Endnotes

1. Rai, A., 2020, Study of various methods for tokenization. In Advances in Natural Language Processing. Available at: https://doi.org/10.1007/978-981-15-6198-6_18

2. Pennington, J., Socher, R. & Manning, C., 2014, GloVe: Global Vectors for Word Representation. [online] Available at: https://nlp.stanford.edu/pubs/glove.pdf.

3. Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q. V. & Hinton, G., 2016, Swivel: Improving embeddings by noticing what's missing. ArXiv, abs/1602.02215. Available at: https://arxiv.org/abs/1602.02215.

4. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. & Dean, J., 2013, Efficient estimation of word representations in vector space. ArXiv, abs/1301.3781. Available at: https://arxiv.org/pdf/1301.3781.pdf.

5. Rehurek, R., 2021, Gensim: open source python library for word and document embeddings. Available at: https://radimrehurek.com/gensim/intro.html.

6. Bojanowski, P., Grave, E., Joulin, A. & Mikolov, T., 2016, Enriching word vectors with subword information. ArXiv, abs/1607.04606. Available at: https://arxiv.org/abs/1607.04606.

7. Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R., 1990, Indexing by latent semantic analysis. Journal of the American Society for Information Science, 41(6), pp. 391-407.

8. Blei, D. M., Ng, A. Y., & Jordan, M. I., 2001, Latent Dirichlet allocation. In T. G. Dietterich, S. Becker, & Z. Ghahramani (Eds.), Advances in Neural Information Processing Systems 14. MIT Press, pp. 601-608. Available at: https://proceedings.neurips.cc/paper/2001/hash/296472c9542ad4d4788d543508116cbc-Abstract.html.

9. Muennighoff, N., Tazi, N., Magne, L., & Reimers, N., 2022, Mteb: Massive text embedding benchmark. ArXiv, abs/2210.07316. Available at: https://arxiv.org/abs/2210.07316.

10. Le, Q. V., Mikolov, T., 2014, Distributed representations of sentences and documents. ArXiv, abs/1405.4053. Available at: https://arxiv.org/abs/1405.4053.

11. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K., 2019, BERT: Pre-training deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 4171-4186. Available at: https://www.aclweb.org/anthology/N19-1423/.

12. Reimers, N. & Gurevych, I., 2020, Making monolingual sentence embeddings multilingual using knowledge distillation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 254-265. Available at: https://www.aclweb.org/anthology/2020.emnlp-main.21/.

13. Gao, T., Yao, X. & Chen, D., 2021, Simcse: Simple contrastive learning of sentence embeddings. ArXiv, abs/2104.08821. Available at: https://arxiv.org/abs/2104.08821.

14. Wang, L., Yang, N., Huang, X., Jiao, B., Yang, L., Jiang, D., Majumder, R. & Wei, F., 2022, Text embeddings by weakly supervised contrastive pre-training. ArXiv. Available at: https://arxiv.org/abs/2201.01279.

15. Khattab, O. & Zaharia, M., 2020, colBERT: Efficient and effective passage search via contextualized late interaction over BERT. In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 39-48. Available at: https://dl.acm.org/doi/10.1145/3397271.3401025.

16. Lee, J., Dai, Z., Duddu, S. M. K., Lei, T., Naim, I., Chang, M. W. & Zhao, V. Y., 2023, Rethinking the role of token retrieval in multi-vector retrieval. ArXiv, abs/2304.01982. Available at: https://arxiv.org/abs/2304.01982.

17. TensorFlow, 2021, TensorFlow hub, a model zoo with several easy to use pre-trained models. Available at: https://tfhub.dev/.

18. Zhang, W., Xiong, C., & Zhao, H., 2023, Introducing BigQuery text embeddings for NLP tasks. Google Cloud Blog. Available at: https://cloud.google.com/blog/products/data-analytics/introducing-bigquery-text-embeddings.

19. Google Cloud, 2024, Get multimodal embeddings. Available at: https://cloud.google.com/vertex-ai/generative-ai/docs/embeddings/get-multimodal-embeddings.

20. Pinecone, 2024, IT Threat Detection. [online] Available at: https://docs.pinecone.io/docs/it-threat-detection.

21. Cai, H., Zheng, V. W., & Chang, K. C., 2020, A survey of algorithms and applications related with graph embedding. In Proceedings of the 29th ACM International Conference on Information & Knowledge Management. Available at: https://dl.acm.org/doi/10.1145/3444370.3444568.

22. Cai, H., Zheng, V. W., & Chang, K. C., 2017, A comprehensive survey of graph embedding: problems, techniques and applications. ArXiv, abs/1709.07604. Available at: https://arxiv.org/pdf/1709.07604.pdf.

23. Hamilton, W. L., Ying, R. & Leskovec, J., 2017, Inductive representation learning on large graphs. In Advances in Neural Information Processing Systems 30. Available at: https://cs.stanford.edu/people/jure/pubs/graphsage -nips17.pdf.

24. Dong, Z., Ni, J., Bikel, D. M., Alfonseca, E., Wang, Y., Qu, C. & Zitouni, I., 2022, Exploring dual encoder architectures for question answering. ArXiv, abs/2204.07120. Available at: https://arxiv.org/abs/2204.07120.

25. Google Cloud, 2021, Vertex AI Generative AI: Tune Embeddings. Available at: https://cloud.google.com/vertex-ai/docs/generative-ai/models/tune-embeddings.

26. Matsui, Y., 2020, Survey on approximate nearest neighbor methods. ACM Computing Surveys (CSUR), 53(6), Article 123. Available at: https://wangzwhu.github.io/home/file/acmmm-t-part3-ann.pdf.

27. Friedman, J. H., Bentley, J. L. & Finkel, R. A., 1977, An algorithm for finding best matches in logarithmic expected time. ACM Transactions on Mathematical Software (TOMS), 3(3), pp. 209-226. Available at: https://dl.acm.org/doi/pdf/10.1145/355744.355745.

28. Scikit-learn, 2021, Scikit-learn, a library for unsupervised and supervised neighbors-based learning methods. Available at: https://scikit-learn.org/.

29. lshashing, 2021, An open source python library to perform locality sensitive hashing. Available at: https://pypi.org/project/lshashing/.

30. Malkov, Y. A., Yashunin, D. A., 2016, Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. ArXiv, abs/1603.09320. Available at: https://arxiv.org/pdf/1603.09320.pdf.

31. Google Research, 2021, A library for fast ANN by Google using the ScaNN algorithm. Available at: https://github.com/google-research/google-research/tree/master/scann.

32. Guo, R., Zhang, L., Hinton, G. & Zoph, B., 2020, Accelerating large-scale inference with anisotropic vector quantization. ArXiv, abs/1908.10396. Available at: https://arxiv.org/pdf/1908.10396.pdf.

33. TensorFlow, 2021, TensorFlow Recommenders, an open source library for building ranking & recommender system models. Available at: https://www.tensorflow.org/recommenders.

34. Google Cloud, 2021, Vertex AI Vector Search, Google Cloud's high-scale low latency vector database. Available at: https://cloud.google.com/vertex-ai/docs/vector-search/overview.

35. Elasticsearch, 2021, Elasticsearch: a RESTful search and analytics engine. Available at: https://www.elastic.co/elasticsearch/.

36. Pinecone, 2021, Pinecone, a commercial fully managed vector database. Available at: https://www.pinecone.io.

37. pgvector, 2021, Open Source vector similarity search for Postgres. Available at: https://github.com/pgvector/pgvector.

38. Weaviate, 2021, Weaviate, an open source vector database. Available at: https://weaviate.io/.

39. ChromaDB, 2021, ChromaDB, an open source vector database. Available at: https://www.trychroma.com/.

40. LangChain, 2021.,LangChain, an open source framework for developing applications powered by language model. Available at: https://langchain.com.

42. Thakur, N., Reimers, N., Ruckl'e, A., Srivastava, A., & Gurevych, I. (2021). BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models. ArXiv, abs/2104.08663.
Available at: https://github.com/beir-cellar/beir

43. Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive Text Embedding Benchmark. In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 2014–2037, Dubrovnik, Croatia. Association for Computational Linguistics.
Available at: https://github.com/embeddings-benchmark/mteb

44. Chris Buckley. trec_eval IR evaluation package. Available from https://github.com/usnistgov/trec_eval

45. Christophe Van Gysel and Maarten de Rijke. 2018. Pytrec_eval: An Extremely Fast Python Interface to trec_eval. In The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR '18). Association for Computing Machinery, New York, NY, USA, 873–876.
Availalbe at: https://doi.org/10.1145/3209978.3210065

46. Boteva, Vera & Gholipour Ghalandari, Demian & Sokolov, Artem & Riezler, Stefan. (2016). A Full-Text Learning to Rank Dataset for Medical Information Retrieval. 9626. 716-722. 10.1007/978-3-319-30671-1_58. Available at https://www.cl.uni-heidelberg.de/statnlpgroup/nfcorpus/

47. Douze, M., Guzhva, A., Deng, C., Johnson, J., Szilvasy, G., Mazaré, P.E., Lomeli, M., Hosseini, L. and Jégou, H., 2024. The Faiss library. arXiv preprint arXiv:2401.08281. Available at https://arxiv.org/abs/2401.08281

48. Lee, J., Dai, Z., Ren, X., Chen, B., Cer, D., Cole, J.R., Hui, K., Boratko, M., Kapadia, R., Ding, W. and Luan, Y., 2024. Gecko: Versatile text embeddings distilled from large language models. arXiv preprint arXiv:2403.20327.
Available at: https://arxiv.org/abs/2403.20327

49. Okapi BM25: a non-binary model" Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze. An Introduction to Information Retrieval, Cambridge University Press, 2009, p. 232.

50. Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res. 21, 1, Article 140 (January 2020), 67 pages.
Available at https://dl.acm.org/doi/abs/10.5555/3455716.3455856

51. Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sashank Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2023. PaLM: scaling language modeling with pathways. J. Mach. Learn. Res. 24, 1, Article 240 (January 2023), 113 pages. Available at https://dl.acm.org/doi/10.5555/3648699.3648939

52. Gemini: A Family of Highly Capable Multimodal Models, Gemini Team, Dec 2023. Available at: https://storage.googleapis.com/deepmind-media/gemini/gemini_1_report.pdf

53. Radford, Alec and Karthik Narasimhan. "Improving Language Understanding by Generative Pre-Training." (2018). Available at: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

54. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F. and Rodriguez, A., 2023. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971. Available at: https://arxiv.org/abs/2302.13971

55. Kusupati, A., Bhatt, G., Rege, A., Wallingford, M., Sinha, A., Ramanujan, V., Howard-Snyder, W., Chen, K., Kakade, S., Jain, P. and Farhadi, A., 2022. Matryoshka representation learning. Advances in Neural Information Processing Systems, 35, pp.30233-30249. Available at: https://proceedings.neurips.cc/paper_files/paper/2022/file/c32319f4868da7613d78af9993100e42-Paper--Conference.pdf

56. Nair, P., Datta, P., Dean, J., Jain, P. and Kusupati, A., 2025. Matryoshka Quantization. arXiv preprint arXiv:2502.06786. Available at: https://arxiv.org/abs/2502.06786

57. Faysse, M., Sibille, H., Wu, T., Omrani, B., Viaud, G., Hudelot, C. and Colombo, P., 2024. Colpali: Efficient document retrieval with vision language models. arXiv preprint arXiv:2407.01449. Available at: https://arxiv.org/abs/2407.01449

58. Aumüller, M., Bernhardsson, E. and Faithfull, A., 2020. ANN-Benchmarks: A benchmarking tool for approximate nearest neighbor algorithms. Information Systems, 87, p.101374.