DQ Prototype & SolrJ Utils

Mark Bennett / Feb 2014

Contents

- DQ Prototype Goals
- Secondary Goals
- Sample Reports
- SolrJ Lessons Learned
- Tech Notes
- The Future ...
- Wrap-up / Links

DQ Prototype Goals

- What **General** things can we do for you ?
  - Understand your data, Overall and **Outliers**
  - Compare 2 Solr systems! Eg: Dev vs. Staging
- NOT in Scope: Full Data Quality
  - Domain Specific Rules
  - Judgement: is this data **good** or **bad**?
  - Fixing Data

Secondary Goals / Side Effects

- Scalability not primary, but ...
  - Find out what "breaks" at Million+ mark
- Python vs. Java -> Java
  - More STABLE
  - faster
  - Better Collections (yes, Java!)
- Exercise SolrJ

Sample Reports

DQ: Single Core

- **Populated Fields**: Fully populated, partially, empty
  - Indexed terms (fast, default) and Stored Values (slow)
  - Show IDs of missing docs (suggested by Fidelity)
- Term **Length** Stats
  - show terms **+/- 3 Standard Deviations**
- **Corrupted** / Mis-Encoded data
  - Theory: random bytes span more Unicode buckets!
- **Dates** Analysis
  - Automatically spot date fields
  - Histogram of Dates
  - Curve-fitting: Your data vs. ideal Exponential Growth

DQ: Between Cores

- Doc IDs only in Core A / only in Core B
- Populated Fields / stats between cores
- Schema Differences:
  - Between running cores or against schema.xml
  - Or running core and default Solr 4.6.1
- LLR / G2 stats: most significant search term diffs

Populated Fields

Total Active Docs: 1,275,077

All Fields: [_root_, _version_, accessories, albumLabel, albumTitle, ... sku, ... url, weight]

Populated at 100%: [_version_, id, regularPrice, salePrice, store_id, text, type]

No Indexed Values / 0% [_root_, author, cat, category, categoryPath, comments, content, content_type, inStock, includes, keywords, last_modified, links, manu, manu_exact, payloads, popularity, price, resourcename, shippingWeight, sku, store, subject, text_rev, title, url]

Partially Populated Fields / Percentages:

- accessories: 11,460 (0.9%)
- albumLabel: 876,821 (68.77%)
- albumTitle: 876,845 (68.77%)
- artistName: 871,477 (68.35%)
- ...
- mpaaRating: 123,899 (9.72%)
- name: 1,274,453 (99.95%)
- ...
- startDate: 1,273,615 (99.89%)
- studio: 256,401 (20.11%)
- subclass: 1,258,757 (98.72%)
- weight: 67,072 (5.26%)

Term Lengths

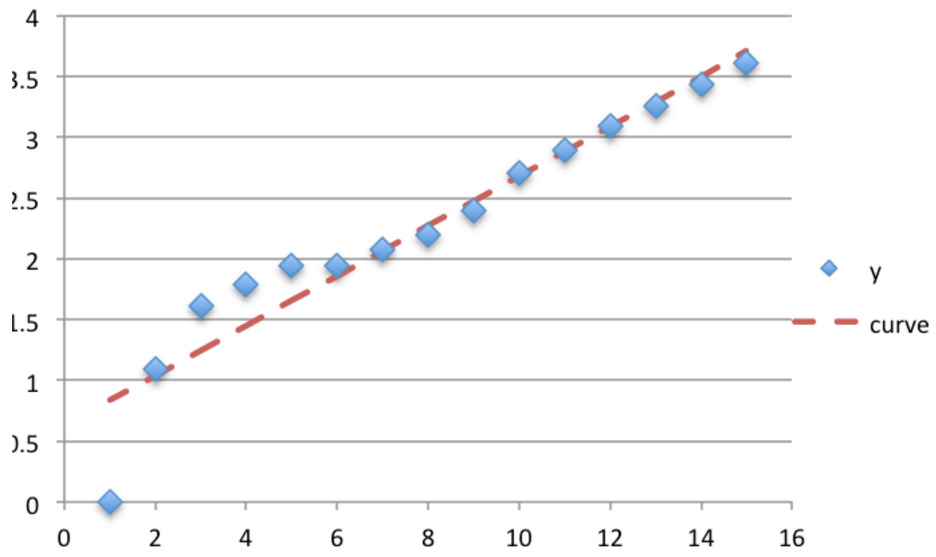
```
Unique Term Length Stats by Field, min/max/avg/std (terms include deleted docs):
    _version_: 19 / 19 / 19.0 / 0.0   (1,476,194 entries)
    ...
    albumLabel: 1 / 35 / 14.1 / 5.89   (42,158 entries)
        Expected Length Range, raw: -4 to 32 (inclusive)
        Expected Length Range, clamped: 1 to 32 (inclusive)
        Unusually Long Terms:
            67: BCI Music (Brentwood Communication), len=35
            87: Warner Elektra Atlantic Corp. (Japa, len=35
            117: Columbia River Entertainment Group, len=34
            122: Warner Bros. Records (Record Label), len=35
            267: Musical Productions Inc./MP Online, len=34
    ...
    text: 1 / 46 / 7.63 / 1.38   (1,479,550 entries)
        Expected Length Range, raw: 3 to 12 (inclusive)
        Expected Length Range, clamped: 3 to 12 (inclusive)
        Unusually Short Terms:
            2: cd, len=2
            5: of, len=2
            8: in, len=2
            10: a, len=1
            11: to, len=2
            ...
        Unusually Long Terms:
            389: automatically, len=13
            595: multimediacard, len=14
            598: compatibility, len=13
            867: entertainment, len=13
            1,165: environmental, len=13
```

Unicode Buckets

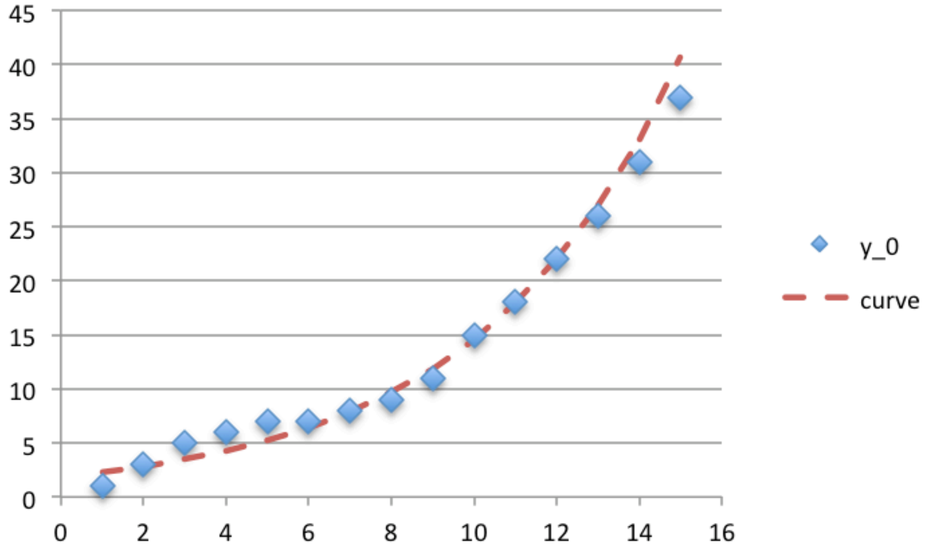
```
Field: color
    Character Classes: [Com-L1Sup-OtherP, UPPER, lower]
        Satina¿
    Character Classes: [Com-L1Sup-OtherP, UPPER, lower, space]
        Titanium¿ color
    Character Classes: [Dash1, Digit, UPPER, lower, space]
        3-Tone Sunburst
        2-Tone Sunburst
    Character Classes: [Dash1, OtherPunct, UPPER, lower]
        Black/Stainless-Steel
        Stainless-Steel/Black
        Black/Tri-Color
    Character Classes: [Dash1, Start, Stop, UPPER, lower, space]
        Stainless-Steel (Special Order)
        White-on-White (Special Order)
    Character Classes: [Dash1, lower, space]
        dark-slate-gray leather
        medium-dark-pewter cloth
    Character Classes: [Digit, OtherPunct, UPPER, lower]
        Cr&#232;me
    Character Classes: [Digit, OtherPunct, UPPER, lower, space]
        Titanium&#153; color
        Cr&#232;me with black trim
    Character Classes: [Digit, OtherPunct, lower, space]
        4 oz.
    Character Classes: [Digit, UPPER]
        3TS
        2TS
```

Curve Fitting

Fitted curve: y = mx + b



Fitted curve: y = A e^kx



Checking Dates / Exp Curve Fit

```
Date Field: startDate
Start/Stop: 1884-01-01T00:00:00Z / 2012-07-29T00:00:00Z
    2012-01-01: ==========#===
    2011-01-01: =========#==========================================
    2010-01-01: ========#=======================================
    2009-01-01: =======#================================
    2008-01-01: ======#============================================
    2007-01-01: =====#=================================================
    2006-01-01: ====#============================================
    2005-01-01: ====#=============================================
    2004-01-01: ===#================================================
    2003-01-01: ===#============================================
    2002-01-01: ==#=================================
    2001-01-01: ==#================================
    2000-01-01: =#=====================================
    1999-01-01: =#=====================================
    1998-01-01: =#===========================
    1997-01-01: =#======================
    1996-01-01: #==================
    1995-01-01: #=======================
    1994-01-01: #====================
    1993-01-01: #================
    1992-01-01: #==========
    1991-01-01: #========
    1990-01-01: #==============
    1989-01-01: #==
    1988-01-01: #
```

Diff: Schemas

```
Schema A = Default Solr 4.6.1 Schema
Schema B = Apollo demo plus local changes
Key Field: Both = 'id'
...

Fields:

    In both = '[_version_, _root_, id, sku, name, ... text, text_rev, manu_exact, payloads]'

    B only = '[accessories, albumLabel, albumTitle, ... categoryNames, categoryPath, ... department, depthCategoryIds, depthCategoryNames, ...
genre, ... mpaaRating, plot ... releaseDate, salePrice ... shippingWeight, shortDescription, softwareGrade, startDate, store_id, studio,
subclass, type]'

...
Types:
    In both = '[string, boolean, int, float, long, double, tint, tfloat, tlong, tdouble, date, tdate, binary, pint, plong, pfloat, pdouble,
pdate, random, text_ws, text_general, text_en, text_en_splitting, text_en_splitting_tight, text_general_rev, alphaOnlySort, phonetic,
payloads, lowercase, descendent_path, ancestor_path, ignored, point, location, location_rpt, currency, text_ar, text_bg, text_ca, text_cjk,
text_cz, text_da, text_de, text_el, text_es, text_eu, text_fa, text_fi, text_fr, text_ga, text_gl, text_hi, text_hu, text_hy, text_id,
text_it, text_ja, text_lv, text_nl, text_no, text_pt, text_ro, text_ru, text_sv, text_th, text_tr]'

...

Copy Field Sources:
    In both = '[cat, name, manu, features, includes, price, title, author, description, keywords, content, content_type, resourcename, url]'
    B only = '[id]'

Copy Field Destinations:
    In both = '[text, manu_exact, price_c, author_s]'
```

Diff: LLR of Indexed Terms

```
---------- A -> B ----------

Corpus A unique / total words: 398 / 579.0
Corpus B unique / total words: 385 / 593.0
Combined unique / total words: 418 / 1172.0
Number of p log(p) calculations: 0

Term Changes, first 5 entries:
    acme: -4.09515240975383
    any: -4.09515240975383
    box: -4.09515240975383
    cardboard: -4.09515240975383
    fits: -4.09515240975383
Term Changes, last 5 entries:
    silentseek: 1.4112036109151607
    sp2514n: 1.4112036109151607
    spinpoint: 1.4112036109151607
    ultra: 1.4112036109151607
    cache: 2.824159489031562
    hard: 2.824159489031562
```

new.xml

```
<add><doc>
  <field name="id">NEW111</field>
  <field name="name">New Sample Product</field>
  <field name="manu">Acme, Inc.</field>
  <!-- Join -->
  <field name="manu_id_s">acme</field>
  <field name="cat">electronics</field>
  <field name="cat">gadget</field>
  <field name="features">Rocket powered, sugar-free, fits in any tackle box!</field>
  <field name="includes">cardboard box</field>
  <field name="weight">10.5</field>
  <field name="price">19.95</field>
  <field name="popularity">101</field>
  <field name="inStock">true</field>
  <!-- Buffalo store -->
  <field name="store">43.17614,-90.57341</field>
</doc></add>
```

Tech: SolrJ, Utils, notes...

SolrJ Utils

- Request Handlers
  - query.setRequestHandler("**/terms**");
  - query.setRequestHandler("**/admin/luke**");
  - query.setRequestHandler("**/schema/...**");
- Response Logic & Data Types
  - QueryResponse res = server.query(query);
  - **NamedList<Object>** res2 = res.getResponse();
  - **SimpleOrderedMap** res3 = (**SimpleOrderedMap**) res2.get("...");
  - ... **NamedList** res4 = (**NamedList**) res3.get(fieldName); ... etc etc ....
- **Sometimes need to re-parse data from Strings**

Other Utils

- DateUtils - to / from strings in various formats, TIMEZONES
- SetUtils
  - inAOnly, inBOnly, union, intersection
  - Stable Maps: head, tail, reverse, sortMapByValues
- StatsUtils
  - Lists: sum, min, max, average, standardDeviation
  - leastSquares_Line, leastSquares_Exponential
- LLR

Tech Notes

- Maven project, builds unified jar
  - use with java -jar (or script)
- Command Line:
  - -h/--host, [-p/--port, -c/--collection]
  - Or -u/--url, diff uses --url_a/--url_b, etc.
  - -f/--field: specific target fields to analyze
  - Run w/ no args to see other options

Current Status / Plan / Wrap-up

The Future ...

- Alpha and internal use
- Integrate into Apollo / LW 5 ?
  - **Other** "metrics" and reporting efforts already underway
  - Coordinate, integration design pending...
- Broader DQ Scope if there's interest:
  - **Search Terms** vs. Document Terms
  - Rules Engine ?
  - Integration into Indexing Pipeline ?
- Blog about SolrJ wrapper examples

Links

- Code and SAMPLE REPORTS:
  - https://github.com/LucidWorks/data-quality
  - https://github.com/LucidWorks/data-quality/tree/master/src/main/resources/sample-reports
- Curve Fitting: Linear and Exponential
  - http://hotmath.com/hotmath_help/topics/line-of-best-fit.html
  - http://math.stackexchange.com/questions/350754/fitting-exponential-curve-to-data
- G2 / LLR: Log Likelihood Ratio (Warning: Dunning post leaves "incomplete")
  - http://tdunning.blogspot.com/2008/03/surprise-and-coincidence.html
  - http://scg.unibe.ch/archive/papers/Kuhn09aLogLikelihoodRatio.pdf