Vector databases

embeddings & indices

Valter Uotila

doctoral researcher

Agenda

Gain familiarity with generating and using vector embeddings with techniques such as Word2Vec, GloVe, BERT, and other deep learning models.

Understand the indexing mechanisms used in vector databases, such as HNSW (Hierarchical Navigable Small World), LSH (Locality-sensitive hashing), and IVFFlat (Inverted File with Flat compression).

Embeddings

vectorization

Word2Vec[1,2]

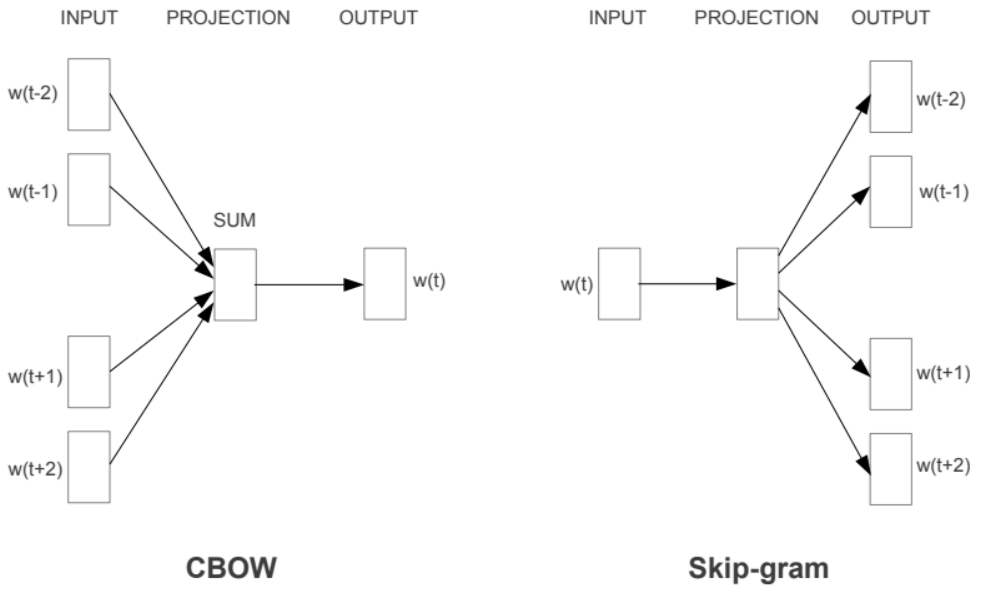


Figure 1: New model architectures. The CBOW architecture predicts the current word based on the context, and the Skip-gram predicts surrounding words given the current word.

Table 1: *Examples of five types of semantic and nine types of syntactic questions in the Semantic-Syntactic Word Relationship test set.*

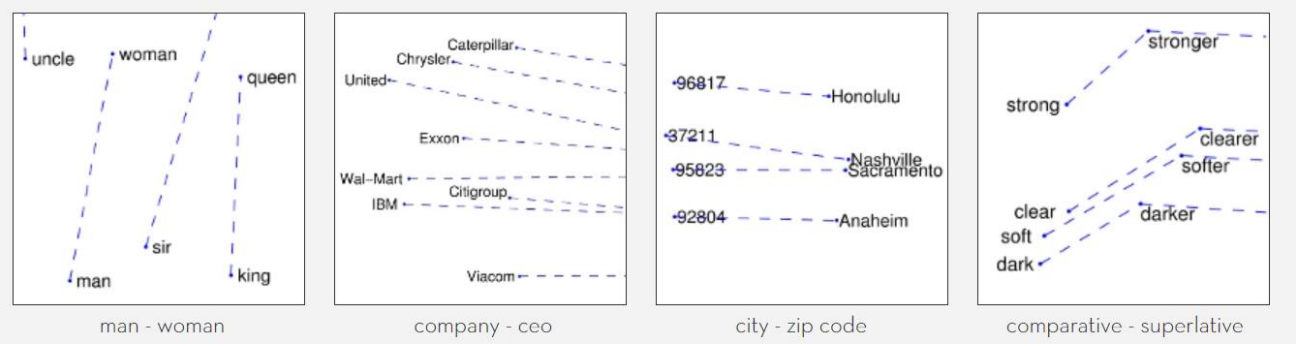
Type of relationship	Word Pair 1		Word Pair 2	
Common capital city	Athens	Greece	Oslo	Norway
All capital cities	Astana	Kazakhstan	Harare	Zimbabwe
Currency	Angola	kwanza	Iran	rial
City-in-state	Chicago	Illinois	Stockton	California
Man-Woman	brother	sister	grandson	granddaughter
Adjective to adverb	apparent	apparently	rapid	rapidly
Opposite	possibly	impossibly	ethical	unethical
Comparative	great	greater	tough	tougher
Superlative	easy	easiest	lucky	luckiest
Present Participle	think	thinking	read	reading
Nationality adjective	Switzerland	Swiss	Cambodia	Cambodian
Past tense	walking	walked	swimming	swam
Plural nouns	mouse	mice	dollar	dollars
Plural verbs	work	works	speak	speaks

[1]Efficient Estimation of Word Representations in Vector Space. Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean. 2013.
[2]Distributed Representations of Sentences and Documents. Quoc Le, Tomas Mikolov. 2014.

GloVe

- GloVe is an unsupervised learning algorithm for obtaining vector representations for words. Training is performed on aggregated global word-word co-occurrence statistics from a corpus, and the resulting representations showcase interesting linear substructures of the word vector space.
- Nearest neighbors
- Linear substructures

source: https://nlp.stanford.edu/projects/glove/

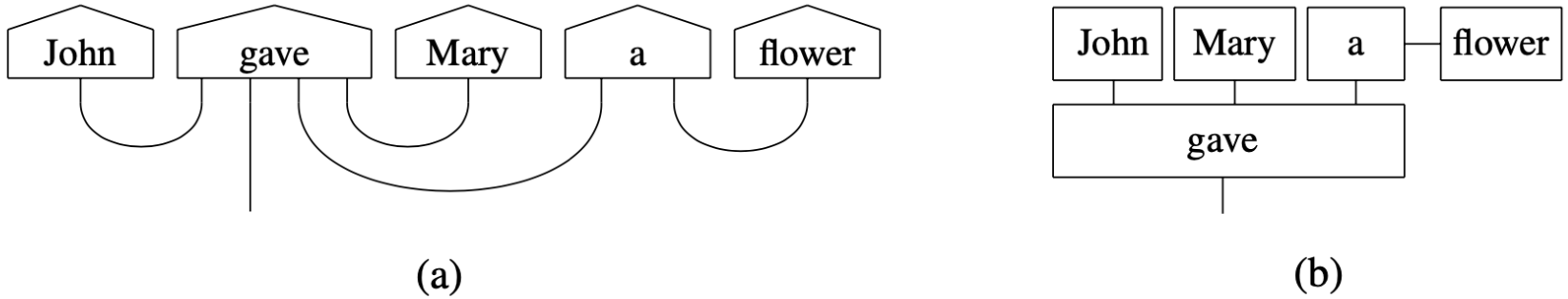


BERT[1,2] (Bidirectional Encoder Representations from Transformers)

- BERT a bidirectional transformer pretrained using a combination of masked language modelling objective and next sentence prediction on a large corpus comprising the Toronto Book Corpus and Wikipedia
- jointly conditioning on both left and right context in all layers
- can be fine-tuned with just one additional output layer

[1]https://huggingface.co/docs/transformers/en/model_doc/bert
[2]BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. 2018.

Other learning models: Vectorization in quantum natural language processing



images: https://cqcl.github.io/lambeq

Indexing in VDBMs

approximate nearest neighbor search

Hierarchical Navigable Small World[1]

- Graph-based approximate nearest neighbor search technique
- Fully graph-based, without any need for additional search structures
- Consists of layers of graphs

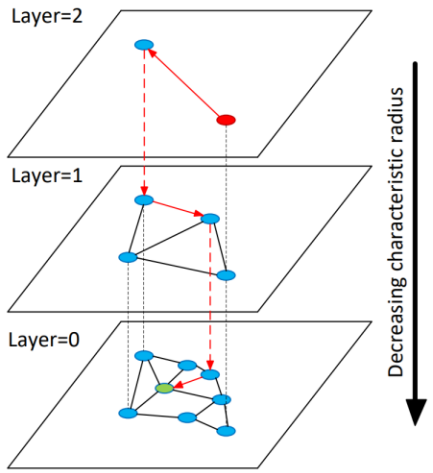


Fig. 1. Illustration of the Hierarchical NSW idea. The search starts from an element from the top layer (shown red). Red arrows show direction of the greedy algorithm from the entry point to the query (shown green).

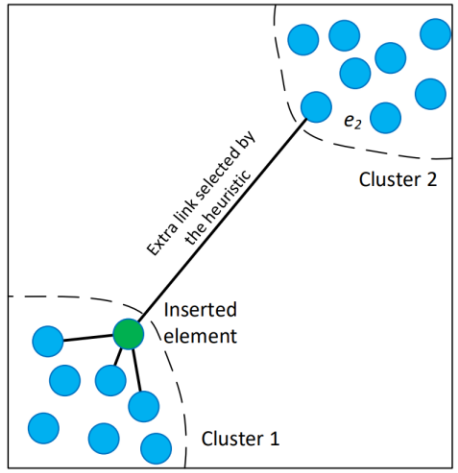


Fig. 2. Illustration of the heuristic used to select the graph neighbors for two isolated clusters. A new element is inserted on the boundary of Cluster 1. All of the closest neighbors of the element belong to the Cluster 1, thus missing the edges of Delaunay graph between the clusters. The heuristic, however, selects element e_2 from Cluster 2, thus, maintaining the global connectivity in case the inserted element is the closest to e_2 compared to any other element from Cluster 1.

[1]Efficient and robust approximate nearest neighbor search using Hierarchical Navigable Small World graphs. Yu. A. Malkov, D. A. Yashunin. 2020.

Locality-sensitive hashing (LSH)

- Hashing technique that hashes similar input items into the same "buckets" with high probability

A finite family $\mathcal{F}$ of functions $h\colon M \to S$ is defined to be an LSH family for a metric space $\mathcal{M} = (M, d)$, a treshold $r > 0$, an approximate factor $c > 1$ and probabilities $p_1 > p_2$ if it satisfies the following condition. For any two points $a, b \in M$ and a hash function h chosen uniformly at random from $\mathcal{F}$:

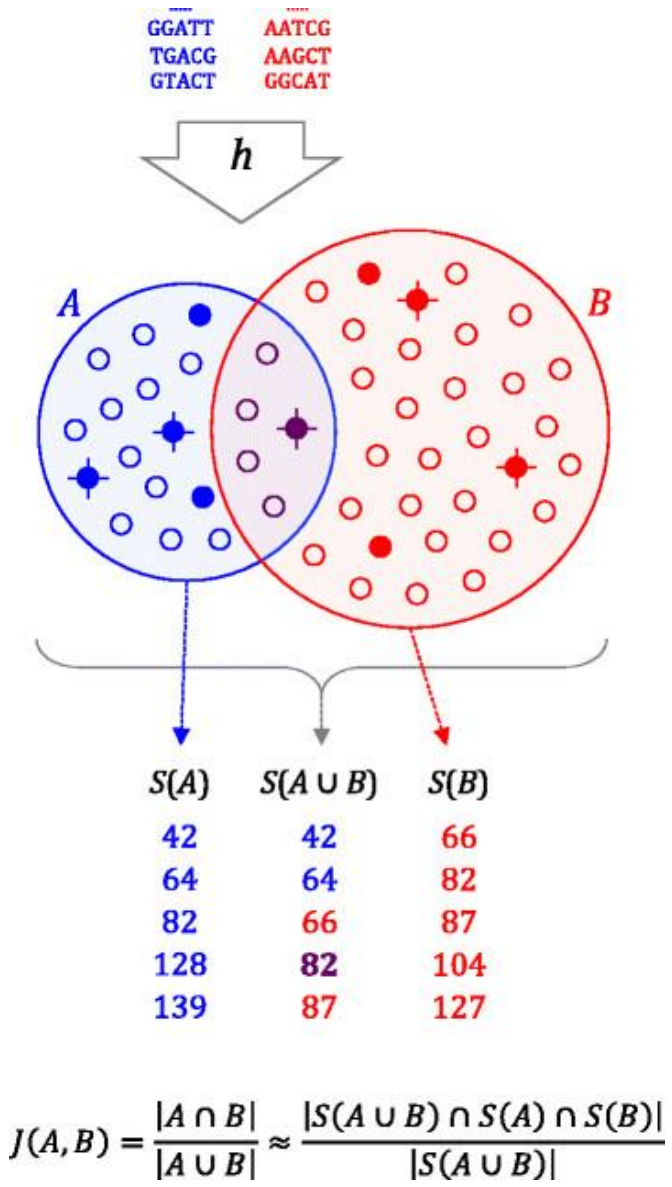
- If $d(a, b) \leq r$, then $h(a) = h(b)$ with probability at least p_1
- If $d(a, b) \geq cr$, then $h(a) = h(b)$ with probability at most p_2.

MinHash

- MinHash is a method for estimating how similar two sets are
- MinHash takes the minimum hash value of all the elements in the set

Often estimated with Jaccard similarity:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

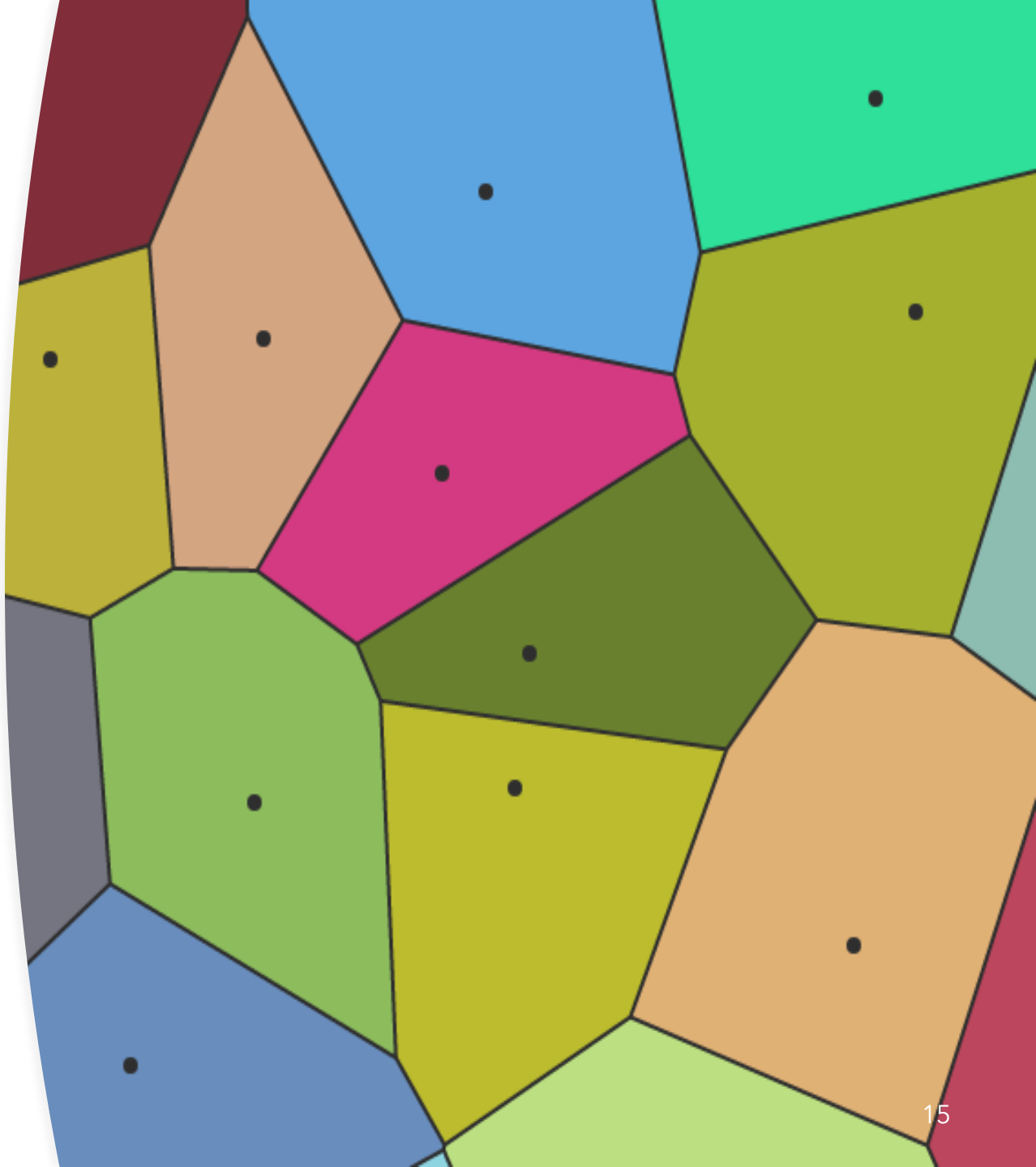


Inverted file with flat compression

- When querying, find a centroid, then the closest data point in the centroid
- PostgreSQL's pgvector extension to speed up similarity searches

https://www.timescale.com/blog/nearest-neighbor-indexes-what-are-ivfflat-indexes-in-pgvector-and-how-do-they-work/

https://towardsdatascience.com/similarity-search-with-ivfpq-9c6348fd4db3



Conclusion

- Word2Vec, GloVe, Hierarchical Navigable Small Worlds, Locality Sensitive Hashing, and Inverted files with flat compression share the common goal of **efficient representation, storage, and retrieval** of high-dimensional vector data