Chinese, Japanese & Korean Searching in Quicksearch

SearchFest! 2018
Harvey Cushing / John Hay Whitney Medical Library
Yale University
Before ...
□ □ = nursing, traditional Chinese

172 results in Orbis

2 results in Quicksearch
查看 = 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)

- 看護雑誌 (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)

<table>
<thead>
<tr>
<th></th>
<th>acquisition</th>
<th>hello</th>
<th>integrated</th>
<th>magazine</th>
<th>nursing</th>
<th>research</th>
<th>science</th>
<th>university</th>
</tr>
</thead>
<tbody>
<tr>
<td>Doc1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Doc2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Doc3</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Doc4</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Doc5</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>
Document term index (Chinese)

Doc1: 看護雑誌

Doc2: 看護科学研究

<table>
<thead>
<tr>
<th></th>
<th>看護</th>
<th>護雑</th>
<th>雑誌</th>
<th>護科</th>
<th>科学</th>
<th>学研</th>
<th>研究</th>
</tr>
</thead>
<tbody>
<tr>
<td>Doc1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Doc2</td>
<td>1</td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Index overlapping bigrams.
An added enhancement ...
Folding equivalent characters

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

Not available in Orbis!
Relevancy ranking across folded characters

The first hit is the same. □ □ 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 here