zilliz

Vector Database 101: A Crash Course

Yujian Tang | Zilliz

Speaker

Yujian Tang

Developer Advocate, Zilliz

yujian@zilliz.com
https://www.linkedin.com/in/yujiantang
https://www.twitter.com/yujian_tang

 |

zilliz

Zilliz at a Glance

Founded	2017
Headquarters	Redwood Shores, CA
Focus	Vector database company for enterprise-grade AI built on Milvus, the popular open-source vector database that helps organizations quickly create AI applications.

Key maintainer of the following Open-Source projects

milvus · Attu · towhee · GPT-Cache · VDB Benchmark · Knowhere

zilliz

CONTENTS

 |

zilliz

01

Vector Databases 101 Review

zilliz

Why Vector Databases?

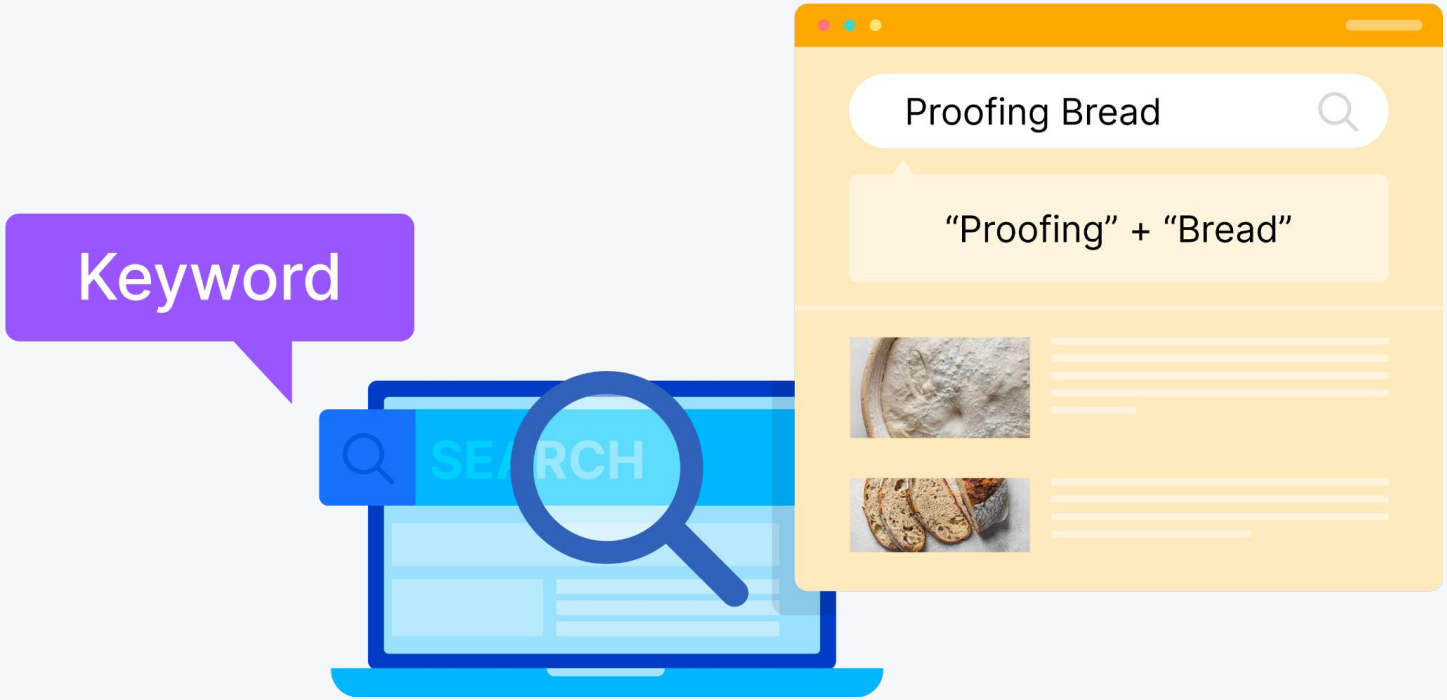
- Unstructured Data is 80% of data

- Vector Databases are the only type of database that can work with unstructured data

- Examples of Unstructured Data include text, images, videos, audio, etc

zilliz

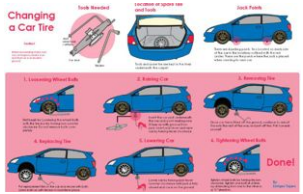
Traditional databases were built on exact search



 |

zilliz

...which misses context, semantic meaning, and user intent

Apple	Rising dough	Change car tire
	Rising Dough ✔	
VS.	*VS.*	*VS.*
	Proofing Bread ❌	

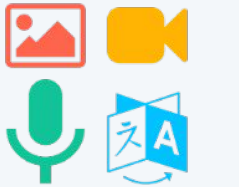
zilliz

But wait! There’s more!

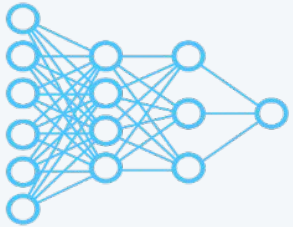
 |

zilliz

Where do Vectors Come From?



Knowledge Base (Documents)



Deep Learning Models

$$\begin{bmatrix} V_{1,1} & V_{1,2} & \cdots & V_{1,n} \\ V_{2,1} & V_{2,2} & \cdots & V_{2,n} \\ V_{3,1} & V_{3,2} & \cdots & V_{3,n} \\ \vdots & \vdots & \vdots & \vdots \\ V_{n,1} & V_{n,2} & \cdots & V_{n,n} \end{bmatrix}$$

Vectors

zilliz

milvus

Vector Databases

 |

zilliz

02

Vector Database Use Cases

zilliz

Common AI Use Cases

Retrieval Augmented Generation (RAG)

Expand LLMs' knowledge by incorporating external data sources into LLMs and your AI applications.

Recommender System

Match user behavior or content features with other similar ones to make effective recommendations.

Text/ Semantic Search

Search for semantically similar texts across vast amounts of natural language documents.

Image Similarity Search

Identify and search for visually similar images or objects from a vast collection of image libraries.

Video Similarity Search

Search for similar videos, scenes, or objects from extensive collections of video libraries.

Audio Similarity Search

Find similar audios in large datasets for tasks like genre classification or speech recognition

Molecular Similarity Search

Search for similar substructures, superstructures, and other structures for a specific molecule.

Anomaly Detection

Detect data points, events, and observations that deviate significantly from the usual pattern

Multimodal Similarity Search

Search over multiple types of data simultaneously, e.g. text and images

 |

Reverse Image Search - Paintings

 |

zilliz

Text Search on Wikipedia

```
pprint(res_512_plot.response)
```

```
('The plot of "The Nightmare Before Christmas" revolves around Jack '
 'Skellington, the Pumpkin King of Halloween Town, who becomes tired of the '
 'same routine of Halloween and discovers Christmas Town. Intrigued by the '
 'concept of Christmas, Jack decides that Halloween Town will take over '
 'Christmas this year. He assigns the residents various Christmas-themed '
 "tasks, but his efforts lead to disastrous consequences. Jack's love "
 'interest, Sally, warns him about the potential disaster, but he dismisses '
 'her warnings. Eventually, Jack realizes his mistake and sets out to fix the '
 'chaos he has caused. With the help of Santa Claus, Jack saves Christmas and '
 'learns the true meaning of the holiday. The film ends with Jack and Sally '
 'declaring their love for each other [6][7][8][9].')
```

 |

zilliz

03

How do Vector Databases Work?

zilliz

Example Entry

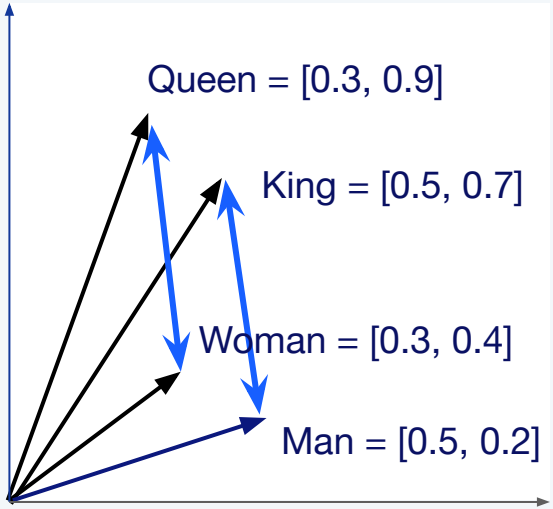
"id": "https://towardsdatascience.com/detection-of-credit-card-fraud-with-an-autoencoder-9275854
"embedding": [-0.042092223,-0.0154002765,-0.014588429,-0.031147376,0.03801204,0.013369046,(
"date": "2023-06-01"
"paragraph": "We define an anomaly as follows:"
"reading_time": "11"
"subtitle": "A guide for the implementation of an anomaly..."
"publication": "Towards Data Science"
"responses": "1"
"article_url": "https://towardsdatascience.com/detection-of-credit-card-fraud-with-an-autoencoder-
"title": "Detection of Credit Card Fraud with an Autoencoder"
"claps": "229"

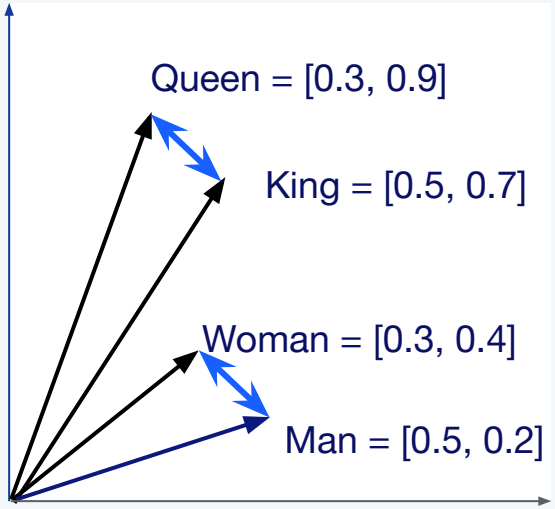
↑ Hide 6 fields 🔍 Vector search

 |

zilliz

Semantic Similarity





Queen - Woman + Man = King

Queen = [0.3, 0.9]
\- Woman = [0.3, 0.4]

[0.0, 0.5]
\+ Man = [0.5, 0.2]

King = [0.5, 0.7]

Image from Sutor et al

 |

zilliz

Vector Similarity Measures: L2 (Euclidean)

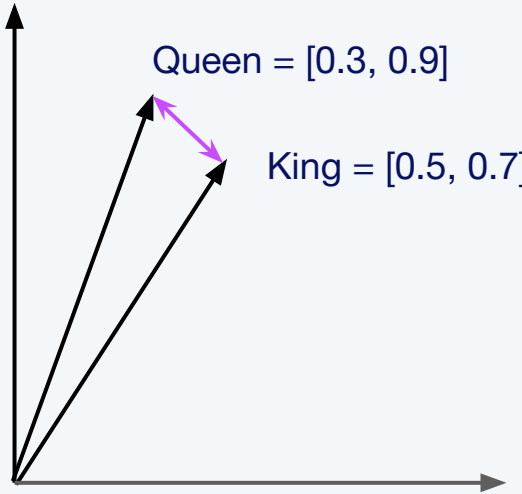
$$d(\mathbf{p}, \mathbf{q}) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$

d(Queen, King) = $\sqrt{(0.3-0.5)^2 + (0.9-0.7)^2}$

= $\sqrt{(0.2)^2 + (0.2)^2}$

= $\sqrt{0.04 + 0.04}$

= $\sqrt{0.08} \cong 0.28$



 |

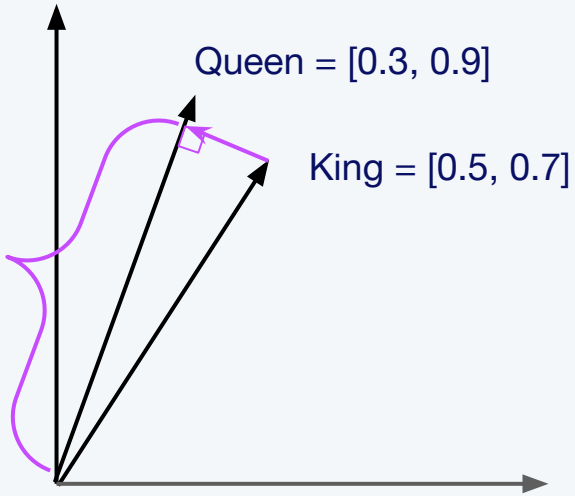
zilliz

Vector Similarity Measures: Inner Product (IP)

$$a \cdot b = \sum_{i=1}^{n} a_i b_i$$

Queen · King = (0.3*0.5) + (0.9*0.7)

= 0.15 + 0.63 = 0.78

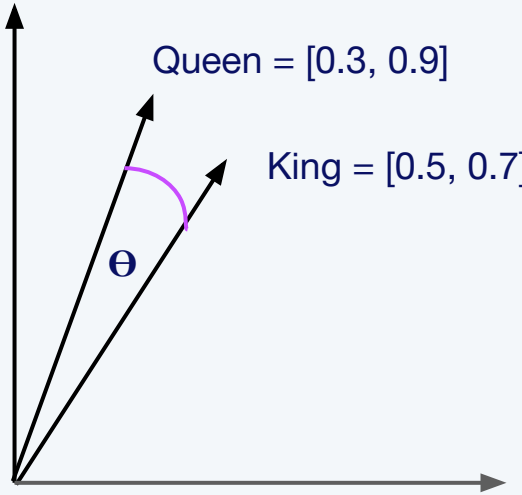


 |

zilliz

Vector Similarity Measures: Cosine

$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$



cos(Queen, King) = $\frac{(0.3*0.5)+(0.9*0.7)}{\sqrt{0.3^2+0.9^2} * \sqrt{0.5^2+0.7^2}}$

= $\frac{0.15+0.63}{\sqrt{0.9} * \sqrt{0.74}}$

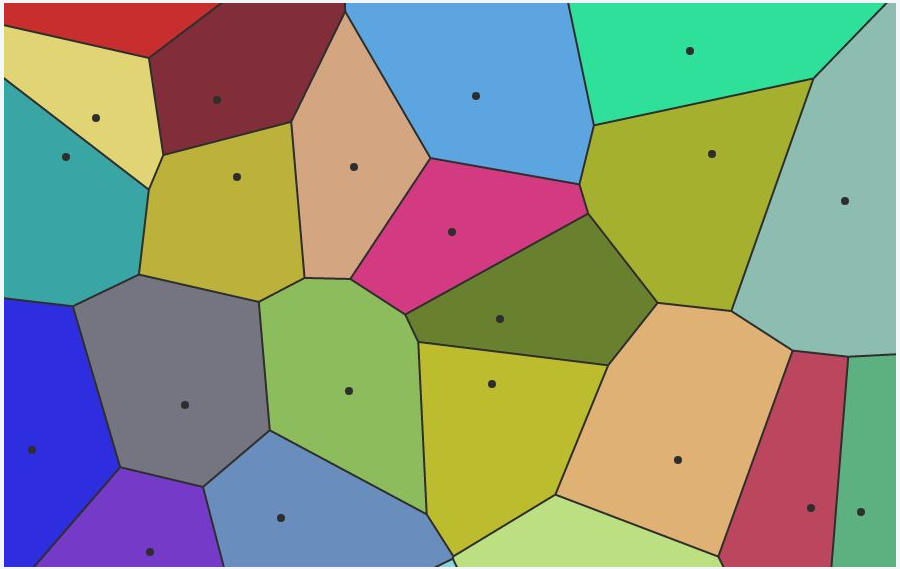
= $\frac{0.78}{\sqrt{0.666}}$

≅ 0.03

 |

zilliz

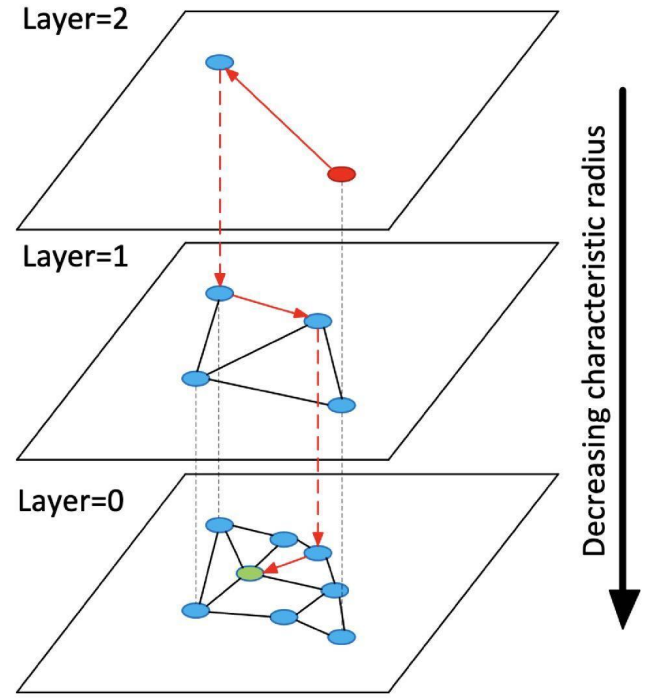
Inverted File Index



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

zilliz

HNSW

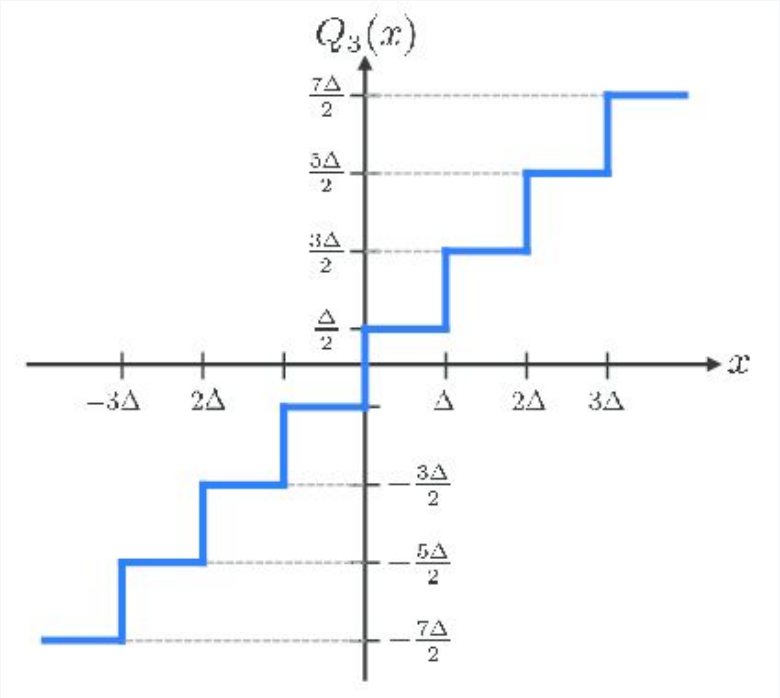


Source:
https://arxiv.org/ftp/arxiv/papers/1603/1603.09320.pdf

 |

zilliz

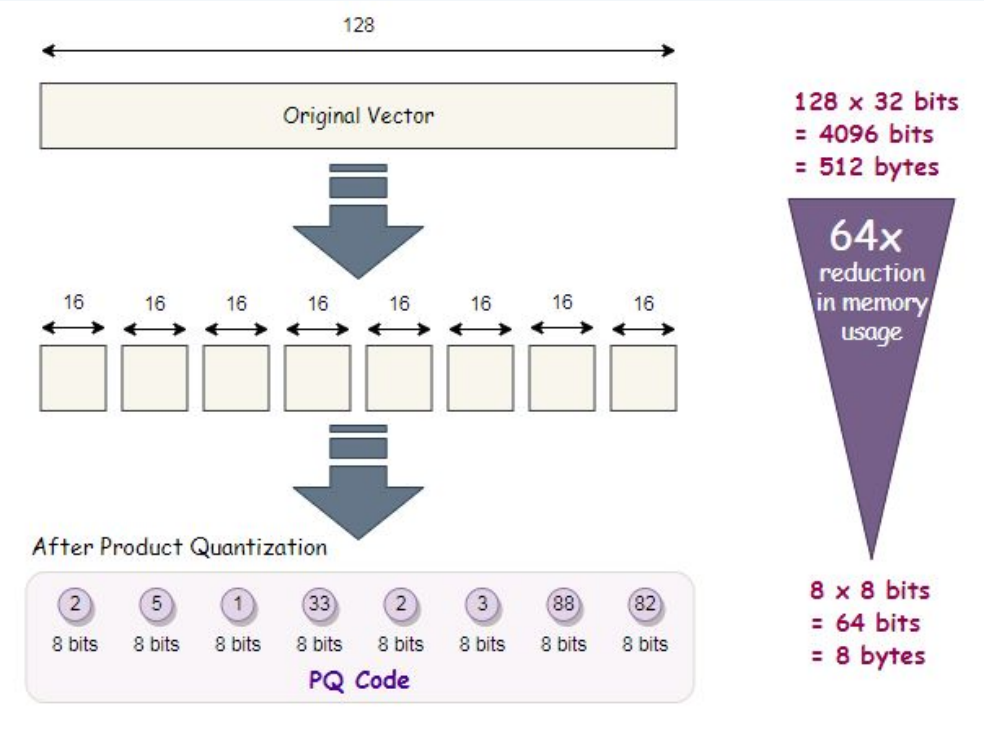
SQ



 |

zilliz

PQ



Source:
https://towardsdatascience.com/product-quantization-for-similarity-search-2f1f67c5fddd

zilliz

04

Demo

zilliz

What is the Demo?

- Download three embeddings models
- Make up a data set
- Embed your data set with two models
- Ingest your data sets into Milvus
- Query and compare the third set of embeddings against your loaded data
- What do you think will happen?

zilliz

Demo

 |

zilliz

Start building
with Zilliz Cloud today!
zilliz.com/cloud