Indexing and Searching Chinese, Japanese, and Korean text in Solr

BROWN

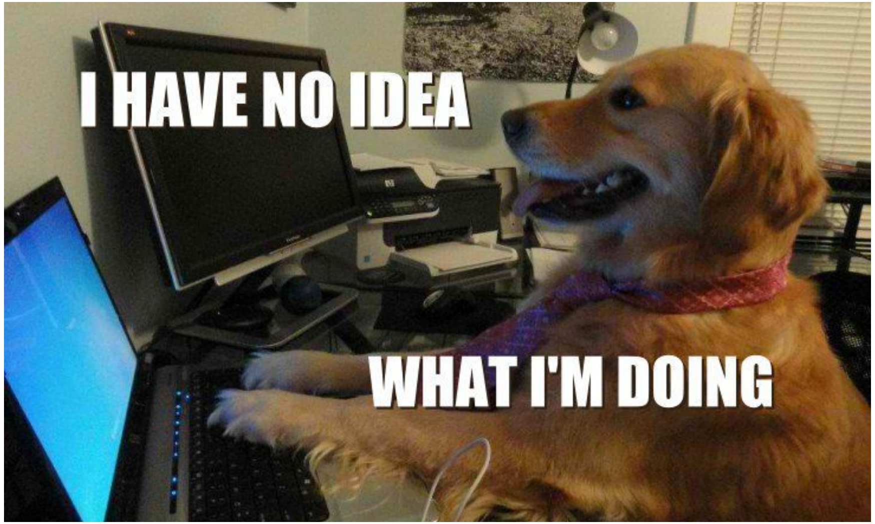
Hector Correa
hector_correa@brown.edu

September 30, 2020

Disclaimer

I do not speak Chinese, Japanese, or Korean…

...this session is about how to index CJK text in Solr from the perspective of a software developer that knows a little bit of Solr but nothing about CJK languages.



Agenda

- The problem we were facing
- Indexing CJK text in Solr
- Searching for CJK text
- Questions and Answers

The problem: precision

Precision

Search	Expected Matches	Total Returned	Precision
莫言 (Mo Yan)	72	300	0.240
柳美里 (Yu Miri)	15	1000	0.015
ふくろうの本 (Fukuro no hon)	11	4800	0.002

- Precision = Number correct matches / Total results returned
- "When I issue a search, are the documents that come back the ones I was looking for?"
- Values close to zero are bad

The root of the problem

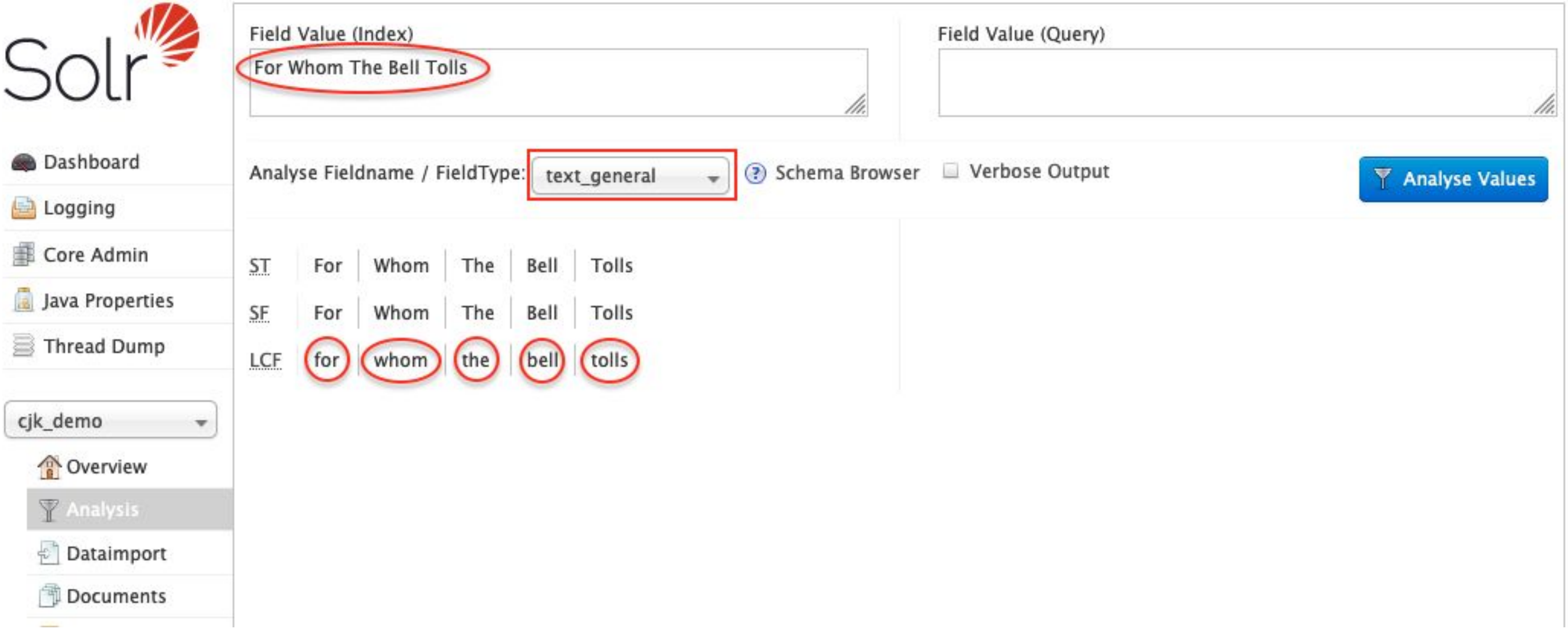
- Solr has many text fields
  - General: **text_general**
  - Language specific: text_ar, text_cjk, **text_en**, text_es, text_fr, …

- **text_general** works OK-ish for several* languages

- But **text_general** field does not work for CJK languages

* some exceptions apply

text_general field with text in English



Field Value (Index)

For Whom The Bell Tolls

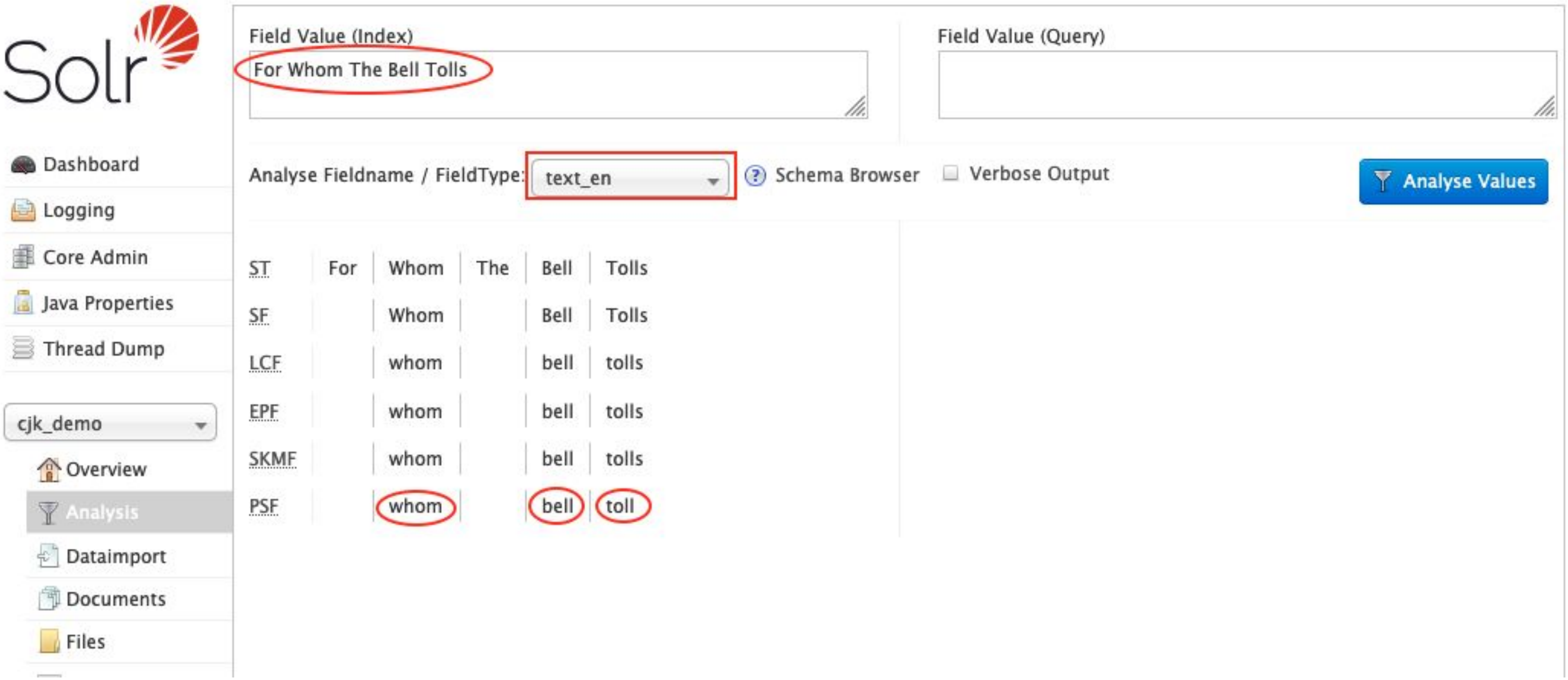
Field Value (Query)

Analyse Fieldname / FieldType: text_general

□ Verbose Output

ST	For	Whom	The	Bell	Tolls
SF	For	Whom	The	Bell	Tolls
LCF	for	whom	the	bell	tolls

text_en field with text in English



Field Value (Index): For Whom The Bell Tolls

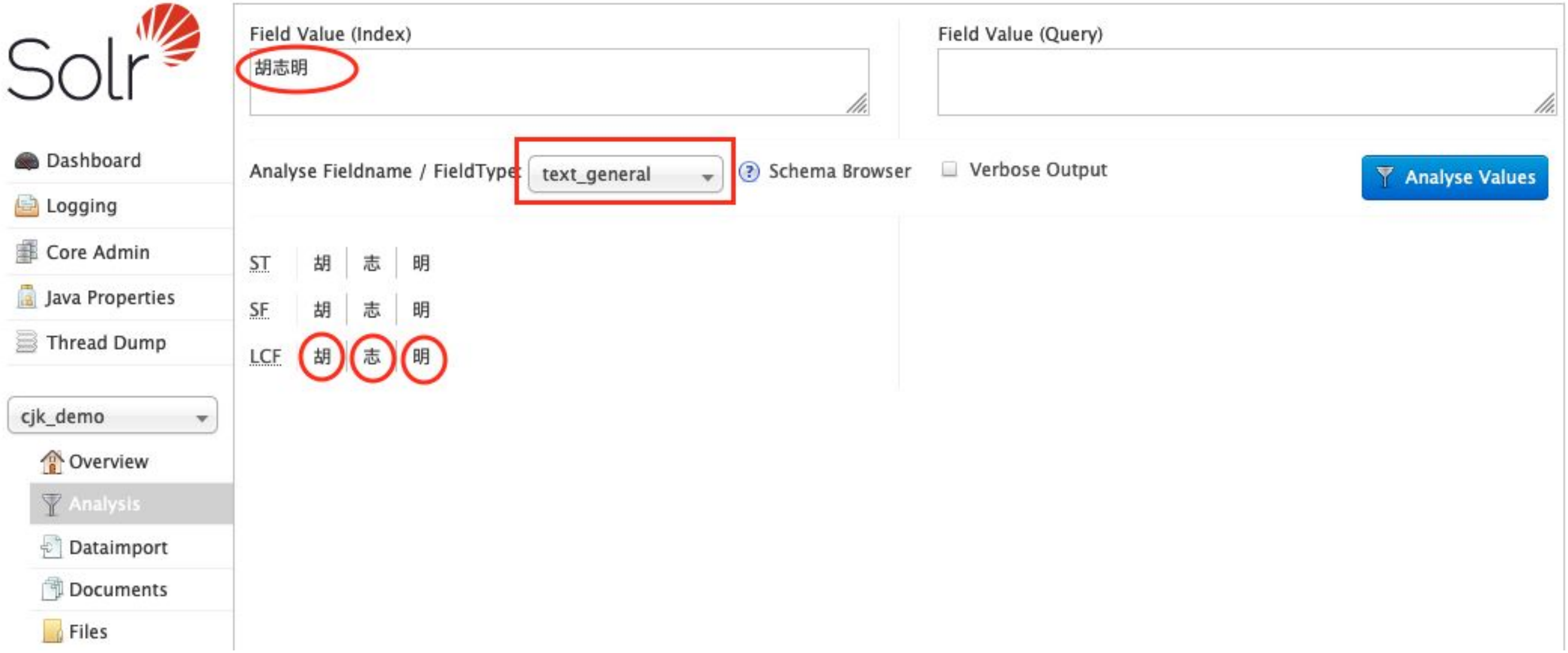
Field Value (Query):

Analyse Fieldname / FieldType: text_en

□ Verbose Output

ST	For	Whom	The	Bell	Tolls
SF		Whom		Bell	Tolls
LCF		whom		bell	tolls
EPF		whom		bell	tolls
SKMF		whom		bell	tolls
PSF		whom		bell	toll

text_general field with text in Chinese
胡志明 (Hồ Chí Minh)



Field Value (Index): 胡志明

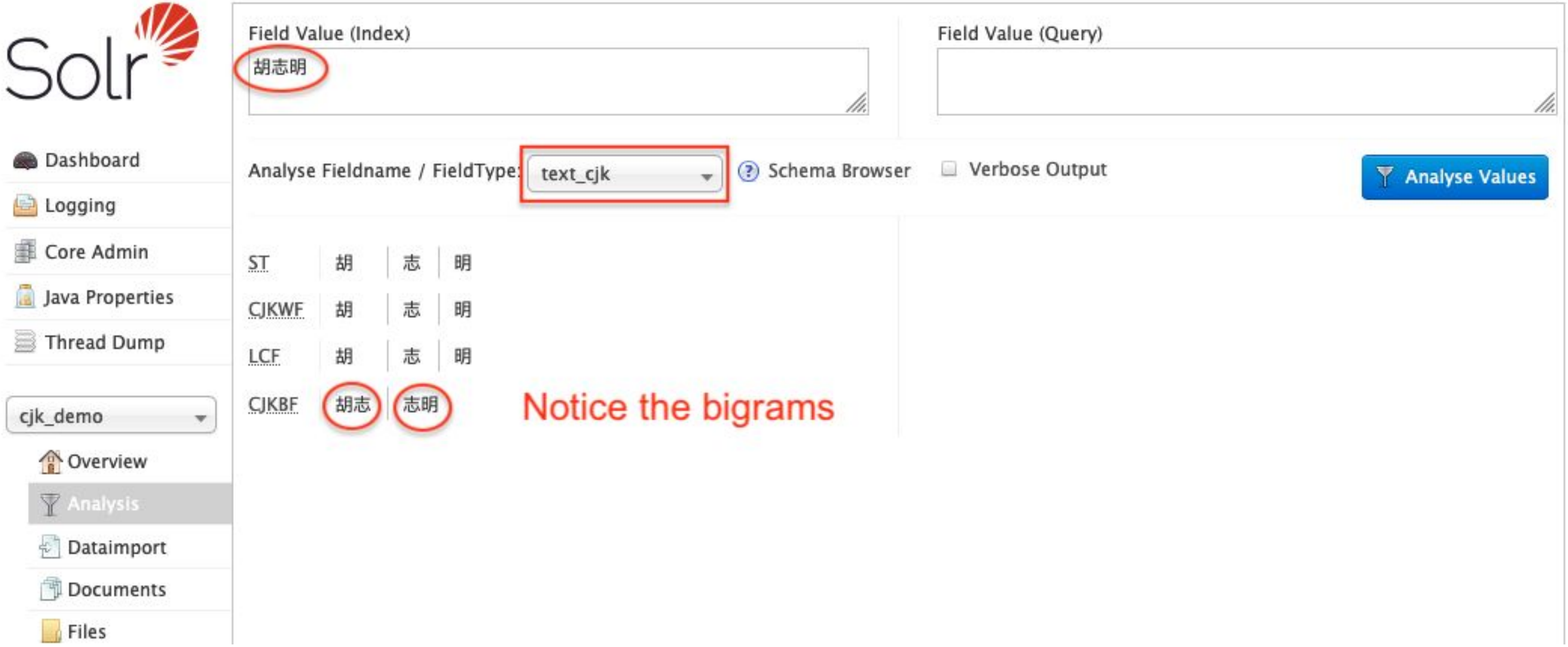
Field Value (Query):

Analyse Fieldname / FieldType: text_general

□ Verbose Output

ST	胡	志	明
SF	胡	志	明
LCF	胡	志	明

text_cjk field with text in Chinese 胡志明 (Hồ Chí Minh)



Field Value (Index)

胡志明

Field Value (Query)

Analyse Fieldname / FieldType: text_cjk

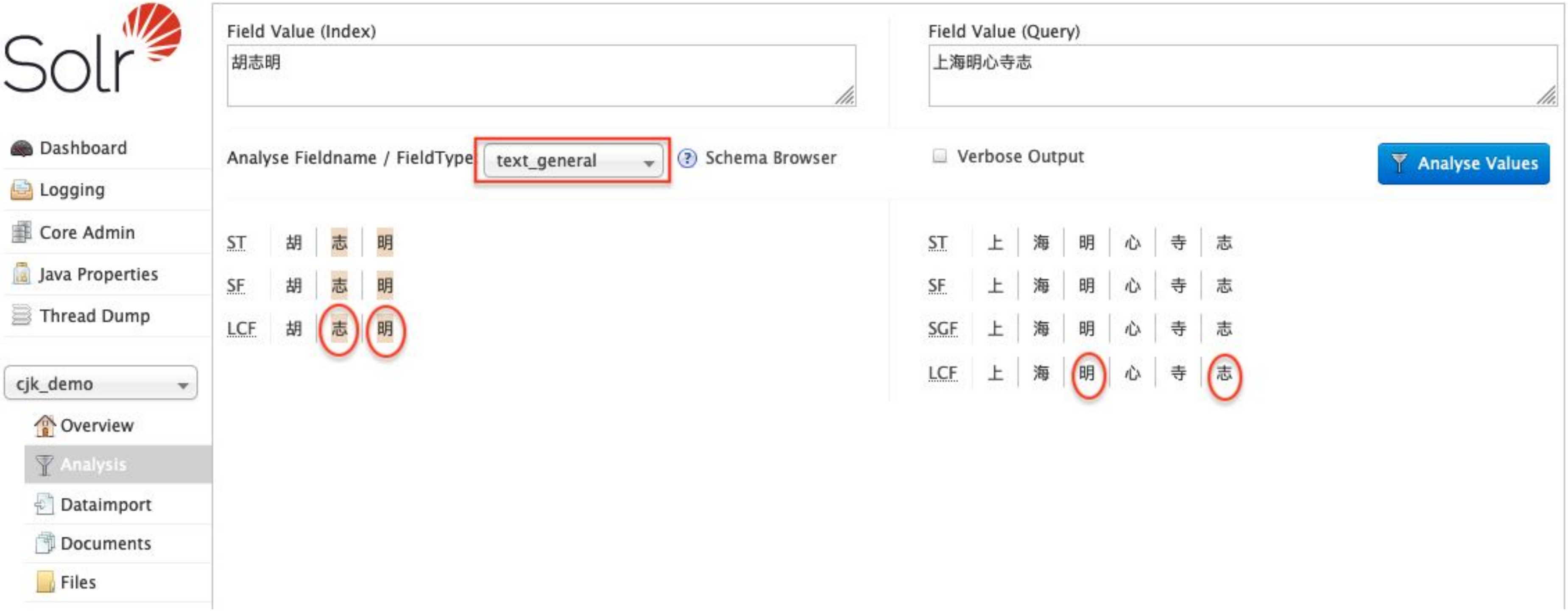
□ Verbose Output

ST	胡	志	明
CJKWF	胡	志	明
LCF	胡	志	明
CJKBF	胡志	志明	

Notice the bigrams

Nonsensical match with text_general

胡志明 (Hồ Chí Minh) vs 上海明心寺志 (Shanghai Ming xin si zhi)



Field Value (Index)

胡志明

Field Value (Query)

上海明心寺志

Analyse Fieldname / FieldType: text_general

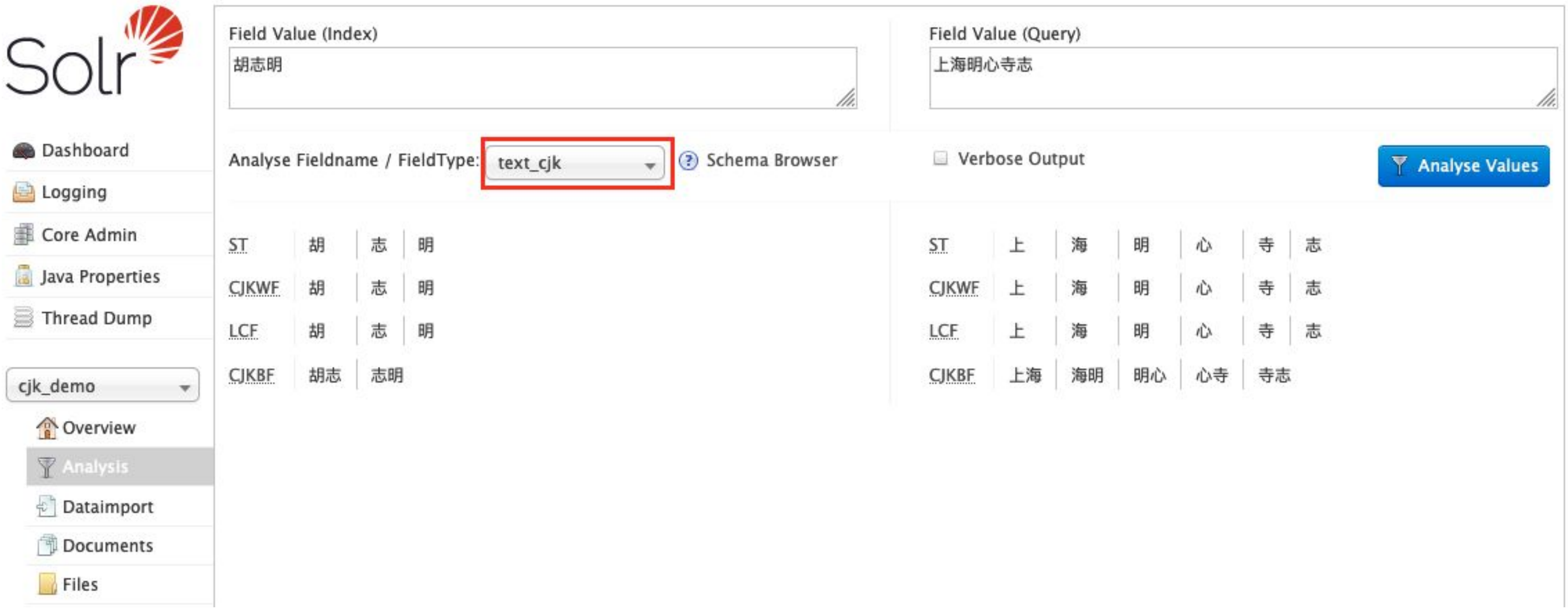
□ Verbose Output

ST	胡	志	明
SF	胡	志	明
LCF	胡	志	明

ST	上	海	明	心	寺	志
SF	上	海	明	心	寺	志
SGF	上	海	明	心	寺	志
LCF	上	海	明	心	寺	志

Nonsensical match avoided with text_cjk

胡志明 (Hồ Chí Minh) vs 上海明心寺志 (Shanghai Ming xin si zhi)



Field Value (Index): 胡志明

Field Value (Query): 上海明心寺志

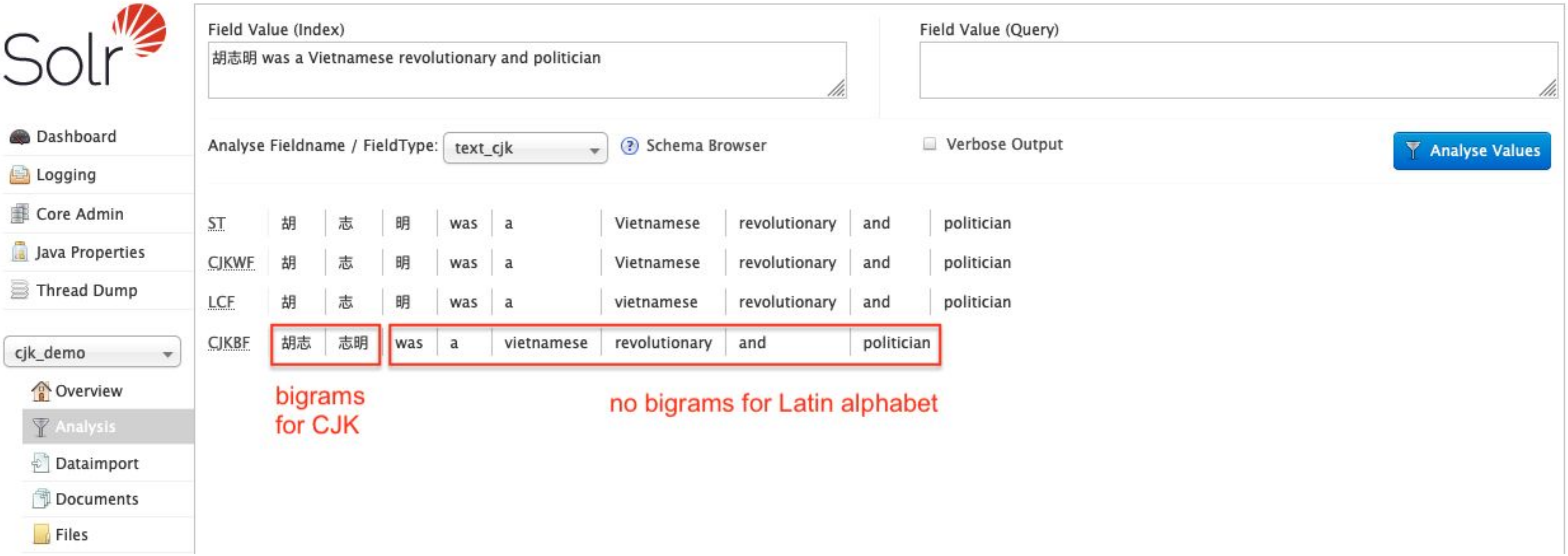
Analyse Fieldname / FieldType: text_cjk

□ Verbose Output

ST	胡	志	明
CJKWF	胡	志	明
LCF	胡	志	明
CJKBF	胡志	志明	

ST	上	海	明	心	寺	志
CJKWF	上	海	明	心	寺	志
LCF	上	海	明	心	寺	志
CJKBF	上海	海明	明心	心寺	寺志	

text_cjk field with text in Chinese & Latin characters 胡志明 (Hồ Chí Minh)



Field Value (Index): 胡志明 was a Vietnamese revolutionary and politician

Field Value (Query):

Analyse Fieldname / FieldType: text_cjk

□ Verbose Output

```
ST     胡    志    明    was   a           Vietnamese   revolutionary   and          politician
CJKWF  胡    志    明    was   a           Vietnamese   revolutionary   and          politician
LCF    胡    志    明    was   a           vietnamese   revolutionary   and          politician
CJKBF  胡志  志明  was   a     vietnamese  revolutionary  and          politician
```

bigrams for CJK

no bigrams for Latin alphabet

Bigrams

"Segmenting into character unigrams or bigrams is computationally easy and requires no knowledge of the language

[...]

Information retrieval research has generally found that simple approaches such as indexing overlapping character bigrams have comparable performance with more sophisticated word based approaches. As an example of overlapping bigrams, if the characters “ABCD” were Chinese characters the tokenizer would split them up into “AB” “BC” and “CD.”"

Source: https://www.hathitrust.org/blogs/large-scale-search/multilingual-issues-part-1-word-segmentation

Solr's CJK Bigram Filter

"Forms bigrams (overlapping 2-character sequences) of CJK characters that are generated from Standard Tokenizer or ICU Tokenizer.

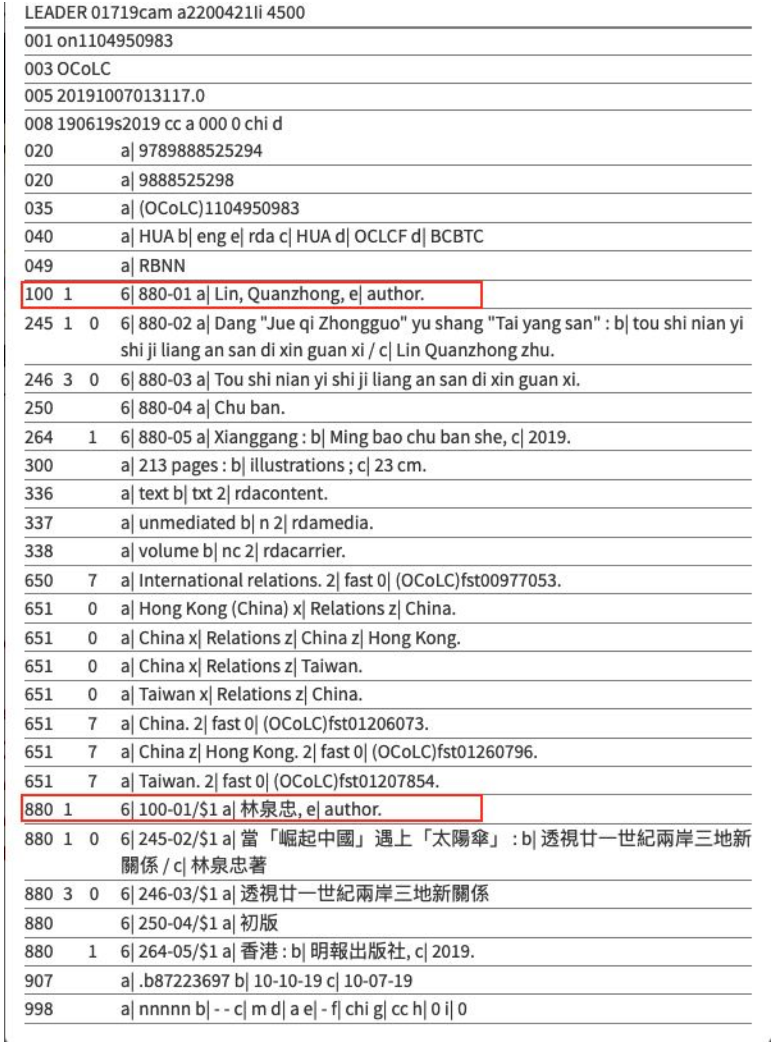
By default, all CJK characters produce bigrams, but finer grained control is available by specifying orthographic type arguments han, hiragana, katakana, and hangul. When set to false, characters of the corresponding type will be passed through as unigrams, and will not be included in any bigrams."

https://lucene.apache.org/solr/guide/8_6/language-analysis.html#cjk-bigram-filter

Indexing our CJK text

CJK text in our data

- Source data is in MARC
- Sample record
- **Author** in MARC 100:
  - Lin, Quanzhong
- Subfield $6 indicates **author in original script** in MARC 880:
  - 林泉忠



```
LEADER 01719cam a2200421li 4500
001 on1104950983
003 OCoLC
005 20191007013117.0
008 190619s2019 cc a 000 0 chi d
020          a| 9789888525294
020          a| 9888525298
035          a| (OCoLC)1104950983
040          a| HUA b| eng e| rda c| HUA d| OCLCF d| BCBTC
049          a| RBNN
100  1       6| 880-01 a| Lin, Quanzhong, e| author.
245  1  0    6| 880-02 a| Dang "Jue qi Zhongguo" yu shang "Tai yang san" : b| tou shi nian yi
             shi ji liang an san di xin guan xi / c| Lin Quanzhong zhu.
246  3  0    6| 880-03 a| Tou shi nian yi shi ji liang an san di xin guan xi.
250          6| 880-04 a| Chu ban.
264     1    6| 880-05 a| Xianggang : b| Ming bao chu ban she, c| 2019.
300          a| 213 pages : b| illustrations ; c| 23 cm.
336          a| text b| txt 2| rdacontent.
337          a| unmediated b| n 2| rdamedia.
338          a| volume b| nc 2| rdacarrier.
650     7    a| International relations. 2| fast 0| (OCoLC)fst00977053.
651     0    a| Hong Kong (China) x| Relations z| China.
651     0    a| China x| Relations z| China z| Hong Kong.
651     0    a| China x| Relations z| Taiwan.
651     0    a| Taiwan x| Relations z| China.
651     7    a| China. 2| fast 0| (OCoLC)fst01206073.
651     7    a| China z| Hong Kong. 2| fast 0| (OCoLC)fst01260796.
651     7    a| Taiwan. 2| fast 0| (OCoLC)fst01207854.
880  1       6| 100-01/$1 a| 林泉忠, e| author.
880  1  0    6| 245-02/$1 a| 當「崛起中國」遇上「太陽傘」: b| 透視廿一世紀兩岸三地新
             關係 / c| 林泉忠著
880  3  0    6| 246-03/$1 a| 透視廿一世紀兩岸三地新關係
880          6| 250-04/$1 a| 初版
880     1    6| 264-05/$1 a| 香港 : b| 明報出版社, c| 2019.
907          a| .b87223697 b| 10-10-19 c| 10-07-19
998          a| nnnnn b| - - c| m d| a e| - f| chi g| cc h| 0 i| 0
```

Indexing our data

- Specific CJK fields in addition to our existing fields
  - Existing: title_txt and author_txt for values in Latin alphabet
  - New: **title_txt_cjk** and **author_txt_cjk** for values using CJK characters

- Example
  - "Lin, Quanzhong" => author_txt (text_general)
  - "林泉忠" => author_txt_cjk (text_cjk)

Indexing author into **author_txt_cjk**

We use Traject (a Ruby gem) to process our MARC files
(source: https://github.com/Brown-University-Library/bul-traject/blob/master/config.rb#L397)

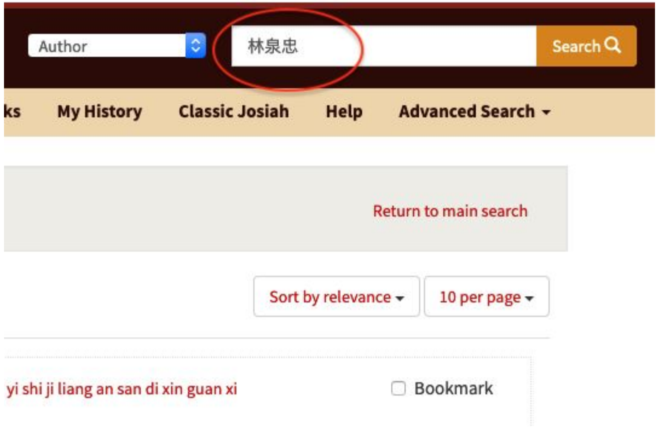
```
# Authors for CJK languages
author_vern_lambda = extract_marc('100abcdq:110abcd:111abcd', :alternate_script=>:only)
to_field "author_txt_cjk" do |rec, acc, context|
     ...
     authors_cjk = []
     author_vern_lambda.call (rec,authors_cjk,nil)
     authors_cjk.each do |author|
       acc << author
     end
end
```

Searching for CJK text

CJK searches

- Since we created separated fields
  - **author_txt** is text_general
  - **author_txt_cjk** is text_cjk

- ...now we need to decide when to use each ¯_(ツ)_/¯



- When searching for "Lin, Quanzhong" use field **author_txt**

- When searching for "林泉忠" use field **author_txt_cjk**

Is text in CJK?

- We are using a regular expression to detect CJK text

```
/\p{Han}|\p{Katakana}|\p{Hiragana}|\p{Hangul}/
```

- `\p{}` matches a character's Unicode script. (source)

```
if regex is a match
    use author_txt_cjk
else
    use author_txt
End
```

- Our code (Ruby): the controller and the regex

Works in PHP 7 too

```
<?php

// outputs 0
echo preg_match_all("/\p{Han}/u", "Lin, Quanzhong");

// outputs 3
echo preg_match_all("/\p{Han}/u", "林泉忠");

?>
```

Notice: The `/u` modifier is required

Current results

Precision for CJK searches has improved significantly

Search	Expected* Matches	Total Returned (before CJK)	Precision (before CJK)	Total Returned (with CJK)	Precision (with CJK)
莫言 (Mo Yan)	72	300	0.240	**56**	**1.285**
柳美里 (Yu Miri)	15	1000	0.015	**14**	**1.071**
ふくろうの本 (Fukuro no hon)	11	4800	0.002	**12**	**0.916**

* Our original "Expected Matches" values were off. The current Total Returned values are in fact more accurate.

A few other notes...

- Bug SOLR-13336 in older versions of Solr causes "exponential expansion of naive queries" when creating bigrams
  - Fixed in latest versions of Solr.

- Other more robust CJK configurations
  - Stanford and Michigan's Solr configurations

- Chinese Simplified vs Chinese Traditional
  - E.g. author "Zhang, Ailing"
  - 张爱玲 (Simplified) and 張愛玲 (Traditional)
  - Not handled by **text_cjk** field

Thanks

Many people were involved in making this work possible

- Thanks to the people at Discovery Day 2018 for making us aware of the problem, and in particular to Nikitas Tampakis and Michael Gibney.
- And thanks to Jeanette Norris and Toshiyuki Minami for helping document, implement, and test the solution at Brown.

Source and other references

- Naomi Dushay's posts: CJK with Solr for Libraries (11 posts)
- An introduction to indexing Chinese
- HathiTrust post on word segmentation
- Podcast: The Wubi Effect (Radiolab)

- MARC field 880
- Book: Solr in Action by Trey Grainger and Timothy Potter
- Shameless plug for my workshop: Solr for newbies

Questions and (hopefully) Answers

slides: https://tinyurl.com/solr-and-cjk | email: hector_correa@brown.edu | twitter: @hectorjcorrea