Searching & Relevance

Xavier Morera
@xmorera
www.searchtechnologies.com

pluralsight
hardcore dev and IT training

Searching in Solr and in General

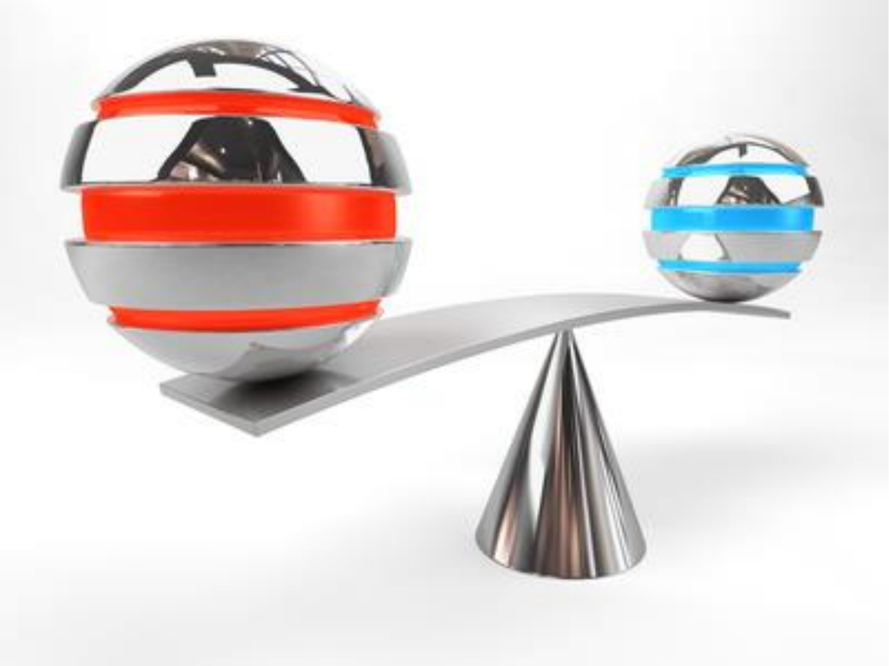
- **People love searching for things!**
- **Not really…**
- **People love finding!**
  - □ Return the results most relevant to their query and let them fine tune
- **Cookie points if *magically* results are what the user wants!**

Relevance

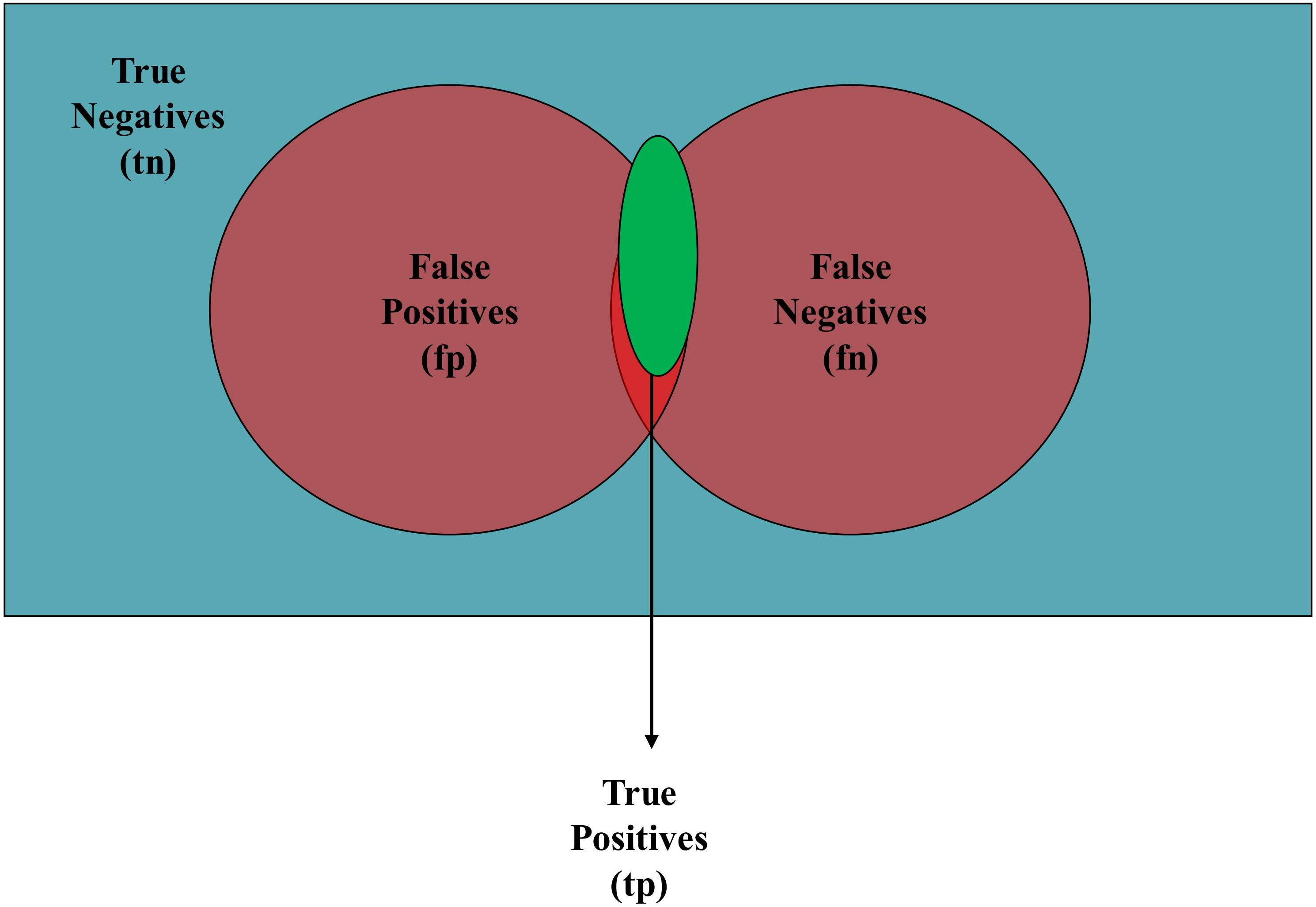
- It is not magic, it is *relevance*!
- Is the degree to which a query response satisfies the user who is searching for information.

Concepts Related to Relevance

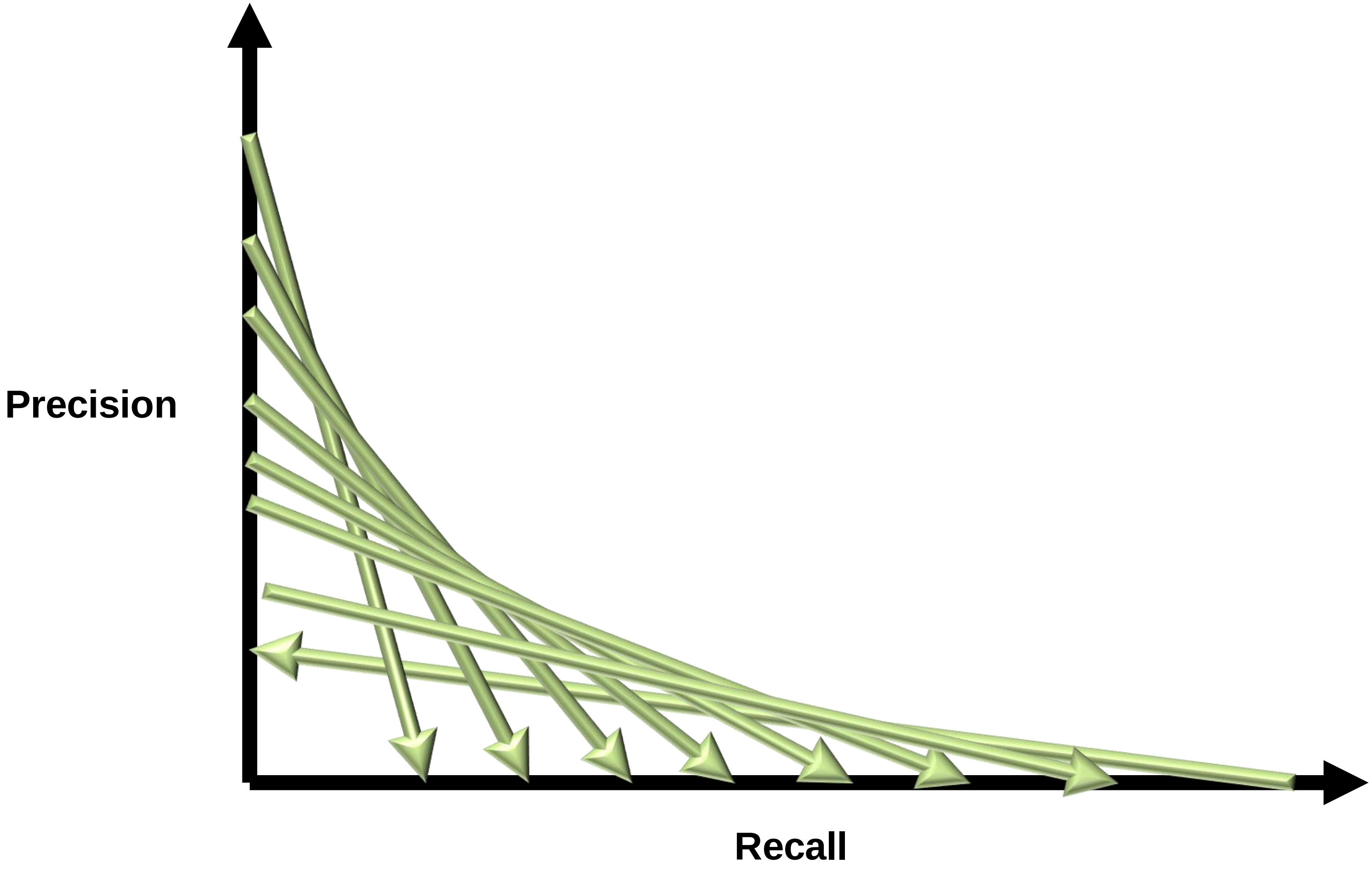
- **Precision is the percentage of documents in the returned results that are relevant.**
- **Recall is the percentage of relevant results returned out of all relevant results in the system.**
  - □ Obtaining perfect recall is trivial: simply return every document in the collection for every query.



The Problem is Relevance



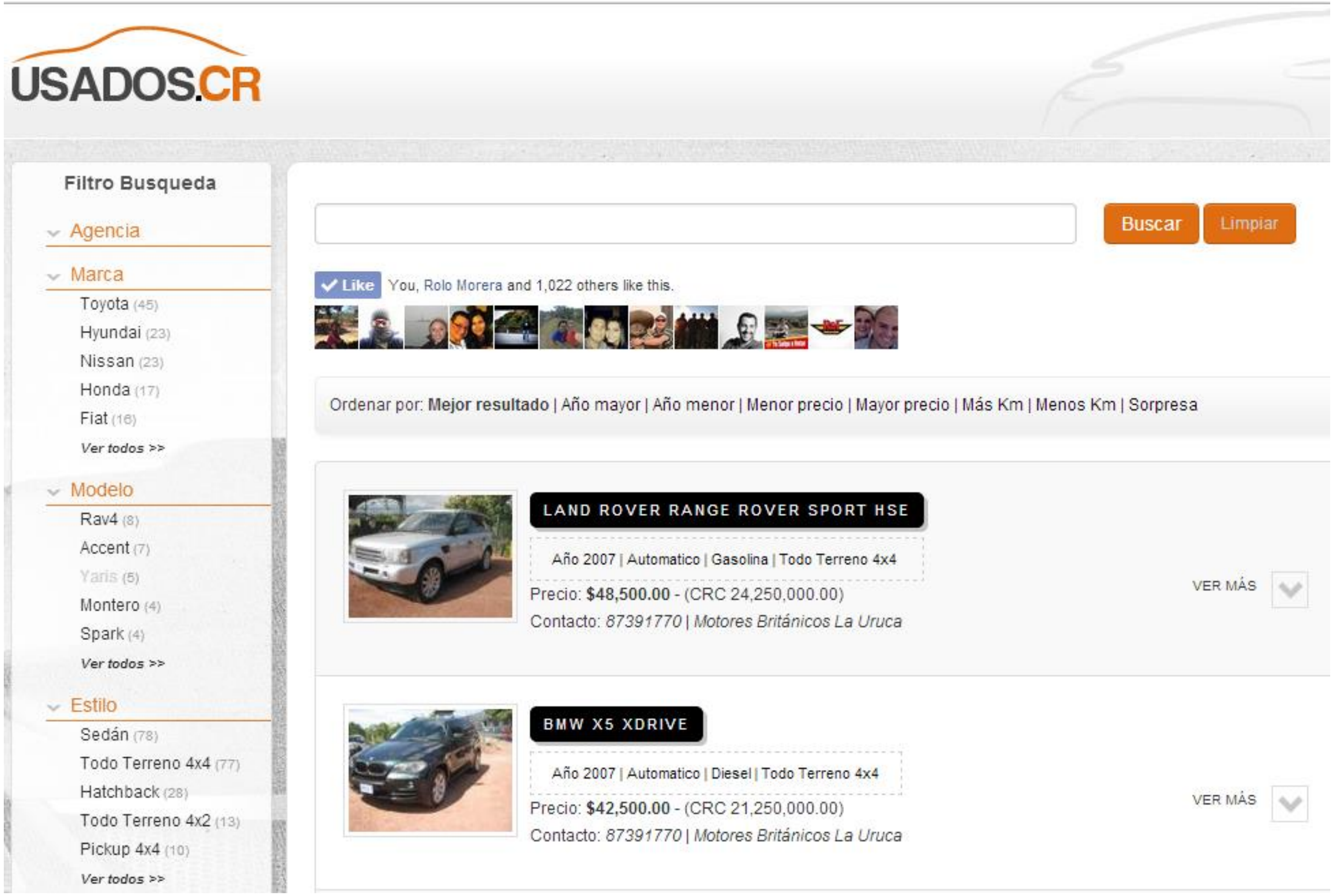
Accuracy Is a Trade off



Not All Results Are Created Equal

- Consider user's needs
- Take into account categories for each context
- Inherent relevance of documents
- Document age
- Security
- (And never forget speed)

Demo: Real Life Searching in Solr



Filtro Busqueda

Agencia

Marca
- Toyota (45)
- Hyundai (23)
- Nissan (23)
- Honda (17)
- Fiat (16)
- Ver todos >>

Modelo
- Rav4 (8)
- Accent (7)
- Yaris (5)
- Montero (4)
- Spark (4)
- Ver todos >>

Estilo
- Sedán (78)
- Todo Terreno 4x4 (77)
- Hatchback (28)
- Todo Terreno 4x2 (13)
- Pickup 4x4 (10)
- Ver todos >>

You, Rolo Morera and 1,022 others like this.

Ordenar por: **Mejor resultado** | Año mayor | Año menor | Menor precio | Mayor precio | Más Km | Menos Km | Sorpresa

LAND ROVER RANGE ROVER SPORT HSE

Año 2007 | Automatico | Gasolina | Todo Terreno 4x4

Precio: **$48,500.00** - (CRC 24,250,000.00)

Contacto: *87391770* | *Motores Británicos La Uruca*

BMW X5 XDRIVE

Año 2007 | Automatico | Diesel | Todo Terreno 4x4

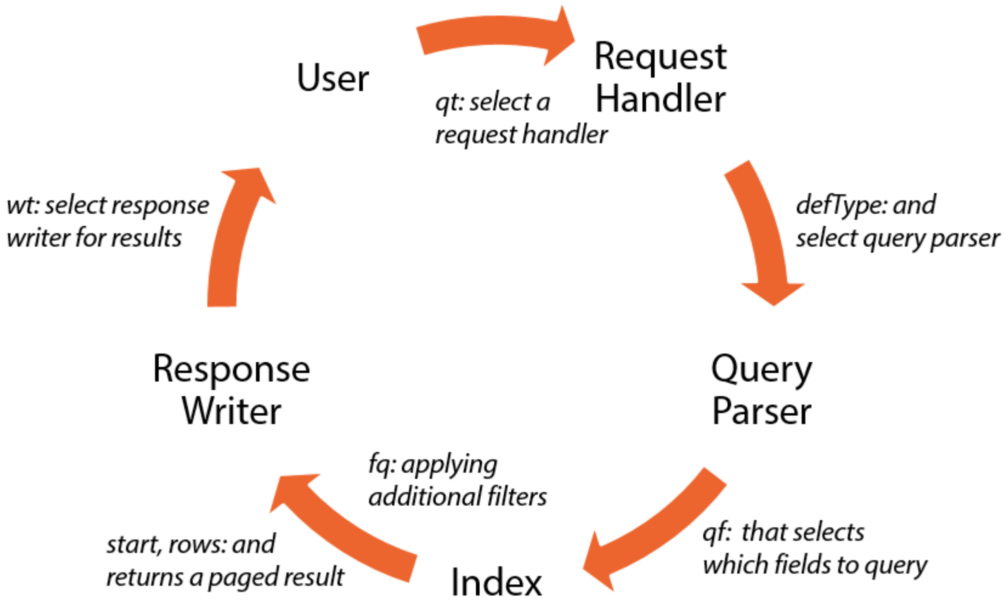
Precio: **$42,500.00** - (CRC 21,250,000.00)

Contacto: *87391770* | *Motores Británicos La Uruca*

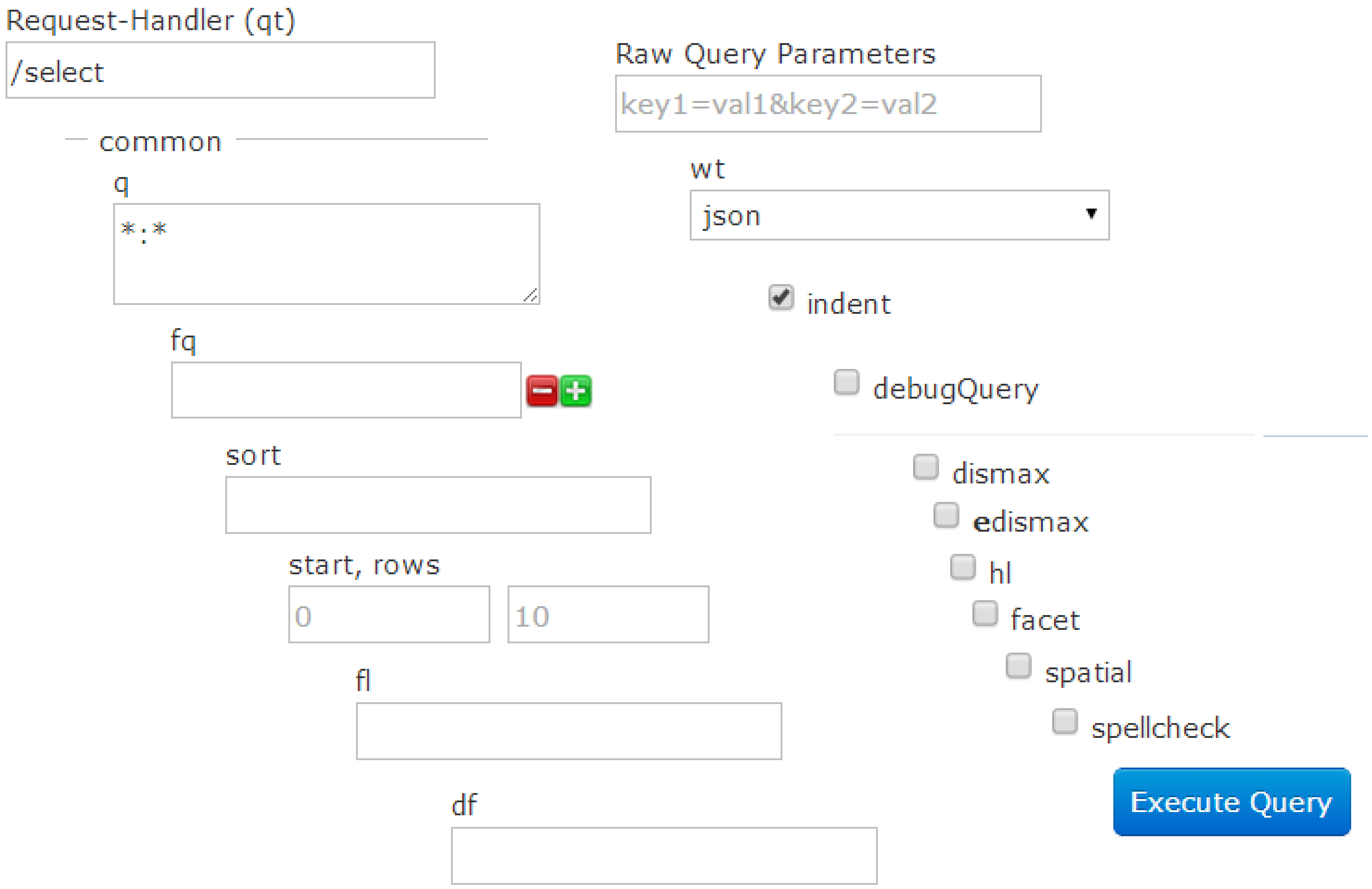
Searching in Solr

Amazing search! Let's see how it is done

- **Query by user**
- **Processed by Request Handler**
- **That calls a Query Parser**
  - □ Standard/DisMax/eDisMax
  - □ Common query parameters
- **Obtain results**
  - □ Paged and in selected response format


User
Request Handler
qt: select a request handler
defType: and select query parser
Query Parser
qf: that selects which fields to query
Index
fq: applying additional filters
start, rows: and returns a paged result
Response Writer
wt: select response writer for results

Searching in Solr



Request-Handler (qt)

/select

common

q

:

fq

sort

start, rows

0

10

fl

df

Raw Query Parameters

key1=val1&key2=val2

wt

json

☑ indent

□ debugQuery

□ dismax

□ edismax

□ hl

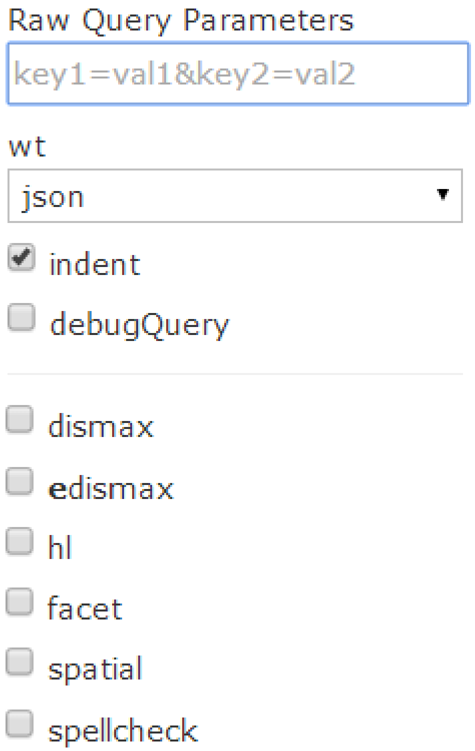
□ facet

□ spatial

□ spellcheck

Raw Query Parameters

- Admin UI shows only a small subset of params
- Use it for all other params
- It is the way to use the entire API
- Admin UI only for humans
- Applications connect via API



Raw Query Parameters

key1=val1&key2=val2

wt

json ▾

☑ indent

□ debugQuery

□ dismax

□ edismax

□ hl

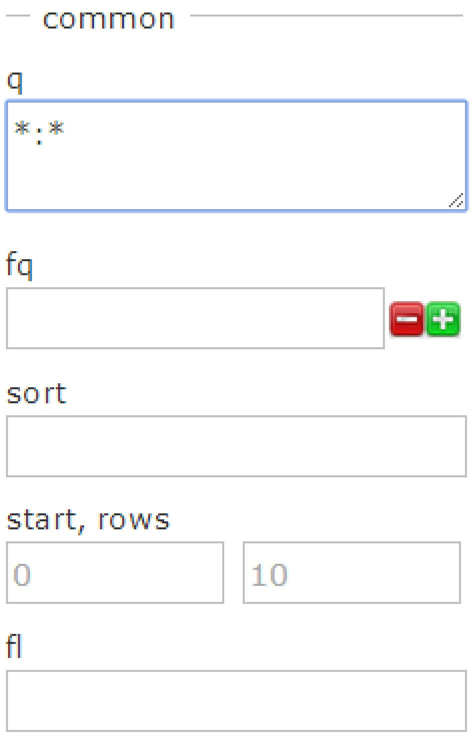
□ facet

□ spatial

□ spellcheck

q

- **Query event (mandatory)**
- **What you are searching for**
- **Results are ordered by relevancy**
  - Score



common

q

:

fq

sort

start, rows

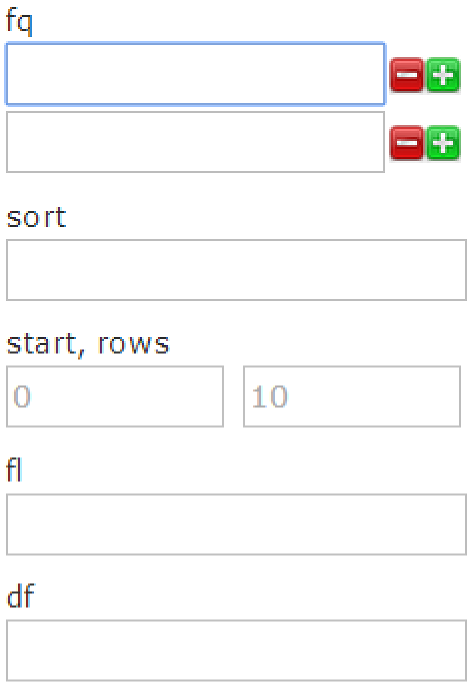
0

10

fl

fq

- **Filter query**
  - Applied to restrict superset results (q)
- **Drill down, without affecting score**
  - Most relevant documents still at the top
  - Caution: some people ignore q and use only fq
    - Relevance might not be appropriate
- **Useful for performance of complex queries**
- **Can specify multiple fq**
  - Together or separate
  - Cached independently
- **Results include only intersection of fq**



fq

sort

start, rows

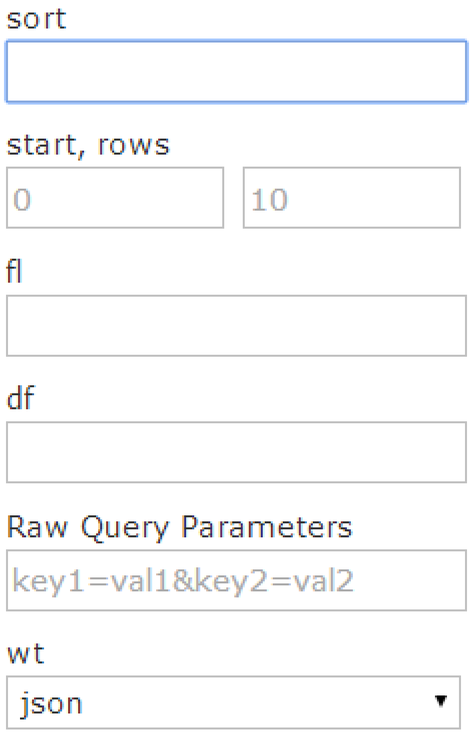
0 10

fl

df

sort

- Sort response in ascending or descending order
- Based on score or other field
- I can sort if
  - Single valued
  - Not tokenized
    - Unless single term
  - Date
  - Numbers
  - Alphabetically



sort

start, rows

0 10

fl

df

Raw Query Parameters

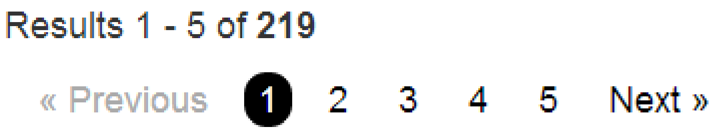
key1=val1&key2=val2

wt

json ▾

start, rows

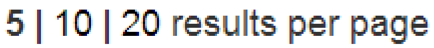
- **Used for pagination**
- **Start is the offset in the query results**
  - i.e. start at document 10
  - Default is 0



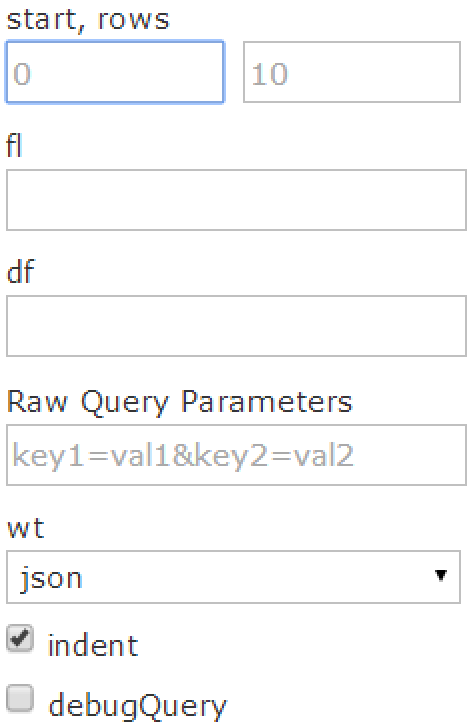
Results 1 - 5 of **219**

« Previous **1** 2 3 4 5 Next »

- **Rows determines how many results to return**



5 | 10 | 20 results per page



start, rows

0 | 10

fl

df

Raw Query Parameters

key1=val1&key2=val2

wt

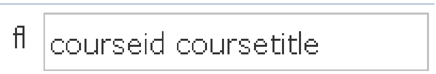
json ▾

☑ indent

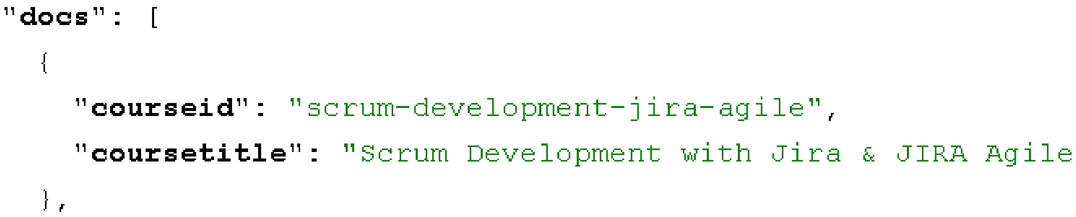
☐ debugQuery

fl

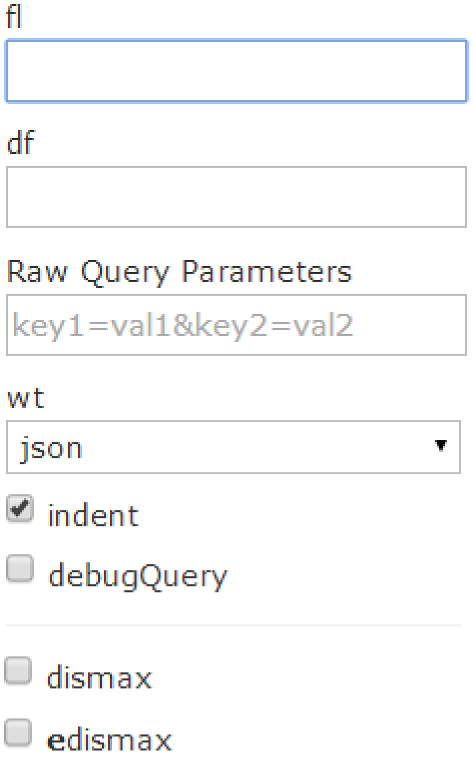
- Fields to return for each document
- Default *
- Can include score
- Separate with comma or space
- Results of functions can be included
- Recommended to avoid returning always everything



fl: courseid coursetitle



```
"docs": [
  {
    "courseid": "scrum-development-jira-agile",
    "coursetitle": "Scrum Development with Jira & JIRA Agile
  },
```



fl:

df:

Raw Query Parameters: key1=val1&key2=val2

wt: json

☑ indent

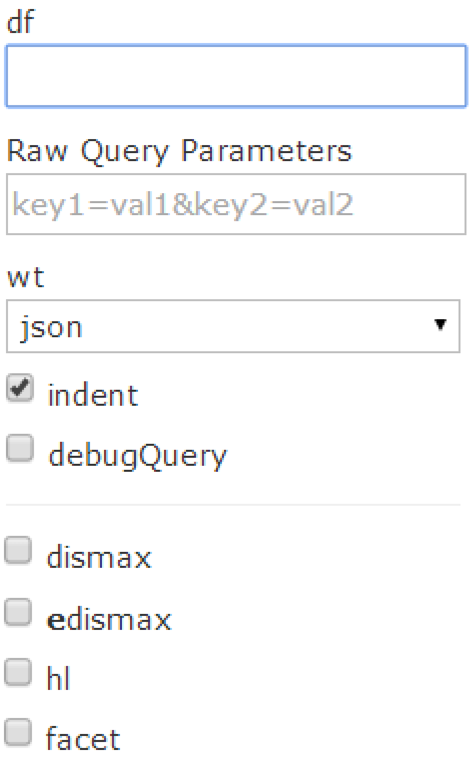
□ debugQuery

□ dismax

□ edismax

df

- **Default search field**
- **Only takes effect if qf not defined**
  - Dismax and eDismax
- **Overrides definition of a default field in the schema.xml**



df

Raw Query Parameters

key1=val1&key2=val2

wt

json ▾

☑ indent

☐ debugQuery

☐ dismax

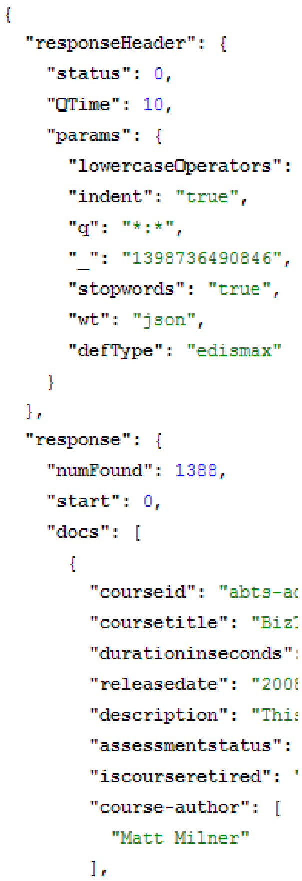
☐ edismax

☐ hl

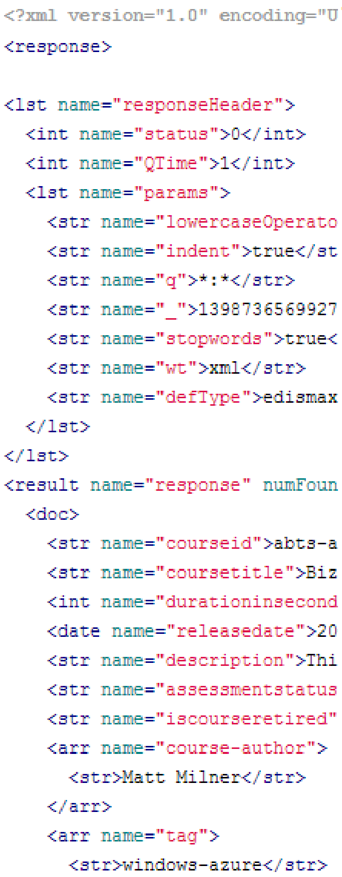
☐ facet

wt

- Response writer
- Xml, json, python, ruby, php, csv, …

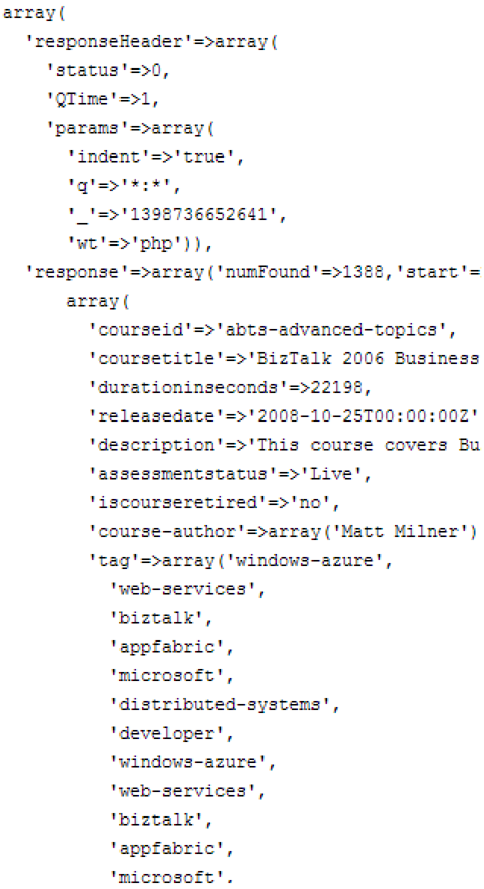


```
{
  "responseHeader": {
    "status": 0,
    "QTime": 10,
    "params": {
      "lowercaseOperators":
      "indent": "true",
      "q": "*:*",
      "_": "1398736490846",
      "stopwords": "true",
      "wt": "json",
      "defType": "edismax"
    }
  },
  "response": {
    "numFound": 1388,
    "start": 0,
    "docs": [
      {
        "courseid": "abts-a
        "coursetitle": "Biz
        "durationinseconds"
        "releasedate": "200
        "description": "Thi
        "assessmentstatus":
        "iscourseretired": '
        "course-author": [
          "Matt Milner"
        ],
```

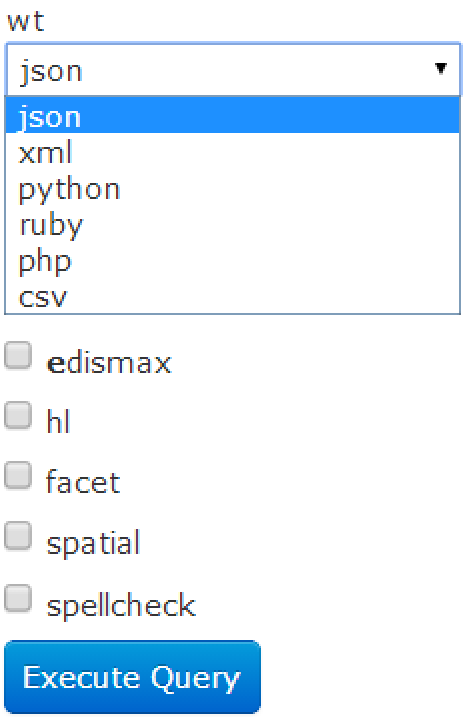


```
<?xml version="1.0" encoding="U
<response>

<lst name="responseHeader">
  <int name="status">0</int>
  <int name="QTime">1</int>
  <lst name="params">
    <str name="lowercaseOperato
    <str name="indent">true</st
    <str name="q">*:*</str>
    <str name="_">1398736569927
    <str name="stopwords">true<
    <str name="wt">xml</str>
    <str name="defType">edismax
  </lst>
</lst>
<result name="response" numFoun
  <doc>
    <str name="courseid">abts-a
    <str name="coursetitle">Biz
    <int name="durationinsecond
    <date name="releasedate">20
    <str name="description">Thi
    <str name="assessmentstatus
    <str name="iscourseretired"
    <arr name="course-author">
      <str>Matt Milner</str>
    </arr>
    <arr name="tag">
      <str>windows-azure</str>
```



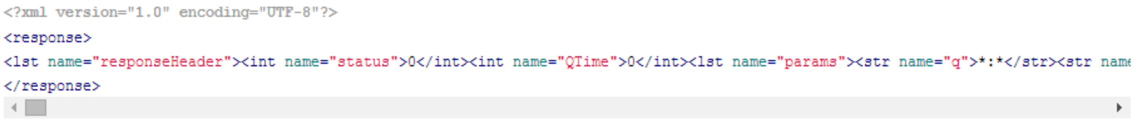
```
array(
  'responseHeader'=>array(
    'status'=>0,
    'QTime'=>1,
    'params'=>array(
      'indent'=>'true',
      'q'=>'*:*',
      '_'=>'1398736652641',
      'wt'=>'php')),
  'response'=>array('numFound'=>1388,'start'=
    array(
      'courseid'=>'abts-advanced-topics',
      'coursetitle'=>'BizTalk 2006 Business
      'durationinseconds'=>22198,
      'releasedate'=>'2008-10-25T00:00:00Z'
      'description'=>'This course covers Bu
      'assessmentstatus'=>'Live',
      'iscourseretired'=>'no',
      'course-author'=>array('Matt Milner')
      'tag'=>array('windows-azure',
        'web-services',
        'biztalk',
        'appfabric',
        'microsoft',
        'distributed-systems',
        'developer',
        'windows-azure',
        'web-services',
        'biztalk',
        'appfabric',
        'microsoft',
```



wt: json

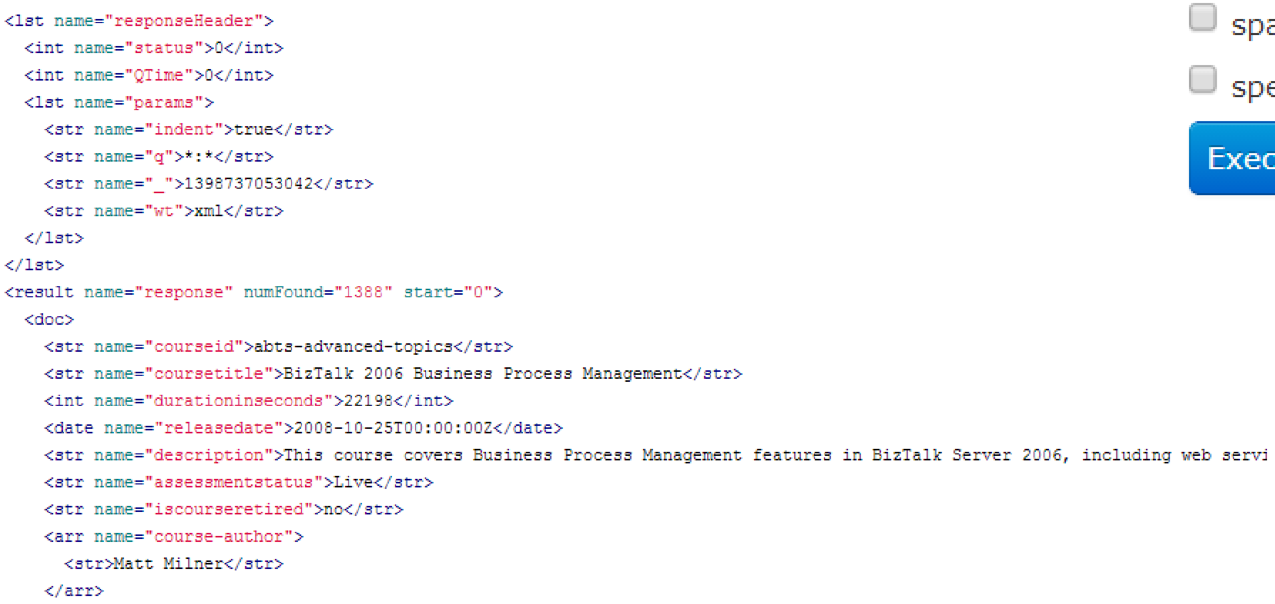
indent

- Request the wt to indent
- More readable for humans

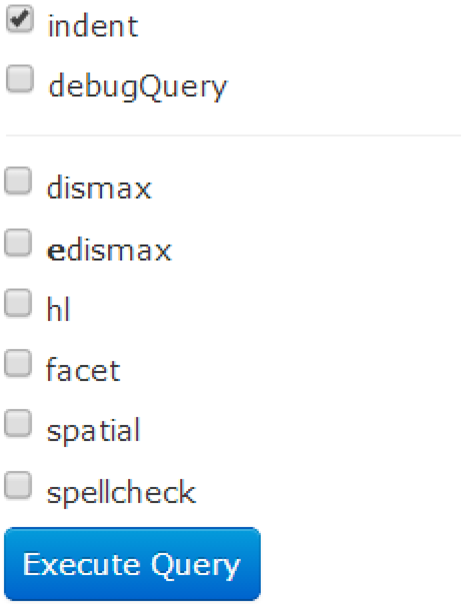


```
<?xml version="1.0" encoding="UTF-8"?>
<response>
<lst name="responseHeader"><int name="status">0</int><int name="QTime">0</int><lst name="params"><str name="q">*:*</str><str nam
</response>
```

VS.



```
<lst name="responseHeader">
  <int name="status">0</int>
  <int name="QTime">0</int>
  <lst name="params">
    <str name="indent">true</str>
    <str name="q">*:*</str>
    <str name="_">1398737053042</str>
    <str name="wt">xml</str>
  </lst>
</lst>
<result name="response" numFound="1388" start="0">
  <doc>
    <str name="courseid">abts-advanced-topics</str>
    <str name="coursetitle">BizTalk 2006 Business Process Management</str>
    <int name="durationinseconds">22198</int>
    <date name="releasedate">2008-10-25T00:00:00Z</date>
    <str name="description">This course covers Business Process Management features in BizTalk Server 2006, including web servi
    <str name="assessmentstatus">Live</str>
    <str name="iscourseretired">no</str>
    <arr name="course-author">
      <str>Matt Milner</str>
    </arr>
```

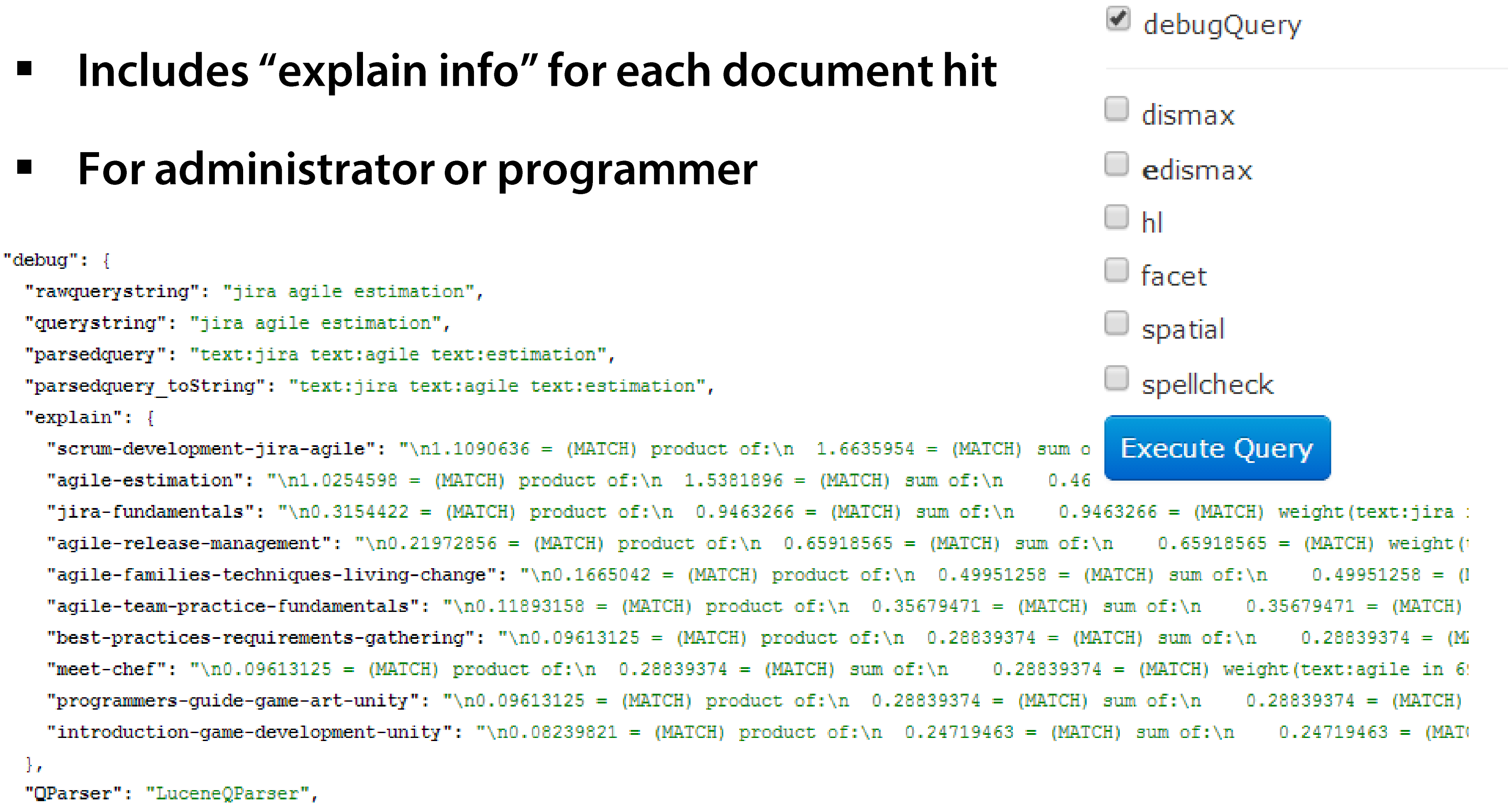


☑ indent
□ debugQuery

□ dismax
□ edismax
□ hl
□ facet
□ spatial
□ spellcheck

debugQuery

- Augment query response with debug info
- Includes “explain info” for each document hit
- For administrator or programmer



☑ debugQuery

□ dismax

□ edismax

□ hl

□ facet

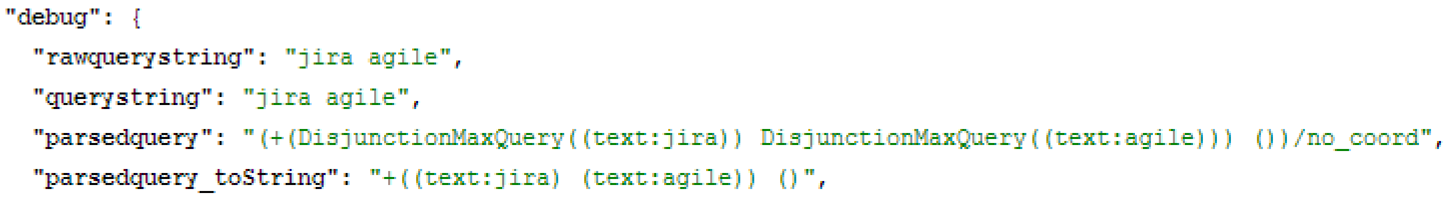
□ spatial

□ spellcheck

```
"debug": {
  "rawquerystring": "jira agile estimation",
  "querystring": "jira agile estimation",
  "parsedquery": "text:jira text:agile text:estimation",
  "parsedquery_toString": "text:jira text:agile text:estimation",
  "explain": {
    "scrum-development-jira-agile": "\n1.1090636 = (MATCH) product of:\n  1.6635954 = (MATCH) sum o
    "agile-estimation": "\n1.0254598 = (MATCH) product of:\n  1.5381896 = (MATCH) sum of:\n    0.46
    "jira-fundamentals": "\n0.3154422 = (MATCH) product of:\n  0.9463266 = (MATCH) sum of:\n    0.9463266 = (MATCH) weight(text:jira
    "agile-release-management": "\n0.21972856 = (MATCH) product of:\n  0.65918565 = (MATCH) sum of:\n    0.65918565 = (MATCH) weight(
    "agile-families-techniques-living-change": "\n0.1665042 = (MATCH) product of:\n  0.49951258 = (MATCH) sum of:\n    0.49951258 = (
    "agile-team-practice-fundamentals": "\n0.11893158 = (MATCH) product of:\n  0.35679471 = (MATCH) sum of:\n    0.35679471 = (MATCH)
    "best-practices-requirements-gathering": "\n0.09613125 = (MATCH) product of:\n  0.28839374 = (MATCH) sum of:\n    0.28839374 = (M
    "meet-chef": "\n0.09613125 = (MATCH) product of:\n  0.28839374 = (MATCH) sum of:\n    0.28839374 = (MATCH) weight(text:agile in 6
    "programmers-guide-game-art-unity": "\n0.09613125 = (MATCH) product of:\n  0.28839374 = (MATCH) sum of:\n    0.28839374 = (MATCH)
    "introduction-game-development-unity": "\n0.08239821 = (MATCH) product of:\n  0.24719463 = (MATCH) sum of:\n    0.24719463 = (MAT
  },
  "QParser": "LuceneQParser",
```

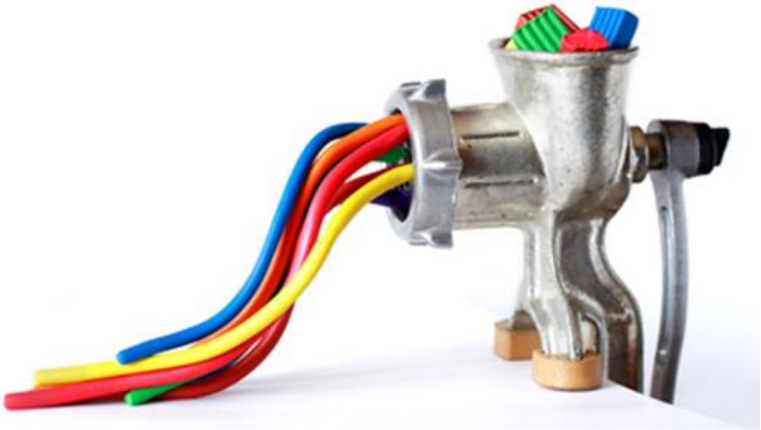
Query Parsers

- Component responsible for parsing the textual query and converting it to a Lucene Query Object



```
"debug": {
  "rawquerystring": "jira agile",
  "querystring": "jira agile",
  "parsedquery": "(+(DisjunctionMaxQuery((text:jira)) DisjunctionMaxQuery((text:agile))) ())/no_coord",
  "parsedquery_toString": "+((text:jira) (text:agile)) ()",
```

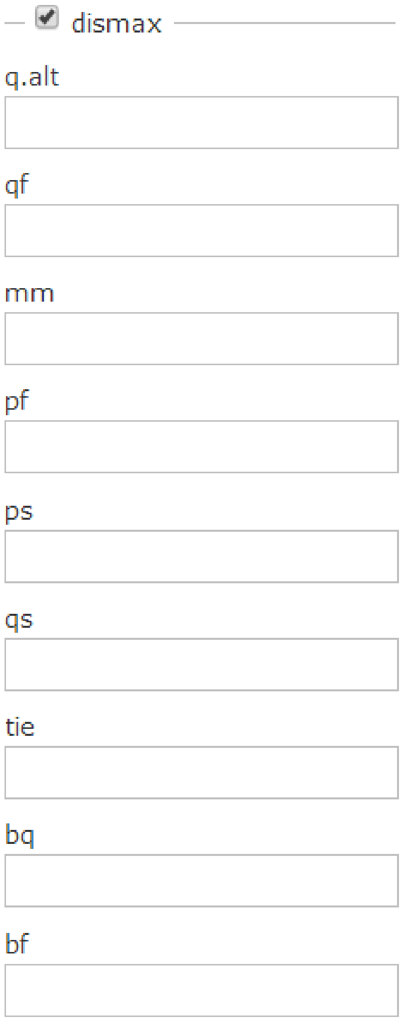
- Three main built in
  - Standard
  - DisMax
  - eDisMax



- Many more, some extremely specific

dismax

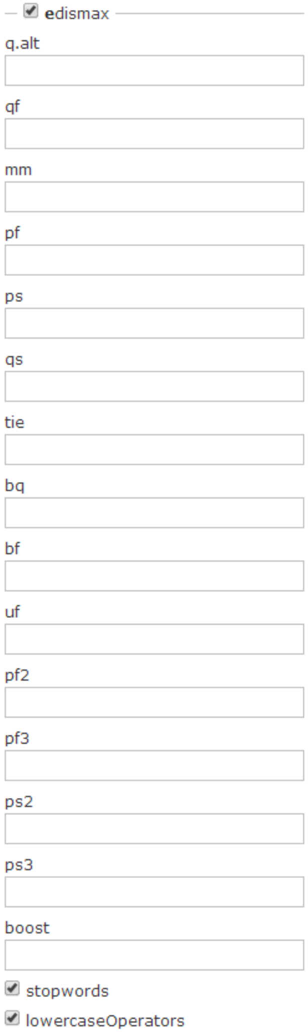
- **Maximum Disjunction**
- **Designed to process simple phrases**
- **More of a Google-like search experience**
- **Simpler, but with advanced searching capabilities**
  - Different fields
  - Different weights or boosts
- **Easy to use, accepting great deal of input and less strict. Returns few errors**



☑ dismax

edismax

- Extended Dismax
- Improved version of Dismax
- Full Lucene query syntax
- Respect magic fields names _val_ and _query_
  - Function queries or nested queries
- Improved boost function
- Improves proximity boosting by using shingles
- Supports pure negative queries



☑ edismax

q.alt

qf

mm

pf

ps

qs

tie

bq

bf

uf

pf2

pf3

ps2

ps3

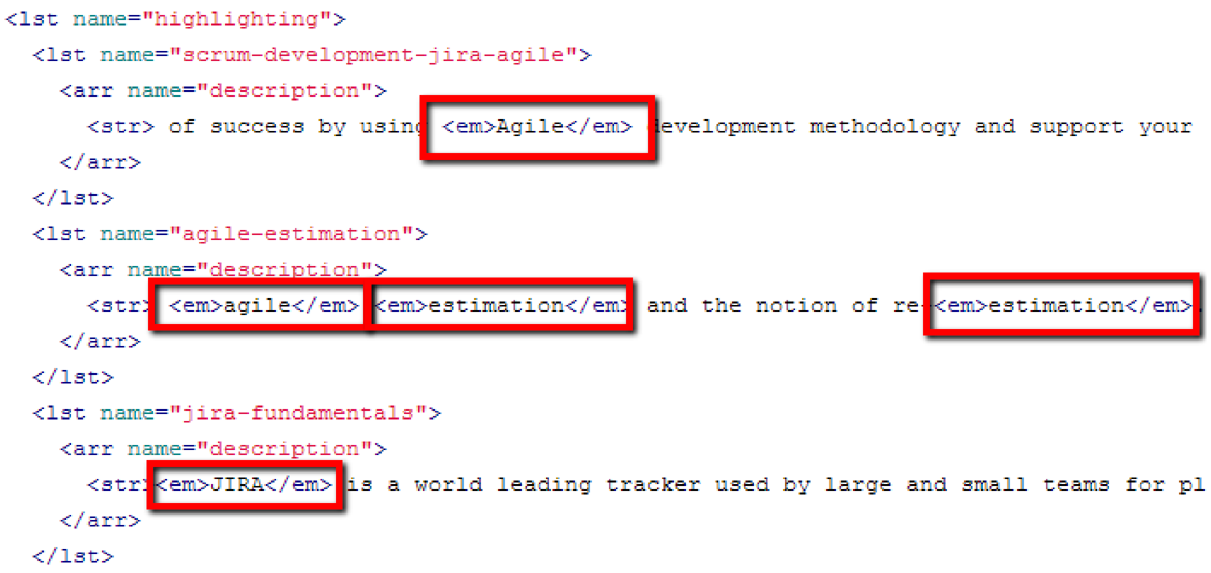
boost

☑ stopwords

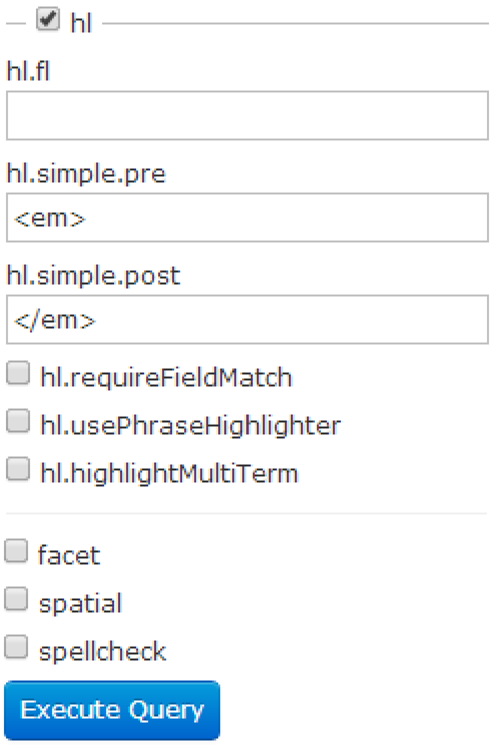
☑ lowercaseOperators

hl

- **Enable highlighting in query response**
- **Three implementations available**
  - Standard highlighter
  - FastVector highlighter
  - Postings Highlighter



```
<lst name="highlighting">
  <lst name="scrum-development-jira-agile">
    <arr name="description">
      <str> of success by using <em>Agile</em> development methodology and support your
    </arr>
  </lst>
  <lst name="agile-estimation">
    <arr name="description">
      <str> <em>agile</em> <em>estimation</em> and the notion of re-<em>estimation</em>
    </arr>
  </lst>
  <lst name="jira-fundamentals">
    <arr name="description">
      <str><em>JIRA</em> is a world leading tracker used by large and small teams for pl
    </arr>
  </lst>
```



☑ hl

hl.fl

hl.simple.pre: <em>

hl.simple.post: </em>

☐ hl.requireFieldMatch

☐ hl.usePhraseHighlighter

☐ hl.highlightMultiTerm

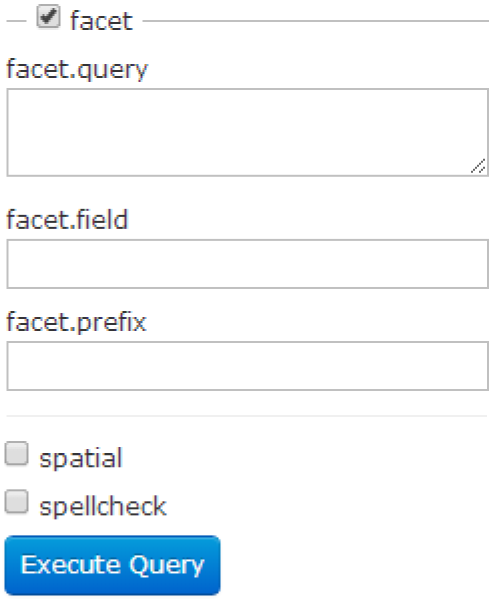
☐ facet

☐ spatial

☐ spellcheck

facet

- **Arrangement of search results into categories**
  - Based on indexed terms
  - Include numerical counts
- **Allow users to drill down and narrow results**
- **facet : true enables faceting**
  - facet.query: Lucene query to generate facet count
  - facet.field: field to be treated as facet
  - facet.prefix: only terms that begin with this prefix
  - Many other options



☑ facet

facet.query

facet.field

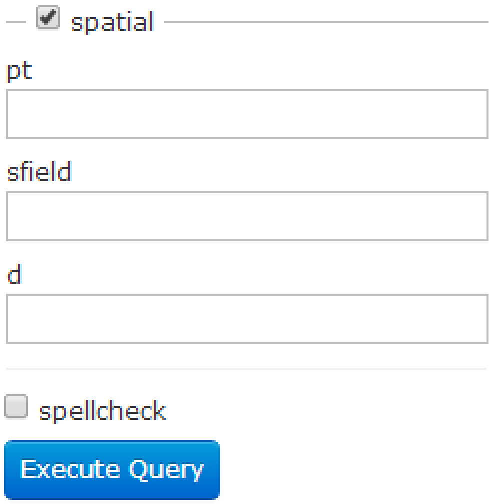
facet.prefix

□ spatial

□ spellcheck

spatial

- **Location search**
  - Called spatial or geo-spatial search
  - Units in km, points of latitude/longitude
- **Sort or score/boost by distance**
- **Also bound by shape**



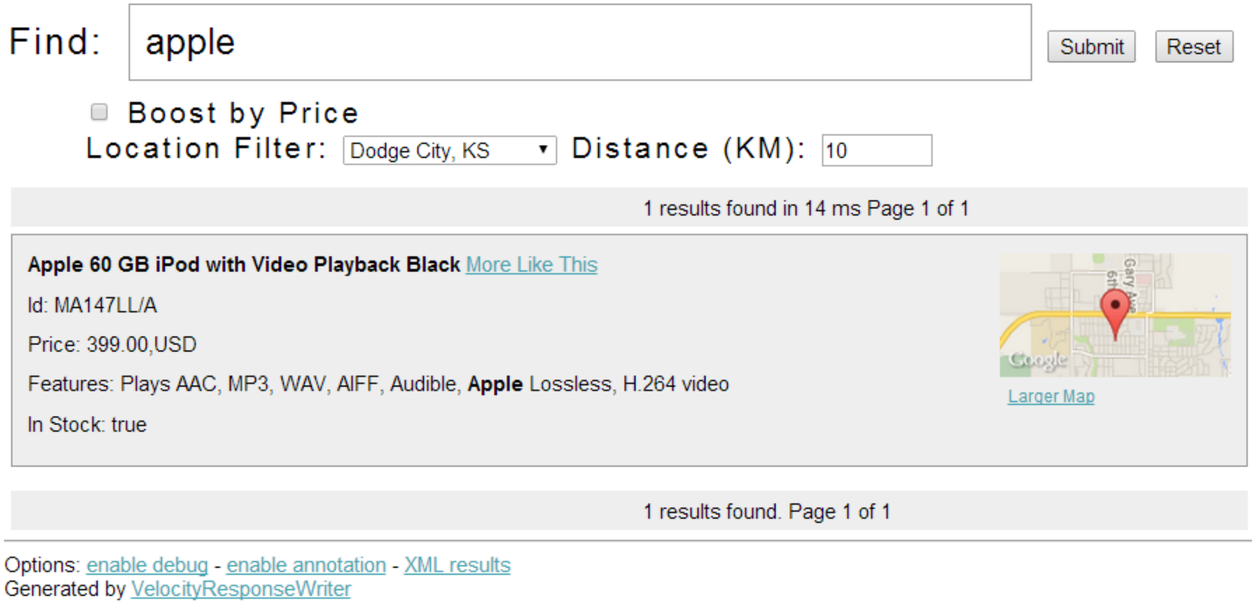
☑ spatial

pt

sfield

d

□ spellcheck



Find: apple

□ Boost by Price

Location Filter: Dodge City, KS Distance (KM): 10

1 results found in 14 ms Page 1 of 1

Apple 60 GB iPod with Video Playback Black More Like This

Id: MA147LL/A

Price: 399.00,USD

Features: Plays AAC, MP3, WAV, AIFF, Audible, **Apple** Lossless, H.264 video

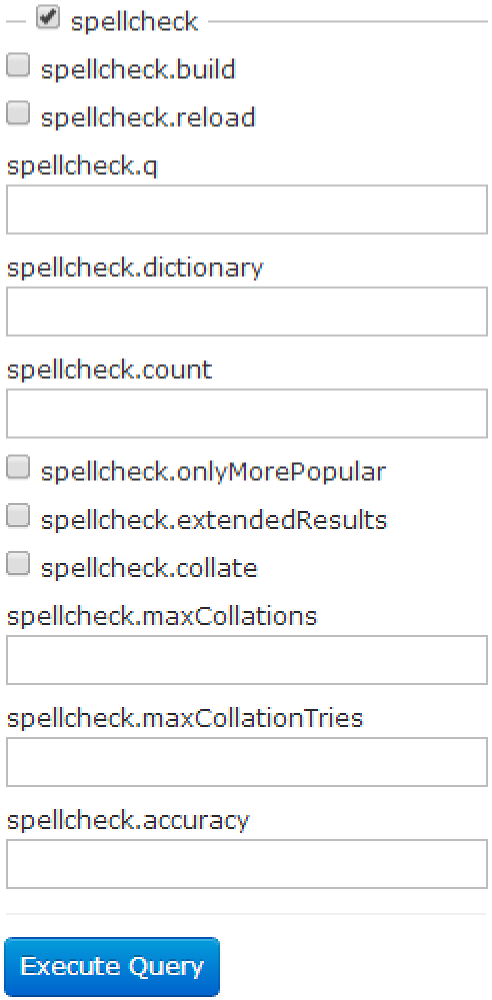
In Stock: true

1 results found. Page 1 of 1

Options: enable debug - enable annotation - XML results
Generated by VelocityResponseWriter

spellcheck

- **Provide inline query suggestions based on other similar terms**
- **Basis can be:**
  - Terms in a field in Solr
  - External text files
  - Fields in other Lucene indexes
- **Collation, max tries, …**
- **Let me show in an application**



☑ spellcheck

☐ spellcheck.build

☐ spellcheck.reload

spellcheck.q

spellcheck.dictionary

spellcheck.count

☐ spellcheck.onlyMorePopular

☐ spellcheck.extendedResults

☐ spellcheck.collate

spellcheck.maxCollations

spellcheck.maxCollationTries

spellcheck.accuracy

Synonyms

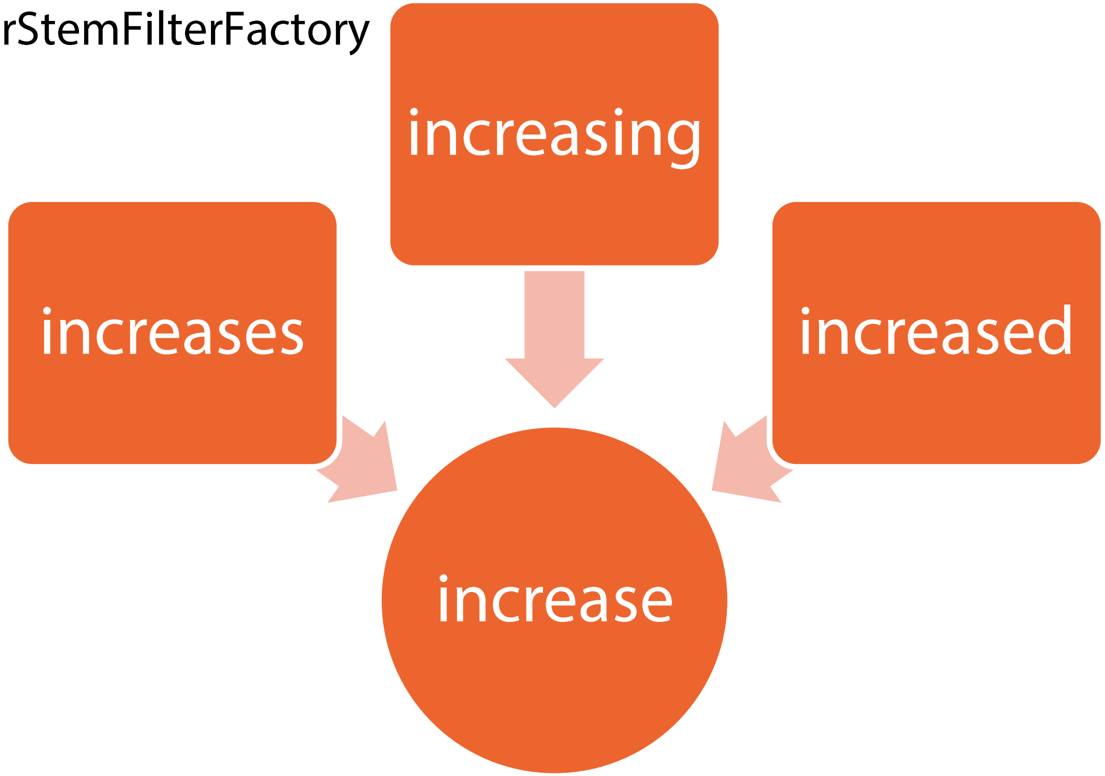
- **Word or phrases that mean the same**
- **Match strings of tokens and replace with other strings of tokens**
- **Help increase recall**
- **Example: Toyota Echo & Toyota Yaris**
  - Same car! Different name!

- **Query time vs. Index time**
  - Query time means longer execution times
  - Index time means bigger index size

- **Configure via SynonymFilterFactory**
- **Synonym Dictionary**

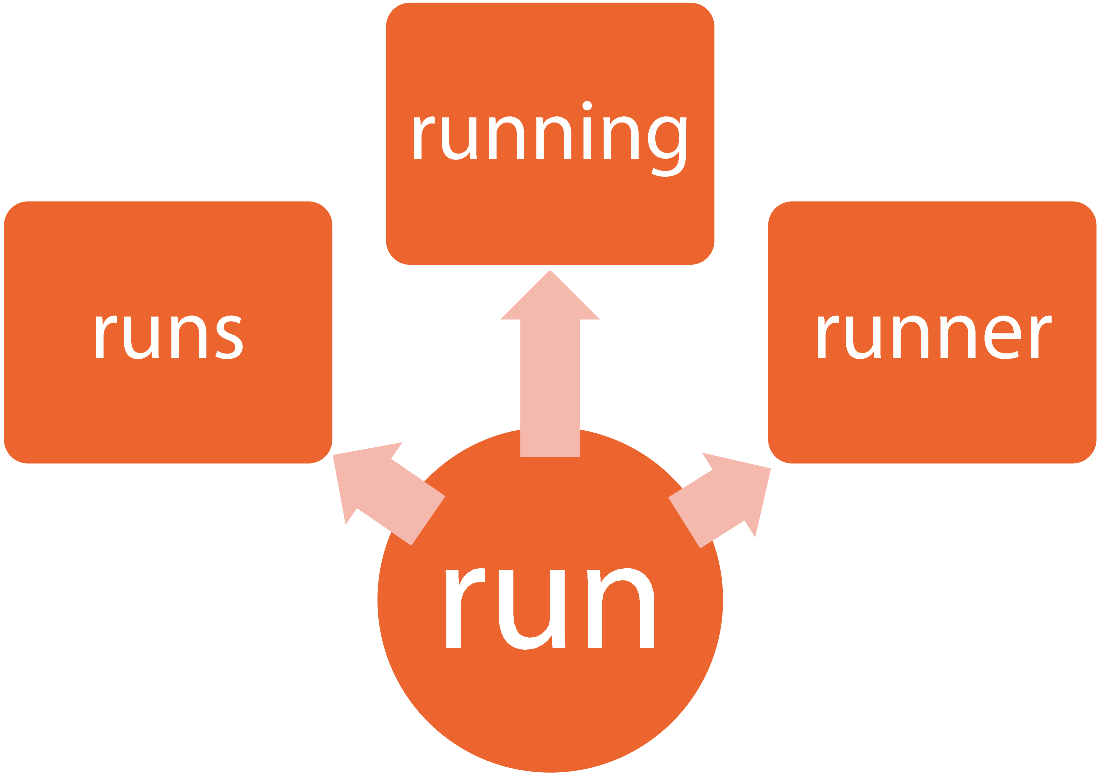
Stemming

- Reducing a word to a shorter base form
- Stemming helps increase recall, but makes your index much bigger
- Via Analyzer, some more aggressive than others:
  - i.e. solr.PorterStemFilterFactory



Lemmatisation

- Expanding a root word to all its various forms
- Use dictionary + synonym filter factory


running

runs

runner

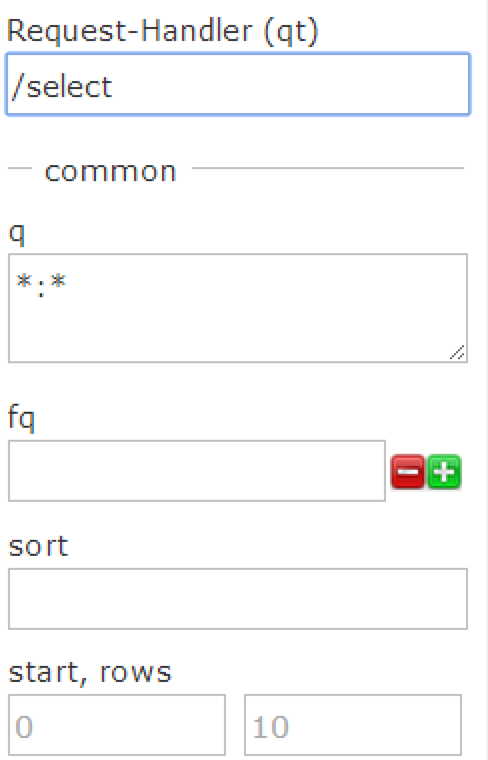
run

Stop Words

- Discards common words
- Standard English stop words included in the list
  - A, an, and, are, as, at, …
- Specify in a file → stopwords.txt
- Ignore case true | false
- Query and Index time
- solr.StopFilterFactory

Request-Handler (qt)

- **Defines logic executed for any request**
  - Filters or facets
  - Append/Invariant
- **Multiple can be specified in same Solrconfig**
- **Named request handlers for cores**
  - …/solr/**psdemo**/select?q=…
- **Many available: DIH, CSV, Spellcheck, Update, …**
- **Create request handlers to specify configurations**
  - But don’t abuse!



Request-Handler (qt)
/select

common

q
:

fq

sort

start, rows
0
10

Takeaway

- Searching in Solr is extremely complex
- Endless options, possibilities and parameters
- But low bar for getting started with simple applications with minimal configuration

Next module:

- Putting a UI together in a few minutes!