miniCOIL
Sparse Neural Retrieval Done Right

Evgeniya Sukhodolskaya
Developer Advocate,
Qdrant

About Me

- Developer Advocate @ Qdrant
- Data Engineering & Analytics
  TU München Masters
- Organizer of a "Bavaria, Advancements in SEarch Development" (#BASED) meetup in Munich

Textual Retriever: 2-for-1

Two Workhorses

🔵 Keyword Matching-Based Search

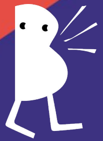
🔍 ” fruit **bat**”
📄 ” **bats** are **fruit** and leaves eaters”

🔍 ”**vectors** in **medicine**”
📄 ”**vector**: in **medicine**, a carrier of disease”

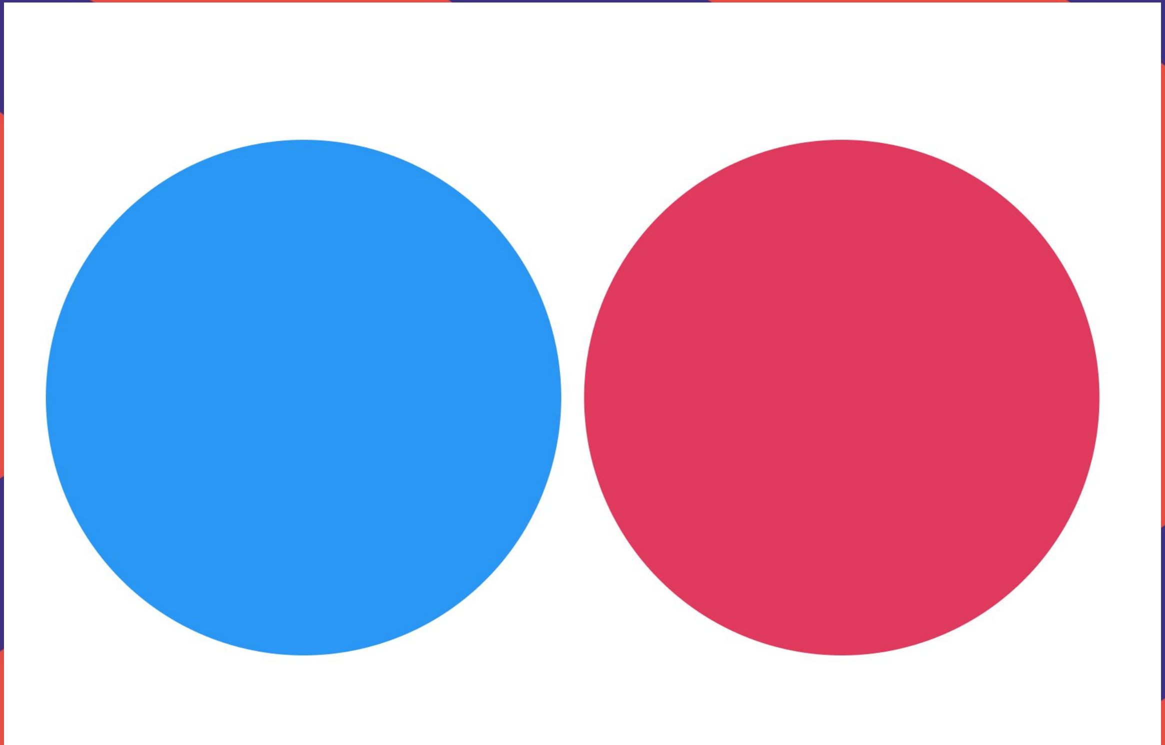
🔴 Semantic Similarity Search

🔍 ” fruit bat”
📄 ” flying foxes are the largest among their species”

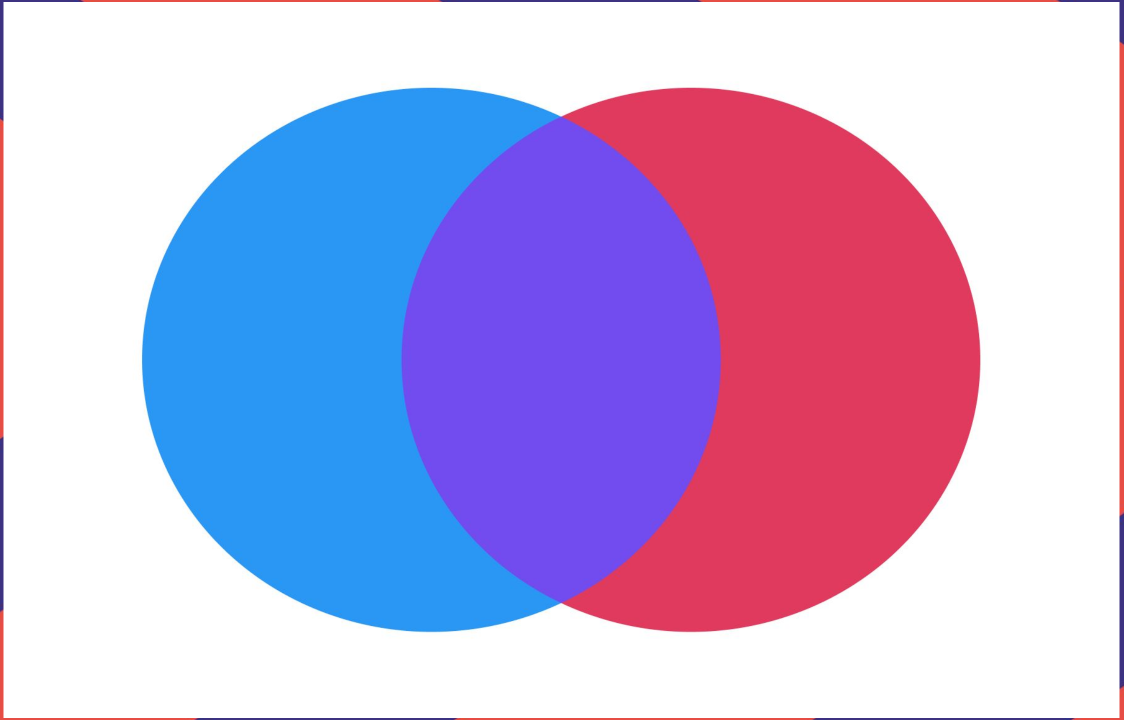
🔍 ” vectors in medicine”
📄 ”mosquito that transfers the infectious agent”



2-for-1: Hybrid Search

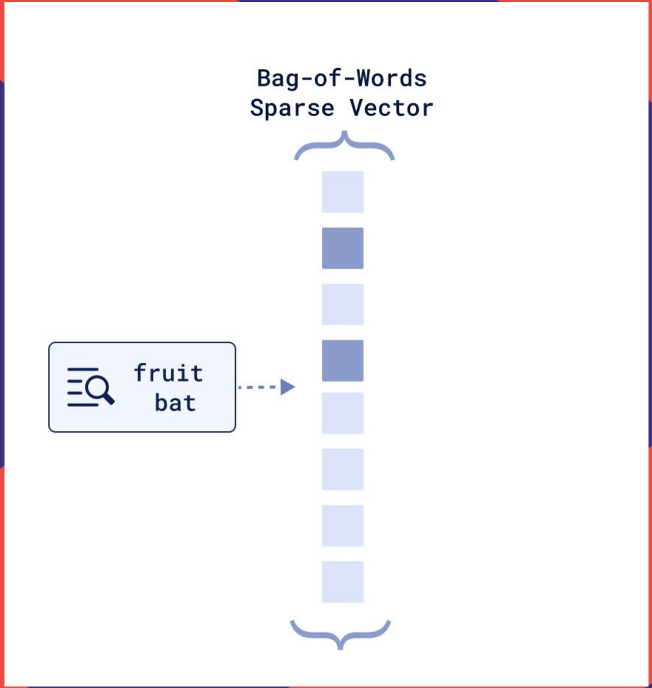


2-for-1: Sparse Neural Retrieval

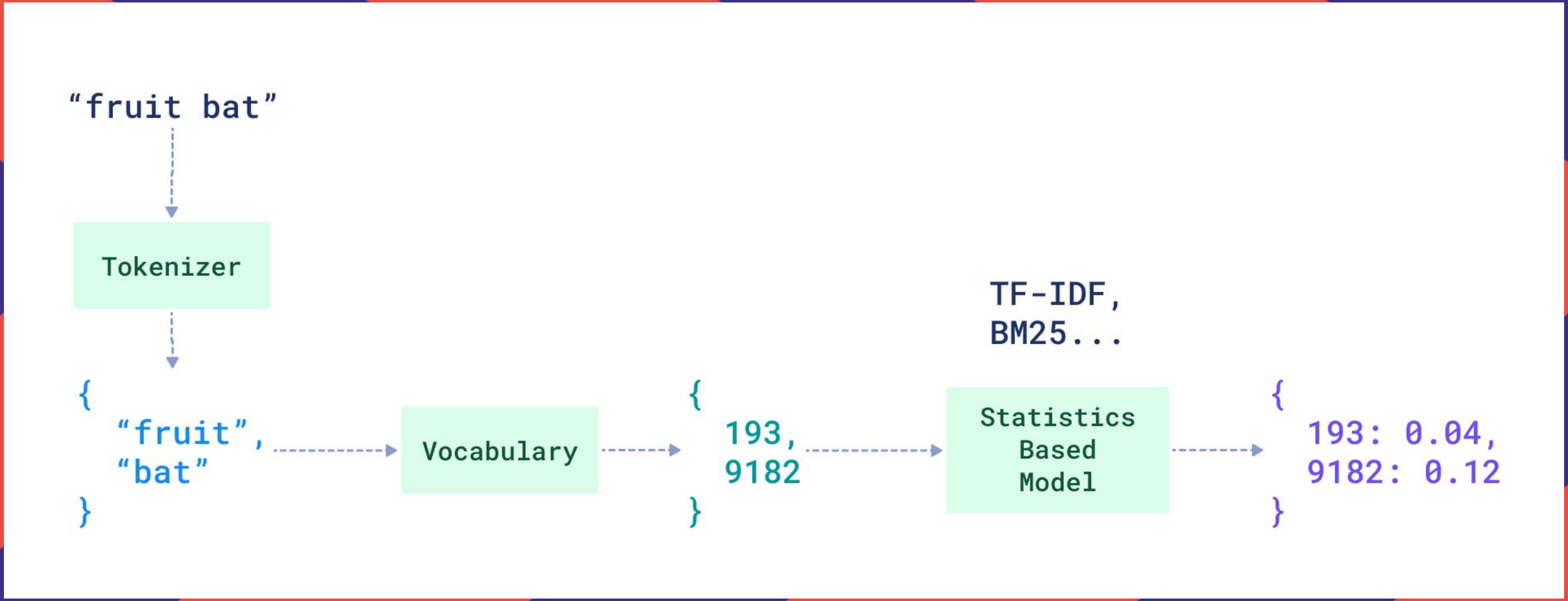


Sparse Neural Retrieval

Sparse Vectors



Sparse Retrieval



"fruit bat"

Tokenizer

```
{
  "fruit",
  "bat"
}
```

Vocabulary

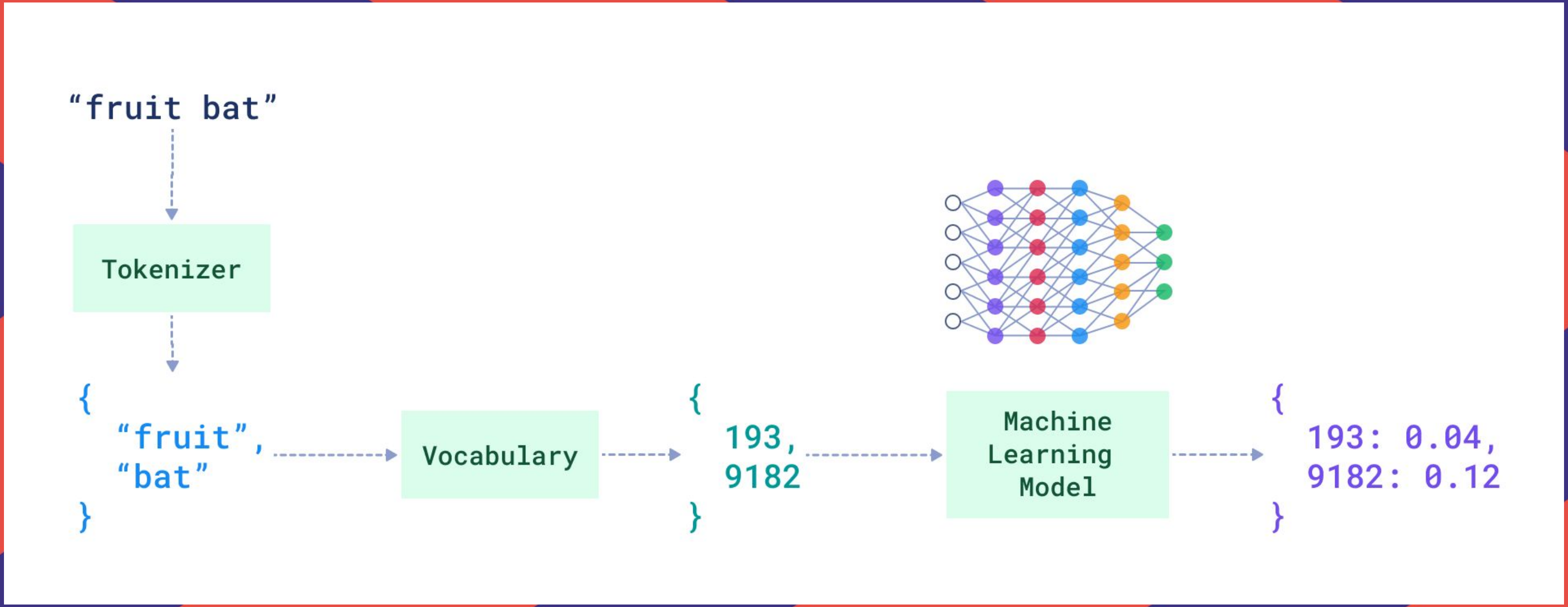
```
{
  193,
  9182
}
```

TF-IDF,
BM25...

Statistics
Based
Model

```
{
  193: 0.04,
  9182: 0.12
}
```

Sparse Neural Retrieval



"fruit bat"

Tokenizer

{ "fruit", "bat" }

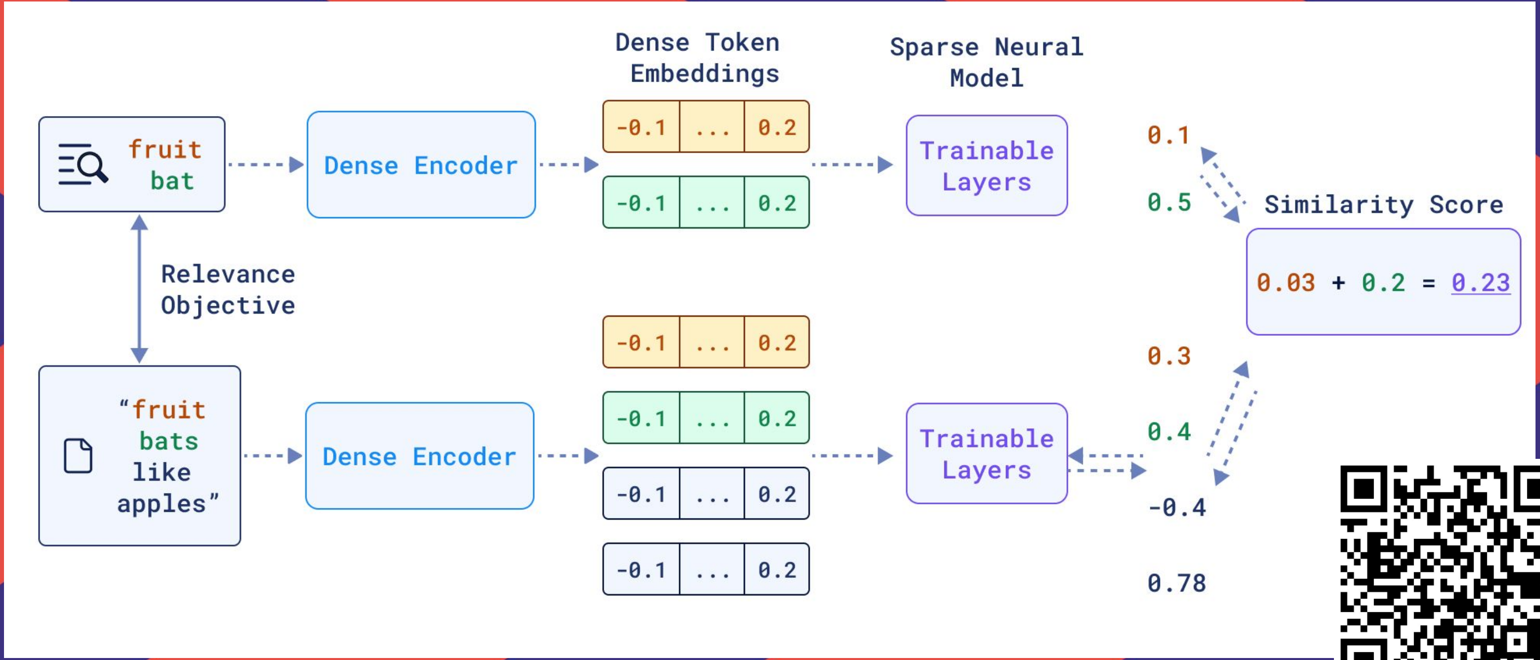
Vocabulary

{ 193, 9182 }

Machine Learning Model

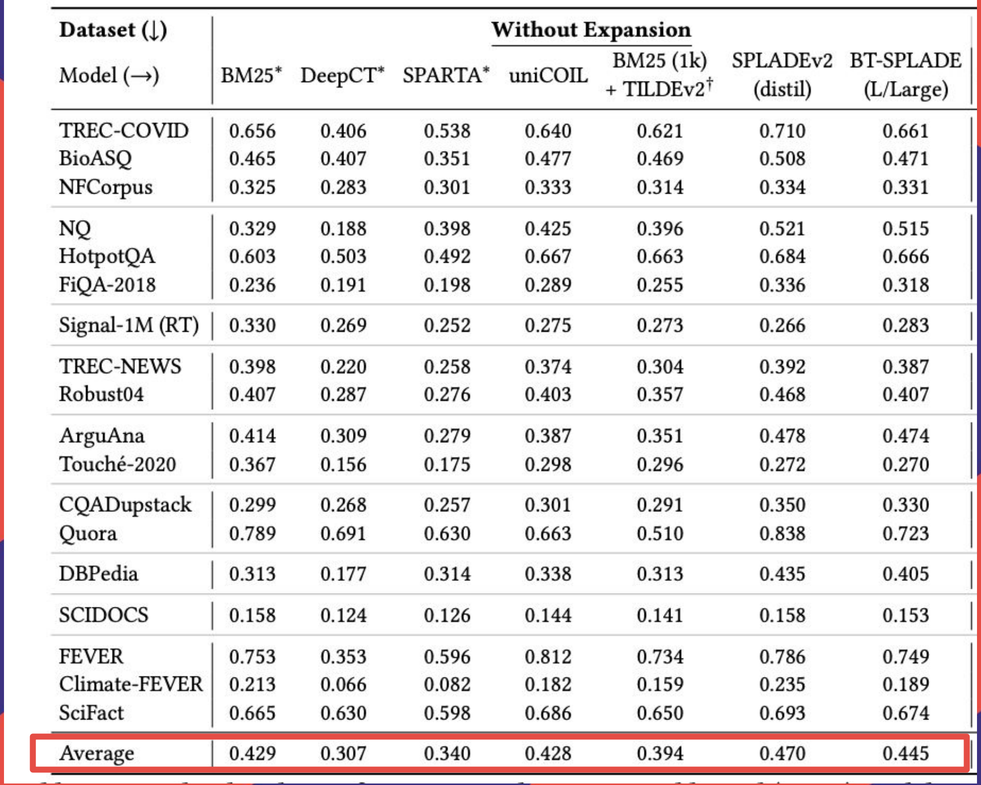
{ 193: 0.04, 9182: 0.12 }

Typical Sparse Neural Retriever



Modern Sparse Neural Retrieval

... Usually Doesn't Work



Dataset (↓)	Without Expansion						
Model (→)	BM25*	DeepCT*	SPARTA*	uniCOIL	BM25 (1k) + TILDEv2†	SPLADEv2 (distil)	BT-SPLADE (L/Large)
TREC-COVID	0.656	0.406	0.538	0.640	0.621	0.710	0.661
BioASQ	0.465	0.407	0.351	0.477	0.469	0.508	0.471
NFCorpus	0.325	0.283	0.301	0.333	0.314	0.334	0.331
NQ	0.329	0.188	0.398	0.425	0.396	0.521	0.515
HotpotQA	0.603	0.503	0.492	0.667	0.663	0.684	0.666
FiQA-2018	0.236	0.191	0.198	0.289	0.255	0.336	0.318
Signal-1M (RT)	0.330	0.269	0.252	0.275	0.273	0.266	0.283
TREC-NEWS	0.398	0.220	0.258	0.374	0.304	0.392	0.387
Robust04	0.407	0.287	0.276	0.403	0.357	0.468	0.407
ArguAna	0.414	0.309	0.279	0.387	0.351	0.478	0.474
Touché-2020	0.367	0.156	0.175	0.298	0.296	0.272	0.270
CQADupstack	0.299	0.268	0.257	0.301	0.291	0.350	0.330
Quora	0.789	0.691	0.630	0.663	0.510	0.838	0.723
DBPedia	0.313	0.177	0.314	0.338	0.313	0.435	0.405
SCIDOCS	0.158	0.124	0.126	0.144	0.141	0.158	0.153
FEVER	0.753	0.353	0.596	0.812	0.734	0.786	0.749
Climate-FEVER	0.213	0.066	0.082	0.182	0.159	0.235	0.189
SciFact	0.665	0.630	0.598	0.686	0.650	0.693	0.674
Average	0.429	0.307	0.340	0.428	0.394	0.470	0.445

“SPRINT: A Unified Toolkit for Evaluating and Demystifying Zero-shot Neural Sparse Retrieva”

What about SPLADE?

Beats BM25 due to its internally done **query & document expansion***

🔎 **”fruit bat fly fox vampire apple”**

📄 ”**fly**ing **fox**es are the largest among their species”

📄 ”**bats** are **fruit** and leaves eaters”

📄 ”**vampire**s hate **apple**s”

*which is HEAVY

Unusable Sparse Neural Retrieval

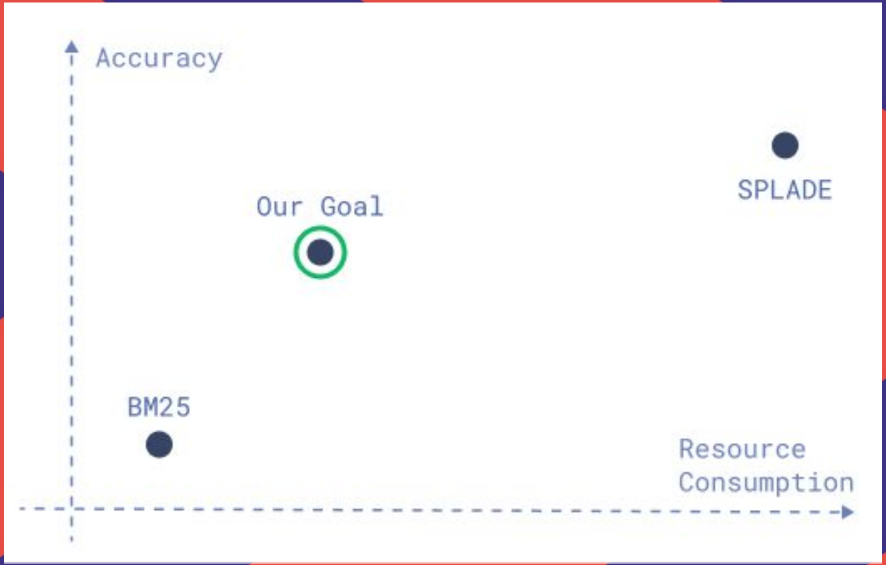
So, **most modern sparse neural retrievers** are:

- **not performant out-of-domain**
  (usually due to dependency on the **relevance objective**)

- or **not lightweight/truly sparse/explainable**
  (usually due to dependency on **document expansion**)

On the Road to Usable Sparse Neural Retrieval

Direction

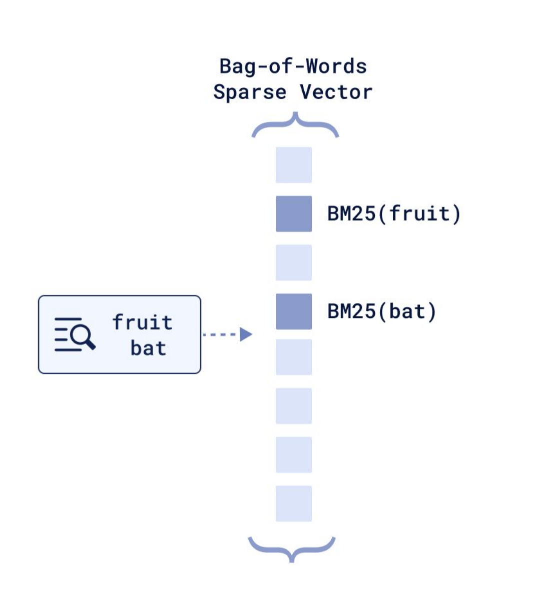


❎ Search Broader

✅ Rank Better

If it Works, it Works

$$\mathrm{BM25}(\mathrm{word}) = \mathrm{IDF}(\mathrm{word}) \times f_{\mathrm{corpus\ params}}(\mathrm{TF}(\mathrm{word}))$$



Bag-of-Words
Sparse Vector

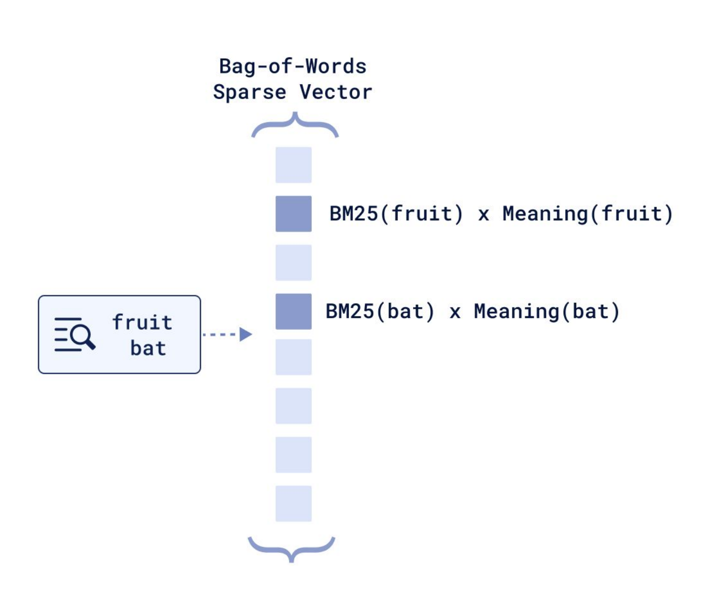
fruit
bat

BM25(fruit)

BM25(bat)

What if BM25 Understood Meaning?

$$\text{BM25(word)} = \text{IDF(word)} \times f_{\text{corpus params}}(\text{TF(word)}) \times \text{Meaning(word)}$$



Bag-of-Words Sparse Vector

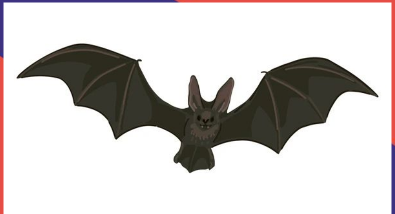
fruit bat

BM25(fruit) x Meaning(fruit)

BM25(bat) x Meaning(bat)

BAT
GERMANY

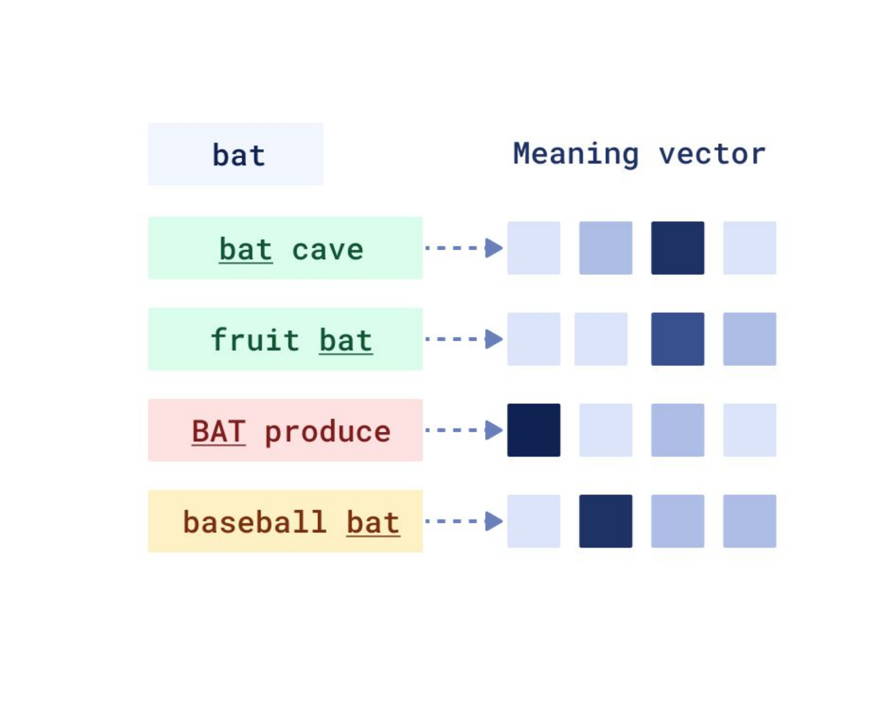
How to Express Meaning?

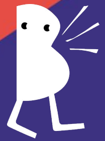


BAT
GERMANY

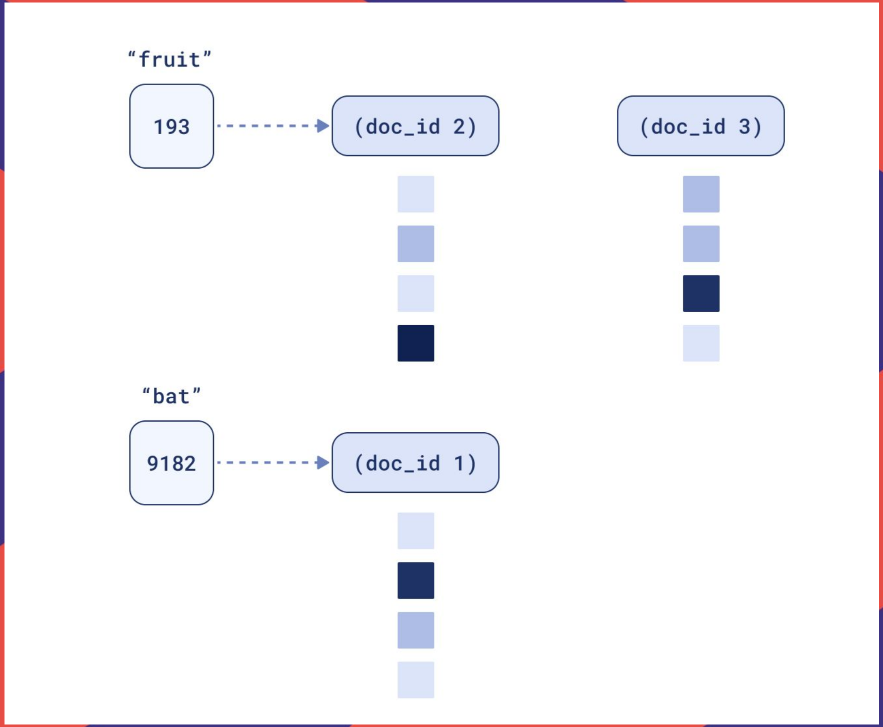
“bat”: 0.08

How to Express Meaning?

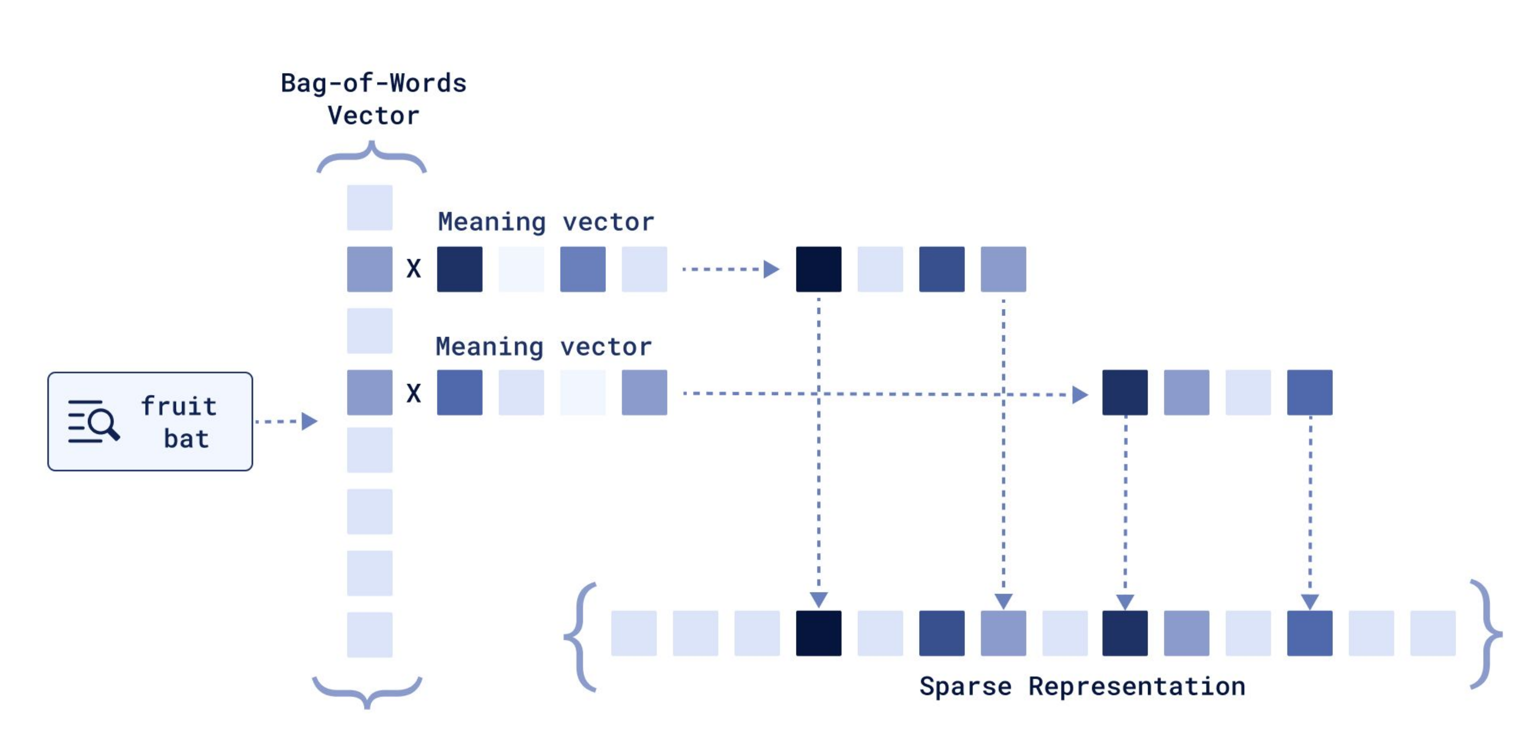




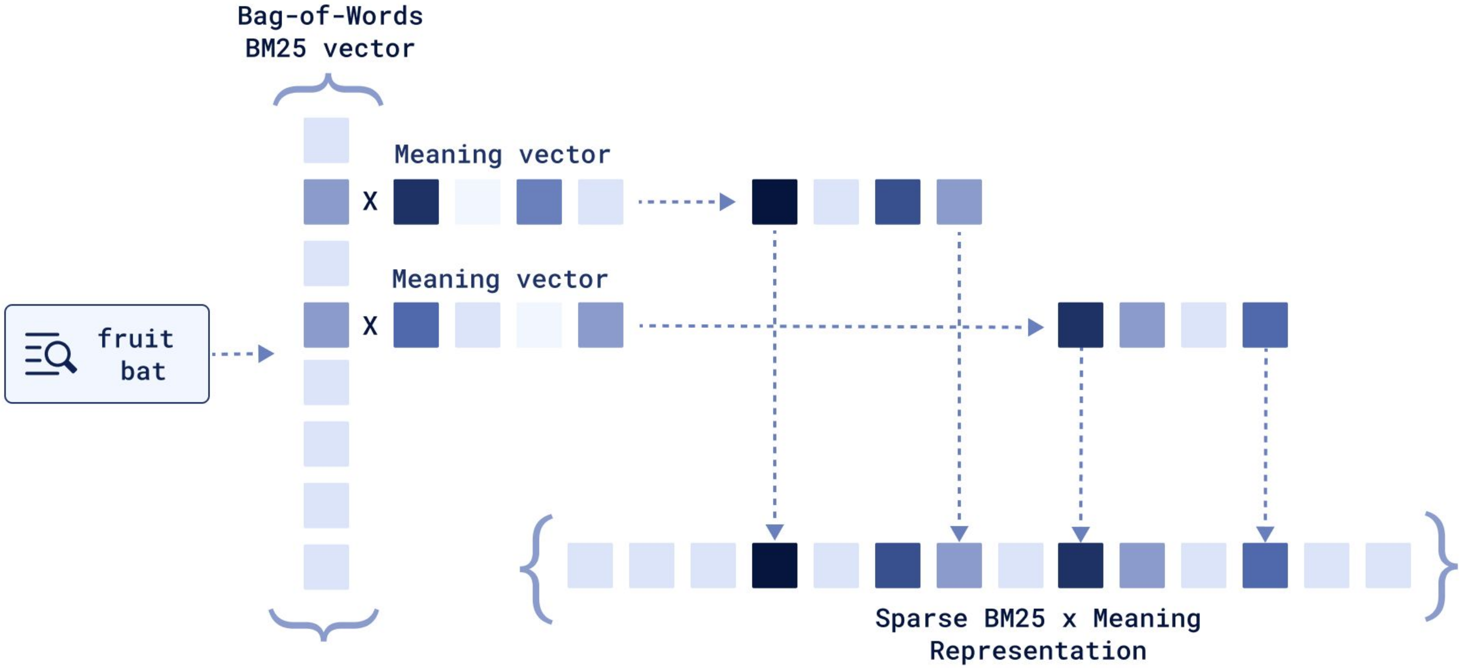
Contextualized Inverted Lists



Flattening COIL

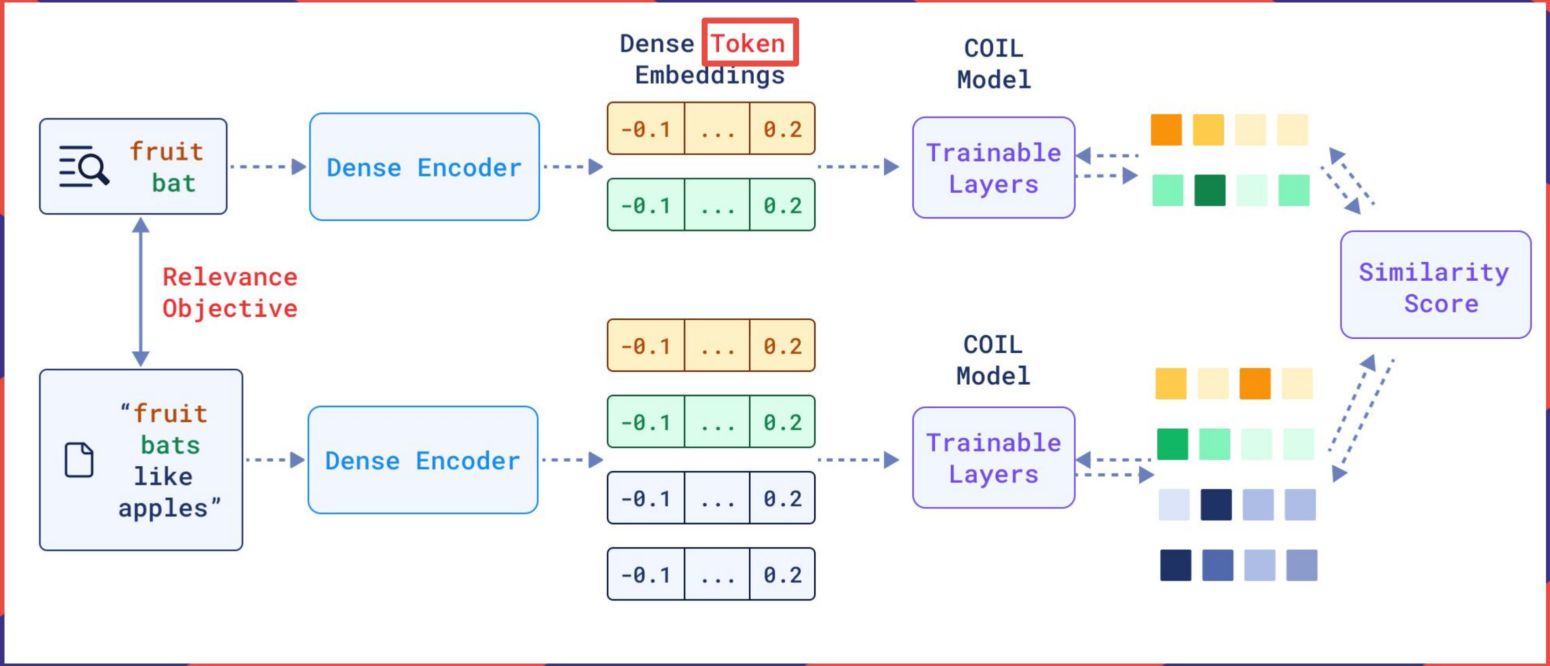


Adding Meaning to BM25



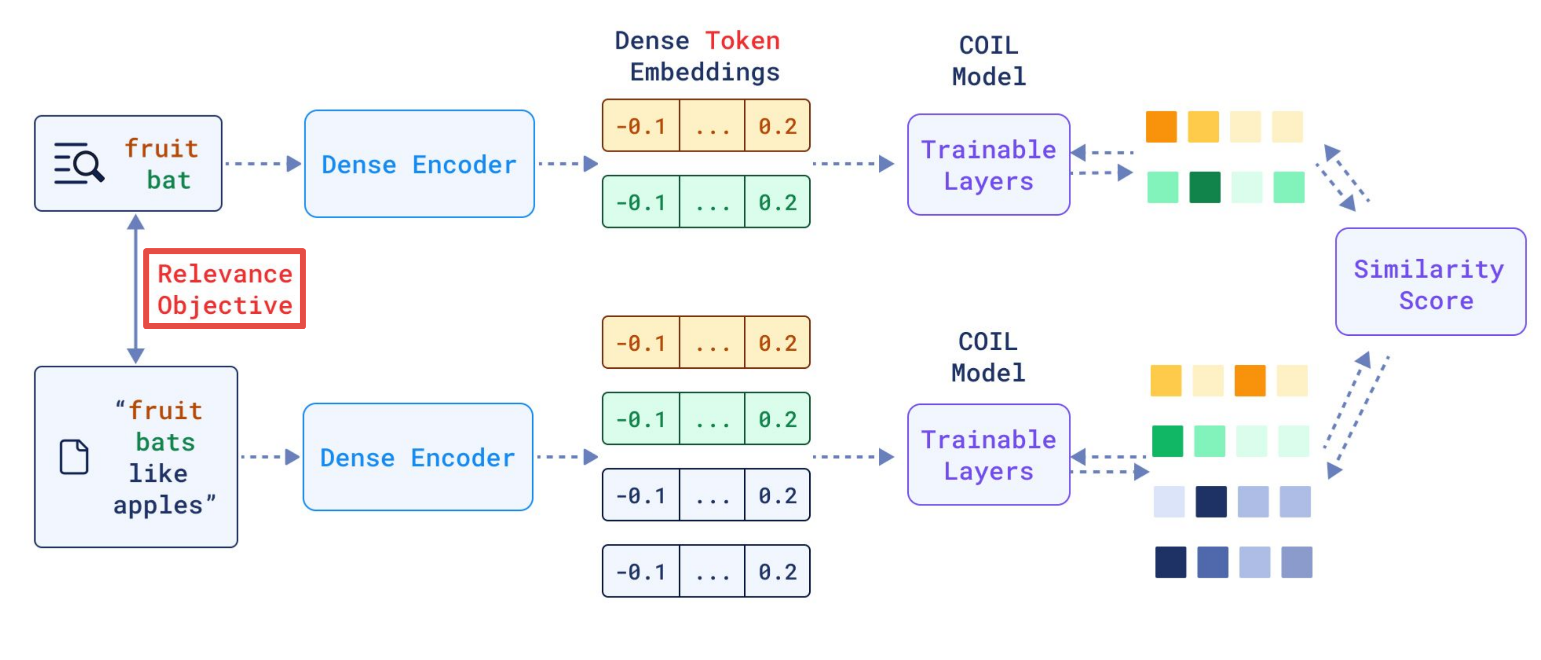
mini COIL != miniCOIL

miniCOIL != mini COIL #1



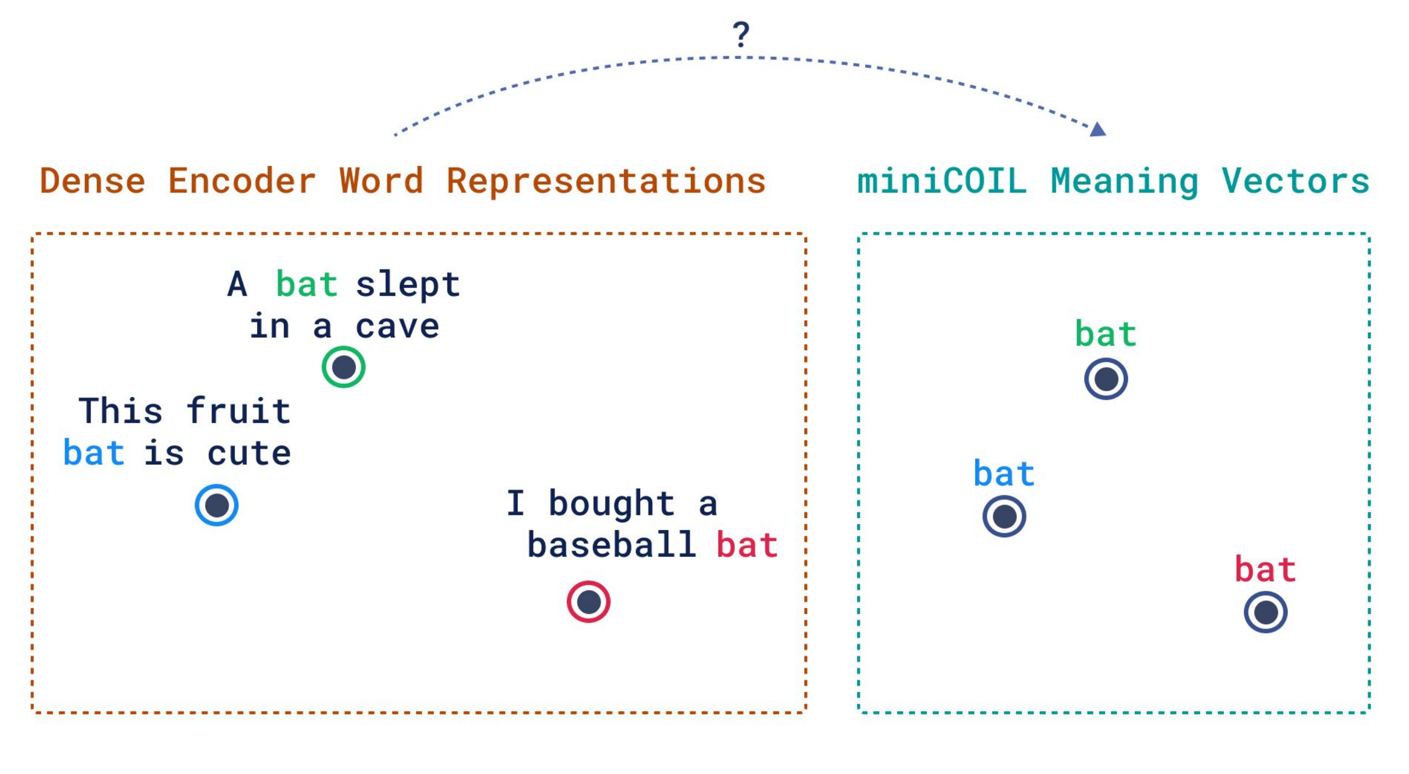
tokens -> words (word stems)

miniCOIL != mini COIL #2

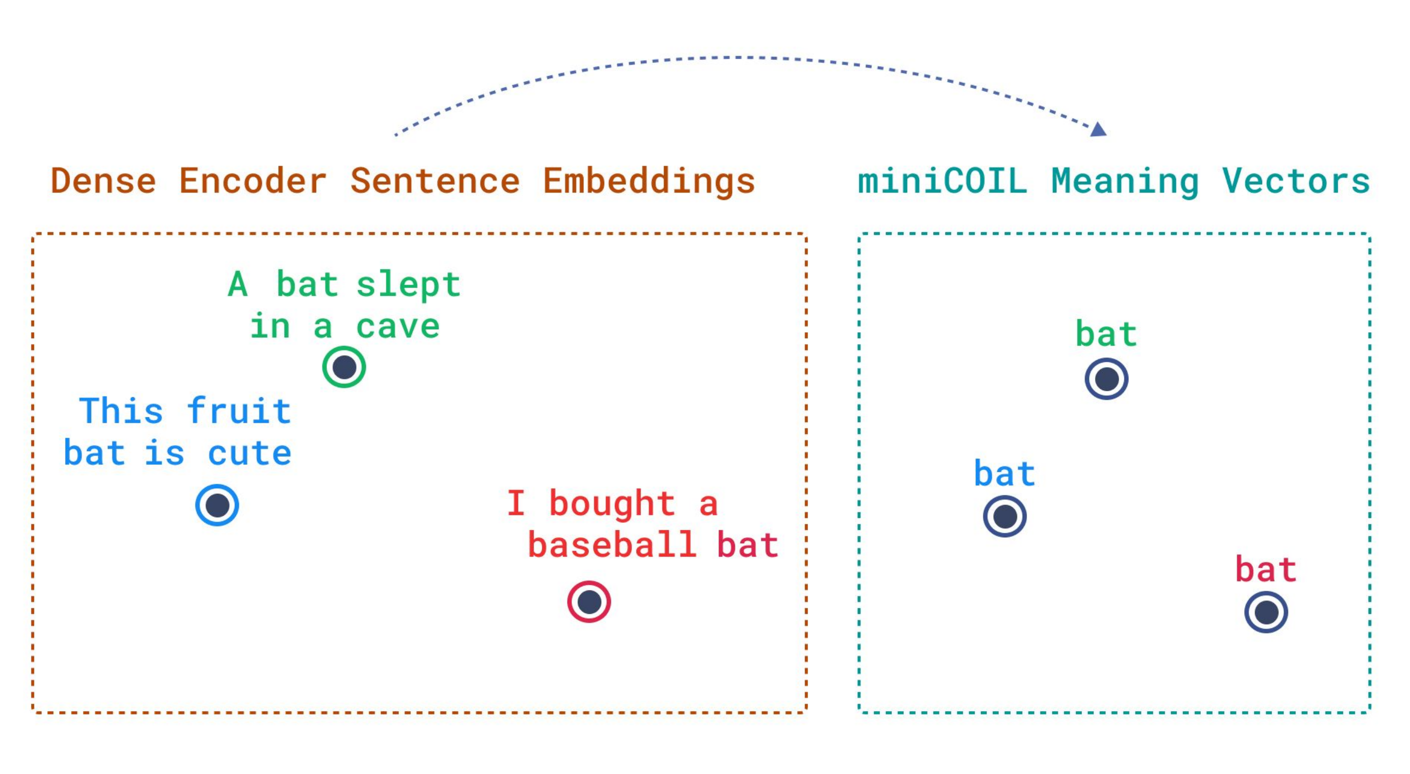


Getting rid of relevance objective

Getting Rid of Relevance Objective



Getting Rid of Relevance Objective



Dense Encoder Sentence Embeddings

A bat slept in a cave

This fruit bat is cute

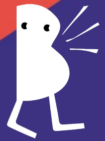
I bought a baseball bat

miniCOIL Meaning Vectors

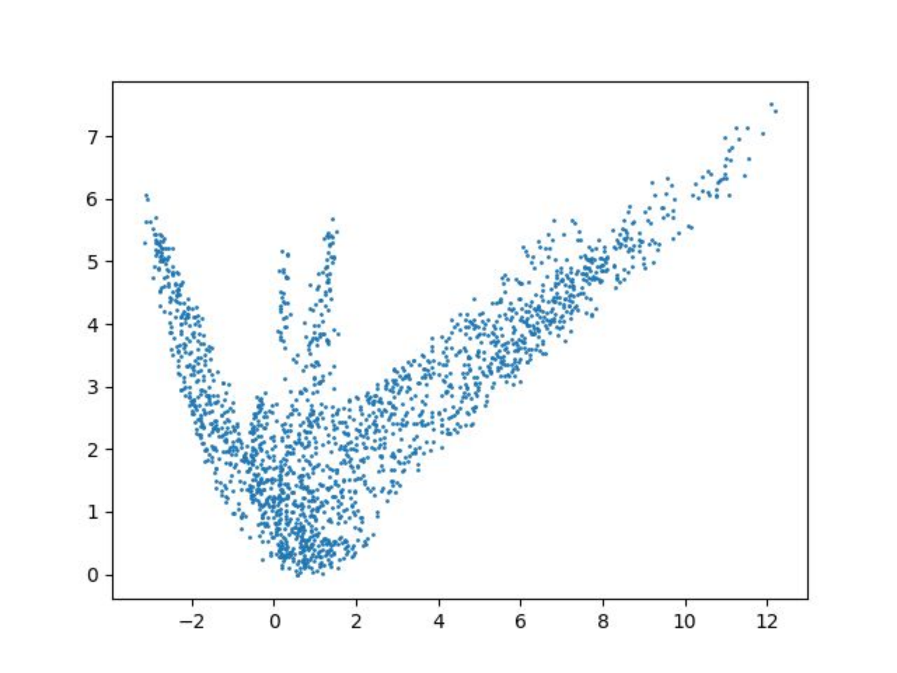
bat

bat

bat

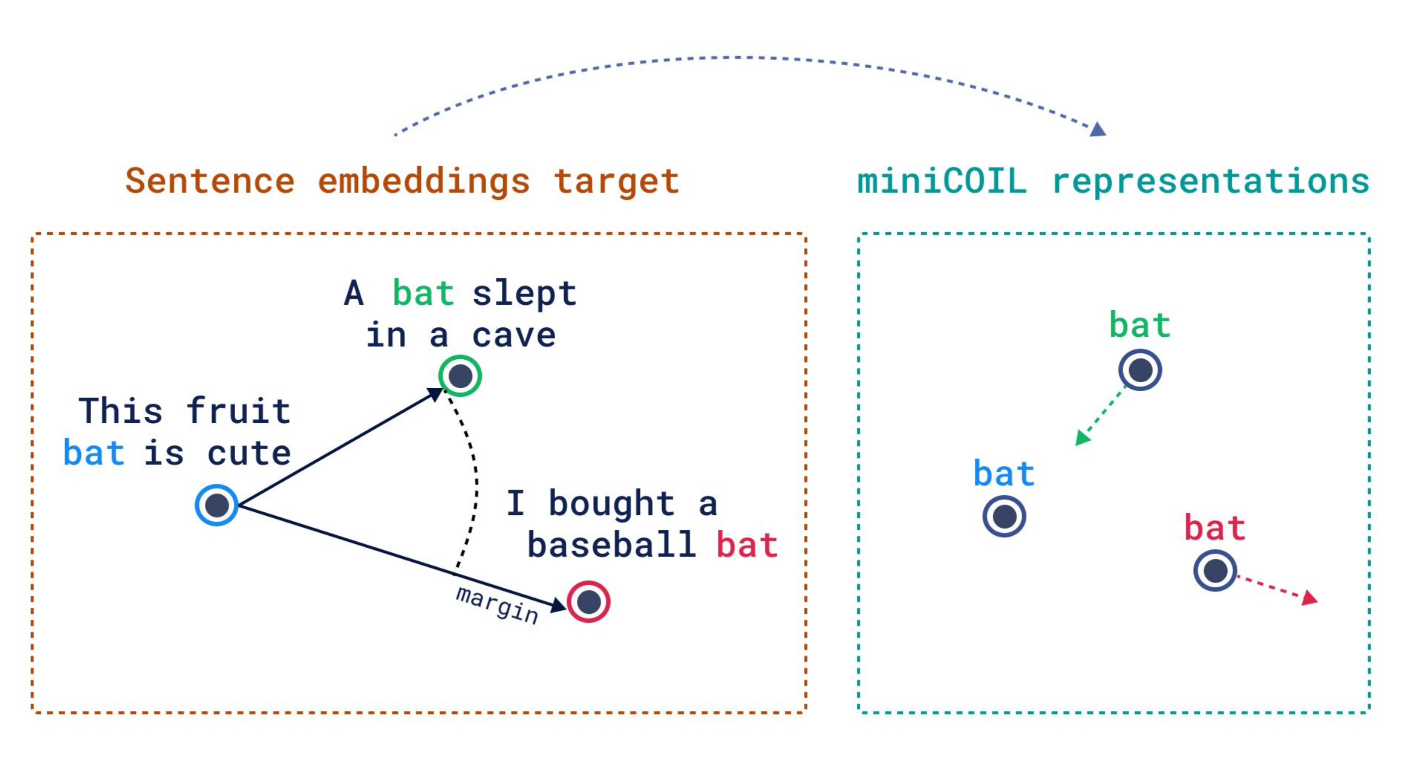


Will it Work? I Bat!

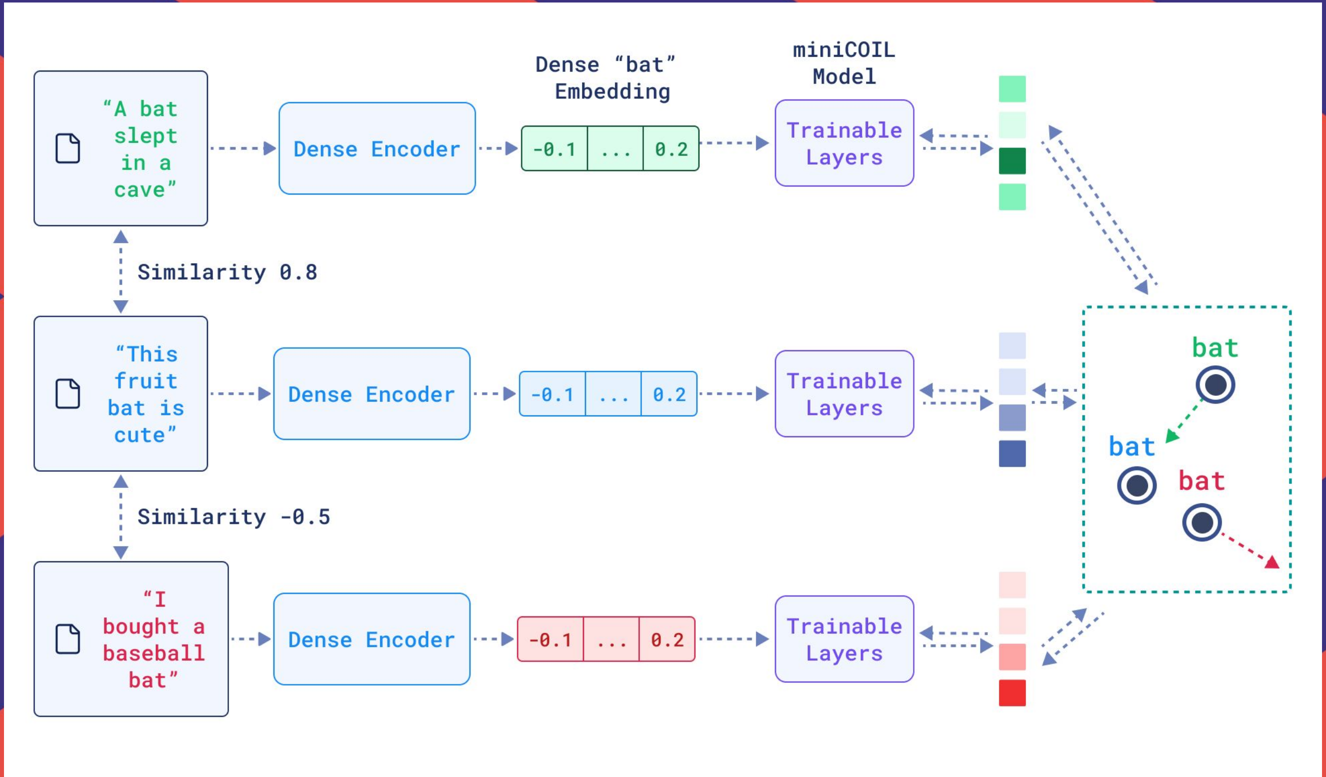


~4k "*bat*" sentences embedded with **mxbai-embed-large-v1-large**

miniCOIL Training Objective



miniCOIL Training Objective



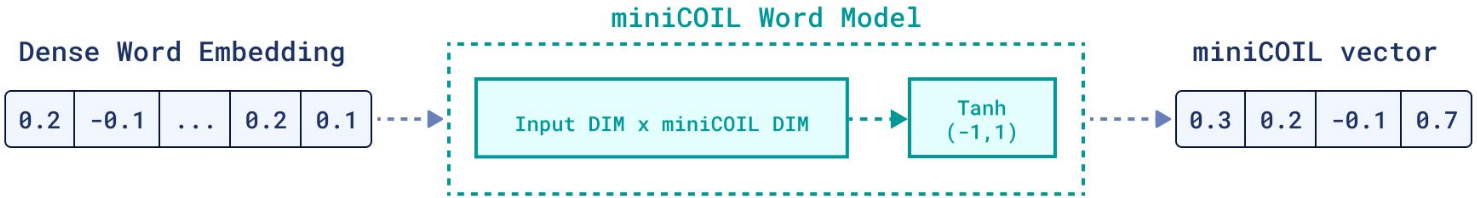
miniCOIL Training Data

40 mln Sentences from OpenWebText

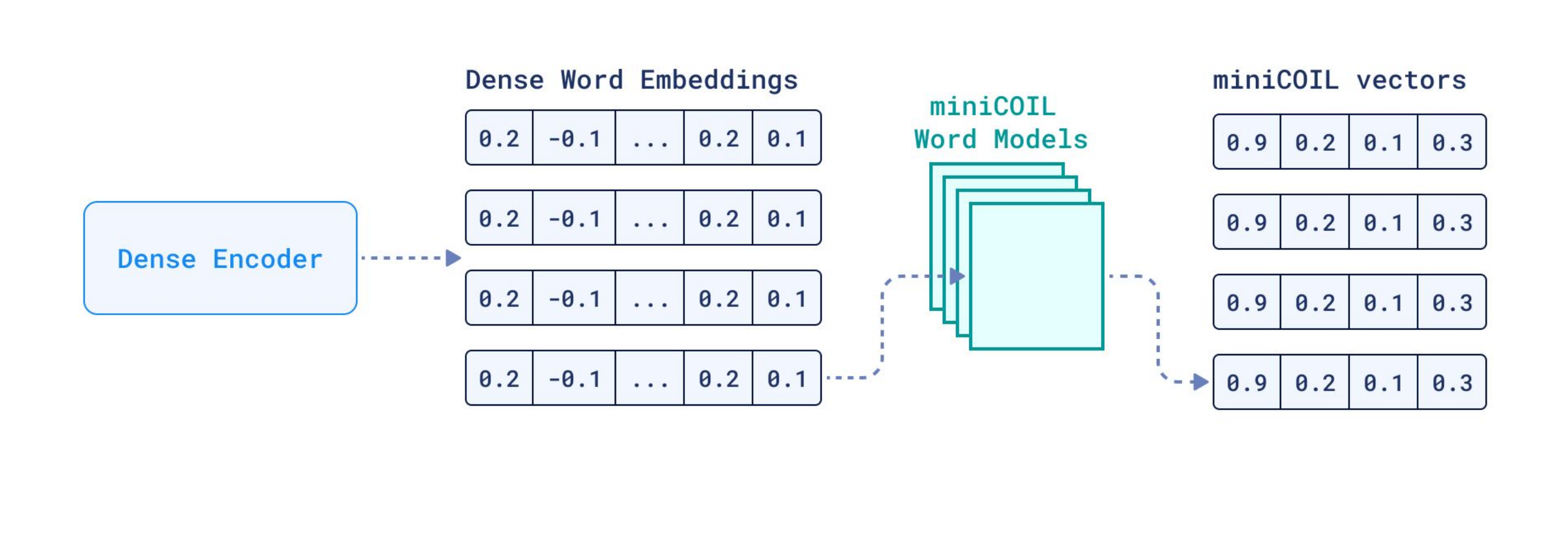
Stored in **Qdrant** to **sample triplets** for training,
using **Distance Matrix API** and **full-text index filtering**

miniCOIL Architecture

Deal with One Word



... and Repeat

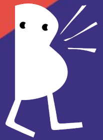

Dense Encoder
Dense Word Embeddings
0.2 | -0.1 | ... | 0.2 | 0.1
0.2 | -0.1 | ... | 0.2 | 0.1
0.2 | -0.1 | ... | 0.2 | 0.1
0.2 | -0.1 | ... | 0.2 | 0.1
miniCOIL
Word Models
miniCOIL vectors
0.9 | 0.2 | 0.1 | 0.3
0.9 | 0.2 | 0.1 | 0.3
0.9 | 0.2 | 0.1 | 0.3
0.9 | 0.2 | 0.1 | 0.3

miniCOIL v1

miniCOIL v1

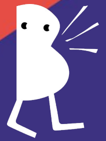
miniCOIL v1 on HuggingFace

Component	Description
Input Dense Encoder	jina-embeddings-v2-small-en
miniCOIL Vectors Size	4 dimensions
miniCOIL Vocabulary	30,000 of the most common English words
Training Data per Word	8,000 sentences per word + augmentation



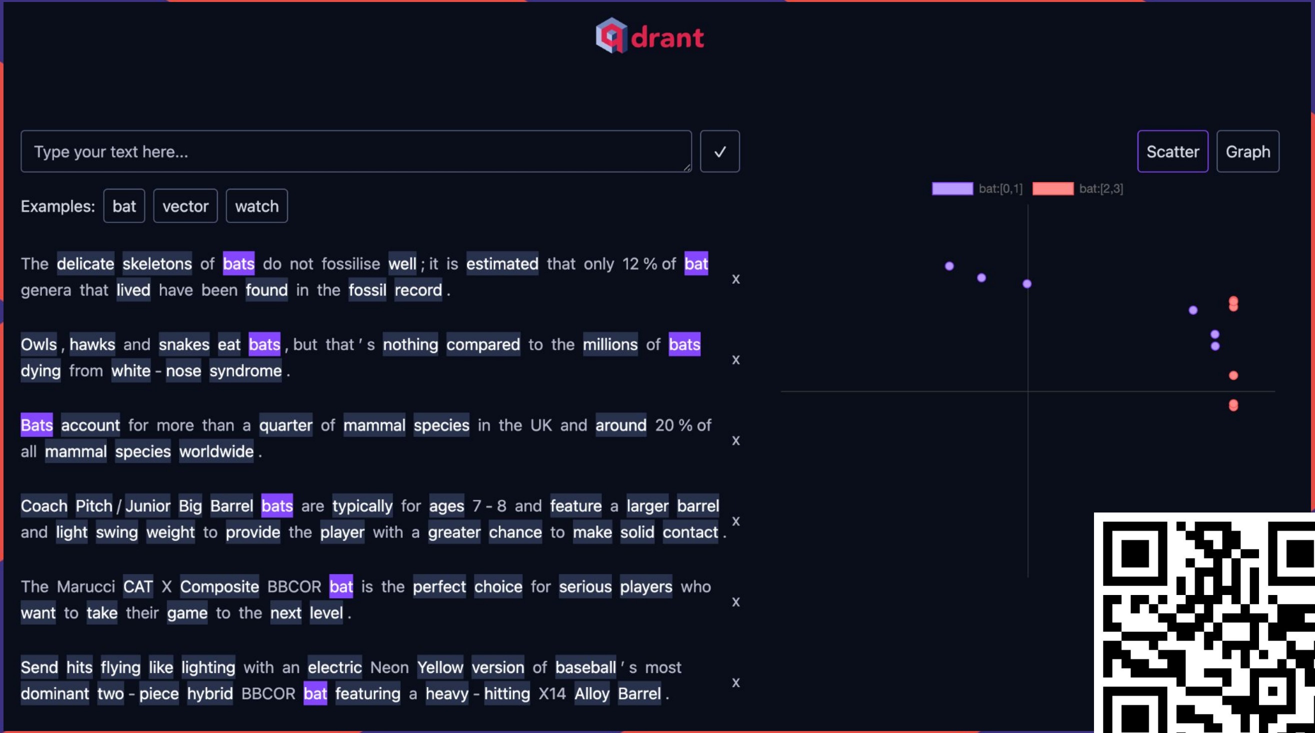
miniCOIL v1 Benchmarks

Dataset	**BM25 (NDCG@10)**	**MiniCOIL (NDCG@10)**
MS MARCO	0.237	**0.244**
NQ	0.304	**0.319**
Quora	0.784	**0.802**
FiQA-2018	0.252	**0.257**
HotpotQA	**0.634**	0.633



Qdrant

miniCOIL v1 Demo



The delicate skeletons of bats do not fossilise well ; it is estimated that only 12 % of bat genera that lived have been found in the fossil record . x

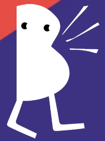
Owls , hawks and snakes eat bats , but that ' s nothing compared to the millions of bats dying from white - nose syndrome . x

Bats account for more than a quarter of mammal species in the UK and around 20 % of all mammal species worldwide . x

Coach Pitch / Junior Big Barrel bats are typically for ages 7 - 8 and feature a larger barrel and light swing weight to provide the player with a greater chance to make solid contact . x

The Marucci CAT X Composite BBCOR bat is the perfect choice for serious players who want to take their game to the next level . x

Send hits flying like lighting with an electric Neon Yellow version of baseball ' s most dominant two - piece hybrid BBCOR bat featuring a heavy - hitting X14 Alloy Barrel . x



minicoil.qdrant.tech

Using miniCOIL v1

When to Use miniCOIL v1

✅ You need something **like BM25 but ranking with semantic understanding**

🔍 ”**vectors** in medicine”
⬆️ 📄 ”**vector** control strategies in public health”
⬇️ 📄 “advanced **vector** calculus for engineers”

❎ You need matches that are similar in meaning but expressed differently

🔍 “vectors in medicine”
📄 ”mosquito that transfers the infectious agent”

How to Use miniCOIL v1

Hybrid Search with miniCOIL v1

Takeaways

Justifying “Done Right”

Benefits of the miniCOIL approach:

- Built on top of reliable BM25, so a safe fallback for untrained words;
- Easy to train (1 CPU for a word-level model),
  easy to extend (train only needed words), lightweight;
- Outputs fit the classic inverted index;
- Makes sense in a hybrid search scenario – reuses dense inference output, adding a one lightweight trained layer on top.

Q&A Time (*if There's Any*😅)

Thank you for bearing with me!

(and my QR code madness)
(here is one more, though -> -> ->)

My LinkedIn, in case you want the presentation to deal with QR codes in peace:D