drant

miniCOIL

A Sparse Neural Retriever

Haystack 2025

qdrant

Vector Search

An essential part of the AI Transformation

INDUSTRIES

HR-Tech | Ad-Tech | Online Dating | Gig Economy

E-Commerce | Law-Tech | Fashion | Med-Tech

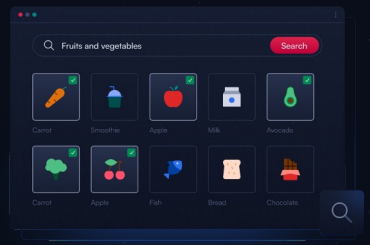
Ed-Tech | Media & News | Biometrics

Agriculture | Manufacturing | Streaming Services

Marketplaces | Anti-fraud

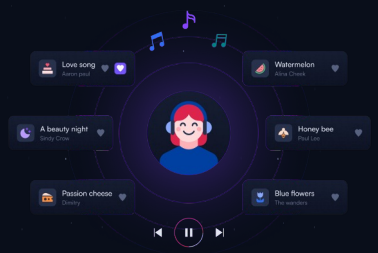
NN Encoders + Vector Database

SEARCH SYSTEMS

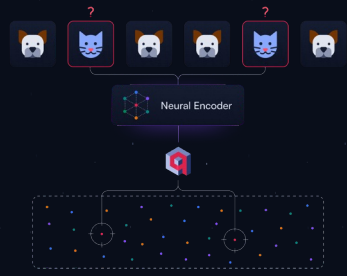


RAG / INFORMATION ASSISTANTS

RECOMMENDATIONS



ANOMALY DETECTION



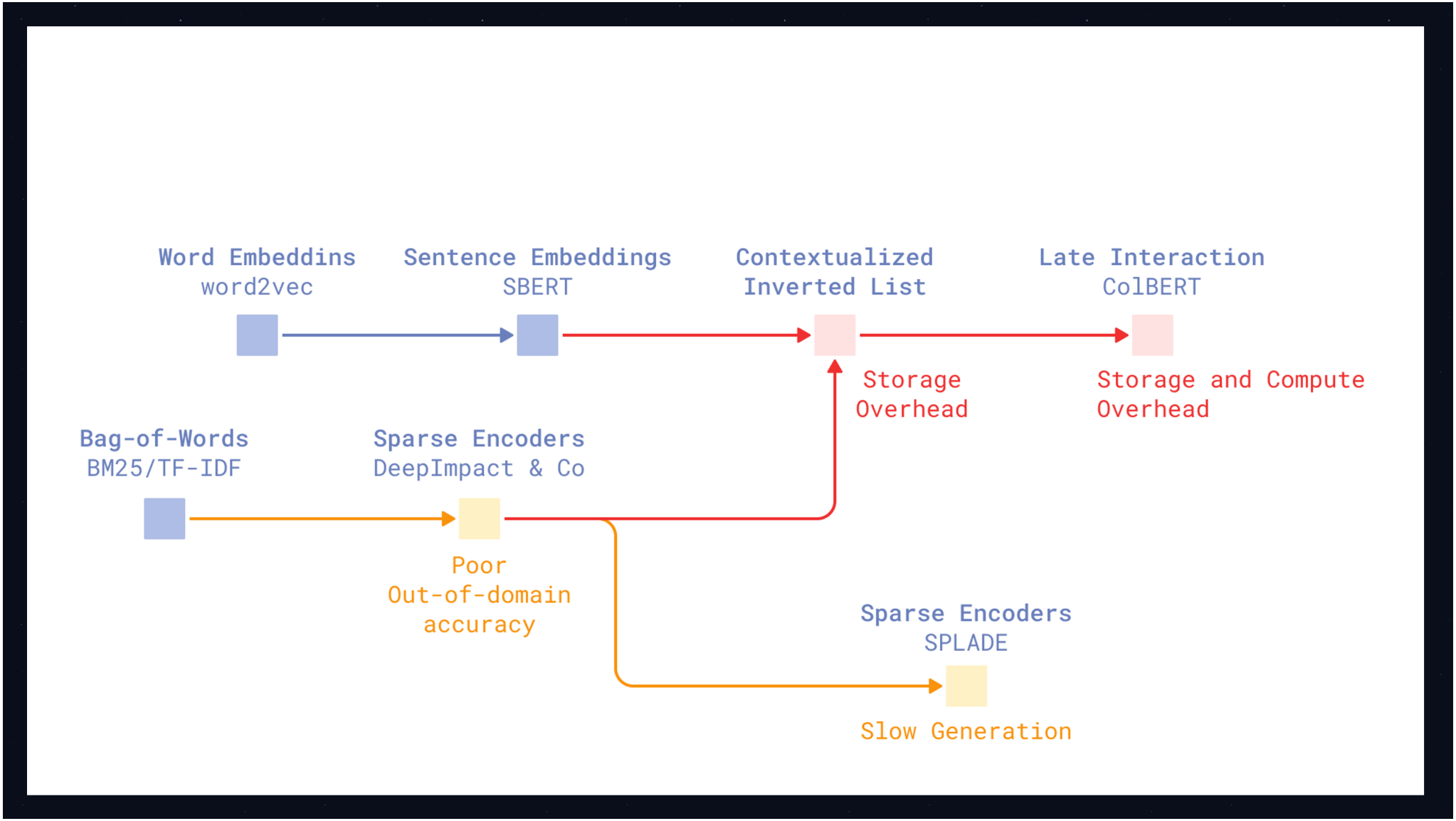
drant

You shall know a word

by the company it keeps

Agenda

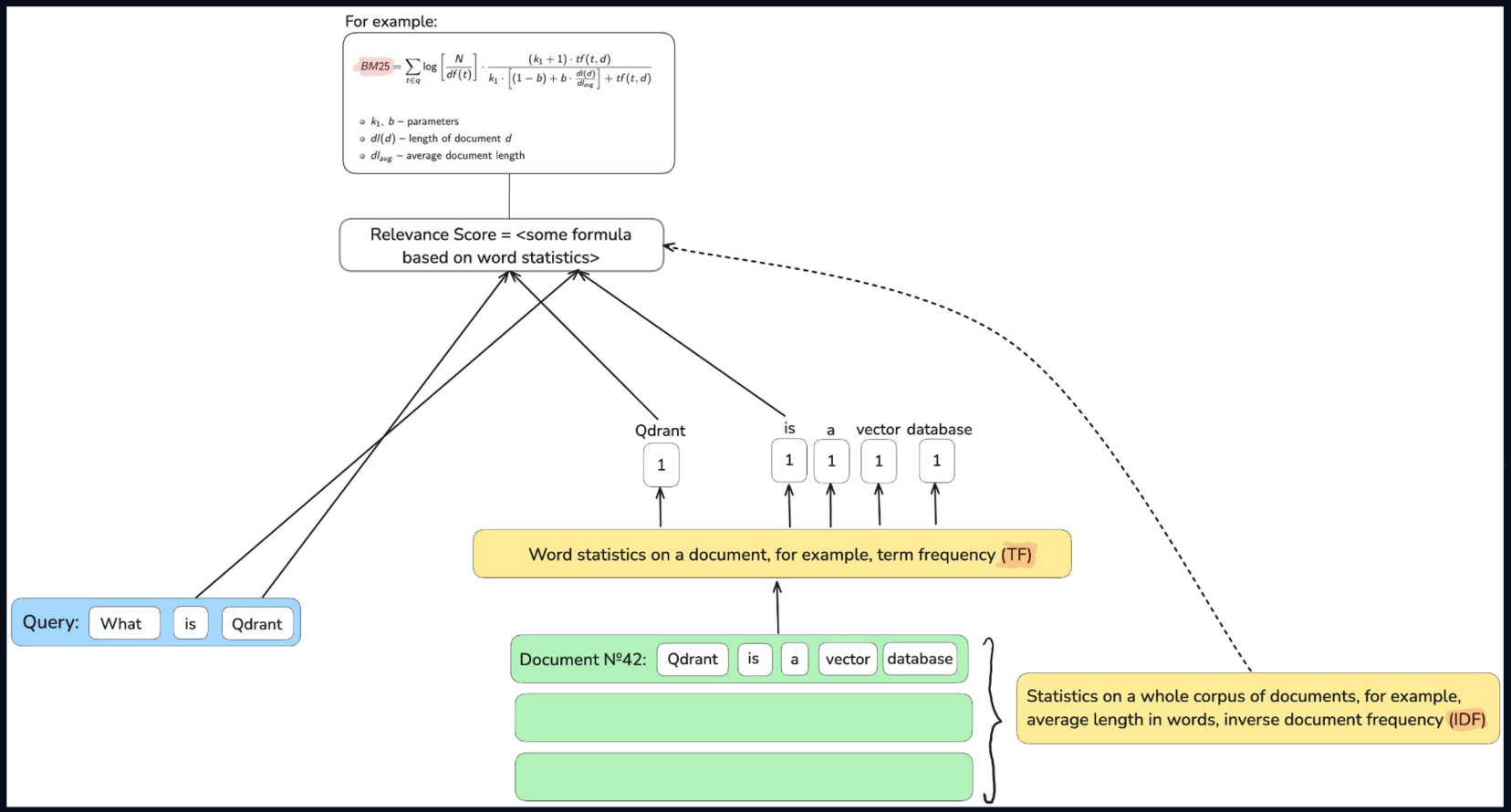
1. Motivation & BM25 recap
2. miniCOIL concept & 4-D vectors
3. Architecture & training pipeline
4. Advantages & benchmarks
5. Demo + how to extend
6. Q&A

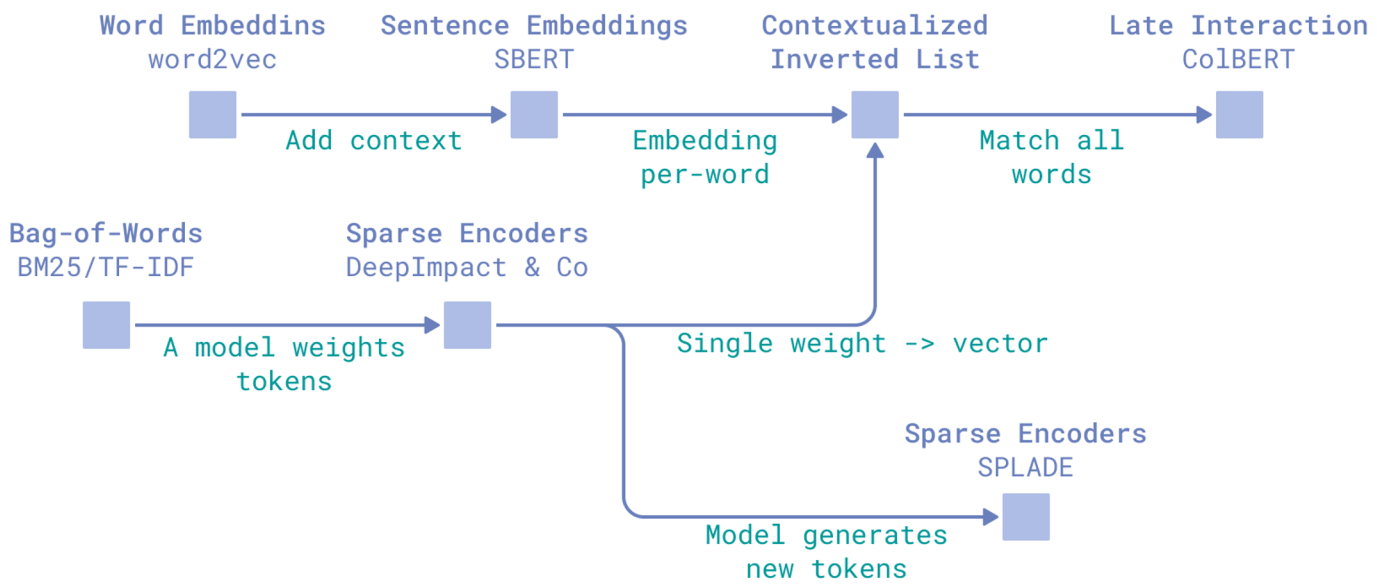

Word Embeddins
word2vec
Sentence Embeddings
SBERT
Contextualized
Inverted List
Late Interaction
ColBERT
Storage
Overhead
Storage and Compute
Overhead
Bag-of-Words
BM25/TF-IDF
Sparse Encoders
DeepImpact & Co
Poor
Out-of-domain
accuracy
Sparse Encoders
SPLADE
Slow Generation

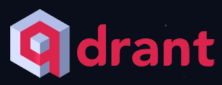
qdrant

BM25

$$\text{score}(D,Q) = \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \left[\frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot \left(1 - b + b \cdot \frac{|D|}{\text{avgdl}}\right)} + \delta \right]$$


For example:
BM25 = $\sum_{t \in q} \log\left[\frac{N}{df(t)}\right] \cdot \frac{(k_1+1)\cdot tf(t,d)}{k_1 \cdot \left[(1-b)+b\cdot\frac{dl(d)}{dl_{avg}}\right] + tf(t,d)}$
k_1, b – parameters
$dl(d)$ – length of document d
dl_{avg} – average document length
Relevance Score = <some formula based on word statistics>
Qdrant
is
a
vector
database
1
1
1
1
1
Word statistics on a document, for example, term frequency (TF)
Query:
What
is
Qdrant
Document №42:
Qdrant
is
a
vector
database
Statistics on a whole corpus of documents, for example, average length in words, inverse document frequency (IDF)


Word Embeddins
word2vec
Sentence Embeddings
SBERT
Contextualized
Inverted List
Late Interaction
ColBERT
Add context
Embedding
per-word
Match all
words
Bag-of-Words
BM25/TF-IDF
Sparse Encoders
DeepImpact & Co
A model weights
tokens
Single weight -> vector
Sparse Encoders
SPLADE
Model generates
new tokens

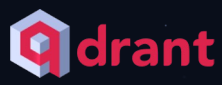


miniCOIL

BM25 formula

×

4-D context vectors

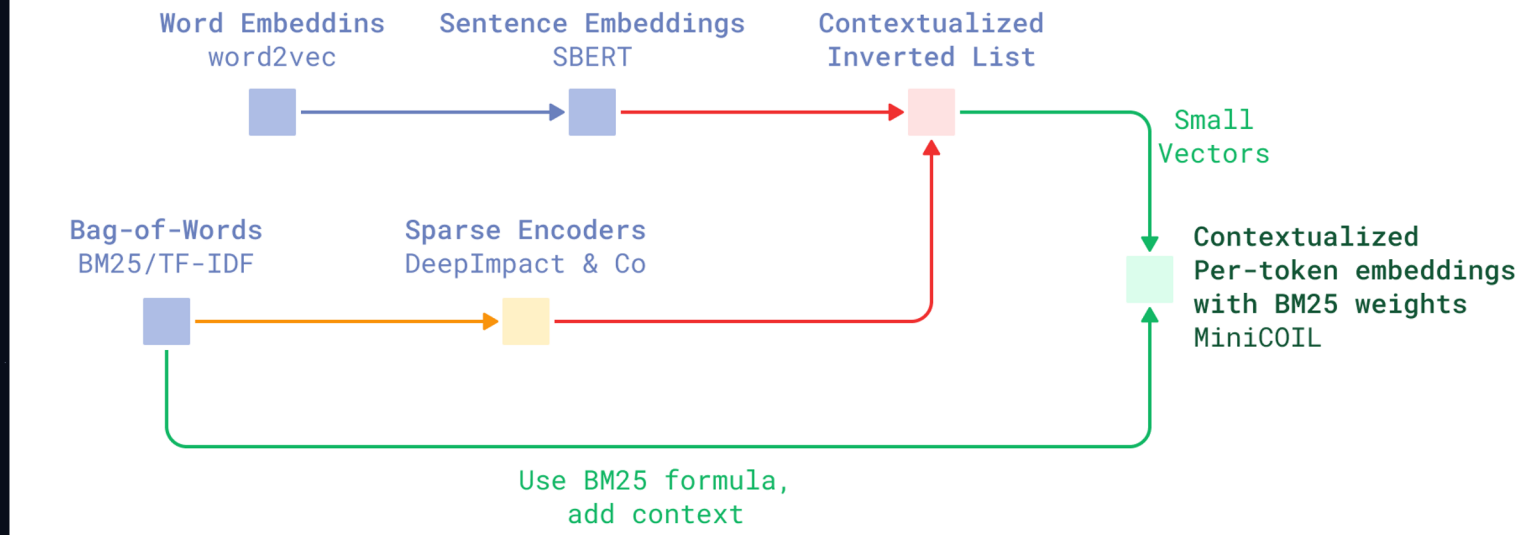


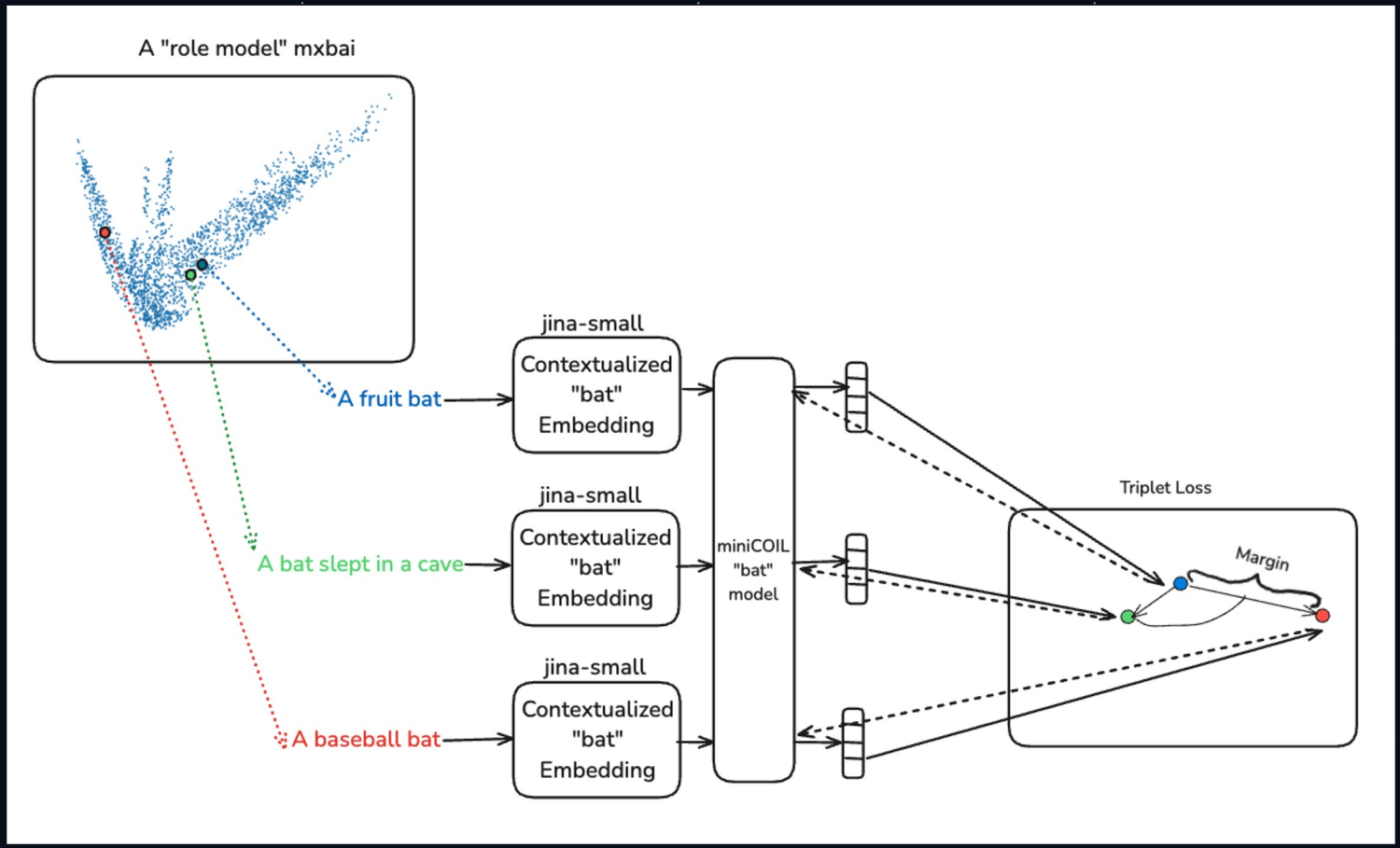
Homographs

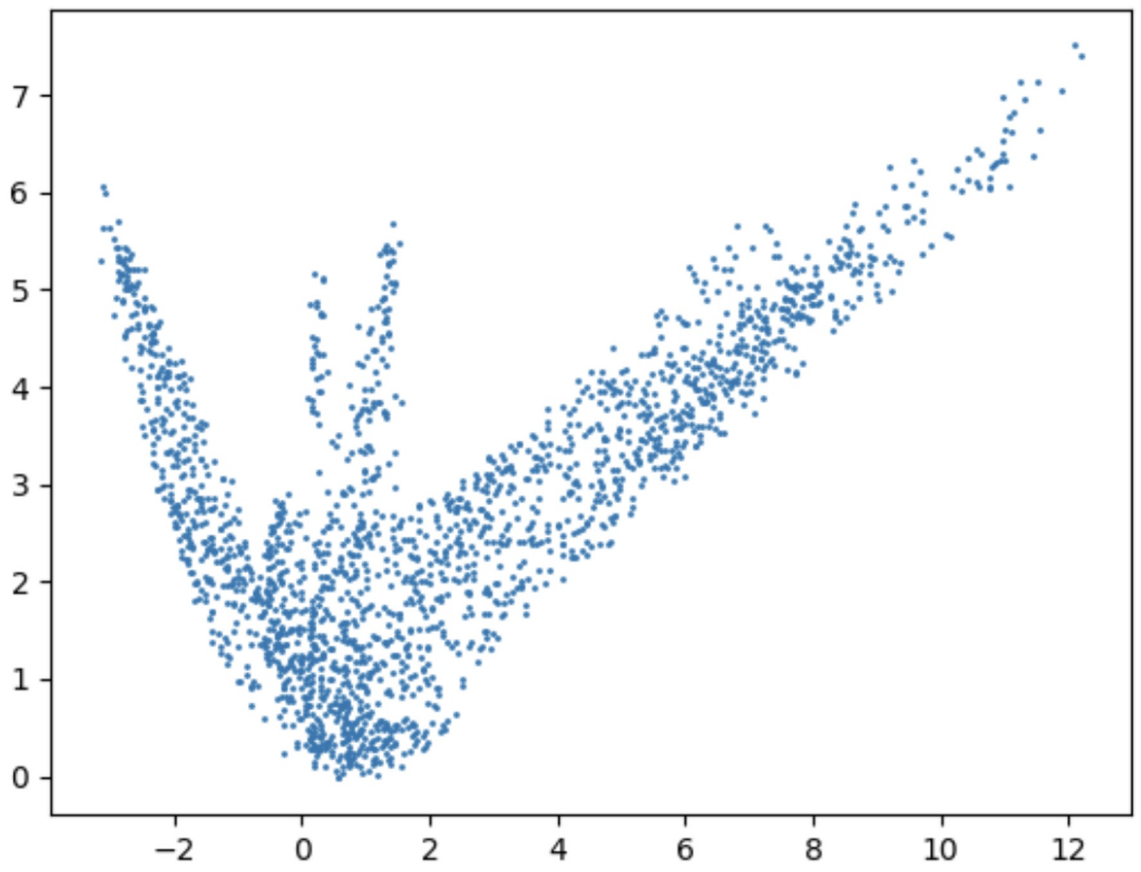
“fruit bat” ≈ “baseball bat”

“calculator” ≈ “calculating”


WELCOME TO THE FAMILY
MARVEL STUDIOS
the Fantastic 4
FIRST STEPS
ONLY IN CINEMAS JULY


Word Embeddins
word2vec
Sentence Embeddings
SBERT
Contextualized
Inverted List
Small
Vectors
Bag-of-Words
BM25/TF-IDF
Sparse Encoders
DeepImpact & Co
Contextualized
Per-token embeddings
with BM25 weights
MiniCOIL
Use BM25 formula,
add context


A "role model" mxbai
A fruit bat
A bat slept in a cave
A baseball bat
jina-small
Contextualized "bat" Embedding
jina-small
Contextualized "bat" Embedding
jina-small
Contextualized "bat" Embedding
miniCOIL "bat" model
Triplet Loss
Margin


7
6
5
4
3
2
1
0
−2
0
2
4
6
8
10
12

Architecture

Context Encoder:

We use a transformer model(we used Jina, but it is agnostic) to generate contextualized token embeddings.

Architecture

Word Mapping:

miniCOIL works with real words or stems, not subword tokens – like “retriever” instead of “re ##trie ##ver”.

Tokens

“retriever”

→ [re, ##trie, ##ver] (tokens)

→ “retriever” (miniCOIL)

Architecture

Downprojection Model:

Each word in the vocabulary has its own lightweight model that projects the contextualized embedding to a 4-dimensional vector.

Architecture

Sparse Scoring:

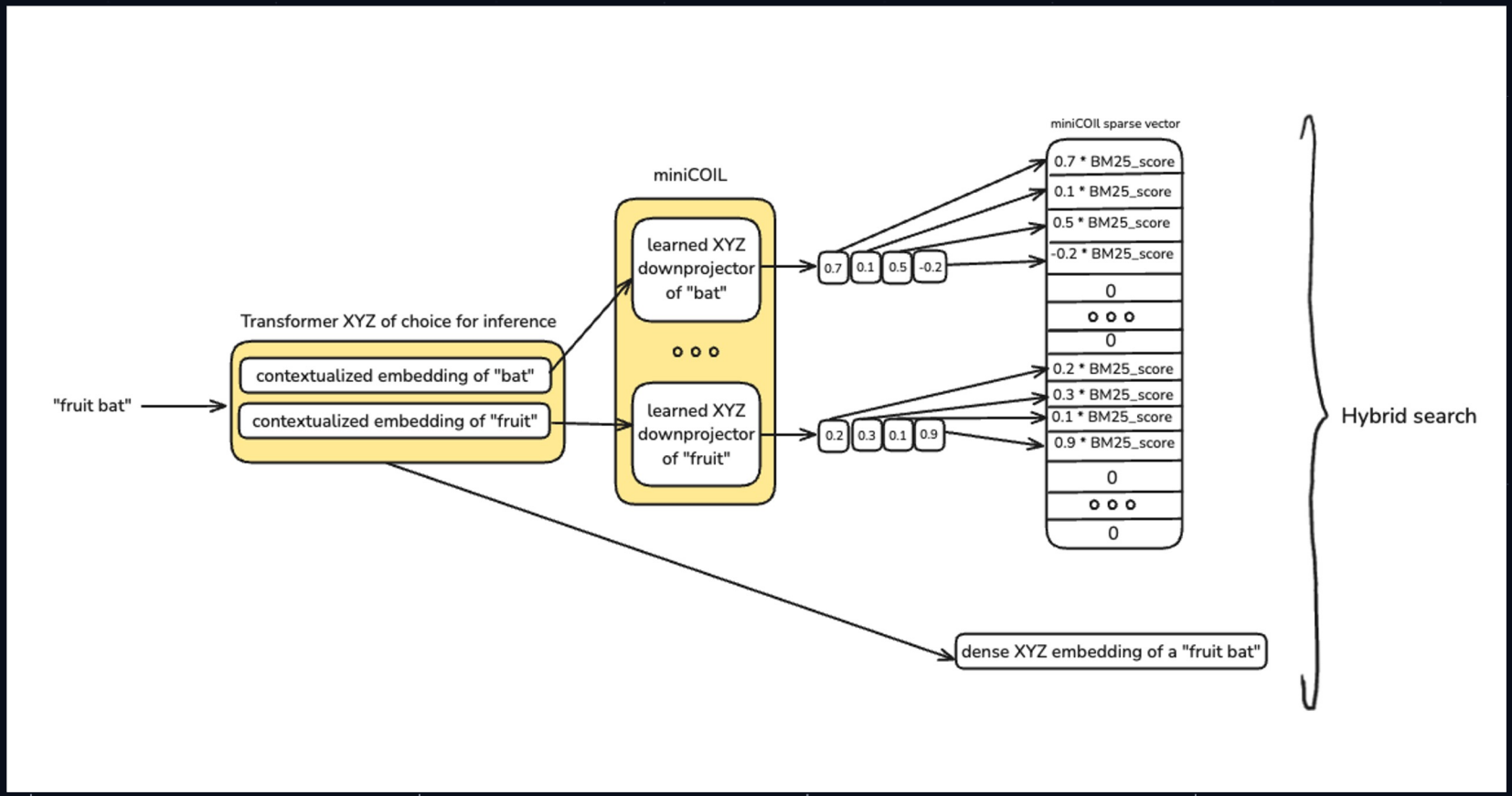
Each 4D vector is multiplied by the BM25 score for that word, creating a new sparse representation that’s both semantic and interpretable.

Architecture

Hybrid Use:

The resulting miniCOIL vector works seamlessly in hybrid search alongside dense vectors.

- Validation loss closely matches the theoretical lower bound
- Model generalizes across examples


miniCOIl sparse vector
miniCOIL
learned XYZ downprojector of "bat"
learned XYZ downprojector of "fruit"
0.7
0.1
0.5
-0.2
0.2
0.3
0.1
0.9
0.7 * BM25_score
0.1 * BM25_score
0.5 * BM25_score
-0.2 * BM25_score
0
0
0.2 * BM25_score
0.3 * BM25_score
0.1 * BM25_score
0.9 * BM25_score
0
0
Transformer XYZ of choice for inference
contextualized embedding of "bat"
contextualized embedding of "fruit"
"fruit bat"
dense XYZ embedding of a "fruit bat"
Hybrid search

MS MARCO

BM25 vs miniCOIL

NDCG@10

BM25: 0.237

miniCOIL: 0.244

Not about benchmarks...

DEMO

miniCOIL Strengths

Flexible Transformer Support

Compatible with various transformers used for inference in hybrid search

Simple Architecture

1 word = 1 small trainable model, enabling extremely fast inference

No Labeled Training Data

Avoids relevance objectives, reducing data collection and overfitting risks

Vocabulary Extendable

New words can be added to the vocabulary without retraining existing models

qdrant

Thierry Damiba

DATA SCIENTIST AT QDRANT

@ptdamiba

thierrypdamiba

thierrydamiba

td@qdrant.com