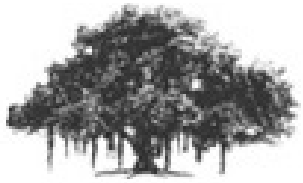


IIIT-H

Web Mining
Lecture 13: Analysis of Microblogs (Part 1): Event Detection

Manish Gupta

14th Sep 2013

Recap of Lecture 12: Social Influence Analysis (Part 2)

- Models for Social Influence Analysis
  - Decision Based Models
  - Probabilistic Models
- Influence Maximization
- Applications of Social Influence Analysis
  - Social Advertising
  - Opinion Leader Finding
  - Social Recommendation
  - Emotion Analysis

Announcements

- Course Project
  - Deadline for choice + optionally your project is Sep 14, 9pm
  - Final allocation of projects by Sep 21
  - Students should ideally start working on the project from Oct 15
  - Mid-project review deadline: Oct 27
  - Final report+presentation+code with posters will be on Nov 20
- Schedule change
  - 25th Sep class moved to 20th Sep 6-7:30pm
  - 28th Sep class moved to 30th Sep 6-7:30pm

Today's Agenda

- Event Detection in Twitter
- Generating Event Descriptions/Annotations
- Application of Event Detection from Twitter

Today's Agenda

- **Event Detection in Twitter**
- Generating Event Descriptions/Annotations
- Application of Event Detection from Twitter

Characteristics of Twitter Data

- 140 characters – short documents
- SMS kind of language
- Code mixing (mix of multiple languages)
- Tweets, Retweets, Mentions, Hashtags
- Very fresh news from human sensors
- Large amount of data with huge data rate
  - Many irrelevant messages
  - Many redundant messages
- Self-contained
- Simple discourse structure

Noisy Twitter Text: Challenges

- Lexical Variation (misspellings, abbreviations)
  - `2m', `2ma', `2mar', `2mara', `2maro', `2marrow', `2mor', `2mora', `2moro', `2morow', `2morr', `2morro', `2morrow', `2moz', `2mr', `2mro', `2mrrw', `2mrw', `2mw', `tmmrw', `tmo', `tmoro', `tmorrow', `tmoz', `tmr', `tmro', `tmrow', `tmrrow', `tmrrw', `tmrw', `tmrww', `tmw', `tomaro', `tomarow', `tomarro', `tomarrow', `tomm', `tommarow', `tommarrow', `tommoro', `tommorow', `tommorrow', `tommorw', `tommrow', `tomo', `tomolo', `tomoro', `tomorow', `tomorro', `tomorrw', `tomoz', `tomrw', `tomz‘
- Unreliable Capitalization
  - “The Hobbit has FINALLY started filming! I cannot wait!”
- Unique Grammar
  - “watchng american dad.”

NLP in News vs. Twitter: Thought Experiment

- **Task 1**
  - Read each sentence from today's New York times
  - Except, first randomly permute the sentences
  - Answer basic questions about today's news
- **Task 2**
  - Read a random sample of tweets
    - From high-quality sources
  - Order is picked randomly
  - Answer basic questions about today's news
- **Claim:**
  - Task 2 is easier than task 1.

Why Detect Events from Twitter?

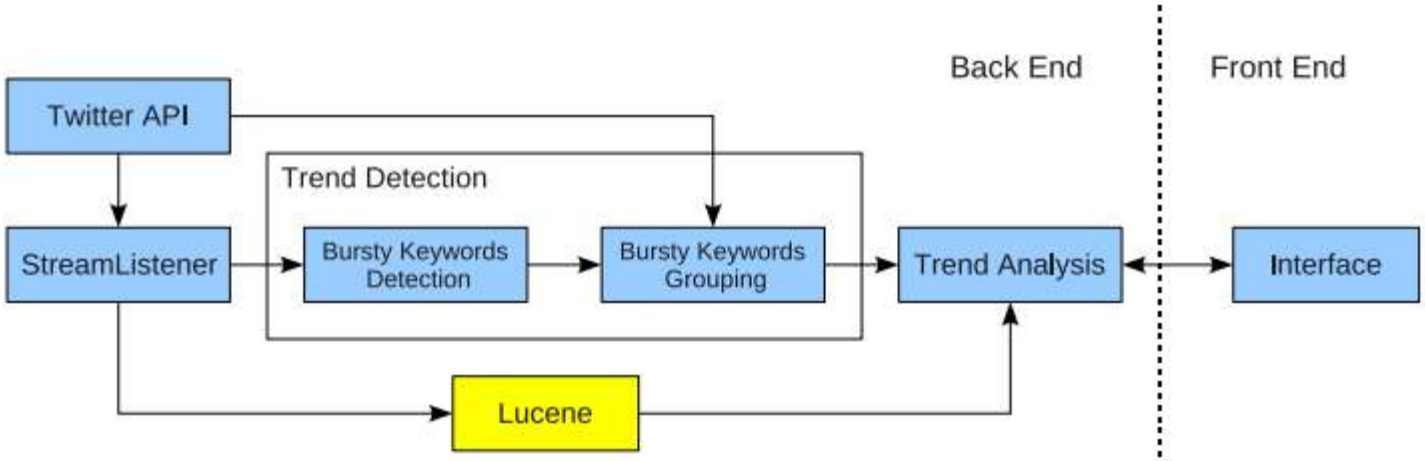
- Twitter is a great news source
  - Human sensors report very quickly
  - Tweet waves travel faster than earthquake waves!
- Overload of information
  - Show only ranked important events
- Showing 10 relevant tweets is not a great idea, since very few real information needs can be satisfied by a single short piece of text
- Will look at applications of event detection later

Manual Event Detection

- Twitter partnered with the third-party website WhatTheTrend to provide definitions of trending topics
- WhatTheTrend allows users to manually enter descriptions of why a topic is trending
- Problems
  - Spam
  - Manual (significant efforts)
  - Time lag

Twitter Monitor

- Michael Mathioudakis and Nick Koudas. TwitterMonitor: trend detection over the twitter stream. SIGMOD '10
- Identifies ‘bursty’ keywords, i.e., keywords that suddenly appear in tweets at an unusually high rate.
- Groups bursty keywords into trends based on their co-occurrences.
- Extracts additional information from the tweets that belong to the trend, aiming to discover interesting aspects of it.



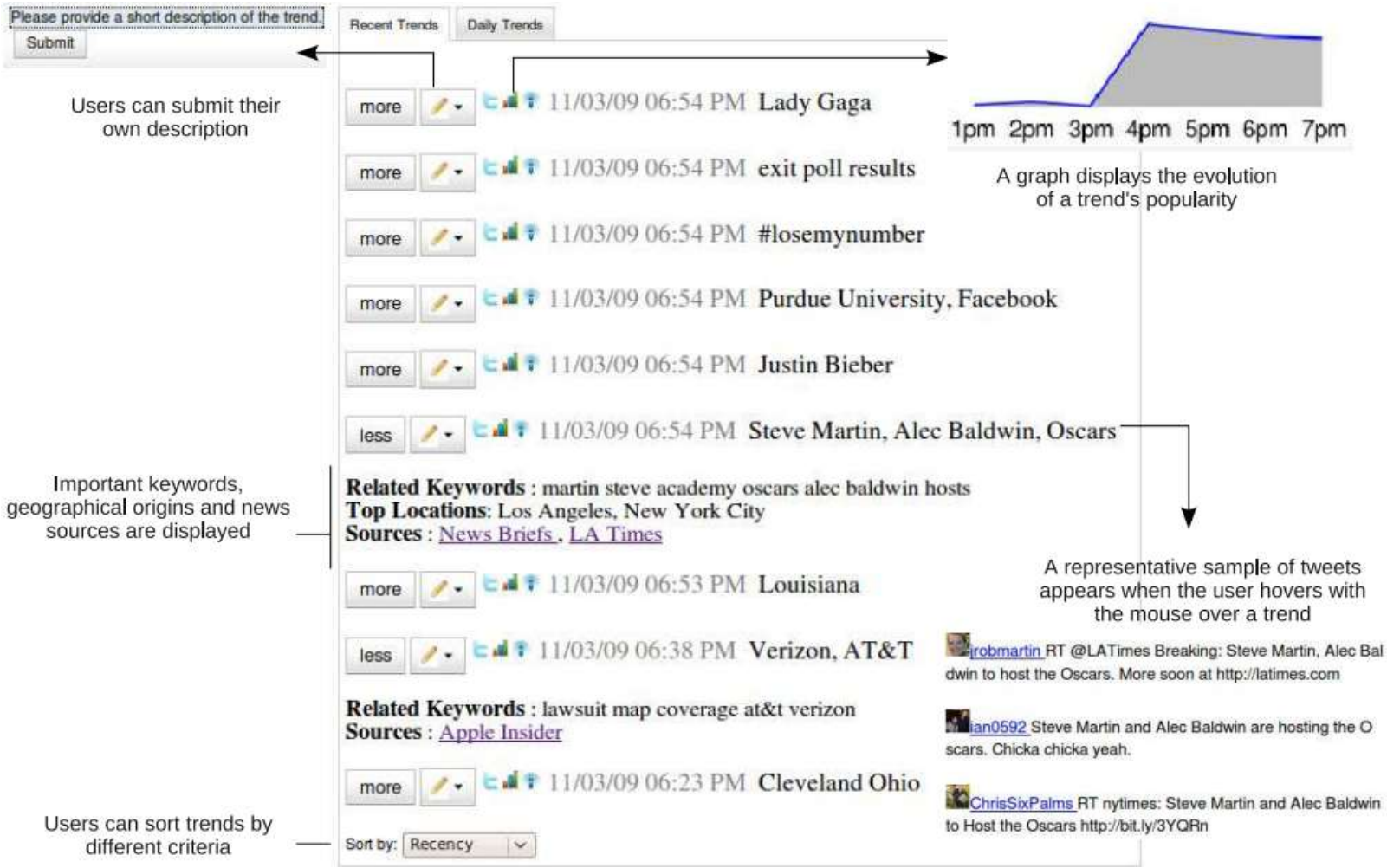
Twitter Monitor: Finding Bursty Keywords

- QueueBurst algorithm
  - One-pass
  - Real-time
  - Adjustable against ‘spurious’ bursts
  - Adjustable against spam
  - Theoretically sound (based on queuing theory)

Twitter Monitor: Trend Analysis

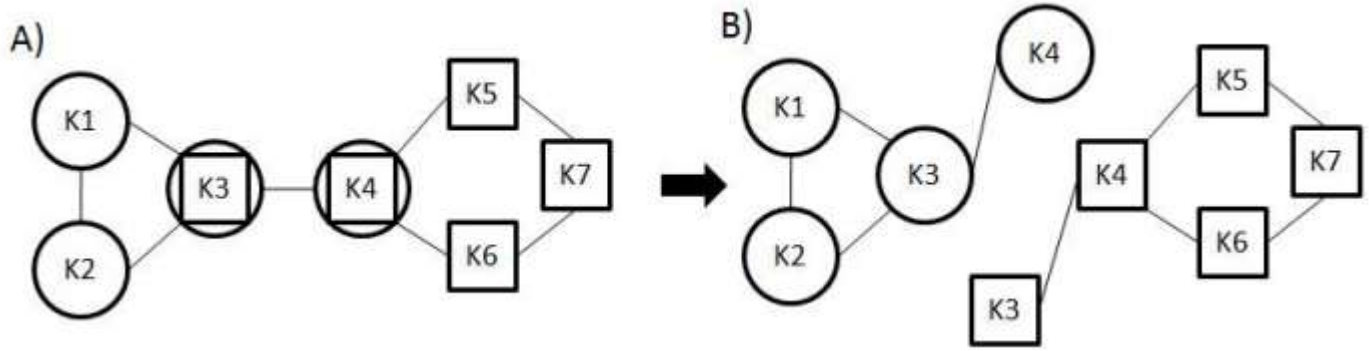
- Group bursty keywords based on their co-occurrences to get trends (keyword clusters)
- For every trend
  - Identify more keywords which may not be bursty but provide the context of the event
    - Using SVD
  - Identify frequently mentioned entities in tweets containing the trend keywords
  - Links in related tweets
  - Frequent geographical origins of related tweets

Twitter Monitor



Detecting Events using Graph Community Analysis

- Sayyadi, H., Hurst, M., and Maykov, A. (2009). Event Detection and Tracking in Social Streams. ICWSM.
- Extract names entities and noun phrases from tweets
- A KeyGraph is created where nodes are keywords and edges between the nodes are formed when those terms co-occur in a document
  - Nodes are created only if both TF and IDF are high for that node
  - Edge is created between two nodes k_i and k_j if co-occurrence prob is high, $p(k_i|k_j)$ is high and $p(k_j|k_i)$ is high
- Apply community analysis techniques to this graph to discover events (communities of keywords)
  - Remove edges with high betweenness centrality iteratively
  - A keyword can belong to more than 1 event
    - Before removing an edge, if the edge's conditional probability is high, edge and corresponding nodes are duplicated



Detecting Events using LSH (1)

- Saša Petrović, Miles Osborne, and Victor Lavrenko. Streaming first story detection with application to Twitter. HLT '10.
- Locality sensitive hashing can be used with the cosine similarity based distance function to compute nearest neighbor documents given a document d.
  - Cosine similarity based nearest neighbor finding can be done using hash functions which project the document vector onto a random hyperplane
  - Increasing the number of such hyperplanes (k) decreases the prob. of chance collisions.
  - But that also decreases prob. of colliding with nearest neighbor.
  - Hence maintain multiple (L) hash tables.

Detecting Events using LSH (2)

- LSH has a problem
  - If the nearest neighbor is far away, LSH does not work
  - So, if the minimum distance of new document d with all documents in colliding bucket (as per LSH) is higher than a threshold, check distance of d with most recent 1000 documents and update min distance if needed.
- Two improvements
  - In each hash table, maintain a constant number of documents per bucket. Remove old documents
  - On collision with buckets in L hash tables, don't compare with all documents in all L hash tables. Instead compare to the 3L documents that collide most frequently with the new document.

Algorithm 2: Our LSH-based approach.

```
input: threshold t
foreach document d in corpus do
    add d to LSH
    S ← set of points that collide with d in LSH
    dis_min(d) ← 1
    foreach document d' in S do
        c = distance(d, d')
        if c < dis_min(d) then
            dis_min(d) ← c
        end
    end
    if dis_min(d) >= t then
        compare d to a fixed number of most
        recent documents as in Algorithm 1 and
        update dis_min if necessary
    end
    assign score dis_min(d) to d
    add d to inverted index
end
```

Detecting Events using LSH (3)

- Specifically for Twitter
  - Tweet a links to tweet b if b is the nearest neighbor of a and $1 - \cos(a, b) < t$, where t is a user-specified threshold
  - Then, for each tweet a we either assign it to an existing thread if its nearest neighbor is within distance t, or say that a is the first tweet in a new thread.
  - Once we have threads of tweets, we are interested in threads which grow fastest, as this will be an indication that news of a new event is spreading. Therefore, for each time interval we only output the fastest growing threads. This growth rate also gives us a way to measure a thread's impact.

Detecting Events using CRFs (1)

- Given a repository of tweets, first find named entities using a CRF-based NER on tweets
- Next, find entity-referring phrases.
- Useful to display in connection with events
  - E.g. "**Steve Jobs**" + [ ] + "**October 6**"
- Helpful in categorizing Events into Types
- Examples
  - Apple to Announce iPhone 5 on October 4th! YES!
  - iPhone 5 announcement coming Oct 4th
  - WOOOHOO NEW IPHONE TODAY! CAN'T WAIT!

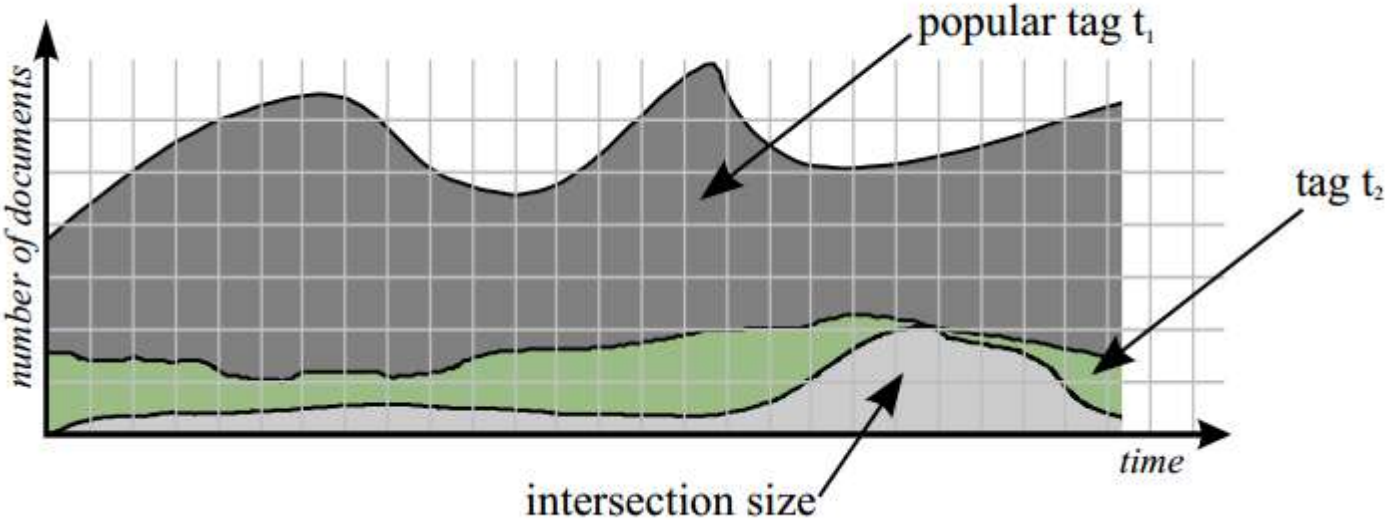
Detecting Events using CRFs

- Using CRF to identify event-referring phrases
  - Contextual features
    - POS tags
    - Adjacent words
  - Dictionary Features
    - Event words gathered from WordNet
    - Brown Clusters
  - Orthographic Features
    - Prefixes, suffixes

Entity	Event Phrase	Date
Steve Jobs	died	10/6/11
iPhone	announcement	10/4/11
GOP	debate	9/7/11
...	...	...

Detecting Events using Tag Correlations (1)

- Foteini Alvanaki, Michel Sebastian, Krithi Ramamritham, and Gerhard Weikum. EnBlogue: emergent topic detection in web 2.0 streams. SIGMOD '11.
- Compared to Mathioudakis and Koudas's Twitter Monitor system, this work considers shifts in tag correlations

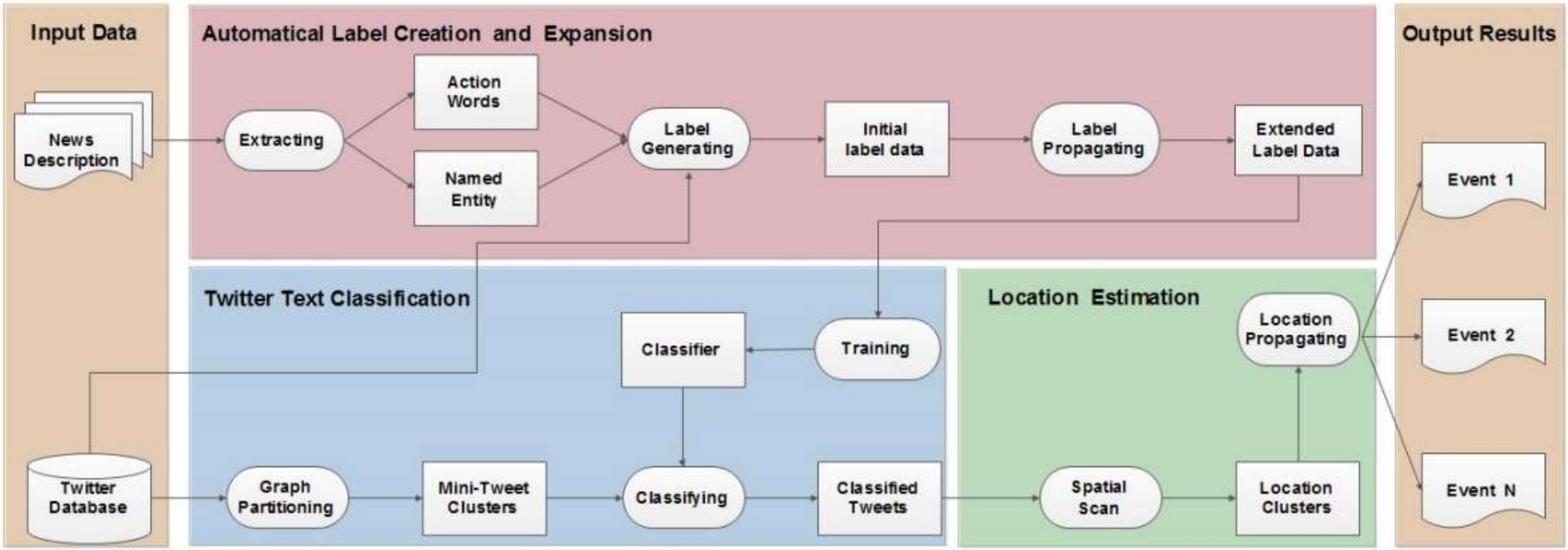


Detecting Events using Tag Correlations (2)

- Framework
  - Seed tag selection
    - Based on popularity
  - Correlation tracking
    - For each tag pair that contains at least one seed tag, track correlations
  - Shift detection
    - sudden (but significant) increases in the correlation of tag pairs
    - If current correlation is significantly different from the prediction based on the previous correlation values

Detecting Events by Label Propagation from News (1)

Ting Hua, Feng Chen, Liang Zhao, Chang-Tien Lu, and Naren Ramakrishnan. STED: semi-supervised targeted-interest event detectionin in twitter. KDD '13



Detecting Events by Label Propagation from News (2)

- From news articles, extract named entities and action words
- Tweets containing at least 1 named entity and 1 action word is labeled as positive
- Label propagation
  - Identify social ties terms from labeled tweets: Mentions(@), Hashtag(#)
  - Remove infrequent terms
  - Get more tweets from database which contain these terms
  - Label new tweets for a term as positive if #newly discovered tweets<#already labeled tweets for term t
  - Iterate label propagation until no new tweets are found

Detecting Events by using Information from Knowledge Bases

- Heather S. Packer, Sina Samangooei, Jonathon S. Hare, Nicholas Gibbins, and Paul H. Lewis. Event detection using Twitter and structured semantic query expansion. CrowdSens '12
- Given an event query, extract tweets containing the query
- From these tweets extract entities
- Find related entities from knowledge bases thereby extending the query
- Use these new entities to retrieve more tweets relevant to the event, thereby summarizing the event in a more comprehensive manner.

Today's Agenda

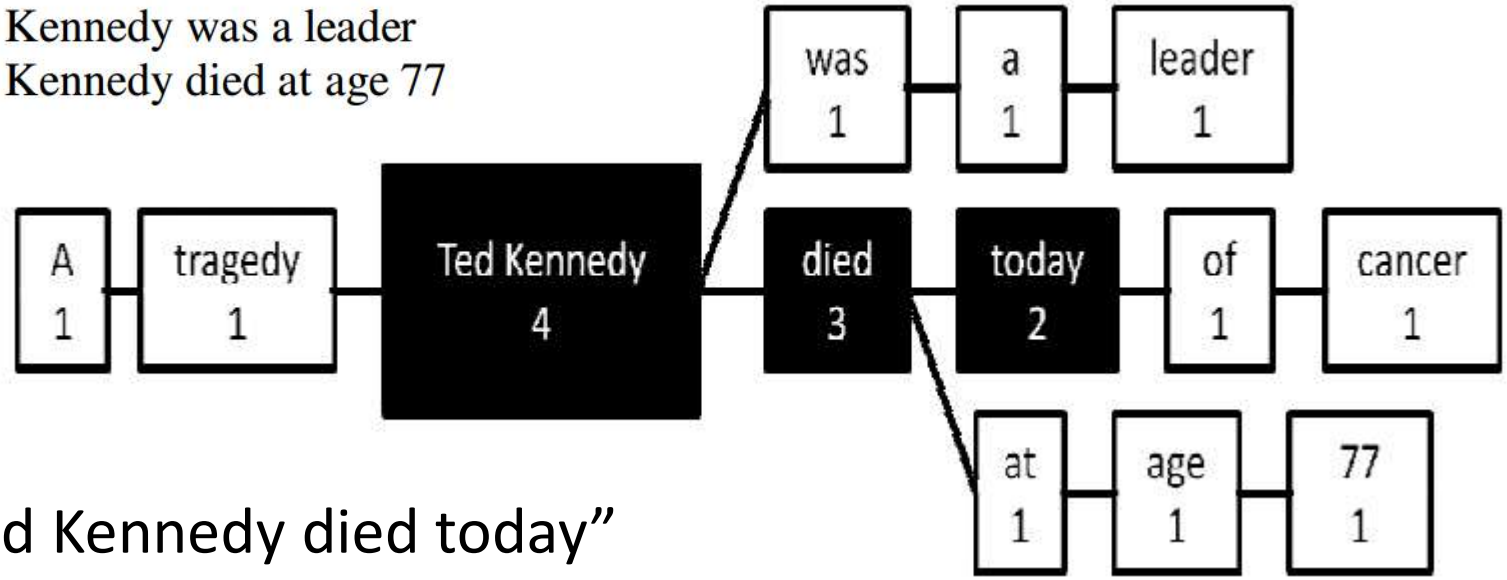
- Event Detection in Twitter
- **Generating Event Descriptions/Annotations**
- Application of Event Detection from Twitter

Finding the Best Phrase to Describe an Event

- Beaux Sharifi, Mark-Anthony Hutton, and Jugal Kalita. Summarizing microblogs automatically. HLT '10

Event p="Ted Kennedy"

1. A tragedy: Ted Kennedy died today of cancer
2. Ted Kennedy died today
3. Ted Kennedy was a leader
4. Ted Kennedy died at age 77



- "Ted Kennedy died today"

Finding the Best Phrase to Describe an Event

- How to describe these events corresponding to a phrase p?
  - Phrase Reinforcement Algorithm
    - Get all tweets containing p
    - Remove spam and non-English tweets
    - Get the longest sentence from each post which contains p
    - Build a graph representing common sequences of words that occur both before and after p
    - Partial sentence = path with max total weight beginning from root and ending at a non-root node and containing nodes that occur>T times
    - Build graph again by setting p as the partial path
    - Full sentence = path with max total weight beginning from root and ending at a non-root node and containing nodes that occur>T times

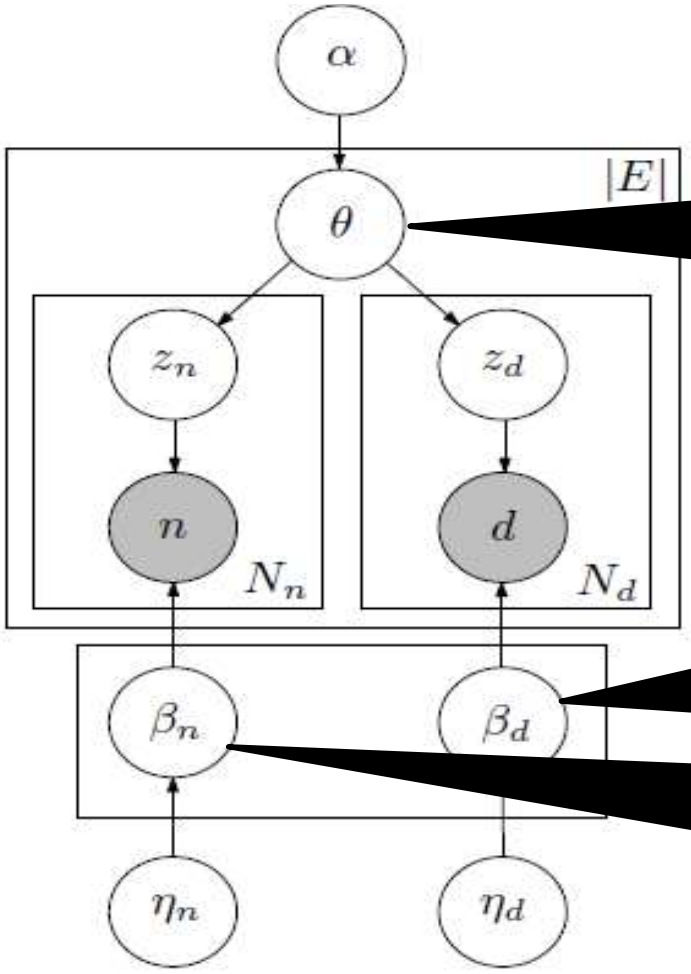
Finding Event Types (1)

- Alan Ritter, Mausam, Oren Etzioni, and Sam Clark. Open domain event extraction from twitter. KDD '12.
- Would like to categorize events into types, for example:
  - Sports
  - Politics
  - Product releases
  - …
- Benefits:
  - Allow more customized Twitter event calendars
  - Could be useful in upstream tasks

Finding Event Types (2)

- Challenges
  - Many Different Types
  - Not sure what is the right set of types
  - Set of types might change
    - Might start talking about different things
    - Might want to focus on different groups of users
- Solution: Unsupervised Event Type Induction
  - Latent Variable Models
    - Generative Probabilistic Models
  - Advantages:
    - Discovers types which **match the data**
    - No need to annotate individual events
    - Don't need to commit to a specific set of types
    - Modular, can integrate into various applications

Finding Event Types (3)



Each **Event Phrase** is modeled as a mixture of types

P(**SPORTS**|**cheered**)= 0.6
P(**POLITICS**|**cheered**)= 0.4

Each **Event Type** is Associated with a Distribution over Entities and Dates

Finding Event Types (4)

Label	Top 5 Event Phrases	Top 5 Entities
Sports	tailgate - scrimmage - tailgating - homecoming - regular season	espn - ncaa - tigers - eagles - varsity
Concert	concert - presale - performs - concerts - tickets	taylor swift - toronto - britney spears - rihanna - rock
Perform	matinee - musical - priscilla - seeing - wicked	shrek - les mis - lee evans - wicked - broadway
TV	new season - season finale - finished season - episodes - new episode	jersey shore - true blood - glee - dvr - hbo
Movie	watch love - dialogue theme - inception - hall pass - movie	netflix - black swan - insidious - tron - scott pilgrim
Sports	inning - innings - pitched - homered - homer	mlb - red sox - yankees - twins - dl
Politics	presidential debate - osama - presidential candidate - republican debate - debate performance	obama - president obama - gop - cnn - america

Finding Event TimeSpans (1)

Donald Metzler, Congxing Cai, and Eduard Hovy. Structured event retrieval over microblog archives. HLT '12.

July 16 2010 at 17 UTC, for **11** hours

Summary tweets:

i. *Ok a 3.6 "rocks" nothing. But boarding a plane there now, Woodward ho! RT @todayshow: 3.6 magnitude #earthquake rocks Washington DC area.*

ii. *RT @fredthompson: 3.6-magnitude earthquake hit DC. President Obama said it was due to 8 years of Bush failing to regulate plate tectonic ...*

iii. *3.6-magnitude earthquake wakes Md. residents: Temblor centered in Gaithersburg felt by as many as 3 million people... http://bit.ly/9iMLEk*

Figure 1: Example structured event representation retrieved for the query "earthquake".

- A list of timespans during which an instance of the event occurred and was actively discussed within the microblog stream.
- For each timespan, a small set of relevant messages are retrieved for the purpose of providing a highlevel summary of the event that occurred during the timespan

Finding Event TimeSpans (2)

- Framework
  - Timespan retrieval
  - Summarization
- Query expansion is needed because
  - Microblog messages that are highly related to the query might not contain any of the query keywords
  - Vocabulary mismatch: Keyword might be expressed in another form: possibly shortened or slang. E.g., earthquake may be written as quake or #eq

Finding Event TimeSpans (3)

- Temporal Query Expansion
  - Given query q, extract N timespans (hours) for which q was heavily discussed
  - Rank timespans based on proportion of messages posted during the timespan that contain q
  - Top few timespans are then considered to be pseudo-relevant.
  - For each word in all pseudo-relevant timespans, compute their burstiness score: $burstiness(w, TS_i) = \frac{P(w|TS_i)}{P(w)}$
    - $P(w|TS_i) = \frac{tf_{w,TS_i} + \frac{\mu tf_w}{N}}{|TS_i| + \mu}$ and $P(w) = \frac{tf_w + K}{N + K|V|}$
    - tf_{w,TS_i} is the number of occurrences of w in timespan TS_i
    - tf_w is the number of occurrences of w in the entire microblog archive
    - $|TS_i|$ is the number of terms in timespan TS_i
    - N is the total number of terms in the microblog archive, V is the vocabulary size, and μ and K are smoothing parameters
    - By smoothing P(w), we dampen the effect of overweighting very rare terms.
  - Score of a word is geometric mean of burstiness scores across all pseudo-relevant timespans.
    - Geometric mean ensures that the highest weighted terms are those that have large weights in a large number of the timespans, thereby eliminating spurious terms
  - The k highest weighted terms are then used as expansion terms

Finding Event TimeSpans (4)

- Timespan Ranking
- We have expanded query q'
- Identify the 1000 highest scoring timespans (with respect to q')
- Merge contiguous timespans into a single, longer timespan, where the score of the merged timespan is the maximum score of its component timespans.
- The final ranked list consists of the merged timespans.
- Two scoring functions
  - Coverage Scoring Function
    - $s(q', TS) = \sum_{w \in q'} \beta_w . tf_{w,TS}$
    - Where $tf_{w,TS}$ is the term frequency of w_i in timespan TS and β_w is the expansion weight of term w.
  - Burstiness Scoring Function
    - $s(q', TS) = \cos(\beta_{q'}, \beta_{TS})$ where β_{TS} is the burstiness score for all terms in the interval TS.

Finding Event TimeSpans (5)

- Timespan Summarization
  - Provide a quick overview of the event to the user
    - Retrieve a small set of microblog messages posted during the timespan that are the most relevant to the expanded representation of the original query
    - $s(q', M) = \sum_{w \in q'} \beta_w . \log P(w|M)$
      - β_w is burstiness score of w
      - $P(w|M)$ is Dirichlet smoothed language modeling estimate for term w in message M

Finding Event TimeSpans (6): Examples

Category	Events
Business	layoffs, bankruptcy, acquisition, merger, hostile takeover
Celebrity	wedding, divorce
Crime	shooting, robbery, assassination, court decision, school shooting
Death	death, suicide, drowned
Energy	blackout, brownout
Entertainment	awards, championship game, world record
Health	recall, pandemic, disease, flu, poisoning
Natural Disaster	hurricane, tornado, earthquake, flood, tsunami, wildfire, fire
Politics	election, riots, protests
Terrorism	hostage, explosion, terrorism, bombing, terrorist attack, suicide bombing, hijacked
Transportation	plane crash, traffic jam, sinks, pileup, road rage, train crash, derailed, capsizes

July 16 2010 at 17 UTC, for **11** hours

Ok a 3.6 *"rocks" nothing. But boarding a plane there now, Woodward ho! RT @todayshow:* 3.6 *magnitude #earthquake rocks Washington DC area.*

September 28 2010 at 11 UTC, for **6** hours

RT @Quakeprediction: 2.6 earthquake (possible foreshock) hits E of Los Angeles; `http://earthquake.usgs.gov/earthquakes/recenteqscanv/Fau` ...

September 04 2010 at 01 UTC, for **3** hours

7.0 quake strikes New Zealand - A 7.0-magnitude earthquake has struck near New Zealand's second largest city. Reside... `http://ht.ly/18R2rw`

October 27 2010 at 01 UTC, for **5** hours

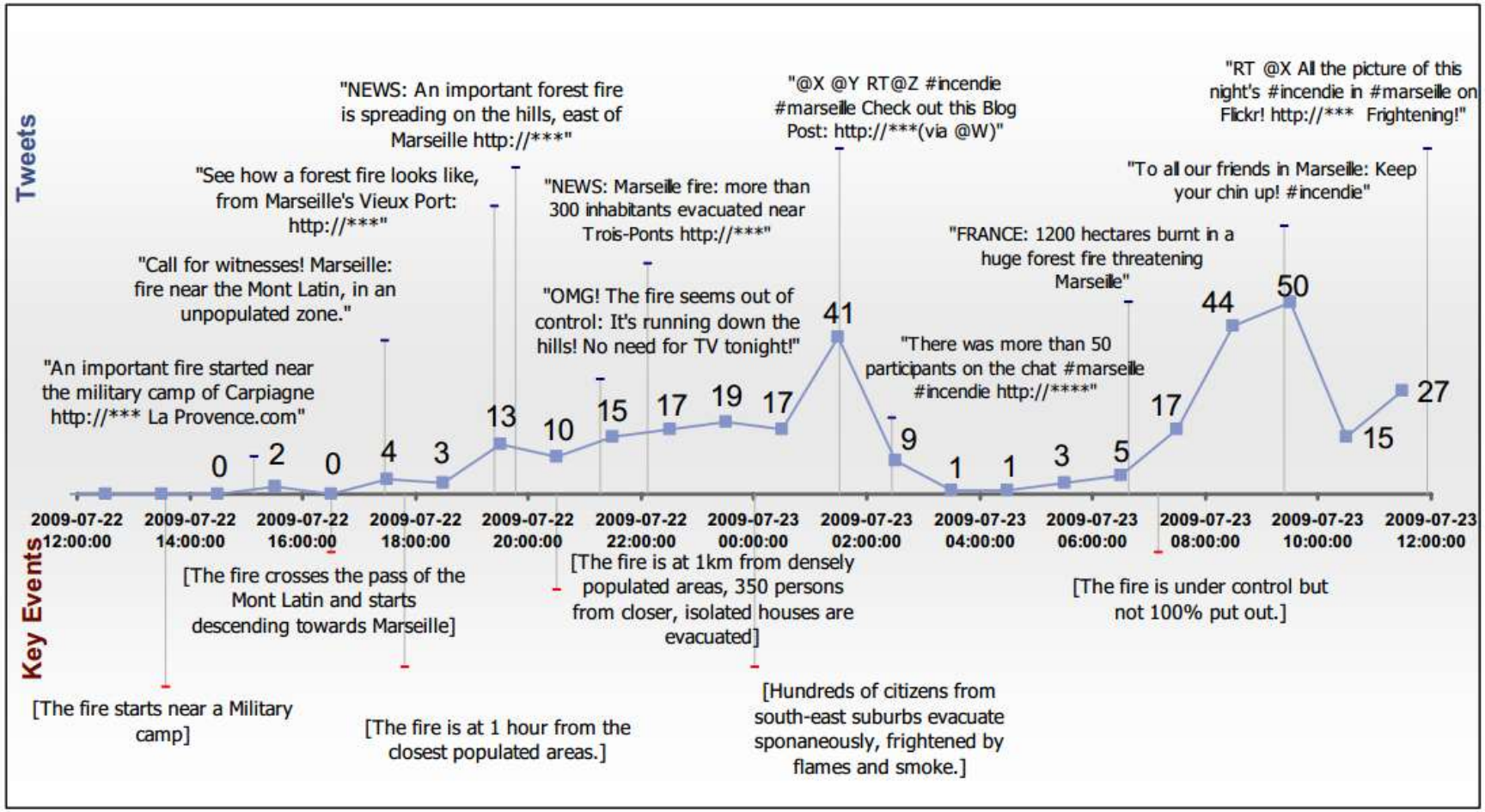
RT @SURFER_Magazine: Tsunami Strikes Mentawais: Wave Spawned By A 7.5-Magnitude Earthquake Off West Coast Of Indonesia `http://bit.ly/8Z9Lbv`

Today's Agenda

- Event Detection in Twitter
- Generating Event Descriptions/Annotations
- **Application of Event Detection from Twitter**

Detecting Forest-fires

Detecting Forest-fires: Bertrand De Longueville, Robin S. Smith, and Gianluca Luraschi. "OMG, from here, I can see the flames!": a use case of mining location based social networks to acquire spatio-temporal data on forest fires. LBSN '09



Detecting Sporting Events (1)

- Jeffrey Nichols, Jalal Mahmud, and Clemens Drews. Summarizing sporting events using twitter. IUI '12
- Generate a journalistic summary of events from tweets
- Spikes are used to identify important moments to describe an event
- Sporting events consist of a sequence of moments, each of which may contain actions by players, the referee, the fans, etc.

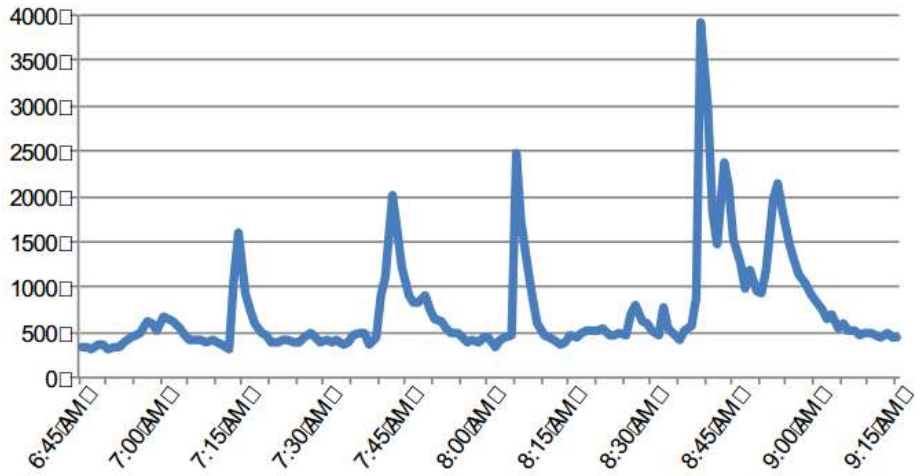


Figure 1. Twitter volume graph for the 2010 World Cup game of US vs. Slovenia. The x-axis is time and the y-axis is volume as measured in tweets/minute.

Detecting Sporting Events (2)

Game	Actual Events	Detected		R	P
		Moments	Events		
US vs. Slovenia	13	9	8	0.62	0.89
Germany vs. Serbia	16	8	11	0.69	0.92
Australia vs. Serbia	11	9	10	0.91	0.91

Key Event Type	Recall
Goal	1.0
Red Card	1.0
Yellow Card	0.53
Penalty	1.0
Game Start	0.67
Game End	1.0
Half Time	0.33
Disallowed Goals	1.0

Detecting Sporting Events (3)

Game	Spike	Manual Summary	Our Summary
US vs. Slovenia	1	In the first 15 mins of the soccer game between USA and Slovenia, Slovenia is leading with a goal by Birsa. Birsa scored an easy goal from midfield to the right of the goal, as USA left that shot wide open. Terrible defense by USA team, too much space left open.	Good goal for Slovenia and the USA once again starts a game terrible. Birsa gives #SVN 1-0 lead with smart shot. Howard didn't even look like he saw that one coming.
Germany vs. Serbia	3	Klose argues with referee, gets second yellow cards and is out of the game. Germany down to 10 men. 1-0 Serbia.	Germany screwed by the refs and a red card for Klose; seconds later, a pretty goal by the Serbs. yellow seems to be a very popular colour in this game.
Australia vs. Serbia	9	Serbia Australia match ends with 2-1. With a result of 1:0 between Germany and Ghana this means that Ghana and Germany will advance to the knock out rounds and serbia and australia will be out.	Australia won 2-1 on serbia, Germany won 1-0 vs Ghana, Germany and Ghana goes on to the next round. Great win by #aus but not good enough to go through. Final score #Aus 2 #Srb 1.

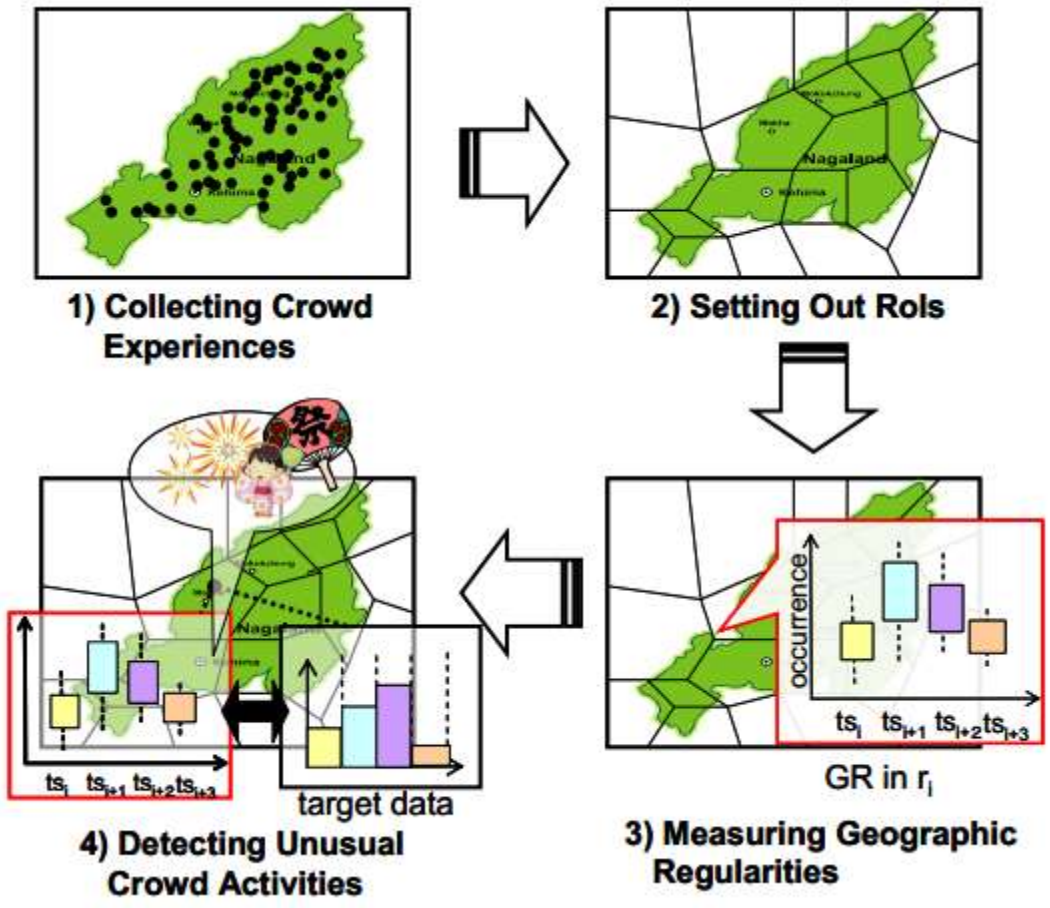
	Mean
Readability	6.01
Grammaticality	5.60
Content	5.19

7 point Likert scale

Detecting Local Festivals (1)

- Ryong Lee and Kazutoshi Sumiya. Measuring geographical regularities of crowd behaviors for Twitter-based geo-social event detection. LBSN '10
- To detect such unusual geo-social events, they depend on geographical regularities deduced from the usual behavior patterns of crowds with geo-tagged microblogs.
- By comparing these regularities with the estimated ones, they decide whether there are any unusual events happening in the monitored geographical area.

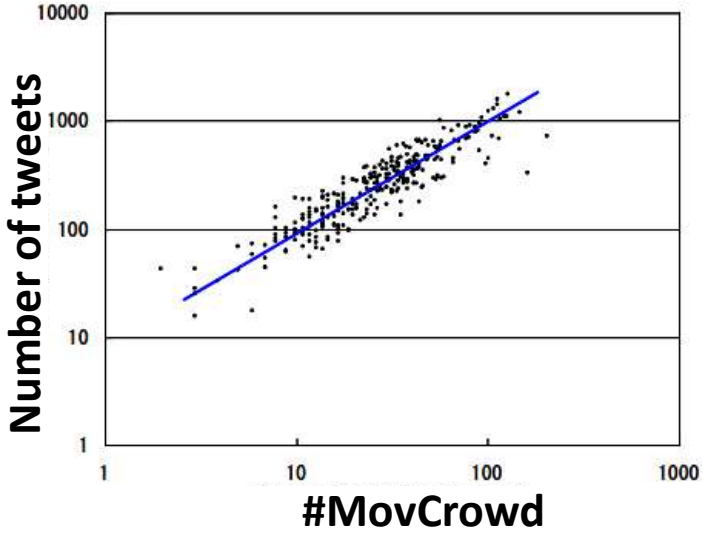
Detecting Local Festivals (2)



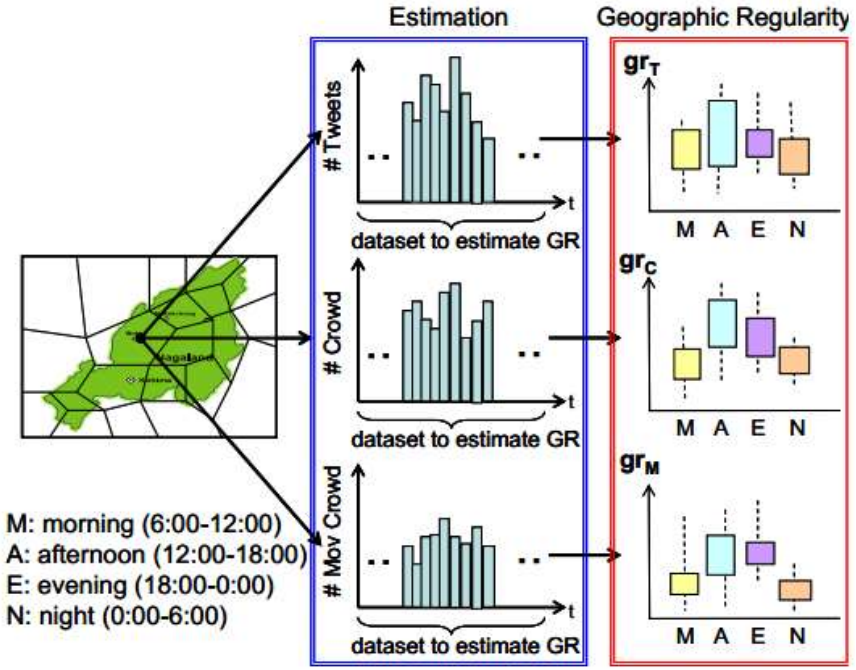
RoIs = Regions of Interest
RoIs (sub-regions) are computed from the region by running K-Means on geographically distributed points

Detecting Local Festivals (3)

- Measuring Geographical Regularity
  - #Tweets: the number of tweets that were written in an RoI within a specific period of time.
  - #Crowd: the number of Twitter users found in an RoI within a specific time period.
  - #MovCrowd: : 1) Inner: A crowd in an RoI moves only inside the region without going outside. 2) Incoming: There are some people coming from outside, and 3) Outgoing: Conversely, some people move outside the RoI. For simplification, they consider #MovCrowd as inner+incoming

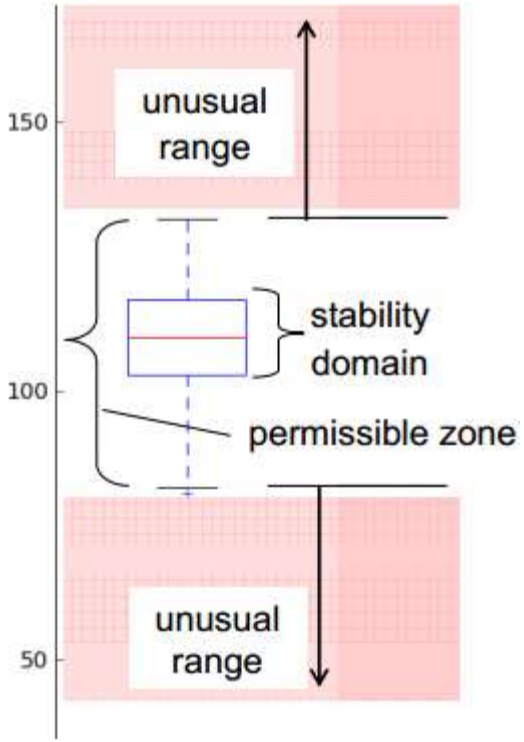


Detecting Local Festivals (4)



	#Tweets (T)	#Crowd (C)	#MovCrowd (M)	FinalDecision (F)
(a)	N	N	N	N
(b)			A	N
(c)		A	N	N
(d)			A	A
(e)	A	N	N	N
(f)			A	A
(g)		A	N	N
(h)			A	A

N : normal A : abnormal



Detecting Local Festivals (5)

Table 3 Town festivals held in Japan for 7/17–7/19

No.	Event name	Place	Event day(s)
1	Kyoto Gion Festival	Sakyo, Kyoto, Kyoto	7/17
2	ShishkuiGion Festival	Shiskui, Kaiyoumachi, Tokushima	7/17, 7/18
3	Towada KosuiFestival	Towada, Aomori	7/17
4	Tamamura Firework	Tamura, Gunmu	7/17
5	Ise Firework	Nakajima, Ise, Mie	7/17
6	AkiyoshiFirework	Syoho, Mine, Yanaguchi	7/17
7	KanonjiFestival	Kannonji, Kagawa	7/17, 7/18
8	Muroto Festival	Muroto, Kouchi	7/18
9	SanoyoiCarnival	Arao, Kumamoto	7/18
10	UminohiFestival in Nagoya	Nagoya, Aichi	7/19
11	HousuiFestival	Noboribetsu, Hokkaido	7/17
12	Shinmatsudo Festival	Matudo, Chiba	7/17, 7/18
13	Nanao Festival	Nanao, Ishkawa	7/17
14	Oota Festival	Oota, Gunma	7/17
15	Toukou Festival	Arita, Wakayama	7/18

The algorithm could detect 13 of the 15 festivals

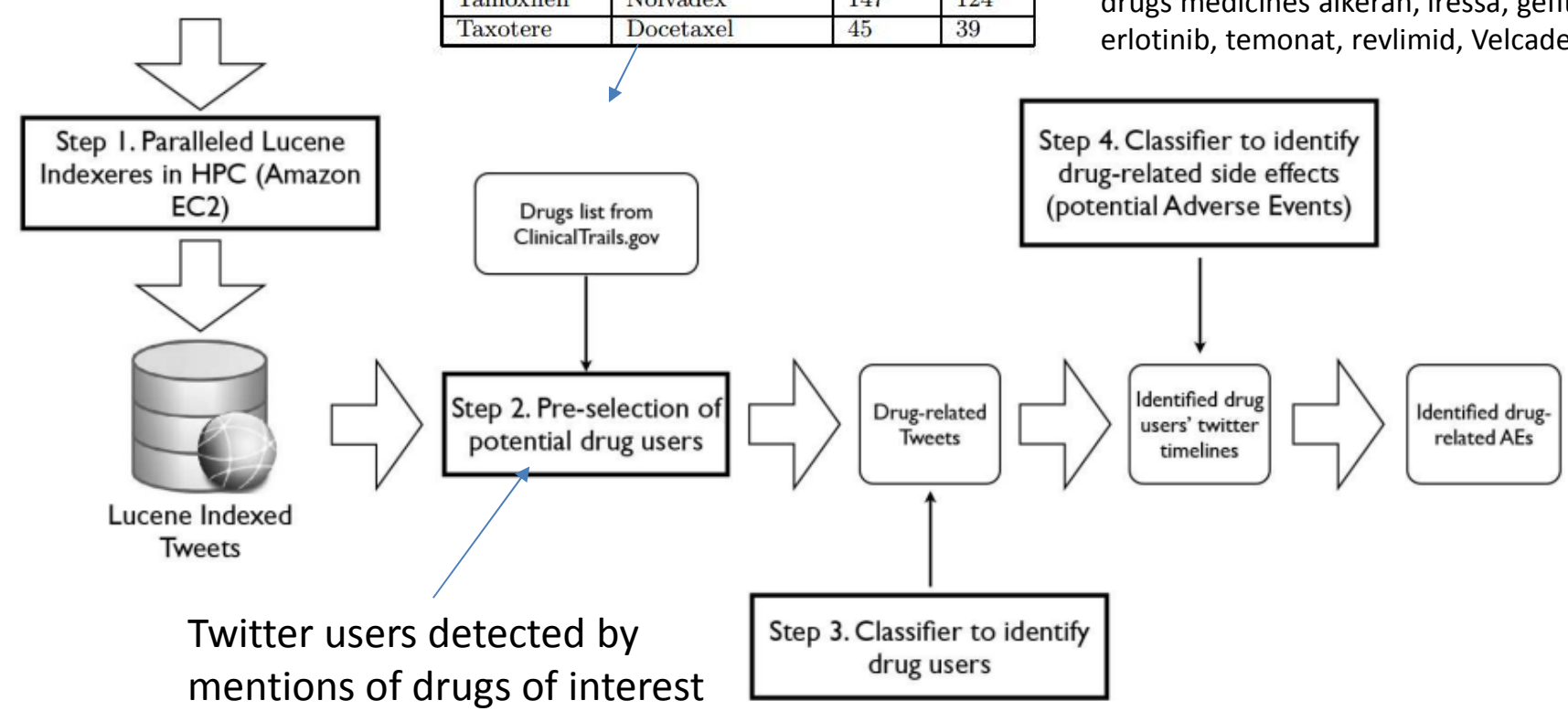
Detecting Drug Related Adverse Events (1)

- Jiang Bian, Umit Topaloglu, and Fan Yu. Towards large-scale twitter mining for drug-related adverse events. SHB '12
- Given the high frequency of user updates, mining Twitter messages can lead us to real-time pharmacovigilance.
- To mine Twitter messages for AEs, the process can be separated into two parts: 1) identifying potential users of the drug; 2) finding possible side effects mentioned in the users' Twitter timeline that might be caused by the use of the drug concerned.

Detecting Drug Related Adverse Events (2)

Drug name	Synonym(s)	# of tweets	# of users
Avastin	Bevacizumab	264	236
Melphalan	ALKERAN	23	15
Rupatadin	Rupafin, Urtimed	10	10
Tamoxifen	Nolvadex	147	124
Taxotere	Docetaxel	45	39

Positive: "No more tamoxifen for me - finished 5 yrs of post cancer drug therapy"
Negative: "Please visit us at www.genglob.com For generic anti cancer drugs medicines alkeran, iressa, gefitinib, erlotinib, temonat, revlimid, Velcade"



Twitter users detected by mentions of drugs of interest in their tweets

Detecting Drug Related Adverse Events (3)

- Features for identifying drug users
  - Textual features that construct a specific meaning in the text:
    - Bag-of-words features that indicate an action or a state that the user has taken the drug
    - Number of hash-tags occurred in the document
    - Number of reply-tags occurred in the document
    - Number of words that indicate negation
    - Number of URLs
    - Number of pronouns
    - Number of occurrences of the drug name or its synonyms
  - Semantic features that express the existence of semantic properties (i.e., based on Unified Medical Language System (UMLS) Concept Unique Identifiers (CUIs) extracted from the Tweets)
    - Number of CUIs in each Semantic Type
    - Number of CUIs in each Semantic Group

Detecting Emerging Controversial Events

- Ana-Maria Popescu and Marco Pennacchiotti. Detecting controversial events from twitter. CIKM '10
- Controversial event: An event is controversial if it provokes a public discussion in which audience members express opposing opinions or disbelief
- Twitter snapshot=target entity+time period+set of tweets
- Goal: Rank Twitter snapshots by controversy score
- Regression model to detect event snapshots and to compute controversy scores
- Features
  - Twitter-based features: snapshots' linguistic properties, structural and social graph information, the intensity of the discussion about the entity, the distribution of sentiment words in the snapshot, and the level of controversy
  - News buzz features: if an entity is buzzy in news articles at the same time it is buzzy in a Twitter snapshot, then the snapshot is likely to refer to a real-world event.
  - Web and news controversy features: assess the past and present levels of controversy surrounding the target entity in the snapshot.

Take-away Messages

- Detecting events from Twitter is difficult because of the unique characteristics of the microblogs
- We saw various ways of detecting interesting events from Twitter
- We also discussed ways of extracting event descriptions from Twitter
- Finally, we discussed various applications of event detection on Twitter

Further Reading (1)

- Michael Mathioudakis and Nick Koudas. TwitterMonitor: trend detection over the twitter stream. SIGMOD '10
- Sayyadi, H., Hurst, M., and Maykov, A. (2009). Event Detection and Tracking in Social Streams. ICWSM.
- Saša Petrovic, Miles Osborne, and Victor Lavrenko. Streaming first story detection with application to Twitter. HLT '10.
- Arpit Khurdiya, Lipika Dey, Diwakar Mahajan, and Ishan Verma. Extraction and Compilation of Events and Sub-events from Twitter. WI-IAT '12
- Anqi Cui, Min Zhang, Yiqun Liu, Shaoping Ma, and Kuo Zhang. Discover breaking events with popular hashtags in twitter. CIKM '12.
- Foteini Alvanaki, Michel Sebastian, Krithi Ramamritham, and Gerhard Weikum. EnBlogue: emergent topic detection in web 2.0 streams. SIGMOD '11.
- Chenliang Li, Aixin Sun, and Anwitaman Datta. Twevent: segment-based event detection from tweets. CIKM '12.
- Ting Hua, Feng Chen, Liang Zhao, Chang-Tien Lu, and Naren Ramakrishnan. STED: semi-supervised targeted-interest event detectionin in twitter. KDD '13
- Heather S. Packer, Sina Samangooei, Jonathon S. Hare, Nicholas Gibbins, and Paul H. Lewis. Event detection using Twitter and structured semantic query expansion. CrowdSens '12
- Beaux Sharifi, Mark-Anthony Hutton, and Jugal Kalita. Summarizing microblogs automatically. HLT '10
- Alan Ritter, Mausam, Oren Etzioni, and Sam Clark. Open domain event extraction from twitter. KDD '12

Further Reading (2)

- Donald Metzler, Congxing Cai, and Eduard Hovy. Structured event retrieval over microblog archives. HLT '12.
- Chung-Hong Lee, Hsin-Chang Yang, Tzan-Feng Chien, and Wei-Shiang Wen. A Novel Approach for Event Detection by Mining Spatio-temporal Information on Microblogs. ASONAM '11
- Detecting Forest-fires: Bertrand De Longueville, Robin S. Smith, and Gianluca Luraschi. "OMG, from here, I can see the flames!": a use case of mining location based social networks to acquire spatio-temporal data on forest fires. LBSN '09
- Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. Information credibility on twitter. WWW '11.
- Jeffrey Nichols, Jalal Mahmud, and Clemens Drews. Summarizing sporting events using twitter. IUI '12
- Ryong Lee and Kazutoshi Sumiya. Measuring geographical regularities of crowd behaviors for Twitter-based geo-social event detection. LBSN '10
- Jiang Bian, Umit Topaloglu, and Fan Yu. Towards large-scale twitter mining for drug-related adverse events. SHB '12
- Sílvio S. Ribeiro, Jr., Clodoveu A. Davis, Jr., Diogo Rennó R. Oliveira, Wagner Meira, Jr., Tatiana S. Gonçalves, and Gisele L. Pappa. Traffic Observatory: A System to Detect and Locate Traffic Events and Conditions using Twitter. LBSN '12
- Vasileios Lampos, Tijl De Bie, and Nello Cristianini. Flu detector: tracking epidemics on twitter. ECML PKDD'10
- Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes Twitter users: real-time event detection by social sensors. WWW '10
- Ana-Maria Popescu and Marco Pennacchiotti. Detecting controversial events from twitter. CIKM '10

Preview of Lecture 14: Analysis of Microblogs (Part 2)

- Location Estimation on Twitter

Disclaimers

- This course represents opinions of the instructor only. It does not reflect views of Microsoft or any other entity (except of authors from whom the slides have been borrowed).
- Algorithms, techniques, features, etc. mentioned here might or might not be in use by Microsoft or any other company.
- Lot of material covered in this course is borrowed from slides across many universities and conference tutorials. These are gratefully acknowledged.

Thanks!