



Anserini Gets Dense Retrieval: Integration of Lucene's HNSW Indexes

Xueguang Ma
University of Waterloo
Waterloo, Canada
x93ma@uwaterloo.ca

Tommaso Teofili
Roma Tre University
Rome, Italy
tommaso.teofili@uniroma3.it

Jimmy Lin
University of Waterloo
Waterloo, Canada
jimmylin@uwaterloo.ca

ABSTRACT

Anserini is a Lucene-based toolkit for reproducible information retrieval research in Java that has been gaining traction in the community. It provides retrieval capabilities for both “traditional” bag-of-words retrieval models such as BM25 as well as retrieval using learned sparse representations such as SPLADE. With Pyserini, which provides a Python interface to Anserini, users gain access to *both* sparse and dense retrieval models, as Pyserini implements bindings to the Faiss vector search library alongside Lucene inverted indexes in a uniform, consistent interface. Nevertheless, hybrid fusion techniques that integrate sparse and dense retrieval models need to stitch together results from two completely different “software stacks”, which creates unnecessary complexities and inefficiencies. However, the introduction of HNSW indexes for dense vector search in Lucene promises the integration of both dense and sparse retrieval within a single software framework. We explore exactly this integration in the context of Anserini. Experiments on the MS MARCO passage and BEIR datasets show that our Anserini HNSW integration supports (reasonably) effective and (reasonably) efficient approximate nearest neighbor search for dense retrieval models, using only Lucene.

CCS CONCEPTS

• Information systems → Retrieval efficiency;

KEYWORDS

Neural IR; Dense Retrieval; HNSW indexes; Lucene

ACM Reference Format:

Xueguang Ma, Tommaso Teofili, and Jimmy Lin. 2023. Anserini Gets Dense Retrieval: Integration of Lucene's HNSW Indexes. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM '23)*, October 21–25, 2023, Birmingham, United Kingdom. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/3583780.3615112>

1 INTRODUCTION

Of practical concern, dense retrieval models based on the bi-encoder architecture require an entirely different “software stack” for efficient top-*k* retrieval [10, 11] compared to sparse (e.g., bag-of-words)

models. These methods take advantage transformers [21] to encode both queries and documents into dense representation vectors [6, 11, 13, 22], where ranking is framed as a nearest neighbor search problem in vector space. Today, efficient execution in an operational setting is usually performed using software libraries such as Faiss [10], which implements hierarchical navigable small-world network (HNSW) indexes [17]. This query evaluation infrastructure is quite distinct and largely incompatible with sparse retrieval using inverted indexes, which has been the mainstay of search systems dating back many decades [19].

Despite the effectiveness of dense retrieval models, we do not anticipate inverted indexes becoming obsolete in the foreseeable future. Thus, achieving state-of-the-art results today requires maintaining two separate “software stacks”. For dense retrieval, the Faiss library [10] has emerged as the most popular solution, and for sparse retrieval, the Lucene search library (which underlies popular platforms such as Elasticsearch) remains the go-to solution. While there exist integrated systems that manage both sparse and dense retrieval models such as Vespa, hybrid fusion combining Faiss and Lucene results via custom implementations remains a common architectural pattern [13, 16]. In real-world applications, such a design creates non-trivial engineering challenges, for example, keeping disparate indexes in sync and scheduling different types of queries. While not insurmountable, these issues add complexity to deployed systems and friction during system development.

A new feature in the latest major release of Lucene (version 9) is the implementation of HNSW indexes [17]. This means that Lucene now provides support for *both* inverted and HNSW indexes. However, these indexes remain under-explored using formal text retrieval benchmarks such as MS MARCO passage [1]. How do effectiveness and efficiency of Lucene's HNSW implementation compare to that of Faiss? This question remains unanswered.

To aid in this exploration, we integrate Lucene's HNSW indexes into the open-source Anserini IR toolkit [23] alongside existing inverted indexes. With this addition, Anserini provides a unified interface for dense and sparse retrieval. We compare this “experience” with the previous approach that requires Pyserini, a Python interface to Anserini that splits dense and sparse retrieval across Faiss and Lucene, respectively [12]. We provide empirical evaluations in terms of effectiveness and various aspects of efficiency—query throughput, indexing throughput, and index size—on modern retrieval test collections, MS MARCO passage [1] and BEIR [20]. We view this work as having three contributions:

- We provide integration of Lucene's HNSW indexing and search capabilities in Anserini, which now offers a “one-stop shop” for reproducible research on common IR test collections for dense as well as sparse representations.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

CIKM '23, October 21–25, 2023, Birmingham, United Kingdom.

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 979-8-4007-0124-5/23/10...\$15.00

<https://doi.org/10.1145/3583780.3615112>

- Due to an implementation choice in Lucene, proper use of HNSW indexes requires training new models that use cosine similarity as the similarity metric (instead of the more common inner product). We have trained such a dense retrieval model, which we share on the Huggingface Model Hub.¹
- With new capabilities in Anserini and a new dense retrieval model, we provide a fair evaluation of tradeoffs between result quality, time (indexing and query evaluation throughput), and space (index size), comparing Lucene vs. Faiss.

At a high level, this work brings two practical benefits. For the researcher: integration of Lucene HNSW indexes into Anserini simplifies the “research toolchain” and streamlines experimental development. For the practitioner: experimental results in this paper provide a fair comparison between the HNSW implementations in Lucene and Faiss, which provides concrete guidance on the time/space/quality tradeoffs that are important in deployed production systems.

2 BACKGROUND AND RELATED WORK

Inverted indexes have been the “go-to” solution for top- k retrieval based on “bag-of-words” models such as BM25 for many decades [2, 19, 23]. However, one of the limitations of traditional inverted indexes is that they are unable to capture semantic relationships. Recently, dense retrieval models based on vector representations generated by transformers [21] have emerged as a popular alternative [11, 13]. In contrast to document vectors that are generated by traditional bag-of-words models, dense representations aim to capture semantic information contained in texts.

While dense retrieval models have proven to be effective, they are not as efficient as traditional inverted indexes when it comes to fast exact match algorithms [7, 14]. Approximate nearest neighbor methods offer a way to speed up dense retrieval models [10] while retaining their effectiveness. One of the most popular algorithms for efficient nearest neighbor search is the hierarchical navigable small-world (HNSW) graph [17]. An HNSW index comprises a multi-layered graph where each layer holds a proximity graph for a subset of the stored vectors. During the search process, the algorithm starts from the top layer and descends through the graph layers to locate the approximate k -nearest neighbors; HNSW indexes can accomplish this with a high degree of accuracy, while also being relatively efficient and scalable.

Despite the popularity of dense retrieval models, lexical retrieval is still necessary in practice, we argue. This claim is based on at least two observations: (1) First, dense retrieval models still exhibit issues such as out-of-domain generalization [20]. (2) Sparse learned representations demonstrate competitive effectiveness compared to dense retrieval models [5]. Generally, results from the literature consistently report that hybrid retrieval architectures—combining dense and sparse results using fusion techniques such as linear score weighting or reciprocal rank fusion—are more effective than their individual constituents [15].

Despite the existence of several dense retrieval libraries such as Faiss [10], there is still a need for a comprehensive *single* framework that can facilitate both dense and sparse retrieval. Recently, Lucene accomplished this, and we have added bindings to Lucene’s HNSW

implementation in Anserini. This integration bridges the software deployment gap between dense and sparse retrieval models, providing a (potentially) unified solution for researchers and practitioners, perhaps enabling more extensive collaborations between academia and industry [4].

3 IMPLEMENTATION

In this section, we describe the integration of HNSW indexing and search in Anserini. Following the same design pattern (exposed via `IndexCollection` and `SearchCollection` classes) for lexical retrieval based on inverted indexes, we introduce the class `IndexHnswDenseVectors` for indexing and the corresponding class `SearchHnswDenseVectors` for search.

For indexing, we implemented the `IndexHnswDenseVectors` class, which takes a collection of dense vectors (in `jsonl` format) and creates an HNSW index using Lucene. The index is constructed in a segmented fashion, where each segment represents a subset of the input collection; orchestration of the segments is handled by Lucene behind the scenes. Same as the description in the original HNSW paper [17], our indexing implementation exposes (i) the `efConstruction` parameter for selecting the number of nearest neighbor candidates to track while searching the graph for each newly inserted node and (ii) the `M` parameter to control the number of bi-directional links created for a new node during index construction. The `IndexHnswDenseVectors` class provides options for multi-threaded indexing, where the number of threads is specified by the user. Additionally, Anserini exposes an `optimize` option to control whether all index segments should be merged into a single segment once all documents have been ingested. Changing the parameters has the following effects:

- `efConstruction`: Increasing this value enhances retrieval effectiveness at the cost of longer indexing times.
- `M`: Increasing this value enhances retrieval effectiveness at the cost of longer indexing times and bigger indexes.
- `threads`: Increasing this value speeds up indexing but consumes more resources.
- `optimize`: This option will yield faster search at the cost of additional time to merge the index segments.

Our implementation provides flexibility to explore the tradeoffs of different settings. For brevity, we do not provide a concrete invocation here, which can be found in Anserini documentation.

For retrieval, we provide the `SearchHnswDenseVectors` class, which takes query vectors and searches the HNSW index for the top- k most similar vectors. The `efSearch` parameter is exposed to control the number of candidate neighbors that are explored during search time. Increasing `efSearch` increases effectiveness but increases query latency as well. Similar to indexing, there are effectiveness–efficiency tradeoffs for search. The search class also provides options for multi-threaded searching; the current implementation provides inter-query parallelism, where multiple queries are processed in parallel. Overall, with this class, users have the ability to perform search under different performance tradeoff scenarios. Again, we refer readers to Anserini documentation for an example of a concrete invocation. Once retrieval finishes, the results are stored in a text file that can be evaluated using standard tools such as `trec_eval`.

¹<https://huggingface.co/castorini/cosdpr-distil/>

4 A NEW DENSE RETRIEVAL MODEL

Dense retrieval models for the MS MARCO passage dataset, such as ANCE [22], TAS-B [9], and TCT_ColBERT [13], are typically trained using dot product similarity. However, the Lucene HNSW implementation requires that the dense representations have an L2 norm of one, which means that we need a model that is trained to generate vector representations for cosine similarity. In this work, we use “best practices” such as distillation from a cross encoder [18] and hard negative mining [6, 18, 22] to train a dense retrieval model that is effective for cosine similarity and adaptable to our Anserini HNSW implementation. We call this model cosDPR-distil.

For our model, we first train a dense retriever using the MS MARCO passage training data with BM25 hard negatives. We use the model to mine hard negative passages for queries in the training set and combine the hard negative passages with the original BM25 hard negatives as augmented training data. Then, we use the augmented training data to train a cross-encoder reranker. We finally use the reranker to teach a dense retriever student, which is the final cosDPR-distil model, with knowledge distillation over the augmented training data.

We used Tevatron [8] to train our model following the process outlined above on a server equipped with $8 \times$ A6000 GPUs. Specifically, we initialized the cross-encoder teacher with the Electra-large model [3] and trained it using a batch size of 16×16 (each batch contains 16 queries and each query contains 1 positive passage and 15 negative passages) for 6 epochs. For the dense retriever student, we initialized the model with Co-Condenser [6] and trained it using a batch size of 16×16 for 10 epochs. During the process of hard negative mining, we randomly selected 30 documents for each query from the top 200 retrieval results of the dense retriever that were not labeled as relevant.

5 EXPERIMENT SETUP

We evaluated our Anserini HNSW integration on the MS MARCO passage and BEIR datasets.

- MS MARCO passage is a web search dataset that consists of a training set with approximately 500k relevant query–passage pairs and a corpus with 8.8 million passages. For evaluation, we used the development set, which includes 6,980 queries. Our main objective is to compare the efficiency and effectiveness of the HNSW implementation in Lucene and Faiss. Effectiveness is measured by two metrics, MRR@10 and Recall@1k, while efficiency is measured in terms of QPS (queries per second).
- BEIR is a collection of datasets specifically created for zero-shot text retrieval evaluation. This collection consists of 18 datasets covering a wide range of domains. We used BEIR to analyze out-of-domain effectiveness generalization. We ran our experiments on the 13 publicly accessible datasets.

To evaluate the time required for index construction and search efficiency for both Lucene and Faiss HNSW indexes, we used a Google Cloud VM with the e2-standard-16 configuration, which contains 16 vCPU cores, 64 GB RAM and 1TB SSD. Experiments were performed with Lucene 9.4.2. All experiments with Faiss were performed in Pyserini. In all cases, we used pre-encoded queries and documents—that is, our timings do *not* include neural inference.

	MRR@10	Recall@1k
BM25	0.184	0.853
ANCE	0.330	0.959
TAS-B	0.344	0.977
CoCondenser	0.382	0.984
cosDPR-distil (Flat)	0.390	0.980
cosDPR-distil (HNSW)	0.389	0.974

Table 1: Comparing our cosDPR-distil model with existing retrieval models on the MS MARCO passage ranking dev set.

	BM25	DPR	ANCE	TASB	cosDPR-distil	
					Flat	HNSW
TREC-COVID	0.656	0.332	0.654	0.481	0.497	0.497
NFCorpus	0.325	0.189	0.237	0.319	0.319	0.319
NQ	0.329	0.474	0.446	0.463	0.468	0.467
HotpotQA	0.603	0.391	0.456	0.584	0.517	0.510
FiQA-2018	0.236	0.112	0.295	0.300	0.238	0.237
ArguAna	0.414	0.175	0.415	0.429	0.472	0.472
Touche-2020 (v2)	0.367	0.127	0.240	0.162	0.180	0.177
Quora	0.789	0.248	0.852	0.835	0.859	0.858
DBpedia	0.313	0.263	0.281	0.384	0.375	0.371
SCIDOCs	0.158	0.077	0.122	0.149	0.143	0.143
FEVER	0.753	0.562	0.669	0.700	0.591	0.580
Climate-FEVER	0.213	0.148	0.198	0.228	0.196	0.197
SciFact	0.665	0.318	0.507	0.643	0.561	0.561
Avg.nDCG@10	0.448	0.263	0.413	0.416	0.417	0.415

Table 2: Comparing our cosDPR-distil model with existing retrieval models on BEIR.

6 RESULTS

Evaluation of model effectiveness. In Tables 1 and 2, we show the effectiveness of our cosDPR-distil model compared to existing representative dense retrieval models on the MS MARCO passage and BEIR datasets. For our model, we provide results using both “flat” indexes (i.e., brute-force comparisons) and HNSW indexes. The HNSW index here was built with $M=16$, $efC=100$, using a single thread with optimization. Results for the other models are with “flat” indexes.

For the MS MARCO passage dataset (in-domain), our model demonstrates competitive effectiveness compared to other dense retrieval baselines. Under the zero-shot setting on BEIR (out-of-domain), our model is on par with other dense retrieval models. Overall, these experiments verify that our cosDPR-distil model is at least on par with existing reference dense retrieval models. Comparing Anserini HNSW results with brute-force exact nearest neighbor search, we see that it is able to retain a high level of effectiveness. We explore additional HNSW settings below.

Evaluation of indexing and search. Table 3 presents the results of a comparison between the Anserini (Lucene) and Faiss implementations of the HNSW indexing algorithm. Performance was measured by varying the efC graph construction parameter, the number of threads used to build the index, and whether or not we optimize the Lucene index through segment merging. Here, we fixed the search parameter $efSearch=1000$ for all the index variants, focusing on the tradeoffs induced by the indexing parameters.

	Index	efC	Threads	Optimize	Time (h)	Size (GB)	MRR@10	Recall@1k	QPS
(1)	Anserini	100	1	no	13.6	27	0.3887	0.9737	3.6
(2)	Anserini	100	16	no	2.9	27	0.3890	0.9743	1.5
(3)	Anserini	1000	16	no	19.6	27	0.3894	0.9784	1.5
(4)	Anserini	100	1	yes	21.6	27	0.3842	0.9600	35.8
(5)	Anserini	100	16	yes	10.7	27	0.3795	0.9546	31.1
(6)	Anserini	1000	16	yes	96.5	27	0.3869	0.9737	23.6
(7)	Faiss FlatIP	-	-	-	-	26	0.3896	0.9796	0.4
(8)	Faiss HNSW	100	1	-	3.8	27	0.3806	0.9522	60.5
(9)	Faiss HNSW	100	16	-	0.3	27	0.3809	0.9584	60.0
(10)	Faiss HNSW	1000	16	-	3.3	27	0.3866	0.9730	44.3

Table 3: HNSW search with different indexing configurations. We set M=16 for indexing and efSearch=1000 for search.

Row	Index	efSearch	threads	MRR@10	QPS
(1)	Anserini (HNSW)	10	1	0.2551	587.8
(2)	Anserini (HNSW)	100	1	0.3612	234.0
(3)	Anserini (HNSW)	1000	1	0.3842	33.0
(4)	Anserini (HNSW)	1000	16	0.3842	373.0
(5)	Faiss (HNSW)	10	1	0.1906	1733.8
(6)	Faiss (HNSW)	100	1	0.3539	522.4
(7)	Faiss (HNSW)	1000	1	0.3806	60.5
(8)	Faiss (HNSW)	1000	16	0.3806	164.5
(9)	Faiss (FlatIP)	-	1	0.3896	0.4
(10)	Faiss (FlatIP)	-	16	0.3896	24.8

Table 4: HNSW search with different search configurations; in these experiments, the index was constructed with M=16, efC=100, single thread, optimized.

We first focus on the Anserini (Lucene) implementation. Overall, compared to Faiss FlatIP (brute-force nearest neighbor search), we see that the HNSW implementation successfully retains retrieval effectiveness while improving retrieval efficiency. Additional observations:

- Comparing rows (1) and (2) or (4) and (5), where the only difference is the number of threads used during indexing: we observe a significant reduction in indexing time, while achieving similar effectiveness in terms of MRR@10 and Recall@1k.
- Comparing rows (1) and (4) or (2) and (5), where the only difference is the optimization of the index through segment merging: although optimization increases indexing time, segment merging brings significant improvements to query latency.
- Comparing rows (2) and (3) or (5) and (6), where the efC parameter is increased from 100 to 1000: we see a slight improvement in effectiveness, but at the cost of significantly longer indexing times. This comparison highlights the tradeoffs between indexing time and retrieval effectiveness.

When comparing the Anserini (Lucene) HNSW implementation with the Faiss implementation, the retrieval effectiveness achieved by both methods appears to be similar. However, the indexing speed of Faiss is much faster than Anserini.

Further explorations of search. In Table 4, we further explore how the search parameters influence performance tradeoffs. For index construction, we used M=16, efC=100, single thread, optimized. Here, we focus on varying two key parameters: the efSearch parameter and the number of threads used during the search.

For the Anserini (Lucene) implementation, comparing rows (1) to (3), we see that as the efSearch parameter increases, the MRR@10 metric also increases (slightly), reaching a peak of 0.3842 at efSearch=1000, which is very close to exact nearest neighbor search. However, QPS decreases significantly as the efSearch parameter increases, as expected. Faiss exhibits the same behavior, shown in rows (5) to (7).

Compared to the HNSW implementation in Faiss, the Lucene implementation appears to be slower with a single query thread. This can be seen in row (3) vs. row (7). Effectiveness is comparable, but Faiss appears to be roughly twice as fast as Lucene.

For the Anserini (Lucene) implementation, we can see from rows (3) and (4) that increasing the number of threads from 1 to 16 while holding constant efSearch=1000 increases QPS from 33 to 373; multi-threaded search improves effectiveness, as expected. The comparable Faiss results are shown in rows (7) and (8). Interestingly, Faiss does not exhibit anywhere close to the increase in QPS from multi-threaded search that we observe in Lucene.

In summary, it appears that effectiveness is comparable between Lucene and Faiss. Using a single thread, Faiss is substantially faster, but Lucene appears to achieve better multi-threaded performance. We are quick to concede that we have not sufficiently delved in the underlying implementation details to understand the source of these observed performance differences.

7 CONCLUSION

In this work, we explored the integration of Lucene’s HNSW indexes in the open-source Anserini IR toolkit. Our approach provides a unified interface for both dense and sparse retrieval in Java, using only Lucene. Experiments on the MS MARCO passage and BEIR datasets show that our cosDPR-distill model achieves reasonable effectiveness, but comparisons to the popular Faiss library reveal differences in performance that we have yet to fully explore. Nevertheless, Anserini provides an intriguing value proposition: a “single stack”, unified Lucene-based framework for *both* dense and sparse retrieval models, which potentially simplifies the deployment of production search applications that combine the best of both worlds in terms of the hybrid fusion of dense and sparse retrieval models.

ACKNOWLEDGEMENTS

This research was supported in part by the Natural Sciences and Engineering Research Council (NSERC) of Canada.

REFERENCES

- [1] Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2018. MS MARCO: A Human Generated MACHine Reading COMprehension Dataset. *arXiv:1611.09268v3* (2018).
- [2] Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to Answer Open-Domain Questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada, 1870–1879.
- [3] Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. In *Proceedings of the 8th International Conference on Learning Representations (ICLR 2020)*. Addis Ababa, Ethiopia.
- [4] Josh Devins, Julie Tibshirani, and Jimmy Lin. 2022. Aligning the Research and Practice of Building Search Applications: Elasticsearch and Pyserini. In *Proceedings of the 15th ACM International Conference on Web Search and Data Mining (WSDM 2022)*. 1573–1576.
- [5] Thibault Formal, C. Lassance, Benjamin Piwowarski, and Stéphane Clinchant. 2021. SPLADE v2: Sparse Lexical and Expansion Model for Information Retrieval. *arXiv:2109.10086* (2021).
- [6] Luyu Gao and Jamie Callan. 2021. Unsupervised Corpus Aware Language Model Pre-training for Dense Passage Retrieval. *arXiv:2108.05540* (2021).
- [7] Luyu Gao, Zhuyun Dai, and Jamie Callan. 2021. COIL: Revisit Exact Lexical Match in Information Retrieval with Contextualized Inverted List. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online, 3030–3042.
- [8] Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2023. Tevatron: An Efficient and Flexible Toolkit for Neural Retrieval. In *Proceedings of the 46th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2023)*. Taipei, Taiwan, 3120–3124.
- [9] Sebastian Hofstätter, Sheng-Chieh Lin, Jheng-Hong Yang, Jimmy Lin, and Allan Hanbury. 2021. Efficiently Teaching an Effective Dense Retriever with Balanced Topic Aware Sampling. In *Proceedings of the 44th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2021)*. 113–122.
- [10] Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data* 7, 3 (2019), 535–547.
- [11] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online, 6769–6781.
- [12] Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021. Pyserini: A Python Toolkit for Reproducible Information Retrieval Research with Sparse and Dense Representations. In *Proceedings of the 44th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2021)*. 2356–2362.
- [13] Sheng-Chieh Lin, Jheng-Hong Yang, and Jimmy Lin. 2021. In-Batch Negatives for Knowledge Distillation with Tightly-Coupled Teachers for Dense Retrieval. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*. 163–173.
- [14] Xueguang Ma, Minghan Li, Kai Sun, Ji Xin, and Jimmy Lin. 2021. Simple and Effective Unsupervised Redundancy Elimination to Compress Dense Vectors for Passage Retrieval. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Punta Cana, Dominican Republic, 2854–2859.
- [15] Xueguang Ma, Ronak Pradeep, Rodrigo Nogueira, and Jimmy Lin. 2022. Document Expansion Baselines and Learned Sparse Lexical Representations for MS MARCO V1 and V2. *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (2022).
- [16] Xueguang Ma, Kai Sun, Ronak Pradeep, Minghan Li, and Jimmy Lin. 2022. Another Look at DPR: Reproduction of Training and Replication of Retrieval. In *Proceedings of the 44th European Conference on Information Retrieval (ECIR 2022)*, Part I. Stavanger, Norway, 613–626.
- [17] Yu A. Malkov and D. A. Yashunin. 2020. Efficient and Robust Approximate Nearest Neighbor Search Using Hierarchical Navigable Small World Graphs. *Transactions on Pattern Analysis and Machine Intelligence* 42, 4 (2020), 824–836.
- [18] Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Wayne Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. 2021. RocketQA: An Optimized Training Approach to Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online, 5835–5847.
- [19] Stephen E. Robertson and Hugo Zaragoza. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends in Information Retrieval* 3, 4 (2009), 333–389.
- [20] Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- [21] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Advances in Neural Information Processing Systems* 30. Long Beach, California, 5998–6008.
- [22] Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, and Arnold Overwijk. 2021. Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval. In *Proceedings of the 9th International Conference on Learning Representations (ICLR 2021)*.
- [23] Peilin Yang, Hui Fang, and Jimmy Lin. 2017. Anserini: Enabling the Use of Lucene for Information Retrieval Research. In *Proceedings of the 40th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2017)*. Tokyo, Japan, 1253–1256.