

Multilingual Search Support in European E-commerce: A Journey with Apache Lucene

Haystack Europe
2023-09-20
@lucianprecup
@a2lean

Haystack On Tour Paris - november 2022



2024 - the year of ...

- 2022 - Vector Search
- 2023 - Large Language Models
- 2024 - ...

2024 - the year of ...

- 2022 - Vector Search
- 2023 - Large Language Models
- 2024 - ... Apache Lucene?

Apache Lucene - 22 years and counting



Apache Lucene - 22 years and counting



Doug Cutting
@cutting

Lucene's FuzzyQuery is 100 times faster!
blog.mikemccandless.com/2011/03/lucene...

9:58 PM · Mar 24, 2011



Uwe Schindler 🇩🇪 🇧🇪 🇫🇷 🇮🇹 🇯🇵
@thetaph1

Mike McCandless talks about [#Apache #Lucene](#) that helps to squash [#Java #JVM](#) bugs: elastic.co/blog/lucene-jv...

9:45 PM · Jul 17, 2015 from Bremen, Germany



Adrien Grand
@jpountz

I ran some benchmarks between Lucene 9.7 and 9.8 (soon to be released), as well as with recursive graph partitioning enabled (-bp): jpountz.github.io/lucene-9.7-vs-... There's a nice speedup on 9.8 alone, and then recursive graph bisection gives another great speedup.

11:46 PM · Sep 13, 2023 · 3,107 Views



Uwe Schindler 🇩🇪 🇧🇪 🇫🇷 🇮🇹 🇯🇵
@thetaph1

[#Apache #Lucene](#) can much faster execute kNN vector queries by calculating dot products / cosine distances using SIMD instructions on AVX2 (x86) and NEON (ARM). It will only work with [#Java20](#) on coming Lucene 9.7 with "--add-modules jdk.incubator.vector":

apache/lucene

#12311 Integrate the Incubating Panama Vector API

🗨️ 165 comments 🗳️ 62 reviews 📁 16 files **+1025 -179** 🇩🇪 🇧🇪 🇫🇷 🇮🇹 🇯🇵

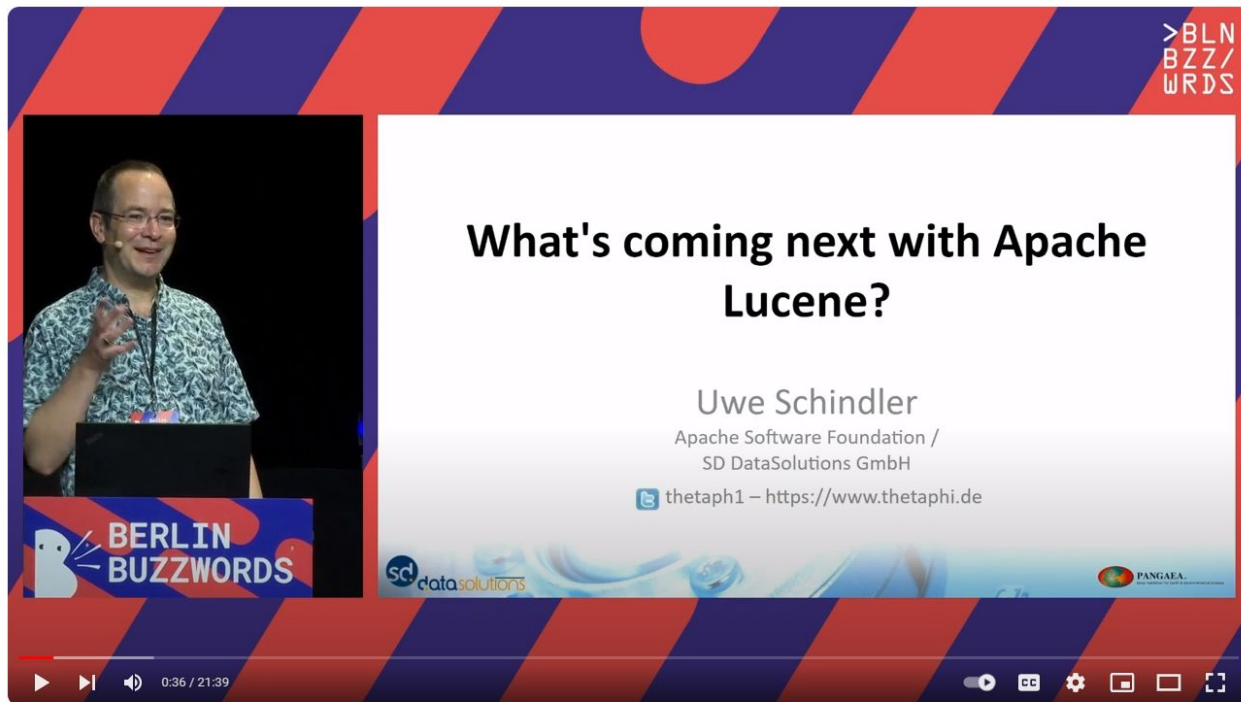
ChrisHegarty · May 18, 2023 -> 59 commits

github.com

Integrate the Incubating Panama Vector API by ChrisHegarty · Pull Request #...
Leverage accelerated vector hardware instructions in Vector Search. Lucene already has a mechanism that enables the use of non-final JDK APIs, currentl...

7:26 PM · May 26, 2023 · 23.1K Views

2031 at Berlin Buzzwords



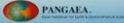


The video player shows a presentation slide with the following content:

- Top Right:** >BLN
BZZ/
WRDS
- Center:**

What's coming next with Apache Lucene?
- Below Title:**

Uwe Schindler
Apache Software Foundation /
SD DataSolutions GmbH
- Below Name:**

 thetaph1 – <https://www.thetaphi.de>
- Bottom Left:** BERLIN BUZZWORDS logo
- Bottom Center:**  data solutions
- Bottom Right:**  PANGAEA

The video player interface includes a progress bar at 0:36 / 21:39 and standard playback controls (play, volume, full screen, etc.) at the bottom.

Uwe Schindler - What's coming next with Apache Lucene?



Plain Schwarz
2.44K subscribers

Subscribe



12



Share

Download

Clip

Save



Thanks to Apache Lucene

Apache Nutch – provides web crawling and HTML parsing

Apache Solr – an enterprise search server

Elasticsearch – an enterprise search server released in 2010

MongoDB Atlas Search – a cloud-native enterprise search application based on MongoDB and Apache Lucene

OpenSearch – an open source enterprise search server based on a fork of Elasticsearch 7

Adelean a2 - an e-commerce and community search server

Who am I ?

Q Adelean

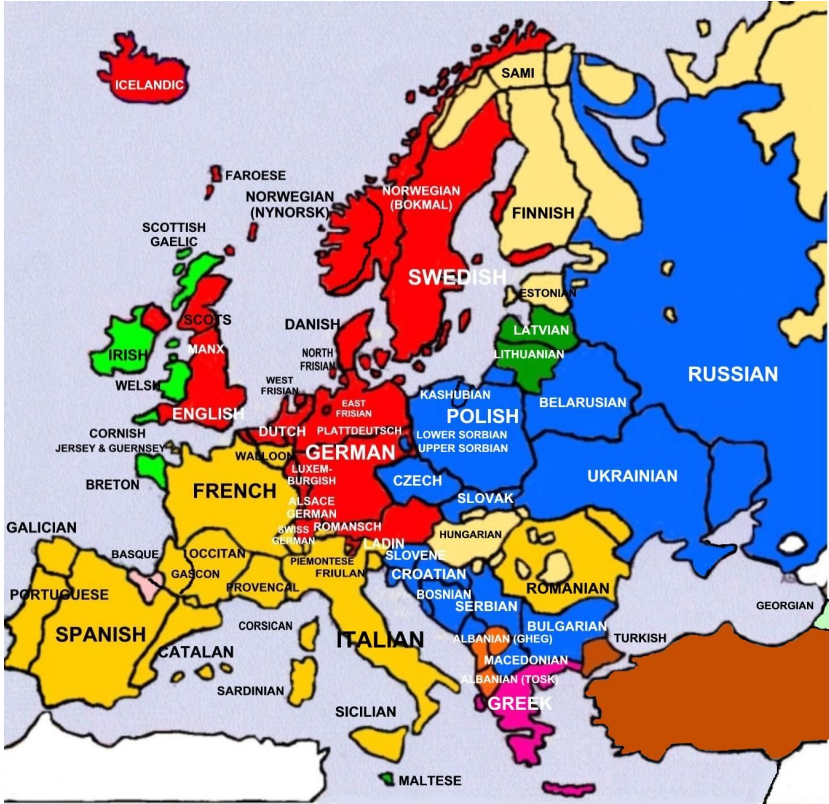
- Q Experts in **Search** technologies
- Q Integrators of **Elasticsearch**, **OpenSearch** and **Solr**
- Q **Consulting** and **Training** providers
- Q Developers of **a2 - E-commerce** and **Enterprise Search** solution
- Q Developers of **all.site** - your **Collaborative** Search Engine



Multilingual Search Support in European E-commerce: A Journey with Apache Lucene

- Lucene levers for processing linguistic specificities
- Tips, tricks, caveats and hacks
- E-commerce specific use cases
- Dealing with person names specific use cases
- Language specific libraries and optimizations
- Getting things done in production today

European languages





OUR ACTIVITY ▾

OUR VISION ▾

JOIN US ▾

PRESS ▾

OUR E-COMMERCE WEBSITES ▾



Casal Sport

ElectricalDirect

Ikaros Cleantech
(Sweden)

Ikaros (Finland)

IronmongeryDirect

Kruizinga

Manutan Belgium

Manutan Collectivités

Manutan Czech Republic

Manutan Denmark

Manutan Finland

Manutan France

Manutan Germany

Manutan Hungary

Manutan Italy

Manutan Netherlands

Manutan Norway

Manutan Poland

Manutan Portugal

Manutan Slovakia

Manutan Spain

Manutan Sweden

Manutan Switzerland

Manutan United
Kingdom

Papeteries Pichon

Rapid Racking



Production et Gestion

Indemnis

 Créer Prospect

RECHERCHES

M. Mle **Dujardin** A CHANCEREL WILLY (A0004 / 000000300)
 M. Melle **Dujardin** PIERRE & (00200 / 000303000)
 M. Denis **DUJARDIN** (00020 / 000100000)
 Mme Patricia **DUJARDIN** (A0001 / 000090000)
 M. Philippe **DUJARDIN** (00500 / 010000000)
 M. Matthieu **DUJARDIN** (A1000 / 000000200)
 Mme Edith **DUJARDIN** (A0000 / 010100000)
 M. Robert **DUJARDIN** (00000 / 000000090)
 M. Xavier **DUJARDIN** (A0003 / 010201200)
 Mme Sylvie **DUJARDIN** (00090 / 010200202)

2201 RESULTATS TROUVES

Trier les résultats par

Pertinence ▾

Essayez avec cette orthographe : DUJARDIN



M. Denis DUJARDIN - Prospect Particulier - A1001 / 010101011
 2 RUE DE PARIS, 75000 PARIS
 Intermédiaire : 00000A



Mme Patricia DUJARDIN - Prospect Particulier - A2001 / 010101012
 1 RUE DE VERSAILLES, 78000 VERSAILLES
 Né(e) le: 31/12/2012 | Intermédiaire : 10001A



M. Philippe DUJARDIN - Prospect Particulier - A0006 / 010202020
 25 B RUE DU VILLAGE, 59000 METZ
 Né(e) le: 01/01/1970 | Intermédiaire : 10000A



DUJARDIN - Prospect Entreprise - A0007 / 012012012
 17 BOULEVARD DU GENERAL DE GAULLE, 75000 Paris
 Intermédiaire : 600000



SARL DUJARDIN - Prospect Entreprise - A1234 / 101010101
 12 RUE DE LA MONTAGNE, 93000 SAINT DENIS
 SIRET : 11008800000010 | Intermédiaire : 15000A

AFFINER LA RECHERCHE

▼ Type client

- Prospect (1647)
 Client (397)
 Ancien client (157)

▼ Nature personne

- Particulier (1939)
 Professionnel (116)
 Entreprise (114)
 Autre (28)
 Copropriété (3)
 Association (1)

[Effacer les filtres](#)Afficher résultats par page

1 2 ... 221

Suivant >

Haut de page ↕

1786 Results

Language

- English 1736
- French 30
- 12
- pt 7
- ko 1



- Lisa Jung est une développeuse avocate chez Elastic. Elle aime se plonger dans les données et enseigner aux développeurs comment rechercher, analyser et visualiser les données avec Elastic Stack. Pour plus d'infos, [cliquez ici](#).
- Il y a une erreur de propriété manquante dans une réponse de l'API Java Elasticsearch. Pour plus d'infos sur la résolution de ce problème, [cliquez ici](#).
- 정수 김 est un conférencier qui a remporté le prix du meilleur projet de données à la conférence Elastic. Pour plus d'infos, [cliquez ici](#).
- Il y a des informations sur les listes et les cartes dans le client Java Elasticsearch. Pour plus d'infos, [cliquez ici](#).
- Il y a une session sur les nouvelles fonctionnalités du client Python Elasticsearch. Pour plus d'infos, [cliquez ici](#).

**Lisa Jung** | 450bb3d9-26e9-4c37-88c2-de5f28cbb93d

Overview Schedule Lisa Jung Developer Advocate | Elastic Lisa Jung is a developer advocate at Elastic. She loves geeking out over data and teaching developers how they can search, analyze, and visualize data with Elastic Stack. Her journey to landing a dream job as a developer advocate has been an u ...

**MissingRequiredPropertyException in a response | Elasticsearch Java API Client [7.17] | Elastic** | [missing-required-property.html](#)

A newer version is available. For the latest information, see the current release documentation. Elastic Docs > Elasticsearch Java API Client [7.17] > Troubleshooting < TroubleshootingNoSuchMethodError RequestOptions\$Builder.removeHeader when creating a client > MissingRequiredPropertyException in a respons ...

**정수 김** | df68d309-431b-4611-8f9e-e7a09c556021

Overview Schedule 정수 김 상명대학교 2016 ~ 2022 상명대학교 게임전공, 인공지능융합전공 (서울, 종로구) 2021 상명 데이터콘서트 최우수상 2021 데이터야놀자 세션 발표 2021 Elastic Fundamental Training Talks 대학생의 우당탕탕 엘라스틱 고군분투 이야기

**Lists and maps | Elasticsearch Java API Client [7.17] | Elastic** | [lists-and-maps.html](#)

Lucene text analysis

Index time

Input documents →

Id	Nom
1	Céline
2	Celia

Ascii folding →

Celine, Celia

Lowercase →

celine, celia

Index

Key	Document id
celine	1
celia	2

Search time

Nom
CÉLINE

← Search term

CELINE ← Ascii folding

celine ← Lowercase

Lucene text analysis

Index time

Input document →

Id	Nom
1	I'm a developer in Berlin
2	She develops software

Lowercase →

i'm a developer in berlin
she develops software

Stop →

i'm developer berlin
she develops software

Stemmer →

i'm develop berlin
she develop softwar

Search time

Nom
Software Development

← Search term

software development ← Lowercase

software development ← Stop

softwar develop ← Stemmer

Index

Key	Document id
softwar	2
develop	2
...	...

_analyze API (Elasticsearch and OpenSearch)

```
GET _analyze
{
  "text": [
    "I'm a developer in Berlin"
  ],
  "tokenizer": "standard",
  "filter": [
    "lowercase",
    {
      "type": "stop",
      "stopwords": "_english_"
    },
    {
      "type": "stemmer",
      "language": "english"
    }
  ]
}
```

```
"tokens": [
  {
    "token": "i'm",
    "start_offset": 0,
    "end_offset": 3,
    "type": "<ALPHANUM>",
    "position": 0
  },
  {
    "token": "develop",
    "start_offset": 6,
    "end_offset": 15,
    "type": "<ALPHANUM>",
    "position": 2
  },
  {
    "token": "berlin",
    "start_offset": 19,
    "end_offset": 25,
    "type": "<ALPHANUM>",
    "position": 4
  }
]
```

Analysis menu in the Solr Admin dashboard

Solr Admin

localhost:8983/solr/#/alias/analysis?analysis.fieldvalue=Berlin&analysis.query=Paris&analysis.fieldname=

Solr

Logout solr
Dashboard
Logging
Security
Cloud
Schema Designer
Collections
Java Properties
Thread Dump
Suggestions

alias

Overview
Analysis
Dataimport
Documents
Files
Query
Stream
Schema

Core Selector

Field Value (Index)
Berlin

Field Value (Query)
Paris

Analyse Fieldname / FieldType: title Schema Browser

Verbose Output Analyse Values

SI	text	Berlin
	raw_bytes	[42 65 72 6c 69 6e]
	start	0
	end	6
	positionLength	1
	type	<ALPHANUM>
	termFrequency	1
	position	1

SE	text	Berlin
	raw_bytes	[42 65 72 6c 69 6e]
	start	0
	end	6
	positionLength	1
	type	<ALPHANUM>
	termFrequency	1
	position	1

LCF	text	berlin
	raw_bytes	[62 65 72 6c 69 6e]
	start	0
	end	6
	positionLength	1
	type	<ALPHANUM>
	termFrequency	1
	position	1

SI	text	Paris
	raw_bytes	[50 61 72 69 73]
	start	0
	end	5
	positionLength	1
	type	<ALPHANUM>
	termFrequency	1
	position	1

SE	text	Paris
	raw_bytes	[50 61 72 69 73]
	start	0
	end	5
	positionLength	1
	type	<ALPHANUM>
	termFrequency	1
	position	1

SGF	text	Paris
	raw_bytes	[50 61 72 69 73]
	start	0
	end	5
	positionLength	1
	type	<ALPHANUM>
	termFrequency	1
	position	1

LCF	text	paris
	raw_bytes	[70 61 72 69 73]
	start	0

Language analyzers - the simple (and wrong) solution

Language analyzer

OpenSearch supports the following language values with the `analyzer` option: `arabic`, `armenian`, `basque`, `bengali`, `brazilian`, `bulgarian`, `catalan`, `czech`, `danish`, `dutch`, `english`, `estonian`, `finnish`, `french`, `galician`, `german`, `greek`, `hindi`, `hungarian`, `indonesian`, `irish`, `italian`, `latvian`, `lithuanian`, `norwegian`, `persian`, `portuguese`, `romanian`, `russian`, `sorani`, `spanish`, `swedish`, `turkish`, and `thai`.

To use the analyzer when you map an index, specify the value within your query. For example, to map your index with the French language analyzer, specify the `french` value for the analyzer field:

```
"analyzer": "french"
```

<https://opensearch.org/docs/latest/analyzers/language-analyzers/>

<https://www.elastic.co/guide/en/elasticsearch/reference/8.7/analysis-lang-analyzer.html>

Stemming

- Lemma : canonical form - dictionary (développeuse → développer)
- Stem : reduced form - algorithmical (développeuse → developeu)

- Stemming is faster and shipped out of the box
- Lemmatization may require a license

- Stemming is largely recommended by Elastic
 - <https://www.elastic.co/guide/en/elasticsearch/reference/8.7/stemming.html#dictionary-stemmers>
 - <https://www.elastic.co/guide/en/elasticsearch/reference/8.7/stemming.html#algorithmic-stemmers>

But stemming is not perfect

Brands (Gillette = gilet, Barbie = barbe with the light_french stemmer)

Technical limitations (travail = travail, travaux = traval with the light_french stemmer)





Irregular words

Similar spelling (broker = broken)

gilette

493 résultats pour "gilette", Voulez vous rechercher **gilet** ou **galette** ou **galettes** ?



Trier par : **Pertinence** Prix

 <p>Manutan Gilet haute visibilité - Manutan 3.75€ HT l'unité</p>	 <p>Portwest Gilet haute visibilité réversible orange - Portwest 70.90€ HT l'unité</p>	 <p>Blaklader Gilet haute visibilité fluorescent 22.90€ HT l'unité</p>	 <p>Portwest Gilet haute visibilité Executive Berlin jaune - Portwest 56.75€ HT l'unité</p>
---	--	--	--

barbie

2 résultats pour "barbie", Voulez vous rechercher **barre** ou **baie** ou **baril** ?

Trier par : **Pertinence** Prix

 <p>- Lot de 25 batonnets pour machine barbe à papa - Scrapcooking 34.50€ HT l'unité</p>	 <p>MP Hygiene Cache barbe jetable 4.35€ HT l'unité</p>
---	---

Details of the convenience language analyzers

```
"analyzer": {  
  "rebuilt_german": {  
    "tokenizer": "standard",  
    "filter": [  
      "lowercase",  
      "german_stop",  
      "german_keywords",  
      "german_normalization",  
      "german_stemmer"  
    ]  
  }  
}
```

```
"filter": {  
  "german_stop": {  
    "type": "stop",  
    "stopwords": "german"  
  },  
  "german_keywords": {  
    "type": "keyword_marker",  
    "keywords": ["Beispiel"]  
  },  
  "german_stemmer": {  
    "type": "stemmer",  
    "language": "light_german"  
  }  
}
```

Lucene levers for algorithmic stemmers

Protected words and stemmer override

- Keyword marker
- Stemmer override

And also

- Conditional token filter
- stem_exclusion parameter for language analyzers

Keyword marker

```
GET /_analyze
{
  "tokenizer": "standard",
  "filter": [
    {
      "type": "stop",
      "ignore_case": true,
      "stopwords": [
        "_french_"
      ]
    },
    {
      "type": "stemmer",
      "language": "light_french"
    },
    "lowercase"
  ],
  "text": "Barbie barbe, Gillette gilet, La Croix croissant"
}
```

```
{
  "tokens" : [
    {
      "token" : "barb",
      "start_offset" : 0,
      "end_offset" : 6,
      "type" : "<ALPHANUM>",
      "position" : 0
    },
    {
      "token" : "barb",
      "start_offset" : 7,
      "end_offset" : 12,
      "type" : "<ALPHANUM>",
      "position" : 1
    },
    {
      "token" : "gilet",
      "start_offset" : 14,
      "end_offset" : 21,
      "type" : "<ALPHANUM>",
      "position" : 2
    },
    {
      "token" : "gilet",
      "start_offset" : 22,
      "end_offset" : 27,
      "type" : "<ALPHANUM>",
      "position" : 3
    },
    {
      "token" : "croi",
      "start_offset" : 32,
      "end_offset" : 37,
      "type" : "<ALPHANUM>",
      "position" : 5
    },
    {
      "token" : "croi",
      "start_offset" : 38,
      "end_offset" : 47,
      "type" : "<ALPHANUM>",
      "position" : 6
    }
  ]
}
```


Keyword marker

```
GET /_analyze
{
  "tokenizer": "standard",
  "filter": [
    {
      "type": "stop",
      "ignore_case": true,
      "stopwords": [
        "_french_"
      ]
    },
    {
      "type": "keyword_marker",
      "keywords": [
        "Barbie",
        "Gillette",
        "La Croix"
      ],
      "ignore_case": true
    },
    {
      "type": "stemmer",
      "language": "light_french"
    },
    "lowercase"
  ],
  "text": "Barbie barbe, Gillette gilet, La Croix croissant"
}
```

```
{
  "tokens": [
    {
      "token": "barbie",
      "start_offset": 0,
      "end_offset": 6,
      "type": "<ALPHANUM>",
      "position": 0
    },
    {
      "token": "barb",
      "start_offset": 7,
      "end_offset": 12,
      "type": "<ALPHANUM>",
      "position": 1
    },
    {
      "token": "gilette",
      "start_offset": 14,
      "end_offset": 21,
      "type": "<ALPHANUM>",
      "position": 2
    },
    {
      "token": "gilet",
      "start_offset": 22,
      "end_offset": 27,
      "type": "<ALPHANUM>",
      "position": 3
    },
    {
      "token": "croi",
      "start_offset": 32,
      "end_offset": 37,
      "type": "<ALPHANUM>",
      "position": 5
    },
    {
      "token": "croi",
      "start_offset": 38,
      "end_offset": 47,
      "type": "<ALPHANUM>",
      "position": 6
    }
  ]
}
```

Stemmer override

```
PUT /my-index-000001
{
  "settings": {
    "analysis": {
      "analyzer": {
        "my_analyzer": {
          "tokenizer": "standard",
          "filter": [ "lowercase", "custom_stems", "porter_stem" ]
        }
      },
      "filter": {
        "custom_stems": {
          "type": "stemmer_override",
          "rules": [
            "running, runs => run",
            "stemmer => stemmer"
          ]
        }
      }
    }
  }
}
```

Dictionary stemmers

- Well suited for
 - Stemming irregular words
 - Discerning between words that are spelled similarly but not related conceptually (broker // broken)
- In practice
 - Algorithmic stemmers typically outperform dictionary stemmers (via Elastic)
<https://www.elastic.co/guide/en/elasticsearch/reference/8.7/stemming.html#dictionary-stemmers>
 - Dictionary quality
 - Size and performance

Lemmatization in e-commerce

- Performs better than stemming if ...
- You have a custom dictionary
- Example for French

```
"lemmagen_fr" : {  
  "type" : "lemmagen",  
  "lexicon" : "fr"  
}
```

<https://github.com/adelean/elasticsearch-analysis-lemmagen>

- IMPORTANT! - see [License](#) chapter.
(<https://github.com/hlavki/jlemmagen#markdown-header-license>)

Stop words

- Exemples: the, in, a, for, ...
- Side effects when words are not removed :
 - “case for xylophone” should return zero results but might return “case for ...” if a rule like « 2<70% » was used for `minimum_should_match`
 - another example (French) : « masque pour bébés »
- Issues with the default Elasticsearch implementation
 - Documented here : [Meetup ElasticFR #61 - Transition du filtre Synonym vers SynonymGraph - YouTube](#) and here: [Bug: When using graph synonym and stop token filter together · Issue #28838 · elastic/elasticsearch · GitHub](#)
 - Workaround : implementing stop words like `char_filter`


The screenshot shows a search results page for the query "masque pour bebés". The search bar at the top right contains the text "masque pour bebés". Below the search bar, it indicates "63 résultats pour 'masque pour bebés', Veuillez vous rechercher masque pour buses ou masque pour...". The results are sorted by "Pertinence" (Relevance) and "Prix" (Price). Three product cards are visible:

- SAM**: Masque protection pour intervention électrique, priced at 118.25€ HT l'unité.
- Deltaplus**: Masque de soudure pour casque de chantier CASOUD2HE, priced at 29.90€ HT l'unité.
- A partially visible card on the right for "Masque se soudage".

Issues with stopwords and graphs in detail


YouTube FR


Search

elastic | 

Transition du filtre
Synonym vers SynonymGraph

06/05/2021


Vincent Bosc
Développeur Java Sénior

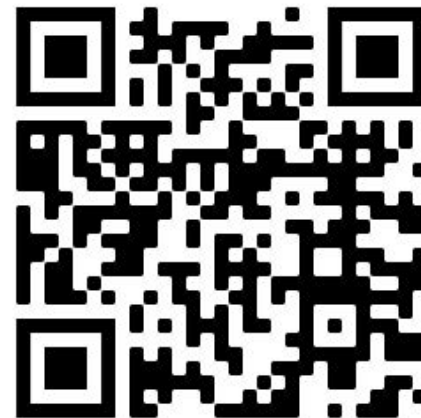

adelean
ÉVÉNEMENTS D'ÉCHANGE

David @ Elastic

Vincent Bosc

1:40 / 39:29

<https://www.youtube.com/watch?v=DcjmhkeQt-I>



Stop words

Contournement : implementer les stop words comme char_filter :

```
"mdf_stop_words": {  
  "type": "pattern_replace",  
  "pattern":  
  "(\\s|^|\\.|-|\\s,)(de|la|des|du|pour|le|les|un|une|au|et) (  
  ?=\\s|$|\\.|-|\\s,)",  
  "replacement": " ",  
  "flags": "CASE_INSENSITIVE"  
}
```

Searching for phrases: detecting the most important word











- orange à jus 🍊 versus jus d'orange 🥤
- The problem:
 - when searching for a single word (ex. oranges) : making sure that the right products pop up first
 - when searching for a phrase and getting zero results : making sure the new search is launched with the most meaningful word
- Simple hack for Latin languages:
 - The most important word is always the first word 😊

The problem with oranges

Accueil > oranges

Promo (48) Affiner par rayon Bio (9) Marque TRIER PAR

Ma recherche : "oranges" 107 résultats trouvés

 <p>ORIGINE C.E.E.</p>	 <p>ESPAGNE</p>	 <p>ESPAGNE</p>	<p>PRENEZ EN 3 = PAYEZ EN 2</p> 	<p>2=10%, 3=15%, 4=20%...</p> 
<p>Orange Navel Cat 1 Cal 43319 Carrefour le filet de 750g</p>	<p>Oranges à dessert Navel Cat.1 le filet de 2 kg</p>	<p>Oranges à jus Salustiana Cat.1 le filet de 2 kg</p>	<p>Jus d'orange avec pulpe Tropicana la bouteille de 2L</p>	<p>Jus d'orange pulpé Carrefour la brique de 2l</p>
<p>0,99€ 1.32 € / Kilogramme</p>	<p>2,79€ 1.40 € / Kilogramme</p>	<p>2,99€ 1.50 € / Kilogramme</p>	<p>3,49€ 1.75 € / L</p>	<p>2,67€ 1.34 € / L</p>
				

The problem, explained

The screenshot shows the Kibana Dev Tools console with the following content:

```
21 - }
22 - }
23 - }
24 -
25 PUT orange_test/product/2
26 {
27   "title" : "jus d'orange"
28 - }
29 -
30 PUT orange_test/product/1
31 {
32   "title" : "oranges à jus"
33 - }
34 -
35 GET orange_test/_analyze
36 {
37   "field" : "title",
38   "text" : "oranges à jus"
39 - }
40 -
41 GET orange_test/_analyze
42 {
43   "field" : "title",
44   "text" : "jus d'orange"
45 - }
46 -
47 GET orange_test/_search
48 {
49   "query": {
50     "match_all": {}
51 - }
52 - }
53 -
54 GET orange_test/_search
55 {
56   "explain": true,
57   "query": {
58     "multi_match": {
59       "query": "oranges",
60       "fields": [
61         "title"
62 - ]
63 - }
64 - }
65 - }
66 -
```

```
1 - {
2   "took": 2,
3   "timed_out": false,
4 - "_shards": {
5     "total": 1,
6     "successful": 1,
7     "skipped": 0,
8     "failed": 0
9 - },
10 - "hits": {
11   "total": 2,
12   "max_score": 0.18232156,
13 - "hits": [
14 - {
15   "_shard": "[orange_test][0]",
16   "_node": "ZudBSfc3THEBnHTDet7bBQ",
17   "_index": "orange_test",
18   "_type": "product",
19   "_id": "2",
20   "_score": 0.18232156,
21 - "_source": {
22     "title": "jus d'orange"
23 - },
24   "_explanation": {
25     "value": 0.18232156,
26     "description": "weight(title:orang in 0) [PerFieldSimilarity], result of:",
27     "details": [
28       {
29         "value": 0.18232156,
30         "description": "score(doc=0,freq=1.0 = termFreq=1.0\n, product of:",
31         "details": [
32           {
33             "value": 0.18232156,
34             "description": "idf, computed as log(1 + (docCount - docFreq + 0.5) / (docFreq + 0.5)) from:",
```

The solution

Indices > getstarted_actors

+ NEW INDEX

MANAGE CURRENT INDEX

PLAY QUERY RULES REPLICAS LOGS STATS UI DEMOS

searchableAttributes

List of attributes you want to search in. For ranking purpose, you must order added attributes from the most to the least important.

Tip: If you want to give the same ranking priority to several attributes, just put them on the same line, separated by a comma.

THE → ORDERED OPTION

By default, matches at the beginning of an attribute end up with a better rank than the ones at the middle/end of it: the query "iPhone", "iPhone 6S" will be sorted before "Case for iPhone". You can disable this behavior by selecting the **Unordered** option.

(ordered by importance)

name → ORDERED

alternative_name → ORDERED

+ ADD A SEARCHABLE ATTRIBUTE

List of attributes you want to search in. Order them by importance, with the most important on top.

Need help? →

The solution with Apache Lucene (via Adelean a2)

Searchable fields ENREG.

Ajouter un nouveau champ de recherche :

Nom du champ +

Hierarchies.Nationale.Level1.Level...
3 🗑️

Hierarchies.Nationale.Level1.Level...
3 🗑️

Hierarchies.Nationale.Level1.Level...
3 🗑️

brandName
2 🗑️

pdctProductSubBrand
2 🗑️

productSimpleView.origin
1 🗑️

GÉNÉRALE **RECHERCHE** FACETTE SYNONYMES BOOST AUTOCOMPLÉTION EXCLUSIONS DICTIONNAIRE

Est une référence ?

Peut être cherché ?

Pondération
80

Recherche approximative

Entête	Pondération en entête
1	1

Phrase Pondération en phrase

2 0,5

Langue Pondération de la lan...

french 2





The solution with Apache Lucene (via Adelean a2)

oranges X Entrez un produit RECHERCHE ⚙

Voulez-vous dire [orange](#) [oranger](#) [orange](#) ?

332 résultats en 183 ms Afficher 10 par page Trier par Pertinence « ‹ 1 2 3 4 5 › »

Score : 2.59362 ▾

1		alcohol_by_volume_label - assortiments.assortiment.AssortStartDate - 2017-07-07 ean - 3000000034460 eligibleProduct - true ES_SansGluten - false flagAOP - false flagDeconseilleFemmesEnceintes - false flagEngagementQualCarrefour - false flagHalal - false flagGP - false flagLabelRouge - false flagSpecialiteTradGarant
2		alcohol_by_volume_label - assortiments.assortiment.AssortStartDate - 2016-03-03 ean - 3276552299644 eligibleProduct - true ES_SansGluten - false flagAOP - false flagDeconseilleFemmesEnceintes - false flagEngagementQualCarrefour - false flagHalal - false flagGP - false flagLabelRouge - false flagSpecialiteTradGarant
3		alcohol_by_volume_label - assortiments.assortiment.AssortStartDate - 2016-03-03 ean - 3276557103861 eligibleProduct - true ES_SansGluten - false flagAOP - false flagDeconseilleFemmesEnceintes - false flagEngagementQualCarrefour - false flagHalal - false flagGP - false flagLabelRouge - false flagSpecialiteTradGarant
4		alcohol_by_volume_label - assortiments.assortiment.AssortStartDate - 2017-10-10 ean - 3276552299347 eligibleProduct - true ES_SansGluten - false flagAOP - false flagDeconseilleFemmesEnceintes - false flagEngagementQualCarrefour - false flagHalal - false flagGP - false flagLabelRouge - false flagSpecialiteTradGarant

Valeur du boost: 1 MODIFIER

- 2.59362 - f(x)
- 2.59362 - max(a, b)
- 2.58482 - Σ
- 0.70143 - Π
- 5.26071 - Σ
 - 0.06536 - Δ -
functionalName.value.simple:oranges^2.0
 - 5.19536 - Δ -
pdctProductNature.value.simple:oranges^80.0
- 0.13333 - coord(2/15)
- 0.77063 - Σ
- 0.00294 - Π
- 0.00588 - Σ
 - 0.00588 - Δ -
functionalName.value.first:orang^0.2
- 0.50000 - coord(1/2)
- 0.76769 - Π
- 1.11276 - Π
- 2.59362 - Σ

Promo ▾



- RD_CRESCENDO (67)
- RI (22)
- RD (2)
- PROMO (1)
- PF (1)

Affiner par rayon ▾

- Fruits et Légumes (60)
- Crèmerie (3)
- Suralés (3)

How about non latin languages?

case for *iphone* versus *iphone case*

With our Czech colleagues we produced the following rule (valid for Czech , English  and other similar languages):

Place the accent (boost) on the second word UNLESS the phrase contains one of the connecting particles (for, about, on, to, etc. - *specific list to be given in the Business Console for each language*)

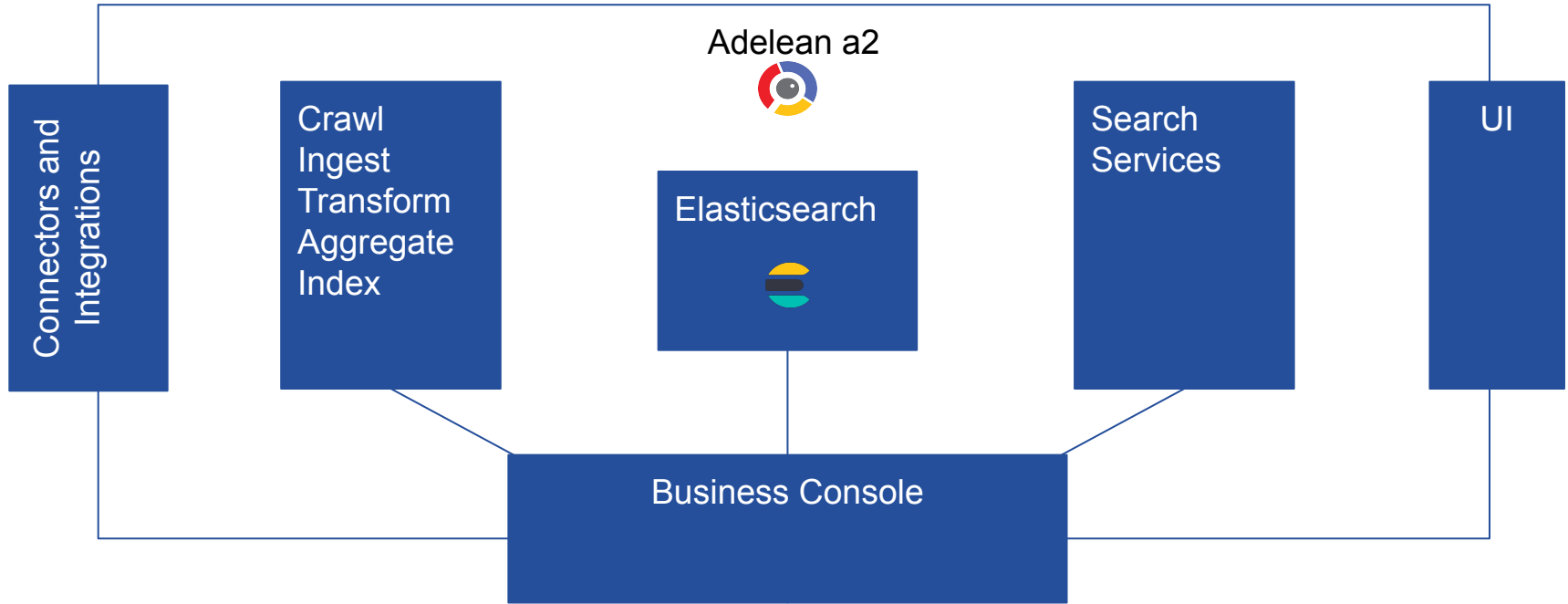
boxy *na zavěšení*

boxy - boxes - main subject

na - for - particle

zavěšení - hanging - use

The Business Console, window to Lucene config



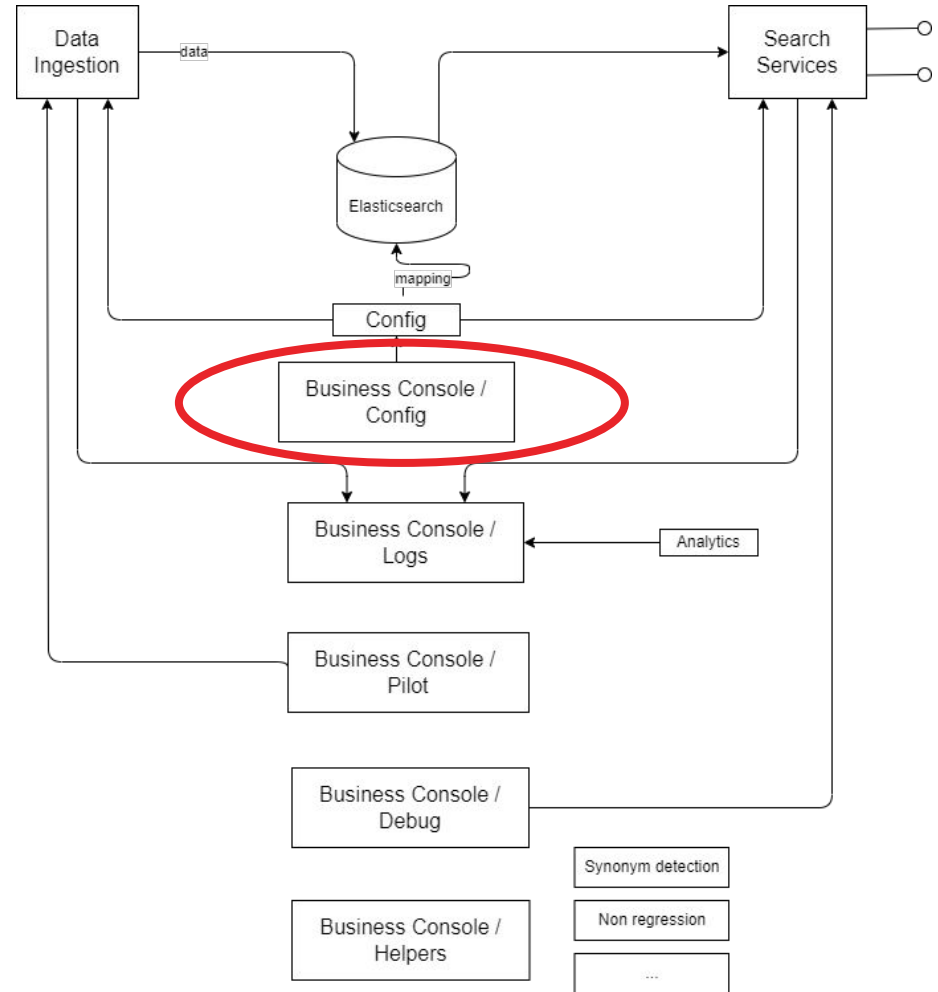
Business Console

Monitoring tools

Actions and configurations

Merchandising

...



<https://all.site/management/>



OBSERVE

Dashboard

↑ ENHANCE

Synonyms

Landing page

Replacements

Stop words

Protected words



Search terms		Number of search queries	Exit rate after a search	Refinement rate	Number of search results	Status	
security shoes		Value ↑ 0.5%	Value ↑ 0.5%	Value ↑ 0.5%	Value ↑ 0.5%	No rule	Add rule
anti-slip adhesive tape for stairs		Value ↑ 0.5%	Value ↑ 0.5%	Value ↑ 0.5%	Value ↑ 0.5%	No rule	Add rule
rack for jars		Value ↑ 0.5%	Value ↑ 0.5%	Value ↑ 0.5%	Value ↑ 0.5%	No rule	Add rule
adi blu		Value ↑ 0.5%	Value ↑ 0.5%	Value ↑ 0.5%	Value ↑ 0.5%	No rule	Add rule
warehouse broom		Value ↑ 0.5%	Value ↑ 0.5%	Value ↑ 0.5%	Value ↑ 0.5%	No rule	Add rule
warehouse broom		Value ↑ 0.5%	Value ↑ 0.5%	Value ↑ 0.5%	Value ↑ 0.5%	No rule	Add rule
warehouse broom		Value ↑ 0.5%	Value ↑ 0.5%	Value ↑ 0.5%	Value ↑ 0.5%	No rule	Add rule
warehouse broom		Value ↑ 0.5%	Value ↑ 0.5%	Value ↑ 0.5%	Value ↑ 0.5%	No rule	Add rule
warehouse broom		Value ↑ 0.5%	Value ↑ 0.5%	Value ↑ 0.5%	Value ↑ 0.5%	No rule	Add rule
warehouse broom		Value ↑ 0.5%	Value ↑ 0.5%	Value ↑ 0.5%	Value ↑ 0.5%	No rule	Add rule

MAF



Log out

1 2 3

50 / 1 500

But NLP is better at this task

SIMPLE_SENTENCE = "I want to buy an iPhone case";

SIMPLE_PHRASE_WITH_PARTICLE = "case for iPhone";

SIMPLE_PHRASE = "iPhone case";

COMPLEX_SENTENCE = "I want to buy a small orange Apple iPhone case";

COMPLEX_SENTENCE_YELLOW = "I want to buy a small yellow Apple iPhone case";



<https://github.com/adelean/opennlp-learning>

Training data : [opennlp-learning/src/main/resources/en-ner-products.train at master · adelean/opennlp-learning \(github.com\)](https://github.com/adelean/opennlp-learning/blob/master/src/main/resources/en-ner-products.train) 42

Special use cases - Hungarian

```
"analyzer": {  
  "rebuilt hungarian": {  
    "tokenizer": "standard",  
    "filter": [  
      "lowercase",  
      "hungarian stop",  
      "hungarian keywords",  
      "hungarian stemmer"  
    ]  
  }  
}
```

```
"analysis": {  
  "filter": {  
    "hungarian_stop": {  
      "type": "stop",  
      "stopwords": " hungarian "  
    },  
    "hungarian keywords": {  
      "type": "keyword_marker",  
      "keywords": ["példa"]  
    },  
    "hungarian_stemmer": {  
      "type": "stemmer",  
      "language": "hungarian"  
    }  
  }  
}
```

Hunspell performs better

Special use cases - Hungarian

Hunspell token filter

Provides [dictionary stemming](#) based on a provided [Hunspell dictionary](#). The `hunspell` filter requires [configuration](#) of one or more language-specific Hunspell dictionaries.

This filter uses Lucene's [HunspellStemFilter](#).



TIP

If available, we recommend trying an algorithmic stemmer for your language before using the `hunspell` token filter. In practice, algorithmic stemmers typically outperform dictionary stemmers. See [Dictionary stemmers](#).

However, `hunspell` performs better for Hungarian 😊
Dictionaries (in [LibreOffice](#), [LibreOffice extensions](#), [Mozilla Add-Ons](#))

Special use cases - how about Finnish ?

lentokonesuihkuturbiinimoottoriapumekaanikkoaliupseerioppilas

"airplane jet turbine engine auxiliary mechanic non-commissioned officer student"

https://en.wikipedia.org/wiki/Longest_words

Hunspell does not perform well for Finnish

Voikko Analysis for Elasticsearch

<https://github.com/EvidentSolutions/elasticsearch-analysis-voikko> deprecated

Raudikko Analysis for Elasticsearch

<https://github.com/EvidentSolutions/elasticsearch-analysis-raudikko> works with

Elasticsearch 8.x

Raudikko Analysis for Elasticsearch

<https://github.com/EvidentSolutions/elasticsearch-analysis-raudikko>

Parameter	Default value	Description
analyzeAll	true	Use all analysis possibilities or just the first
splitCompoundWords	false	Split analysed compound words to its parts
minimumWordSize	3	minimum length of words to analyze
maximumWordSize	100	maximum length of words to analyze
analysisCacheSize	1024	number of analysis results to cache

Searching for names in a multi-language context

Phonetic analysis plugin

<https://www.elastic.co/guide/en/elasticsearch/plugins/8.7/analysis-phonetic.html>: brings too much noise

- Madonna → MTN
- mouton → MTN

Alternative : IPA or ARPAbet encoding

- https://en.wikipedia.org/wiki/International_Phonetic_Alphabet
- <https://en.wikipedia.org/wiki/ARPABET>

What is language...



Berlin Buzzwords: **bɜl'ɪn b'ʌzɹɜdɜ**

Tour Eiffel: **tʊɹ ɛfɛɪ**

Berliner Fernsehturm : **bɛɹ-'li:-nɛ 'fɛɹn-ze:-tʊɹm**

Berlin Buzzwords: B ER L IH N B AH Z W ER D Z

Tour Eiffel: T UW R EH F EH L


Berliner Fernsehturm : B ER L IH N ER F EH ER N Z EH T UH R M

ARPAbet encoding

Sophie Carboni & Lucian Precup – Speech to text with Elasticsearch

>BLN
BZZ/
WRDS

What is language...



Berlin Buzzwords: bɜl'ɪn b'ʌzɜdɜz

Tour Eiffel: tʊɜ ɛfɛl

Berliner Fernsehturm : bɛɜ-'li:-nɜ 'fɛɜn-ze:-tʊɜm

Berlin Buzzwords: B ER L IH N B AH Z W ER D Z

Tour Eiffel: T UW R EH F EH L

Berliner Fernsehturm : B ER L IH N ER F EH ER N Z EH T UH R M

16

19:11 / 36:57 • Cms >

Scroll for details



How about compound words?

Solution in two steps:

- 1/ Create a keywords dictionary for auto-completion
- 2/ Also index with ngram
- 3/ Provide a zero results fallback

Step 1 : keywords and key phrases

The screenshot shows a search interface for the term 'armoire'. The search bar at the top contains the text 'armoire' and a magnifying glass icon. Below the search bar, the results are organized into three main sections:

- Produits 1**: A list of five product items, each with a small image, a title, a model number, and a price. The items are:
 - armoire à rideaux en bois (A129088, 529,00 € HT)
 - armoire haute portes battantes (A106922, 435,00 € HT)
 - armoire haute avec réhausse (A141375, 529,00 € HT)
 - Tablettes supplémentaires pour armoire (A745825, 52,90 € HT)
 - armoire haute avec réhausse (A755571, 529,00 € HT)
- Catégories 2**: A list of five category suggestions, each with a small icon and text. The categories are:
 - Armoire bois (dans Armoire)
 - Armoire métal (dans Armoire)
 - Armoire à compartimentage (dans Armoire d'atelier)
 - Armoire à portes battantes (dans Armoire d'atelier)
 - Armoire à portes coulissantes (dans Armoire d'atelier)
 - Armoire à portes transparentes (dans Armoire d'atelier)
- Mots clés**: A list of three keyword suggestions:
 - armoire
 - Armoire et rangement
 - armoire en bois (marked with a yellow circle containing the number 3)
 - armoire à double portes
 - armoire à battantes

Solution :

- based on auto-completion #3
- separate index with keywords and keyphrases

Problem : how to generate relevant keywords and keyphrases? Especially when the input data is unstructured.

Auto-completion 3: keywords and key phrases

Extract keywords and key phrases from a non-structured text :

- NLP, Machine Learning
- Shingles
- Custom code and usage of analyzers to normalize the data

A community search engine to organize the Internet?

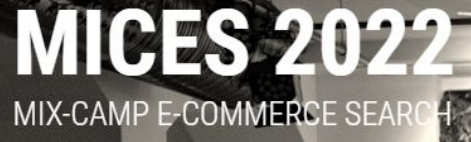
Shingle 2 \Rightarrow A community, community search, search engine, engine *to, to*
organize, organize *the, the* Internet

Shingle 3 \Rightarrow A community search, community search engine, search engine *to,*
engine *to* organize, *to* organize *the,* organize the Internet

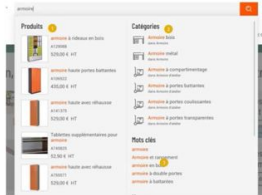
A story of auto-completions



Search



Auto-completion 3: keywords and key phrases



Solution : separate index with keywords and keyphrases

Problem : how to generate relevant keywords and keyphrases? Especially when the input data is unstructured.



A story of auto-completions - Lucian Precup & Radu Pop - MICES 2022 (US Edition)



Subscribe



142 views Jul 15, 2022

A story of auto-completions - Lucian Precup & Radu Pop - MICES 2022 (Mix-camp e-commerce search, US Edition)

https://youtu.be/Lo262_tiv9M?t=2114

Step 2: index with ngram

```
"fields" : {  
  "ngram" : {  
    "type" : "text",  
    "store" : true,  
    "term_vector" : "with_positions_offsets",  
    "analyzer" : "a2_ngram"  
  },  
  ...  
}
```

```
"analysis" : {  
  "filter" : {  
    "a2_ngram" : {  
      "type" : "ngram",  
      "min_gram" : "3",  
      "max_gram" : "4"  
    },  
    ...  
  }  
}
```

Step 3: Provide a zero results fallback

```
POST _a2/search
{
  "catalog": [
    | "hyperu"
  ],
  "text": [
    | "tomatecerise"
  ]
}
```

```
{
  "engineTimeInMillis" : 61,
  "queryRelaxing" : {
    "isRelaxed" : true,
    "parameters" : [ {
      "fuzzy" : {
        "fuzziness" : "auto",
        "threshold" : 1,
        "enabledSmartCompletion" : false
      }
    }, {
      "operator" : {
        "operator" : "OR"
      }
    }, {
      "a2_ngram" : {
        "threshold" : 1,
        "enabledSmartCompletion" : false
      }
    }
  ]
},
  "sortQuery" : {←},
  "alimentaryType" : "ALIMENTARY",
  "facetRankingStrategy" : "defaultStrategy",
  "products" : {
    "totalHits" : 38,
```

Thank you for your attention

Questions / Feedback / More ...

@lucianprecup

@a2lean

info@adelean.com

<http://www.adelean.com>

<http://www.linkedin.com/company/adelean>

<http://www.meetup.com/fr-FR/search-and-data>

