Есть что новое в поиске?
А если найду?

Михаил Хлуднев

ASF PMC

t.me/MUST_SEARCH

mkhl@apache.org

Joker<?>

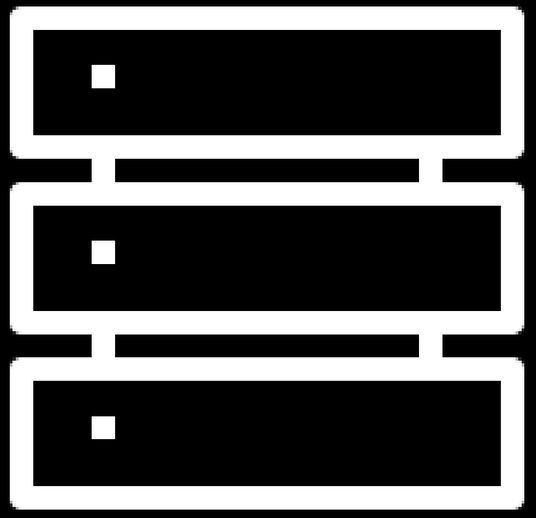
1. Ctrl+F
2. SELECT … WHERE foo LIKE %bar%
3. q=*:*
4. {"combined_fields": { "query": "foo", …}}

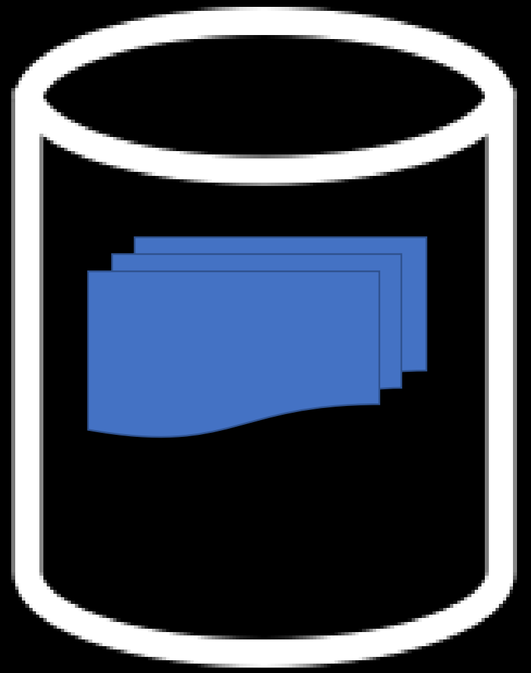
I. Java Search Overview

II. Suggestions in Elasticsearch

III. Extended Boolean Search

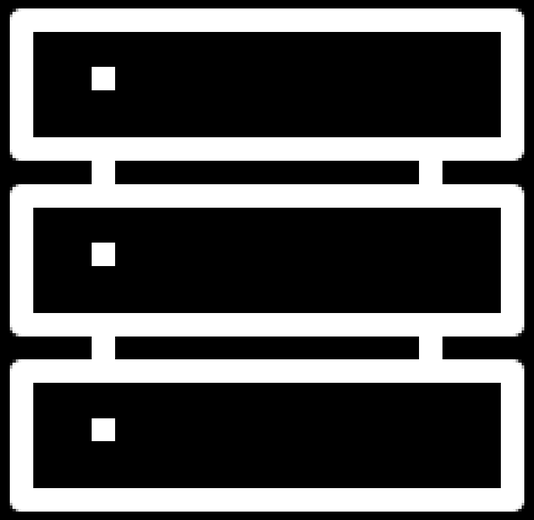
APACHE LUCENE™





ASL

Solr



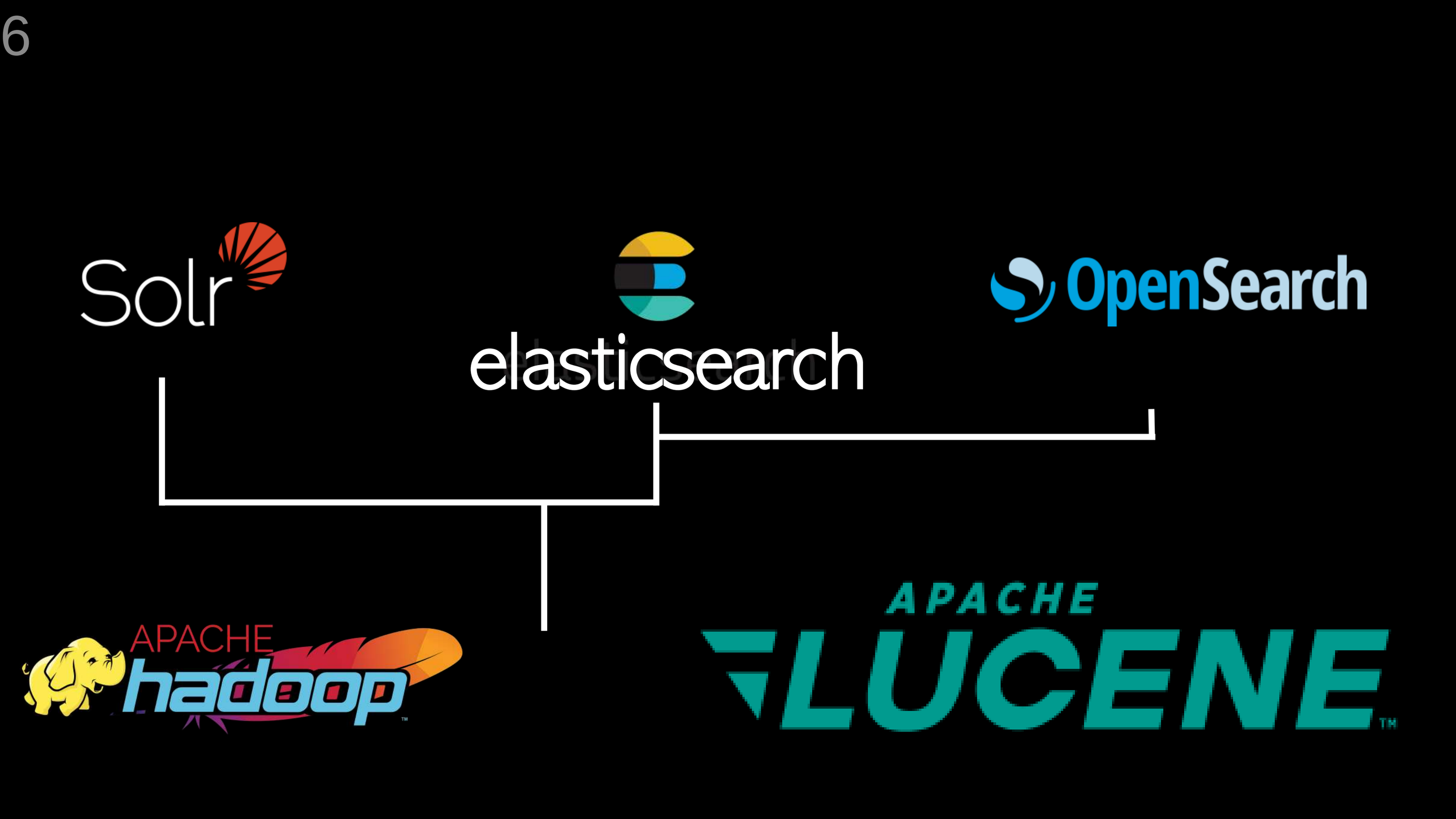
EL, SSPL, $

elasticsearch

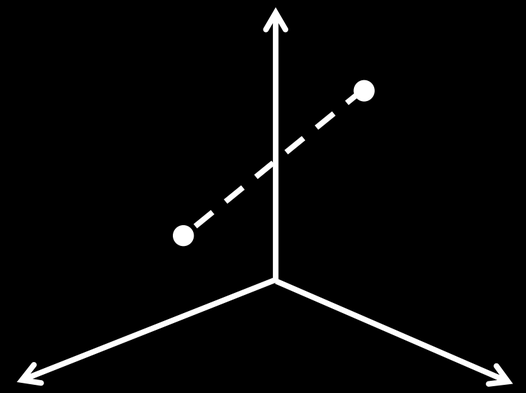
ASL

OpenSearch

APACHE LUCENE™


Solr
elasticsearch
OpenSearch
APACHE
hadoop
APACHE
LUCENE

vespa

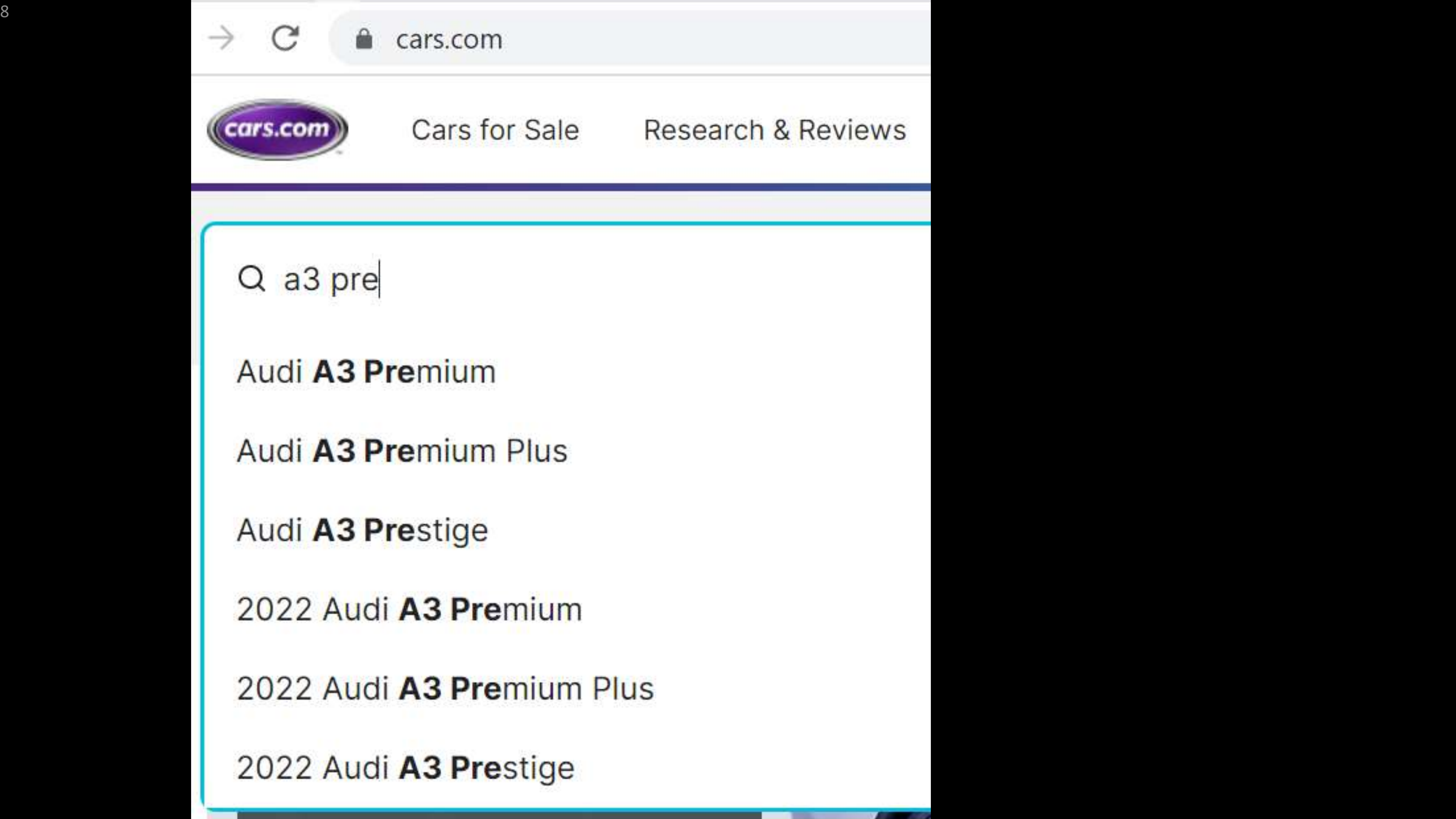


Solr

elasticsearch

OpenSearch

APACHE LUCENE™


cars.com
cars.com
Cars for Sale
Research & Reviews
a3 pre
Audi **A3 Pre**mium
Audi **A3 Pre**mium Plus
Audi **A3 Pre**stige
2022 Audi **A3 Pre**mium
2022 Audi **A3 Pre**mium Plus
2022 Audi **A3 Pre**stige

SEARCH_AS_YOU_TYPE

A build-in field type for you

PUT /cars

```
"mappings":{
    "properties": {
        "make":{"type":"text"}, "model":{"type":"text"},
        "year":{"type":"integer"},"color":{"type":"text"},

        "suggestion": {
              "type": "search_as_you_type"
        }
}  }
```

Indexing Titles

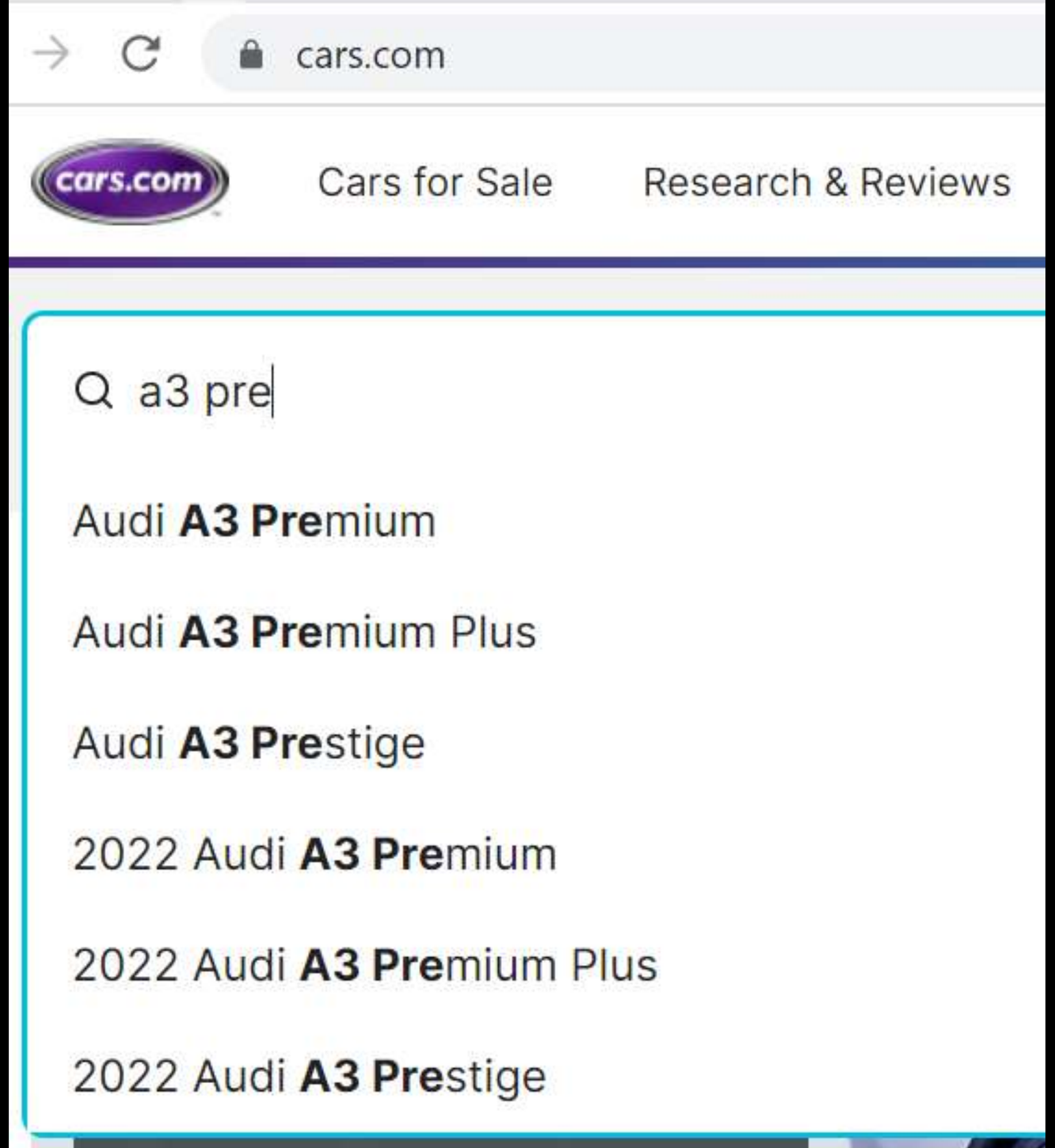
```
Map.of(   "make", row.make,    "model", row.model,
          "year", row.year,  "color",row.color, "trim", row.trim,

              "suggestion", Arrays.asList(
                    join(" ", row.make, row.model),
                    join(" ", row.make, row.model, row.trim ),
                    join(" ", row.make, row.model, row.color),
                  join(" ", row.make, row.model, row.trim, row.color ),
                    join(" ", row.year, row.make, row.model),
                    join(" ", row.year, row.make, row.model, row.color)
));
    }
```

Suggesting "au"

```
GET cars/_search
{
    "size":20,"query":{
    "multi_match":{       "query":"au",
        "fields":[
           "suggestion^1.0",
           "make", "suggestion._index_prefix"
         ],"type":"bool_prefix"
     }},
     "highlight":{"pre_tags":["["],
     "post_tags":["]"], …
     "fields":{"suggestion":{},"model":{},
             "trim":{},"make":{},"color":{}}
}}
```

```
"_source": {
    "color": "Black",
    "year": 2019,          "model": "A3",
    "trim": "Premium", "make": "Audi",
    "suggestion": [
      "Audi A3", "Audi A3 Premium",
      "Audi A3 Black", "Audi A3 Premium Blac
      "2019 Audi A3",  "2019 Audi A3 Black"
        ]
      },
     "highlight": {
        "suggestion": [
          "[Audi] A3", "[Audi] A3 Black",
          "2019 [Audi] A3",
          "[Audi] A3 Premium",
          "2019 [Audi] A3 Black"
        ],
        "make": [ "[Audi]"  ]
      }
```


cars.com
cars.com
Cars for Sale
Research & Reviews
a3 pre
Audi **A3 Pre**mium
Audi **A3 Pre**mium Plus
Audi **A3 Pre**stige
2022 Audi **A3 Pre**mium
2022 Audi **A3 Pre**mium Plus
2022 Audi **A3 Pre**stige

That’s Why

```
GET /cars/_analyze
{
 "field": "suggestion._index_prefix",
 "text": "Audi A3 Premium Silver"
}
```

```
{
    "token": "a3 p",
    "start_offset": 5,
    "end_offset": 22,
    "type": "shingle",
    "position": 1
  },
  {
    "token": "a3 pr",
    "start_offset": 5,
    "end_offset": 22,
    "type": "shingle",
    "position": 1
  },
  {
    "token": "a3 pre",
    "start_offset": 5,
    "end_offset": 22,
    "type": "shingle",
    "position": 1
  },
  {
    "token":
        "a3 prem",
    "start_offset": 5,
    "end_offset": 22,
    "type": "shingle",
    "position": 1
  },
```

EDGE_NGRAM TOKENIZER to rescue

PUT cars/_mapping

```
"make": {
     "type": "text",
     "analyzer": "autocomplete",

        },

"model": {
     "type": "text",
     "analyzer": "autocomplete",


        },
```

```
"trim": {
  "type": "text",
  "analyzer": "autocomplete",


},
```

PUT cars/_settings

```
"analysis": {
"analyzer": {
     "autocomplete": {
        "filter": ["lowercase"],
        "tokenizer": "autocomplete"},
},
"tokenizer": {
  "autocomplete": {
     "token_chars": [
       "letter",  "digit"
     ],
     "min_gram": "1",
     "type": "edge_ngram",
     "max_gram": "10"
  }
```

GET cars/_analyze


```
{
      "token": "p",
      "start_offset": 0,
      "end_offset": 1,
      "type": "word",
      "position": 0
    },
    {
      "token": "pr",
      "start_offset": 0,
      "end_offset": 2,
      "type": "word",
      "position": 1
    },
    {
      "token": "pre",
      "start_offset": 0,
      "end_offset": 3,
      "type": "word",
      "position": 2
    },
    {
      "token": "prem",
      "start_offset": 0,
      "end_offset": 4,
      "type": "word",
      "position": 3
    },
```

GET cars/_search

```
{"query":{"multi_match":{
    "query":"a3 pre",
"fields":[…],
"type":"cross_fields","operator":"A
ND"}},
"highlight":{"pre_tags":["["],"post
_tags":["]"],
"fields":{"year":{},"make":{},"mode
l":{},"trim":{},"color":{}}}
}
```

```
 {
          "color": "Silver",
"year": "2017", "make": "Audi",
          "model": "A3",  "trim":
"Premium"
        },
        "highlight": {
          "trim": [ "[Pre]mium"  ],
          "model": [ "[A3]"     ]
```

Summary

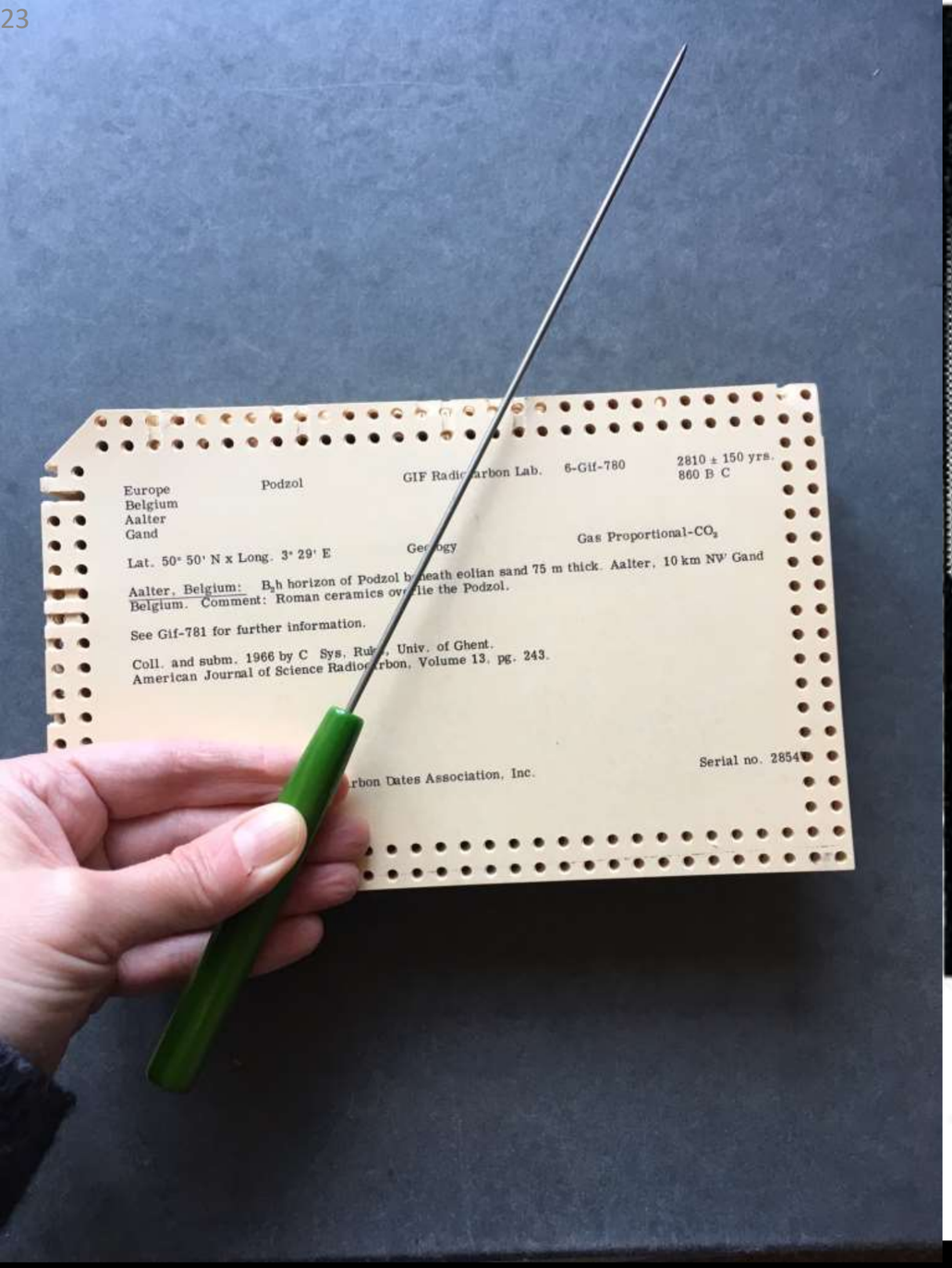
- Elasticsearch exposes REST API
- Mapping a kinda SCHEMA
- Field type, analyzer, tokenizer you know
- Ngrams rocks

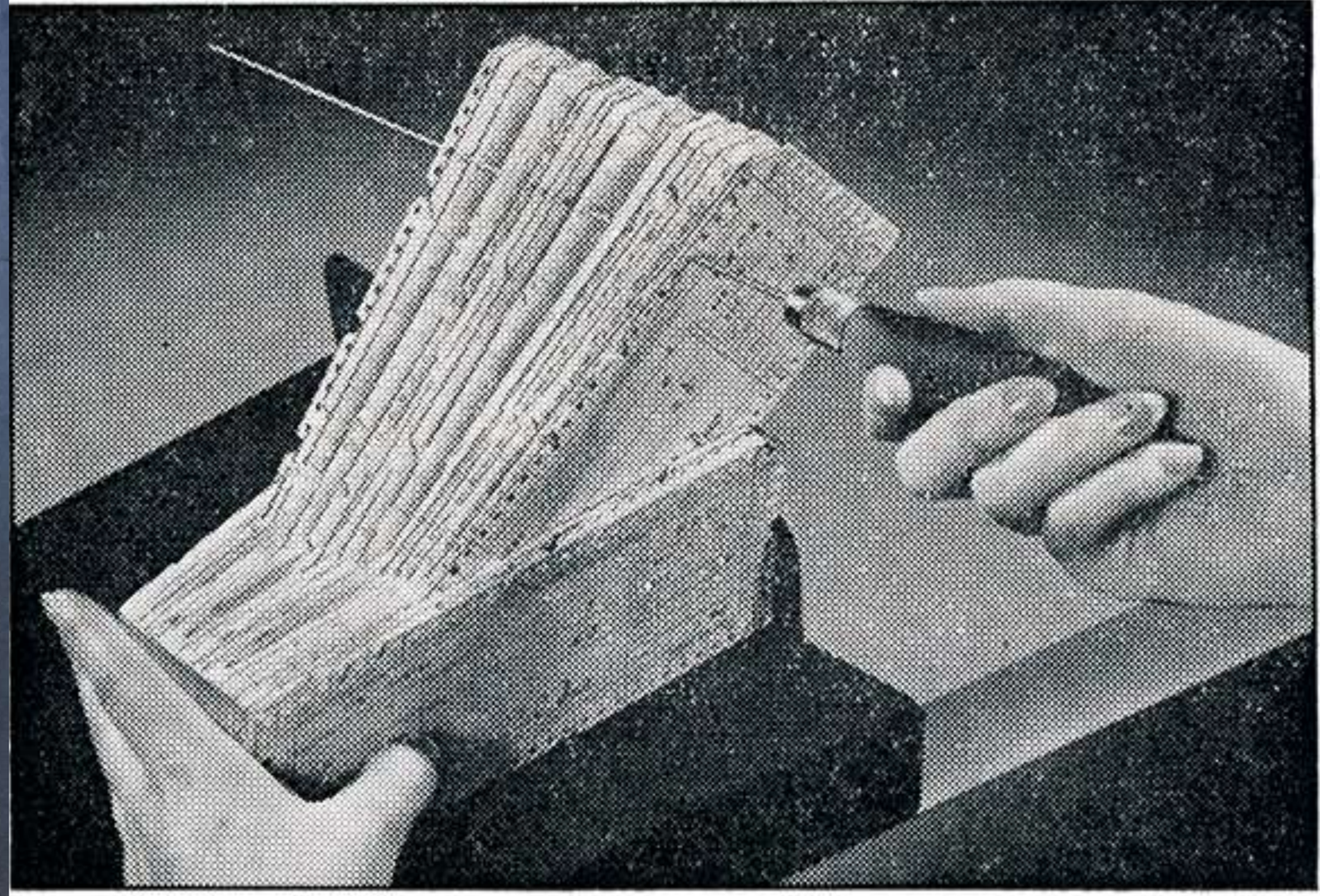
Apache

Lucene

Boolean Search

foo AND bar



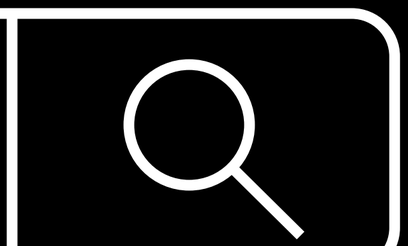


McBee Systems,
Royal Typewriter Company

Figure 9-11. McBee Key-Sort

Document	X	Y	+X +Y ***X ∧ Y***	X Y ***X ∨ Y***	X +Y
doc1: { X }	1.0	n/a	n/a	1.0	n/a
doc2: { Y }	n/a	1.0	n/a	1.0	1.0
doc3: { X Y }	1.0	1.0	2.0	2.0	2.0

Red shirt



Red pants black shirt

NOT (only) Boolean Search

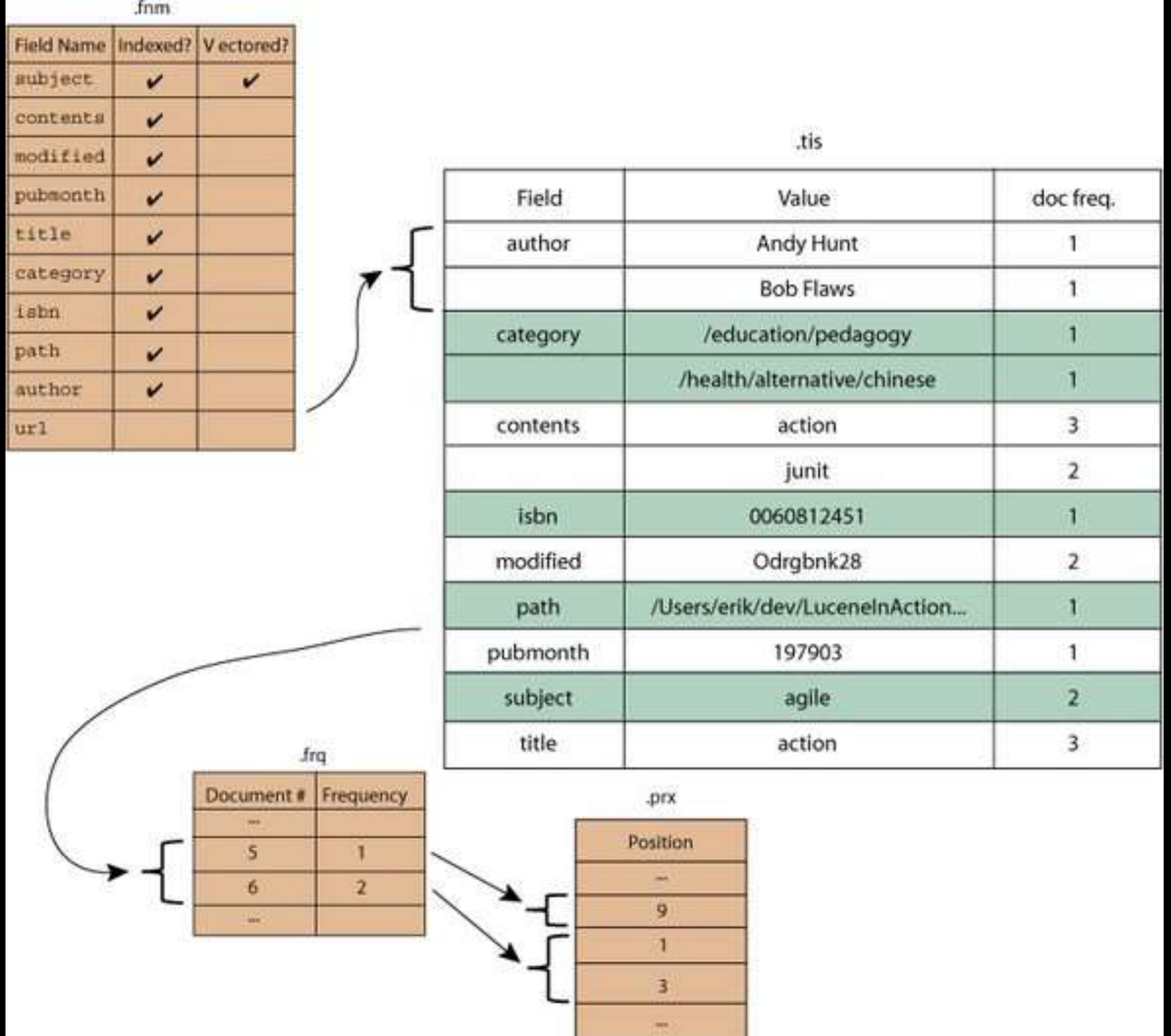
foo AND bar

"foo bar"

"foo bar"~2 → "foo … … bar"

"foo bar" OR "bar foo"

"foo (bar OR baz)"


.fnm
Field Name
Indexed?
Vectored?
subject
contents
modified
pubmonth
title
category
isbn
path
author
url
.tis
Field
Value
doc freq.
author
Andy Hunt
1
Bob Flaws
1
category
/education/pedagogy
1
/health/alternative/chinese
1
contents
action
3
junit
2
isbn
0060812451
1
modified
Odrgbnk28
2
path
/Users/erik/dev/LuceneInAction...
1
pubmonth
197903
1
subject
agile
2
title
action
3
.frq
Document #
Frequency
5
1
6
2
.prx
Position
9
1
3


(опция сбережения или опция забери жжения и надбавка или опус сбережения или (вознаграждения и не (вознаграждения со дня или без взимания вознаграждения или без внимания вознаграждения)) или опция опереженис
коррупции сбережения или функцию сбережения или опция бери женек или опции забери Женей или услуга сбережения или кэшбек движение или опция забери Женей или собственно забери Женя или по сто сбережения или дейс
или опцию переживания или системе сбережения или сбережения подключен или оборот выполнен или стоит сбережения или акции сбережения или бонусную акциз Женя или акции Женя или у тебя сбережения или что-то сбереж
процент забери жене или ((пользуюсь картой или пользуюсь дебетовой картой) и не (не пользуюсь)) или пути сбережения или опции решения или опция без движения или сберегательный Ладу или сейчас опцию подключить
условия по мультикарте или Усов цербер Женя или "пятнадцатого" или сбережения поступают или найдется сбережения или самое сбережения или бонус сбережений или с карты тратишь более или пойдет в оборот или пути сбережен
или опции решения или опция без движения или опция движения или бонусная опция или оптике движения и улица сбережения или полки сбережения или чтобы повышен было или дополнительные один процент или "сбережени
"плюс сбережения" или "функцию надо подключить" или подключил сбережения или дополнительный процент у вас на мастер счет или дополнительный процент у вас на мастер счёт или дополнительный процент на мастер счёт или
дополнительный процент на мастер счет или как подключить накопительный счет копилка или доп опция активировать или дополнительный процент или "по сто сбережения" или опция сбережения или функция сбережения или статус сб
услуга сбережения или подключена в избежание или "волосы сбережения" или "от себя сбережения" или акции сбережения или полки сбережения или палки сбережения или бонусы сбережения или плюс сбережения или от связь сб
пути сбережения или опыта сбережения или категории сбережения или улице сбережения или мерина сбережения или программы сбережения или хлопцы сбережения или опыта сбережения или окна сбережения или полцены сбе
по-волчьи сбережения или программа сбережения или бонус сбережения или опция если покупки или если трачу процентная ставка или сбережения и не "открыть" или сбережения один процент) и не (банк москвы или в раздел
попытки списания средств или какому курсу или пинкод или пин код или аэропорт или снять деньги или карта готова или блокируем карту или заблокируем карту или недостаточно средств или история операций или Грейс перио
карту заблокировать или карту разблокировать или закрытие кары или не могу пользоваться картой или ипотечные счета или заблокировать личный кабинет или реквизиты или между своими счетами или между собственными счетами
"не могу пользоваться картой" или коллекция или зачисления на текущий или ((пристав или арест или судебный пристав или служба судебных или взыскание или исполнительное дело или исполнительное производство или взыс
взыскали счета или "исполнительное дело" или приставы списывают или приставы заблокировали или приставы наложили или постановление приставов или арест приставов или приставы сняли или задолженность приставами или п
приставы отправили или судебный пристав) и не (налог или "сотых приставом" или арест не наложен или исполнительный лист или не связано с приставами или сумма должна была назад вернуться или обращения взыскания или
ограничения у вас на счету нет или ничего не переводил или категория сбережения или личном приставы или бесплатно опция или Ева арестом или открыток арест или приставы политических или облагаться не будет или без комис
или приставами внесения или приставами миллиона или приставами тысяч или пристава бабушка или вознаграждение или новые опцию или активировать счет или мастер счет переводил или мастер счёт переводил или приставам
признак мир или ограничения работа системы или бессрочный или заказать деньги или накопительная пенсия или Банк Москвы или просроченная задолженность или первая страничка или досрочное погашение или блокировка расход
"блокировка расходах операций" или чтобы активировали или блокировка расходных средств или ((выгодный процент или высокий процент или максимальный процент или лучший процент или удобный процент или большой процент
или актуальный процент или актуально процент или дополнительный процент или низкий процент или плюс один процент или процент годовых или какой процент сейчас или дополнительно процент или дополнительный процент или
или нулевые проценты или более выгодный или идет процентная ставка или высвечивается процентная ставка или отражается процентная ставка или указана процентная ставка или отражена процентная ставка или процент действовать
или проценты меньше или процентная ставка какая или повышен процент или повышена ставка или канализация или капитализация или ставка процентная по моему счету или ставка по-моему счету или ставка по моему счету или сам
процентная ставка на сегодняшний день или повыше ставки по вкладам или какие у вас проценты оклады или какой процент по копилке идет или актуальные ставки или действующие ставки или выгодный вариант или первые три месяца
начисляются ежедневно или начисляются каждодневно или проценты на остаток или капитализация процентов или капитализация процентов или как ставка рассчитывается или проценты капиталистического или проценты перечисляются на
проценты перечисляются на матери или условия выплаты процентов или начисление процентов в конце месяца или каждодневная или хорошо процент или хороший процент или начисления как производятся или в течение месяца следую
проценты по депозитным вкладам или зависимости от срока нахождения денежных средств или как идет начисление процентов или начисляется на остаток или сколько у вас денежные средств находятся на счете или когда начисление лежит
как происходит начисление процентов или начисление происходит каждый день или последний день месяца или не терая процент или не терять процент или не потерать процент или какой там процент будет или поступление происходит на м
на минимальный остаток или какой процент или "есть какой-то вклад" или какой процент или самый выгодный или по какому принципу начисляются проценты или ставка на сегодня или большая ставка или лучшая ставка или выгодная с
актуальная ставка или высокая ставка или максимальная ставка или каждый день начисляются или доходность или средней взвешенное или ставка выше или ставкой хотела поинтересоваться или каждый день начисляются или действуют
ставка интересует или процент интересует или взвешанной ставке или "ставка не подтягивается" или "какая у меня ставка" или действует ставка или ставки максимальные или начиная с третьего месяца или при первичном открытии или в
или ставка лучше или максимальный процент или считается три месяца или процентов годовых или при первичном открытии или за количество дней или повышаться процент или процентов на ежедневно остаток или "почему ставка" и
"проценты будут начисляться" или ознакомиться вкладам на нашем сайте или "три месяца проходит" или тарифы на депозит или проценты начислялись на дебетовую карту или "повышенный процент" или поднять процент или первые три ме
или последний календарный день или "система фиксирует остаток" или "объяснить как начислются проценты" или изменилась процентная или "ставка по нулям" или "проценты в конце месяца" или каким образом начисление происходит или
или наиболее выгодно или наиболее выгодный или максимально удобная или максимальная удобная или максимально удобно или лучшее предложение или вклад для пенсионеров или ставка или фактический процент или средневзве
номинальная или эффективная или номинально или эффективно или взвешенная или среднее или впервые или "в первый раз" или первого третий или ставка срок или минимальный срок) и не (заказать деньги или ранее пр
просроченные проценты или кредитная карта или комиссия или бонусные рубли или арест или декларация или закрытие карты или "могу подключить" или "подключаем")))) или ((выгодный процент или высокий процент
или лучший процент или удобный процент или большой процент или доходный процент или актуальный процент или актуально процент или дополнительный процент или низкий процент или плюс один процент или процент годовых

Lucene - Core / LUCENE-7398

Nested Span Queries are buggy

Lucene - Core / LUCENE-5331

nested SpanNearQuery with repeating groups does not find match

Lucene - Core / LUCENE-2861

Search doesn't return document via query

Edit | Add comment | Assign | More | Enable Patch Review | Attach Files | Resolve Issue | Export

Details

Type:	Bug	Status:	OPEN
Priority:	Major	Resolution:	Unresolved
Affects Version/s:	2.9.1, 2.9.4, 3.0.3	Fix Version/s:	None
Component/s:	core/search		
Labels:	None		
Environment:	Doesn't depend on enviroment		
Lucene Fields:	New		

Description

The query doesn't return document that contain all words from query in correct order.

The issue might be within mechanism how do SpanQuerys actually match results (http://www.lucidimagination.com/blog/2009/07/18/the-spanquery/)

Please refer for details below. The example text wasn't passed through snowball analyzer, however the issue exists after analyzing too

People

Assignee:	Unassigned
	Assign to me
Reporter:	Zenoviy Veres
Votes:	1 Vote for this issue
Watchers:	2 Start watching this issue

Dates

Created:	12/Jan/11 13:59
Updated:	05/Sep/16 12:40

Agile

View on Board

Slack

Efficient Optimally Lazy Algorithms for Minimal-Interval Semantics*

Paolo Boldi Sebastiano Vigna
Dipartimento di Informatica, Università degli Studi di Milano

Abstract

Minimal-interval semantics [8] associates with each query over a document a set of intervals, called *witnesses*, that are incomparable with respect to inclusion (i.e., they form an antichain): witnesses define the minimal regions of the document satisfying the query. Minimal-interval semantics makes it easy to define and compute several sophisticated proximity operators, provides snippets for user presentation, and can be used to rank documents. In this paper we provide algorithms for computing conjunction and disjunction that are linear in the number of intervals and logarithmic in the number of operands; for additional operators, such as ordered conjunction and Brouwerian difference, we provide linear algorithms. In all cases, space is linear in the number of operands. More importantly, we define a formal property of *optimal laziness*, and either prove it, or prove its impossibility, for each algorithm. We cast our results in a general framework of finite antichains of intervals on total orders, making our algorithms directly applicable to other domains.

Lucene Query Syntax

```
fn:ordered(quick fn:or(fox dog))
fn:unordered(fn:phrase(brown fox) fn:phrase(fox jumps))
fn:within(fn:or(lazy quick) 1 fn:or(dog fox))
fn:atLeast(2 fn:unordered(furry dog) fn:unordered(brown dog)
                          lazy quick)
```

ELASTICSEARCH

```
{
  "query": {
    "intervals" : {
      "my_text" : {
        "all_of" : {
          "ordered" : true,
          "intervals" : [
            {
              "match" : {
                "query" : "my
favorite food",
                "max_gaps" : 0,
                "ordered" : true
              }
            },
            {
              "any_of" : {
                "intervals" : [
                  { "match" : {
"query" : "hot water" } },
                  { "match" : {
"query" : "cold porridge" } }
                ]
```

XML Retrieval

```
new IntervalQuery(field,
    Intervals.unordered(
        Intervals.phrase(Intervals.term("<family>"),
                Intervals.term("doe"),
                Intervals.term("</family>")
        ),
        Intervals.phrase(Intervals.term("<name>"),
                Intervals.term("joe"),
                Intervals.term("</name>"))
            )
```

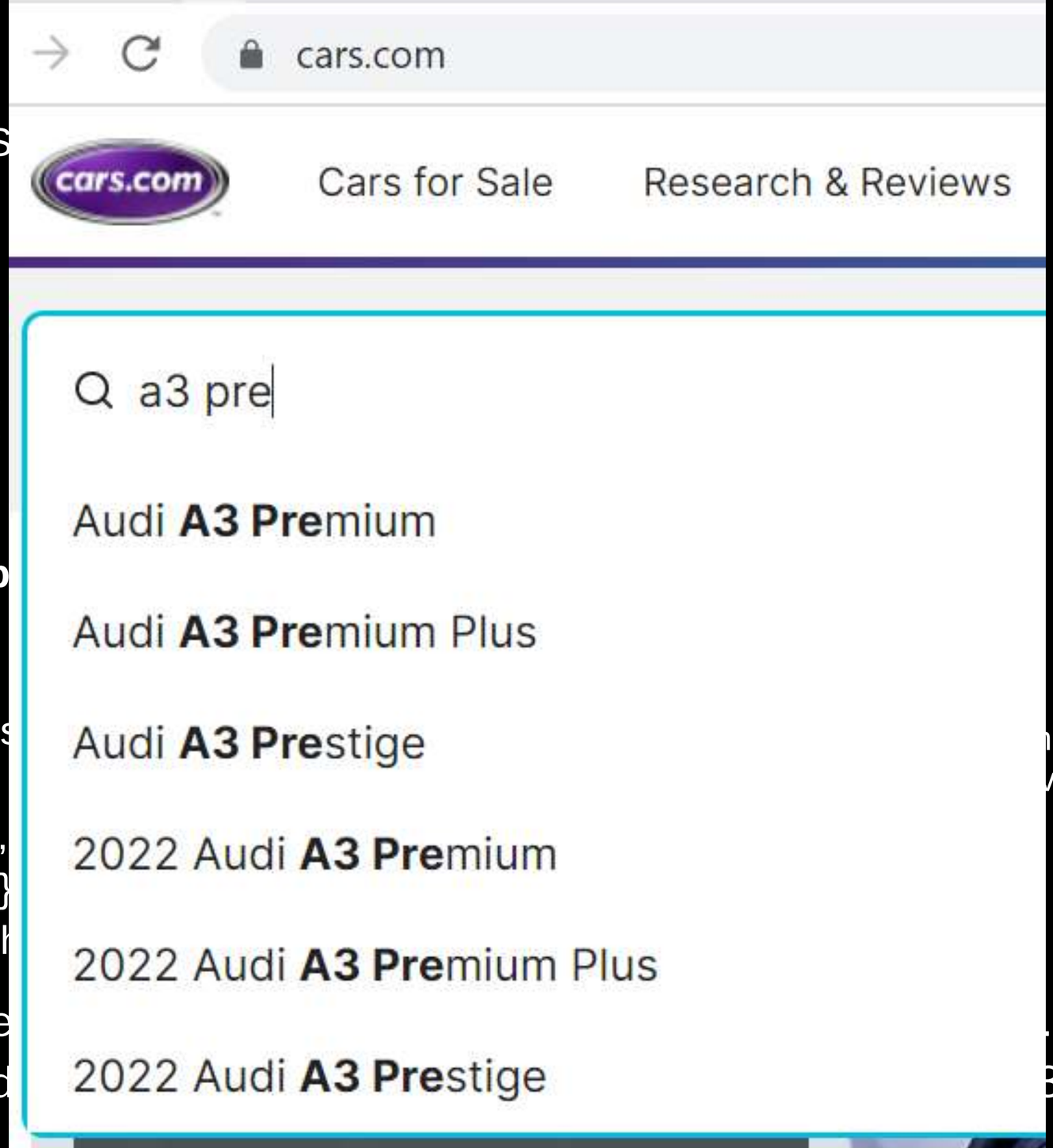
Done

- **Lucene's gang overview**
- `search_as_you_type` **in Elasticsearch**
- **Ancient Boolean Search**
- **Spans vs Intervals**

t.me/MUST_SEARCH

Follow up Discussion

- NGrams are deadly heavy
- Lucene parties
- One more thing about ES subscription
- What makes Lucene look like Hadoop
- Solr vs Hadoop drama




```
"highlight": {
    "suggestion._index_prefix
     "[Audi A3]",
     "2017 [Audi A3]",
     "[Audi A3 Silver]",
    "[Audi A3 Premium]",
     "2017 [Audi A3 Silver]"
   ]
 }
```

Not good highlightin g

3 pre in 0) ..."

TERM DICTIONARY

The **.tim** file contains the list of terms in each field along with per-term statistics (such as docfreq) and **per-term metadata (typically pointers to the postings list** for that term in the inverted index).

TermsDict (.tim) --> Header, FieldDict^{NumFields}, Footer

FieldDict --> *PostingsHeader*, NodeBlock^{NumBlocks}

NodeBlock --> (OuterNode | InnerNode)

OuterNode --> EntryCount, SuffixLength, Byte^{SuffixLength}, StatsLength, < TermStats >^{EntryCount}, MetaLength, <*TermMetadata*>^{EntryCount}

TermMetadata --> (**DocFPDelta**|SingletonDocID), **PosFPDelta**?, PosVIntBlockFPDelta?, PayFPDelta?, SkipFPDelta?

DocFPDelta determines the position of this term's TermFreqs within the .doc file.

PosFPDelta determines the position of this term's TermPositions within the .pos file

POSITIONS

The .pos file contains the lists of positions that each term occurs at within documents. ..

PosFile(.pos) --> Header, <TermPositions> TermCount, Footer
Header --> IndexHeader
TermPositions --> <PackedPosDeltaBlock> PackedPosBlockNum, VIntBlock?
VIntBlock --> <**PositionDelta**[, PayloadLength?], PayloadData?, OffsetDelta?, OffsetLength?>PosVIntCount
PackedPosDeltaBlock --> PackedInts
PositionDelta, OffsetDelta, OffsetLength --> VInt