Unit8™

Building a domain specific search engine

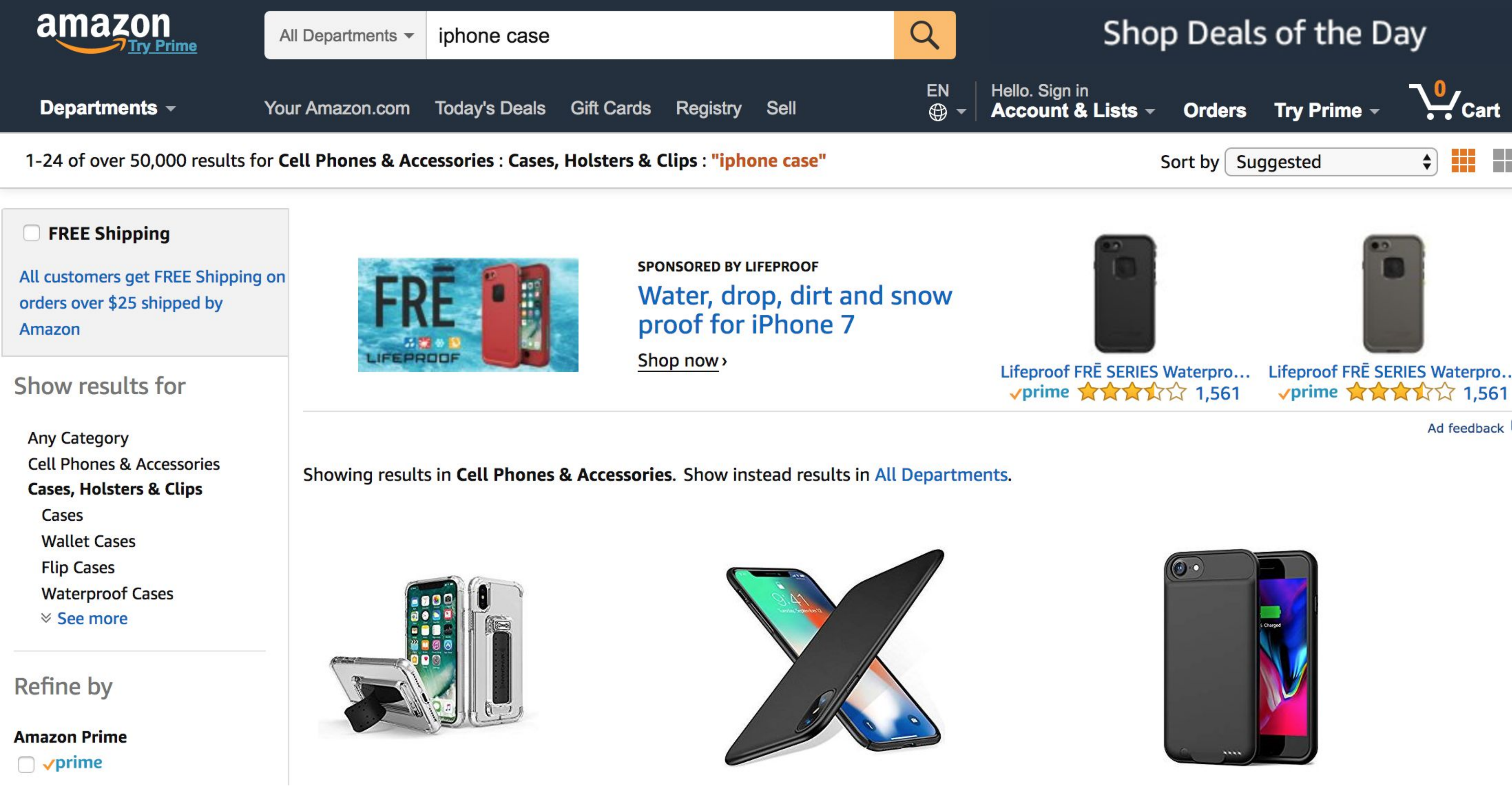
digital natives

unit8.co

⬛ Great search engines are everywhere

(since quite some time now …)

Google
Google Search
I'm Feeling Lucky



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

☐ **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

SPONSORED BY LIFEPROOF

Water, drop, dirt and snow proof for iPhone 7

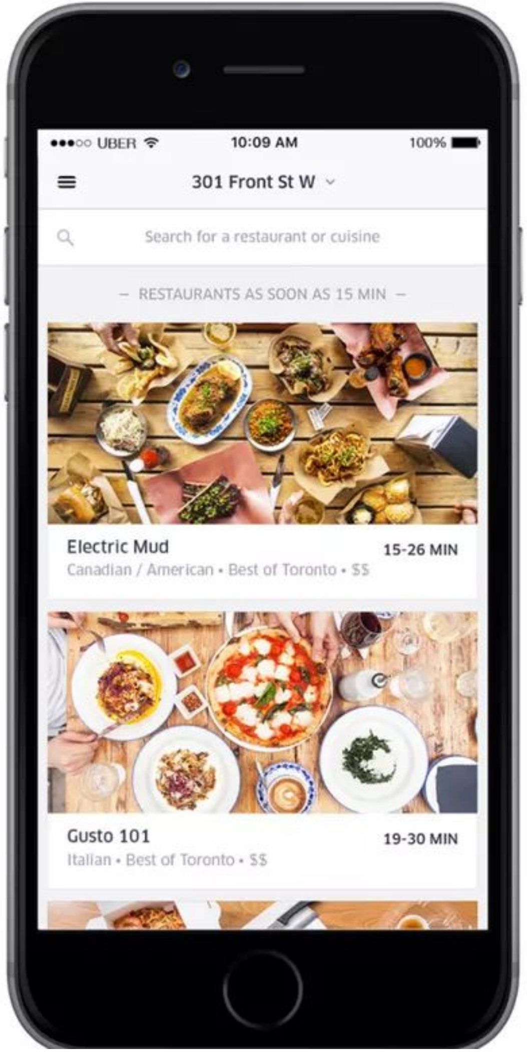
Shop now ›

Lifeproof FRĒ SERIES Waterpro...
✓prime ★★★½☆ 1,561

Lifeproof FRĒ SERIES Waterpro...
✓prime ★★★½☆ 1,561

Ad feedback

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



●●●○○ UBER 10:09 AM 100%

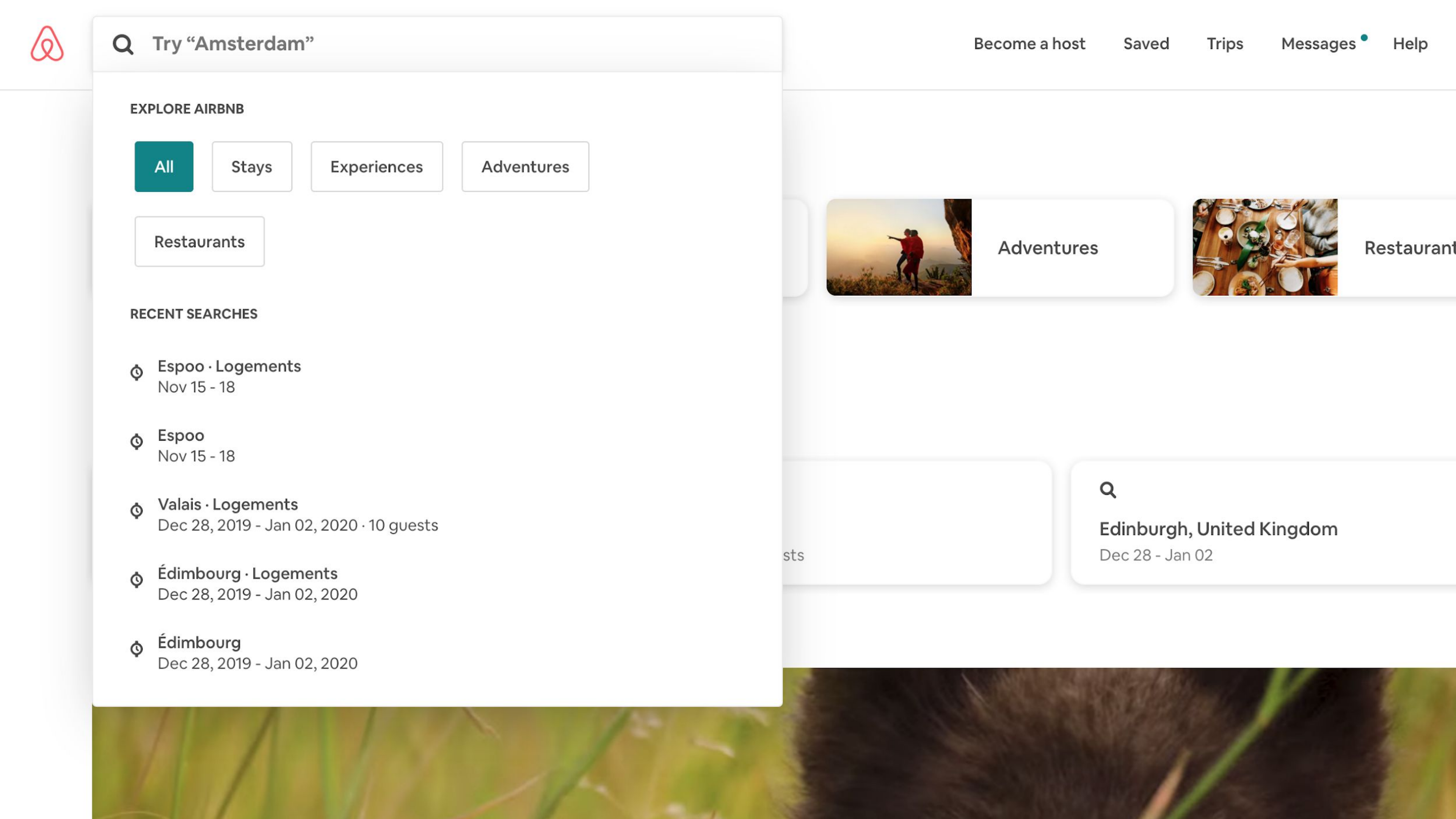
301 Front St W

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 • $$



EXPLORE AIRBNB

RECENT SEARCHES

Espoo · Logements
Nov 15 - 18

Espoo
Nov 15 - 18

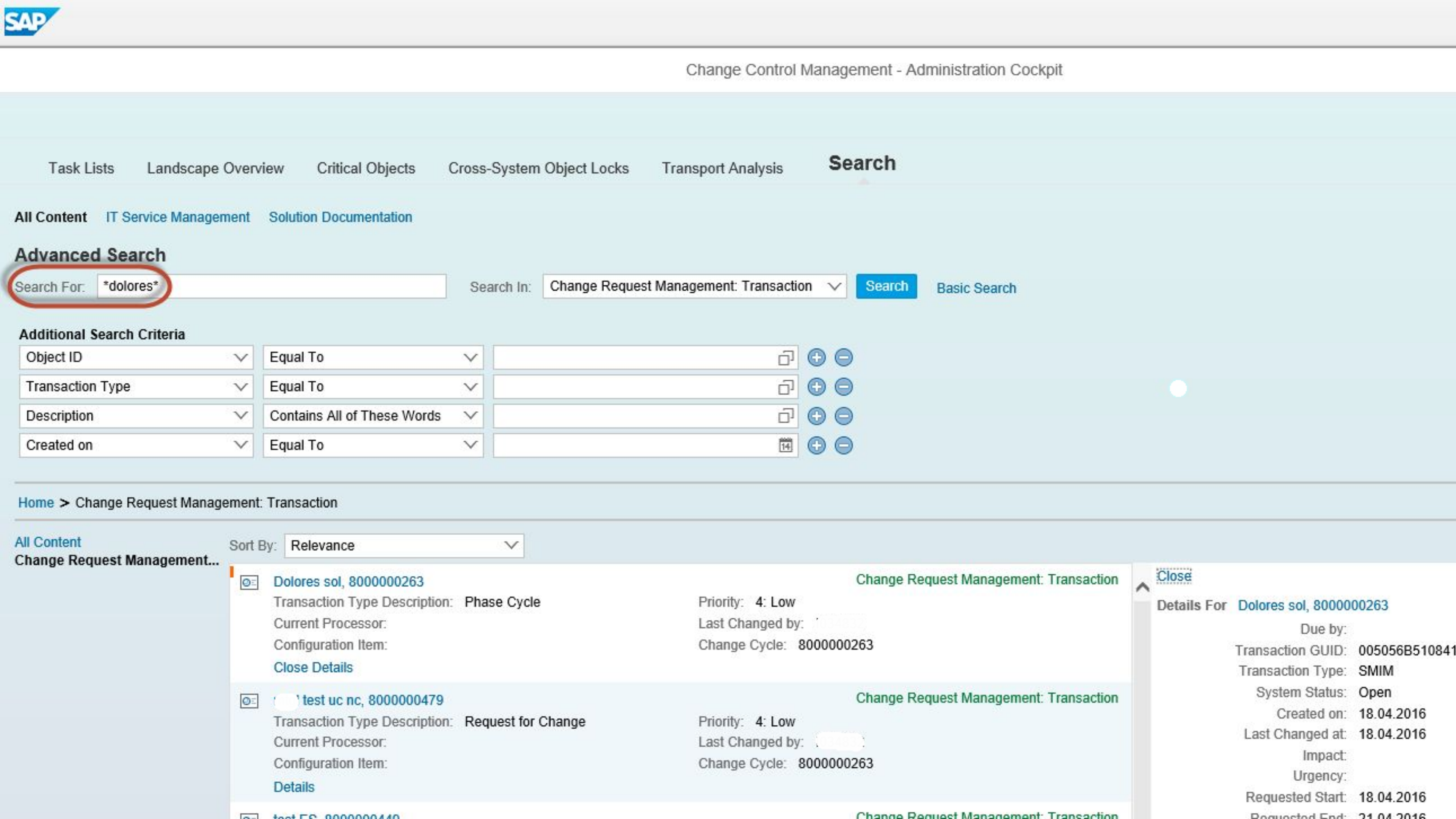
Valais · Logements
Dec 28, 2019 - Jan 02, 2020 · 10 guests

Édimbourg · Logements
Dec 28, 2019 - Jan 02, 2020

Édimbourg
Dec 28, 2019 - Jan 02, 2020

Edinburgh, United Kingdom
Dec 28 - Jan 02

⬛ Yet, many entreprise tools seem to neglect it



Change Control Management - Administration Cockpit

Advanced Search

Search For: *dolores*

Search In: Change Request Management: Transaction

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

Sort By: Relevance

Dolores sol, 8000000263 — Change Request Management: Transaction

Transaction Type Description: Phase Cycle
Priority: 4: Low
Current Processor:
Last Changed by: [redacted]
Configuration Item:
Change Cycle: 8000000263

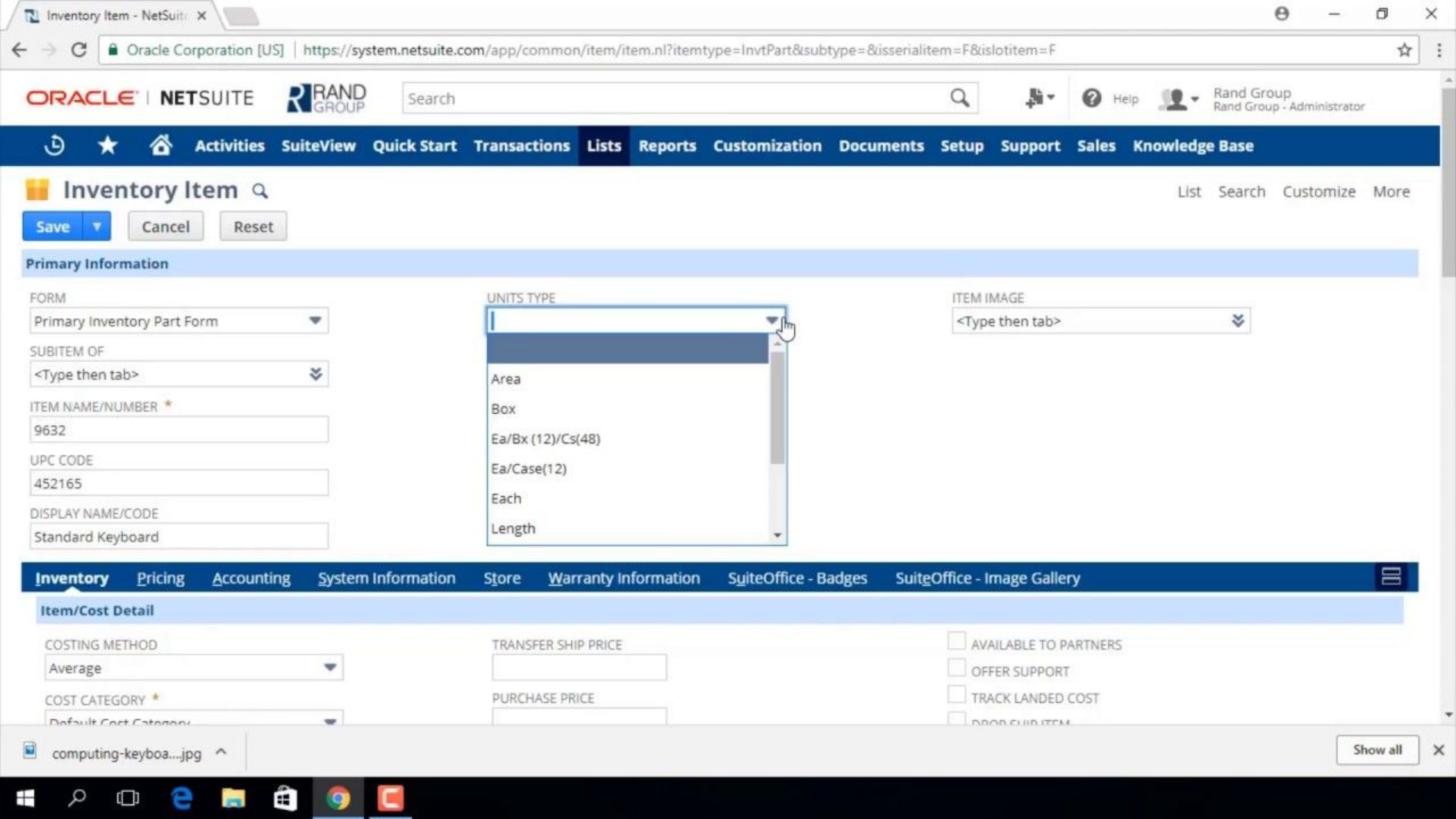
[redacted] test uc nc, 8000000479 — Change Request Management: Transaction

Transaction Type Description: Request for Change
Priority: 4: Low
Current Processor:
Last Changed by: [redacted]
Configuration Item:
Change Cycle: 8000000263

test FS, 8000000440 — Change Request Management: Transaction

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
Requested End: 21.04.2016



Inventory Item

Primary Information

FORM: Primary Inventory Part Form

SUBITEM OF: <Type then tab>

ITEM NAME/NUMBER *: 9632

UPC CODE: 452165

DISPLAY NAME/CODE: Standard Keyboard

UNITS TYPE:

- Area
- Box
- Ea/Bx (12)/Cs(48)
- Ea/Case(12)
- Each
- Length

ITEM IMAGE: <Type then tab>

Item/Cost Detail

COSTING METHOD: Average

COST CATEGORY *: Default Cost Category

TRANSFER SHIP PRICE

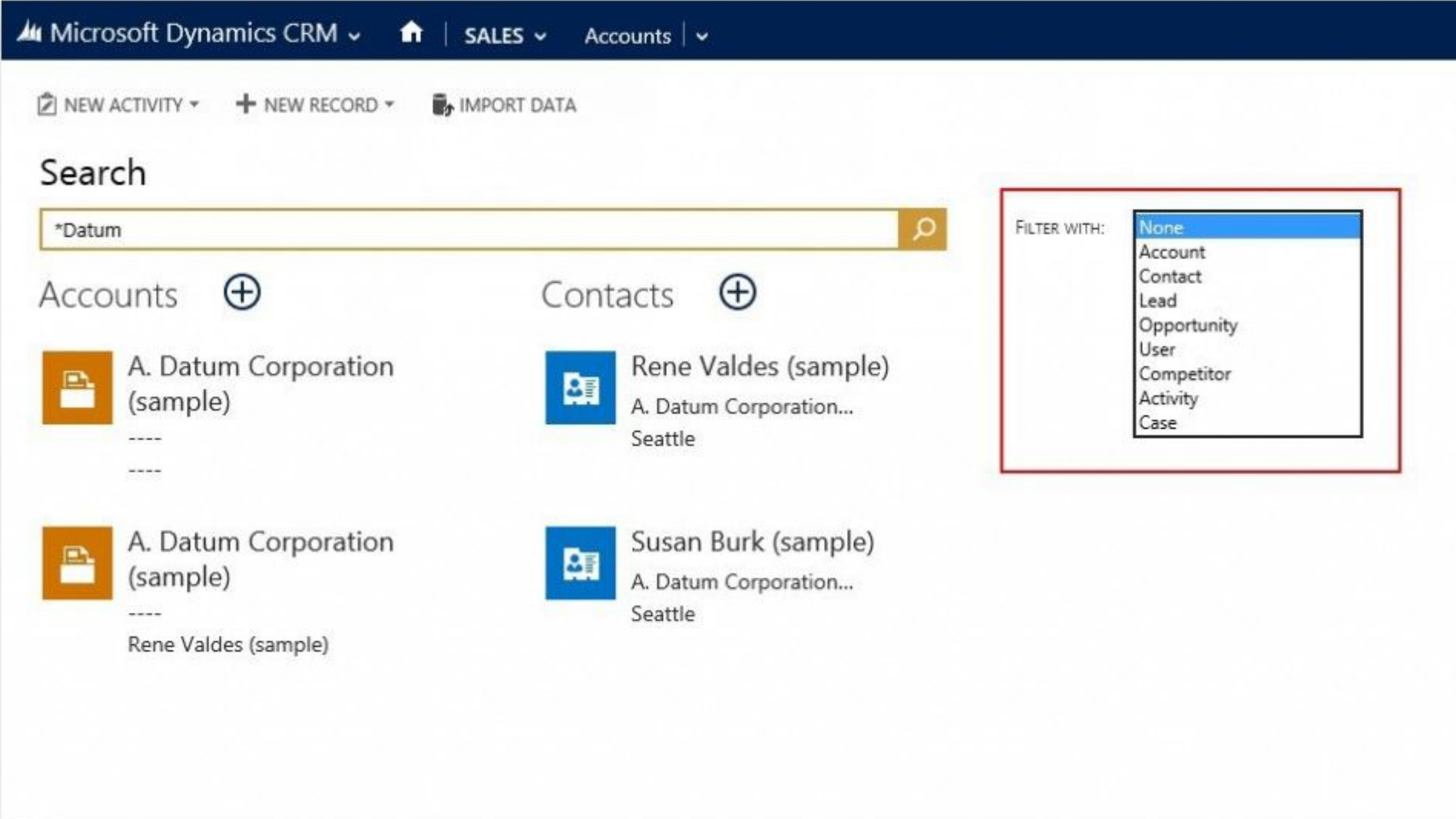
PURCHASE PRICE

□ AVAILABLE TO PARTNERS

□ OFFER SUPPORT

□ TRACK LANDED COST

□ DROP SHIP ITEM



Search

*Datum

Accounts

A. Datum Corporation (sample)

A. Datum Corporation (sample)

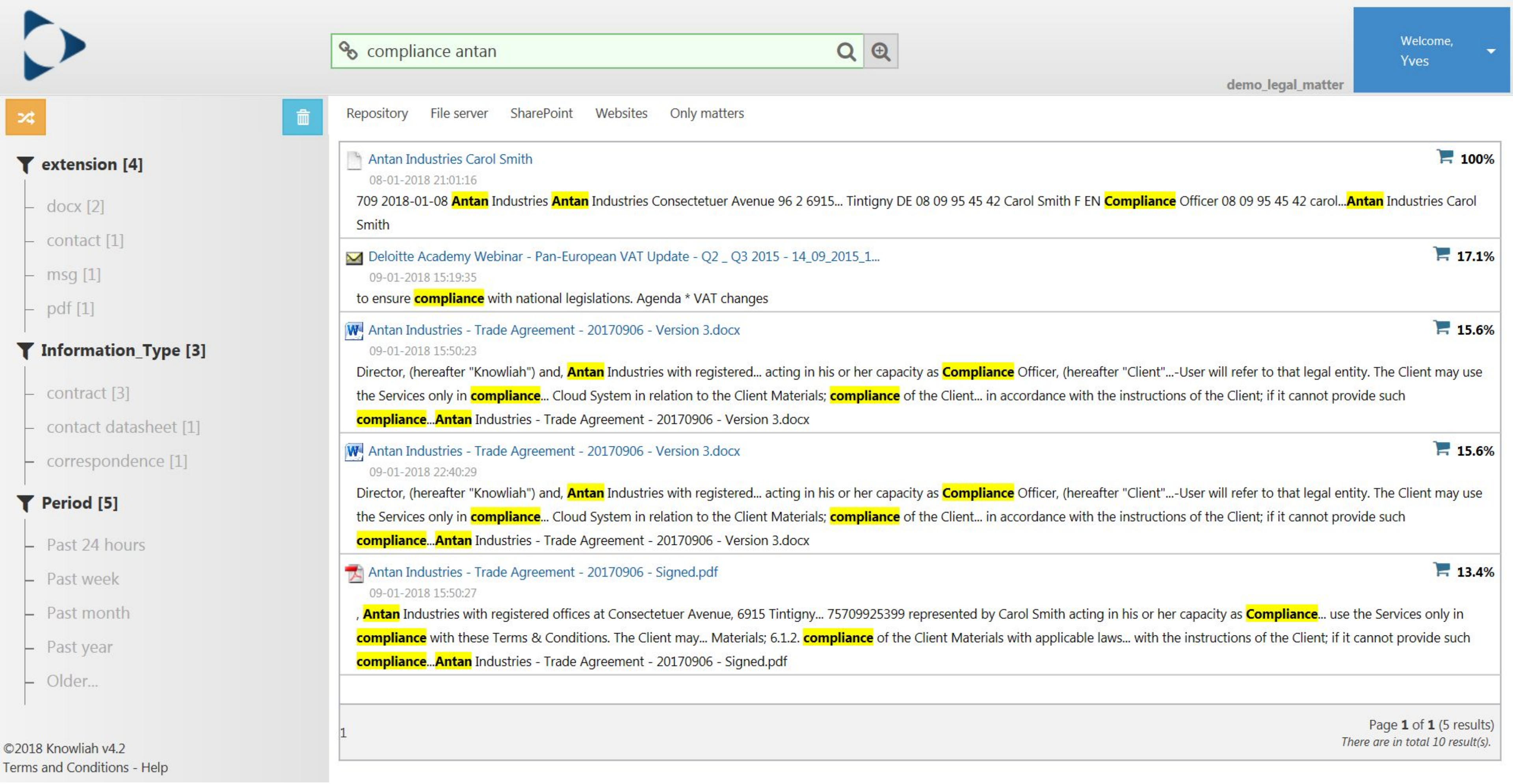
Rene Valdes (sample)

Contacts

Rene Valdes (sample)
A. Datum Corporation...
Seattle

Susan Burk (sample)
A. Datum Corporation...
Seattle

FILTER WITH: None



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

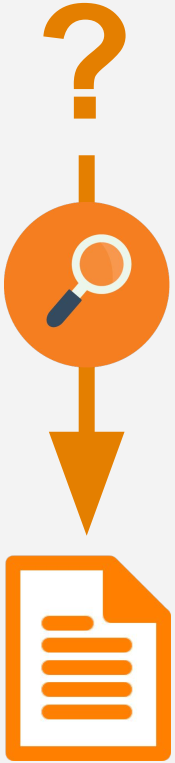
1

Page **1** of **1** (5 results)
There are in total 10 result(s).

©2018 Knowliah v4.2
Terms and Conditions - Help

Why companies should care?

- Many processes rely on quickly finding the relevant pieces of information (Customer support, Sales, Ops, ...)
- The average company is using >1000 cloud services which can be hard to unify and search the relevant information
- It has become easy to deploy and customize existing open-source search engines

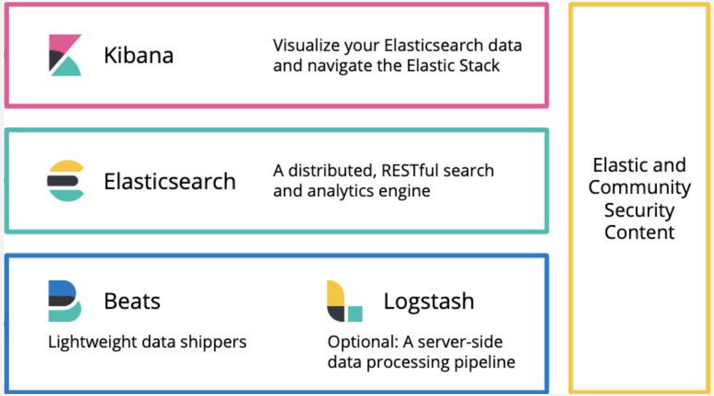


Building your own search engine

(using Elasticsearch)

10sec. intro to Elasticsearch

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



Kibana
Visualize your Elasticsearch data and navigate the Elastic Stack
Elasticsearch
A distributed, RESTful search and analytics engine
Beats
Lightweight data shippers
Logstash
Optional: A server-side data processing pipeline
Elastic and Community Security Content

Terminology in Elasticsearch

Table →	Index
Row →	Document
Column →	Field
Schema →	Mapping

⬛ Understand your search engine

(& how to get the most out of it)

Major parts of a search engine

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

~~⬛~~ Let’s build ours!

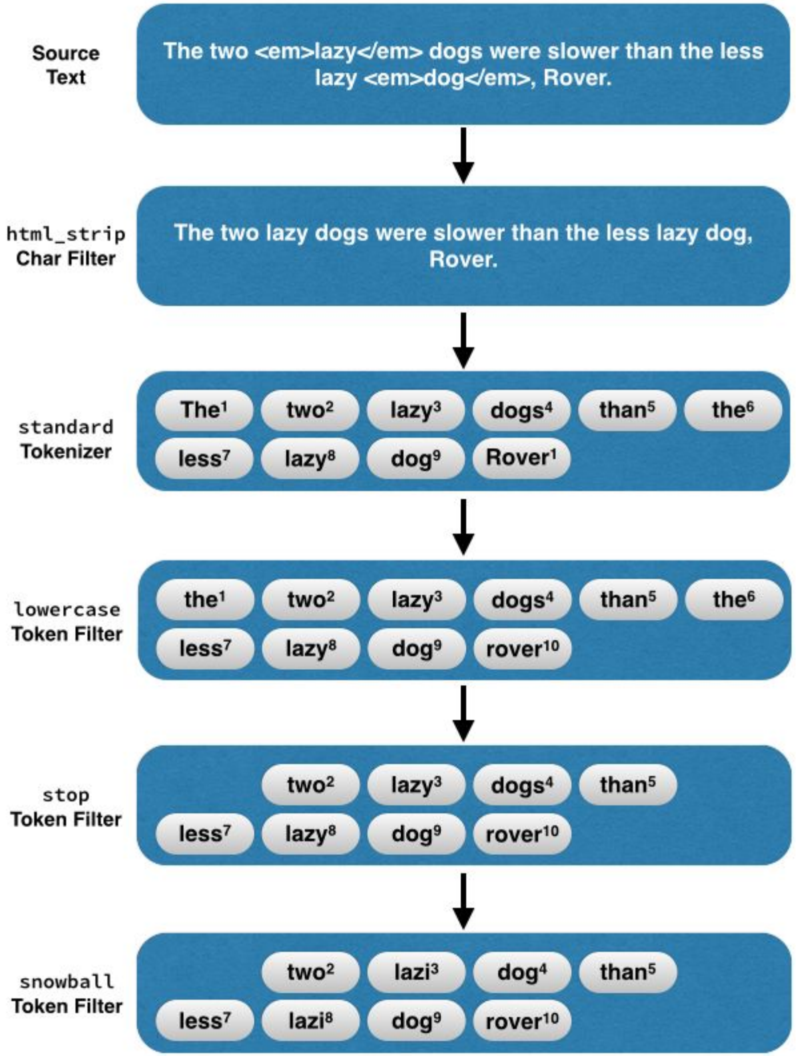
(https://github.com/unit8co/dmz-workshop-er-search)

⬛ Tips to customize your Tokenizer

(or should we say analyzer)

Tips #1: Analyzers are powerful

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



Source Text

The two <em>lazy</em> dogs were slower than the less lazy <em>dog</em>, Rover.

html_strip Char Filter

The two lazy dogs were slower than the less lazy dog, Rover.

standard Tokenizer

The^{1} two^{2} lazy^{3} dogs^{4} than^{5} the^{6} less^{7} lazy^{8} dog^{9} Rover^{1}

lowercase Token Filter

the^{1} two^{2} lazy^{3} dogs^{4} than^{5} the^{6} less^{7} lazy^{8} dog^{9} rover^{10}

stop Token Filter

two^{2} lazy^{3} dogs^{4} than^{5} less^{7} lazy^{8} dog^{9} rover^{10}

snowball Token Filter

two^{2} lazi^{3} dog^{4} than^{5} less^{7} lazi^{8} dog^{9} rover^{10}

Tips #2: Synonyms

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

```
{
  "region_synonym": {
    "type": "synonym",
    "synonyms": [
      "europe, eu, eur",
      "middle-east, me, moyen-orient",
      "india middle east africa, ima",
      "latin america, south america, lam, latam",
      "north america, nam",
      "south east asia, sea",
      "north east asia, nea",
      "united states of america, usa, us",
      "united kingdom, uk",
      "all regions, world, monde"
    ]
  }
}
```

A few examples of real-life issues

- “-” (dash) vs “–” (em-dash)
  - **Possible solution**: normalize characters in analyzers
- Special Characters in IDs
  - DEF#265/312
  - ABC*234/124
  - **Possible solution**: Change the analyzer for that field
- Language specific issues
  - “L'Équipe” vs “L Equipe”
  - “Balletttänzerin”
  - **Possible solution**: Pick the proper language analyzer
    - For German, have a look at “Word decomposers”

What we don’t have time to talk about

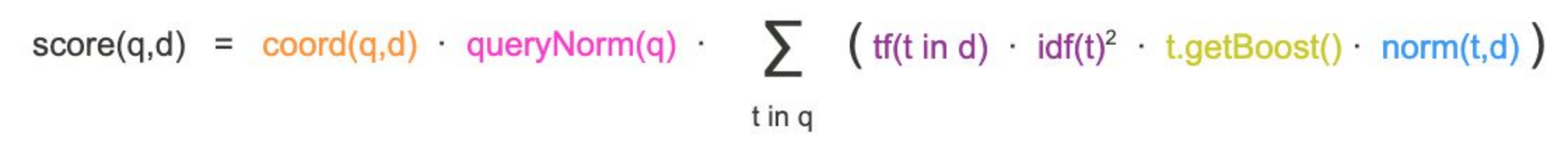
- Shingle (index groups of words for faster queries)
- Phonetic analyzers
- Pattern Capture
- ...

If you want to know more, please have a look at:

https://www.elastic.co/blog/found-text-analysis-part-1

Tips to customize your Ranking

Tips #1: Understanding ranking is key



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

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)

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 #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

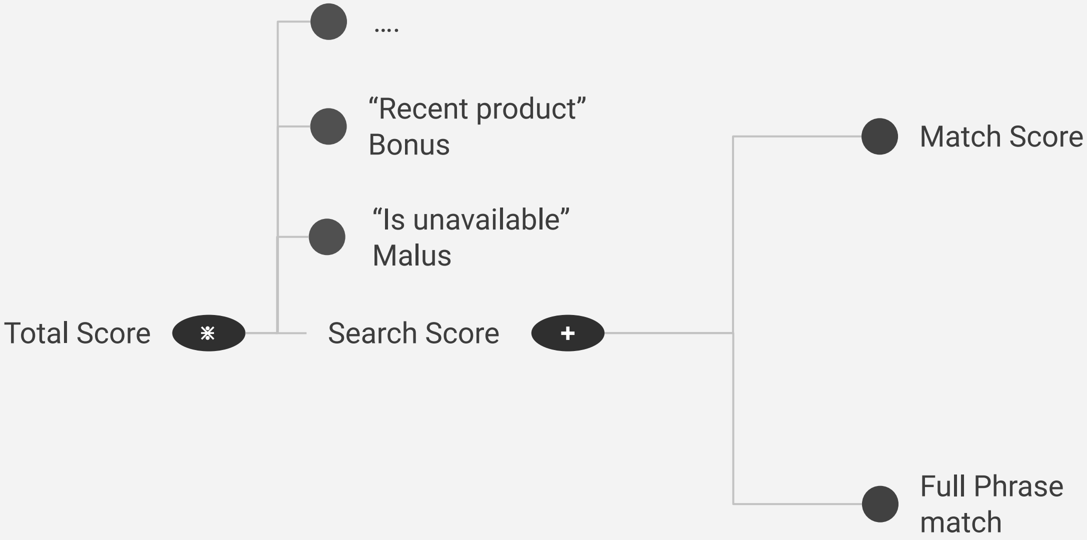
event sourcing using kafka 🔎

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

Tips#3: Function score

- Use the function score which allows to perform extra computation on document score:
    - Boost more recent products (e.g. last Mercedes S)
    - Promote “popular” items based on your fancy machine learning model
    - ...

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



....

“Recent product” Bonus

“Is unavailable” Malus

Total Score ⋇ Search Score +

Match Score

Full Phrase match

⬛ **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

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

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

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

What’s next ?

(In the world of search engines)

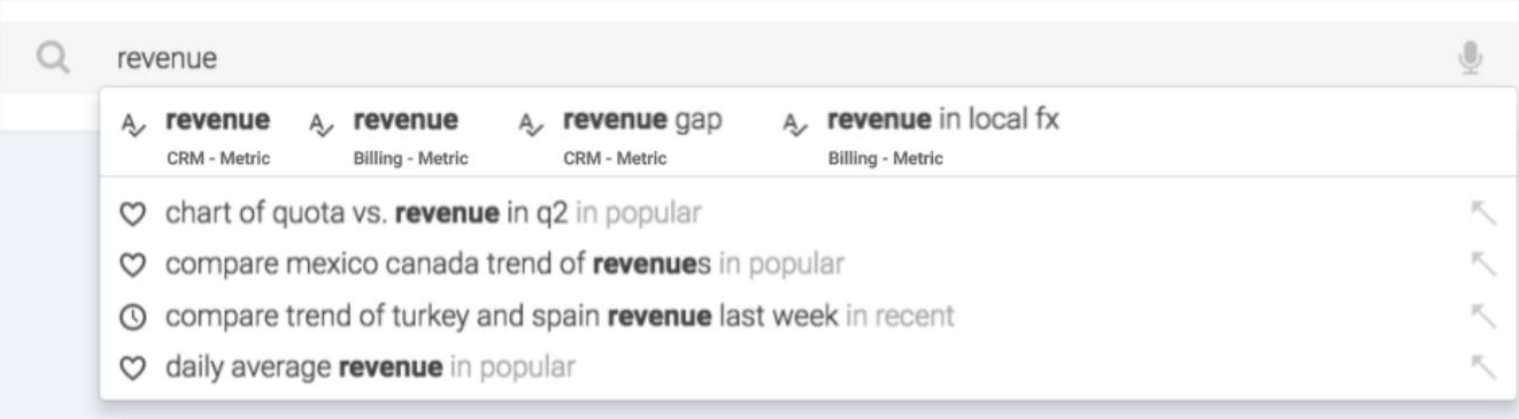
Entity Based Search

- Improve search by disambiguating some user keywords
- Highlight words belonging to an ontology

```
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

Entity Based Search



revenue

revenue
CRM - Metric

revenue
Billing - Metric

revenue gap
CRM - Metric

revenue in local fx
Billing - Metric

chart of quota vs. **revenue** in q2 in popular

compare mexico canada trend of **revenue**s in popular

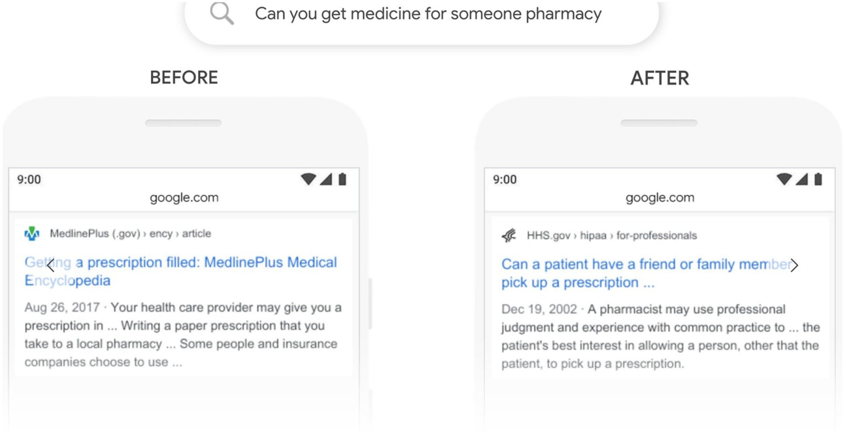
compare trend of turkey and spain **revenue** last week in recent

daily average **revenue** in popular

Unit8™

Understand user intent

- Integrate neural networks embedding (BERT) in ranking to return better results
- Easy to transfer learning to other languages than English



Can you get medicine for someone pharmacy

BEFORE

9:00
google.com

MedlinePlus (.gov) › ency › article

Getting a prescription filled: MedlinePlus Medical Encyclopedia

Aug 26, 2017 · Your health care provider may give you a prescription in ... Writing a paper prescription that you take to a local pharmacy ... Some people and insurance companies choose to use ...

AFTER

9:00
google.com

HHS.gov › hipaa › for-professionals

Can a patient have a friend or family member pick up a prescription ...

Dec 19, 2002 · A pharmacist may use professional judgment and experience with common practice to ... the patient's best interest in allowing a person, other that the patient, to pick up a prescription.

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

Conclusions

- Setting up (e.g. managed instance on the cloud) and configuring a basic search engine is relatively easy
- Including your expertise or knowledge can help providing more relevant results
- With the recent progress in NLP, probably lots of improvements coming in the next years!

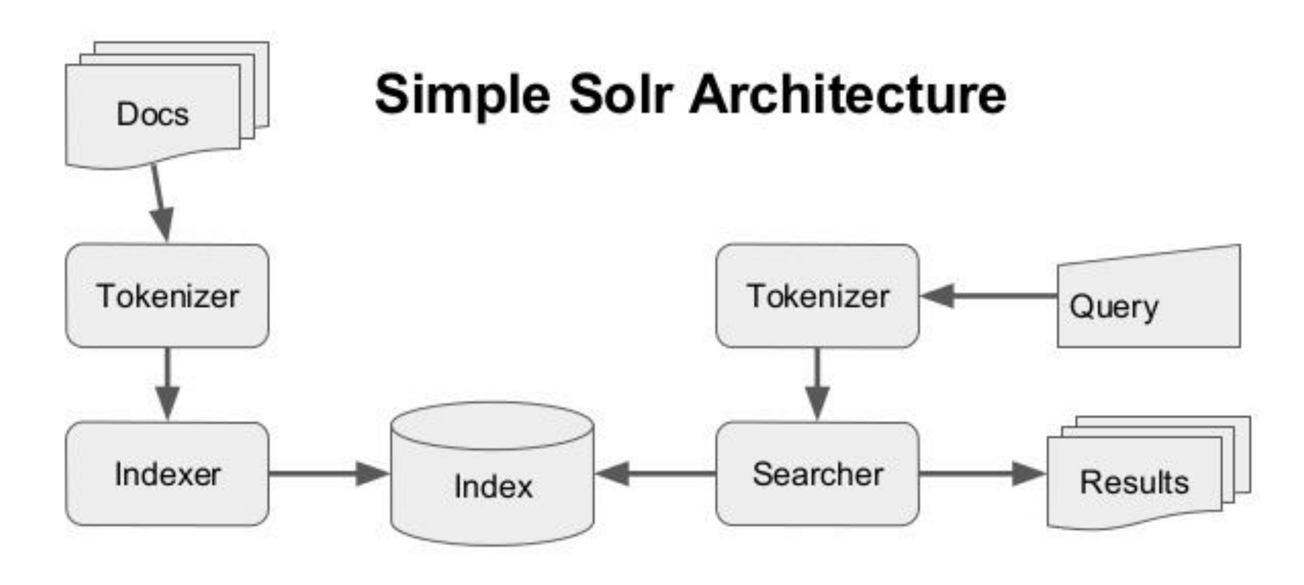
unit8.co

thank you

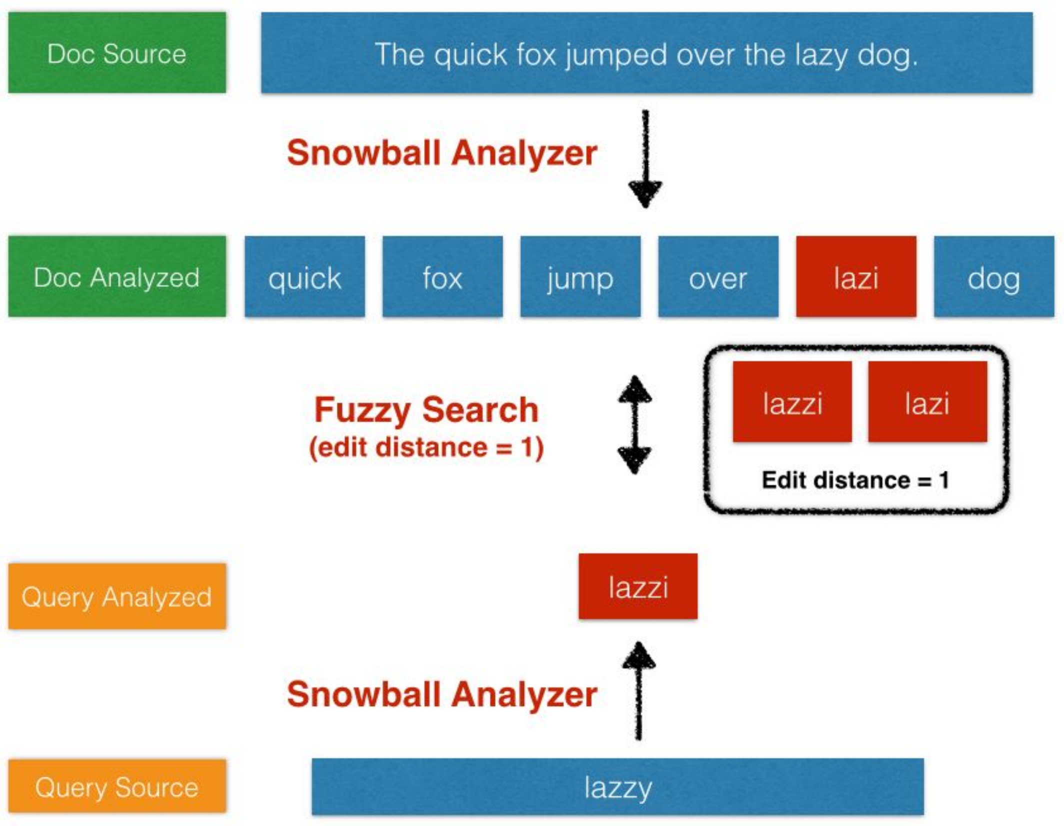
Useful Links

- 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

Steps to search engines



Fuzziness


Doc Source
The quick fox jumped over the lazy dog.
Snowball Analyzer
Doc Analyzed
quick
fox
jump
over
lazi
dog
Fuzzy Search
(edit distance = 1)
lazzi
lazi
Edit distance = 1
Query Analyzed
lazzi
Snowball Analyzer
Query Source
lazzy

Limitations of Full-Text Search over SQL

+ Easy to setup over traditional SQL DBs

- Typically Slow response time (>1 sec.)
- Doesn’t scale well
- Little control over indexing

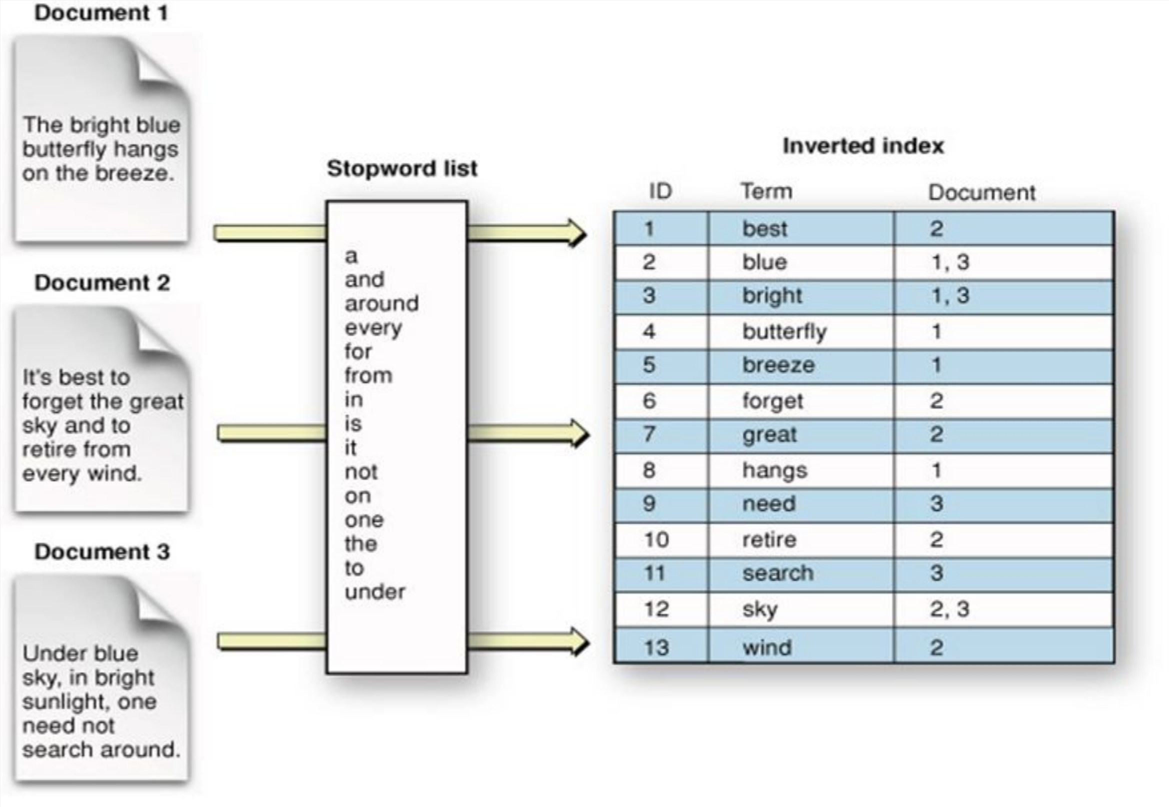
Limitation of general purpose search engines

+ Minimal configuration
+ Fast and scales well

– Does not always return the most relevant results (especially when dealing with other languages than english)

– Does not leverage specialized knowledge

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.

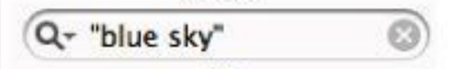
Stopword list

a
and
around
every
for
from
in
is
it
not
on
one
the
to
under

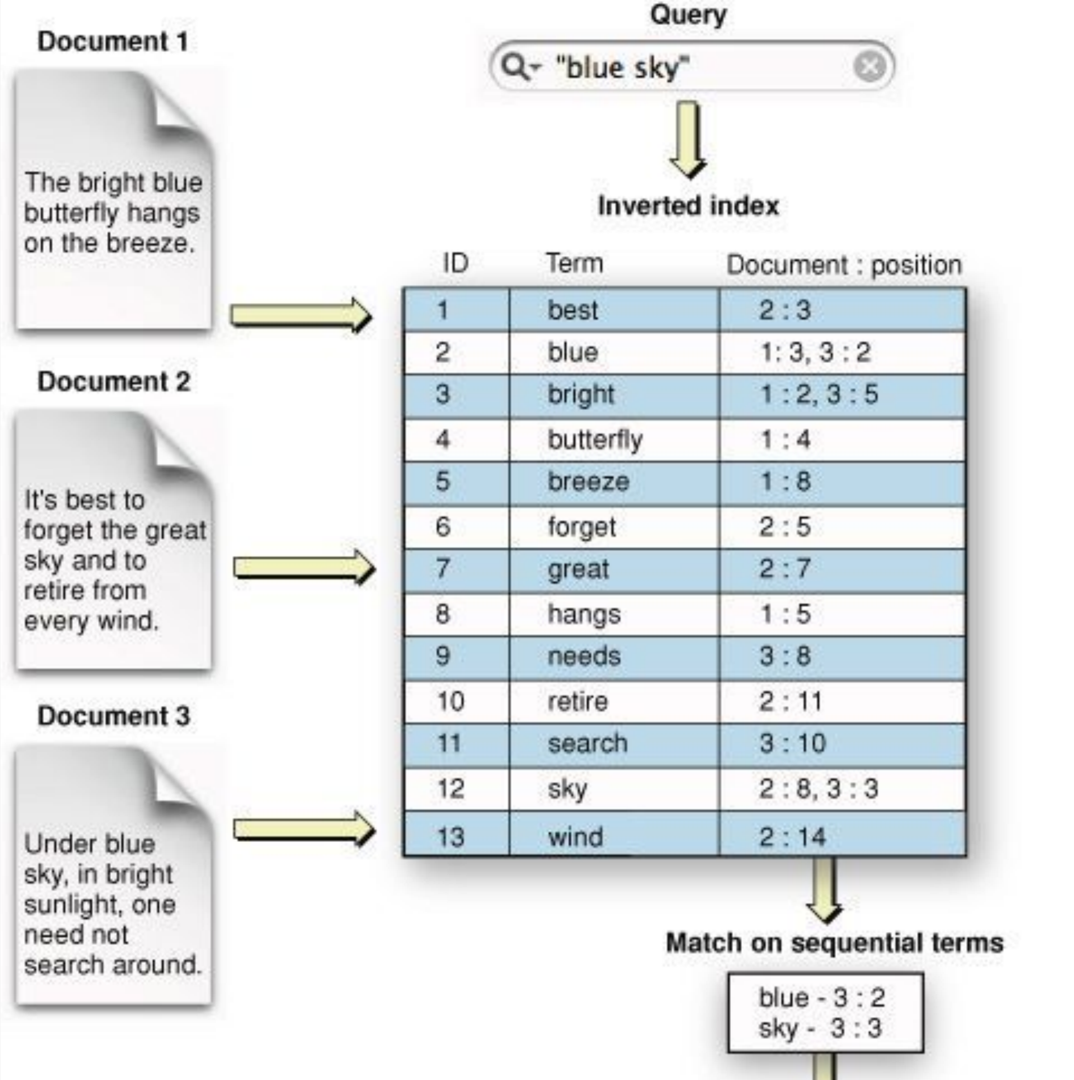
Inverted index

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?



"blue sky"



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