Improving Relevancy
It’s Harder Than You Think

Xavier Morera

PASSIONATE ABOUT TEACHING

@xmorera   www.xaviermorera.com

Improving Relevance

When results are spot on

- It is not magic or mind reading…
- It is relevance! (tf-idf)

Relevance is incredibly important

- Don’t believe me?
- Let’s take a quick quiz

Quiz

Where do you hide a dead body?

The second page of Google, because no one ever looks there!

That’s why relevance is so important!

Relevance

Degree to which query response satisfies the user who is searching for information.

It Is like a Popularity Contest

Multiple factors taken into account

All at once at query time

- Usually, except multiple queries per user query

Think in terms of search engine, not database

- Avoid “take this and put it first, then take this…”

Solr ranks results by calculating a score

- “Relevant”

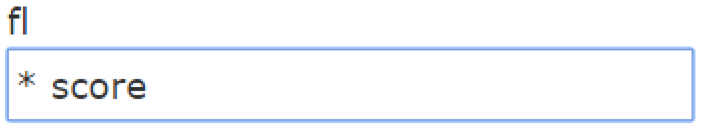
If results are not expected tune search engine until desired results

Precision vs. recall

Score?

I Want to See the Score

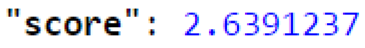
Add score as a field



fl

```
* score
```

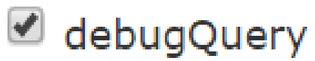
Part of returned fields



```
"score": 2.6391237
```

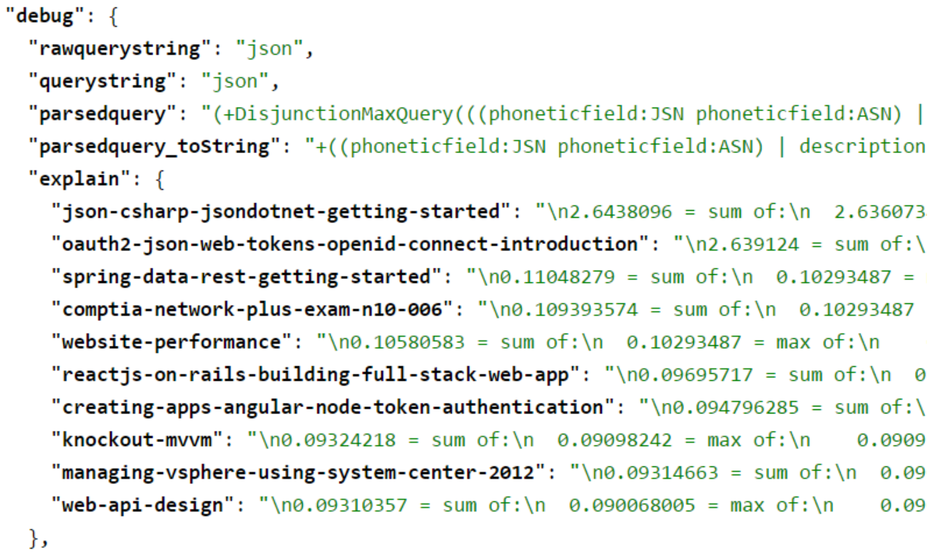
How Is the Score Calculated?

Enable debugQuery



☑ debugQuery

Inspect the response



```
"debug": {
  "rawquerystring": "json",
  "querystring": "json",
  "parsedquery": "(+DisjunctionMaxQuery(((phoneticfield:JSN phoneticfield:ASN) |
  "parsedquery_toString": "+((phoneticfield:JSN phoneticfield:ASN) | description
  "explain": {
    "json-csharp-jsondotnet-getting-started": "\n2.6438096 = sum of:\n  2.636073
    "oauth2-json-web-tokens-openid-connect-introduction": "\n2.639124 = sum of:\
    "spring-data-rest-getting-started": "\n0.11048279 = sum of:\n  0.10293487 =
    "comptia-network-plus-exam-n10-006": "\n0.109393574 = sum of:\n  0.10293487
    "website-performance": "\n0.10580583 = sum of:\n  0.10293487 = max of:\n
    "reactjs-on-rails-building-full-stack-web-app": "\n0.09695717 = sum of:\n  0
    "creating-apps-angular-node-token-authentication": "\n0.094796285 = sum of:\
    "knockout-mvvm": "\n0.09324218 = sum of:\n  0.09098242 = max of:\n    0.0909
    "managing-vsphere-using-system-center-2012": "\n0.09314663 = sum of:\n  0.09
    "web-api-design": "\n0.09310357 = sum of:\n  0.090068005 = max of:\n    0.09
  },
```

Precision vs. Recall

Precision

Results returned are what the user was looking for

Matching docs/Total docs returned

Recall

Results include all documents that should be returned

Matching docs/(Matching docs + Missed docs)

Some Ways of Improving Relevance

Boosting, boost queries and boost functions

Stemming & lemmatization

Stop words

Synonyms

Spell check

Phonetic search

Multilingual search

What Happens Frequently

Fix This

"Can you fix this scenario please?"

"We always want this type of doc first!"

"Just boost this metadata^1 000 000"

You do it

And then you get big bucks and gifts…

Well… maybe not always….

Break That

By fixing one thing, you may break many

Measure overall impact (side effects)

Your users might feel like this…

Avoid “Fix This… Break That”

Measure overall impact

Analyze carefully each change

Test multiple scenarios

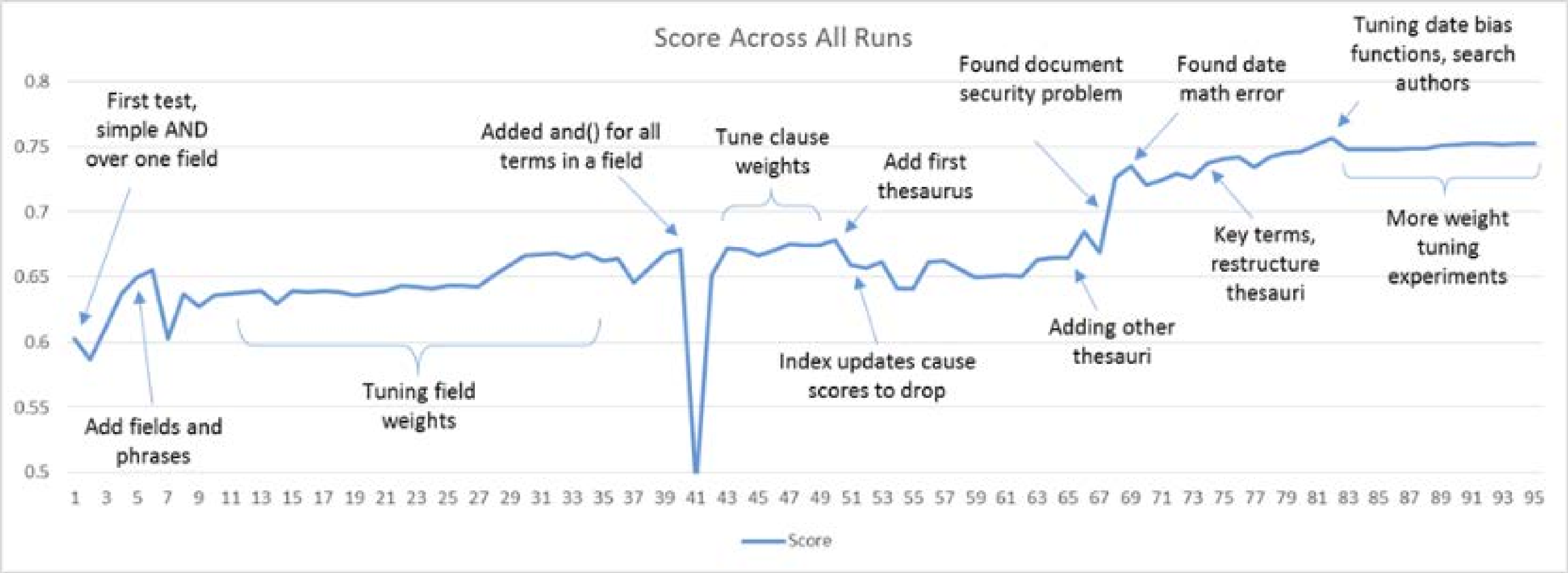
If possible implement a process

- Repeatable
- Measurable

Many ways of doing this

A recommended approach is

Search Engine Scoring: Building a Smarter Search Engine



Recommend:

http://www.searchtechnologies.com/search-engine-relevance-scoring

http://www.searchtechnologies.com/search-engine-scoring-webinar

Demo

Improving Relevancy

M7 D1

Query Fields

Specify which fields in the index should be used to perform the query

If no parameter, defaults to df

Boost

Ability to assign higher importance to fields or field values

Numeric

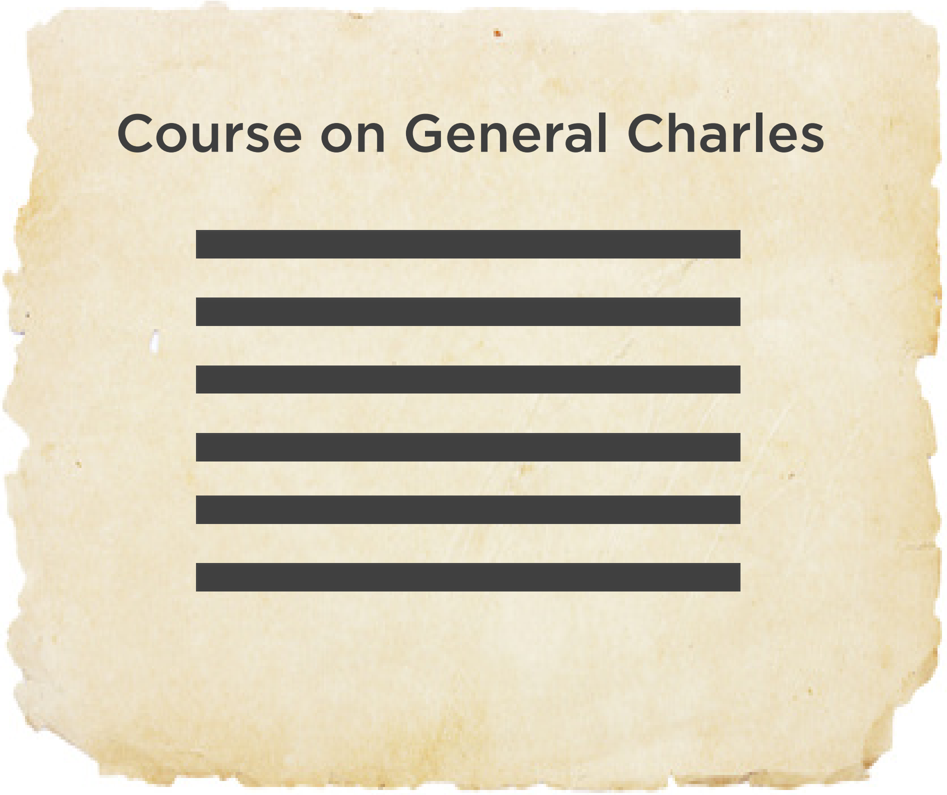
Index time or query time

Boosting: Not All Fields are Created Equal

General Charles

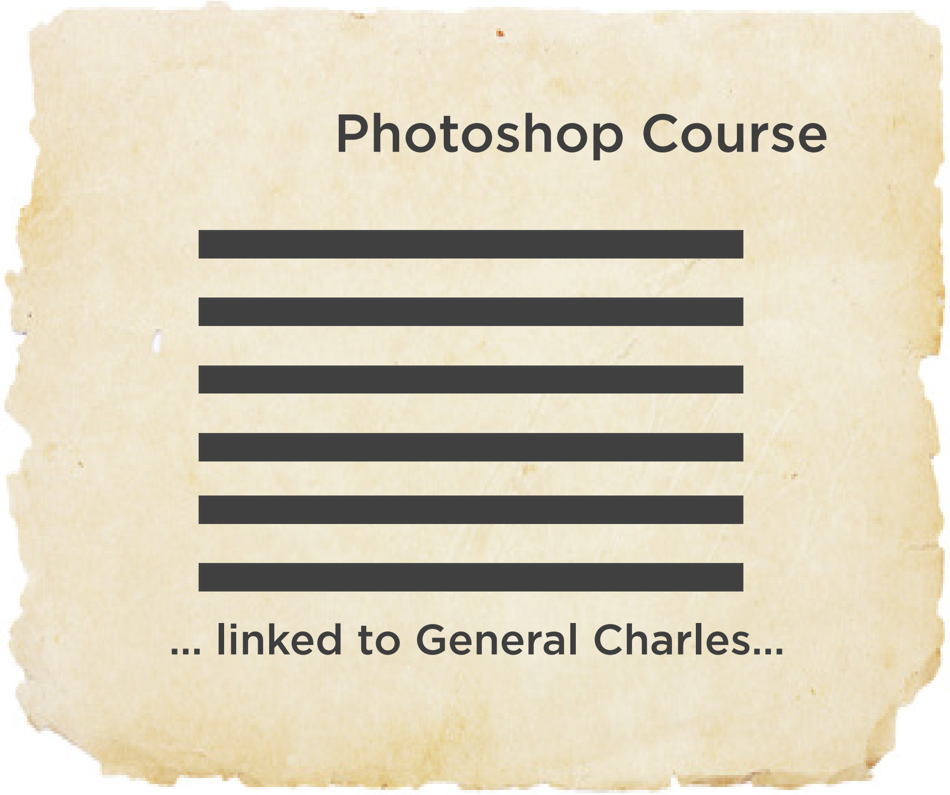
Boosting: Not All Fields are Created Equal

Course on General Charles



Course on General Charles

Mentions General Charles



Photoshop Course

... linked to General Charles...

Boosting: Not All Fields are Created Equal

Is Title Text More Important?

Course on General Charles

Photoshop Course

... linked to General Charles...

Is Text in The Body More Important?

Photoshop Course

... linked to General Charles...

Course on General Charles

```
var q = new SolrQuery(“coursetitle:query").Boost(10d) +
          new   SolrQuery(“coursedescription:query”);

var courses = solr.Query(q);
```

Boosting at Query Time

Many ways of applying boosts (query or query fields)

Can be applied at query time or index time

Using SolrNet or in Solr

Demo Summary

Not all fields are created equal

Boost applies higher search value

Important fields

Several ways

- SolrQuery().Boost()
- qf
- Solrconfig.xml

Boost Queries

Specify boost factor for certain types of documents

- Regardless of user query (i.e. Sponsored)
- Ran after initial user query
- Boost documents from results of first search that match the applied boost
- Specific values

Via bq in Solr and ExtraParams in SolrNet

Boost Queries

- User searches for JSON
- Two courses, 1 webinar and 1 play by play
- All have very similar
- We want to boost courses *(just an example)*

```
bq=coursetype:”course”^10.0
```

- Solr selects all documents that match
- And then boosts based on course type
- Requires SME knowledge

1. JSON & Json.NET
2. MongoDB
3. Json vs. XML Webinar
4. Json Play by Play

Demo Summary

Push to the top specific types of results

Ran after initial result set

Specific boost

bq or ExtraParams

Boost Functions

Very similar to Boost Queries

- Helps achieve same objectives

Function instead of specific values

Typical example is boost by date

- Usually newer content more up to date
- And perhaps more important

Boost Functions

Does not influence “hits”

Only boost applied on initial result set

Via bf parameter in Solr

ExtraParams in SolrNet

Demo Summary

Boost functions are similar to boost queries

Functions

Not static

Typical: Boost by date

Stemming & Lemmatization

Stemming

Helps match words even if not exact matches

Crude heuristic process to reduce words

At both index and query time

Walking vs. Walk

Aim of increasing recall

Porter stemmer is one of the most common

Tradeoff → false positives vs. false negatives

Lemmatization

Helps match words even if not exact matches

Similar to stemming (increase recall)

Aide of morphological analysis and vocabulary

Reduce words to lemma

Saw to See or Saw depending on context

Better lemmatizes to Good

Natural language processing

Harder to implement and more expensive

Demo Summary

Stemming helps improve relevancy

Not only exact matches

Stem

Not 100% accurate (algorithm)

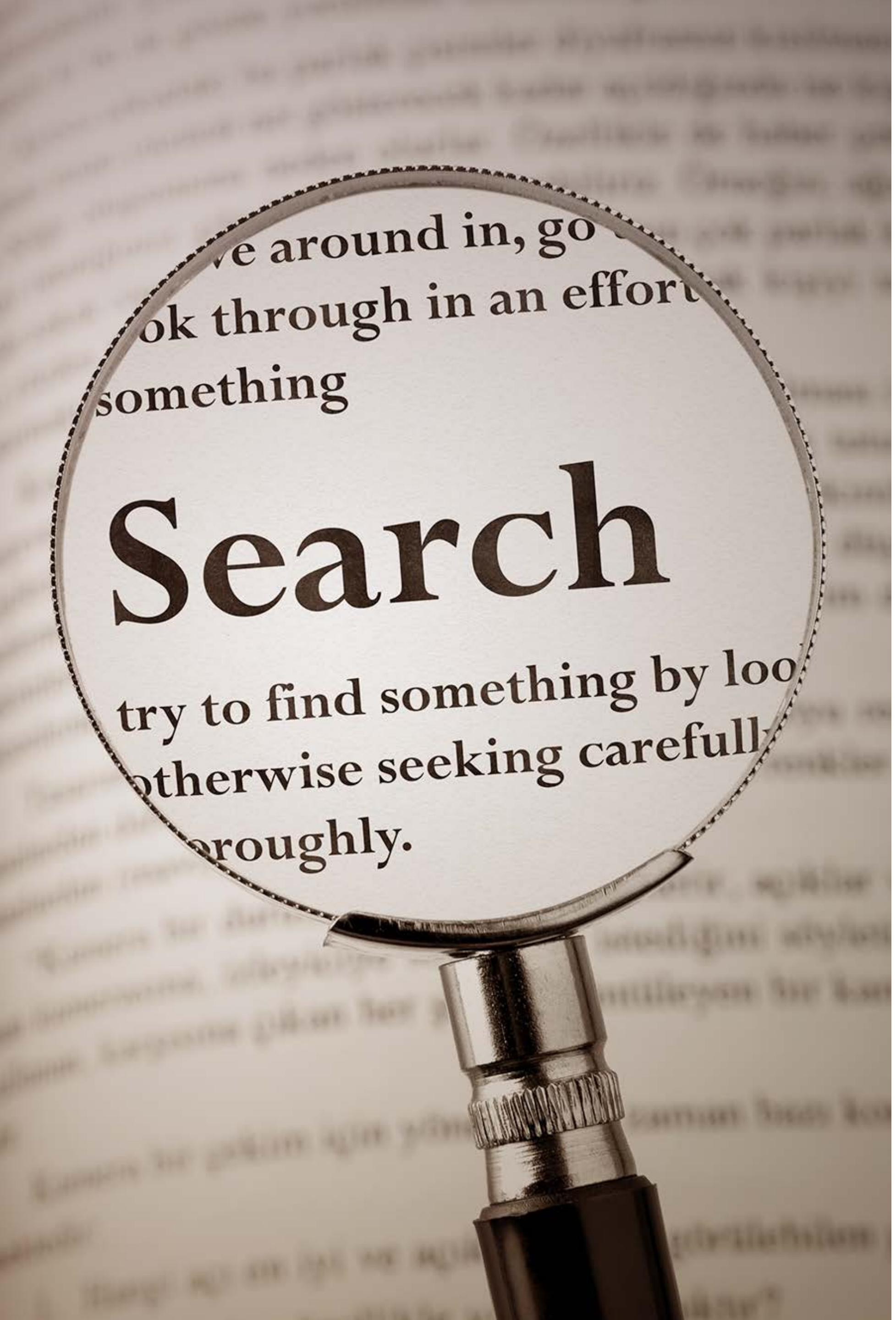
Lemmatization uses morphological analysis and vocabulary

Lemma

Much more precise but harder

syn·o·nym

a word or phrase that means exactly or nearly the same as another word or phrase in the same language, for example *fight* is a synonym of *battle*.



Synonyms

Help increase recall

- Same idea

Scenarios

- Real synonyms : clash,battle,fight
- Acronyms: gb,gigabyte
- Spelling corrections: solar,solr

Create dictionary (or can buy them!)

Defined in synonyms.txt in Solr

Synonyms

Toyota Yaris

Toyota Echo

Demo Summary

Synonyms help increase recall

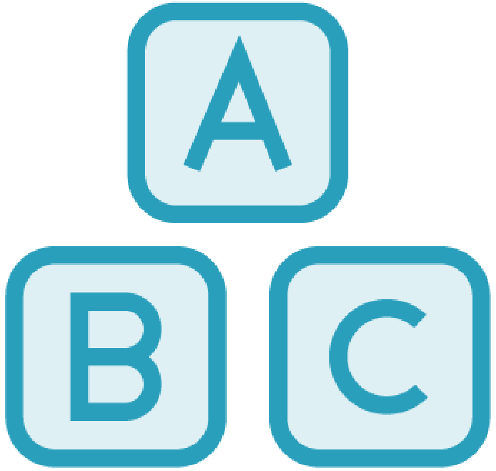
- "It is the same idea"
- Products with different names

Very easy to define in Solr

- In synonyms.txt

Order of analysis chain matters

Spellcheck



Mistakes happen

Typos happen

Solr helps by spellchecking

Offers query suggestions

Based on

- Terms in current fields
- Externally created files
- Fields in other indexes

Multiple spellcheck components in Solr

Demo Summary

Did you mean?

Helps by providing alternatives to users

Give suggestions

Improve relevancy

Spellcheck component

Phonetic Search

Match different words

With the same pronunciation

Common scenario for non native speakers

For example:

- Someone said: *learn angularjs*
- Someone heard: *learn angiularjs*

Solr can help which increases recall

Other example:

- Swardsenegger matches Schwarzenegger

Demo Summary

Match words with different spellings

But with same pronunciation

Common in non native languages

Helps increase recall

And improve search experience

Multilingual Search

Many languages

Each has their own rules

Solr’s built in support for multiple languages

Define fields with language specific type

- Use text_en vs. text_es
- Use appropriate qf

Fields have their own rules

- i.e. Language specific stemming
- Managing compound words

Multilingual Search

In English field

rents will be analyzed to rent

Let's take a word in another language

alquiler means rent in Spanish

alquileres should be analyzed to alquiler

- In Spanish field

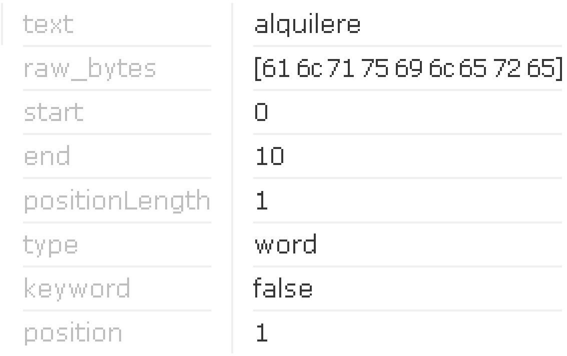
What happens if you use other field type?

- i.e. English

Multilingual Search

English

alquileres

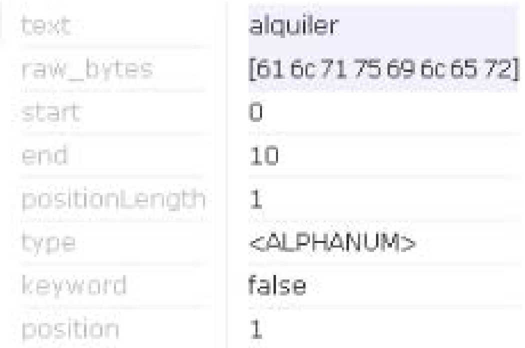


text	alquilere
raw_bytes	[61 6c 71 75 69 6c 65 72 65]
start	0
end	10
positionLength	1
type	word
keyword	false
position	1

👎

Spanish

alquileres



text	alquiler
raw_bytes	[61 6c 71 75 69 6c 65 72]
start	0
end	10
positionLength	1
type	<ALPHANUM>
keyword	false
position	1

👍

Demo Summary

Built in support for multiple languages

Use corresponding field types

Use qf to search in specific languages

- User mapped to particular language
- Map to several

Solr can try to autodetect languages

- Tika or LangDetect

Takeaway

First page results

Relevancy

Improving relevancy is hard

Precision vs. recall

Boosting, boost queries and boost functions, Stemming & lemmatization, Stop words, Synonyms, Spell check, Phonetic search, Multilingual search and more

Avoid fix this, break that

- Repeatable and measureable
- Search engine scoring