Why BASIS tokenization is not working for Chinese records? What can we do?

Alan Ng
Systems Librarian
The Chinese University of Hong Kong Library

Agenda

- about CUHK Library
- Chinese character 101
- how it all started
- about BASIS tokenization
- problems of BASIS
- a workaround solution
- BASIS problem part 2

CUHK Library

- established in 1965
- 7 branches
- ~220 staff
- ~60K current patrons
- 127K journal, 4.6M ebooks, ~2.6M printed vol.
- special collections includes oracle bones, Chinese rare books, modern Chinese literary

Main Library

Main Library

Chinese character 101

- in English, each word is composed of a string of characters
- but in Chinese language, each character (ideograph) has its own meaning and usage
- each Chinese character can be used independently
- in Chinese context, each character is actually a word

Chinese word 101

- single Chinese word examples:
- 我 (me) 你 (you) 大 (big) 小 (small) 書 (book)
- individual Chinese words combine to become a Chinese term or phrase, e.g.
- 我們 (we) 圖書館 (library) 電子書 (ebook)

Chinese word 101

- forming Chinese phrase from a stem
- 圖書 (book)
- 圖書館 (library)
- 圖書館員 (librarian)

How it all started?

- in 2013, I got a simple user enquiry:
- why searching “陪審” would obtain fewer results than “陪審團” in Primo?
- “陪審” means “the act of being a jury member”
- “陪審團” means the jury
- it just didn’t make sense, started a 13-month journey

What is BASIS?

- not mentioned widely
- in “Primo Version 4.x Highlights”, it is for improving search result relevancy for Chinese language
- implemented since Primo 4.3
- not much further info

BASIS tokenization

- analysis is done during indexing and searching
- the PNX record <search> fields are subject to tokenization
- e.g. the TITLE “諮詢文件 : 出任陪審員的準則” is broken into:
- “諮詢” “文件” “出任” “陪審員” “的” “準則”
- Primo only indexes these six tokens
- n-gram tokenization

BASIS tokenization

- the idea was to only index specific Chinese phrases instead of individual word

- based on assumption Chinese users would search for Chinese phrases, instead of arbitrary single Chinese word

- leading to fewer but more relevant search results

- but also ignoring a lot of matches

How a match is found under BASIS?

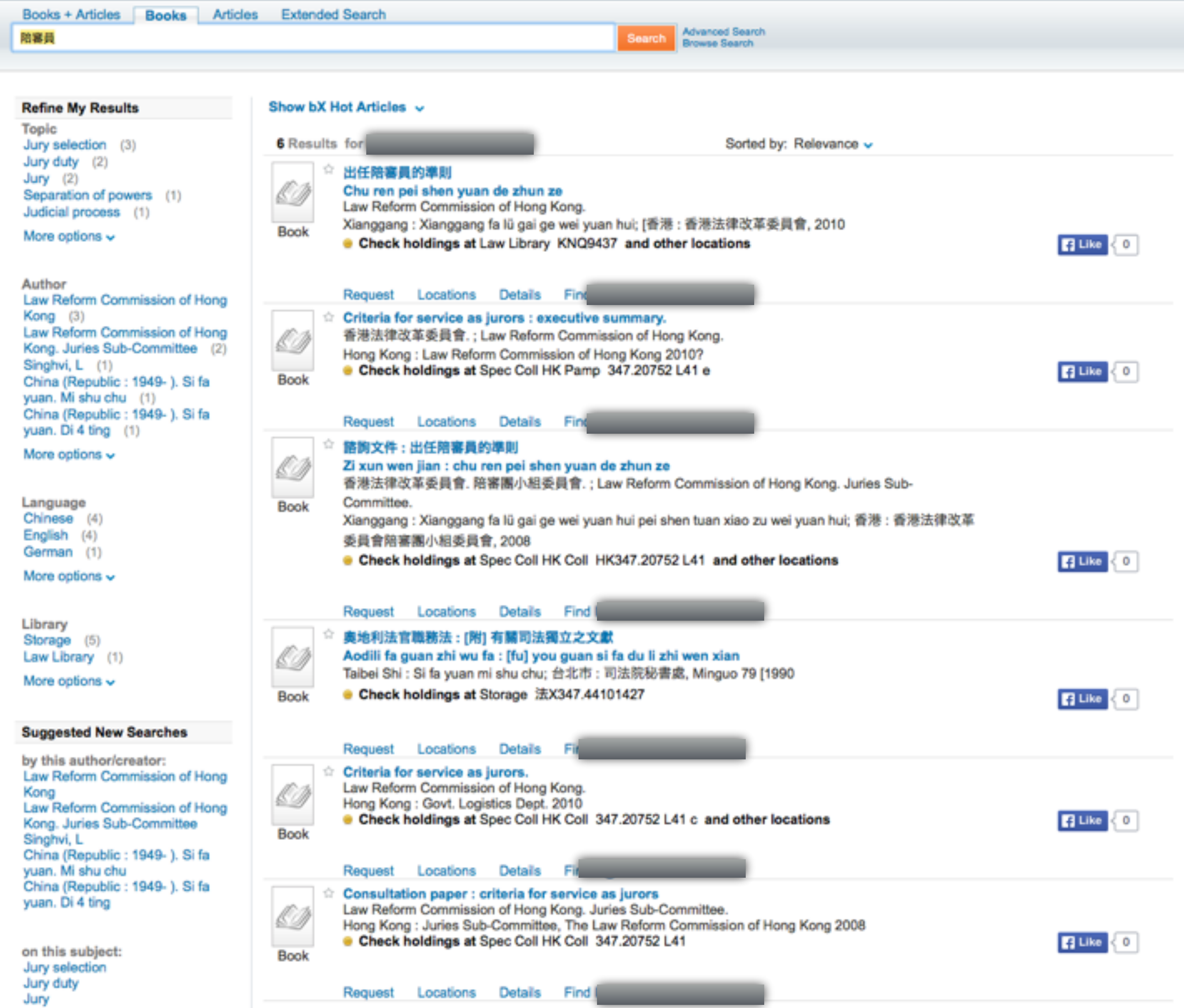
- not just the PNX data, the search terms will go through the same BASIS tokenization

- to retrieve an entry, you must have a matching token in your tokenized search terms

How a match is found under BASIS?

- to find the TITLE: “諮詢文件：出任陪審員的準則”
- your tokenized search teams must have either: “諮詢” “文件” “出任” “陪審員” “的” “準則”
- if you search with a stem of the token (e.g. “陪審”), you won’t get it
- it is problematic and unacceptable

Case 1 (6 entries for“陪審員”)



Refine My Results

Topic
- Jury selection (3)
- Jury duty (2)
- Jury (2)
- Separation of powers (1)
- Judicial process (1)
- More options ⌄

Author
- Law Reform Commission of Hong Kong (3)
- Law Reform Commission of Hong Kong. Juries Sub-Committee (2)
- Singhvi, L (1)
- China (Republic : 1949-). Si fa yuan. Mi shu chu (1)
- China (Republic : 1949-). Si fa yuan. Di 4 ting (1)
- More options ⌄

Language
- Chinese (4)
- English (4)
- German (1)
- More options ⌄

Library
- Storage (5)
- Law Library (1)
- More options ⌄

Suggested New Searches

by this author/creator:
- Law Reform Commission of Hong Kong
- Law Reform Commission of Hong Kong. Juries Sub-Committee
- Singhvi, L
- China (Republic : 1949-). Si fa yuan. Mi shu chu
- China (Republic : 1949-). Si fa yuan. Di 4 ting

on this subject:
- Jury selection
- Jury duty
- Jury

Show bX Hot Articles ⌄

6 Results for [redacted]   Sorted by: Relevance ⌄

☆ 出任陪審員的準則
Chu ren pei shen yuan de zhun ze
Law Reform Commission of Hong Kong.
Xianggang : Xianggang fa lü gai ge wei yuan hui; [香港 : 香港法律改革委員會, 2010
● **Check holdings at** Law Library KNQ9437 **and other locations**
Book
Like 0
Request Locations Details Fin [redacted]

☆ Criteria for service as jurors : executive summary.
香港法律改革委員會. ; Law Reform Commission of Hong Kong.
Hong Kong : Law Reform Commission of Hong Kong 2010?
● **Check holdings at** Spec Coll HK Pamp 347.20752 L41 e
Book
Like 0
Request Locations Details Fin [redacted]

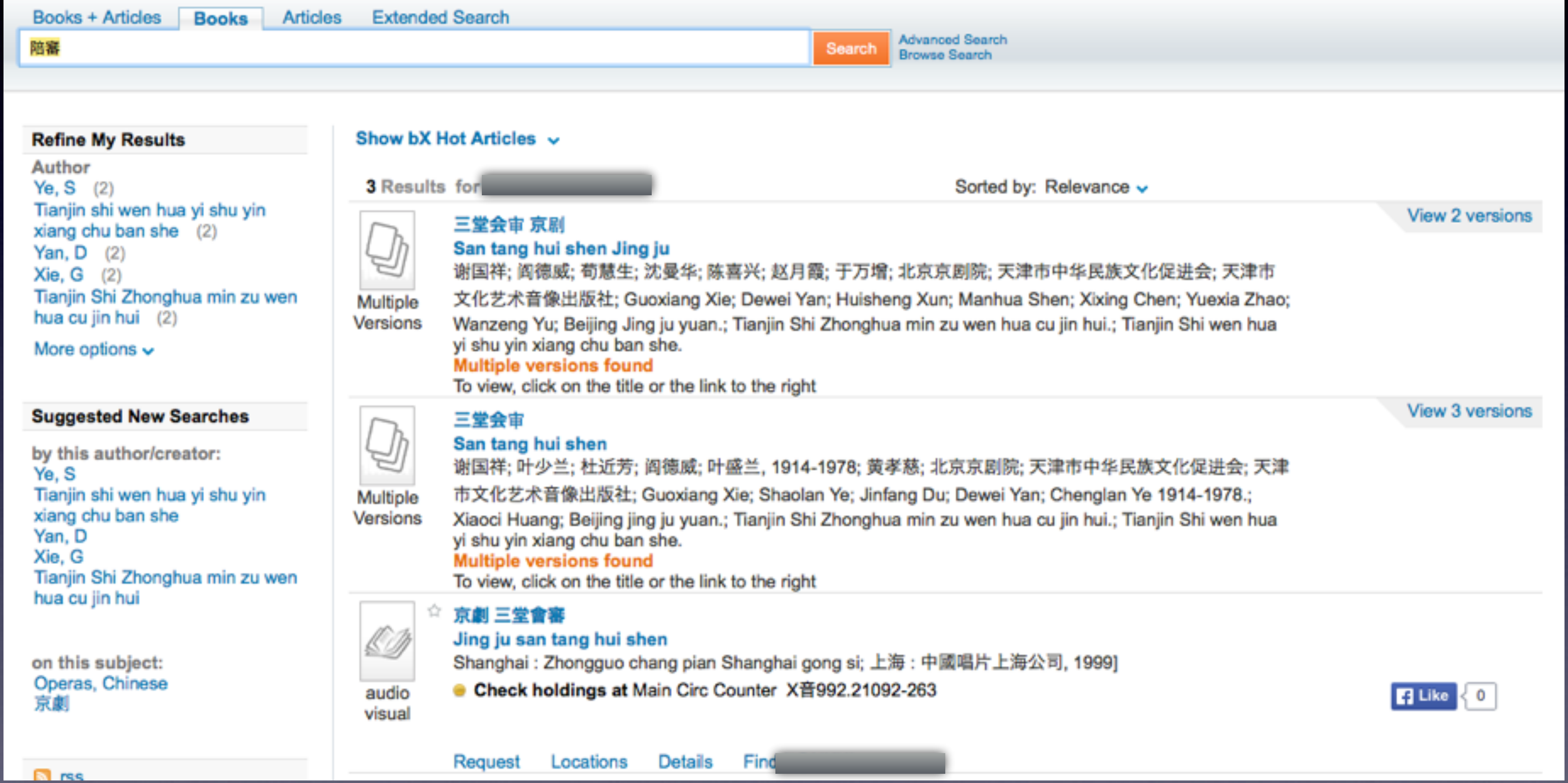
☆ 諮詢文件 : 出任陪審員的準則
Zi xun wen jian : chu ren pei shen yuan de zhun ze
香港法律改革委員會. 陪審團小組委員會. ; Law Reform Commission of Hong Kong. Juries Sub-Committee.
Xianggang : Xianggang fa lü gai ge wei yuan hui pei shen tuan xiao zu wei yuan hui; 香港 : 香港法律改革委員會陪審團小組委員會, 2008
● **Check holdings at** Spec Coll HK Coll HK347.20752 L41 **and other locations**
Book
Like 0
Request Locations Details Find [redacted]

☆ 奧地利法官職務法 : [附] 有關司法獨立之文獻
Aodili fa guan zhi wu fa : [fu] you guan si fa du li zhi wen xian
Taibei Shi : Si fa yuan mi shu chu; 台北市 : 司法院秘書處, Minguo 79 [1990
● **Check holdings at** Storage 法X347.44101427
Book
Like 0
Request Locations Details Fi [redacted]

☆ Criteria for service as jurors.
Law Reform Commission of Hong Kong.
Hong Kong : Govt. Logistics Dept. 2010
● **Check holdings at** Spec Coll HK Coll 347.20752 L41 c **and other locations**
Book
Like 0
Request Locations Details Fi [redacted]

☆ Consultation paper : criteria for service as jurors
Law Reform Commission of Hong Kong. Juries Sub-Committee.
Hong Kong : Juries Sub-Committee, The Law Reform Commission of Hong Kong 2008
● **Check holdings at** Spec Coll HK Coll 347.20752 L41
Book
Like 0
Request Locations Details Find [redacted]

Case 1 (3 entries for “陪審”)



陪審

Refine My Results

Author

Ye, S (2)

Tianjin shi wen hua yi shu yin xiang chu ban she (2)

Yan, D (2)

Xie, G (2)

Tianjin Shi Zhonghua min zu wen hua cu jin hui (2)

More options

Suggested New Searches

by this author/creator:

Ye, S

Tianjin shi wen hua yi shu yin xiang chu ban she

Yan, D

Xie, G

Tianjin Shi Zhonghua min zu wen hua cu jin hui

on this subject:

Operas, Chinese

京劇

3 Results for [redacted]

Sorted by: Relevance

三堂会审 京剧

San tang hui shen Jing ju

谢国祥; 阎德威; 荀慧生; 沈曼华; 陈喜兴; 赵月霞; 于万增; 北京京剧院; 天津市中华民族文化促进会; 天津市文化艺术音像出版社; Guoxiang Xie; Dewei Yan; Huisheng Xun; Manhua Shen; Xixing Chen; Yuexia Zhao; Wanzeng Yu; Beijing Jing ju yuan.; Tianjin Shi Zhonghua min zu wen hua cu jin hui.; Tianjin Shi wen hua yi shu yin xiang chu ban she.

Multiple versions found

To view, click on the title or the link to the right

三堂会审

San tang hui shen

谢国祥; 叶少兰; 杜近芳; 阎德威; 叶盛兰, 1914-1978; 黄孝慈; 北京京剧院; 天津市中华民族文化促进会; 天津市文化艺术音像出版社; Guoxiang Xie; Shaolan Ye; Jinfang Du; Dewei Yan; Chenglan Ye 1914-1978.; Xiaoci Huang; Beijing jing ju yuan.; Tianjin Shi Zhonghua min zu wen hua cu jin hui.; Tianjin Shi wen hua yi shu yin xiang chu ban she.

Multiple versions found

To view, click on the title or the link to the right

☆ **京劇 三堂會審**

Jing ju san tang hui shen

Shanghai : Zhongguo chang pian Shanghai gong si; 上海 : 中國唱片上海公司, 1999]

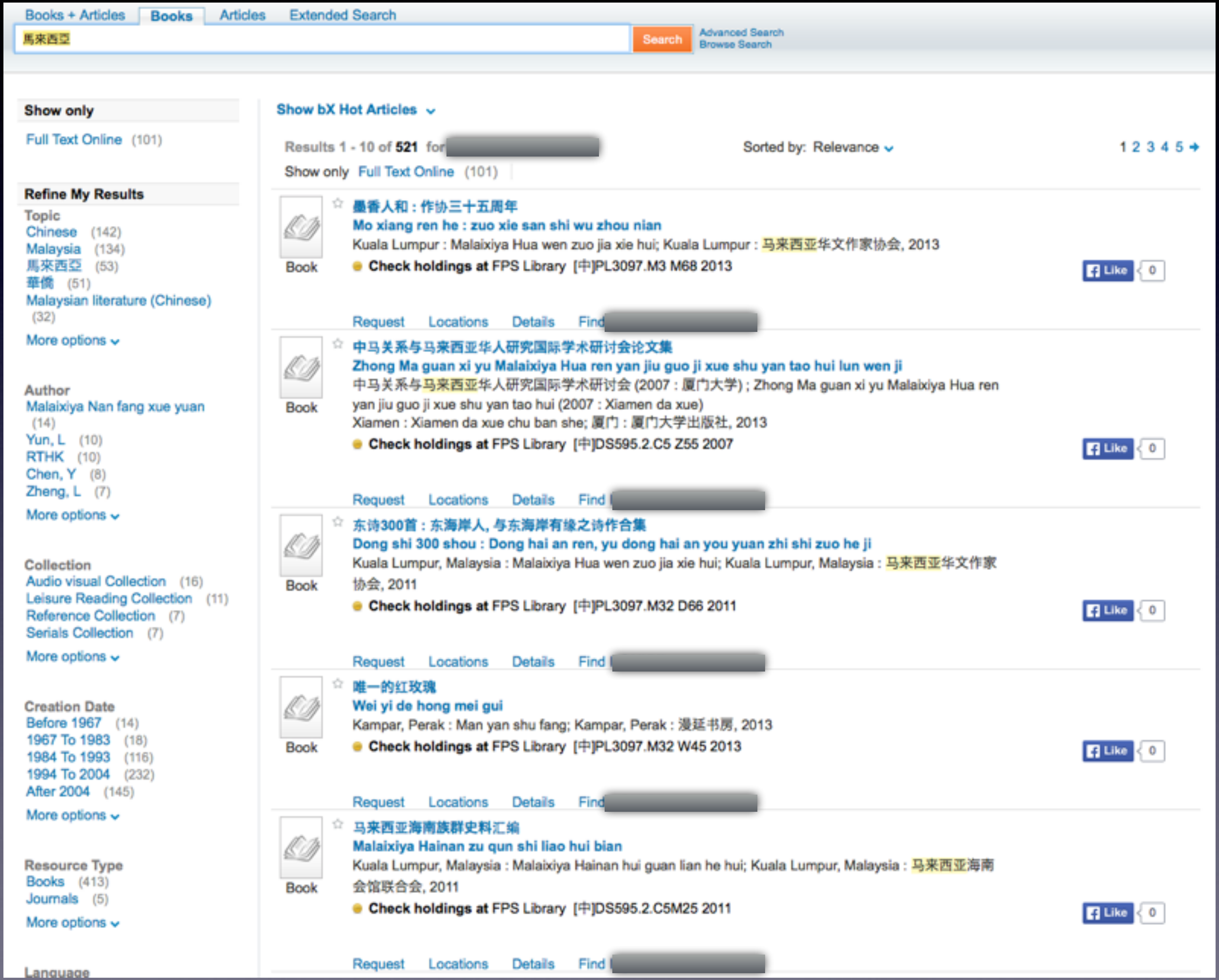
● **Check holdings at** Main Circ Counter X音992.21092-263

Find [redacted]

What is wrong?

- stem “陪審” gets far fewer matches (6) than the longer term“陪審員” (3)
- both result sets mutually exclusive
- but the search terms share common characters

Case 2 (521 entries for “馬來西亞”)



Books + Articles | **Books** | Articles | Extended Search

馬來西亞 | Search | Advanced Search
Browse Search

Show only

Full Text Online (101)

Refine My Results

Topic
- Chinese (142)
- Malaysia (134)
- 馬來西亞 (53)
- 華僑 (51)
- Malaysian literature (Chinese) (32)
- More options ⌄

Author
- Malaixiya Nan fang xue yuan (14)
- Yun, L (10)
- RTHK (10)
- Chen, Y (8)
- Zheng, L (7)
- More options ⌄

Collection
- Audio visual Collection (16)
- Leisure Reading Collection (11)
- Reference Collection (7)
- Serials Collection (7)
- More options ⌄

Creation Date
- Before 1967 (14)
- 1967 To 1983 (18)
- 1984 To 1993 (116)
- 1994 To 2004 (232)
- After 2004 (145)
- More options ⌄

Resource Type
- Books (413)
- Journals (5)
- More options ⌄

Language

Show bX Hot Articles ⌄

Results 1 - 10 of **521** for [redacted] Sorted by: Relevance ⌄ 1 2 3 4 5 →

Show only Full Text Online (101)

Book ☆ **墨香人和 : 作协三十五周年**
Mo xiang ren he : zuo xie san shi wu zhou nian
Kuala Lumpur : Malaixiya Hua wen zuo jia xie hui; Kuala Lumpur : 马来西亚华文作家协会, 2013
● **Check holdings** at FPS Library [中]PL3097.M3 M68 2013
Like 0
Request Locations Details Find [redacted]

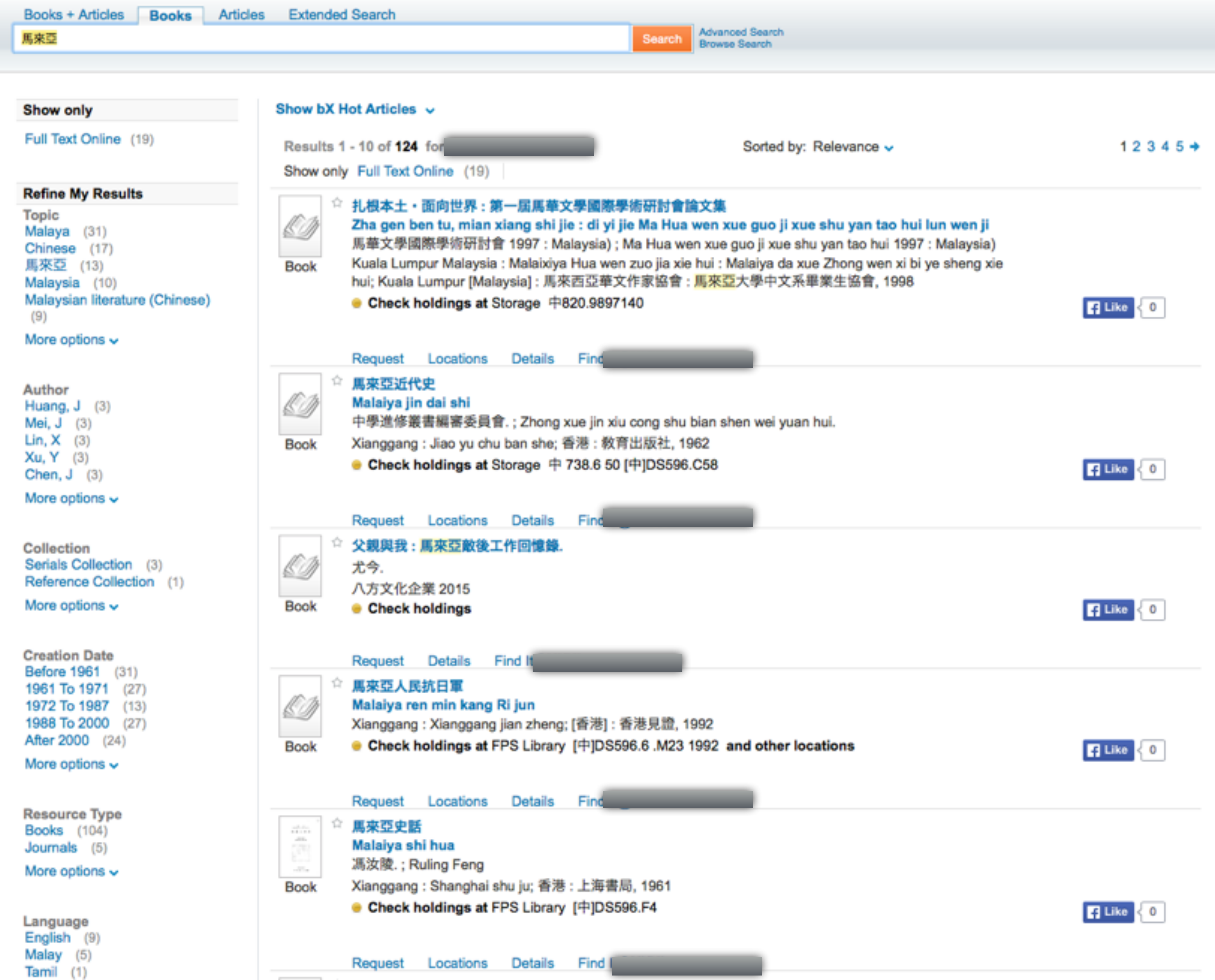
Book ☆ **中马关系与马来西亚华人研究国际学术研讨会论文集**
Zhong Ma guan xi yu Malaixiya Hua ren yan jiu guo ji xue shu yan tao hui lun wen ji
中马关系与马来西亚华人研究国际学术研讨会 (2007 : 厦门大学) ; Zhong Ma guan xi yu Malaixiya Hua ren yan jiu guo ji xue shu yan tao hui (2007 : Xiamen da xue)
Xiamen : Xiamen da xue chu ban she; 厦门 : 厦门大学出版社, 2013
● **Check holdings** at FPS Library [中]DS595.2.C5 Z55 2007
Like 0
Request Locations Details Find [redacted]

Book ☆ **东诗300首 : 东海岸人, 与东海岸有缘之诗作合集**
Dong shi 300 shou : Dong hai an ren, yu dong hai an you yuan zhi shi zuo he ji
Kuala Lumpur, Malaysia : Malaixiya Hua wen zuo jia xie hui; Kuala Lumpur, Malaysia : 马来西亚华文作家协会, 2011
● **Check holdings** at FPS Library [中]PL3097.M32 D66 2011
Like 0
Request Locations Details Find [redacted]

Book ☆ **唯一的红玫瑰**
Wei yi de hong mei gui
Kampar, Perak : Man yan shu fang; Kampar, Perak : 漫延书房, 2013
● **Check holdings** at FPS Library [中]PL3097.M32 W45 2013
Like 0
Request Locations Details Find [redacted]

Book ☆ **马来西亚海南族群史料汇编**
Malaixiya Hainan zu qun shi liao hui bian
Kuala Lumpur, Malaysia : Malaixiya Hainan hui guan lian he hui; Kuala Lumpur, Malaysia : 马来西亚海南会馆联合会, 2011
● **Check holdings** at FPS Library [中]DS595.2.C5M25 2011
Like 0
Request Locations Details Find [redacted]

Case 2 (124 entries for “馬來亞”)



馬來亞

Show only

Full Text Online (19)

Refine My Results

Topic
Malaya (31)
Chinese (17)
馬來亞 (13)
Malaysia (10)
Malaysian literature (Chinese) (9)
More options

Author
Huang, J (3)
Mei, J (3)
Lin, X (3)
Xu, Y (3)
Chen, J (3)
More options

Collection
Serials Collection (3)
Reference Collection (1)
More options

Creation Date
Before 1961 (31)
1961 To 1971 (27)
1972 To 1987 (13)
1988 To 2000 (27)
After 2000 (24)
More options

Resource Type
Books (104)
Journals (5)
More options

Language
English (9)
Malay (5)
Tamil (1)

Show bX Hot Articles

Results 1 - 10 of **124** for [redacted]

Sorted by: Relevance

1 2 3 4 5 →

Show only Full Text Online (19)

Book
☆ 扎根本土・面向世界：第一屆馬華文學國際學術研討會論文集
Zha gen ben tu, mian xiang shi jie : di yi jie Ma Hua wen xue guo ji xue shu yan tao hui lun wen ji
馬華文學國際學術研討會 1997 : Malaysia) ; Ma Hua wen xue guo ji xue shu yan tao hui 1997 : Malaysia)
Kuala Lumpur Malaysia : Malaixiya Hua wen zuo jia xie hui : Malaiya da xue Zhong wen xi bi ye sheng xie hui; Kuala Lumpur [Malaysia] : 馬來西亞華文作家協會 : 馬來亞大學中文系畢業生協會, 1998
● **Check holdings at** Storage 中820.9897140
Like 0
Request Locations Details Find [redacted]

Book
☆ 馬來亞近代史
Malaiya jin dai shi
中學進修叢書編審委員會. ; Zhong xue jin xiu cong shu bian shen wei yuan hui.
Xianggang : Jiao yu chu ban she; 香港 : 教育出版社, 1962
● **Check holdings at** Storage 中 738.6 50 [中]DS596.C58
Like 0
Request Locations Details Find [redacted]

Book
☆ 父親與我：馬來亞敵後工作回憶錄.
尤今.
八方文化企業 2015
● **Check holdings**
Like 0
Request Details Find [redacted]

Book
☆ 馬來亞人民抗日軍
Malaiya ren min kang Ri jun
Xianggang : Xianggang jian zheng; [香港] : 香港見證, 1992
● **Check holdings at** FPS Library [中]DS596.6 .M23 1992 **and other locations**
Like 0
Request Locations Details Find [redacted]

Book
☆ 馬來亞史話
Malaiya shi hua
馮汝陵. ; Ruling Feng
Xianggang : Shanghai shu ju; 香港 : 上海書局, 1961
● **Check holdings at** FPS Library [中]DS596.F4
Like 0
Request Locations Details Find [redacted]

Problems about BASIS

- Chinese language is too complicated
- the tokenization logic is unknown
- the Chinese phrase dictionary used by the BASIS process is unknown
- no recursive tokenization, stems are ignored
- e.g. “陪審員” won’t be further broken down into “陪審”,“員” and “陪” “審” “員”

Problems about BASIS

- also affect Japanese records using the “kanji” (漢字), which shares the same Chinese characters
- Oxford University Library found the same problem in May 2015
- any Primo with Chinese / Japanese record is affected

Chinese / Japanese record discovery fails

- Nevertheless, an exact title match (including punctuations) can always retrieve the record
- because both raw data and search term will go through the exact same tokenization
- it became a “known search”
- purpose of discovery defeated

A workaround solution

- the BASIS current approach simply doesn’t work
- BASIS needs to be OFF
- needs to go back to index every single Chinese character separately (1-gram tokenization)
- Ex Libris acknowledged the problem in Oct 2014
- help CUHK turn off BASIS and reindex
- please refer to support case #00092878 (cuhk)

became a known issue

- listed among the Known Issues for both Primo 4.9 and April 2015 release
- solution still pending

Suggestions

- enhance the tokenization dictionary to handle the stems within the Chinese phrase

- do recursive tokenization, detect the stems

- keep both n-gram and unigram (1-gram) tokenization for enhanced relevancy and complete result set

BASIS problem part 2?

- searching a string of over 30 Chinese character
- huge HPROF file (10+GB) spawned in FE se_bin folder
- “/exlibris/primo/p4_1/ng/jaguar/home/system/bin"
- will hang the Primo FE
- support case #00119959 (cuhk)
- permanent fix expected in Primo 2015 November release

Q & A

Thank you!

Contact:
alanng@cuhk.edu.hk