ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

An end-to-end framework for event detection in social media

Nguyen Thanh Tam, Daniel Gatica-Perez

CSM, June 2016

Event Detection

Event Detection: identification of items and observations that do not conform to an expected patterns or other observations.

Different types of events:

Natural disasters

Social activities

Political campaigns

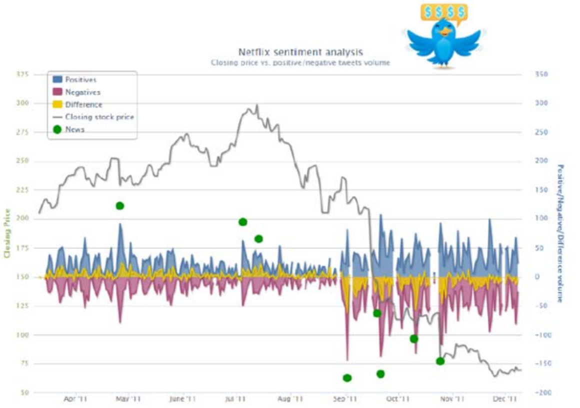
Incidents

Event Detection in Social Media

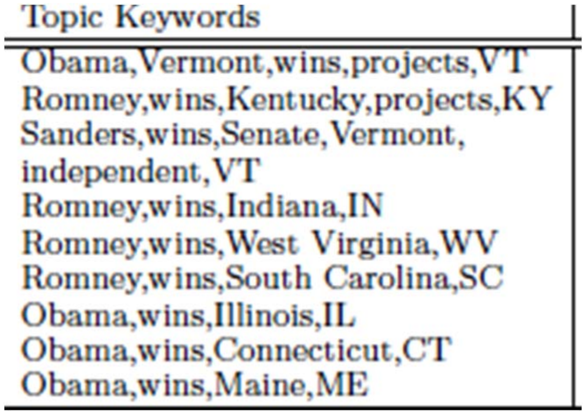
Social Media: a lot of human texts, pictures, videos

Applications of event detection:

- Predict future outcomes
- Understand a known event
- Early warning (e.g. Twitter user is faster than a BBC reporter)

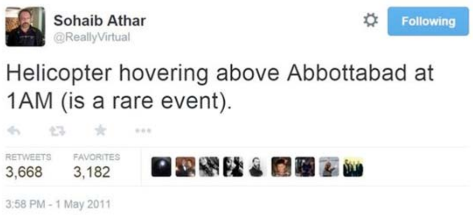


Stock price vs. tweet sentiment



Topic Keywords
Obama,Vermont,wins,projects,VT
Romney,wins,Kentucky,projects,KY
Sanders,wins,Senate,Vermont, independent,VT
Romney,wins,Indiana,IN
Romney,wins,West Virginia,WV
Romney,wins,South Carolina,SC
Obama,wins,Illinois,IL
Obama,wins,Connecticut,CT
Obama,wins,Maine,ME

Top 10 topics in US 2012 election



Sohaib Athar
@ReallyVirtual
Following

Helicopter hovering above Abbottabad at 1AM (is a rare event).

RETWEETS 3,668 FAVORITES 3,182

3:58 PM - 1 May 2011

instant reporter

Framework for Social Media

Limitations:

- Human text has different syntactic/semantic elements (sentiment, term, topic, entity)

  → need text mining techniques

- Events are often hidden or interpreted in different ways (#followers, #burst keywords, anomalies in spatial-temporal dimensions)

  → need event detection techniques

→**Goal:** develop an end-to-end framework for event detection in social media

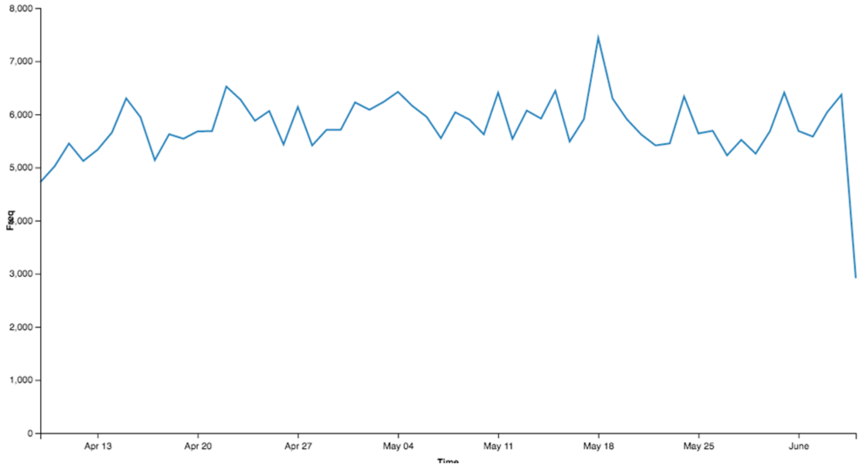
Scope:

- Target user: social media researchers for further analyses ("social good")
- Data: Tweets from Guanajuato, a touristy city in central Mexico, collected as part of the SenseCityVity project at EPFL [1]
- Focuses: ~~Data analysis~~, machine learning, visualization

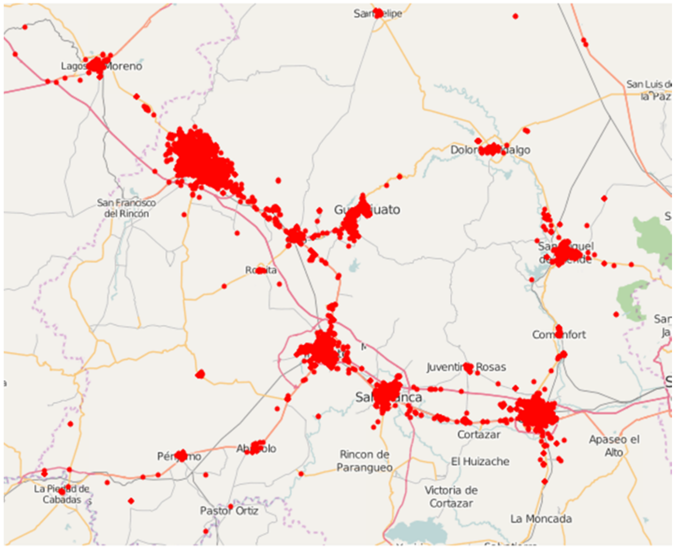
Outline

1. Preliminary Data Statistics
2. Analytical Pipeline
   - 2.1 Data Preparation
   - 2.2 Syntactical Analysis
   - 2.3 Semantic Analysis
   - 2.4 Event Detection
3. Potential Analyses

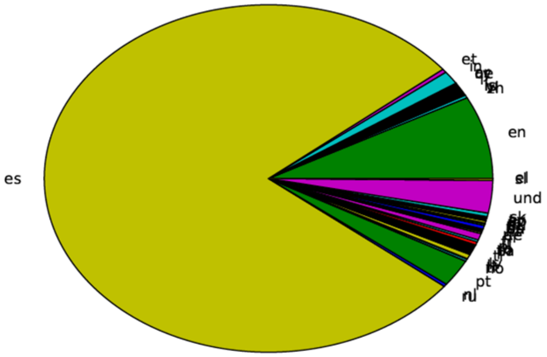
1. Preliminary Data Statistics



Tweet Count (~6K/day)
09/04/2014 -> 05/06/2014

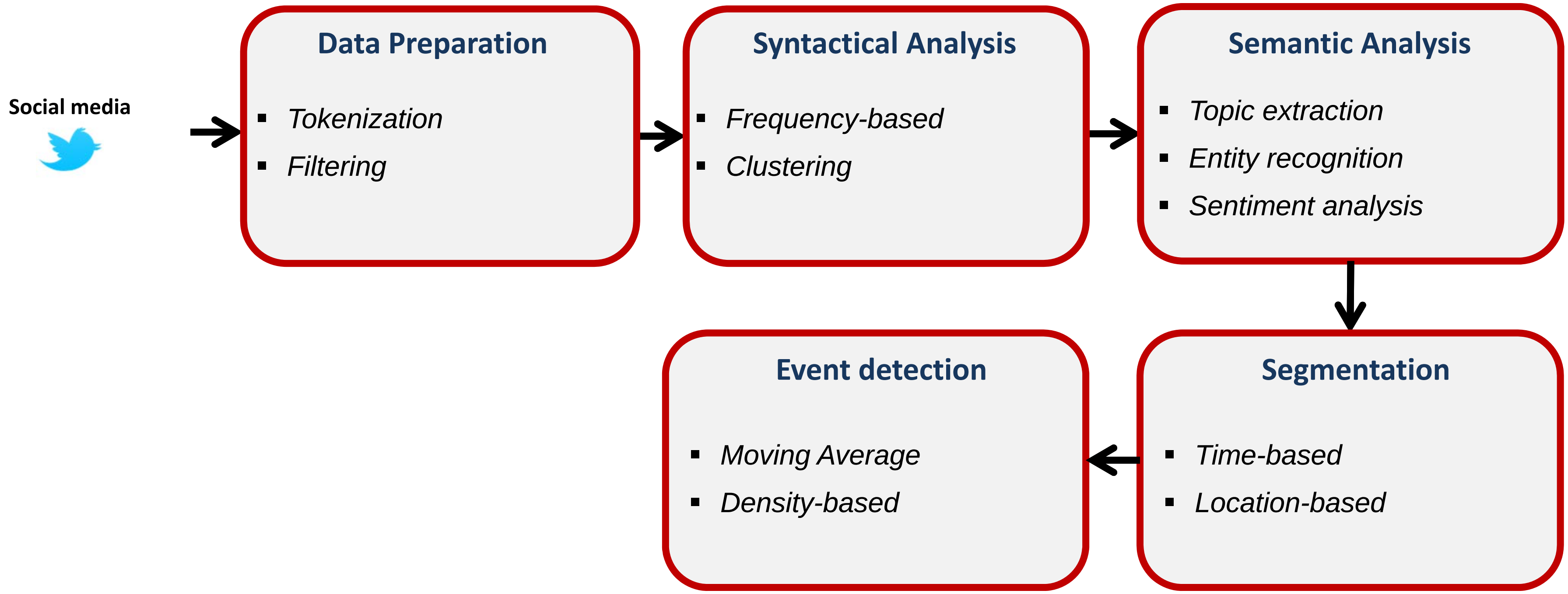


Geographic Distribution



Language Distribution (en = 7.7%, es = 78.52%, total=334836)

2. Analytical Pipeline



Social media

Data Preparation

- *Tokenization*
- *Filtering*

Syntactical Analysis

- *Frequency-based*
- *Clustering*

Semantic Analysis

- *Topic extraction*
- *Entity recognition*
- *Sentiment analysis*

Segmentation

- *Time-based*
- *Location-based*

Event detection

- *Moving Average*
- *Density-based*

2.1. Data Preparation

Tokenization: use regular expression to segment the text into tokens

- Emoticons: eyes [:=;], nose [oO\-]?, mouth [D\)\]\(\]/\\OpP]
- HTML tags: <[^>]+>
- @-mentions: (?:@[\w_]+)
- #hashtags: (?:\#+[\w_]+[\w\'_\-]*[\w_]+)
- #numbers: (?:(?:\d+,?)+(?:\.?\d+)?)
- #words with - and ‘: (?:[a-z][a-z'\-_]+[a-z])
- #other words: (?:[\w_]+)
- #anything else: (?:\S)

Filtering: filter or construct important terms from tokens

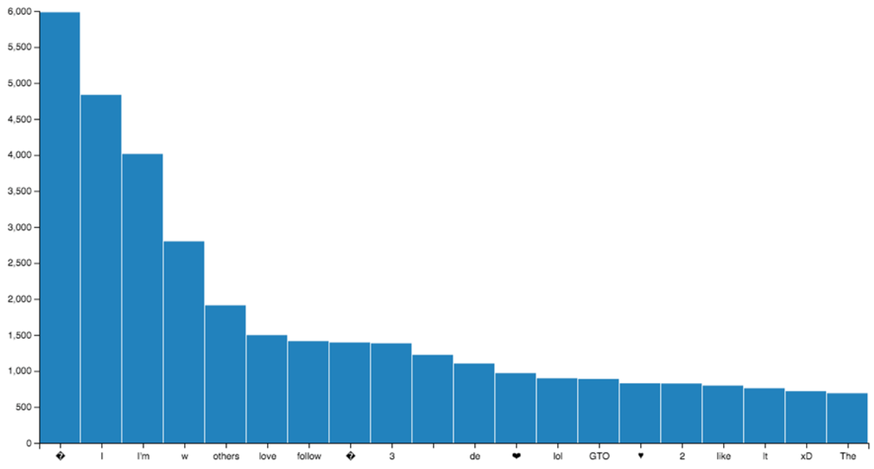
- stop words (NLTK English corpus), punctuation, special phrases (‘rt’, ‘via’)
- n-grams (NLTK library): keep phrases of 2,3, etc. words
- hashtag_only: keep only hashtag

```
terms = terms_nolink(terms_stop(terms_only(tokens)))
```

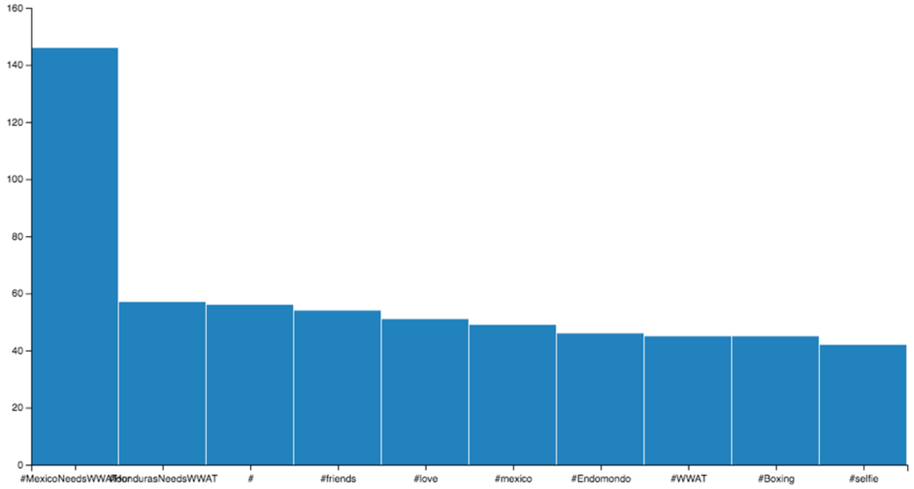
Data Preparation: Example

Input	Output
@KunderaQuotes you're my favorite writer ❤	you're my favorite writer
Everything that kills me makes me feel alive http://t.co/ifdqOkCPou	Everything kills makes feel alive
My queen looks incredible, love her more than anything #katyperry #iheartradio @ Reality http://t.co/jrxEiHxMOa	My queen looks incredible love anything

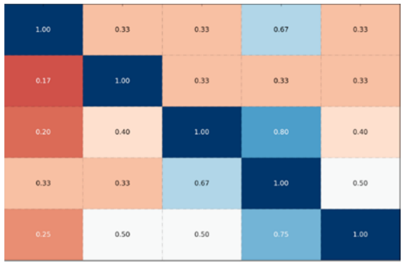
2.2. Syntactical Analysis



Top 20-terms (love, follow, like)



Top-10 hashtags (#WWAT, #Boxing, #Endomondo)



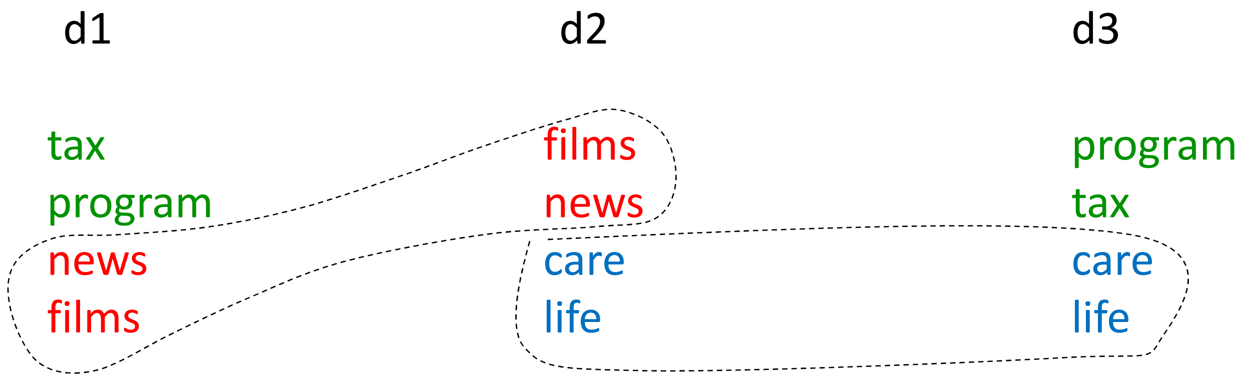
(#birthday, #katyperry)
(#friends, #selfie)
(#mexico, #sanmigueldeallende)

Co-occurrence

2.3. Semantic analysis: topic (Latent Dirichlet Allocation)

Input: tweets (bags of terms) + tf/idf (term frequency)

Output: ‘clusters’ of co-occurring words

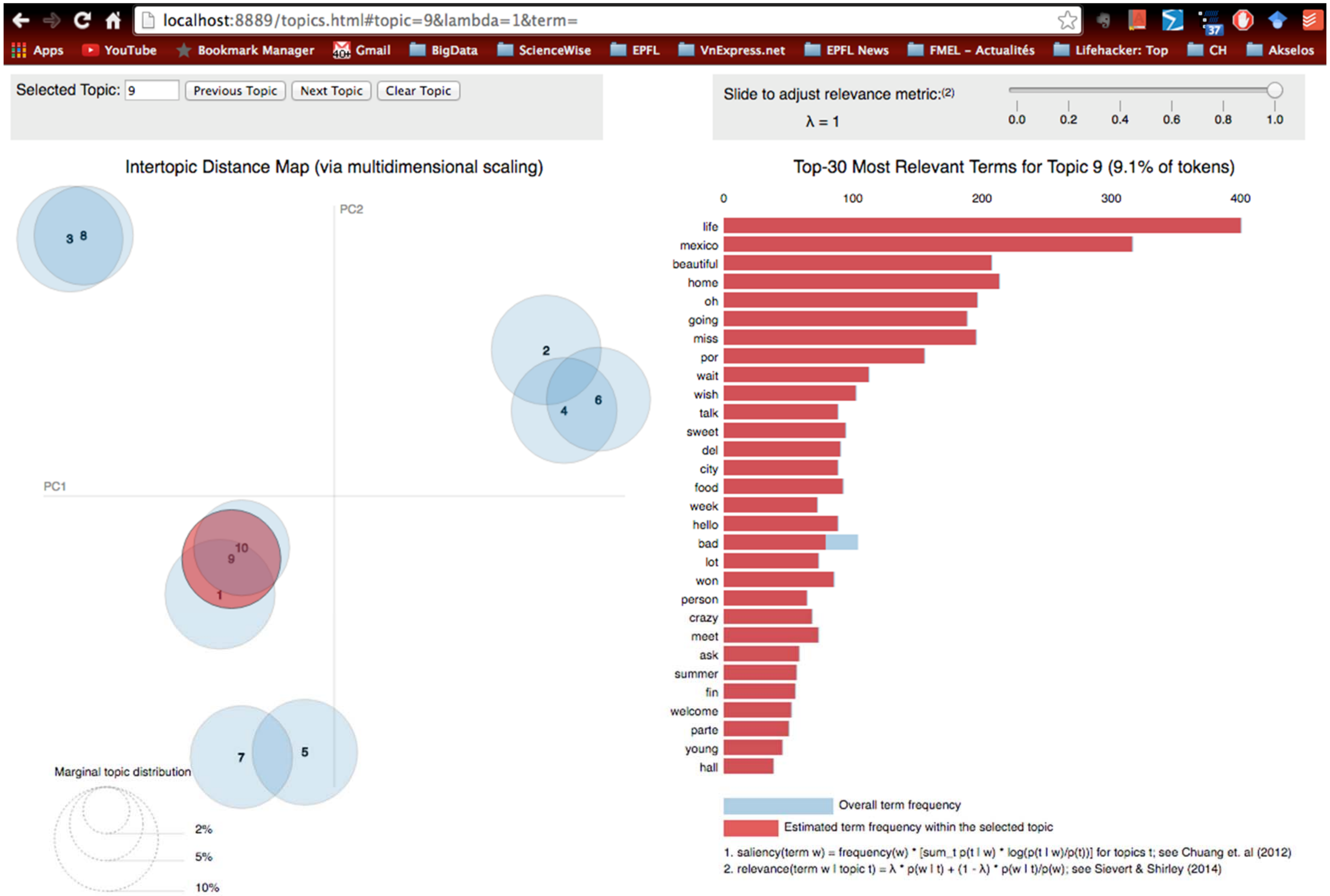


d1	d2	d3
tax	films	program
program	news	tax
news	care	care
films	life	life

Limitations:

- Sensitive to parameter K (#topics) → Hierarchical topic models [4] (non-parametric)
- Sensitive to short text → Twitter-LDA [5,6] (assumption: each tweet has 1 topic)

Topic modeling: user interface



Selected Topic: 9

Slide to adjust relevance metric:(2)

λ = 1

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Topic distance: Jensen-Shannon divergence with multidimensional downscale to 2 [10]

Semantic analysis: sentiment

Polarity (semantic orientation): user opinion in the text (positive, neutral, negative)

- Simple unsupervised method: co-occurrences with pre-defined positive/negative words (English lexicon [2])
- Supervised method: Python NLTK, Stanford classifiers

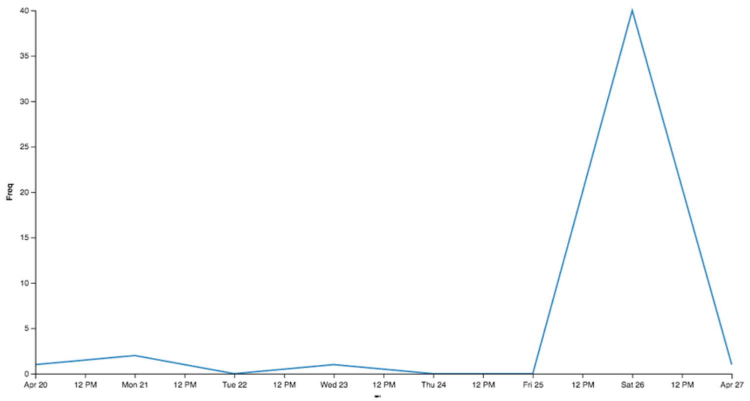
Term	Polarity
birthday	80.34
goodness	53.35
photograph	0
forecast	0
cigarettes	-40.39
bitch	-114.36

Polarity	#Documents
positive	1894
negative	8279
neutral	15617

2.4. Event Detection

Event: different definitions

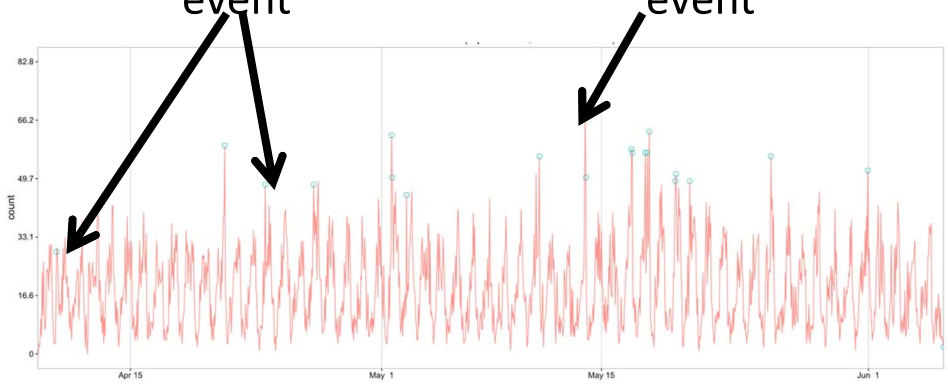
- Time-based: burst period of #topics
  - Manual observation
  - Automatic techniques [3]: moving average, box-and-whisker



Hashtag ‘#WWAT’

Local event

Global event

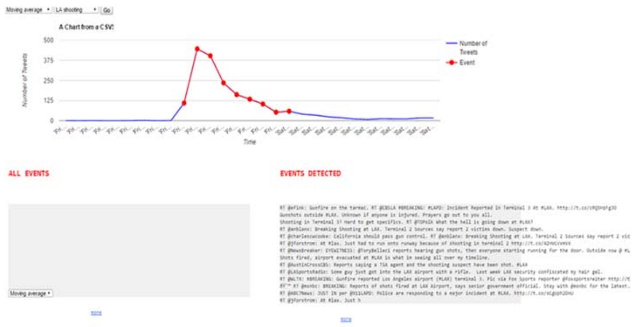


#Topics per tweet

Interactive UI

- Choose different algorithms
- Click and observe important tweets of that event

Moving Average

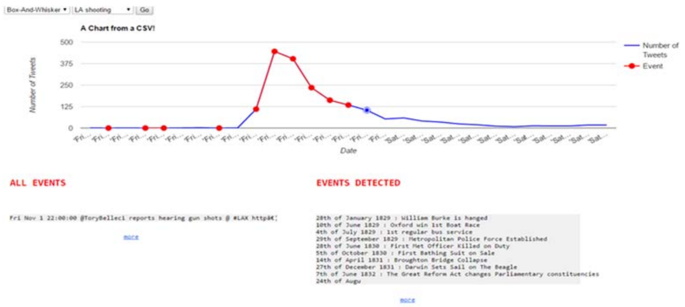


ALL EVENTS

EVENTS DETECTED

```
RT @efink: Gunfire on the tarmac. RT @CBSLA #BREAKING: #LAPD: Incident Reported In Terminal 3 At #LAX. http://t.co/cRQ5HbFgJ0
Gunshots outside #LAX. Unknown if anyone is injured. Prayers go out to you all.
Shooting in Terminal 3? Hard to get specifics. RT @TOPolk what the hell is going down at #LAX?
RT @anblanx: Breaking Shooting at LAX. Terminal 2 Sources say report 2 victims down. Suspect down.
RT @charlescwcooke: California should pass gun control. RT @anblanx: Breaking Shooting at LAX. Terminal 2 Sources say report 2 vic
RT @jforstrom: At #lax. Just had to run onto runway because of shooting in terminal 2 http://t.co/AZHACzvHVX
RT @NewsBreaker: EYEWITNESS: @ToryBelleci reports hearing gun shots, then everyone starting running for the door. Outside now @ #L
Shots fired, airport evacuated at #LAX is what im seeing all over my timeline.
RT @AustinCrossCBS: Reports saying a TSA agent and the shooting suspect have been shot. #LAX
RT @LASportsRadio: Some guy just got into the LAX airport with a rifle.  Last week LAX security confiscated my hair gel.
RT @WLTX: #BREAKING: Gunfire reported Los Angeles airport (#LAX) terminal 3. Pic via Fox Sports reporter @foxsportsreiter http://t
â€™ RT @msnbc: BREAKING: Reports of shots fired at LAX Airport, says senior government official. Stay with @msnbc for the latest.
RT @ABC7News: JUST IN per @911LAPD: Police are responding to a major incident at #LAX. http://t.co/oCgqph2DHU
RT @jforstrom: At #lax. Just h
```

Box And Whisker



ALL EVENTS

```
Fri Nov 1 22:00:00 @ToryBelleci reports hearing gun shots @ #LAX httpâ€¦
```

EVENTS DETECTED

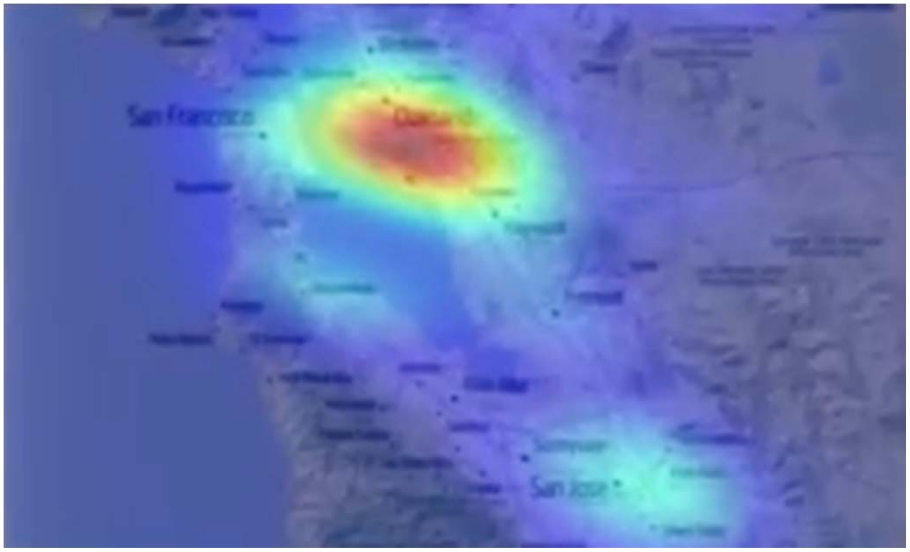
```
28th of January 1829 : William Burke is hanged
10th of June 1829 : Oxford win 1st Boat Race
4th of July 1829 : 1st regular bus service
29th of September 1829 : Metropolitan Police Force Established
28th of June 1830 : First Met Officer Killed on Duty
5th of October 1830 : First Bathing Suit on Sale
14th of April 1831 : Broughton Bridge Collapse
27th of December 1831 : Darwin Sets Sail on The Beagle
7th of June 1832 : The Great Reform Act changes Parliamentary constituencies
24th of Augu
```

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Event Detection (cont'd)

Event: different definitions

- Time-based: burst period of #topics
- Location-based: areas with high-density of #topics (tentative)
  - Step 1: Gridded topic counts (count unique tweets of given topic in an area using Twitter coordinates)
  - Step 2: Locate high-density areas



Conclusions

Take-home messages:

- An end-to-end and unified framework for event detection in social media (Twitter)
- https://github.com/tamlhp/csm

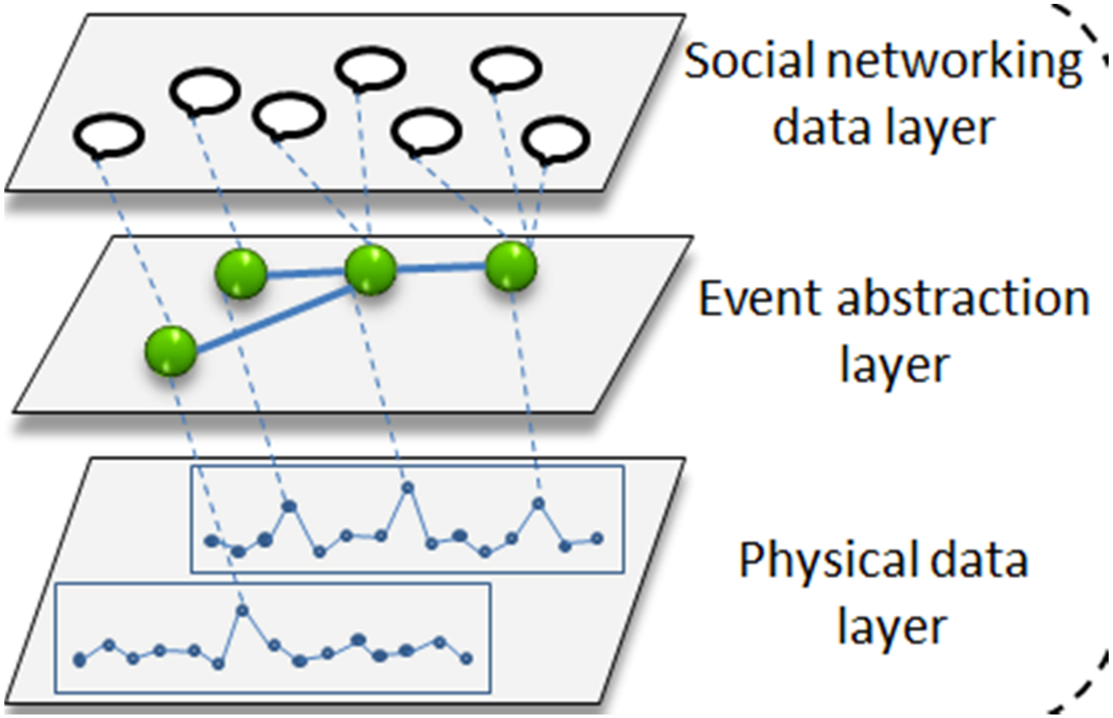
Limitations:

- Only a prototype
- English only (7.7% of Guanajuato data)

Future Work

1. Streaming version for Twitter API (online topic modeling [7], online event detection [8])
2. Complex event processing:
   - Aggregate small events into a complex event → more understanding
   - Techniques: formulation, abstraction, matching [9]

Output of our framework



References

[1] http://www.idiap.ch/project/sensecityvity/

[2] https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#lexicon

[3] http://videolectures.net/icwsm2011_lee_detection/

[4] http://www.cse.ust.hk/~lzhang/teach/6931a/slides/3.HTM.pdf

[5] http://users.cecs.anu.edu.au/~ssanner/Papers/sigir13.pdf

[6] Zhao, Wayne Xin, et al. "Comparing twitter and traditional media using topic models." *European Conference on Information Retrieval*. Springer Berlin Heidelberg, 2011.

[7] Wang, Yu, Eugene Agichtein, and Michele Benzi. "TM-LDA: efficient online modeling of latent topic transitions in social media." *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2012.

[8] Abdelhaq, Hamed, Christian Sengstock, and Michael Gertz. "Eventweet: Online localized event detection from twitter." *Proceedings of the VLDB Endowment* 6.12 (2013): 1326-1329.

[9] Cameron, Mark A., et al. "Emergency situation awareness from twitter for crisis management." *Proceedings of the 21st International Conference on World Wide Web*. ACM, 2012.

[10] https://cran.r-project.org/web/packages/LDAvis/vignettes/details.pdf

Demos

```
python -m SimpleHTTPServer 8889
```

http://localhost:8889/test.en.term_freq.html

http://localhost:8889/test.en.count.time_chart.html

http://localhost:8889/test.en.topic.time_chart.html

http://localhost:8889/test.en.term_freq.html

http://localhost:8889/test.count.time_chart.html

http://localhost:8889/test.en.geo.html

http://localhost:8889/topics.html

THANK YOU