IFLA—the global voice of libraries

IFLA: portavoz mundial de las bibliotecas

IFLA : la voix des bibliothèques dans le monde

IFLA—die globale Stimme der Bibliotheken

ИФЛА - глобальный голос библиотек

الإف إم إل - الصوت الدولي للمكتبات
News Alike – Text analytics to link related NewspaperSG articles

Kia Siang Hock
kia_siang_hock@nlb.gov.sg
About NLB

Libraries & Archives

1 National Library
1 National Archives
25 Public Libraries
Excellence in Singapore Content

NewspaperSG

Infopedia

BookSG

PictureSG

MusicSG

Singapore Memory
Using Text Analytics to Connect Content

Text tokenised; tokens parsed and weighted (TF/IDF)

Similarity = 0.295

Weighted tokens similarity computed

Text tokenised; tokens parsed and weighted (TF/IDF)
Discovery of related content

Providing a seamless & comprehensive contextual discovery experience to promote deeper understanding
About NewspaperSG

- Online resource of current and historic Singapore and Malaya newspapers
  - includes The Straits Times, The Business Times, 星洲日報, 南洋商报, 联合早报, Berita Harian, TODAY
- Over 20,000,000 articles published
- Full articles accessible off-site:
  - Up to 1989: The Straits Time, 星洲日報, Berita Harian
  - Up to 2009: The Business Times, TODAY
PoC on NewspaperSG (The Straits Time, 1989)
The NewspaperSG Challenges

2 key factors that determine the processing resources required:
① size of the data set  ② number of dimensions in the data set

<table>
<thead>
<tr>
<th>Number of articles</th>
<th>Number of dimensions</th>
</tr>
</thead>
<tbody>
<tr>
<td>20 million articles</td>
<td>OCR errors</td>
</tr>
<tr>
<td>20 million X 20 million pairs to compute for similarity values!!</td>
<td># of dimensions for 6.28 m English articles: 5.78m  # of words in WordNet: 155,287</td>
</tr>
</tbody>
</table>

SINGAPORE, Fri. A lion with a fish tail rising from a shimmering pea on which the words "Lion City" appear has been adopted by the Tourist Promotion Board as its emblem. This emblem will be used in all the Board's promotional material.
The Initial Attempts

Increase # of servers

Reduce # of dimensions

<table>
<thead>
<tr>
<th>Minimum support</th>
<th>Minimum document frequency</th>
<th>Max document frequency percentage</th>
<th>No of dimensions</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1</td>
<td>90</td>
<td>5,780,000</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
<td>90</td>
<td>1,521,828</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
<td>25</td>
<td>1,521,828</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>25</td>
<td>1,119,369</td>
</tr>
<tr>
<td>30</td>
<td>10</td>
<td>25</td>
<td>473,883</td>
</tr>
</tbody>
</table>

4 years of English newspaper articles processed in 4 days
Almost there...

Infopedia – close to 1,700 articles on Singapore’s history, culture, people, places and events

Housing and Development Board (HDB)

The Housing and Development Board (HDB) is the national public housing authority of Singapore and it is currently headquartered at HDB Hub in Toa Payoh. It was formed in February 1960, shortly after Singapore attained self-government, to alleviate the severe housing shortage facing the state at the time. Through the years, the emphasis of its housing programmes has shifted from quantity of housing to quality of life. Since 1965, its flats have been housing over 60% of Singapore’s resident population.

Background
HDB was formed as a direct result of the failure of its predecessor, the Singapore Improvement Trust (SIT), to provide adequate housing for the people. SIT was established in 1927 by the colonial government at a time when Singapore was facing an acute housing shortage. By the time SIT was replaced by HDB in 1960, the housing problem had worsened significantly as its building programmes had fallen far short of what was required to keep pace with the fast-growing population.

The first step towards the formation of HDB was taken in August 1958 with the introduction of the Housing and Development Bill in the legislative assembly. The bill was passed in January 1959 and came into effect on 1 February 1960, thus bringing HDB into existence to succeed SIT as Singapore’s housing authority. Its primary function was to build and manage housing units for the low income groups, and it had to provide the maximum return in the minimum time and at the lowest maintenance cost to the government.

Key Phrase Extraction
HDB flats, public housing, new flats, home ownership, home ownership rate, housing shortage, HDB Hub, Background HDB, housing programmes, housing authority, housing problem, housing units, HDB home ownership, housing estate, estate renewal strategy

Search NewspaperSG
Top 20,000 NewspaperSG articles

Similarity Processing
Similar Articles
Almost there…

Major oil spills in the Straits of Singapore

The Straits of Singapore, 48 km long and 3.1 km wide, lies between Singapore and the Riau Archipelago, Indonesia, and links the Straits of Malacca to the South China Sea. It is on the shipping route of the Asia-Pacific region linking West Asia to Europe. This makes it as one of the busiest sea-lanes in the world. Furthermore, Singapore is the busiest port as well as one of the largest oil refining centres in the world. This makes the Straits highly vulnerable to oil spills. It is considered one of the world’s hot spots for oil spills. Singapore and Malaysia together have suffered at least 39 spills of 34 tons or more since 1960

Recognising the high vulnerability of the Straits to incidents leading to oil spills…

Related NewspaperSG articles

- Big oil spill clean-up after tankers collide, 17 Oct 1997
- The man behind the clean up, 15 Oct 2000
- Oil spill off S’pore after ships collide, 6 Dec 2002
- Beaches stay clean as oil spill is contained, 14 Jun 2002
- Prevent further oil spills now, 4 Jul 1998
- Exaggerated? 29 Apr 1999
- Don’t over-dramatise piracy issues, 29 Apr 1999
- Danger of oil spill from tankers, 22 Jan 1971
- All familiar with emergency procedures on oil pollution, 25 Feb 1983

- 1,682 clusters processed
- 1,521,059 articles covered

But…

- Gaps in local news coverage
- Regional and international news not catered for
- Infopedia mainly in English; cannot cater for Chinese and Malay papers
Clustering – Final Piece to the Puzzle

Clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters).

Mahout K-Means Clustering with Cosine Distance
Clustering – Final Piece to the Puzzle

<table>
<thead>
<tr>
<th>Size of cluster</th>
<th>Top 50 stemmed terms in the cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>86,881</td>
<td>olymp, athlet, game, sport, medal, event, team, gold, championship, record, world, singapor, metr, swim, won, year, champion, win, nation, women, time, coach, asian, meet, competit, train, swimmer, two, second, race, compet, first, amateur, bronz, intern, associ, best, finish, yesterdai, silver, old, on, relai, men, set, track, medallist, mark, south, run</td>
</tr>
<tr>
<td>142,289</td>
<td>school, student, educ, teacher, univers, secondari, singapor, children, year, primari, studi, pupil, parent, mr, teach, colleg, cours, ministri, english, languag, chines, on, institut, examin, time, learn, princip, train, programm, work, graduat, help, nation, class, two, govern, girl, scienc, boi, first, level, malai, centr, make, dr, academ, dai, organis, scholarship, junior</td>
</tr>
<tr>
<td>125,629</td>
<td>polic, arrest, offic, suspect, two, men, yesterdai, man, investig, report, found, mr, gang, road, raid, on, station, crime, detain, year, night, arm, robberi, car, believ, peopl, forc, charg, singapor, robber, hous, todai, told, stolen, seiz, spokesman, old, held, four, escap, murder, reuter, chines, detect, member, street, made, drug, dai, home</td>
</tr>
</tbody>
</table>

67 clusters averaging 93,000 articles each
Clustering – Final Piece to the Puzzle

By Language  | # of articles
---|---
English articles | 6,277,574
Chinese articles | 2,280,158
Malay articles | 978,665

Close to 1 billion similarity associations identified!!

News Alike service
What’s Next?

Discover related newspaper articles across languages

- Welcome
- Selamat datang

Machine translation

Clustering

Similarity Processing
The Cenotaph, located at Esplanade Park along Connaught Drive, is a war memorial which commemorates the sacrifice of the men who perished during World War I and World War II. It was unveiled on 31 March 1922 by the Prince of Wales. On 28 December 2010, it was gazetted as a national monument together with two other structures in Esplanade Park, the Lim Bo Seng Memorial and the Tan Kim Seng Fountain...

**Newspaper articles**

- Singapore’s War Memorial to the Glorious Dead (11 Nov 1920)
- Singapore’s War Memorial (26 Sep 1921)
- Arrival of the Prince (31 Mar 1922)
- Lest we forget (8 Nov 1953)
- Singapore students learn to care about history (13 Jul 1997)

**Contextual Discovery – the Journey**

Connecting content to content to connect people to knowledge
Questions?

Kia Siang Hock
kia_siang_hock@nlb.gov.sg