesia in science and art*); **Labels:** Synästhesie (*Synesthesia*) [https://d-nb.info/984370986]

C Samples in Similarity Categories

		LO		
		low	**med**	**high**
MeanSB.	**low**	45	31	5
	med	97	161	15
	high	51	40	40

D Result Set Sizes

Prompt	**# Result set**
low_low	1586
high_low	730
high_high	611
mixed_sim	1638
deducible	1371
combination	2067
many_labels	2141
ensemble ($i \geq 1$)	2920