Full-Text Search Explained

Philipp Krenn @xeraa

elastic


elastic


elastic


Eurovision
elastic

elastic

Infrastructure | Developer Advocate

ViennaDB

Papers We Love Vienna

elastic

Who uses Databases?

Who uses Search?

Database

vs

Full-Text Search

elastic


elastic

But I can do...

```
SELECT *
  FROM my_table
 WHERE my_text LIKE ‘%my_term%’
```

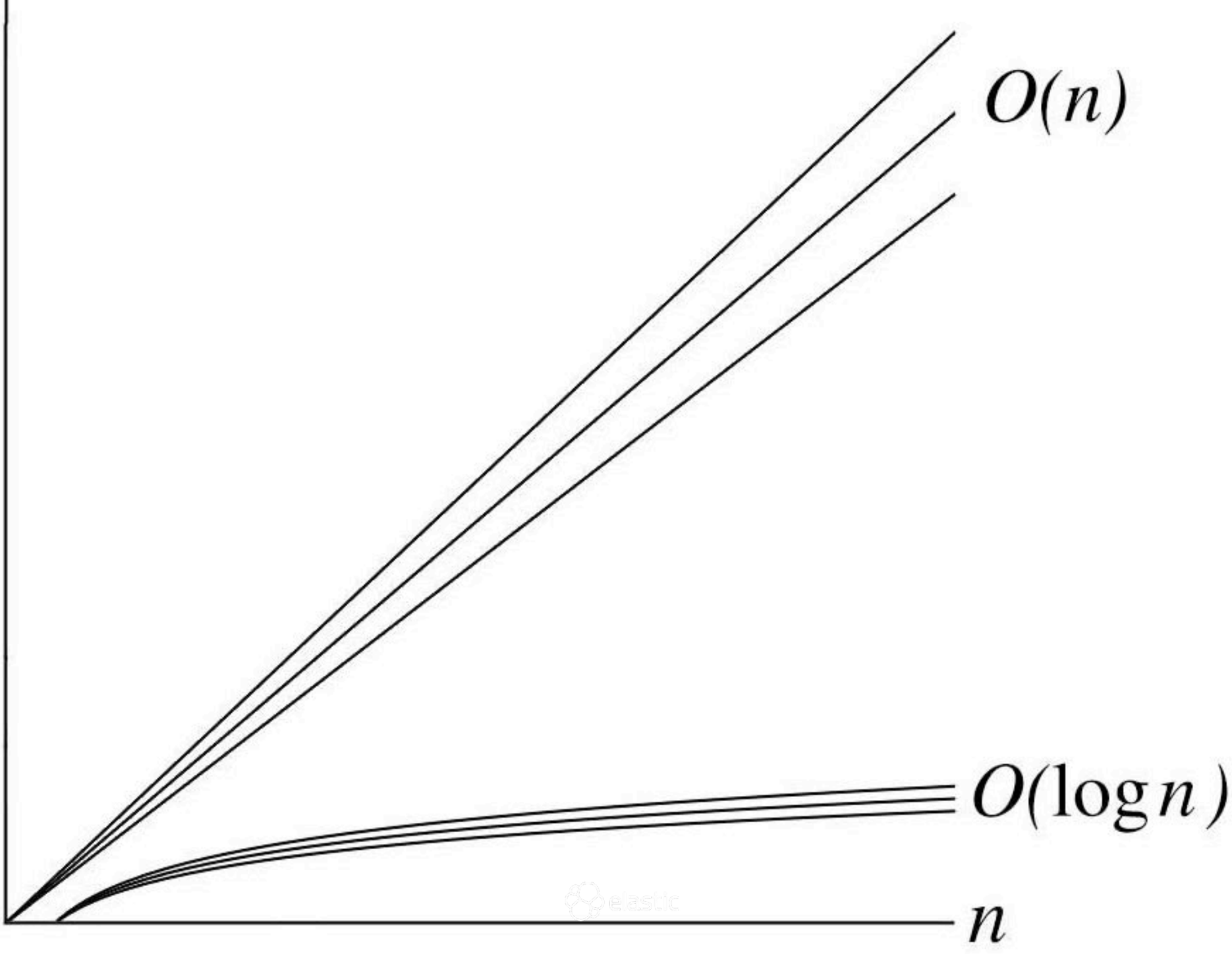
elastic

1. Performance

B-Tree

elastic


Help your customers find you.
KINCAID & SON CONSTRUCTION CO INC
243-8644
569-1290
569-1291
694-3199
688-2633
562-8708
562-5983
245-2399
337-4884
522-1616
277-8972
929-3900
LITTLE SUSITNA CONSTRUCTION COMPANY
694-8565
694-6364
694-6363
644-8323
688-9643
LONGSTAFF ENGINEERS & CONSTRUCTION
274-7571
277-7574
NANUQ INC
NEESER CONSTRUCTION INC
646-
276-10
563-561
563-5615
242-9163
929-2784
229-7979
258-0042
563-3178
276-1444
OSBORNE CONSTRUCTION CO


O(n)
O(log n)
n


AS EXPECTED WORKS
memegenerator.net

2. Features

Fuzziness, synonyms, scoring,...

Store

Indexing

Remove formating

Indexing

Tokenize

Indexing

Stop words

Indexing

Stemming

http://snowballstem.org

Indexing

Synonyms

elastic

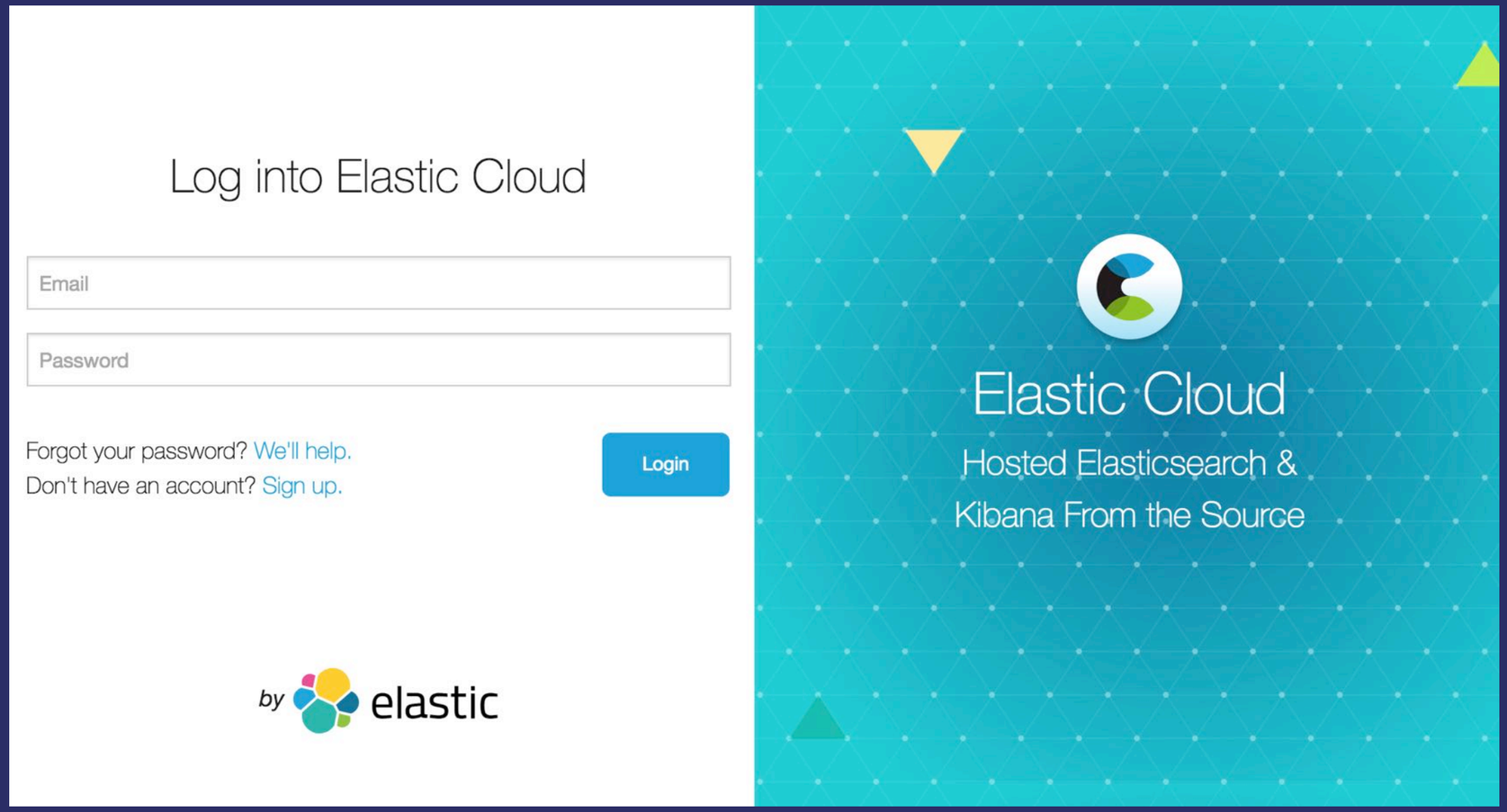
elastic

Apache Lucene
Elasticsearch

cloud

elastic

https://cloud.elastic.co



docker

elastic

```yaml
---
version: '2'
services:
  kibana:
    image: docker.elastic.co/kibana/kibana:5.2.1
    links:
      - elasticsearch
    ports:
      - 5601:5601

  elasticsearch:
    image: docker.elastic.co/elasticsearch/elasticsearch:5.2.1
    cap_add:
      - IPC_LOCK
    volumes:
      - esdata1:/usr/share/elasticsearch/data
    ports:
      - 9200:9200

volumes:
  esdata1:
    driver: local
```


elastic

Example

These are not the droids you are looking for.

`html_strip` Char Filter

These are not the droids you are looking for.

`standard` Tokenizer

These are not the droids you are looking for

`lowercase` Token Filter

these are not the droids you are looking for

`stop` Token Filter

droids you looking

elastic

`snowball` Token Filter

droid you look

```
GET /_analyze
{
  "char_filter": [
    "html_strip"
  ],
  "tokenizer": "standard",
  "filter": [
    "lowercase",
    "stop",
    "snowball"
  ],
  "text": "These are <em>not</em> the droids you are looking for."
}
```

```
{
  "tokens": [
    {
      "token": "droid",
      "start_offset": 27,
      "end_offset": 33,
      "type": "<ALPHANUM>",
      "position": 4
    },
    {
      "token": "you",
      "start_offset": 34,
      "end_offset": 37,
      "type": "<ALPHANUM>",
      "position": 5
    },
    ...
  ]
}
```

Stop Words

a, an, and, are, as, at, be, but, by, for, if, in, into, is, it, no, not, of, on, or, such, that, the, their, then, there, these, they, this, to, was, will, with

Russian

Это не те дроиды, которых ты ищешь

elastic

Russian

эт те дроид котор ищеш

```
GET /_analyze
{
  "analyzer": "russian",
  "text": "Это не те дроиды, которых ты ищешь"
}
```

```
{
  "tokens": [
    {
      "token": "эт",
      "start_offset": 0,
      "end_offset": 3,
      "type": "<ALPHANUM>",
      "position": 0
    },
    {
      "token": "те",
      "start_offset": 7,
      "end_offset": 9,
      "type": "<ALPHANUM>",
      "position": 2
    },
    ...
  ]
}
```

German, please!

Das sind nicht die Droiden, nach denen du suchst.

elastic

German, please!

droid den such

```
GET /_analyze
{
  "analyzer": "german",
  "text": "Das sind nicht die Droiden, nach denen du suchst."
}
```

German with the English Analyzer

das sind nicht die droiden nach denen du suchst

German Stop Words

https://github.com/apache/lucene-solr/blob/master/lucene/analysis/common/src/resources/org/apache/lucene/analysis/snowball/german_stop.txt

Detecting Languages

https://github.com/spinscale/elasticsearch-ingest-langdetect

Another Example

Obi-Wan hat mir nie erzählt, was mit meinem Vater geschehen ist.

Another Example

obi wan nie erzahlt vat gescheh

Another Example

Nein. Ich bin dein Vater

Another Example

nein vat

Languages in 5.0

arabic, armenian, basque, brazilian, bulgarian, catalan, cjk, czech, danish, dutch, english, finnish, french, galician, german, greek, hindi, hungarian, indonesian, irish, italian, latvian, lithuanian, norwegian, persian, portuguese, romanian, russian, sorani, spanish, swedish, turkish, thai

Ukrainian

Lucene 6.2

Elasticsearch 5.1

https://github.com/elastic/elasticsearch/pull/21176

Another Example

Obi-Wan never told you what happened to your father.

Another Example

obi wan never told you what happen your father

Another Example

No. I am your father.

Another Example

i am your father

Language Rules

English: `Philipp's` → `philipp`

French: `l'église` → `eglis`

German: `äußerst` → `ausserst`

`phonetic` Token Filter

Plugin

elastic

```
GET /_analyze
{
  "tokenizer": "standard",
  "filter": [
    {
      "type": "phonetic",
      "encoder": "metaphone",
      "replace": false
    }
  ],
  "text": "Obi-Wan"
}
```

```
{
  "tokens": [
    {
      "token": "OB",
      "start_offset": 0,
      "end_offset": 3,
      "type": "<ALPHANUM>",
      "position": 0
    },
    {
      "token": "Obi",
      "start_offset": 0,
      "end_offset": 3,
      "type": "<ALPHANUM>",
      "position": 0
    },
    {
      "token": "WN",
      "start_offset": 4,
      "end_offset": 7,
      "type": "<ALPHANUM>",
      "position": 1
    },
    {
      "token": "Wan",
      "start_offset": 4,
      "end_offset": 7,
      "type": "<ALPHANUM>",
      "position": 1
    }
  ]
}
```

Elasticsearch

Index, typ, mapping

elastic

Multi-Language Support

One language per index

~~One language per type~~

One language per field

```
PUT /starwars
{
  "settings": {
    "analysis": {
      "filter": {
        "my_synonym_filter": {
          "type": "synonym",
          "synonyms": [
            "droid,machine",
            "father,dad"
          ]
        }
      },
```

```
    "analyzer": {
      "my_analyzer": {
        "char_filter": [
          "html_strip"
        ],
        "tokenizer": "standard",
        "filter": [
          "lowercase",
          "stop",
          "snowball",
          "my_synonym_filter"
        ]
      }
    }
  }
},
```

```
"mappings": {
  "quotes": {
    "properties": {
      "quote": {
        "type": "text",
        "analyzer": "my_analyzer"
      }
    }
  }
}
```

```
GET /starwars/_mapping

GET /starwars/_settings
```

```
PUT /starwars/quotes/1
{
  "quote": "These are <em>not</em> the droids you are looking for."
}
PUT /starwars/quotes/2
{
  "quote": "Obi-Wan never told you what happened to your father."
}
PUT /starwars/quotes/3
{
  "quote": "<b>No</b>. I am your father."
}
```

```
GET /starwars/quotes/1

GET /starwars/quotes/1/_source
```

Inverted Index

	ID 1	ID 2	ID 3
am	0	0	1[2]
droid	1[4]	0	0
father	0	1[9]	1[4]
happen	0	1[6]	0
i	0	0	1[1]
look	1[7]	0	0
never	0	1[2]	0
obi	0	1[0]	0
told	0	1[3]	0
wan	0	1[1]	0
what	0	1[5]	0
you	1[5]	1[4]	0
your	0	1[8]	1[3]

Search

elastic

```
POST /starwars/_search
{
  "query": {
    "match_all": { }
  }
}
```

GET **VS** POST

```
{
  "took": 1,
  "timed_out": false,
  "_shards": {
    "total": 5,
    "successful": 5,
    "failed": 0
  },
  "hits": {
    "total": 3,
    "max_score": 1,
    "hits": [
      {
        "_index": "starwars",
        "_type": "my_type",
        "_id": "2",
        "_score": 1,
        "_source": {
          "quote": "Obi-Wan never told you what happened to your father."
        }
      },
      ...
```

```
POST /starwars/_search
{
  "query": {
    "match": {
      "quote": "droid"
    }
  }
}
```

```
{
  "took": 2,
  "timed_out": false,
  "_shards": {
    "total": 5,
    "successful": 5,
    "failed": 0
  },
  "hits": {
    "total": 1,
    "max_score": 0.39556286,
    "hits": [
      {
        "_index": "starwars",
        "_type": "quotes",
        "_id": "1",
        "_score": 0.39556286,
        "_source": {
          "quote": "These are <em>not</em> the droids you are looking for."
        }
      }
    ]
  }
}
```

```
POST /starwars/_search
{
  "query": {
    "match": {
      "quote": "dad"
    }
  }
}
```

```
...
  "hits": {
    "total": 2,
    "max_score": 0.41913947,
    "hits": [
      {
        "_index": "starwars",
        "_type": "quotes",
        "_id": "3",
        "_score": 0.41913947,
        "_source": {
          "quote": "<b>No</b>. I am your father."
        }
      },
      {
        "_index": "starwars",
        "_type": "quotes",
        "_id": "2",
        "_score": 0.39291072,
        "_source": {
          "quote": "Obi-Wan never told you what happened to your father."
        }
      }
    ]
  }
}
```

```
POST /starwars/_search
{
  "query": {
    "match_phrase": {
      "quote": "I am your father"
    }
  }
}
```

```
{
  "took": 3,
  "timed_out": false,
  "_shards": {
    "total": 5,
    "successful": 5,
    "failed": 0
  },
  "hits": {
    "total": 1,
    "max_score": 1.5665855,
    "hits": [
      {
        "_index": "starwars",
        "_type": "quotes",
        "_id": "3",
        "_score": 1.5665855,
        "_source": {
          "quote": "<b>No</b>. I am your father."
        }
      }
    ]
  }
}
```

```
POST /starwars/_search
{
  "query": {
    "match_phrase": {
      "quote": "I am not your father"
    }
  }
}
```

```
{
  "took": 15,
  "timed_out": false,
  "_shards": {
    "total": 5,
    "successful": 5,
    "failed": 0
  },
  "hits": {
    "total": 0,
    "max_score": null,
    "hits": []
  }
}
```

```
POST /starwars/_search
{
  "query": {
    "match_phrase": {
      "quote": {
        "query": "I am not your father",
        "slop": 1
      }
    }
  }
}
```

```
{
  "took": 5,
  "timed_out": false,
  "_shards": {
    "total": 5,
    "successful": 5,
    "failed": 0
  },
  "hits": {
    "total": 1,
    "max_score": 1.0409548,
    "hits": [
      {
        "_index": "starwars",
        "_type": "quotes",
        "_id": "3",
        "_score": 1.0409548,
        "_source": {
          "quote": "<b>No</b>. I am your father."
        }
      }
    ]
  }
}
```

```
POST /starwars/_search
{
  "query": {
    "match": {
      "quote": {
        "query": "van",
        "fuzziness": "AUTO"
      }
    }
  }
}
```

```
{
  "took": 14,
  "timed_out": false,
  "_shards": {
    "total": 5,
    "successful": 5,
    "failed": 0
  },
  "hits": {
    "total": 1,
    "max_score": 0.18155496,
    "hits": [
      {
        "_index": "starwars",
        "_type": "quotes",
        "_id": "2",
        "_score": 0.18155496,
        "_source": {
          "quote": "Obi-Wan never told you what happened to your father."
        }
      }
    ]
  }
}
```

```
POST /starwars/_search
{
  "query": {
    "match": {
      "quote": {
        "query": "ovi-van",
        "fuzziness": 1
      }
    }
  }
}
```

```
{
  "took": 109,
  "timed_out": false,
  "_shards": {
    "total": 5,
    "successful": 5,
    "failed": 0
  },
  "hits": {
    "total": 1,
    "max_score": 0.3798467,
    "hits": [
      {
        "_index": "starwars",
        "_type": "quotes",
        "_id": "2",
        "_score": 0.3798467,
        "_source": {
          "quote": "Obi-Wan never told you what happened to your father."
        }
      }
    ]
  }
}
```

FuzzyQuery History

http://blog.mikemccandless.com/2011/03/lucenes-fuzzyquery-is-100-times-faster.html

Before: Brute force

Now: Levenshtein Automaton

```
SELECT *
  FROM starwars
 WHERE quote LIKE "?an" OR
       quote LIKE "V?n" OR
       quote LIKE "Va?"
```

Scoring

Term Frequency / Inverse Document Frequency (TF/IDF)

Search one term

elastic

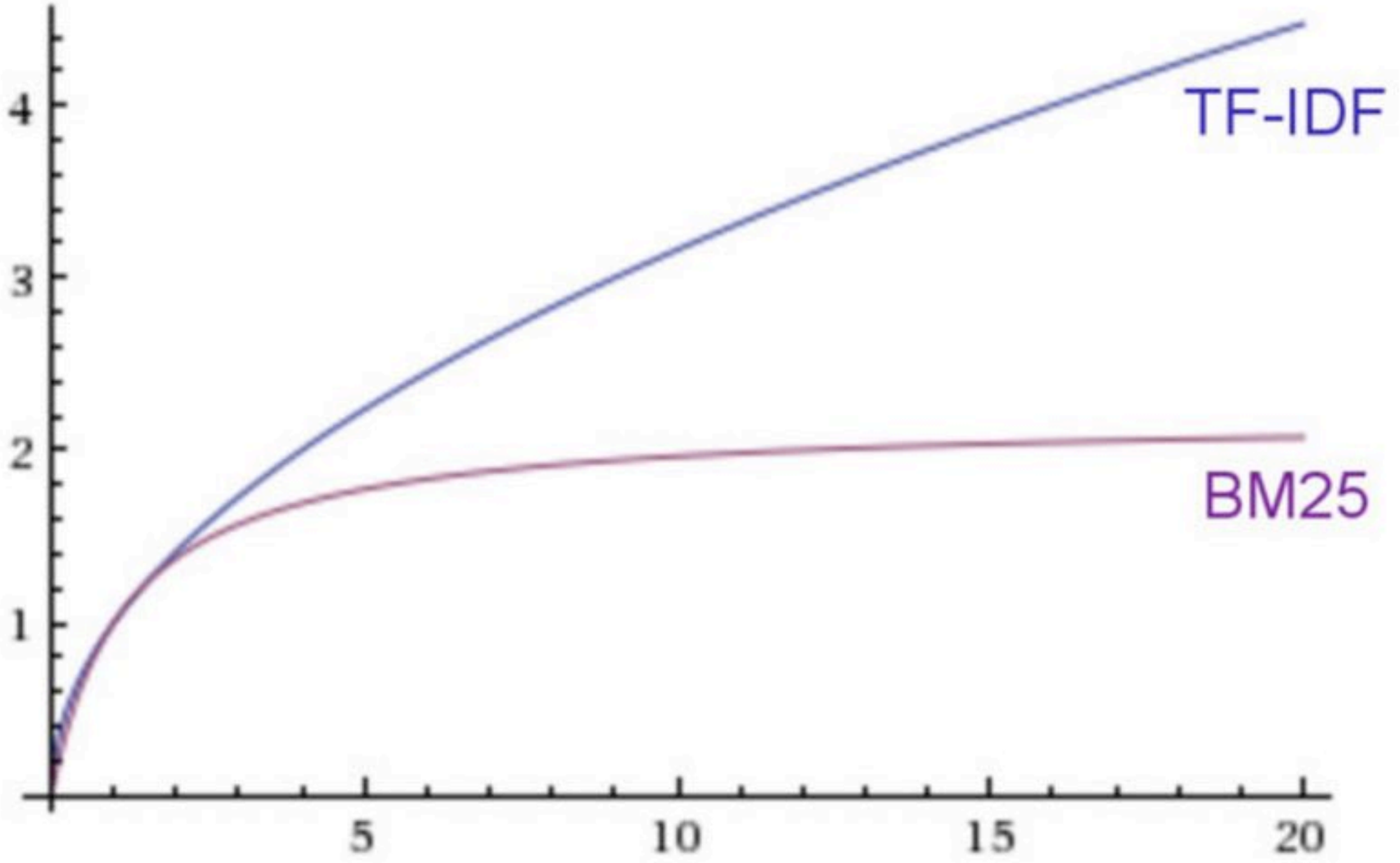
BM25

Default in Elasticsearch 5.0

https://speakerdeck.com/elastic/improved-text-scoring-with-bm25

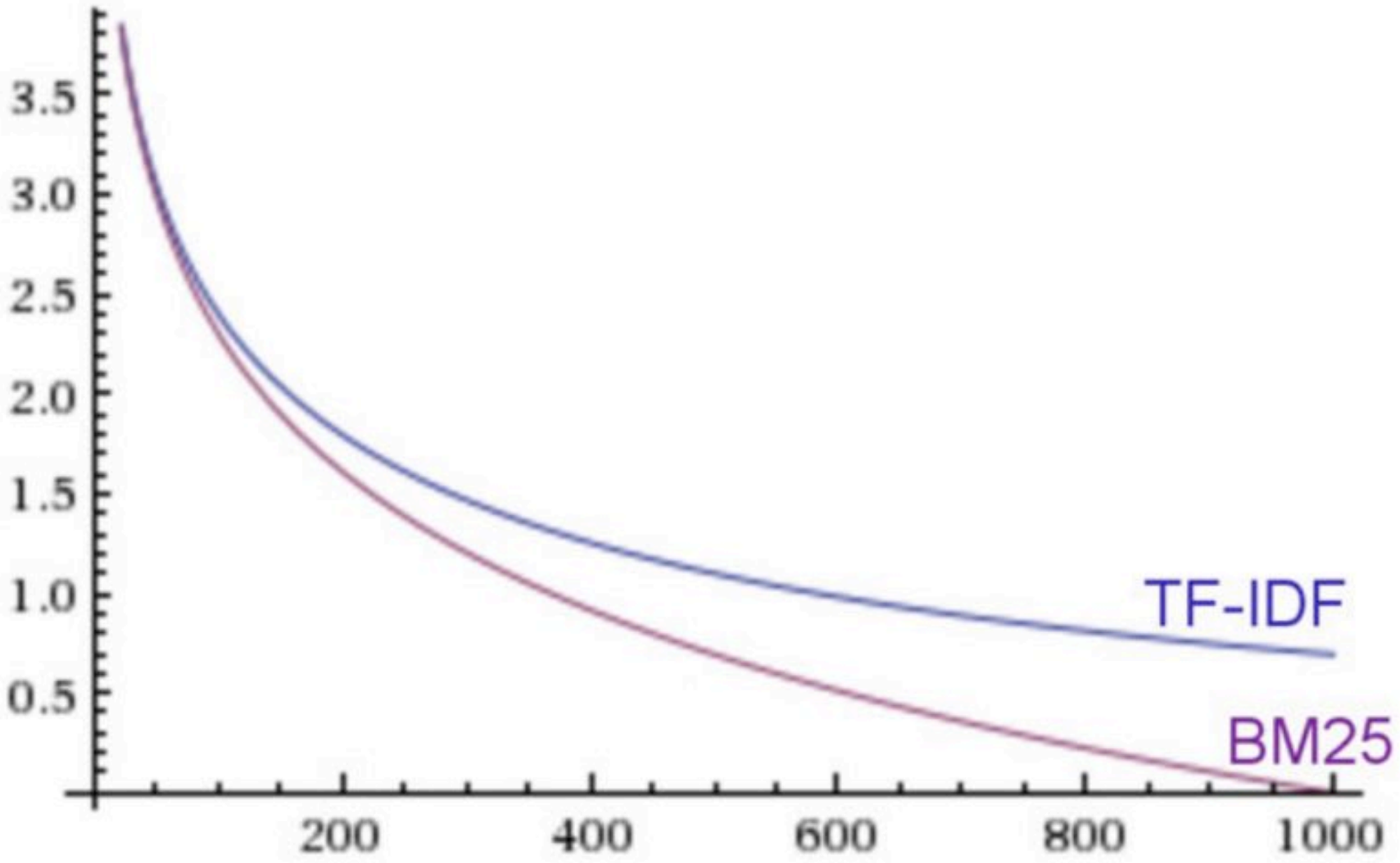
Term Frequency

$$tf(t\ in\ d) = \sqrt{frequency}$$


TF-IDF
BM25
4
3
2
1
5
10
15
20

Inverse Document Frequency

$$idf(t) = 1 + log(\frac{numDocs}{docFreq + 1})$$


3.5
3.0
2.5
2.0
1.5
1.0
0.5
200
400
600
800
1000
TF-IDF
BM25

Field-Length Norm

$$norm(d) = \frac{1}{\sqrt{numTerms}}$$

Putting it Together

```
score(q,d)  =
              queryNorm(q)
            · coord(q,d)
            · Σ (
                tf(t in d)
              · idf(t)²
              · t.getBoost()
              · norm(t,d)
              ) (t in q)
```

```
POST /starwars/_search?explain
{
  "query": {
    "match": {
      "quote": "father"
    }
  }
}
```

```
...
"_explanation": {
  "value": 0.41913947,
  "description": "weight(Synonym(quote:dad quote:father) in 0) [PerFieldSimilarity], result of:",
  "details": [
    {
      "value": 0.41913947,
      "description": "score(doc=0,freq=2.0 = termFreq=2.0\n), product of:",
      "details": [
        {
          "value": 0.2876821,
          "description": "idf(docFreq=1, docCount=1)",
          "details": []
        },
        {
          "value": 1.4569536,
          "description": "tfNorm, computed from:",
          "details": [
            {
              "value": 2,
              "description": "termFreq=2.0",
              "details": []
            },
            ...
```

Score

0.41913947: i am your father

0.39291072: obi wan never told you what happen your father

Vector Space Model

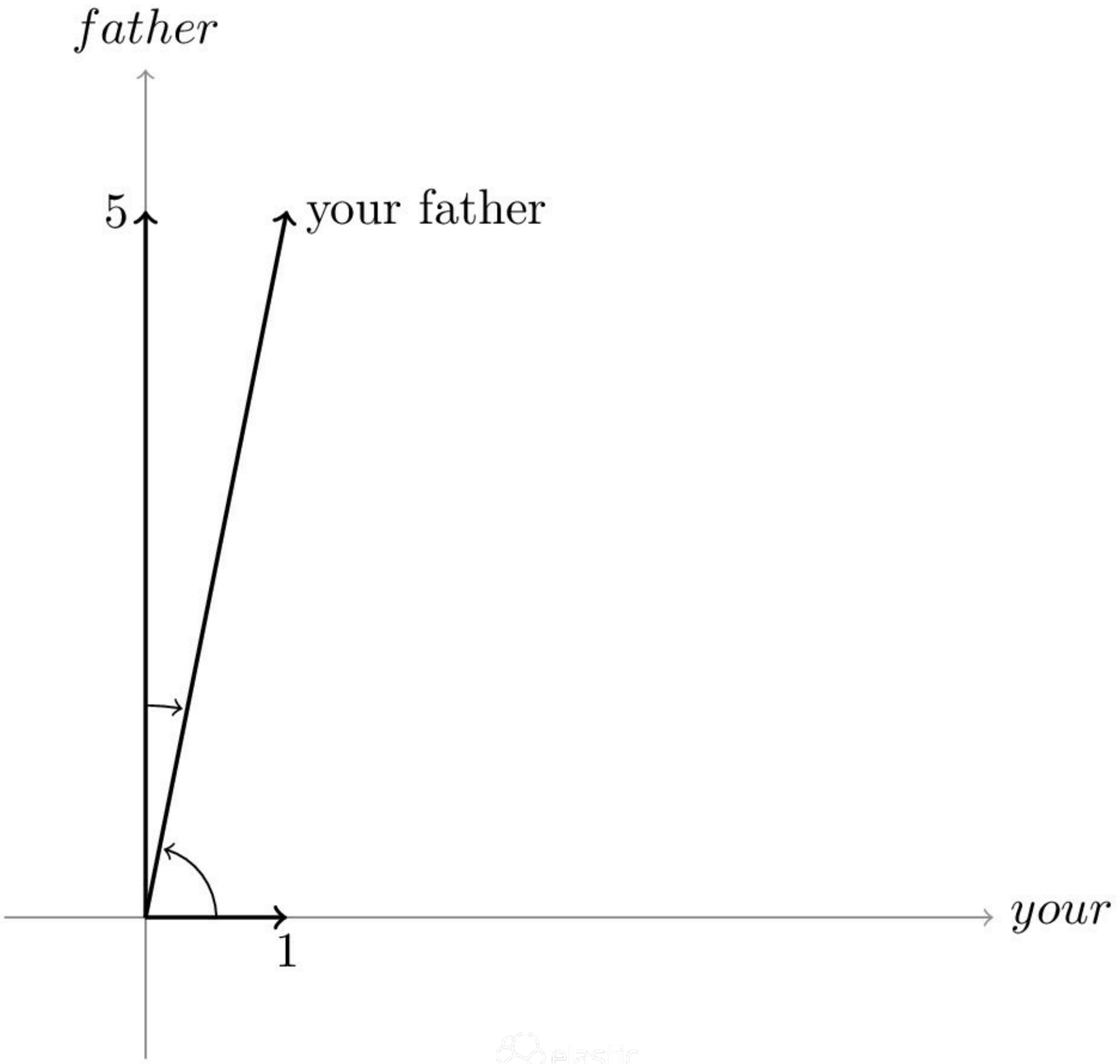
Search multiple terms

elastic

Score each term
Vectorize
Calculate angle

Search your father

elastic


father
5
your father
your
1

Function Score

Script, weight, random, field value, decay (geo or date)

```
POST /starwars/_search
{
  "query": {
    "function_score": {
      "query": {
        "match": {
          "quote": "father"
        }
      },
      "random_score": {}
    }
  }
}
```

More Features

elastic

```
POST /starwars/_search
{
  "query": {
    "match": {
      "quote": "father"
    }
  },
  "highlight": {
    "pre_tags": [
      "<tag>"
    ],
    "post_tags": [
      "</tag>"
    ],
    "fields": {
      "quote": {}
    }
  }
}
```

```
...
"hits": [
  {
    "_index": "starwars",
    "_type": "quotes",
    "_id": "3",
    "_score": 0.41913947,
    "_source": {
      "quote": "<b>No</b>. I am your father."
    },
    "highlight": {
      "quote": [
        "<b>No</b>. I am your <tag>father</tag>."
      ]
    }
  },
  ...
```

Boolean Queries

must must_not should filter

```
POST /starwars/_search
{
  "query": {
    "bool": {
      "must": {
        "match": {
          "quote": {
            "query": "father"
          }
        }
      },
      "should": [
        {
          "match": {
            "quote": {
              "query": "your"
            }
          }
        },
        {
          "match": {
            "quote": {
              "query": "obi"
            }
          }
        }
      ]
    }
  }
}
```

```
...
  "hits": {
    "total": 2,
    "max_score": 0.96268076,
    "hits": [
      {
        "_index": "starwars",
        "_type": "quotes",
        "_id": "2",
        "_score": 0.96268076,
        "_source": {
          "quote": "Obi-Wan never told you what happened to your father."
        }
      },
      {
        "_index": "starwars",
        "_type": "quotes",
        "_id": "3",
        "_score": 0.73245656,
        "_source": {
          "quote": "<b>No</b>. I am your father."
        }
      }
    ]
  }
}
```

```
POST /starwars/_search
{
  "query": {
    "bool": {
      "filter": {
        "match": {
          "quote": {
            "query": "father"
          }
        }
      },
      "should": [
        {
          "match": {
            "quote": {
              "query": "your"
            }
          }
        },
        {
          "match": {
            "quote": {
              "query": "obi"
            }
          }
        }
      ]
    }
  }
}
```

```
...
  "hits": {
    "total": 2,
    "max_score": 0.56977004,
    "hits": [
      {
        "_index": "starwars",
        "_type": "quotes",
        "_id": "2",
        "_score": 0.56977004,
        "_source": {
          "quote": "Obi-Wan never told you what happened to your father."
        }
      },
      {
        "_index": "starwars",
        "_type": "quotes",
        "_id": "3",
        "_score": 0.31331712,
        "_source": {
          "quote": "<b>No</b>. I am your father."
        }
      }
    ]
  }
}
```

```
POST /starwars/_search
{
  "query": {
    "bool": {
      "must": {
        "match": {
          "quote": {
            "query": "father"
          }
        }
      },
      "should": [
        {
          "match": {
            "quote": {
              "query": "your"
            }
          }
        },
        {
          "match": {
            "quote": {
              "query": "obi"
            }
          }
        },
        {
          "match": {
            "quote": {
              "query": "droid"
            }
          }
        }
      ],
      "minimum_number_should_match": 2
    }
  }
}
```

```
...
  "hits": {
    "total": 1,
    "max_score": 0.96268076,
    "hits": [
      {
        "_index": "starwars",
        "_type": "quotes",
        "_id": "2",
        "_score": 0.96268076,
        "_source": {
          "quote": "Obi-Wan never told you what happened to your father."
        }
      }
    ]
  }
}
```

Boosting

Default 1 — greater or smaller

```
POST /starwars/_search
{
  "query": {
    "bool": {
      "must": {
        "match": {
          "quote": {
            "query": "father"
          }
        }
      },
      "should": [
        {
          "match": {
            "quote": {
              "query": "your"
            }
          }
        },
        {
          "match": {
            "quote": {
              "query": "obi",
              "boost": 3
            }
          }
        }
      ]
    }
  }
}
```

```
...
  "hits": {
    "total": 2,
    "max_score": 1.5324509,
    "hits": [
      {
        "_index": "starwars",
        "_type": "quotes",
        "_id": "2",
        "_score": 1.5324509,
        "_source": {
          "quote": "Obi-Wan never told you what happened to your father."
        }
      },
      {
        "_index": "starwars",
        "_type": "quotes",
        "_id": "3",
        "_score": 0.73245656,
        "_source": {
          "quote": "<b>No</b>. I am your father."
        }
      }
    ]
  }
}
```

Suggestion

Suggest a similar text

`_search` **end point**

`_suggest` **deprecated**

```
POST /starwars/_search
{
  "query": {
    "match": {
      "quote": "fath"
    }
  },
  "suggest": {
    "my_suggestion" : {
      "text" : "drui",
      "term" : {
        "field" : "quote"
      }
    }
  }
}
```

```
...
  "hits": {
    "total": 0,
    "max_score": null,
    "hits": []
  },
  "suggest": {
    "my_suggestion": [
      {
        "text": "drui",
        "offset": 0,
        "length": 4,
        "options": [
          {
            "text": "droid",
            "score": 0.5,
            "freq": 1
          }
        ]
      }
    ]
  }
}
```

NGram

Partial matches

Trigram

Edge Gram

elastic

```
GET /_analyze
{
  "char_filter": [
    "html_strip"
  ],
  "tokenizer": {
    "type": "ngram",
    "min_gram": "3",
    "max_gram": "3",
    "token_chars": [
      "letter"
    ]
  },
  "filter": [
    "lowercase"
  ],
  "text": "These are <em>not</em> the droids you are looking for."
}
```

```
{
  "tokens": [
    {
      "token": "the",
      "start_offset": 0,
      "end_offset": 3,
      "type": "word",
      "position": 0
    },
    {
      "token": "hes",
      "start_offset": 1,
      "end_offset": 4,
      "type": "word",
      "position": 1
    },
    {
      "token": "ese",
      "start_offset": 2,
      "end_offset": 5,
      "type": "word",
      "position": 2
    },
    {
      "token": "are",
      "start_offset": 6,
      "end_offset": 9,
      "type": "word",
      "position": 3
    },
    ...
```

```
GET /_analyze
{
  "char_filter": [
    "html_strip"
  ],
  "tokenizer": {
    "type": "edge_ngram",
    "min_gram": "1",
    "max_gram": "3",
    "token_chars": [
      "letter"
    ]
  },
  "filter": [
    "lowercase"
  ],
  "text": "These are <em>not</em> the droids you are looking for."
}
```

```
{
  "tokens": [
    {
      "token": "t",
      "start_offset": 0,
      "end_offset": 1,
      "type": "word",
      "position": 0
    },
    {
      "token": "th",
      "start_offset": 0,
      "end_offset": 2,
      "type": "word",
      "position": 1
    },
    {
      "token": "the",
      "start_offset": 0,
      "end_offset": 3,
      "type": "word",
      "position": 2
    },
    {
      "token": "a",
      "start_offset": 6,
      "end_offset": 7,
      "type": "word",
      "position": 3
    },
    {
      "token": "ar",
      "start_offset": 6,
      "end_offset": 8,
      "type": "word",
      "position": 4
    },
    ...
```

Conclusion

Indexing

Formatting

Tokenize

Lowercase, Stop Words, Stemming

Synonyms

Scoring

Term Frequency

Inverse Document Frequency

Field-Length Norm

Vector Space Model

elastic

Querying

Match: Phrase, Slop, Fuzziness

Boolean

Suggestion

NGram

elastic

Danke!

Questions?

Philipp Krenn @xeraa

PS: Stickers

elastic

Image Credits

Schnitzel https://flic.kr/p/9m27wm

Architecture https://flic.kr/p/6dwCAe

Conchita https://flic.kr/p/nBqSHT