Social Media Mining

Case Study : Event Detection in Twitter

Based on 'Event Detection in Twitter' by Jianshu Weng, Bu-Sung Lee ICWSM 2011

Presented by :
Yue HE & Falitokiniaina RABEARISON

Motivation

social events
- parties
- baseball games
- presidential campaign

disastrous events
- storms
- fires
- traffic jams
- riots
- heavy rain-falls
- earthquakes

Outline

- Related work
- Presentation of EDCoW
- Experiments and Results
- Conclusion
- Discussion

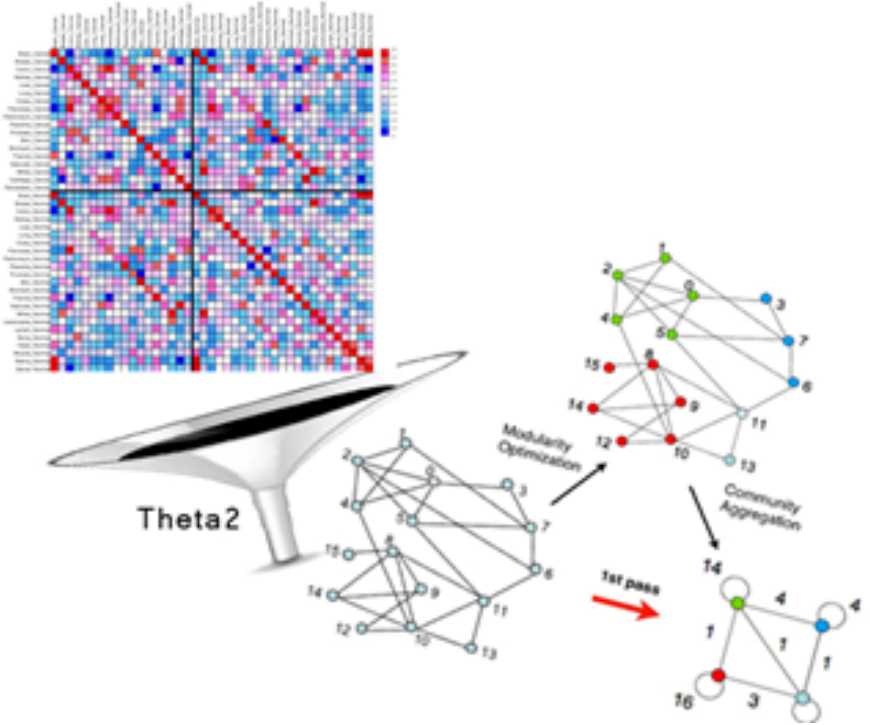
Related Work

Related Work

- Term-weighting-based approaches
- Topic-modeling-based approaches
- Clustering-based approaches

PRESENTATION of EDCoW

EDCoW [Workflow]



EDCoW [Components]

Build signal for individual words → Filter away trivial words → Cluster signals by modularity-based graph partitioning

- DF-IDF
- Wavelet transform

- Auto correlation
- Cross correlation

- Modularity of partitioning
- Newman Algorithm

$$s_w(t) = \frac{N_w(t)}{N(t)} \times \log \frac{\sum_{i=1}^{T_c} N(i)}{\sum_{i=1}^{T_c} N_w(i)}$$

$$S(t) = \sum_{j=1}^{N_J} \sum_{k} C_j(k)\psi_{j,k}(t) = \sum_{j=1}^{N_J} r_j(t).$$

$$(f \star g)(t) = \sum f * (\tau) g(t+\tau)$$

$$MAD(\mathcal{S}^{\mathcal{I}}) = \text{median}(|A_i^{\mathcal{I}} - \text{median}(A_i^{\mathcal{I}})|)$$

$$Q = \frac{1}{2m} \sum_{ij} (w_{ij} - \frac{d_i \cdot d_j}{2m}) \delta_{c_i, c_j}$$

$$\epsilon = (\sum w_{ij}^c) \times \frac{e^{1.5n}}{(2n)!},\ n = |V^c|$$

Parameters:

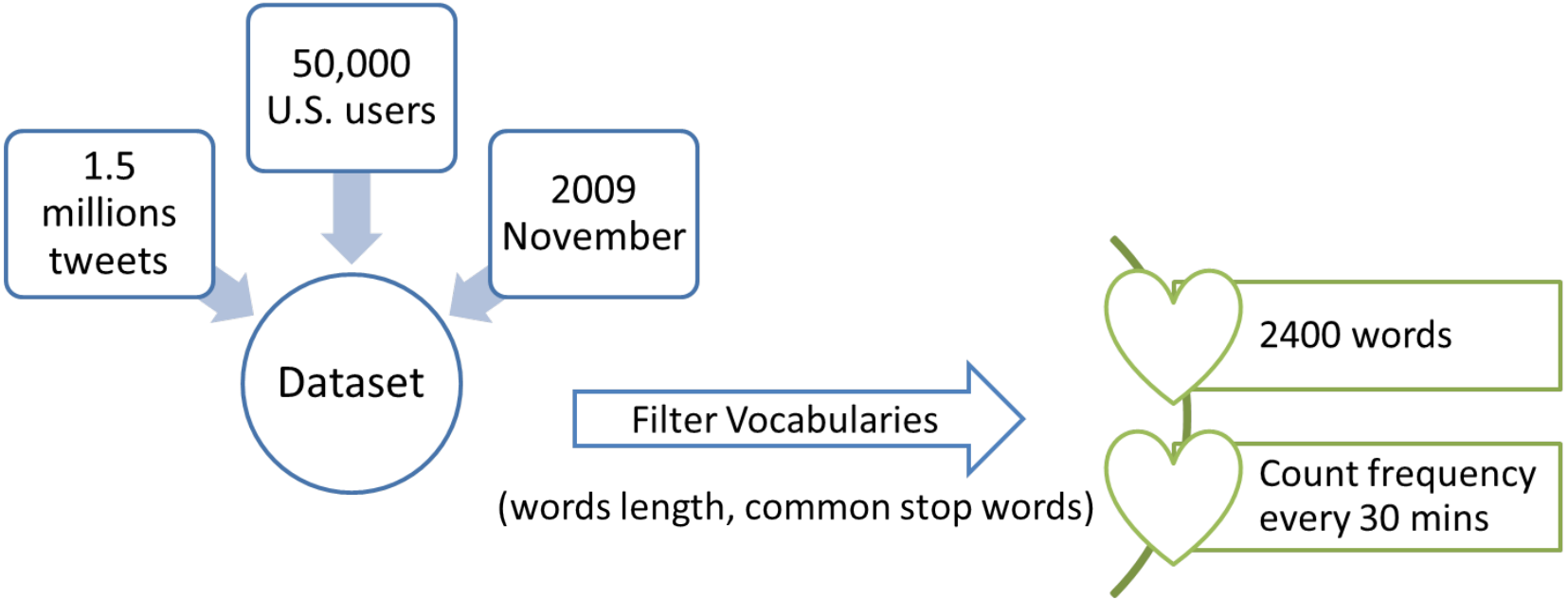
window slide: delta

threshold: gamma1
gamma2

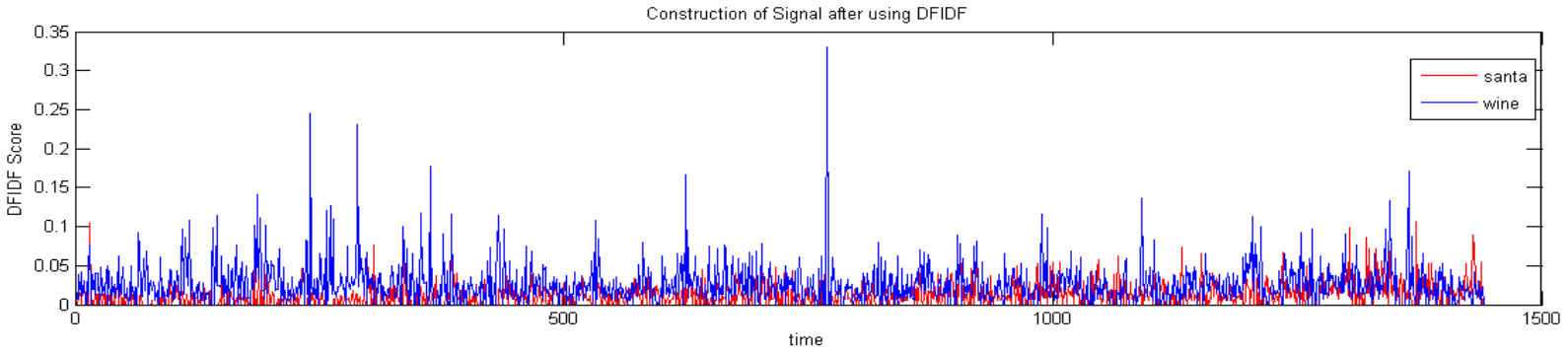
event significance: thesholdE

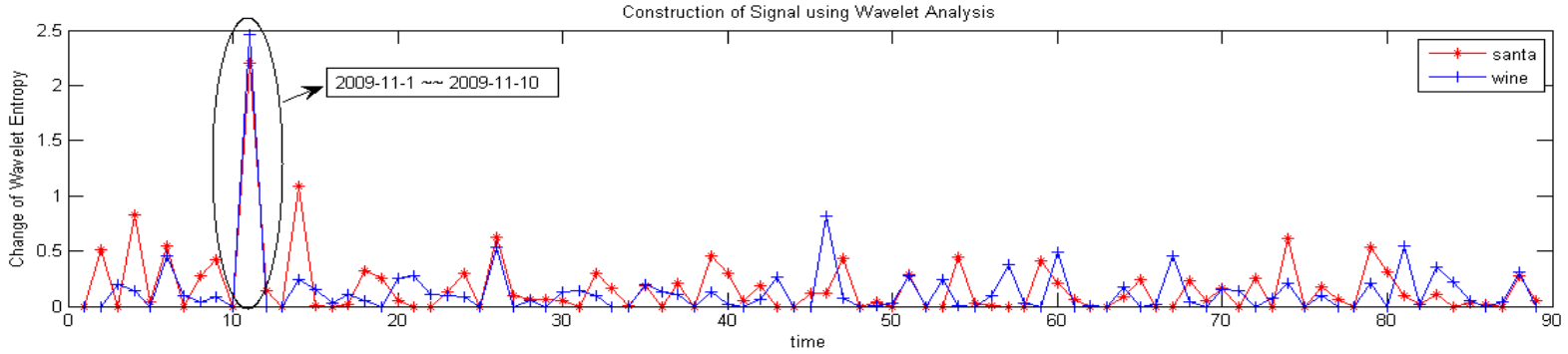
EXPERIMENTS and RESULTS

Dataset Description and Preparation



Signal Construction





“santa + wine”

Santa Rita

Pier and the sun setting over the Pacific Ocean. What a perfect backdrop on a ...

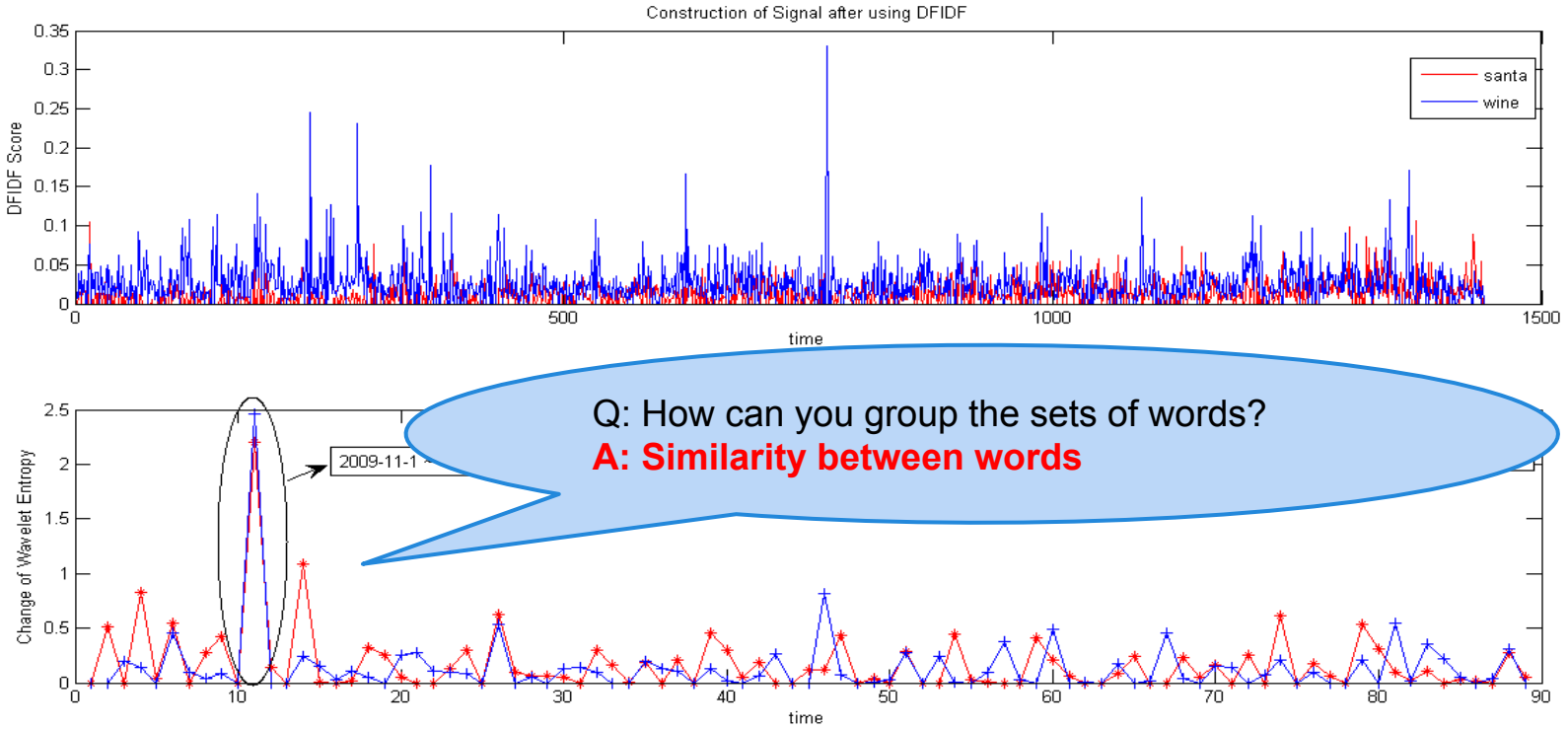
Wine Travel - The **wine** anorak
www.**wine**anorak.com/travel.htm ▾ Traduire cette page
South Africa revisited (**November** 2009 and October 2010) ... The **USA's** most famous **wine** region is in easy striking distance of San Francisco, and .... of two of the more interesting **wine** regions, **Santa** Ynez (near **Santa** Barbara) and Sonoma ...

Sanford Winery
www.sanford**wine**ry.com/ ▾ Traduire cette page
THE HEART OF THE **SANTA** RITA HILLS. May 2010 **Wine** Club Events 1 of 13
November 2009 Harvesting Pinot ... Sure, call **us** biased, but we think we can

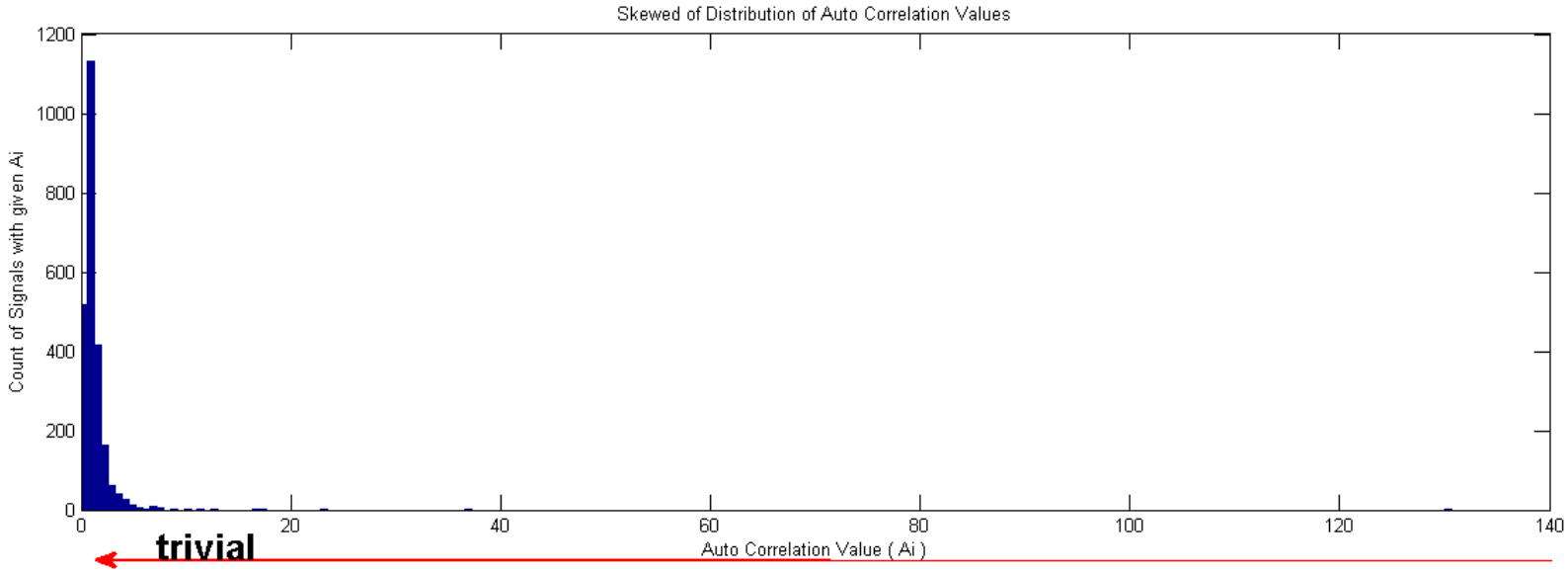
Santa Ynez **Wine** Country in Southern California | **USA** ...
www.wrsol.com/**usa**travelguide/.../**santa**-ynez-**wine**-co... ▾ Traduire cette page
19 oct. 2012 - In the shadows of the rolling **Santa** Ynez and San Rafael Mountain ranges lie a six mile patch of family owned vineyards and wineries.

November 09 Newsletter - Wandering Dog **Wine** Bar

Signal Construction

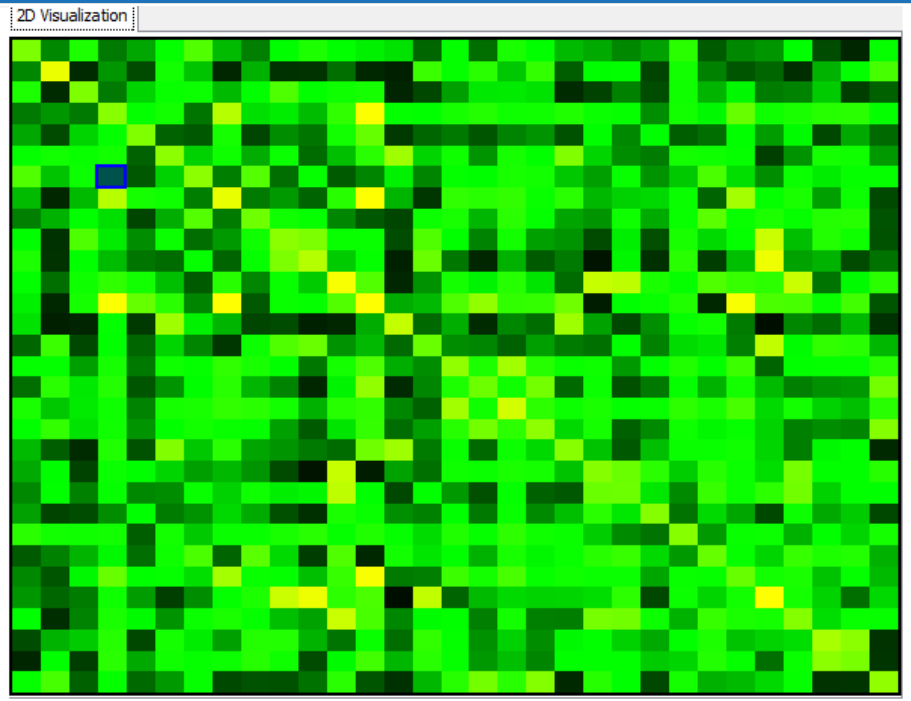


Auto Correlation Computation

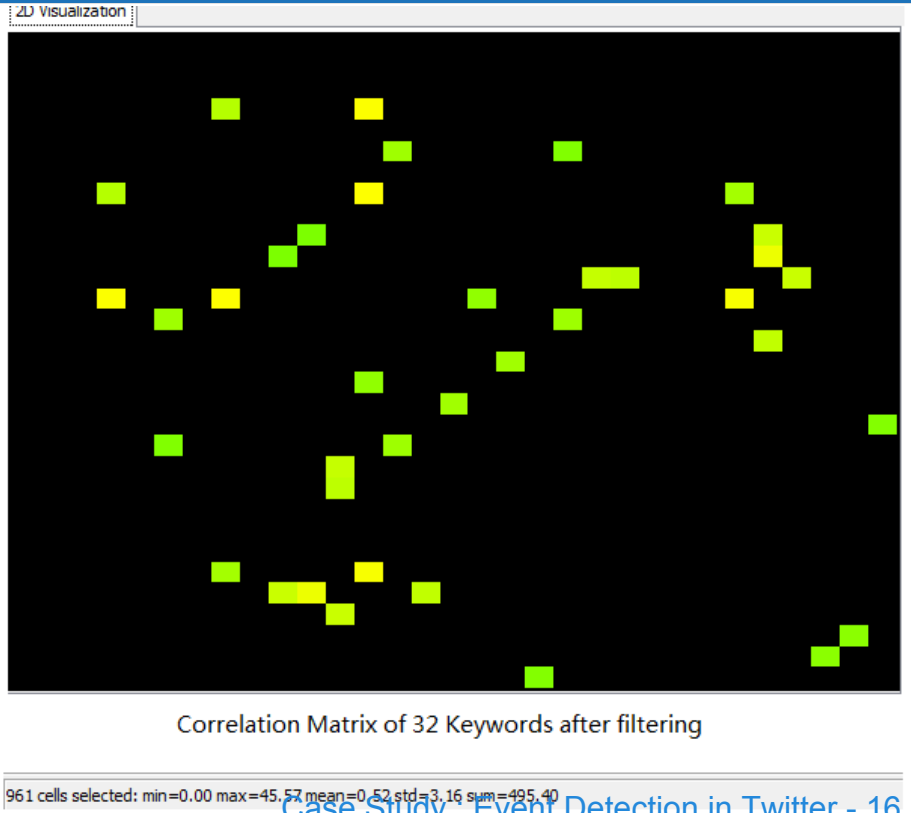


Filter: *Median Absolute Deviation*(MAD) : 2400 words → 32 keywords

Cross Correlation Computation

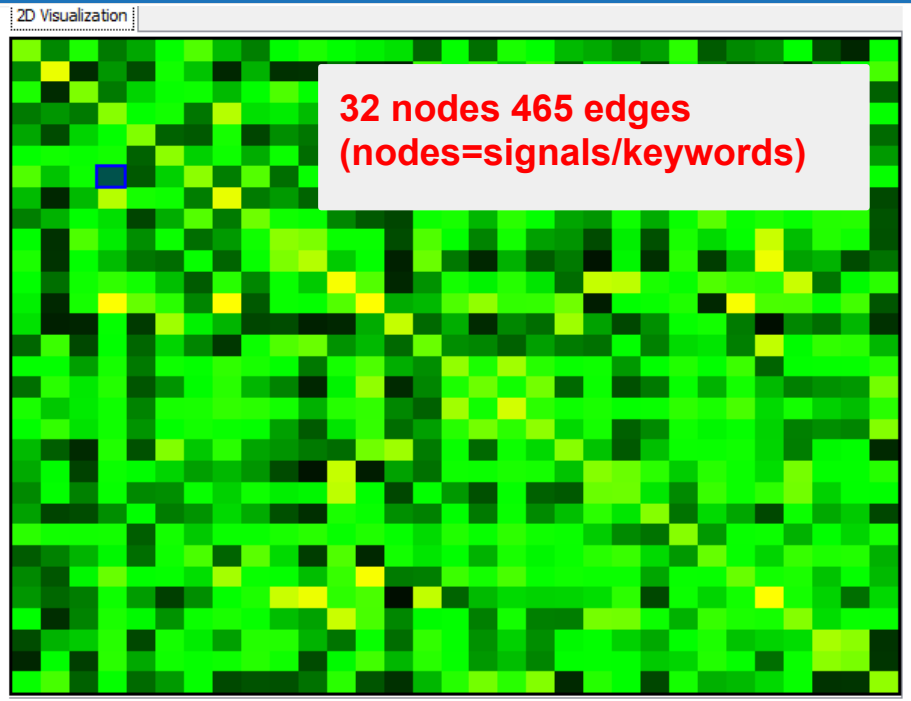


Correlation Matrix for 32 Keywords
(the lighter the color of the cell in the matrix, the higher similarity between signals)



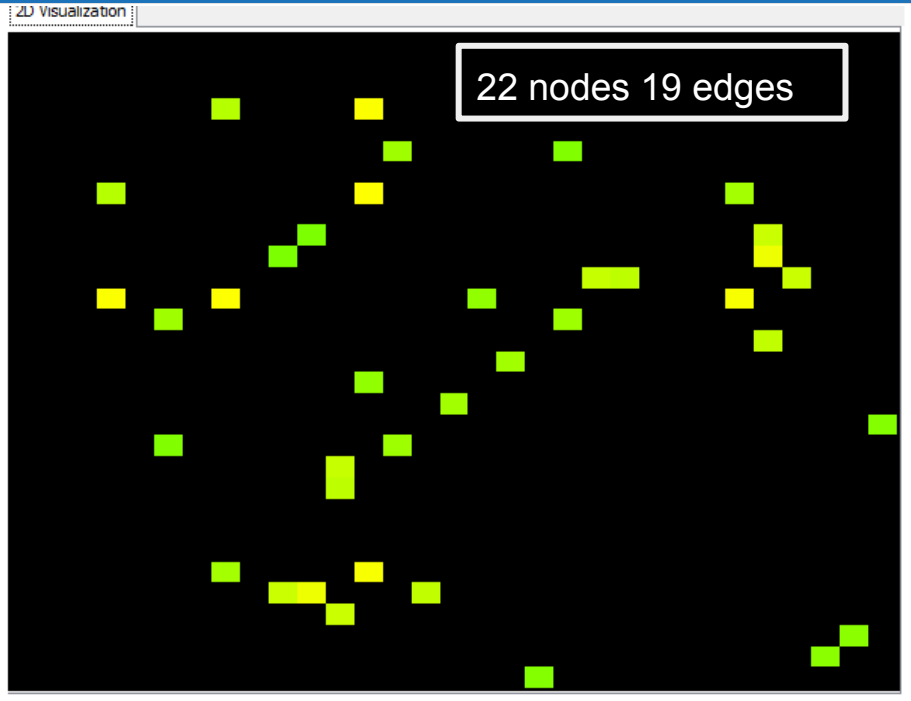
Correlation Matrix of 32 Keywords after filtering

Cross Correlation Computation



Correlation Matrix for 32 Keywords
(the lighter the color of the cell in the matrix, the higher similarity between signals)

1 cells selected: min=0.46 max=0.46 mean=0.46 std=0.00 sum=0.46



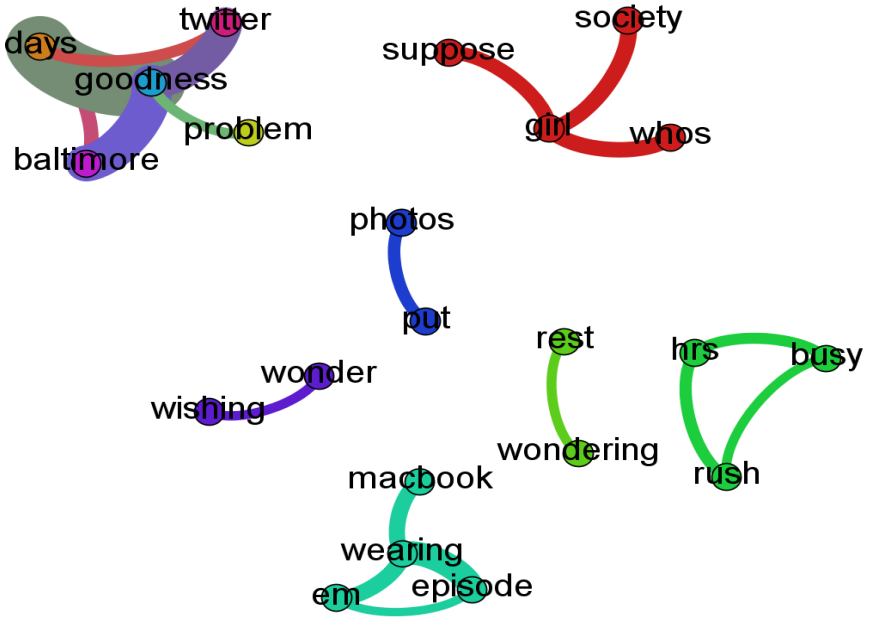
Correlation Matrix of 32 Keywords after filtering

961 cells selected: min=0.00 max=45.57 mean=0.52 std=3.16 sum=495.40

Reduction

words	after filtering by auto correlation (keywords)	after filtering by cross correlation		# of event
		node(keywords)	edge(weight)	
2400	32	20	15	8
2400	47	33	75	2
2400	31	22	19	6
2400	122	67	121	2
2400	137	15	25	2
2400	100	12	15	1

Modularity-based Graph Partition



Time Period		Significance of event	Key Words		
Begin	End				
11/21/2009	11/30/2009	7.094459382	photos	put	
		5.95984265	wishing	wonder	
		5.632178398	rest	wondering	
		2.900670455	busy	hrs	rush
		0.467600004	em	episode	macbook wearing
		0.336489143	girl	society	suppose whos

Network contains **22** nodes and **19** edges.

After using Newman Alg, we find **6** clusters(event).

The events' significance is presented by weights of the edges between keywords and the lengths of the event.

Event

Parameter:	gamma1 = 12 gamma2 = 10 thresholdE = 0.1 delta = 16				
Time	Significance of event	Keywords in Event			
01/11/09	0.905574323	diet	quiet	dream	
02/11/09	0.140348795	bug	house	meet	program
04/11/09	8.843723289	bible	build		
05/11/09	0.138681618	person	stock	video	win
07/11/09	3.942605774	biggest	page		
09/11/09	0.806623543	author	box	flight	
	0.783191902	public	strategy	tried	
10/11/09	18.03409289	hill	table	youve	
12/11/09	4.911412701	kick	yahoo		
14/11/09	0.275774247	blue	flight	ice	nope
15/11/09	5.032902893	flickr	usa		
16/11/09	2.226469888	sleep	una		
17/11/09	7.78609984	freedom	spirit		
18/11/09	1.467770626	latest	risk	streaming	
22/11/09	1.217226836	baby	lights	page	quote
23/11/09	4.582530554	deserve	reason	set	
	0.264357552	creative	dead	easier	ratings
27/11/09	1.278909451	episode	imagine	small	
28/11/09	0.195049121	chair	strong	thought	tuesday
29/11/09	2.761999223	coach	effect	matter	
	0.168426699	developer	gold	ha	lots

Detect Event

Parameter:	gamma1 = 12 gamma2 = 10	thresholdE = 0.1	delta = 16		
Time Period		Significance of event	Keywords in Event			
Begin	End					
01/11/09	10/11/09	10.54879631	tonight	website		
		9.569440935	central	kick		
		5.425693084	santa	wine		
		5.238103263	simply	upgrade		
		5.006220337	hospital	reading		
		4.146540539	development	las	usa	
		4.00962903	11	hour	town	
		0.425628924	bring	gym	kim	standing
11/11/09	20/11/09	10.69636029	home	personal		
		7.428329551	ice	broke		
21/11/09	30/11/09	7.094459382	photos	put		
		5.95984265	wishing	wonder		
		5.632178398	rest	wondering		
		2.900670455	busy	hrs	rush	
		0.467600004	em	episode	macbook	wearing
		0.336489143	girl	society	suppose	whos

Detect Event

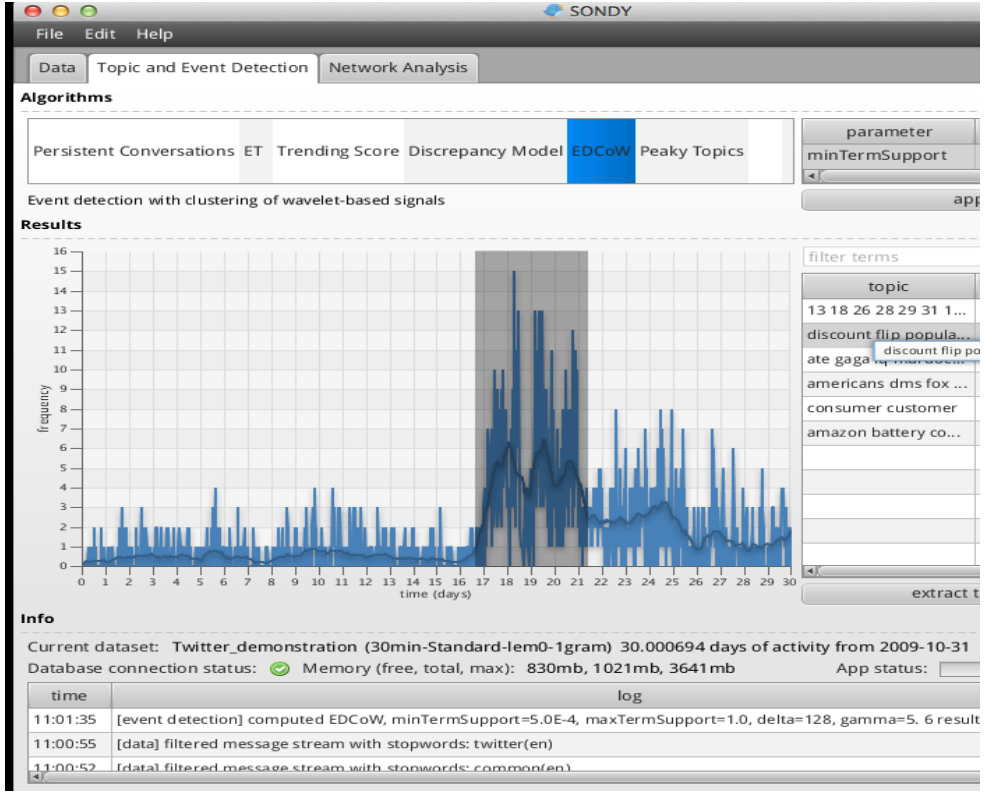
Parameter:	gamma1 = 12 gamma2 = 10	
Time Period		Significance of event
Begin	End	
01/11/09	10/11/09	10.54879631
		9.569440935
		5.425693084
		5.238103263
		5.006220337
		4.146540539
		4.00962903
		0.425628924
11/11/09	20/11/09	10.69636029
		7.428329551
21/11/09	30/11/09	7.094459382
		5.95984265
		5.632178398
		2.900670455
		0.467600004
		0.336489143

CONCLUSION

Conclusion

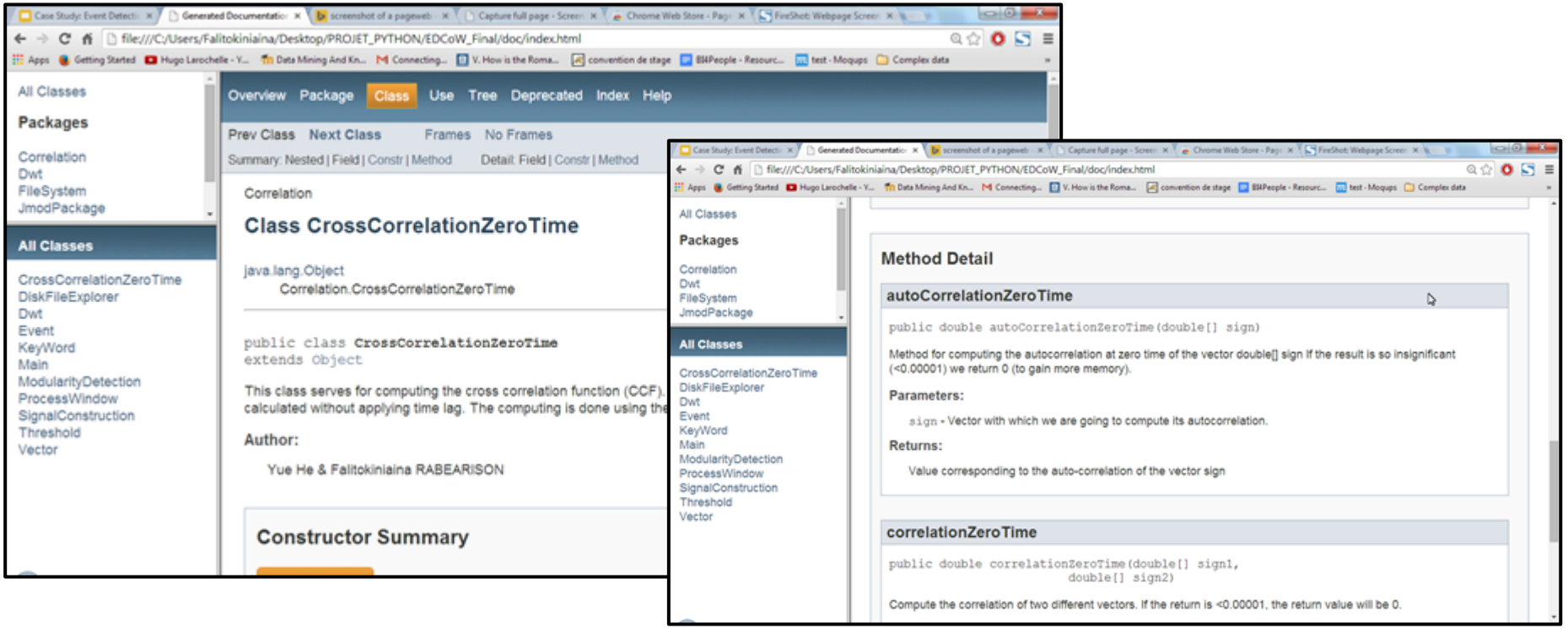
- After studying and implementing the Event Detection with Clustering of Wavelet-based Signals Algorithm in Java, we find it works in different time period. We also try to use different parameters to evaluate it on the twitter dataset.
- Some places to improve:
- How to control the parameters?
- How to let the keywords(events) make sense?
- It need some background to translate the combined keywords into event.
- How to utilize the results from public opinion?

Integrate EDCoW into SONDY



Thanks Adrien :)

EDCoW [Javadoc]



Case Study -- Detect Event in Social Network (20/03/2014 -- 18/06/2014)

Task Schedule

ID	Meeting Time	Target Task	What we did
1	20/03/2014	- present the paper EDCoW - the workload - the work flow - the deadlines - how you will be evaluated	- present the paper EDCoW(1/2) - devide the case study into 3 parts [1. study the EDCoW 2. implement the EDCoW in the platform of SONDY 3.critical analysis of EDCoW]
2	21/03/2014	Learning Basic Wavelet Knowledge from Jairo	- Wavelet Families - CWT - DWT[wavelet expression, quantities, how to measure the similiarity between signals--> shannon wavelet entropy]
3	3/4/2014	- finish the presentation - implement DWT [comupute coeffiecient in each scale/the shannon wavelet entropy in matlab]	- finish the presentation - Implement DWT in Java and MATLAB - Compare the results in different software - use java for the further work - criteria of grading : * Provide clean code * Documents for code
4	11/4/2014	- make the code easy to read - code till SWE function with : input :words signals - do some experiments in real data - send the code on Thursday 10-04-2014 - build function in MATLAB to plot data (Technical meeting for real data / plot some data)	- make the code easy to read - code till SWE function with : input :words signals output: construction of signal in the first step and second step - build function in MATLAB to plot data
5	17/4/2014	- check the devision by 0 (in the first step of construction signal) - change log_e into log_2 - write the load code * aims: load big file with all the keywords, output the result in one file - analysis the real data from pro, try to explain the result (shown in the matlab) - make sure the code to be correct	everything is done
6		- the level number should not be the input - change the wavelet family - change the deltha(the size of sliding window) - find the lib in Java to comput the Cross Correlation	everything is done
7	25/04-2014	- play with the parametrs - [modify the code to generate files for gephi input] - [play with gephi] - screenshots (signals transform, bar of autocorrelation, matrix of crosscorrelation, clusters of the graph) - import Jmod to do the cluster	everything is done
8	9/5/2014	implementation : clean : not write on harddrive [ok] flag : debug true/false [ok] epsylon : number(18)[ok] put 0 instead of 0.000000001 (using JMT) sparse vector, matrix [ok] return (sum<0.00001)?0:sum; not put the tsv in the hard [ok] time window parameter : one day size of the window	everything is done
9	5/2014 -- 6/2014	clean the code; correct mistakes; finish the documentation; finish the slide for presentation	

Acknowledge

Thank *Jairo Cugliari* & *Adrien Guille* for your guides, comments and discussions!

DISCUSSION

Thanks for your attention! ;)