A Benchmark Collection for Assessing Scholarly Search by Non-Educated Users

Simple Text 2022–2024 Content Search Task: results and lessons

Stéphane Huet[1], Éric SanJuan[1]

[1]Avignon Université, LIA, 339 chemin des Meinajariés, BP 1228, 84911 Avignon Cedex 9, France

Abstract

This paper presents a resource that was collected in the context of the CLEF SimpleText track. During three editions, Task 1 has focused on the retrieval scientific abstracts in response to a query derived from a popular science article. Several baseline systems were proposed to participants, leveraging bag-of-words models and dense vector document representations. A key element for evaluation was the development of query-document relationships (Qrels). We describe the collected data and conduct an extensive analysis of these annotations. We evaluate the behaviour of several systems on this resource, which is made available for further assessment.

Keywords

Information Retrieval, Science Popularization, Dense Search

1. Introduction

Science has long been the driving force behind human progress, shaping our understanding of the world and improving the quality of life for individuals around the globe. From the simplest household appliances to the most complex medical treatments, science is an integral part of our daily lives. If the internet has made easier access to scientific papers, harnessing the knowledge and the language of the scientific literature can be challenging for the population. Besides, the abundance of online sources and academic publications can make it challenging for non-experts to find reliable information on complex scientific topics.

The general public tends to prefer easily understandable information on social media and websites prioritizing commercial or political goals rather than correctness and informational value, either of which may be unreliable. Conspiracy and speculative sources are often chosen by users as they provide a single simple idea explained in plain language, seem to be coherent, and do not require prior background knowledge.

During three editions (2022-2024), the CLEF SimpleText Track 1 asks participants to retrieve scientific abstracts in response to a query prompted by a popular science article [1, 2, 3]. In the context of this track, we introduce a new test collection, called *SimpleText-1*, for scientific information access by the general public. This test collection consists of

- a large corpus of scientific abstracts;
- a set of relevance labels (qrels);
- additional automatic judgments based on dense vector representations and LLMs;
- a complete relational framework for efficient storage and retrieval of multiple embeddings for short passage;
- multiple baseline systems for meta evaluation of qrels.

Our implementation is based on open source PostgreSQL and its integration of SQL extensions for textual search within normalized relational schema based on Generalised Inverted Indexes (GIN) and dense vector types. Documents are stored as JSON, since the JSON field type allows for the extraction

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain

stephane.huet@univ-avignon.fr (S. Huet); eric.sanjuan@univ-avignon.fr (É. SanJuan)

of elements from decisions and is compatible with join operations. The pgvector[1] library enables vector operations, such as dot product for snippet search [4] and is compatible with ordering and aggregation operations. With its flexibility, the relational schema supports experimentation with diverse passage aggregation methods, thereby extending the scope of our system from searching specific passages to retrieving complex sets of documents.

The paper is organized as follows. Section 2 discusses related work. Section 3 describes processed data including extended qrels. Section 4 introduces a general resource to reproduce and extend experiments. Section 5 covers extended experiments on long queries. Finally, Section 6 presents conclusions and future perspectives.

2. Related work

For decades, specialized scientific documents have been a core component of IR systems [5]. Not only are they crucial for researchers seeking to stay current with the latest advancements in their field, but also for anyone interested in staying informed about recent scientific breakthroughs and developments. Journalists play a vital role in making complex scientific content more accessible and widely disseminated; their initiative includes *Nature*[2], *The Guardian*[3], *ScienceDaily*[4], *ScienceX*[5].

The constant increase in scientific publications necessitates the use of automated tools for information retrieval and summarization [6]. This trend raises two key challenges: making complex research accessible to non-expert readers who struggle with technical terminology and academic structures; and providing experts with a more detailed understanding without requiring them to read lengthy papers. Furthermore, growing concerns about public misinformation and disinformation campaigns highlight the need for developing technologies that cater to diverse audiences [7].

Dense representations of documents have significantly advanced the state of the art in information retrieval (IR). These dense vectors, often generated by sophisticated language models, capture rich semantic information, enhancing the ability to retrieve relevant documents accurately.

Despite their effectiveness [4, 8], these neural models are resource-intensive, requiring substantial amounts of data for training as well as significant computational power [9]. This high demand for resources often necessitates a hybrid approach to document retrieval. An initial retrieval phase may use a more traditional and less computationally demanding method, such as tf-idf vectorization, which is based on keyword matching [10]. The documents retrieved in this phase are then re-ranked using the dense representations provided by the neural models [11, 12]. This two-step process leverages the strengths of both methods, combining the efficiency and scalability of BM25 with the nuanced understanding of document relevance offered by neural approaches.

The final step in presenting search results to users frequently involves Language Models in Retrieval Augmented Generation (LLMs in RAG) [13]. This innovative approach utilizes large language models to generate answers or summaries for the user, based on the information contained in the top-ranked documents [14]. While powerful, this method carries the risk of "hallucinations" or the generation of plausible but incorrect or unsupported information.

Based on the experience of running the CLEF Simple Text task 1 track, this paper proposes a lightweight complete system to explore the application of these advancements on the use case of scientific search. We demonstrate how relational schemas, enhanced with extended JSON and vector types, can efficiently manage multiple embeddings for scientific references. By integrating passage retrieval techniques with relational database operators, this hybrid approach allows us to combine dense vector search with Boolean search and related symbolic approaches like formal concept analysis [15].

[1] https://pypi.org/project/pgvector/
[2] https://www.nature.com/news
[3] https://www.theguardian.com/science
[4] https://sciencedaily.com/.
[5] https://sciencex.com/

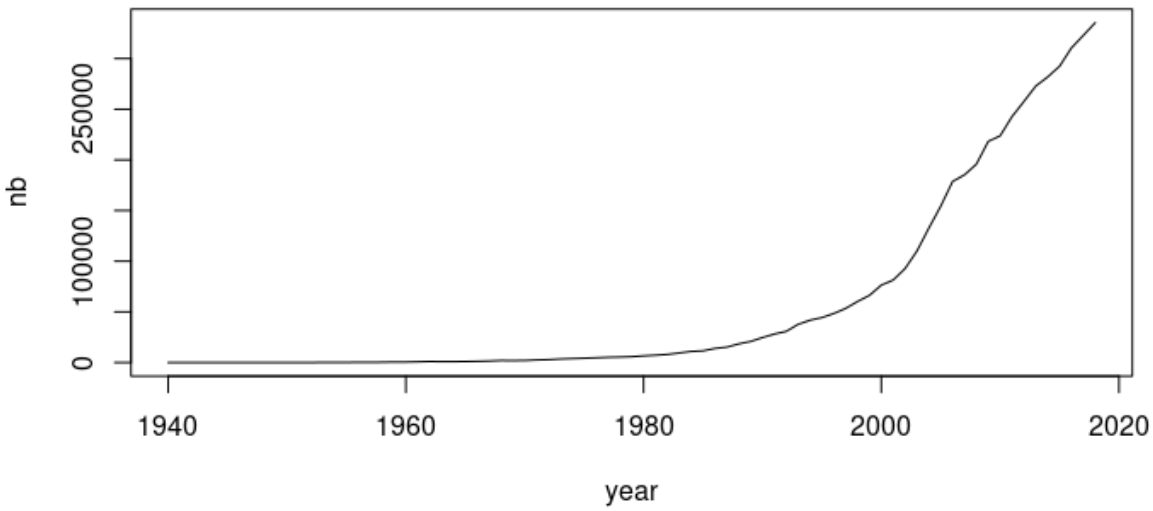

Figure 1: Number of publications per year in the corpus.

3. Data

3.1. Corpus

The corpus considered throughout the three editions of Task 1 is the Citation Network Dataset: DBLP+Citation, ACM Citation network (12th version) [16][6]. This data set is derived from scientific publications within the fields of computer science and related disciplines; it serves here as a source of scientific documents that can be used as reference passages. It provides:

- $4,894,083$ bibliographic references published before 2020,
- $4,232,520$ abstracts in English,
- $3,058,315$ authors with their affiliations, and
- $45,565,790$ ACM citations.

From this corpus can be extracted textual content together with authorship. Figure 1 provides the time span of the corpus.

3.2. Topics and queries

Search requests are derived from a pool of 40 press articles written for a general audience. These news articles, used as *topics* for the task, include 20 articles sources from *The Guardian* (G01-G20 topics), an influential global newspaper featuring a tech section, and 20 from *Tech Xplore*[7] (TG01-T20 topics), a website taking part in the Science X Network to provide a comprehensive coverage of engineering and technology advances [17]. These two sets of articles cover various domains of computer science and electronics (AI, networking, cybersecurity, bioinformatics...). For each topic, the title and full text of the article are provided, with a link to the online page that may contain images and references.

To guide participants through the various facets covered in the original press article, between 1 to 4 *queries* are provided per topic (Table 1). These queries, made of few keywords (e.g. “privacy”, “OTP

[6] https://www.aminer.cn/citation
[7] https://techxplore.com/

Table 1
CLEF SimpleText Task 1 Topics and Queries.

	Topics	Queries	#Queries	Examples
Guardian	G01-G20	G*.[1-4]	42	gene editing, drug discovery, crispr, forensics, advertising, Snowden
		G*.C[1-5]	63	how algorithms are designed with human interaction in mind
Tech Xplore	T01-T20	T*.[1-4]	67	phototransistor, 3G, energy efficiency, empathy, Bayesian approach

memory" or "intelligent parking"), bridge the gap between our task and traditional information retrieval, enabling the application of established relevance metrics. Queries were crafted by two computer scientists to pinpoint essential technical concepts and indicate potentially tricky areas for lay readers. Furthermore, they have manually verified that each query enables participants to access at least 5 relevance excerpts from the abstracts available in the corpus, which can be used as potential sources for citation within a press article.

The two subsets of queries, all related to Information Technology (IT), have distinct characteristics. The Guardian (G) topics are grounded in real-world societal issues such as privacy and misinformation. In contrast, Tech Xplore (T) topics are directly linked to original research papers and focus on technical aspects like neural networks and indoor positioning systems. The task inherently involves disambiguating relevant information within the scope of specific articles; this is particularly critical for the G queries that relate to topics outside the IT domain.

In 2024, the set of queries were enriched by 62 long queries, from 2 to 5 per Guardian topic. These expanded queries (G*.C*) were generated by GPT4 with a prompt asking to list the main subtopics related to computer science and using the integrated press article as context. The queries were manually reviewed to ensure their accuracy, confirming that each was correctly associated with the underlying article and did not duplicate any existing queries. Figure 2 provides the ten long queries that have been evaluated for the 2024 test.

3.3. Qrels

The Qrels collection was built iteratively throughout the track's editions. While the full set of 40 topics and their initial short queries were made available to participants from the first edition, the long queries were only introduced for the 2024 edition. The generation of relevance judgments followed a cumulative process. For each edition, a new set of assessments was created using a pooling method, where a sample of the results submitted by the participants was manually judged. These newly created judgments were then added to the training set for the following edition. To ensure a fair assessment and to limit the effectiveness of pure machine learning approaches that might overfit on the existing labels, the official test for each edition was always conducted on queries that had not yet been evaluated.

More specifically, during the three editions of the track, the relevance documents were judged from their title and their abstract. Judgments were made by assessing how well each retrieved document addressed the query and corresponded to relevant aspects of the original news article, providing meaningful insights into one or more of its key themes.

For the 2022 edition, a scale of 0 to 5 was used. A total of 475 documents were assessed from a pool of documents chosen by at least two participants on several topics. This built qrels was shallow with fewer than 7 documents assessed on average per query. In 2023, we dramatically expanded the qrels with annotations made by students (mainly two master students in computer science) with a more limited range of scores between 0 and 2 (the higher the better) to speed up the process. A subset of qrels was first released for training on the G01-G15 topics; test was carried out on 5 G topics and 5 T topics (see Table 2, lines 1 and 2). A pooling established from the 2023 runs was assessed by two researchers in computer science (line 2). For the 2024 edition, the training qrels included all the previously released qrels (lines 1 and 2) and additional judgments done on 2023 pooling for the first 15 G topics (line 3). We pooled all documents retrieved at depth 10 from all submitted systems in 2024 and judgments were done by two researchers in computer science, one who focused on the G topics and the other on T

Table 2
CLEF SimpleText Task 1 Qrels Collection Statistics.

Qrels	Topics	#Queries	#Assessed abstracts		
			0	**1**	**2**
2023 train	G01–G15	29	672	271	356
2023 test	G16–G20, T01-T05	34	2174	345	1207
2024 train	G01–G15	30	790	130	83
2024 extended test	G01–G20, T01-T05, T12-T20	66	3,681	991	457
2024 test	G01.C1–G10.C1, T06–T11	30	2,775	1,500	579

Figure 2: Evaluated long queries with press article context.

G01.C1 Concerns related to the handling of sensitive information by voice assistants.

G02.C1 How children interact with voice assistants and the design of child-friendly interfaces.

G03.C1 Use of AI to improve success rates and speed in the pharmaceutical research field.

G04.C1 Application of machine learning algorithms to predict genomic features, functions, and the outcomes of gene-editing interventions like Crispr.

G05.C1 How AI systems, especially virtual assistants, can perpetuate gender stereotypes?

G06.C1 Ethical considerations, governance frameworks, and policies for the responsible development and deployment of AI technologies.

G07.C1 Use of NLP techniques to detect and analyze misinformation in textual content on social media platforms.

G08.C1 Understand the cryptographic underpinnings of blockchain technology, which is the foundation of Bitcoin and other cryptocurrencies

G09.C1 Computer science techniques to analyze spatial data and imagery, particularly for reconstructing crime scenes or human rights violation incidents

G10.C1 Study of robotic technologies and automated systems that are replacing human labor in various sectors.

topics. A complete assessment was done at depth 10 on the queries of the T06-T11 topics and 10 long queries (G01.C1-G10.C1) and released as the official 2024 test qrels (line 5). Supplementary judgments were done on other queries from this pool (line 4).

To standardize the annotation process and reduce inconsistencies between annotators and periods, the same two annotators carried out a large proportion of the judgments each year. They reviewed and agreed upon their assessments of a small sample of documents. Given that discrepancies could be significant on a five-point scale, we opted to switch to a more streamlined three-point scale to improve efficiency while maintaining annotation quality.

Throughout the 3 editions, $16,011$ query relevance judgments were assessed on this collection. Table 2 summarizes the size of qrels and when they were released. For future use, we split this resource into two sets: $11,157$ judgments on 95 queries (lines 1 to 4, with an average of 117.4 assessments per query) for training, and $4,854$ judgments on 30 queries for test (line 5, with an average of 161.8 assessments per query).

Let us note that the train dataset can be expanded using documents with similar titles. For example, the simple SQL procedure in appendix B, which is based on the similarity between title embeddings stored in our PostgreSQL database, searches for documents of the dataset collection very closed to documents with a 2 relevance score in the qrels. This procedure finds more than $2,000$ relevant extra documents adding up to a $13,407$ qrels set. For this extension, we focus solely on the topics associated with the training set of our ground truth (qrels).

We also compared our manual evaluations with recent LLMs that can be run internally. Despite a significant correlation with human annotations on the long queries, the gap is too important. Table 3 provides the results for 6 models based on the following prompt:

prefix Here is a societal question and a scientific paper in computer science. Please, do not recommend papers that are off topic. I do not have time to read them all.

prompt Answer only returning a relevance score 0, 1 or 2. 0: Not really relevant, 1: relevant, 2: very relevant.

system You are a journalist writing about a tech topic that raises societal questions. You are looking for scientific publications that could feed your paper for a large audience.

Table 3
Comparison of LLM Models to generate q-rels by tau, P-value, and Accuracy Levels.

Model	tau	P-value	Accuracy (3)	Accuracy (2)
Qwen	-1.25 %	63.29 %	23.24 %	48.44 %
Qwq	30.00 %	***	32.17 %	52.73 %
Gemma3 :small	33.26 %	***	37.00 %	58.45 %
Gemma3:12b	31.95 %	***	38.29 %	59.59 %
Phi4	41.54 %	***	45.16 %	62.79 %
Llama 4	40.04 %	***	52.11 %	69.58 %

These results exhibit a Kendall's tau very close to 0 for Qwen, which indicates there is no correlation with the human judgments. For the other LLMs the conclusion is not so clear-cut, but the agreement is moderate. The accuracy measured in three classes (0, 1, 2) remains low, even for the best LLM LLama 4 which barely exceeds 50 %. Even the number of classes is reduced to 2 (0 and >0), the accuracy is still under 70 %.

3.4. Complexity and credibility scores

Retrieving relevant information to a query asked by a non-educated user is essential. However, in order to be exploited, it has to come from a reliable source or be simple enough to be understood.

For the credibility part of the source, it is reasonable to consider the collection of original documents, which consist of scientific publications, as relatively trustful, at least more than any source from social networks. Besides, the corpus provides additional information that can be used for this purpose. For example, the number of citations of a given document can be obtained to take into account the peer recognition. Similarly, the number of bibliographical references cited in a document can assess that an effort has been made to situate the work within the scientific community [18].

For the simplicity part, it is possible to turn to classic readability indices, such as the well-established Flesch–Kincaid Grade Level test (FKGL). We computed these readability scores from abstracts of all documents and released them. Since these scores can be easily manipulated or usually overestimate difficulty for technical or specialized texts [19], we provide complementary indices determined with the NLTK[8] and Readability[9]:

- the average number of characters per word,
- the average number of syllables per word,
- the number of long words (at least 7 characters),
- the number of complex words (at least 3 syllables, ignore proper nouns and numbers),
- the size of the vocabulary of the abstracts.

[8] https://www.nltk.org/
[9] https://pypi.org/project/readability/#description

Table 4
Number of assessed abstracts by Credibility and Complexity annotations [20].

Annotators	Credibilty			Complexity		
	0	1	2	0	1	2
B. stud. in Humanity	402	907	549	542	815	515
B. stud. in Comp. Sci.				72	114	109
M. Sc. stud. in Comp. Sci.	1172	361	712	1274	548	424

Table 5
Evolution of automatic complexity measures toward human assessments.

Metrics	Bachelor stud.			Master stud.		
	0	1	2	0	1	2
FKGL	15.02	15.16	15.25	15.1	15.38	14.72
#words	116.36	138.24	146.13	128.48	149.49	154.06
#complex words	29.93	36.61	38.11	34.01	39.5	39.98
vocabulary size	76.91	88.76	92.26	83.15	95.02	96.21

To complement these automatic metrics, a small subset of documents have been evaluated by students [20, 21, 22] on a scale of 0-2 across the two dimensions: credibility (the higher, the more credible a document is judged) and complexity (the higher, the more complex). Table 4 summarizes the number of annotations by three groups of annotators: Bachelor of Arts students in Humanities, 1 Bachelor student in Computer Science and 2 Master students in Computer science. It should be noted that the same documents were not annotated by these three populations.

The human assessments can be used to study how automatic metrics correlate with perceptual measures of complexity. Table 5 shows by way of example how four metrics evolve w.r.t. scores given by bachelor or master students. FKGL moderately increases with the judgments made by bachelor students, while it surprisingly drops for the higher score given by master students, which suggest that the indices used by FGKL are not the same taken into account by humans. Other indices, such as the number of word occurrences inside the abstract, the number of complex words, or the size of vocabulary, follow a more expected trend, with a concomitant increase of scores given by humans.

4. Resources

We provide a PostgreSQL relational database with all the data and baseline embeddings. Figure 3 describes its relational schema.

4.1. Relational vector database

We use PostgreSQL as our relational database management system:

1. the JSON type allows for managing the content of the entire documents as unique values on which tree traversal operators can be applied to extract sub-elements;
2. Generalized Inverted Indexes (GIN) allow indexing of textual content for classic textual search and can be used with specialized dictionaries;
3. we use a simple ivvflat index on vectors which corresponds to a k-means and a quadratic reduction in the computation time of the nearest neighbours.

The whole database presented in figure 3 is available for CLEF Simple Text[10] and MADICS[11] participants in standard SQL code or Docker image with integrated services.

[10] https://simpletext-project.com/
[11] https://www.madics.fr/

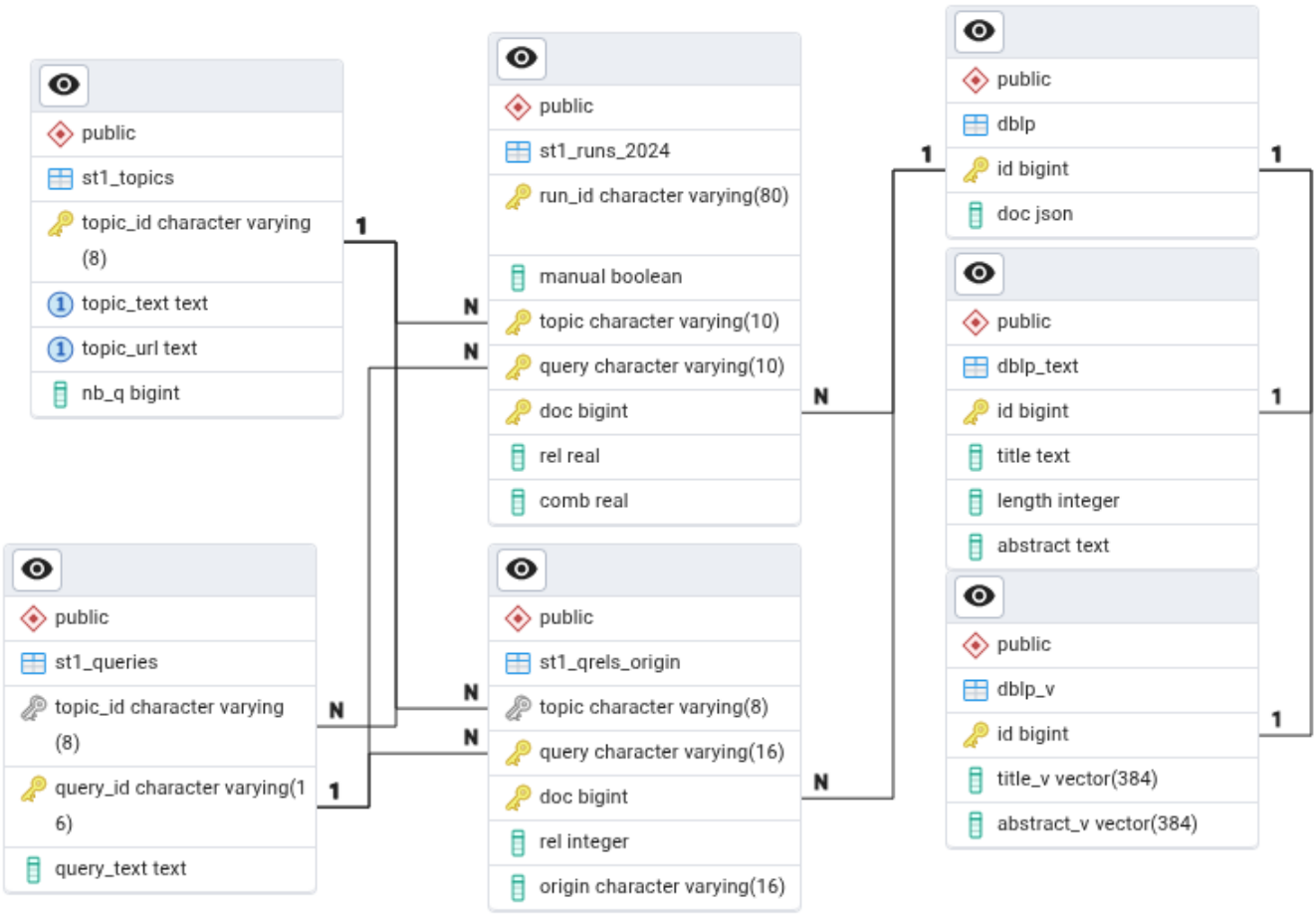

Figure 3: Relational schema including JSON corpora and q-rels.

Documents are stored in JSON format in the central table dblp. Two relations are derived from it:

dblp_text with all textual content extracted from title and abstract fields when available and indexed for full-text search (GIN).

dblp_v with the embeddings generated by this textual content.

This architecture implies the following functional dependencies:

$$\textbf{dblp} \longleftarrow \textbf{dblp\_text} \longleftrightarrow \textbf{dblp\_v}$$

The dblp_text and dblp_v relations could therefore be joined, but this ability to manage text content and vector representation independently has two advantages:

- the addition of a dense representation of the texts does not alter the management of documents nor the indexing of excerpts;
- multiple vector dimensions can be considered; if we proceeded here with dimensions less than 350, it is possible to add representations of higher cardinality in separate relations.

We experimented with the ms-marco-minilm model[12] [23, 24] to generate the embeddings of titles and abstracts. This model has been trained on data before 2020, so it is based on data anterior to publications used to define query topics. It is also frugal enough to be easily refined or even re-learned on specialized corpora, which we consider doing subsequently.

[12] https://huggingface.co/cross-encoder/ms-marco-MiniLM-L-12-v2

Sentence transformers are computationally efficient, needing only a single forward pass to capture nuanced word relationships [25]. Here, we used them as pre-trained on the MS Marco dataset, without further fine-tuning.

Topics, runs and qrels are stored in separated relations with multiple references among them ensuring data integrity and empowering multiple q-rels extensions based on text embeddings. Appendix A shows the SQL code for to retrieve the nearest documents based on title similarities and Figure B applies this function to compute q-rels extensions.

To enrich the analysis, an additional table, `st1_complexity_metrics`, has been added to the schema. This table associates each record from the DBLP corpus with a set of 20 textual complexity metrics. These measures, computed using both the NLTK and `readability` libraries, make it possible to assess the readability of the scientific abstracts.

The table contains the following columns:

NLTK-based indicators `sent`, `tok`, `cmu_tok`, `cmu_syl`, `cmu_wrd_per_sent`, `cmu_syl_per_wrd`, and `cmu_fkgl` (Flesch-Kincaid Grade Level).

Readability-based indicators `sent`, `tok`, `syl`, `read_char`, `wordtypes`, `long_words`, `complex_words`, `char_per_wrd`, `wrd_per_sent`, `syl_per_wrd` and `type_token_ratio`.

The integration of these metrics directly into the relational database allows them to be easily used in SQL queries to filter, sort, or analyse documents based on their perceived complexity, in combination with keyword or dense vector searches.

4.2. Integrated baseline system

We provide a complete online baseline system based on the light paragraph cross-encoder MS MARCO Mini LM (all-MiniLM-L6-v2)[13]. This is done by adding to PostgreSQL database, with pgcurl[14], an integrated online embedding service for user queries written in natural language running in real time on CPU.

We also explore sparse retrieval at the passage level using PostgreSQL GIN indexes with default resources for conjunctive queries, requiring all query tokens to appear in the passage.

Table 6 shows relevance evaluation for scientific document retrieval, with a ranking by NDCG@10 (CLEF 2024 task 1 official measure). We report rankings based on two different embeddings: titles and abstracts. We also report plain BM25 results powered by an external ElasticSearch that was initially used as a baseline for previous editions of the task.

Results in table 6 shows that dense retrieval significantly outperforms all sparse approaches. Title based embeddings outperform full abstract based ones on CLEF official measure, but not on MAP, leaving room for improvement by combining them. A first take away here is that mini LM models facilitate the integration of efficient IR procedures into relational databases with additional vector types, for short documents indexing.

Table 6
Evaluation on CLEF SimplText task 1 2024 test dataset.

Run	MRR	Precision		NDCG		Bpref	MAP
		10	20	10	20		
baseline_vir_title	**0.8454**	**0.6933**	**0.4383**	**0.5090**	**0.4010**	0.3594	0.1534
baseline_vir_abstract	0.7683	0.6000	0.4067	0.4269	0.3539	**0.3857**	**0.1603**
baseline_bool	0.7242	0.5233	0.3633	0.3409	0.2906	0.2661	0.1199
baseline_BM25	0.6173	0.3733	0.2900	0.2818	0.2442	0.3016	0.1325

[13] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2
[14] https://github.com/pandrewhk/pgcurl

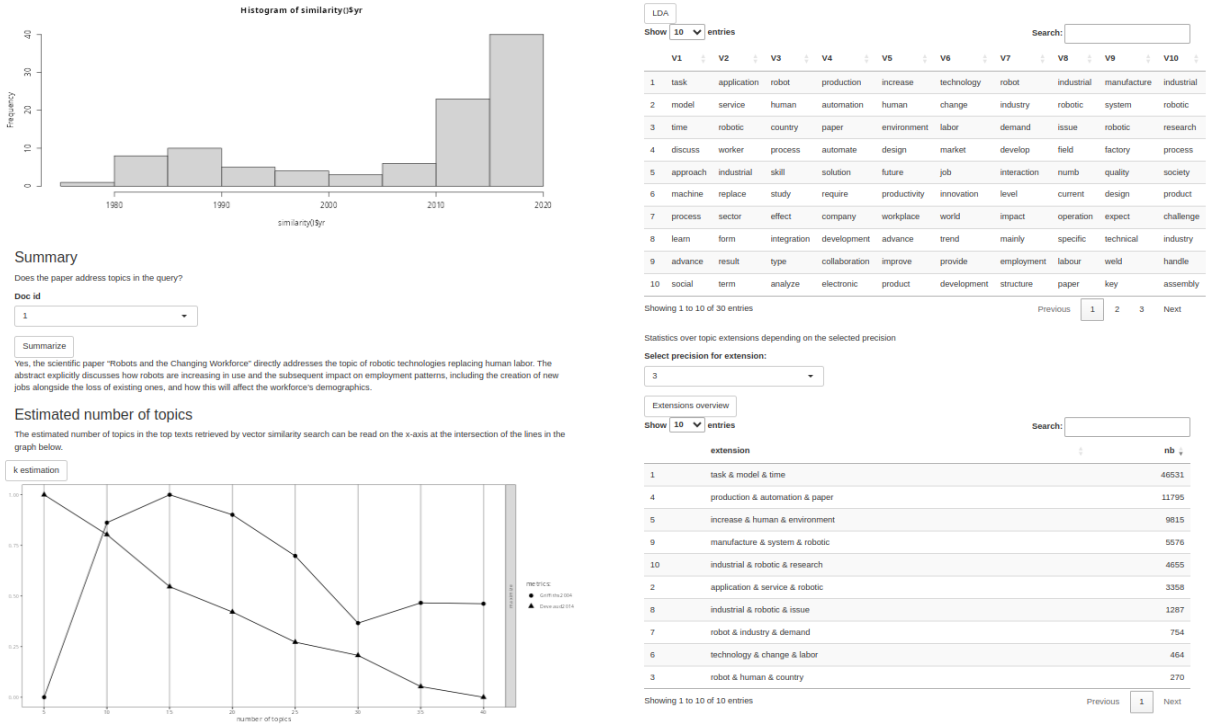

Figure 4: On line shiny interface to analyse topic coverage at clef24st1.iut-avignon.fr.

5. Long query analysis

As a user case of the resource, we investigate their potential for exploring the contents of the corpus in relation to targeted queries. In this section we focus on the 10 long queries from The Guardian presented in figure 2.

The total number of available qrels for these ten topics is 1, 282 with a minimum of 92 and a maximum of 158. The score repartition is (0: 672,1: 311,2:2 99). Table 7 provides baseline results on these ten queries and related qrels. If the relevance scores are higher on these queries than in the whole test dataset (Table 6) the systems relying on embedding vectors still are more effective.

To investigate the corpus's contents, we use a Shiny interface[15] (Figure 4) powered by R and PostgreSQL to carry out this analysis. In a nutshell, a k Nearest Neighbors search based on mini LM models is combined with the Latent Dirichlet Allocation (LDA) and an exhaustive Boolean search for topic coverage analysis. This interface can be experimented for any plain text query and provides insights about subjects covered in the corpus through triplets of terms.

To implement this tool, a sample of 100 documents is retrieved per query using the vector abstract baseline system powered by the PostgreSQL system and the same similarity score as appendix A, but this time using abstract embeddings instead of title embeddings. Then, LDA [26, 27] is applied with the number of topics fixed as 10 on a sample of 100 retrieved abstracts. Each topic is characterized by three words, which provide a gist on the content of the topic and are used as queries to a Boolean search. In this way, frequencies of each triplet of words can be computed in real time from the GIN index over these sets over all available abstracts.

Table 8 shows for each long query, one of the most frequent triplets of words in the corpus (first quartile) and one of the least frequent (last quartile). Results provide a general idea of corpus coverage. For example, for the "Concerns related to the handling of sensitive information by voice assistants" query, related articles of the original collection frequently deal with "agent, technology and people". Other relevant articles, albeit less common, are about "attack, command and google".

[15] http://clef24st1.iut-avignon.fr

Table 7
Evaluation on CLEF SimpleText task 1 2024 long queries test dataset.

Run	MRR	Precision		NDCG		Bpref	MAP
		10	20	10	20		
baseline_vir_title	0.9333	**0.8600**	0.5650	**0.7184**	0.5415	0.5196	0.2633
baseline_vir_abstract	**0.9500**	0.8200	**0.6150**	0.6701	**0.5543**	**0.5533**	**0.2996**
baseline_bool	0.6500	0.4100	0.2550	0.3167	0.2328	0.1216	0.0694
baseline_BM25	0.8192	0.5200	0.4050	0.4434	0.3623	0.3968	0.2000

Our user case can be seen as a revisit of the well-known cluster hypothesis: *closely associated documents tend to be relevant to the same requests* [28]. For each of the long queries, our system retrieves a cluster of documents whose content is close. This collection of documents exhibits a broad range of topics, with several prominent themes emerging from the analysis. Some of these themes are highly salient and widely discussed within the corpus, while others remain relatively understudied.

Table 8
Topics and Item Sets with Corresponding Coverage.

Topic	Item set Q1	Cov Q1	Item set Q4	Cov Q4
G01	agent & technology & people	1731	attack & command & google	32
G02	interaction & adult & interact	3781	child & behavior & conversation	38
G03	intelligence & artificial & medical	64271	disease & trial & level	193
G04	prediction & feature & genetic	27104	tool & crisp & sequence	38
G05	bias & learn & technology	2244	woman & computer & image	309
G06	research & discuss & responsible	57476	education & ethical & principle	402
G07	analysis & content & online	5536	misinformation & topic & understand	98
G08	blockchain & distribute & application	253	currency & cryptographic & security	36
G09	spatial & model & analysis	23966	system & criminal & visualization	43
G10	increase & human & environment	9815	technology & change & labor	464

6. Conclusion

In this paper, we introduce SimpleText-1, a comprehensive test collection designed to advance research in scientific information access for the general public. This collection provides a large corpus of scientific abstracts, topics and queries derived from popular science articles, and an extensive set of relevance labels (Qrels) developed over three years of the CLEF SimpleText track.

A key contribution of this work is the release of a complete, operational benchmark packaged within a PostgreSQL relational database. This framework integrates modern dense vector search alongside traditional textual search and structured SQL querying, offering a practical environment for developing and evaluating hybrid retrieval systems, such as those used in Retrieval-Augmented Generation (RAG), directly within an enterprise-grade database.

Furthermore, the collection's core corpus, containing millions of scientific abstracts published before 2020, represents a significant asset for the research community. It serves as a large-scale, historical benchmark of technical language, notably free from the influence of modern generative LLMs. This provides an uncontaminated playground for authentically experimenting with and evaluating neural approaches on textual content.

Our experiments on this collection show that dense retrieval methods can significantly outperform strong sparse retrieval baselines like BM25. However, their effectiveness heavily depends on the chosen neural model and how content is represented (e.g., titles versus abstracts). This highlights the necessity of managing multiple dense models alongside documents, a task for which the proposed extended SQL

architecture is particularly well suited. By combining relevance labels with textual complexity metrics, this resource paves the way for future research into user-centric information retrieval systems that consider not only the usefulness of information but also its comprehensibility.

Finally, to better analyse the thematic scope of the corpus in relation to specific queries, we implemented a hybrid approach that combines "fuzzy" similarity-based retrieval with "crisp" Boolean search. The process begins with an initial fuzzy search using k-Nearest Neighbors on dense vector models to retrieve a relevant set of documents. A Latent Dirichlet Allocation (LDA) model is then applied to this retrieved subset to discover underlying topics, which are represented as frequent itemsets of co-occurring words [29]. These itemsets are subsequently used to construct crisp, conjunctive Boolean queries. By leveraging the database's efficient GIN indexes, these queries are executed against the entire corpus to compute the absolute frequency of each itemset in real-time. This methodology allows for a quantitative evaluation of the topic coverage of the corpus for a given query, providing clear insights into how well a specific sub-topic is represented within the whole collection.

Acknowledgments

This track would not have been possible without the great support of numerous individuals. We want to thank in particular the colleagues and the students who participated in data construction and evaluation. Please visit the SimpleText website for more details on the track.[16] *We also thank the MaDICS research group.*[17]

Declaration on Generative AI

During the preparation of this work, the authors used Llama 3.1 to help them rephrase a few sentences to improve clarity, conciseness, or style. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] E. SanJuan, S. Huet, J. Kamps, L. Ermakova, Overview of the CLEF 2022 SimpleText Task 1: Passage Selection for a Simplified Summary, in: G. Faggioli, N. Ferro, A. Hanbury, M. Potthast (Eds.), Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th - to - 8th, 2022, volume 3180 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022, pp. 2762–2772. URL: https://ceur-ws.org/Vol-3180/paper-235.pdf.

[2] E. SanJuan, S. Huet, J. Kamps, L. Ermakova, Overview of the CLEF 2023 SimpleText Task 1: Passage Selection for a Simplified Summary, in: M. Aliannejadi, G. Faggioli, N. Ferro, M. Vlachos (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023), Thessaloniki, Greece, September 18th to 21st, 2023, volume 3497 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023, pp. 2823–2834. URL: https://ceur-ws.org/Vol-3497/paper-238.pdf.

[3] E. SanJuan, S. Huet, J. Kamps, L. Ermakova, Overview of the CLEF 2024 SimpleText Task 1: Retrieve Passages to Include in a Simplified Summary, in: G. Faggioli, N. Ferro, P. Galuscáková, A. G. S. d. Herrera (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), Grenoble, France, 9-12 September, 2024, volume 3740 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2024, pp. 3115–3128. URL: https://ceur-ws.org/Vol-3740/paper-305.pdf.

[4] S. Althammer, A. Askari, S. Verberne, A. Hanbury, DoSSIER@COLIEE 2021: Leveraging dense retrieval and summarization-based re-ranking for case law retrieval, 2021. URL: http://arxiv.org/abs/2108.03937. doi:10.48550/arXiv.2108.03937, arXiv:2108.03937 [cs].

[16] https://simpletext-project.com/

[17] https://www.madics.fr/ateliers/simpletext/

[5] E. Garfield, " science citation index"—a new dimension in indexing: This unique approach underlies versatile bibliographic systems for communicating and evaluating information., Science 144 (1964) 649–654.
[6] Y. Guo, W. Qiu, Y. Wang, T. Cohen, Automated lay language summarization of biomedical scientific reviews, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, 2021, pp. 160–168.
[7] D. A. Scheufele, N. M. Krause, Science audiences, misinformation, and fake news, Proceedings of the National Academy of Sciences 116 (2019) 7662–7669.
[8] R. Goebel, Y. Kano, M.-Y. Kim, J. Rabelo, K. Satoh, M. Yoshioka, Summary of the Competition on Legal Information, Extraction/Entailment (COLIEE) 2023, in: Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law, ICAIL '23, Association for Computing Machinery, New York, NY, USA, 2023, pp. 472–480. URL: https://dl.acm.org/doi/10.1145/3594536.3595176. doi:10.1145/3594536.3595176.
[9] M.-Q. Bui, D.-T. Do, N.-K. Le, D.-H. Nguyen, K.-V.-H. Nguyen, T. P. N. Anh, M. Le Nguyen, Data Augmentation and Large Language Model for Legal Case Retrieval and Entailment, The Review of Socionetwork Strategies (2024). URL: https://doi.org/10.1007/s12626-024-00158-2. doi:10.1007/s12626-024-00158-2.
[10] S. Wehnert, V. Sudhi, S. Dureja, L. Kutty, S. Shahania, E. W. De Luca, Legal norm retrieval with variations of the bert model combined with TF-IDF vectorization, in: Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law, ICAIL '21, Association for Computing Machinery, New York, NY, USA, 2021, pp. 285–294. URL: https://dl.acm.org/doi/10.1145/3462757.3466104. doi:10.1145/3462757.3466104.
[11] P. Bhattacharya, S. Poddar, K. Rudra, K. Ghosh, S. Ghosh, Incorporating domain knowledge for extractive summarization of legal case documents, in: Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law, ACM, São Paulo Brazil, 2021, pp. 22–31. URL: https://dl.acm.org/doi/10.1145/3462757.3466092. doi:10.1145/3462757.3466092.
[12] M. Elaraby, Y. Zhong, D. Litman, Towards Argument-Aware Abstractive Summarization of Long Legal Opinions with Summary Reranking, 2023. URL: http://arxiv.org/abs/2306.00672, arXiv:2306.00672 [cs].
[13] A. Shukla, P. Bhattacharya, S. Poddar, R. Mukherjee, K. Ghosh, P. Goyal, S. Ghosh, Legal Case Document Summarization: Extractive and Abstractive Methods and their Evaluation, 2022. URL: http://arxiv.org/abs/2210.07544, arXiv:2210.07544 [cs].
[14] J.-S. Lee, LexGPT 0.1: pre-trained GPT-J models with Pile of Law, 2023. URL: http://arxiv.org/abs/2306.05431. doi:10.48550/arXiv.2306.05431, arXiv:2306.05431 [cs].
[15] D. Poshyvanyk, M. Gethers, A. Marcus, Concept location using formal concept analysis and information retrieval, ACM Transactions on Software Engineering and Methodology 21 (2013) 23:1–23:34. URL: https://dl.acm.org/doi/10.1145/2377656.2377660. doi:10.1145/2377656.2377660.
[16] J. Tang, A. C. Fong, B. Wang, J. Zhang, A Unified Probabilistic Framework for Name Disambiguation in Digital Library, IEEE Transactions on Knowledge and Data Engineering 24 (2012) 975–987. URL: http://ieeexplore.ieee.org/document/5680902/. doi:10.1109/TKDE.2011.13.
[17] L. Ermakova, E. SanJuan, J. Kamps, S. Huet, I. Ovchinnikova, D. Nurbakova, S. Araújo, R. Hannachi, Mathurin, P. Bellot, Overview of the CLEF 2022 SimpleText Lab: Automatic Simplification of Scientific Texts, in: A. Barrón-Cedeño, G. D. S. Martino, M. D. Esposti, F. Sebastiani, C. Macdonald, G. Pasi, A. Hanbury, M. Potthast, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction - 13th International Conference of the CLEF Association, CLEF 2022, Bologna, Italy, September 5-8, 2022, Proceedings, volume 13390 of *Lecture Notes in Computer Science*, Springer, 2022, pp. 470–494. URL: https://doi.org/10.1007/978-3-031-13643-6_28. doi:10.1007/978-3-031-13643-6_28.
[18] C. Lioma, J. G. Simonsen, B. Larsen, Evaluation measures for relevance and credibility in ranked lists, in: Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval, ICTIR '17, Association for Computing Machinery, New York, NY, USA, 2017, p. 91–98. URL: https://doi.org/10.1145/3121050.3121072. doi:10.1145/3121050.3121072.

[19] T. Tanprasert, D. Kauchak, Flesch-kincaid is not a text simplification evaluation metric, in: Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021), 2021, pp. 1–14.

[20] L. Ermakova, I. Ovchinnikova, J. Kamps, D. Nurbakova, S. Araújo, R. Hannachi, Overview of the CLEF 2022 simpletext task 2: Complexity spotting in scientific abstracts, in: G. Faggioli, N. Ferro, A. Hanbury, M. Potthast (Eds.), Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th - to - 8th, 2022, volume 3180 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022, pp. 2773–2791. URL: https://ceur-ws.org/Vol-3180/paper-236.pdf.

[21] L. Ermakova, E. SanJuan, S. Huet, O. Augereau, H. Azarbonyad, J. Kamps, CLEF 2023 SimpleText Track - What Happens if General Users Search Scientific Texts?, in: J. Kamps, L. Goeuriot, F. Crestani, M. Maistro, H. Joho, B. Davis, C. Gurrin, U. Kruschwitz, A. Caputo (Eds.), Advances in Information Retrieval - 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2-6, 2023, Proceedings, Part III, volume 13982 of *Lecture Notes in Computer Science*, Springer, 2023, pp. 536–545. URL: https://doi.org/10.1007/978-3-031-28241-6_62. doi:10.1007/978-3-031-28241-6_62.

[22] L. Ermakova, H. Azarbonyad, S. Bertin, O. Augereau, Overview of the CLEF 2023 simpletext task 2: Difficult concept identification and explanation, in: M. Aliannejadi, G. Faggioli, N. Ferro, M. Vlachos (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023), Thessaloniki, Greece, September 18th to 21st, 2023, volume 3497 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023, pp. 2835–2854. URL: https://ceur-ws.org/Vol-3497/paper-239.pdf.

[23] R. Nogueira, K. Cho, Passage Re-ranking with BERT, 2020. URL: http://arxiv.org/abs/1901.04085. doi:10.48550/arXiv.1901.04085, arXiv:1901.04085 [cs].

[24] X. Ma, R. Pradeep, R. Nogueira, J. Lin, Document Expansion Baselines and Learned Sparse Lexical Representations for MS MARCO V1 and V2, in: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, Madrid Spain, 2022, pp. 3187–3197. URL: https://dl.acm.org/doi/10.1145/3477495.3531749. doi:10.1145/3477495.3531749.

[25] C. Lassance, H. Dejean, S. Clinchant, An Experimental Study on Pretraining Transformers from Scratch for IR, in: J. Kamps, L. Goeuriot, F. Crestani, M. Maistro, H. Joho, B. Davis, C. Gurrin, U. Kruschwitz, A. Caputo (Eds.), Advances in Information Retrieval, Springer Nature Switzerland, Cham, 2023, pp. 504–520. doi:10.1007/978-3-031-28244-7_32.

[26] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent Dirichlet Allocation, Journal of Machine Learning Research 3 (2003) 993–1022. URL: https://jmlr.csail.mit.edu/papers/v3/blei03a.html.

[27] R. Deveaud, E. SanJuan-Ibekwe, P. Bellot, Accurate and effective latent concept modeling for ad hoc information retrieval, Document Numérique 17 (2014) 61–84. URL: https://doi.org/10.3166/dn.17.1.61-84. doi:10.3166/DN.17.1.61-84.

[28] C. Van Rijsbergen, Information Retrieval, Butterworths, 1979.

[29] J. Poelmans, D. I. Ignatov, S. O. Kuznetsov, G. Dedene, Formal concept analysis in knowledge processing: A survey on applications, Expert Systems with Applications 40 (2013) 6538–6560. URL: https://www.sciencedirect.com/science/article/pii/S0957417413002959. doi:10.1016/j.eswa.2013.05.009.

A. SQL function to compute nearest references based on title embeddings

```
CREATE OR REPLACE FUNCTION
    public.dblp_knn_title(
        v_query vector,
        nb numeric,
        sc float)
    RETURNS TABLE(id bigint)
LANGUAGE plpgsql
    AS $$
       BEGIN
        SET LOCAL ivfflat.probes = 65;
        SET LOCAL enable_seqscan = off;
        SET LOCAL min_parallel_table_scan_size = 1;
        SET LOCAL parallel_setup_cost = 1;
        RETURN QUERY
        SELECT P.id FROM (
               SELECT J.id , (J.title_v <#> v_query) AS ip
               FROM dblp_v AS J
               ORDER BY ip LIMIT nb
               ) AS P
        WHERE P.ip <= sc;
       END;
       $$;
```

B. SQL procedure for Q-rel extension based on title embeddings

```
    CREATE VIEW st1_qrels_train_knn AS
    (
     SELECT
     topic,
     query,
     dblp_knn_title(qdblp_v_title(doc),10,-0.8) AS doc,
     rel,
     origin
     FROM  st1_qrels_train WHERE rel=2
    )
    UNION
    (
     SELECT * FROM st1_qrels_train WHERE rel<2
    )
```