Two-Step SPLADE

Carlos Lassance (now at Cohere)
Hervé Dejean
Stéphane Clinchant
Nicola Tonellotto

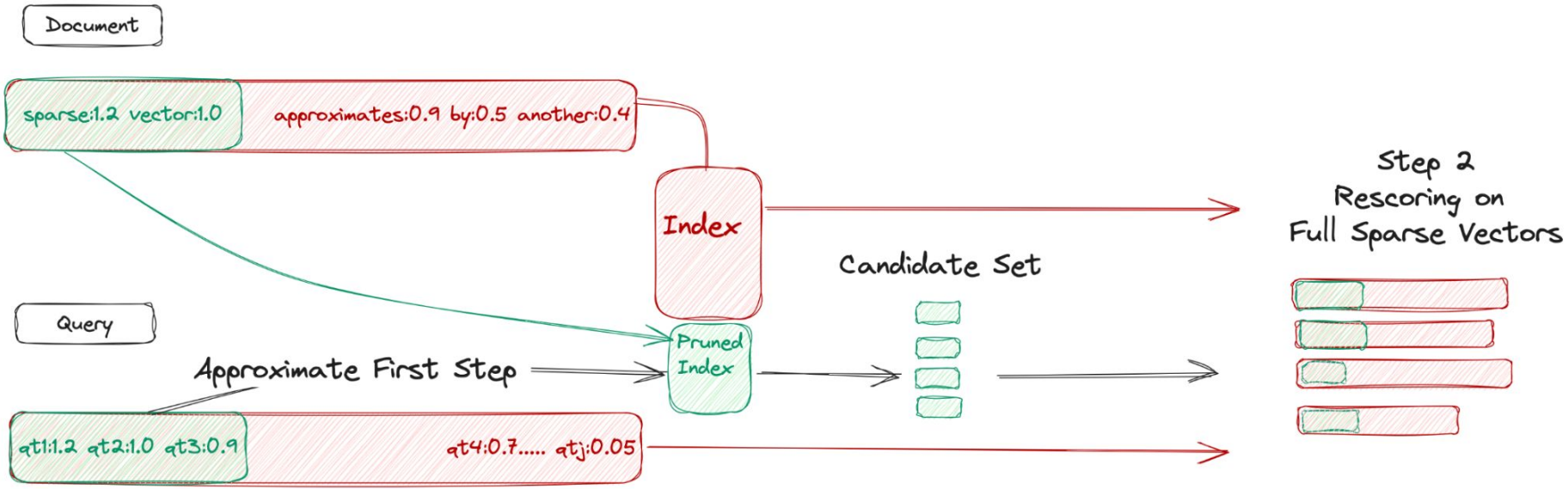
NAVER LA
E

UNIVERSITÀ DI PISA

Two-Step SPLADE

- **What is the best we can do with a fixed model and tools?**
  - Pruning helps, but not so much
- Take lessons from dense
  - Approximate search (First Step) + Rescoring (Second Step)
  - Two-Step
- TLDR:
  - **Pruning** and **saturation** leads to a **good approximation** of the **SPLADE** vector
- So simple, ElasticSearch went with this option concurrently to us
  - https://www.elastic.co/search-labs/blog/articles/introducing-elser-v2-part-2

Two-Step SPLADE



Two-Step SPLADE (ElasticSearch)

```
GET /my-index/_search
{
  "query": {
    "bool": {
      "should": [
        {"term": {"tokens": {"value": "<kept token 1>", "boost": <kept weight 1>}}},
        {"term": {"tokens": {"value": "<kept token 2>", "boost": <kept weight 2>}}},
        …
      ]
    }
  },
  "rescore": {
    "window_size": 50,
    "query" : {
      "rescore_query" : {
        "bool": {
          "should": [
            {"term": {"tokens": {"value": "<dropped token 1>", "boost":<dropped weight 1>}}},
            {"term": {"tokens": {"value": "<dropped token 2>", "boost":<dropped weight 2>}}},
            …
          ]
        }
      },
      "query_weight" : 1,
      "rescore_query_weight" : 1
    }
  }
}
```

non-frequent or critical tokens

frequent and non-critical tokens

score = 1 x ("query kept token 1 weight" x "doc kept token 1 weight" +
"query kept token 2 weight" x "doc kept token 2 weight" + …) +
1 x ("query dropped token 1 weight" x "doc dropped token 1 weight" +
"query dropped token 2 weight" x "doc dropped token 2 weight" + …)

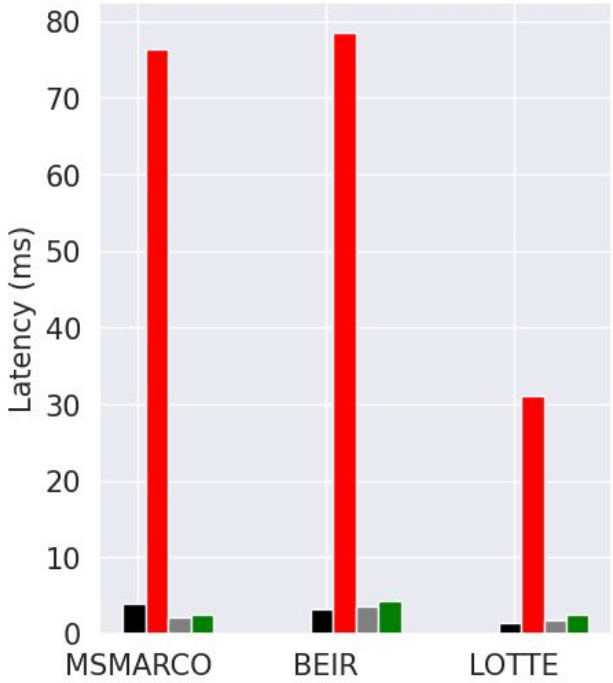
How is it different from the state of art?

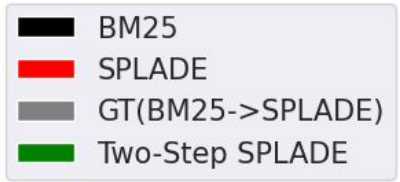
- Completely redesign training for efficiency
  - Efficient-SPLADE (Ours, SIGIR 2022) -> Bad OOD, needs retraining

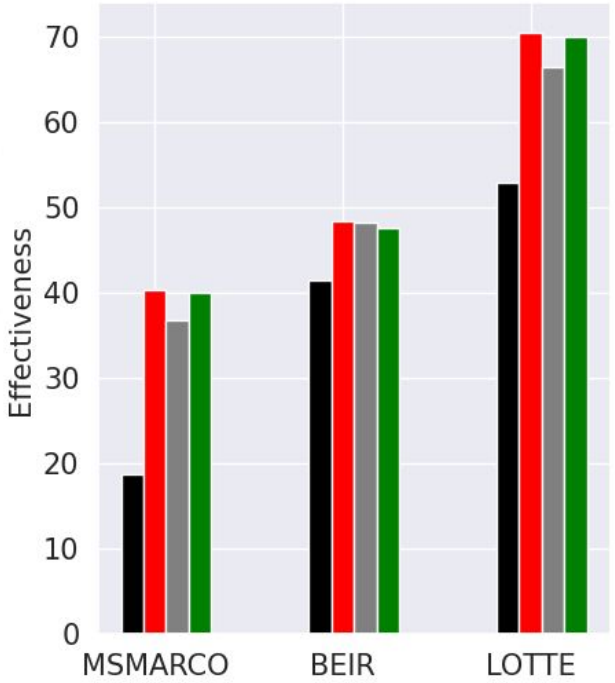
- Approximate Sparse retrieval
  - Pruning (Ours, SIGIR 2023) -> Minor efficiency gains
  - Sketching (Bruch et al ACM 2023) -> Needs new tooling
  - NeurIPS 2023 Big ANN Benchmark -> Hard to add new documents

- Modify the inverted index algorithm -> Need new tooling
  - **Guided Traversal** (Mallia et al, SIGIR 22) -> Our main point of comparison
  - Guided Traversal ++ (Qiao et al, WWW and SIGIR 23)
  - Postings Clippings (Mackenzie et al, EMNLP 22)

Does it work in practice? (average)







Does it work in practice? (relative, 30 datasets)

	Method	Effect size against				MSMARCO		BEIR		Lotte
		SPLADE (b)		GT (d)		Latency		18	>1M	
		≥ (>)	<	≥ (>)	<	Average	p99	AvG	L	AvG L
	Baselines									
a	BM25	7 (1)	23	7 (1)	23	1.0	1.0	1.0	1.0	1.0
b	SPLADE-v3	N/A		27 (16)	3	19.1	12.4	24.8	32.6	22.1
	Advanced Baselines									
c	Approx. First Step (Pruning) over (b)	7 (1)	23	16 (2)	14	0.7	0.4	2.3	2.6	2.6
d	GT (Our Implementation) ($a \rightarrow b$)	14 (3)	16	N/A		1.1	1.0	1.2	1.2	1.3
	This work									
	Two-Step ($c \rightarrow b$)	22 (0)	8	26 (15)	4	0.8	0.4	2.5	2.7	3.0

See you at the poster section

For code:
https://github.com/carlos-lassance/splade/tree/two_step/two_step