



Vector Databases and the Future of AI-powered Search



Sam Partee

Principal Applied AI Engineer, Redis AI & ML

Twitter: @SamPartee



Vector Embeddings

What are vector embeddings and how are they created?

Vectors

- Commonly represent unstructured data
 - Videos, text, images, etc
- Usually of high-dimension in the form of a **dense** embedding.
- Reduce data size for index and search
- Use “off-the shelf” models to create



Hugging Face

Vector Embedding Creation

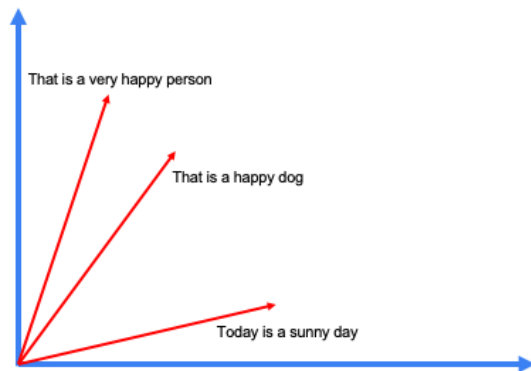
- Simple creation APIs
- Example with HuggingFace Sentence Transformer

```
1 from sentence_transformers import SentenceTransformer
2 model = SentenceTransformer('sentence-transformers/all-MiniLM-L6-v2')
3
4 sentences = [
5     "That is a very happy Person",
6     "That is a Happy Dog",
7     "Today is a sunny day"
8 ]
9 embeddings = model.encode(sentences)
```

Vector Similarity Search

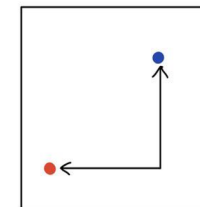
How are vector embeddings used for similarity search?

- Three semantic (text) vectors
 - “today is a sunny day”
 - “that is a very happy person”
 - “that is a very happy dog”

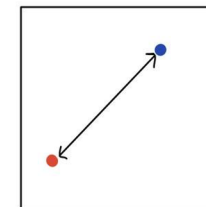


Distance Metrics

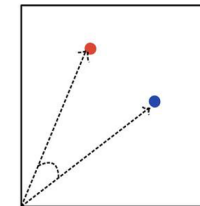
- Calculate the distance between many vectors in a dataset



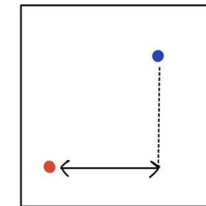
Manhattan



Euclidean



Cosine

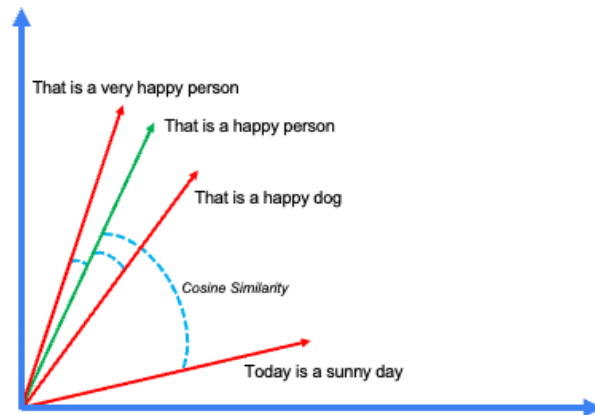


Chebyshev

Vector Similarity Search

How are vector embeddings used for similarity search?

- Query Vector
 - “That is a happy person”



Cosine Similarity

- Cosine distance between vectors

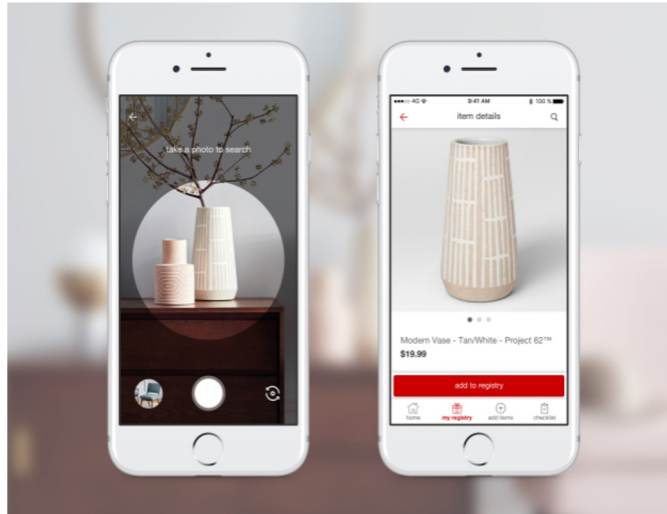
$$\cos(\mathbf{t}, \mathbf{e}) = \frac{\mathbf{t} \cdot \mathbf{e}}{\|\mathbf{t}\| \|\mathbf{e}\|} = \frac{\sum_{i=1}^n t_i e_i}{\sqrt{\sum_{i=1}^n (t_i)^2} \sqrt{\sum_{i=1}^n (e_i)^2}}$$

- Cosine Similarity results

That is a happy dog	0.695
That is a very happy person	0.943
Today is a sunny day	0.257

Vector Search – Use cases

Visual Search



Find similar products through image data

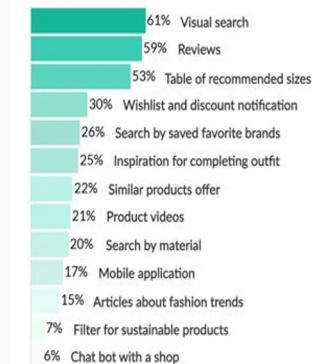
Natural Language Search



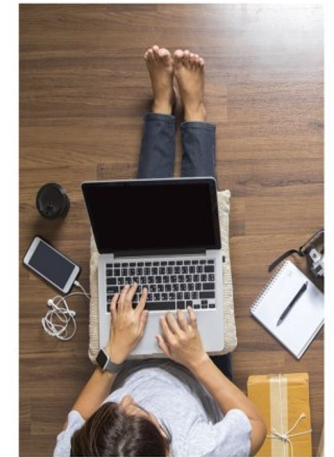
Semantic search, Q&A, document retrieval

Recommenders

Features that customers desire



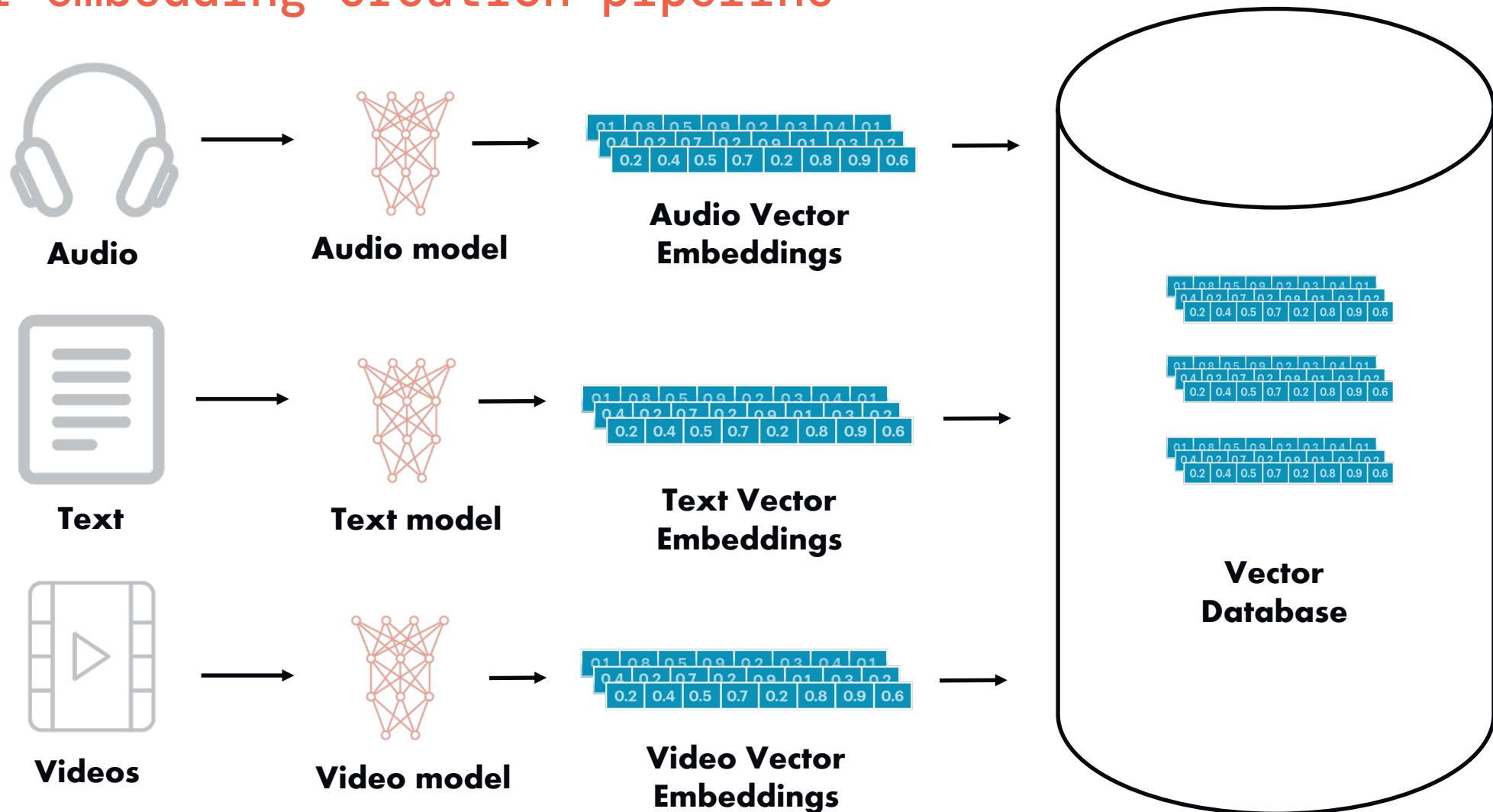
*Choose 3-5 features you consider important while shopping online for fashion**



Recommend products, brands, properties, etc

Vector Embeddings

Vector embedding creation pipeline



Vector Database

What is a vector database?

Vector Database

- Purpose-built database to store, index, and query vector embeddings generated by passing unstructured data through machine learning models
- Common indexing methods
 - Flat (brute-force)
 - HNSW
 - FAISS

Vector Database Requirements

- Fast
 - Low-latency read and writes.
 - Usually vectors stored in-memory
- Reliable
 - Fault tolerant (replication)
 - Highly available
- Scalable
 - Distributed, often tiered memory

Redis: Vector Search



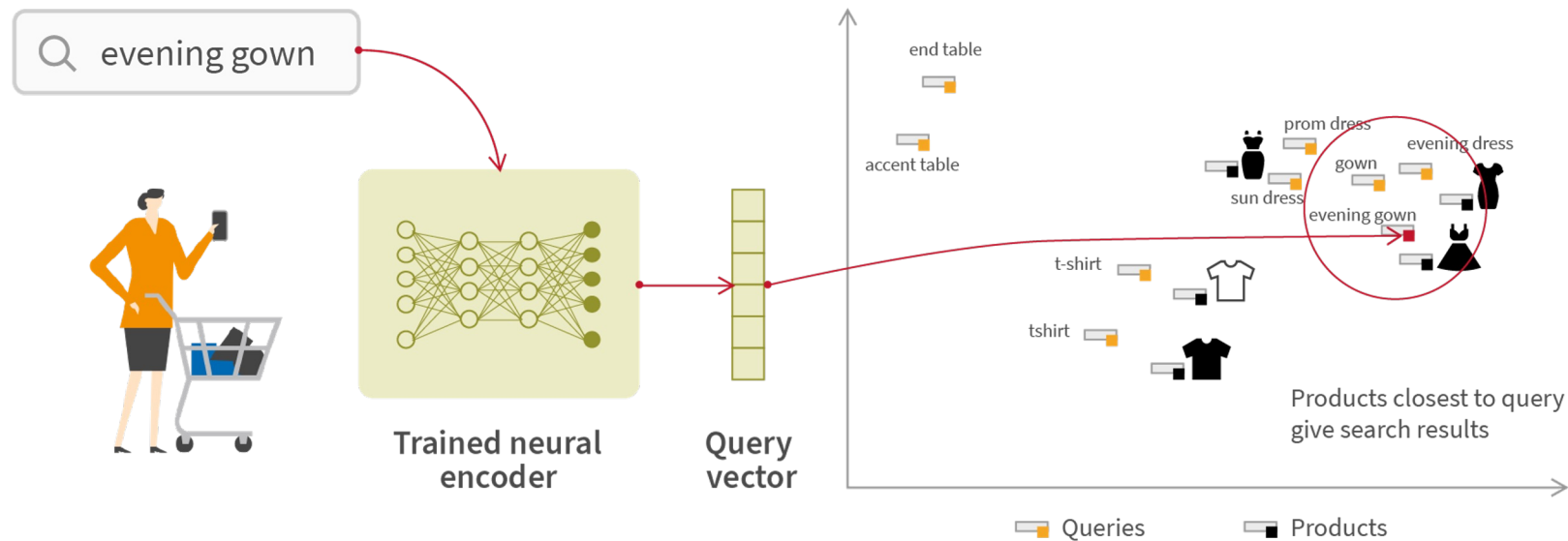
Redis Vector Similarity Search

 RediSearch 2.4

- **Redis:** Low-latency, scalable, in-memory database
- Indexing methods
 - HSNW (ANN)
 - Flat (KNN)
- Distance metrics
 - L2, Cosine, internal product
- Support for hybrid queries
 - Vector search + filtering by text, geo, etc.

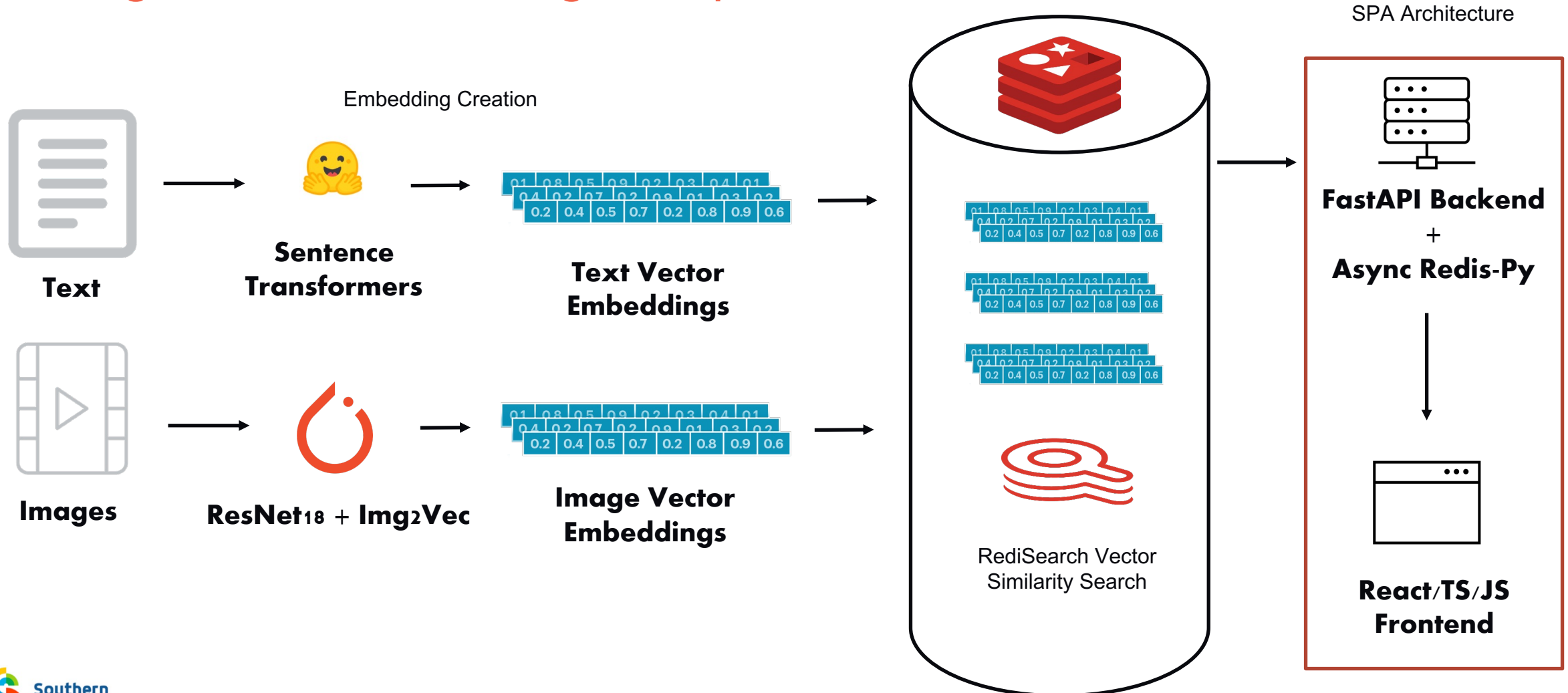
Vector Similarity Search Demos

Visual and Semantic Search demo applications with Redis VSS

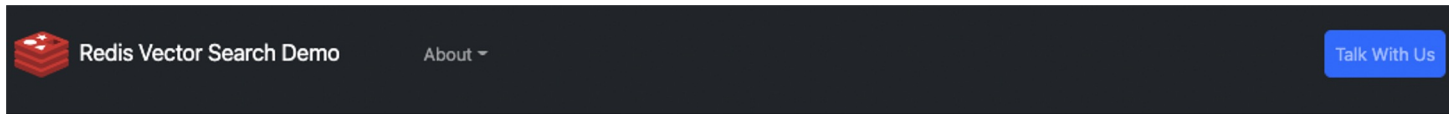


Fashion Product Finder

Using vector embeddings in production



Redis VSS Demo: Visual Vector Search



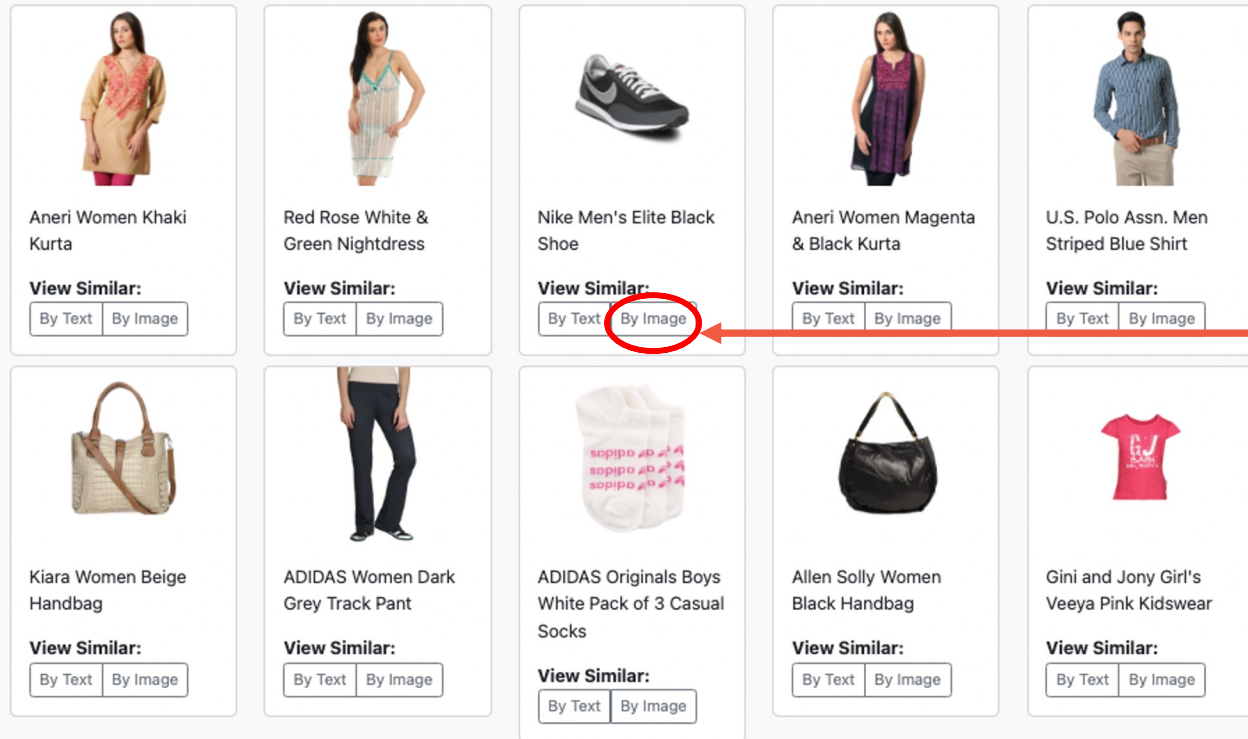
Fashion Product Finder

This demo uses the built in Vector Search capabilities of Redis Enterprise to show how unstructured data, such as images and text, can be used to create powerful search engines.

Apply Filters

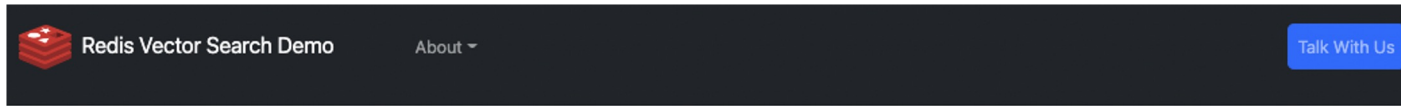
Load More Products

3000 products



Search by Image

Redis VSS Demo: Visual Vector Search



Fashion Product Finder

This demo uses the built in Vector Search capabilities of Redis Enterprise to show how unstructured data, such as images and text, can be used to create powerful search engines.

Apply Filters

Load More Products

Query Image



3000 products



Nike Men's Elite Black Shoe

View Similar:

By Text By Image

1.00



Nike Women's Double Team Lite Black Shoe

View Similar:

By Text By Image

0.94



Nike Men's Incinerate MSL White Blue Shoe

View Similar:

By Text By Image

0.92



Nike Men's LunarFly Blue Shoe

View Similar:

By Text By Image

0.91



Nike Men White Capri II Casual Shoe

View Similar:

By Text By Image

0.91



Nike Women Sweet AC Black Shoe

View Similar:

By Text By Image

0.91



F Sports Men Navy Blue Vito Shoes

View Similar:

By Text By Image

0.91



Puma Men Black Aquil II Sports Shoes

View Similar:

By Text By Image

0.91



Gas Men Mila Navy Blue Shoes

View Similar:

By Text By Image

0.91



Nike Men The Overplay VII Black Sports Shoes

View Similar:

By Text By Image

0.91

Product
"Similarity Score"

Sample Query by Image

FastAPI example route for building Redis VSS enabled apps

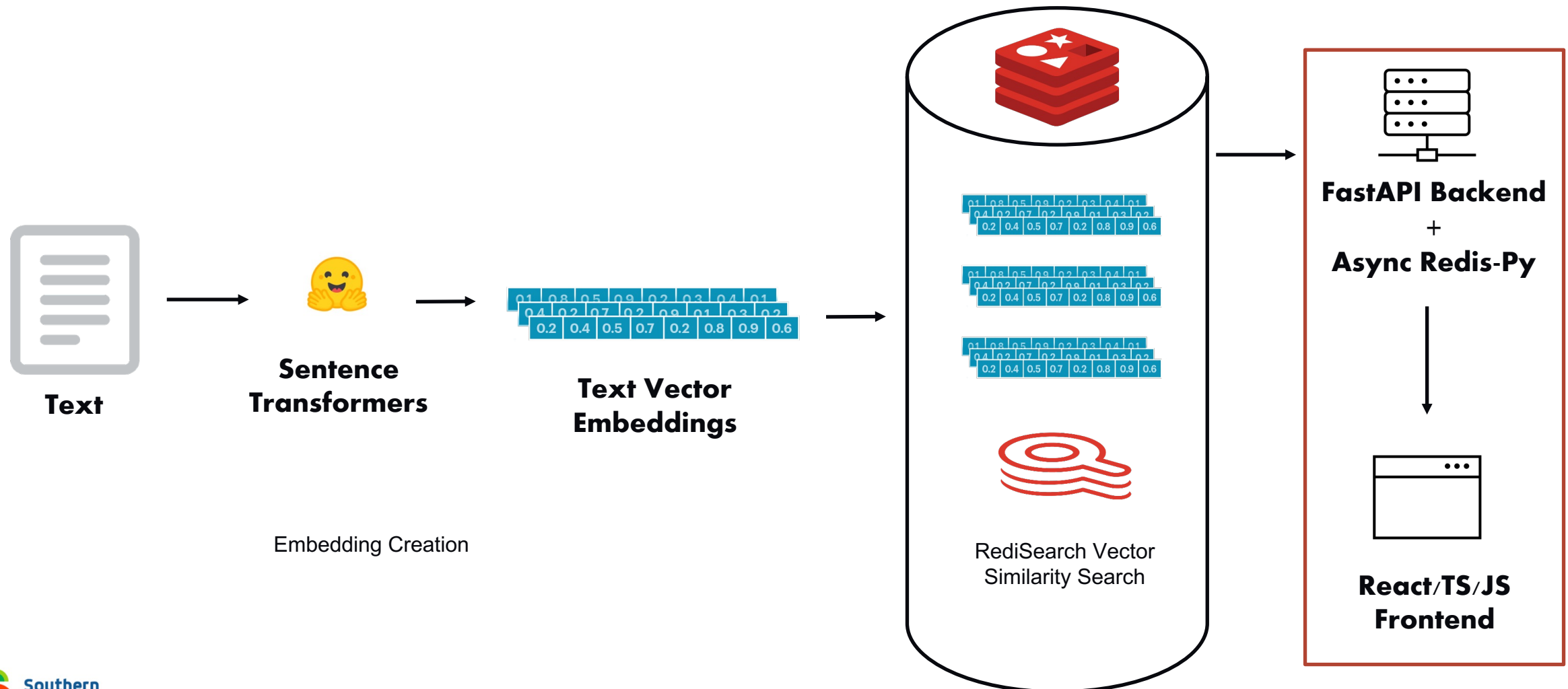
```
1  @r.post("/vectorsearch/image", response_model=t.List[Product])
2  async def find_products_by_image(request: SimilarityRequest):
3
4      # Obtain image vector of the queried product and create query
5      vector = redis_client.hget(request.product_id)
6      q = create_query(request.search_type, request.num_results)
7
8      # Use Redis search capabilities to return visually similar products
9      results = redis_client.ft().search(q, query_params={"vector": vector})
10     return [await Product.get(product.pk) for product in results]
```

Try it out!

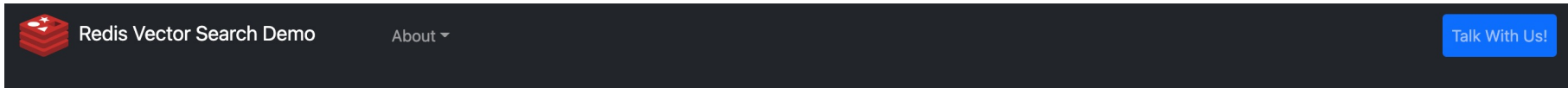
<https://ecommerce.redisventures.com>

Arxiv Paper Search

Document Retrieval with Redis Vector Similarity



Redis VSS Demo: Text Semantic Search



arXiv Paper Search

This demo uses the built in Vector Search capabilities of Redis Enterprise to show how unstructured data, such as paper abstracts (text), can be used to create a powerful search engine.

Enter a search query below to discover scholarly papers hosted by [arXiv](#) (Cornell University).

Search by Text
(Vector)

deep financial models

×

Applications of deep learning in stock market prediction: recent progress
[Read Me](#) [Download](#)
Weiwei Jiang

More Like This

Deep Prediction of Investor Interest: a Supervised Clustering Approach
[Read Me](#) [Download](#)
Baptiste Barreau, Laurent Carlier, Damien Challet

More Like This

Multi-Transformer: A New Neural Network-Based Architecture for Forecasting S&P Volatility
[Read Me](#) [Download](#)
Eduardo Ramos-Piñerez, Pablo J. Alonso-González, Josée Javier Ní~nez-Velázquez

More Like This

Autoencoding Conditional GAN for Portfolio Allocation Diversification
[Read Me](#) [Download](#)
Jun Lu, Shao Yi

More Like This

Stock Portfolio Optimization Using a Deep Learning LSTM Model
[Read Me](#) [Download](#)
Jaydip Sen, Abhishek Dutta, and Sidra Mehtab

More Like This

Prior knowledge distillation based on financial time series
[Read Me](#) [Download](#)
Jie Fang and Jianwu Lin

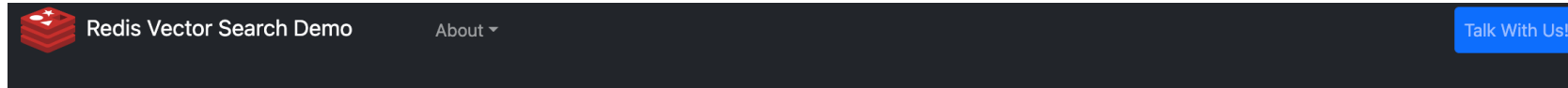
Improving Decision Analytics with Deep Learning: The Case of Financial Disclosures
[Read Me](#) [Download](#)

Deep Factors for Forecasting
[Read Me](#) [Download](#)
Yuyang Wang, Alex Smola, Danielle C. Maddix, Jan Gasthaus, Dean Foster, Tim Januschowski

Equity2Vec: End-to-end Deep Learning Framework for Cross-sectional Asset Pricing
[Read Me](#) [Download](#)

Earnings Prediction with Deep Learning
[Read Me](#) [Download](#)
Lars Elend, Sebastian A. Tideman, Kerstin Lopatta, Oliver Kramer

Redis VSS Demo: Text Semantic Search



arXiv Paper Search

This demo uses the built in Vector Search capabilities of Redis Enterprise to show how unstructured data, such as paper abstracts (text), can be used to create a powerful search engine.

Enter a search query below to discover scholarly papers hosted by [arXiv](#) (Cornell University).

predicting medical diagnoses

Search by Text
(Vector)

Leveraging Implicit Expert Knowledge for Non-Circular Machine Learning in Sepsis Prediction

[Read Me](#) [Download](#)

Shigehiko Schamoni, Holger A. Lindner, Verena Schneider-Lindner, Manfred Thiel, Stefan Riezler

[More Like This](#)

On Classifying Sepsis Heterogeneity in the ICU: Insight Using Machine Learning

[Read Me](#) [Download](#)

Zina Ibrahim and Honghan Wu and Ahmed Hamoud and Lukas Stappen and Richard Dobson and Andrea Agarossi

[More Like This](#)

Predicting Patient COVID-19 Disease Severity by means of Statistical and Machine Learning Analysis of Blood Cell Transcriptome Data

[Read Me](#) [Download](#)

Sakifa Aktar, Md. Martuza Ahamad, Md. Rashed-Al-Mahfuz, AKM Azad, Shahadat Uddin, A H M Kamal, Salem A. Alyami, Ping-I Lin, Sheikh Mohammed Shariful Islam, Julian M.W. Quinn, Valsamma Eapen, and Mohammad Ali Moni

[More Like This](#)

Prognosis Prediction in Covid-19 Patients from Lab Tests and X-ray Data through Randomized Decision Trees

[Read Me](#) [Download](#)

Alfonso Emilio Gerevini, Roberto Maroldi, Matteo Olivato, Luca Putelli, Ivan Serina

[More Like This](#)

Automatically Explaining Machine Learning Prediction Results: A Demonstration on Type 2 Diabetes Risk Prediction

[Read Me](#) [Download](#)

Gang Luo

[More Like This](#)

Interpretable Machine Learning for COVID-19: An Empirical Study on Severity Prediction Task

Multiclass Disease Predictions Based on Integrated Clinical and Genomics Datasets

A Simple and Interpretable Predictive Model for Healthcare

Identifying Cancer Patients at Risk for Heart Failure Using Machine Learning Methods

Predicting Cancer Using Supervised Machine Learning: Mesothelioma

Try it out!

<https://docsearch.redisventures.com>

Redis & Relevance AI

Enabling everyone to benefit from the power of VSS



Relevance + Redis Enterprise VSS

- A collaborative platform to quickly analyze unstructured data.
- Two interfaces:
 - **No-Code:** Tableau-like dashboards, reports, and workflows in a GUI.
 - **Low-Code API:** Use relevance features within your service.





THANK YOU



Sam Partee

Principal Applied AI Engineer, Redis AI & ML

Twitter: @SamPartee

