Chinese, Japanese & Korean Searching in Quicksearch

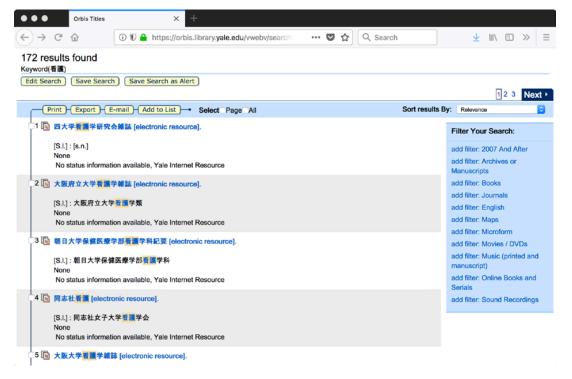
SearchFest! 2018

Harvey Cushing / John Hay Whitney Medical Library
Yale University

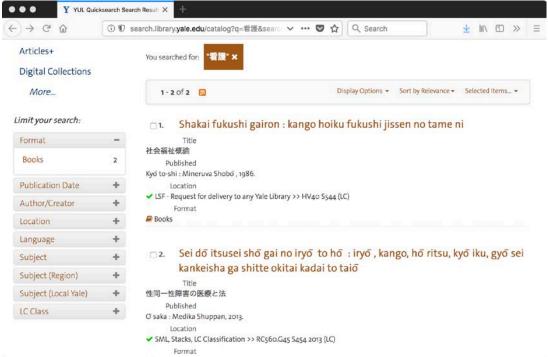
Before ...

$\Box = nursing$, traditional Chinese

172 results in Orbis

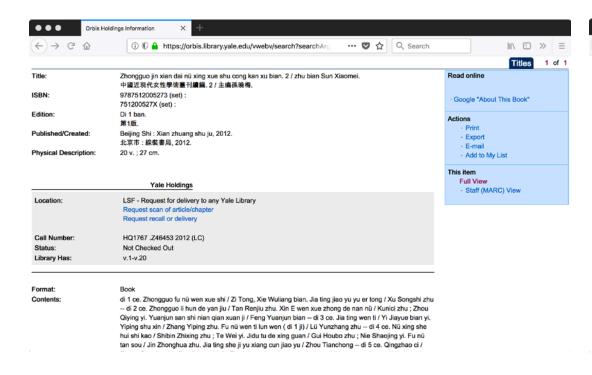


2 results in Quicksearch

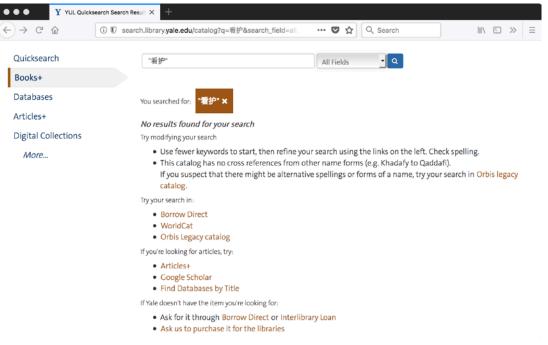


$\Box \Box = nursing$, simplified Chinese

1 result in Orbis

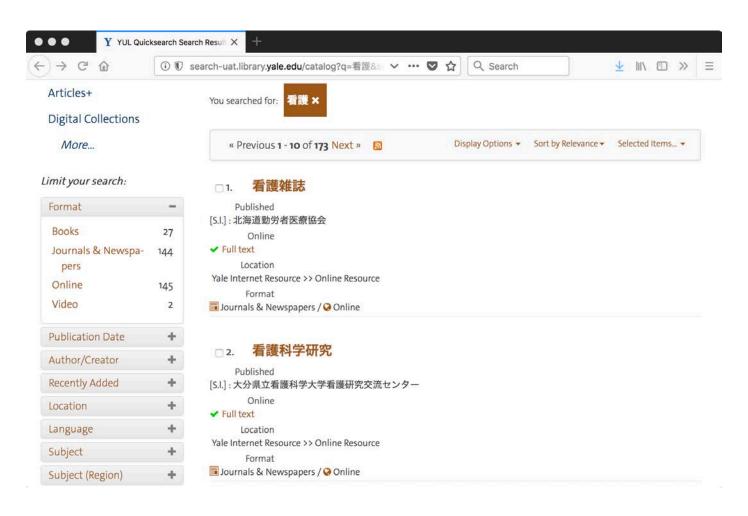


No results in Quicksearch

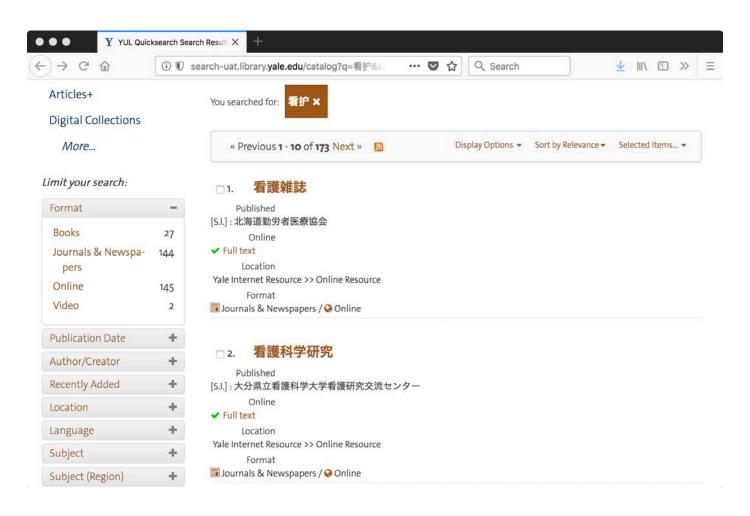


After ...

(nursing, traditional Chinese)173 results



(nursing, simplified Chinese)173 results



The solution?

CJK tokenization using bigrams (not spaces)

- 看護雜誌 (Nursing Magazine)
- 看護科学研究 (Nursing Science)
- ◆ 看護研究集録 (Acquisition of Nursing Research)
- 看護学統合研究 (Nursing Integrated Research)
- 看護大からこんにちは (Hello from Nursing University)

Document term index (English)

| | acquisition | hello | integrated | magazine | nursing | research | science | university |
|------|-------------|-------|------------|----------|---------|----------|---------|------------|
| Doc1 | | | | 1 | 1 | | | |
| Doc2 | | | | | 1 | | 1 | |
| Doc3 | 1 | | | | 1 | 1 | | |
| Doc4 | | | 1 | | 1 | 1 | | |
| Doc5 | | 1 | | | 1 | | | 1 |

Document term index (Chinese)

Doc1: 看護雑誌

Index overlapping bigrams.

Doc2: 看護科学研究

| | 看護 | 護雑 | 雑誌 | 護科 | 科学 | 学研 | 研究 |
|------|----|----|----|----|----|----|----|
| Doc1 | 1 | 1 | 1 | | | | |
| Doc2 | 1 | | | 1 | 1 | 1 | 1 |

An added enhancement ...

Folding equivalent characters

Not available in Orbis!

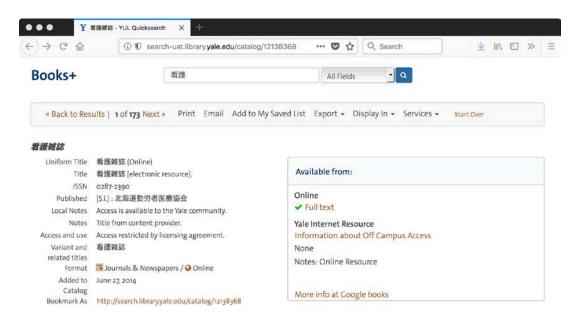
- □ □ (nursing, traditional Chinese) <=> □ □ (nursing, simplified Chinese)
- かんご (nursing, Japanese Hiragana) <=> カンゴ (nursing, Japanese Katakana)

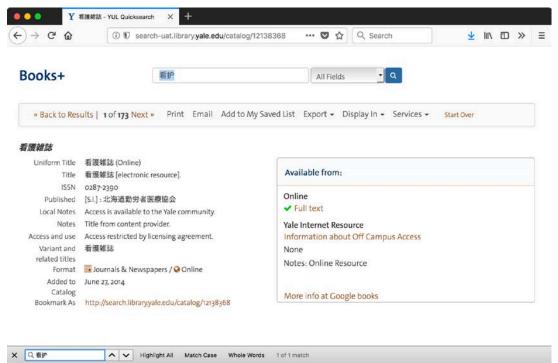
Relevancy ranking across folded characters

The first hit is the same. \square \square does not occur in the record!

□ □ (nursing, traditional Chinese)







Credits

- Yukari Sugiyama (primary stakeholder representative)
 ... check out her test results at http://bit.ly/CJKYALE
- Suzanne Lovejoy and Kalee Sprague (QAG co-chairs)
- Osman Din and Martin Lovell (LIT programmers)
- Dajin Sun and Ji-Young Kim (TS language consultants)
- Code adapted from Stanford University Library
 - ... read more <u>here</u>