Every thursday

New webinar

Unit8™ Talks

We will teach you how to unlock
the potential of data and AI.
Introducing Unit8 webinars.

on
technologies

Unit8 Talks

Thursday, June 4

Building domain specific search engine with Elasticsearch

Webinar series

Webinar 02

Gael talks on technologies

- **Great search engines are everywhere**
  (since quite some time now ...)

Google Search I'm Feeling Lucky

amazon Try Prime

All Departments | iphone case

Shop Deals of the Day

Departments | Your Amazon.com | Today's Deals | Gift Cards | Registry | Sell | EN | Hello. Sign in Account & Lists | Orders | Try Prime | 0 Cart

1-24 of over 50,000 results for **Cell Phones & Accessories** : **Cases, Holsters & Clips** : **"iphone case"**

Sort by Suggested

FREE Shipping

All customers get FREE Shipping on orders over $25 shipped by Amazon

Show results for

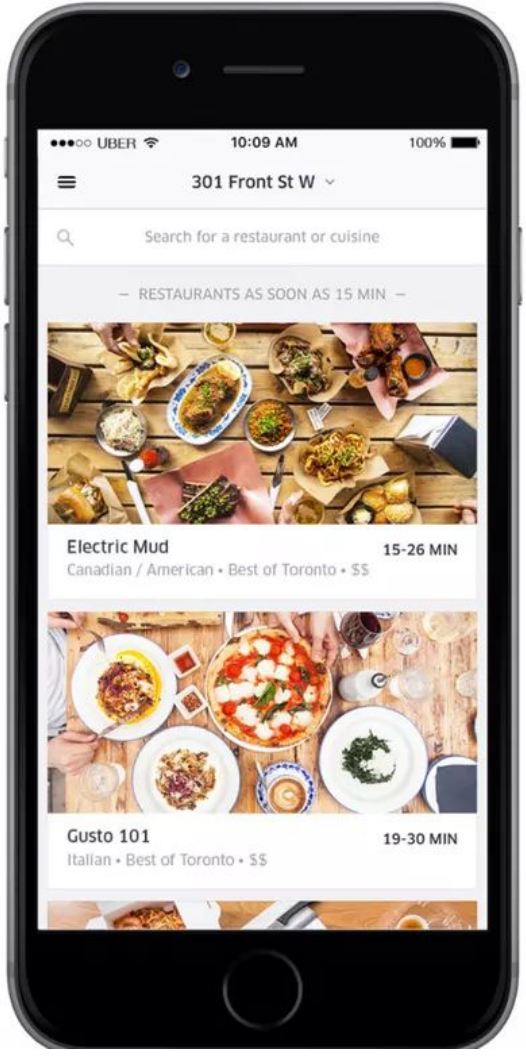
- Any Category
- Cell Phones & Accessories
- **Cases, Holsters & Clips**
  - Cases
  - Wallet Cases
  - Flip Cases
  - Waterproof Cases
  - See more

Refine by

Amazon Prime

prime

Showing results in **Cell Phones & Accessories**. Show instead results in All Departments.


●●●○○ UBER
10:09 AM
100%
301 Front St W
Search for a restaurant or cuisine
— RESTAURANTS AS SOON AS 15 MIN —
Electric Mud
15-26 MIN
Canadian / American • Best of Toronto • $$
Gusto 101
19-30 MIN
Italian • Best of Toronto • $$

Try “Amsterdam”

Become a host | Saved | Trips | Messages | Help

EXPLORE AIRBNB

All | Stays | Experiences | Adventures | Restaurants

RECENT SEARCHES

- Espoo · Logements
  Nov 15 - 18
- Espoo
  Nov 15 - 18
- Valais · Logements
  Dec 28, 2019 - Jan 02, 2020 · 10 guests
- Édimbourg · Logements
  Dec 28, 2019 - Jan 02, 2020
- Édimbourg
  Dec 28, 2019 - Jan 02, 2020

Adventures

Restaurant

sts

Edinburgh, United Kingdom
Dec 28 - Jan 02

- **Yet, many tools in organizations seem to neglect it**

Task Lists | Landscape Overview | Critical Objects | Cross-System Object Locks | Transport Analysis | **Search**

All Content | IT Service Management | Solution Documentation

Advanced Search

Search For: *dolores* Search In: Change Request Management: Transaction Search Basic Search

Additional Search Criteria

Object ID	Equal To	
Transaction Type	Equal To	
Description	Contains All of These Words	
Created on	Equal To	

Home > Change Request Management: Transaction

All Content
Change Request Management...

Sort By: Relevance

Dolores sol, 8000000263 — Change Request Management: Transaction
Transaction Type Description: Phase Cycle
Current Processor:
Configuration Item:
Priority: 4: Low
Last Changed by:
Change Cycle: 8000000263
Close Details

test uc nc, 8000000479 — Change Request Management: Transaction
Transaction Type Description: Request for Change
Current Processor:
Configuration Item:
Priority: 4: Low
Last Changed by:
Change Cycle: 8000000263
Details

Close

Details For Dolores sol, 8000000263

Due by:	
Transaction GUID:	005056B510841
Transaction Type:	SMIM
System Status:	Open
Created on:	18.04.2016
Last Changed at:	18.04.2016
Impact:	
Urgency:	
Requested Start:	18.04.2016

compliance antan

demo_legal_matter

Welcome, Yves

Repository | File server | SharePoint | Websites | Only matters

extension [4]
- docx [2]
- contact [1]
- msg [1]
- pdf [1]

Information_Type [3]
- contract [3]
- contact datasheet [1]
- correspondence [1]

Period [5]
- Past 24 hours
- Past week
- Past month
- Past year
- Older...

Antan Industries Carol Smith — 100%
08-01-2018 21:01:16
709 2018-01-08 **Antan** Industries **Antan** Industries Consectetuer Avenue 96 2 6915... Tintigny DE 08 09 95 45 42 Carol Smith F EN **Compliance** Officer 08 09 95 45 42 carol...**Antan** Industries Carol Smith

Deloitte Academy Webinar - Pan-European VAT Update - Q2 _ Q3 2015 - 14_09_2015_1... — 17.1%
09-01-2018 15:19:35
to ensure **compliance** with national legislations. Agenda * VAT changes

Antan Industries - Trade Agreement - 20170906 - Version 3.docx — 15.6%
09-01-2018 15:50:23
Director, (hereafter "Knowliah") and, **Antan** Industries with registered... acting in his or her capacity as **Compliance** Officer, (hereafter "Client"...-User will refer to that legal entity. The Client may use the Services only in **compliance**... Cloud System in relation to the Client Materials; **compliance** of the Client... in accordance with the instructions of the Client; if it cannot provide such **compliance**...**Antan** Industries - Trade Agreement - 20170906 - Version 3.docx

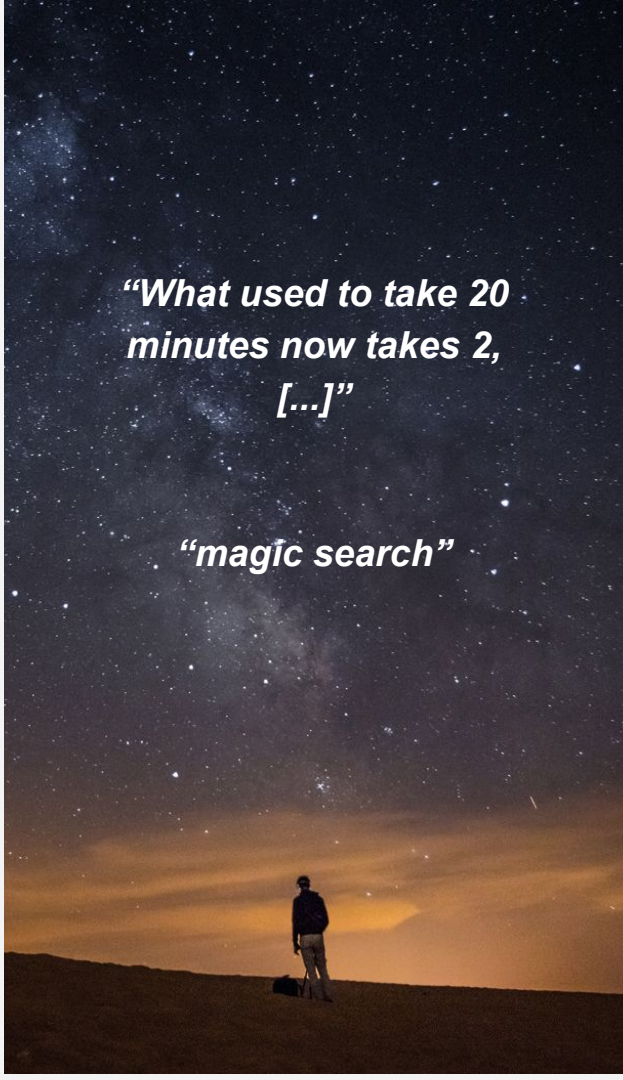
Antan Industries - Trade Agreement - 20170906 - Version 3.docx — 15.6%
09-01-2018 22:40:29
Director, (hereafter "Knowliah") and, **Antan** Industries with registered... acting in his or her capacity as **Compliance** Officer, (hereafter "Client"...-User will refer to that legal entity. The Client may use the Services only in **compliance**... Cloud System in relation to the Client Materials; **compliance** of the Client... in accordance with the instructions of the Client; if it cannot provide such **compliance**...**Antan** Industries - Trade Agreement - 20170906 - Version 3.docx

Antan Industries - Trade Agreement - 20170906 - Signed.pdf — 13.4%
09-01-2018 15:50:27
, **Antan** Industries with registered offices at Consectetuer Avenue, 6915 Tintigny... 75709925399 represented by Carol Smith acting in his or her capacity as **Compliance**... use the Services only in **compliance** with these Terms & Conditions. The Client may... Materials; 6.1.2. **compliance** of the Client Materials with applicable laws... with the instructions of the Client; if it cannot provide such **compliance**...**Antan** Industries - Trade Agreement - 20170906 - Signed.pdf

Page **1** of **1** (5 results)

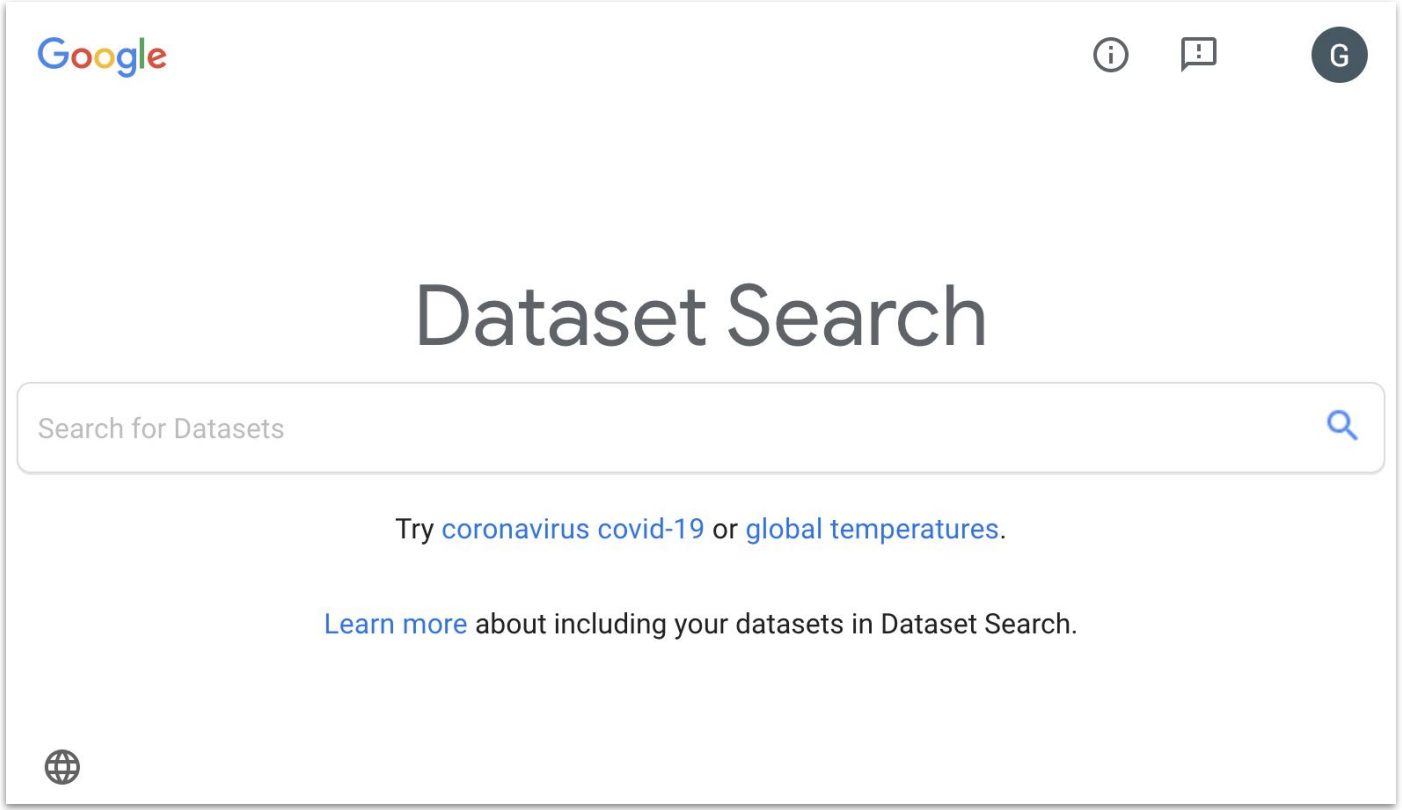
Why companies should care?

- In a fast-evolving world, it is crucial to be able to quickly access the relevant information
- Many processes and roles productivity heavily rely on that:
  - Resolve a customer problem
  - Find the correct product for your client
  - Know who to go to
- It has become easy to deploy and customize existing open-source search engines



Unit8

■ How are domain-specific (or “topical”) search engine different?


Google
G
Dataset Search
Search for Datasets
Try coronavirus covid-19 or global temperatures.
Learn more about including your datasets in Dataset Search.

Google

global temperatures

Last updated | Download format | Usage rights | Topic | Free

100+ datasets found

U

Global surface temperatures: BEST: Berkeley Earth Surface...

climatedataguide.ucar.edu
archive.is

D

Data from: Climate Prediction Center (CPC) Global Temperatu...

catalog.data.gov
datadiscoverystudio.org
+1more

Updated Sep 26, 2015

Global surface temperatures: BEST: Berkeley Earth Surface Temperatures

Explore at climatedataguide.ucar.edu | Explore at archive.is

5 scholarly articles cite this dataset (View in Google Scholar)

Time period covered

Jan 1701 - Jul 2019

Description

The Berkeley Earth Surface Temperatures (BEST) are a set of data products, originally a gridded reconstruction of land surface air temperature records spanning 1701-present, and now including an 1850-present merged land-ocean data set that combines the land analysis with an interpolated version of HadSST3. The land station data are available in an archive, and an experimental version provides daily data for 1880-present. Average

Context is key in a domain-specific search engine

You should know what your user are looking for and what information they care about. With that, there are 3 key aspects you can tune:

1. Curation of content
2. Adding key contextual data
3. **Tune queries and search results**
   a. Data needs to be properly indexed
   b. Relevance of search results need to be tuned

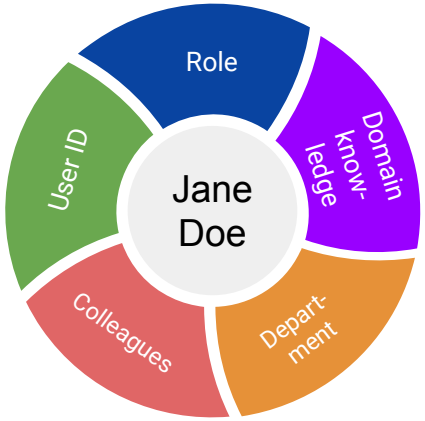


Fig 1. Enrich your content with additional information if possible

"construction worker in orange safety vest is working on road."

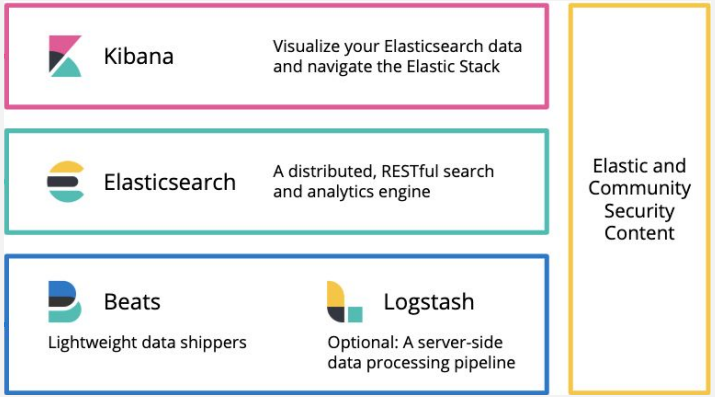
Unit8

- **Build your own search engine**
  (using Elasticsearch)

Unit8

10sec. intro to Elasticsearch

- Document based
- Search and analytics engine
- Distributed and scalable
- Easy to ingest many sources



Unit8

Elasticsearch setup & Basics of indexing and querying

Notebook available to guide you through these steps:
https://github.com/unit8co/dmz-workshop-er-search

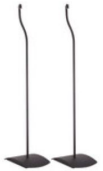
	_index	_type	_id	_score	_source.product_name	_source.description	_source.brand	_source.categories
0	test_index	_doc	4769785	19.982872	Alpine - 12" Dual-Voice-Coil 4-Ohm Subwoofer - Black	Listen to music in your vehicle with rich, clear sound with this Alpine SWS-12D4 12" dual-voice-coil 4-ohm subwoofer that features improved thermal capacity and mechanical strength for reliable performance.	Alpine	[Car Electronics & GPS, Car Audio, Car Subwoofers]
1	test_index	_doc	9597106	19.982872	Alpine - 8" Powered Subwoofer System - Black	Amplify the bass in your vehicle using this Alpine PWE-S8 subwoofer system that features a compact design you can fit beneath or behind a truck seat, in a sedan trunk or in the hatchback of an SUV. The cast-metal construction ensures long-lasting use.	Alpine	[Car Electronics & GPS, Car Audio, Car Subwoofers]

```
In [47]: HTML(search_engine("speaker stands"))
```

Out[47]:

Bell'O - Speaker Stands (Pair) - Black/Brown

Bell'O — **Keep your speakers at the optimal height for ideal sound performance with these Bell'O SP211 speaker stands. A combination of black powdercoat steel and cherry-finish wood inlaid fronts offers a stylish accent to complement your home d&233;cor.**



Bose® - UFS-20 Series II Universal Floor Stands (Pair) - Black

Bose® — **These Bose® UFS-20 Series II universal floor stands are designed to optimize listening. Slender metal stands elegantly display your speakers while hiding the speaker wires.**

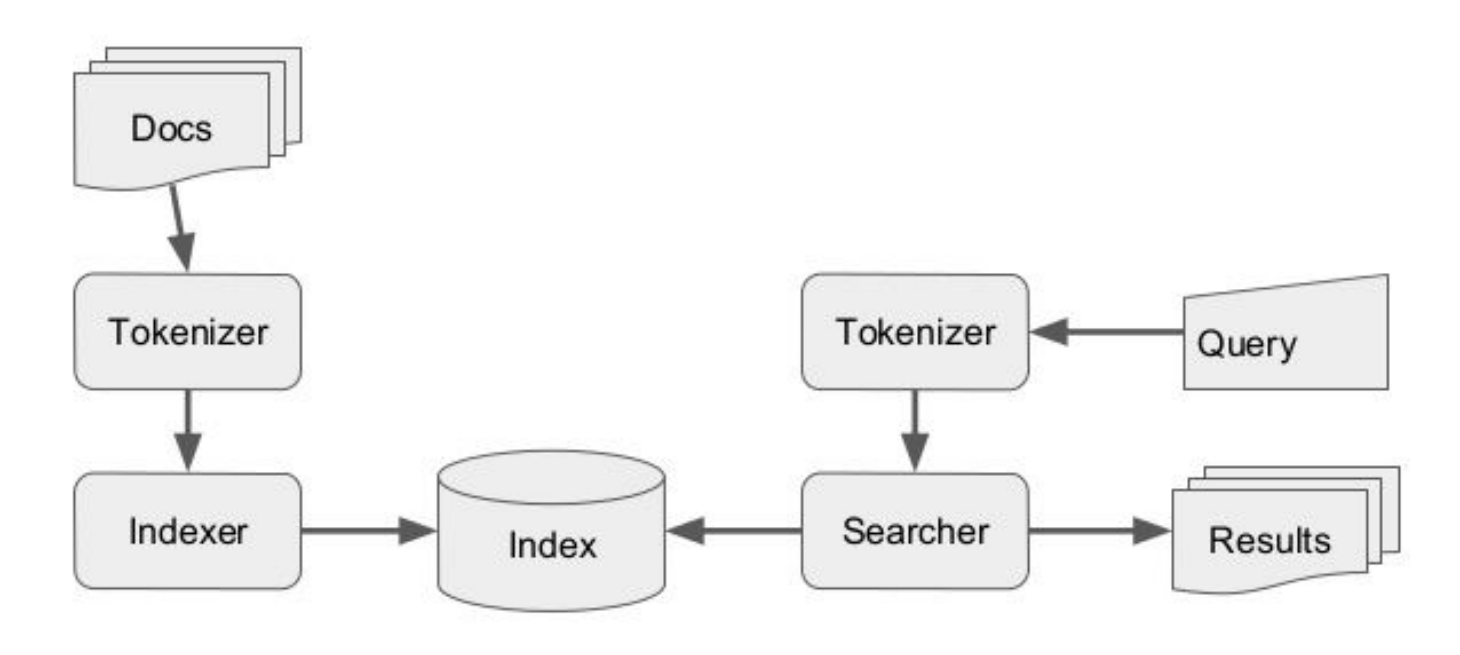
■ **Understand your search engine**
(& how to get the most out of it)

Unit8™

Major parts of a search engine

1. Split sentences into words (**Tokenizer**)
2. Index words in documents (**Indexing**)
3. Rank documents (**Ranking**)

Major parts of a search engine



- **Tokenizers & Indexing document**
  (or should we say analyzer)

A look at how to define analyzers in Elasticsearch

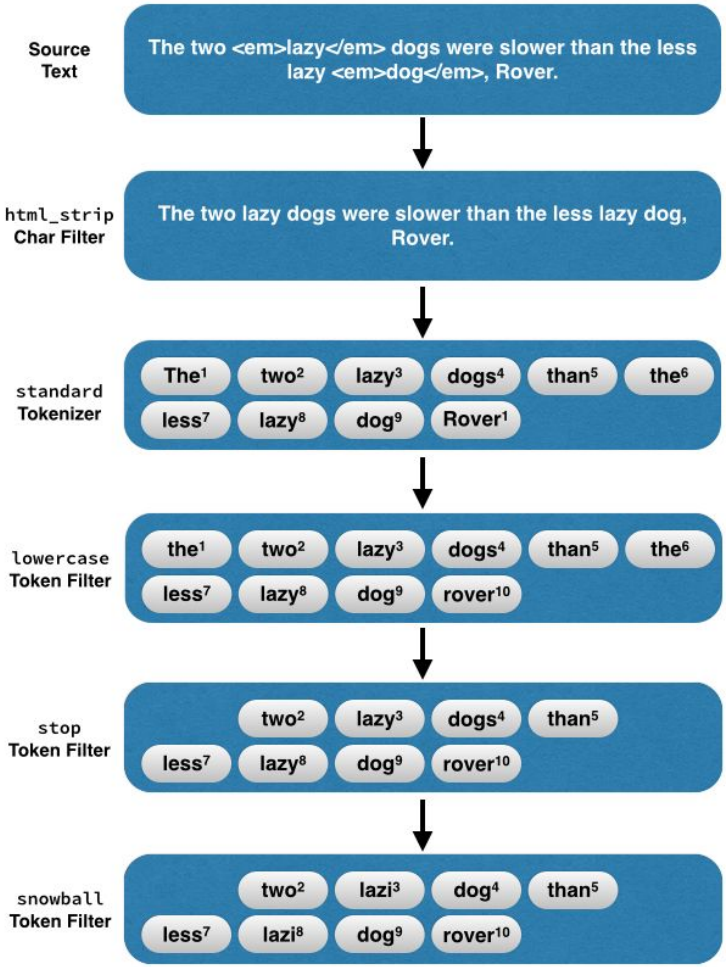
They are called analyzers because they do a bit more than a tokenizer. They are 3 steps you can configure:

1. Char filter (preprocessing)
2. Tokenizer (splitting into words)
3. Filter (post-processing)

```
PUT /YOUR_INDEX/_mapping
{
  "settings" : {
    "analysis": {
      "analyzer": {
        "my_analyzer":  {
          "type": "custom",
          "char_filter": [
            "html_strip"
          ],
          "tokenizer": "standard",
          "filter": [
            "lowercase"
          ]
        }
      }
    }
  },
  "mappings": {
    "properties": {
      "product_name": {"type": "text"},
      "description":  {"type": "text",
                       "analyzer":"my_analyzer"},
      "categories":   { "type": "text" },
      "type":         { "type": "text" },
      "brand":        { "type": "text" },
      "popularity":   { "type": "double"}
      }
  }
}
```

Tip #1: Make sure to configure your analyzers for key fields

- It is worth tailoring the analyzer for your most important fields:
  - Language (e.g. “Balletttänzerin”)
  - Field content (e.g. parsing IDs)
  - Case sensitive?
  - ....



Tip #1: A few real-life examples

1. “-” (dash) vs “–” (em-dash)
   - **Possible solution**: normalize characters in analyzers

2. Language specific issues
   - “Balletttänzerin”
   - **Possible solution**: Pick the proper language analyzer
     - For German, have a look at “Word decomposers”

```
PUT /YOUR_INDEX/_mapping
{
  "settings" : {
    "analysis": {
      "analyzer": {
        "my_analyzer": {
          ...
          "char_filter": [
                {
                    "type": "pattern_replace",
                    "pattern": "&mdash;",
                    "replacement": "-"
                }
          ],
          ...
        }
      }
    }
  },
  "mappings": {
    "properties": {
      ...
      "id":  {"type": "keyword",
              "analyzer":"my_analyzer"},
      ...
      }
  }
}
```

Tip #2: Use synonyms

- Great tool to reconcile different notations across documents:
  - Leverage existing business knowledge and abbreviations
  - Use existing ontologies
    - Departments
    - Region
    - Acronyms

```
PUT /YOUR_INDEX/_mapping
{
  "settings" : {
    "analysis": {
      "analyzer": {
        "my_analyzer": {
          ...
          "filter": {
             "region_synonym": {
               "type": "synonym",
               "synonyms": [
                 "europe, eu, eur",
                 "middle-east, me, moyen-orient",
                 "india middle east africa, ima",
                 "latin america, lam, latam",
                 "north america, nam",
                 "south east asia, sea",
                 "north east asia, nea",
                 "united states of america, usa",
                 "united kingdom, uk",
                 "all regions, world, monde"
               ]
             }
           }
        }
      }
    },
    "mappings": {
      "properties": {
        ...
        "region":  {"type": "text",
                    "analyzer":"my_analyzer"},
        ...
        }
    }
}
```

See that article for a more complete overview: https://www.elastic.co/blog/found-text-analysis-part-1

Customizing your Ranking

Tips #1: Understanding ranking is key

$$\text{score}(q,d) = \text{coord}(q,d) \cdot \text{queryNorm}(q) \cdot \sum_{t \text{ in } q} \left(\text{tf}(t \text{ in } d) \cdot \text{idf}(t)^2 \cdot t.\text{getBoost}() \cdot \text{norm}(t,d) \right)$$

tf = Measure of how often a term appears in the document
idf = Measure of how often the term appears in all documents
norm = Normalization factor based on number of words in field
boost = Boost of the field

coord = Number of terms (in the query) found in the document
queryNorm = Normalization query score (so it can be compared across queries)

Unit8

What this means in practice

- Documents containing all the search terms are good
- Matches on rare words are better than for common words
- Long documents are not as good as short ones
- Documents which mention the search terms many times are good

Tips #1: Boost most relevant fields

- Easy but important step, not all fields are born equal
  - The title or product name may contain more relevant terms than the description

```
PUT /YOUR_INDEX/_mapping
{
  "settings" : {
    "analysis": {
      ...
    }
  },
  "mappings": {
    "properties": {
      "product_name": {"type": "text", “boost”: 3},
      "description":  {"type": "text"},
      "categories":   { "type": "text", “boost”: 2},
      "type":         { "type": "text" },
      "brand":        { "type": "text" },
      "popularity":   { "type": "double"}
      }
  }
}
```

Unit8

Source: Swiftype.com

Tips #2: Match the full sentence

- Without quotes exact matches do not have a special importance
  - Use a query to match the words (***match***) and the the full sentence (***match_phrase***) at the same time

event sourcing using kafka 🔎

1. Kafka — how to configure
2. New release of Apache Kafka
3. Event Sourcing using Kafka
4. Event Store — lessons learnt

```
PUT /YOUR_INDEX/_search
{
  "query": {
    "bool": {
      "should": [
        {"match_phrase": {"query": "cool cars"}},
        {"match": {"query": "cool cars"}}
        }
      ]
    }
  }
}
```

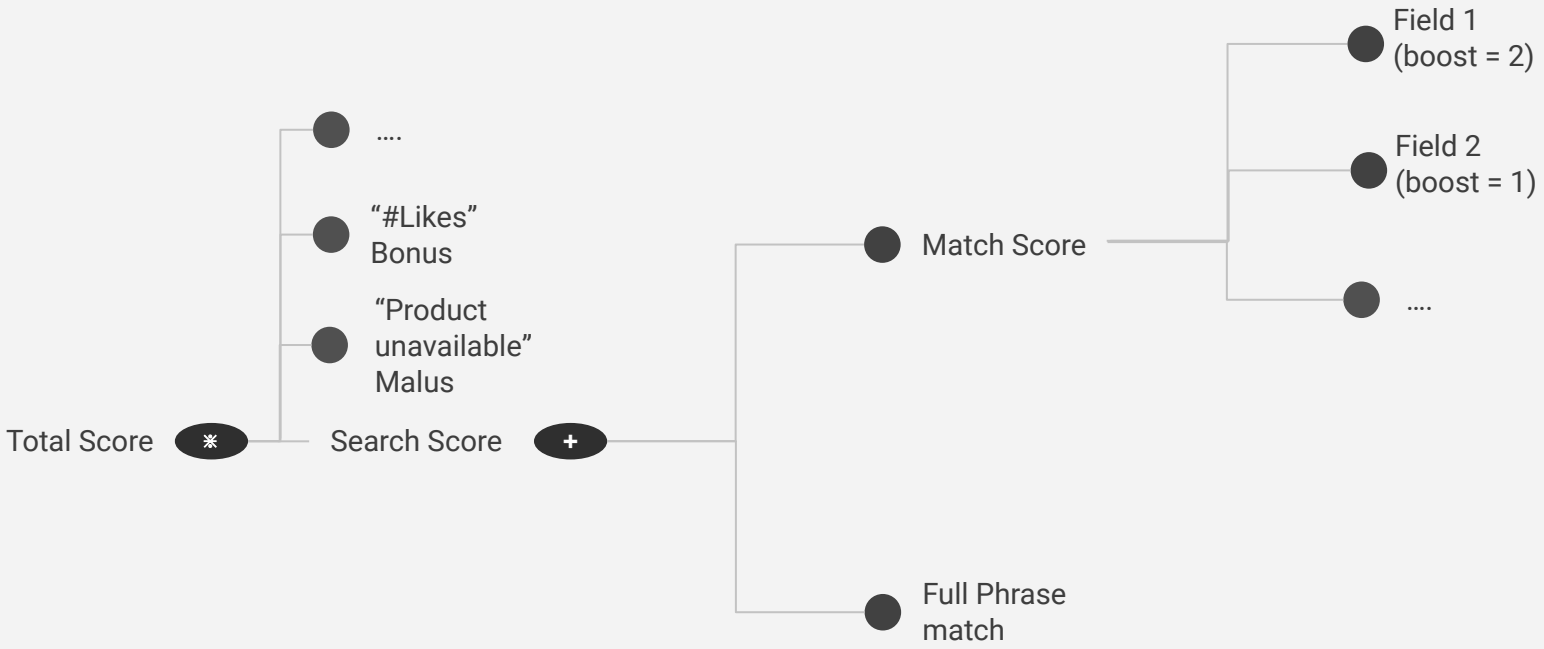
Unit8

Tips#3: Function score

- Use the function score which allows to perform extra computation on document score:
  - Boost more recent products
  - Favor document from a given source
  - Promote “popular” items based on your fancy machine learning model
  - ...

```
PUT /YOUR_INDEX/_search
{
    "query": {
        "function_score": {
            "boost_mode":"multiply",
            "query": {
                "match": {
                    "message": "elasticsearch"}
              },
              "script_score": {
                  "script": {
                    "Source":
                      "Math.log(2+ doc['likes'].value)"
                  }
              }
        }
    }
}
```

Unit8


Field 1
(boost = 2)
Field 2
(boost = 1)
....
Match Score
....
"#Likes"
Bonus
"Product
unavailable"
Malus
Total Score
*
Search Score
+
Full Phrase
match

Unit8

- **Lots of parameters to tune.**
  How do we evaluate our changes?

Ranking Evaluation API

- Allow to give score to specific search results
  - Great way to agree with business on important “top results”
- Great way to validate changes to analyzers/ranking
- Can integrate user feedback directly there

```
POST /enwiki_rank/_rank_eval
...
  "query": {
        "query_string":  { "query": "JFK" }
...
  "ratings": [
        {"_id": "3054546", "rating": 1 },
        {"_id": "5119376", "rating": 3},
        ...
    ]
...
   "metric": { ... }
```

https://www.elastic.co/blog/made-to-measure-how-to-use-the-ranking-evaluation-api-in-elasticsearch

■ What’s next ?

(In the world of search engines)

Entity Based Search

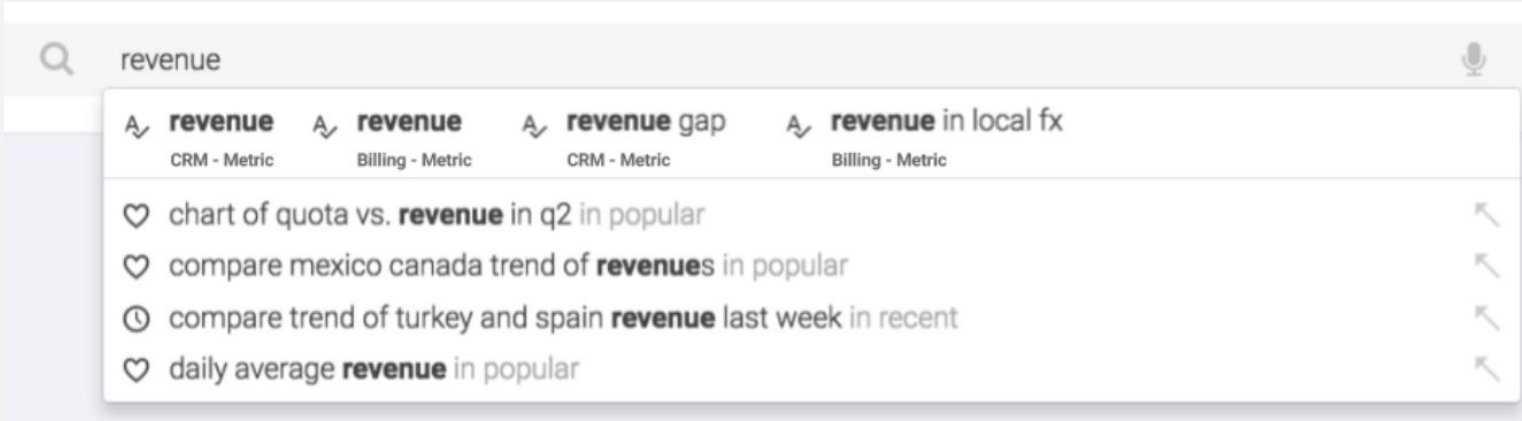
- Improve search by disambiguating some user keywords
- Highlight words belonging to an ontology
- Use entities to improve links with other documents

```
POST my_index/_analyze

{

        "field": "my_rich_text_field",

        "text": "Today [Johannes](Johannes Mueller)
                 announced a new model of
                 electrical car"

}
```

https://www.elastic.co/blog/search-for-things-not-strings-with-the-annotated-text-plugin

Unit8

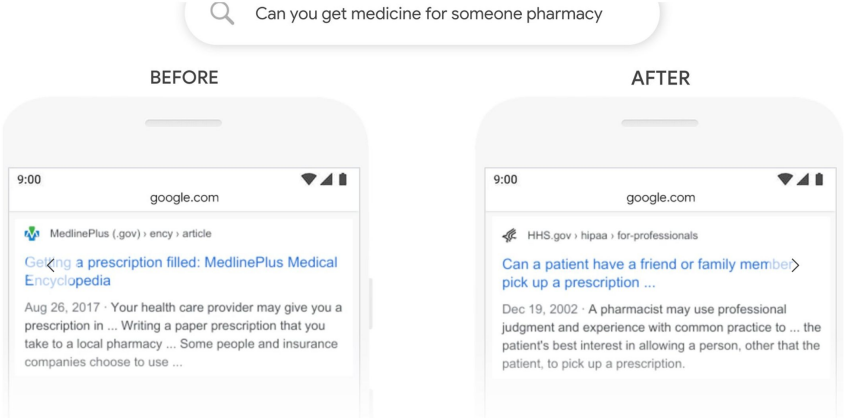
Entity Based Search



Unit8

Language models

- A step towards a better understanding of user queries
- Integrate neural networks embedding (BERT) in ranking to return better results (used in 1 in 10 queries)
- Transfer learning to other languages than English

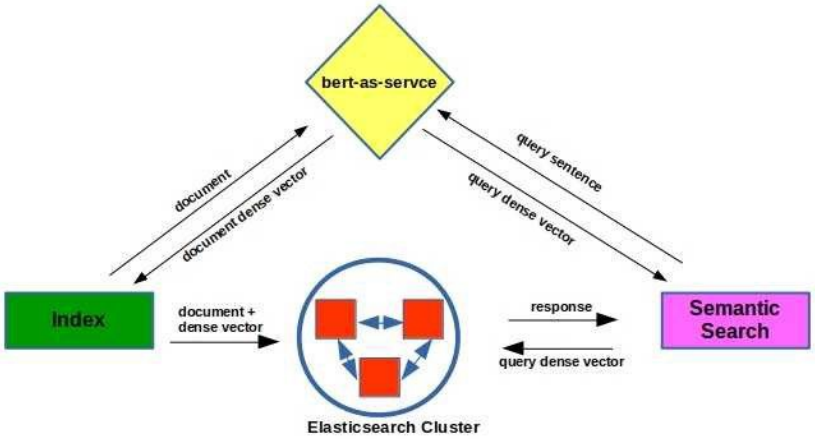


https://blog.google/products/search/search-language-understanding-bert/

Unit8

Use sentence embedding with Elasticsearch

- Compute embedding of titles or documents at indexing time
- When a user type a query, compute its embedding and compare it with your documents
- Can also be used to search for images, sound or other documents

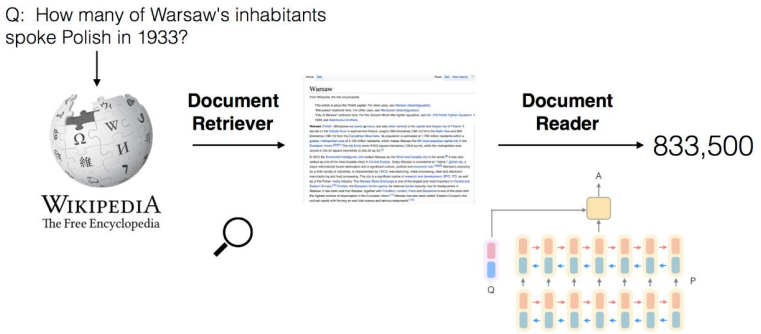


https://www.elastic.co/blog/text-similarity-search-with-vectors-in-elasticsearch

Source: https://towardsdatascience.com/semantics-at-scale-bert-elasticsearch-be5bce877859

And many more...

- Automatically extract answers from documents
  - https://github.com/facebookresearch/DrQA
- Automatic summarization of search results
- User input sense disambiguation
  - https://openai.com/blog/discovering-types-for-entity-disambiguation/
- ...



The prey saw the **jaguar** cross the jungle.

Jaguar Cars 0.60

jaguar 0.29

Conclusions

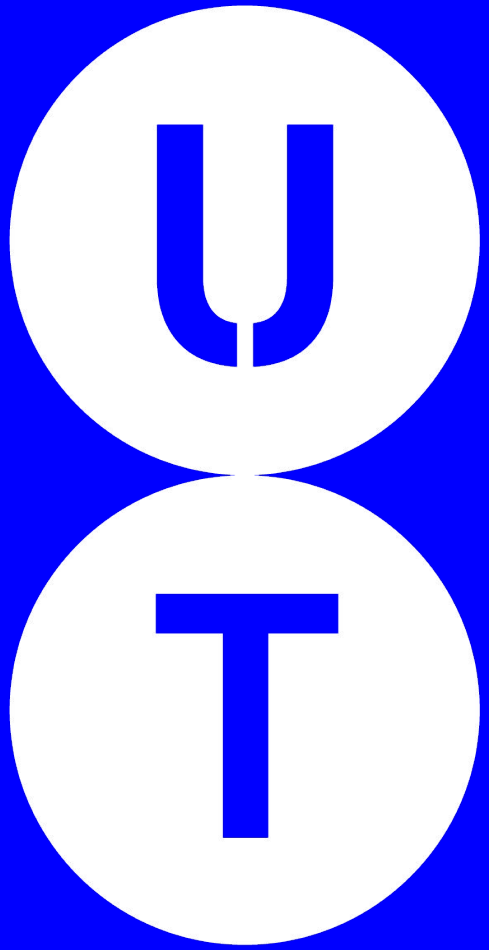
- Setting up (e.g. managed instance on the cloud) and configuring a basic search engine is relatively easy
- Context is key so, including your expertise and knowledge of your users will help providing more relevant results
- A good search engine can have some real practical impact
  - “***What used to take 20 minutes now takes 2, [...]***”
- With the recent progress in NLP, probably lots of improvements coming in the next years!

Unit8 Talks

Q&A

New webinar

Every thursday



Unit8™

We will teach you how to unlock
the potential of data and AI.
Introducing Unit8 webinars.

thank you

Talks

on technologies
on business

Unit8 Talks

Thursday, June 11

Bianca and Radu talk on technologies

Real data stories: Entity Resolution

Webinar series

Webinar 02

Next webinar

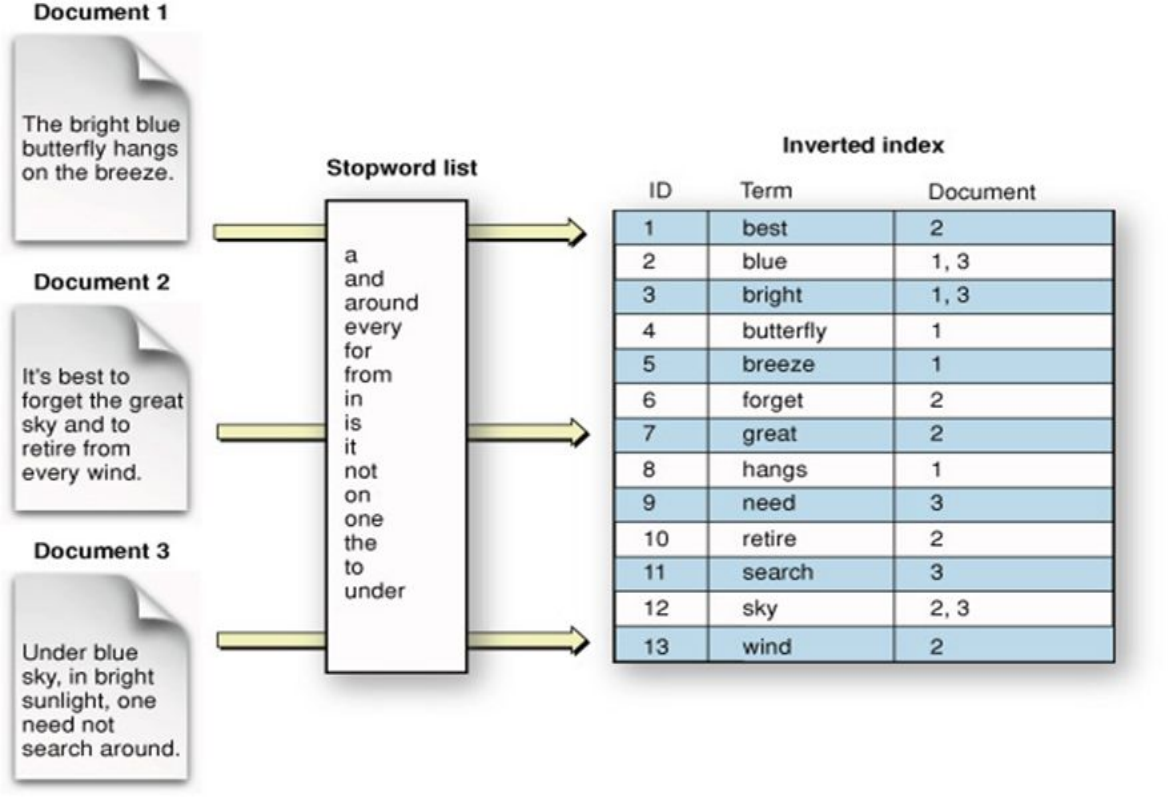
June 11, Thursday 1PM

Useful Links

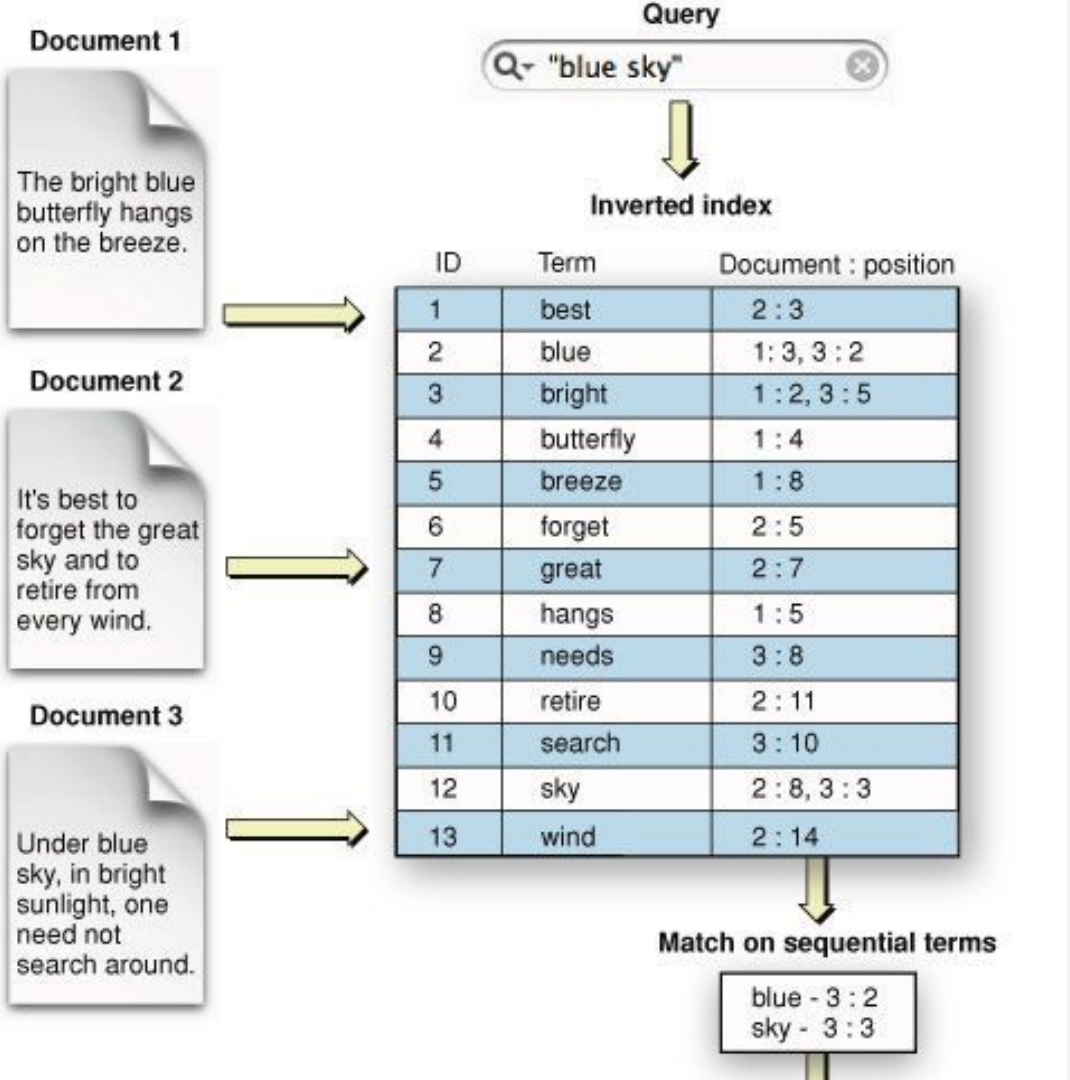
- Build your search engine (Jupyter Notebook)
  https://github.com/unit8co/dmz-workshop-er-search
- Elasticsearch documentation:
  https://www.elastic.co/guide/en/elasticsearch/reference/current/index.html
- Elasticsearch intro to analyzers:
  https://www.elastic.co/blog/found-text-analysis-part-1
- Synonyms in Elasticsearch:
  https://www.elastic.co/blog/boosting-the-power-of-elasticsearch-with-synonyms
- Function score in Elasticsearch:
  https://www.elastic.co/blog/found-function-scoring
- Type-ahead:
  https://www.linkedin.com/pulse/new-elasticsearch-datatype-searchasyoutype-jess%C3%A9-peixoto/

Extra Material

Indexing


Document 1
The bright blue butterfly hangs on the breeze.
Document 2
It's best to forget the great sky and to retire from every wind.
Document 3
Under blue sky, in bright sunlight, one need not search around.
Stopword list
a
and
around
every
for
from
in
is
it
not
on
one
the
to
under
Inverted index
ID
Term
Document
1 best 2
2 blue 1, 3
3 bright 1, 3
4 butterfly 1
5 breeze 1
6 forget 2
7 great 2
8 hangs 1
9 need 3
10 retire 2
11 search 3
12 sky 2, 3
13 wind 2

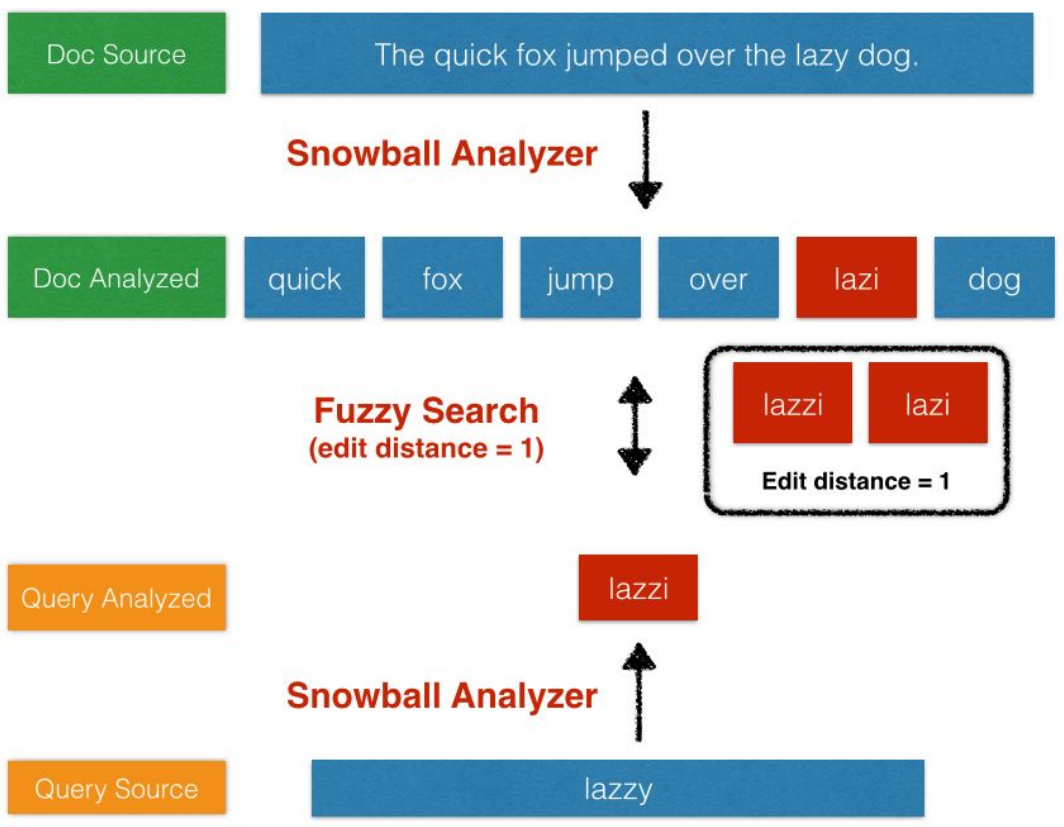
What about expressions and phrases?


Query
"blue sky"
Inverted index
Document 1
The bright blue butterfly hangs on the breeze.
Document 2
It's best to forget the great sky and to retire from every wind.
Document 3
Under blue sky, in bright sunlight, one need not search around.
ID
Term
Document : position
1 best 2 : 3
2 blue 1: 3, 3 : 2
3 bright 1 : 2, 3 : 5
4 butterfly 1 : 4
5 breeze 1 : 8
6 forget 2 : 5
7 great 2 : 7
8 hangs 1 : 5
9 needs 3 : 8
10 retire 2 : 11
11 search 3 : 10
12 sky 2 : 8, 3 : 3
13 wind 2 : 14
Match on sequential terms
blue - 3 : 2
sky - 3 : 3

Unit8

Ranking

Fuzziness



https://www.elastic.co/blog/found-fuzzy-search

Unit8

■ Manage access to data

X-Pack Security features

To operate your elasticsearch cluster, X-Pack has a lot of easy-to-use security features for free (since May 2019):

- TLS encryption
- Role Based access control
- Key management

1. Curation

Understand the use-case and select the relevant document to make search results more pertinent:

- Index only relevant fields
- Select best sources of information
- Up-to-date documents
- Words can have multiple meaning so keep only relevant sources reduces noise

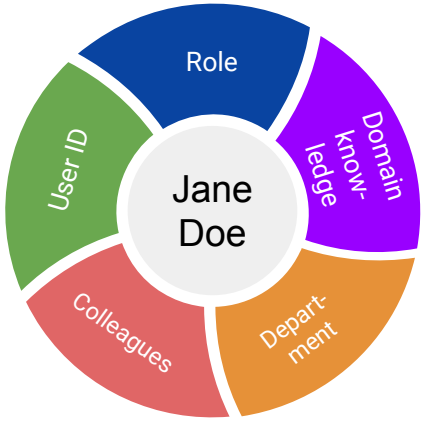
Source: Things Organized Neatly: The Art of Arranging the Everyday

Unit8

2. Enrich with contextual data

It is oftentimes possible to add information that will enrich existing documents and make them easier to search:

- Extra information about document
  - Extra IDs
  - Department / Domain
  - Regulatory
- Generate your own labels
  - “Best-seller” item
  - ML to generate labels for images in documents
  - ...



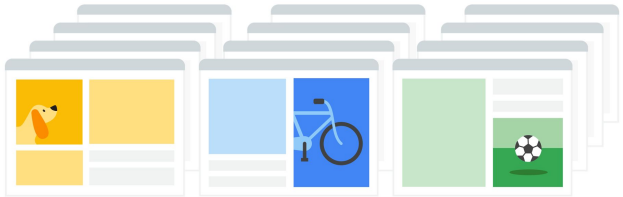
"construction worker in orange safety vest is working on road."

Unit8

3. Tune queries and search results

Last but critical aspect is to configure your search engine properly:

- Data needs to be properly indexed
- Relevance of search results need to be tuned



Unit8

Source: Digitalgenx