Notes on Vector Retrieval Methods and Techniques

compiled by D. Gueorguiev, 8/9/2025

# Vectors as Units of Retrieval

We often intend for the vector representation of two similar objects to be “close” to each other according to some well-defined distance function. Thus, similarity in the vector space must imply similarity between objects.

We must choose the dimensionality of the target space (a subset of ) for embedding along with the distance function . The vector space of embeddings together with the distance function defines a metric space.

Consider the lexical representation of a text document where is the size of the English vocabulary. Let be the distance variant of the Jaccard index, , where with denoting the -th coordinate of vector .

In the resulting space, if vectors and have smaller distance than vectors and , then the document represented by is lexically more similar to the one represented by than it is to the document represents.

Argument with respect to semantic embedding of text documents

Consider sparse embeddings with d being the size of the vocabulary, and more concretely, take SPlade

# References

[1] [Foundations of Vector Retrieval, Sebastian Bruch, 2024](https://github.com/dimitarpg13/rag_architectures_and_concepts/blob/main/articles/Foundations_of_Vector_Retrieval_Bruch_2024.pdf)

[2] [Vector database management systems: Fundamental Concepts, use-cases, and current challenges, Toni Taipalus, 2024](https://github.com/dimitarpg13/rag_architectures_and_concepts/blob/main/articles/vector_db/Vector_database_management_systems-Fundamental_concepts_use-acases_and_current_challenges_Taipalus_2024.pdf)

[3] [Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks, Patrick Lewis, Ethan Perez et al, 2021](https://github.com/dimitarpg13/rag_architectures_and_concepts/blob/main/articles/Retrieval-Augmented_Generation_for_Knowledge-Intensive_NLP_Tasks_Lewis_2021.pdf)

[4] [SPLADE: Sparse Lexical and Expansion Model for First Stage Ranking, Thibault Formal et al, 2021](https://github.com/dimitarpg13/vector_db_intro/blob/main/articles/Splade-Sparse_lexical_and_expansion_model_for_first_stage_ranking_Formal_2021.pdf)

# Appendix

## Jaccard index and Jaccard distance

The *Jaccard index* is a statistic used for gauging the similarity and the lack of diversity of sample. It is defined as the ratio of two sizes (areas or volumes), the intersection size divided by the union size also called *intersection over union*.

Thus, the Jaccard index is defined as the size of the intersection divided by the size of the union of the sample sets

Obviously, . If no elements are common, then obviously . If clearly .

The *Jaccard distance* measures dissimilarity between sample sets and is complementary to the Jaccard index. It is obtained as:

.

## The SPLADE sparse embedding model

The article introducing the SPLADE embedding model is [4].

### Sparse Lexical Representations for First-Stage Ranking

#### The SparTerm model