Two-Step SPLADE

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

Document

sparse1.2 vector:1.0 approximates:0.9 by:0.5 another:0.4

Query

Approximate First Step

Pruned Index

Index

Candidate Set

Step 2
Rescoring on Full Sparse Vectors
Two-Step SPLADE (ElasticSearch)

```json
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
        }
    }
}
```
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)

<table>
<thead>
<tr>
<th>Method</th>
<th>Effect size against</th>
<th>MSMARCO</th>
<th>BEIR</th>
<th>Lotte</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SPLADE (b) ≥ (&gt; ) &lt;</td>
<td>Latency Average p99</td>
<td>18 &gt;1M AvG L</td>
<td>AvG L</td>
</tr>
<tr>
<td></td>
<td>GT (d) ≥ (&gt; ) &lt;</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baselines</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a</td>
<td>BM25</td>
<td>7 (1) 23</td>
<td>7 (1) 23</td>
<td>1.0</td>
</tr>
<tr>
<td>b</td>
<td>SPLADE-v3</td>
<td>N/A 27 (16) 3</td>
<td>19.1</td>
<td>12.4</td>
</tr>
<tr>
<td>Advanced Baselines</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>c</td>
<td>Approx. First Step (Pruning) over (b)</td>
<td>7 (1) 23</td>
<td>16 (2) 14</td>
<td>0.7</td>
</tr>
<tr>
<td>d</td>
<td>GT (Our Implementation) (a → b)</td>
<td>14 (3) 16</td>
<td>N/A</td>
<td>1.1</td>
</tr>
<tr>
<td>This work</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Two-Step (c → b)</td>
<td>22 (0) 8</td>
<td>26 (15) 4</td>
<td>0.8</td>
<td>0.4</td>
</tr>
</tbody>
</table>
See you at the poster section

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