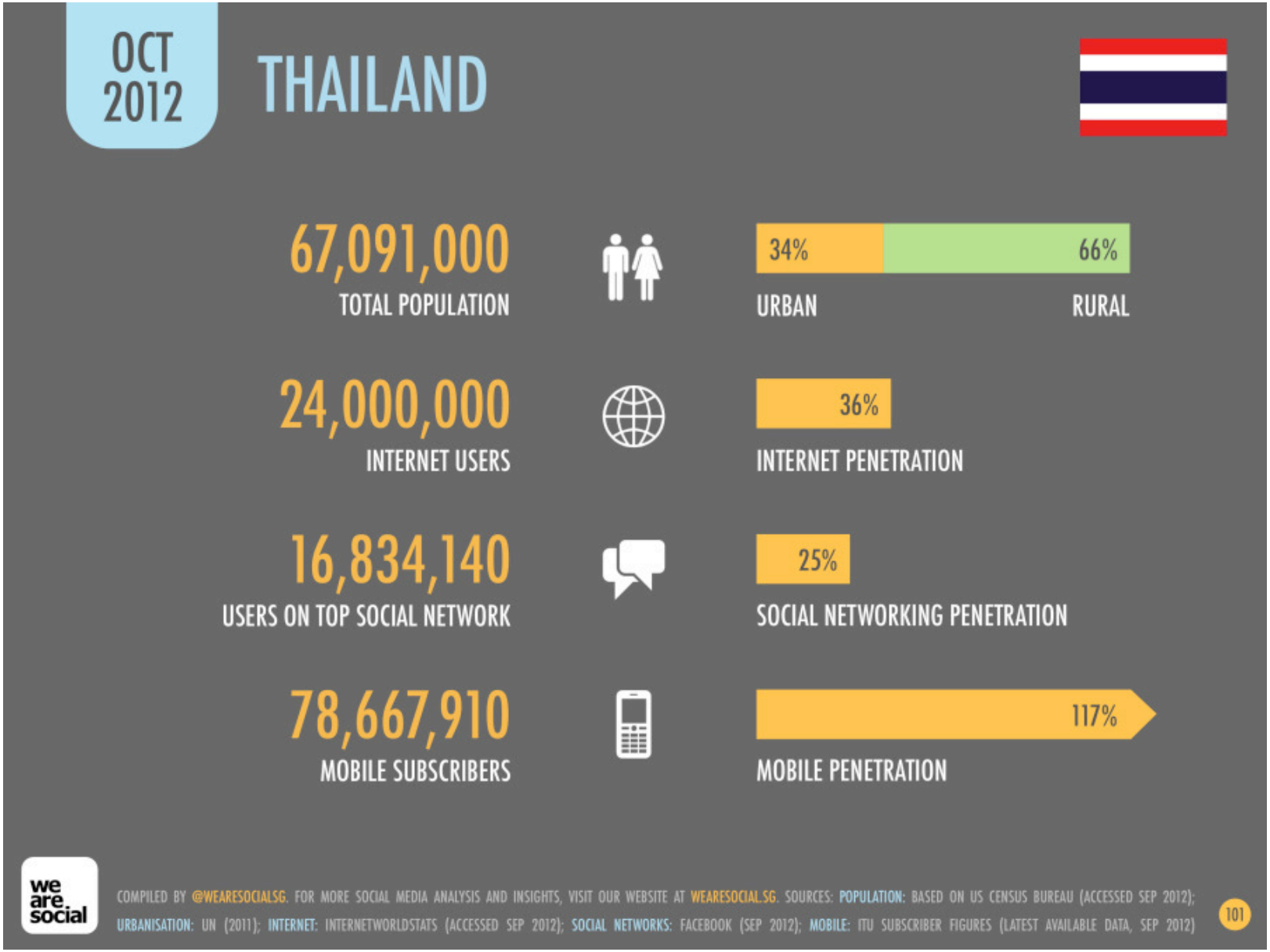
Social Movement Understanding by Keyword Tracking

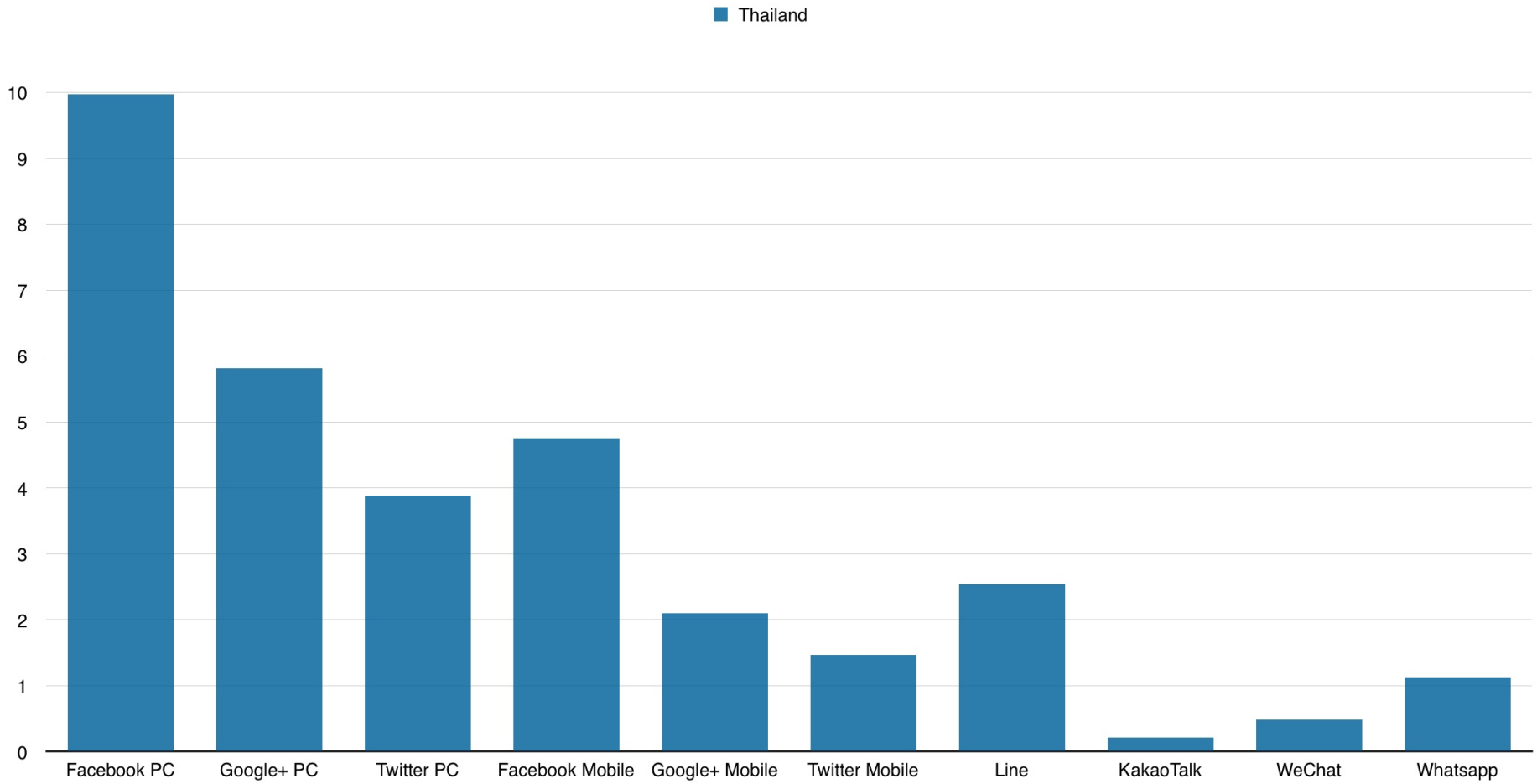
Virach Sornlertlamvanich
virach@gmail.com
Kobkrit Viriyayudhakorn
kobkrit@gmail.com

EJC2014 Panel Session, June 3-6, 2014, Kiel, Germany



http://tulaneict4d.wordpress.com/2013/04/05/social-media-in-thailand/

Thailand Monthly Active Users on Social Network and Messenger App in 2013



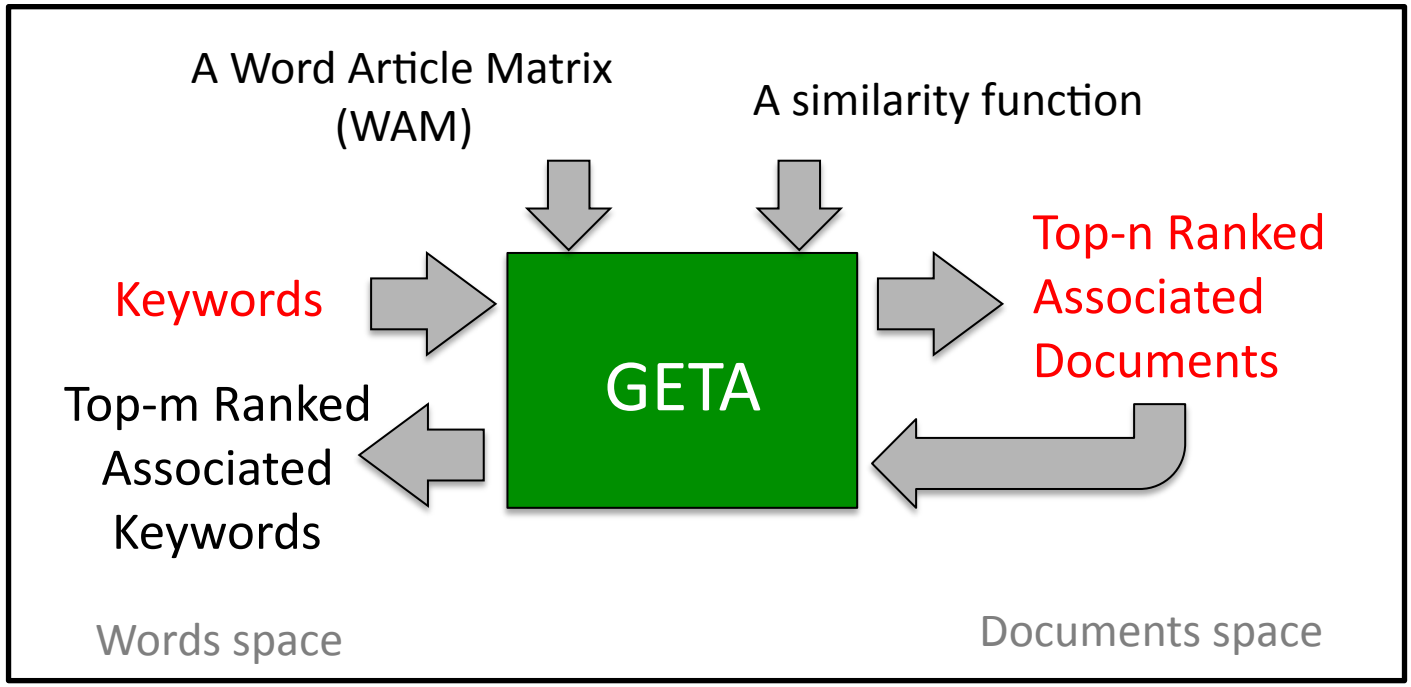
http://www.chandlernguyen.com/2013/12/state-of-social-networks-in-southeast-asia.html

Data Preparation

- Word segmentation
- Keyword
  extracted from topic related documents
  (training set)
- Tweeter inquiry
  using the prepared topic related list of keywords
- Text similarity
  using GETA algorithm

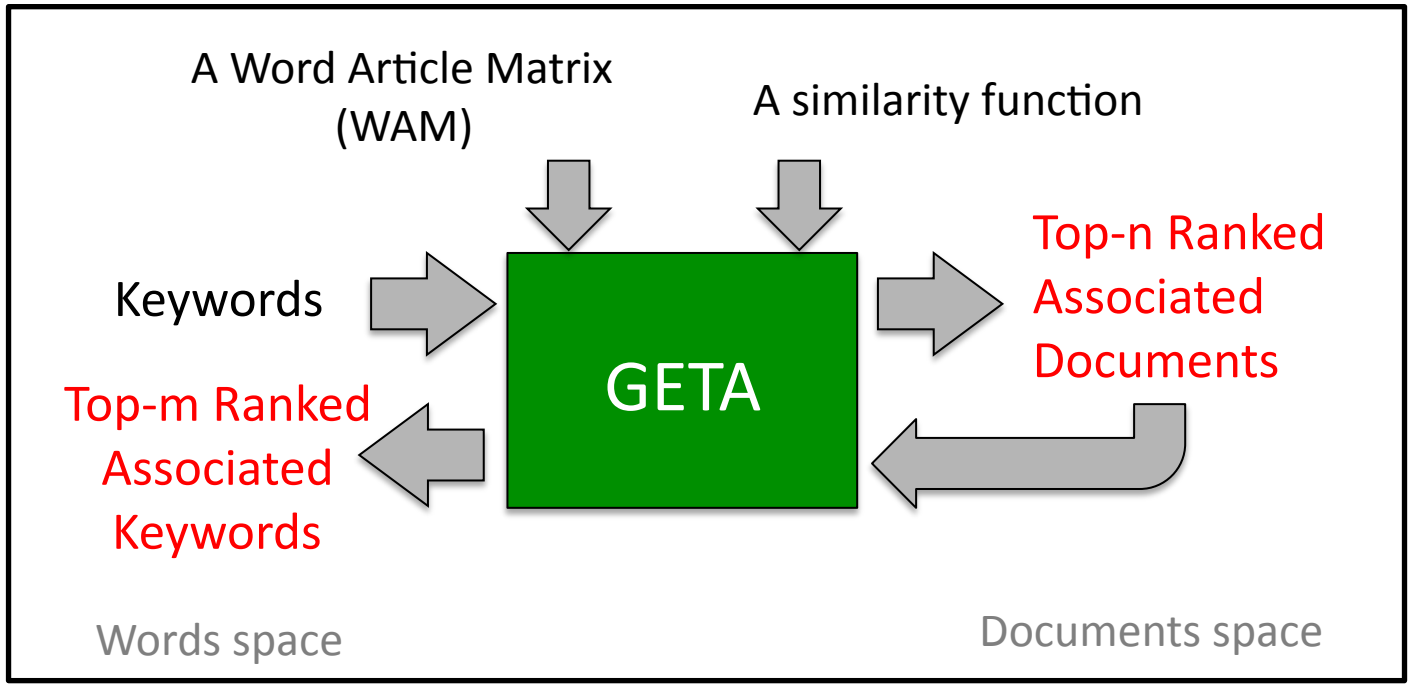
GETA: Association Search Engine

- Proposed by Akihiko Takano (Takano, 2003)
- Back-and-forth searches between words and documents spaces
  - Set of keywords ⇒ The top-n ranks of associated documents
  - Set of documents ⇒ The top-m ranks of associated keywords
- Require a **Word Article Matrix (WAM)**, and a **similarity function**



GETA: Association Search Engine

- Proposed by Akihiko Takano (Takano, 2003)
- Back-and-forth searches between words and documents spaces
  - Set of keywords ⇒ The top-n ranks of associated documents
  - Set of documents ⇒ The top-m ranks of associated keywords
- Require a **Word Article Matrix (WAM)**, and a **similarity function**



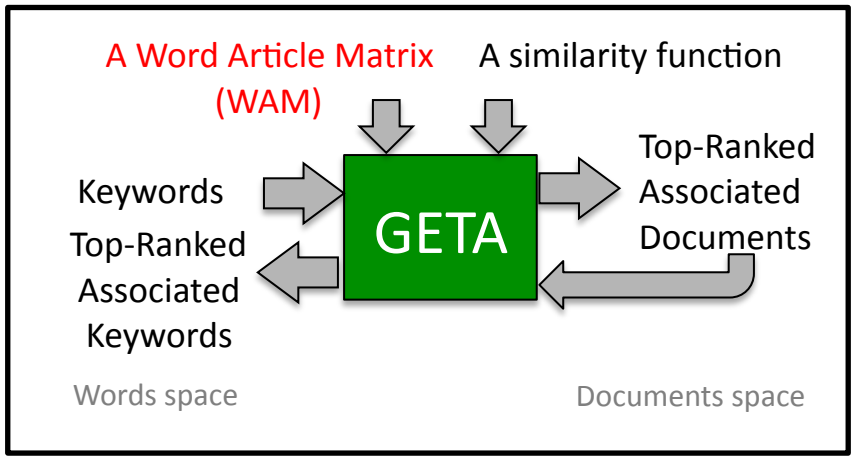
Word Article Matrix (WAM)

Building WAM is indexing

All Wikipedia Articles

All Pages' Contents

Word seg. & Lemmatization
Thinking -> Think



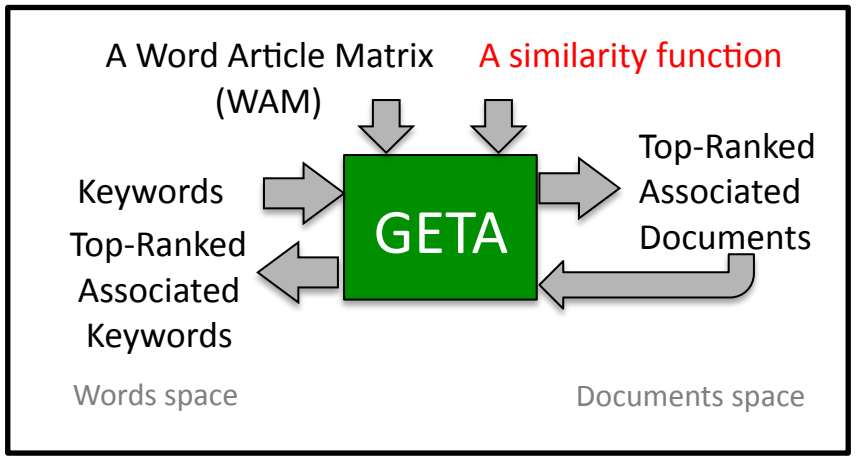
All Words List

All Pages' Titles

Pages\Words	"Twitter"	"Tennis"	"Dollar"	"Google"	…
IT	2	0	1	4	
Sport	0	2	1	0	
Economics	0	0	2	0	
…					

A Wikipedia WAM

Similarity Functions



- Ranking weight in both words and documents spaces
- Equation form

$$\mathrm{SIM}(d,q) = \sum_{t \in q} \frac{\mathrm{wq}(t,q) \cdot \mathrm{wd}(t,d)}{\mathrm{norm}(d,q)}$$

  - wd(t,d) = Weight of term t in document d, $\mathrm{wd}(t,d) = 0$ if $t \notin d$
  - wq(t,q) = Weight of term t in query sentence q, $\mathrm{wq}(t,q) = 0$ if $t \notin q$.
  - norm(d,q) = Normalization function due to the different length of d and q.
- Smart measure (Singhal et.al., 1996)

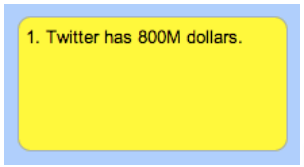
$$\frac{1}{avg(f_d) + \theta(f_d - avg(f_d))} \sum_{t \in q \wedge t \in d} log(\frac{N}{f_t}) \cdot \frac{1 + log(f_{d,t})}{1 + log(avg_{\omega \in d}(f_{d,\omega}))} \cdot \frac{1 + log(f_{q,t})}{1 + log(avg_{\omega \in q}(f_{q,\omega}))}$$

- Dot Product (Wilkinson et.al., 1996)

$$\sum_{t \in q \wedge t \in d} (w_{q,t} \cdot w_{d,t})$$

Example GETA Calculation

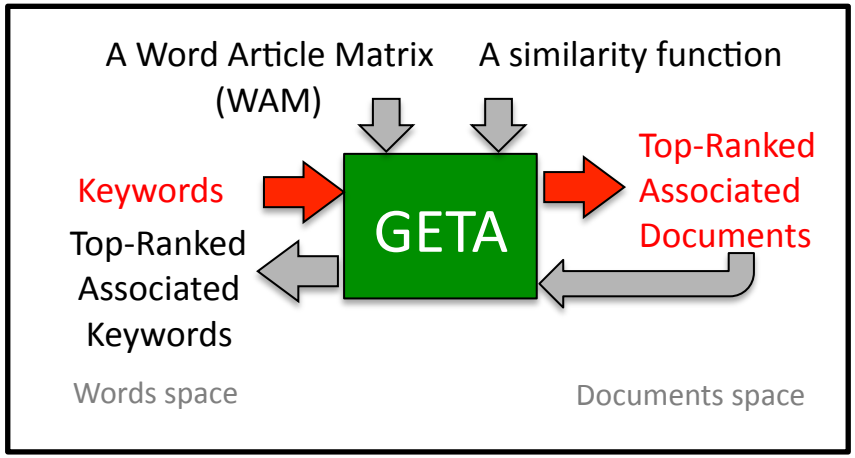
SIM Function: Dot Product $\sum_{t \in q \wedge t \in d} (w_{q,t} \cdot w_{d,t})$



A idea label
"Twitter has 800M dollars"

Word seg. & Lemmatization
dollars -> dollar

"Twitter"	"Tennis"	"Dollar"	"Google"
1	0	1	0



Pages\Words	"Twitter"	"Tennis"	"Dollar"	"Google"	...
IT	2	0	1	4	
Sport	0	2	1	0	
Economics	0	0	2	0	
...					

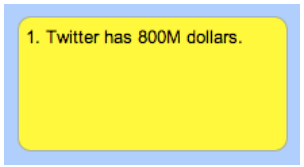
Wikipedia WAM

(n=2) Select Top-n
Most Associated Documents Ranking

Pages	Dot Product Score
IT	3
Sport	1
Economics	2
...	

Example GETA Calculation

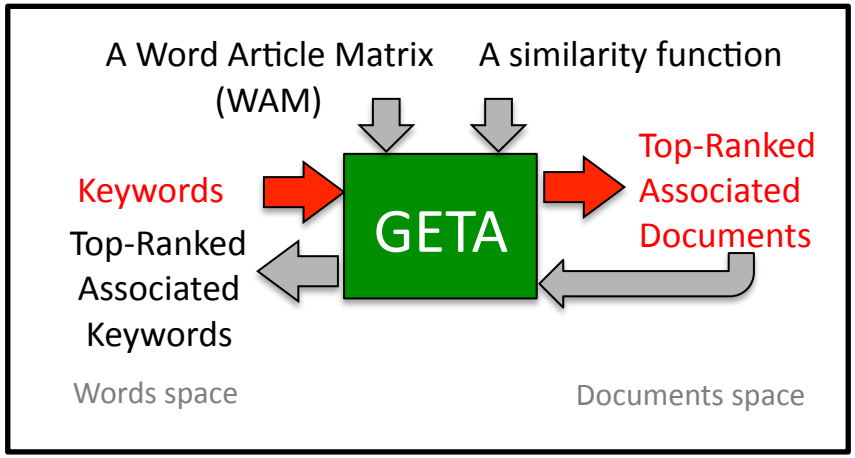
SIM Function: Dot Product $\sum_{t \in q \wedge t \in d} (w_{q,t} \cdot w_{d,t})$



A idea label
"Twitter has 800M dollars"

Word seg. & Lemmatization
dollars -> dollar

"Twitter"	"Tennis"	"Dollar"	"Google"
1	0	1	0



Pages\Words	"Twitter"	"Tennis"	"Dollar"	"Google"	...
IT	2	0	1	4	
Sport	0	2	1	0	
Economics	0	0	2	0	
...					

Wikipedia WAM

(n=2) Select Top-n
Most Associated Documents Ranking

Pages	Dot Product Score
IT	3
Sport	0
Economics	2
...	

Example GETA Calculation (Cont.)

SIM Function: Dot Product $\sum_{t \in q \wedge t \in d} (w_{q,t} \cdot w_{d,t})$

Top-2 Most Associated Documents

Pages	Dot Product Score
IT	3
Sport	0
Economics	2

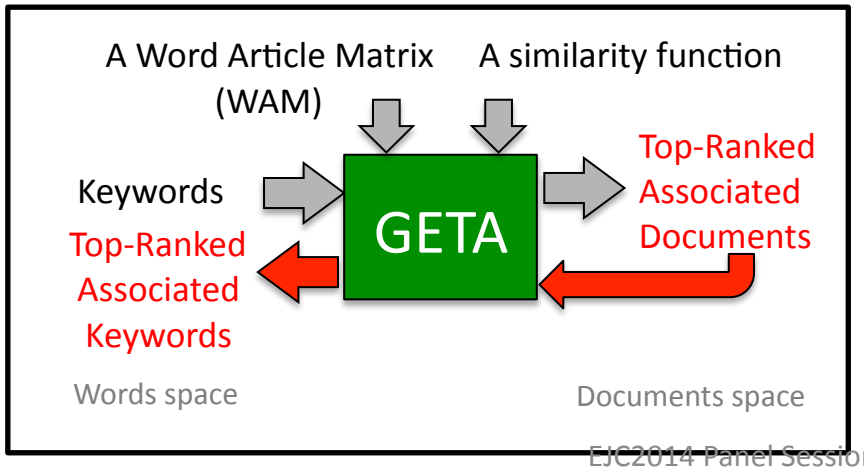
Output of GD

Pages\Words	"Twitter"	"Tennis"	"Dollar"	"Google"	...
IT	2	0	1	4	
Sport	0	2	1	0	
Economics	0	0	2	0	
...					

Wikipedia WAM

"Twitter"	"Tennis"	"Dollar"	"Google"
6	0	7	12

(m=3) Top-m Most Associated Keywords



Example GETA Calculation (Cont.)

SIM Function: Dot Product $\sum_{t \in q \wedge t \in d} (w_{q,t} \cdot w_{d,t})$

A idea label "~~Twitter~~ has 800M ~~dollars~~"

Top-2 Most Associated Documents

Pages	Dot Product Score
IT	3
Sport	0
Economics	2

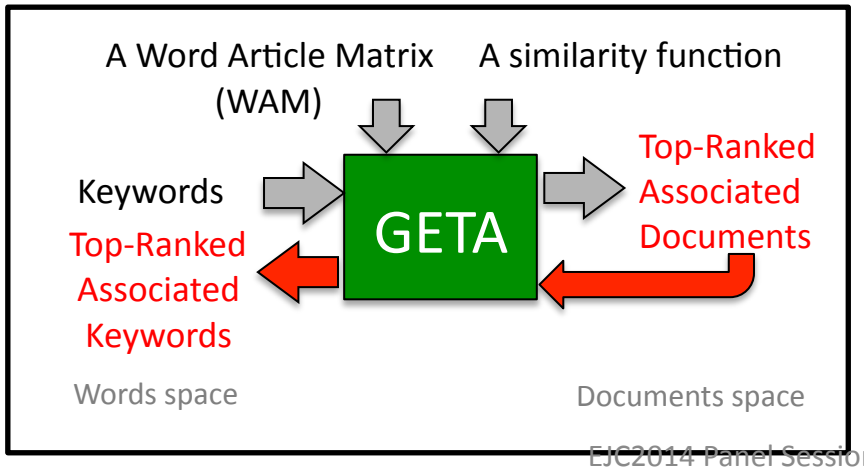
*Sport Pages = 0, since only Top-2 is selected from previous slide.

Pages\Words	"Twitter"	"Tennis"	"Dollar"	"Google"	...
IT	2	0	1	4	
Sport	0	2	1	0	
Economics	0	0	2	0	
...					

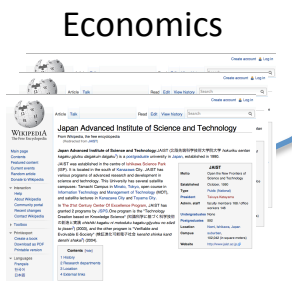
Wikipedia WAM

"Twitter"	"Tennis"	"Dollar"	"Google"
6	0	7	12

(m=3) Top-m Most Associated Keywords



Tweeter Viewer



Word segmentation

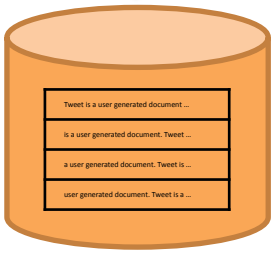
Article

Word

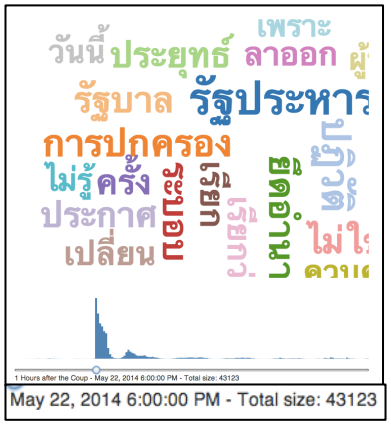
Pages \Words	"Twitter"	"Tennis"	"Dollar"	"Google"	...
IT	2	0	1	4	
Sport	0	2	1	0	
Economics	0	0	2	0	
...					

Word Article Matrix (WAM)

Tweet query by keyword



Domain specific tweets



Social movement timeline viewer

Coup on May 22, 2014

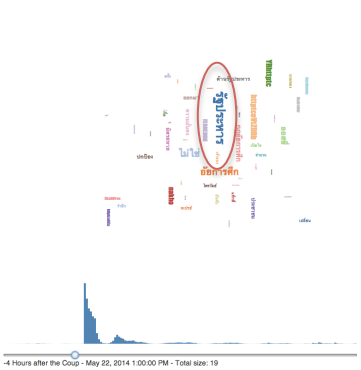
- ทหาร, คสช., ประเทศ, ประกาศ, สงบ, อำนาจ, รัฐบาล, รัฐประหาร, ชุมนุม, ตำรวจ, สถานการณ์, นายก, ควบคุม, ยึด, ประชุม, เศรษฐกิจ, กฎหมาย, ศึก, แกนนำ, รัฐมนตรี, เลือกตั้ง, ประชาธิปไตย, ปฏิวัติ. ยึดอำนาจ. เคอร์ฟิว. กฎอัยการศึก
- military, NCPO, country, announce, peace, power, government, coup d'etat, gathering, police, situation, PM, control, seize, meeting, economy, law, war, leader, minister, election, democracy, revolution, seize the power, curfew, martial law

Tweet Query

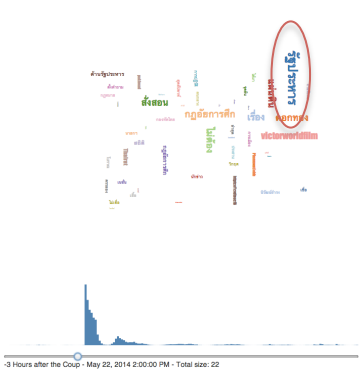
- Search Tweets by using Restful API
  - GET search/tweets. Set q = the keyword set
  - 100 tweets/search limited
  - Repeatedly fetch data until all tweets in the coup periods are discovered
- Be able to search back to 7 days

May 22, 2014 Coup-related tweet : 339,148 tweets

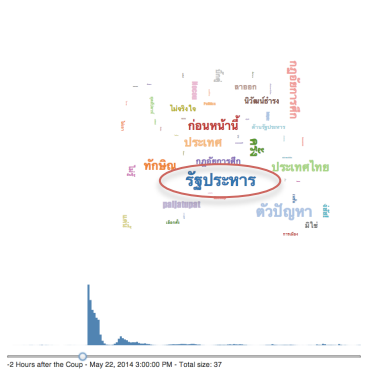
Timeline Word Cloud



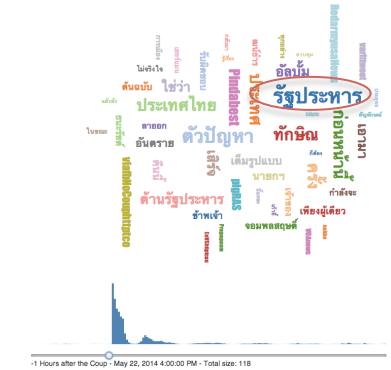
May 22, 2014 1:00:00 PM - Total size: 19



May 22, 2014 2:00:00 PM - Total size: 22

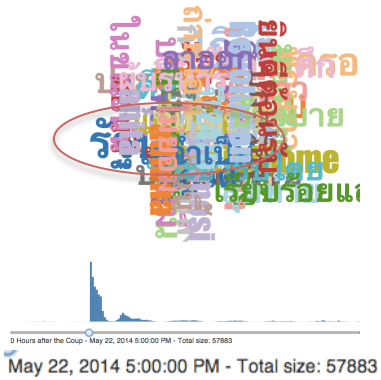


May 22, 2014 3:00:00 PM - Total size: 37

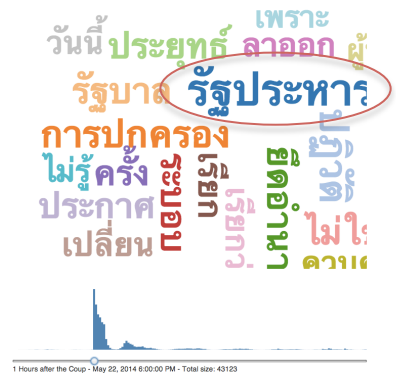


May 22, 2014 4:00:00 PM - Total size: 118

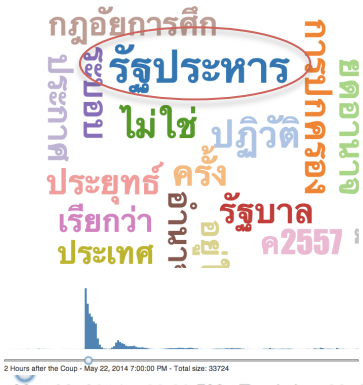
Coup D'etat



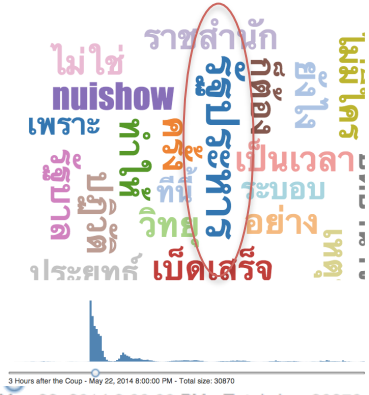
May 22, 2014 5:00:00 PM - Total size: 57883



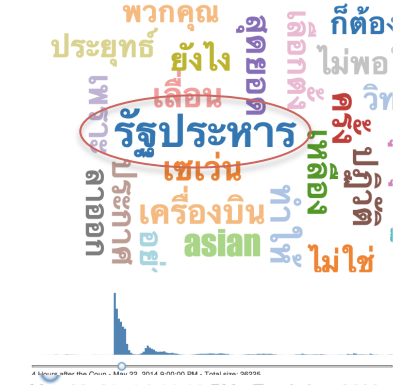
May 22, 2014 6:00:00 PM - Total size: 43123



May 22, 2014 7:00:00 PM - Total size: 33724



May 22, 2014 8:00:00 PM - Total size: 30870



May 22, 2014 9:00:00 PM - Total size: 26225

Conclusion

- Key word expansion is effective to understand the short message i.e. tweet
- Key word expansion can be done using the known training corpus and GETA algorithm
- Timeline word cloud shows the development of the social movement e.g. some events are predictable