

Continuous Geospatial Monitoring of Catastrophic Natural Disasters using Twitter

Eduard Hovy

Don Metzler, Congxing Cai, and Stephan Gouws

University of Southern California
Information Sciences Institute

CCICADA Center for Advanced Data Analytics

hovy@isi.edu

The situation

- People are everywhere and observe their environment
- When they're interconnected and report their findings, we have a distributed 'sensor' network
 - FaceBook, Twitter, MySpace, SMS/texting...
- We can track information flow on the non-private portion of the network to determine what is going on

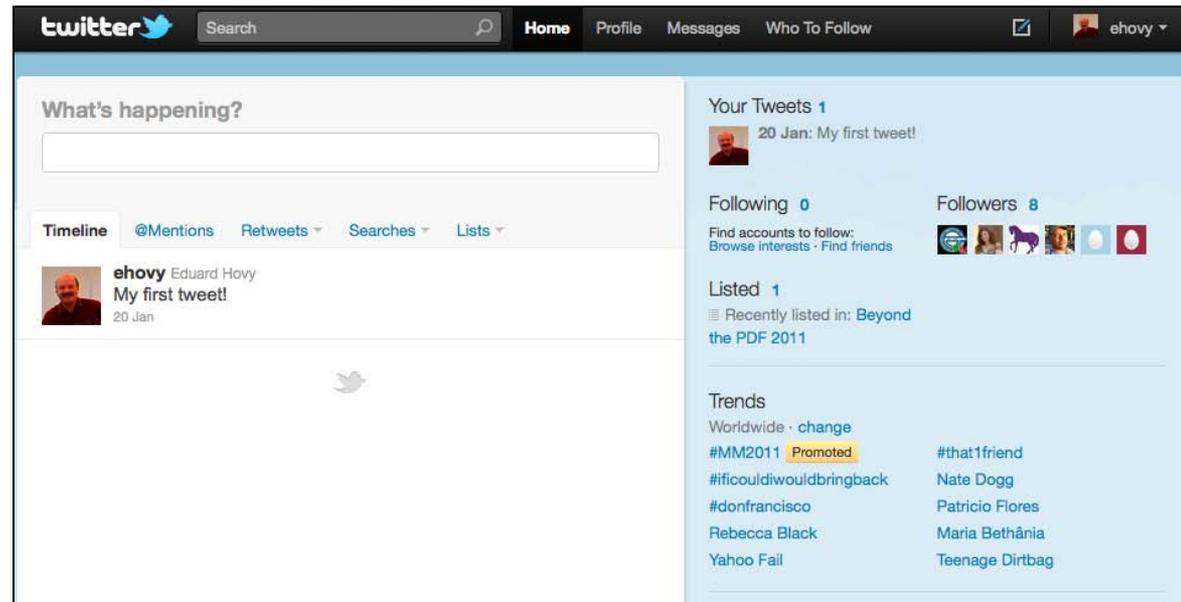
Potential uses

- **Catastrophes:** Situation monitoring and response planning
- **Anomaly Detection:** Recognizing problems before they occur
- **Human Trafficking:** Tracking down perpetrators

Twitter

<http://twitter.com/>

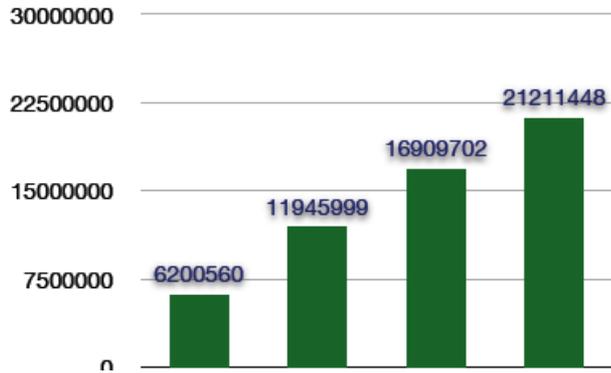
- Registering:
 - It's free
 - You choose a name
- Using Twitter:
 - Twitter users write 'tweets' (text messages, shorter than 140 characters) using a computer or cellphone
 - Anyone can read these messages
 - You can attach yourself to someone so you always are alerted when they write
 - Hashtags: Define a keyword: #sometopic



Monitoring events on Twitter

- Goal: Can we find out when events occur by watching the Twitter stream?
- Approach: Analyze Twitter stream
 - Build model of default Twitter behavior: ‘background noise’
 - Learn models for 50 specific kinds of events:
 - Topic signatures for earthquakes, fires, explosions, etc.
 - Develop methods to find event signals against the background noise
 - Develop methods to pinpoint locations in tweets

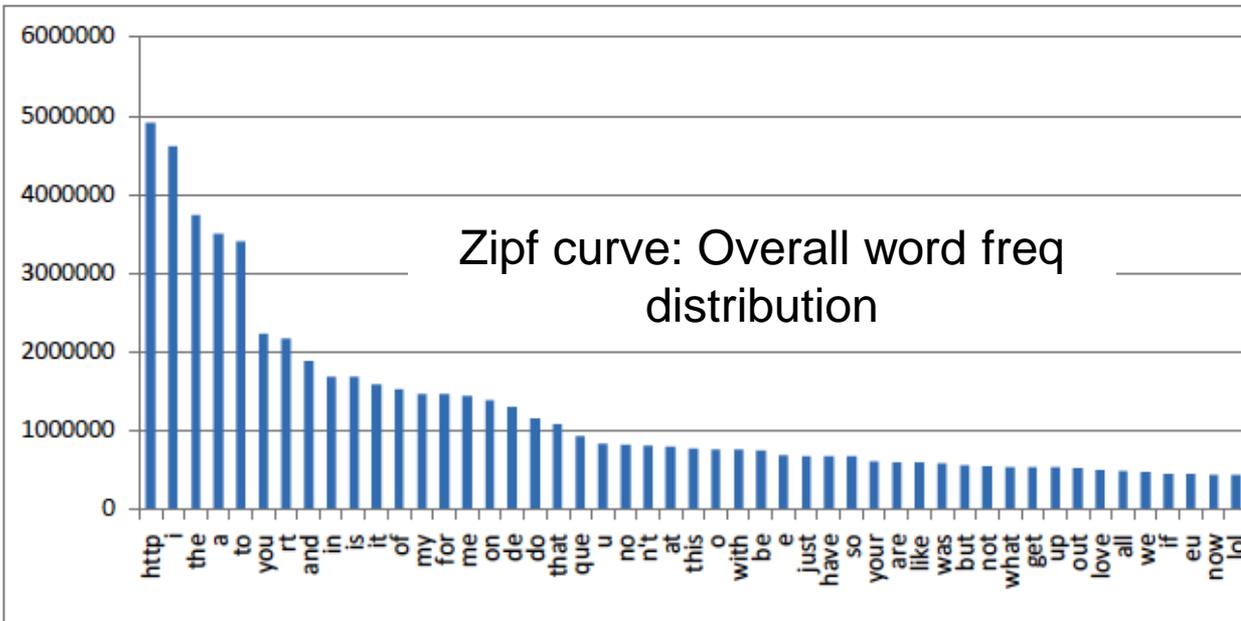
Tweet characteristics 1



Weekly vocab growth

Word frequencies

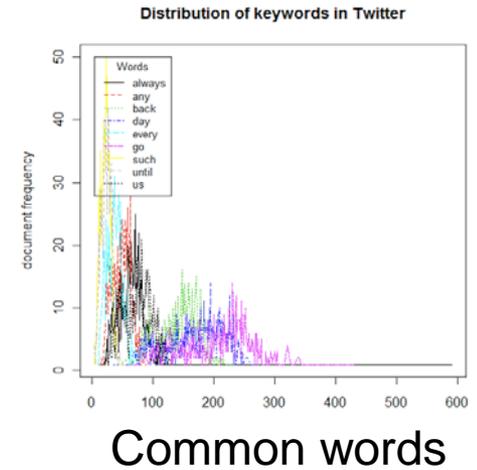
0 term	count	200 hey	118004
1 http	4913344	201 check	117406
2 i	4608795	202 always	117051
3 the	3738805	203 dont	116007
4 a	3505473	204 wanna	115912
5 to	3409943	205 first	112865
6 you	2231102	206 sama	111189
7 rt	2173270	207 were	110779
8 and	1889318	208 minha	110100
9 in	1685091	209 apa	110033
	83047	210 vc	110017
	90624	211 ini	109841
	36010	212 es	109822
	58836	213 r	109582
	54855	214 yeah	109462
	42648	215 hope	109438
	35128	216 ok	109381
	17157	217 morning	108966
	54579	218 watch	108047
	92580	219 tonight	107810
	33071	220 tweet	107609



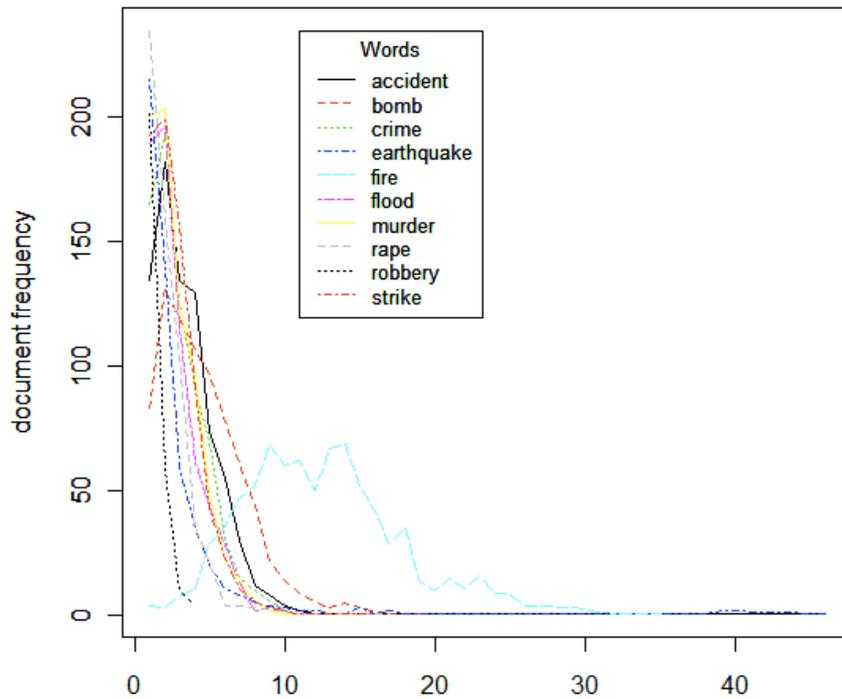
Zipf curve: Overall word frequency distribution

Tweet characteristics 2

Words come in bursts: word distributions per tweet

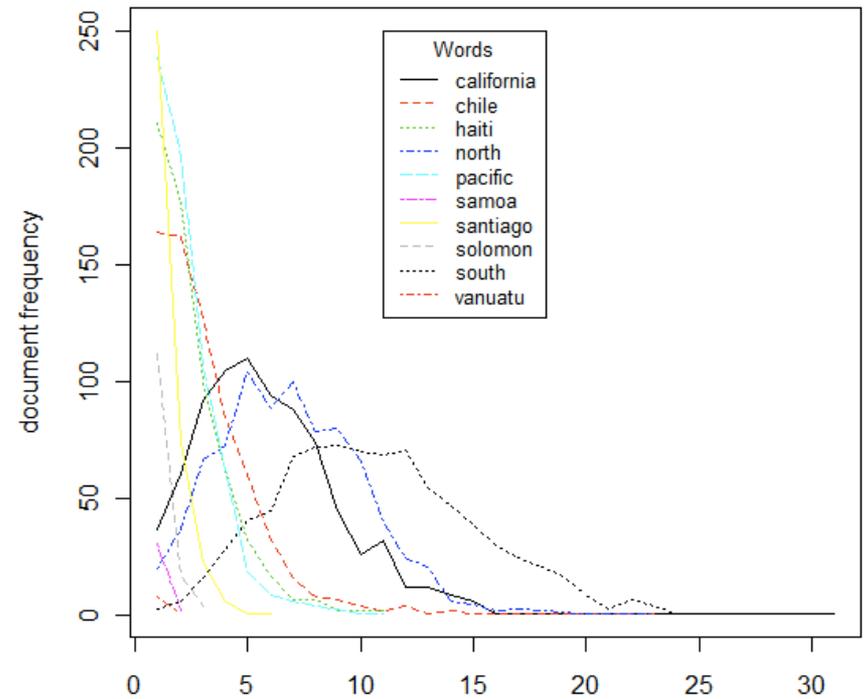


Event words



term frequency
events word plots

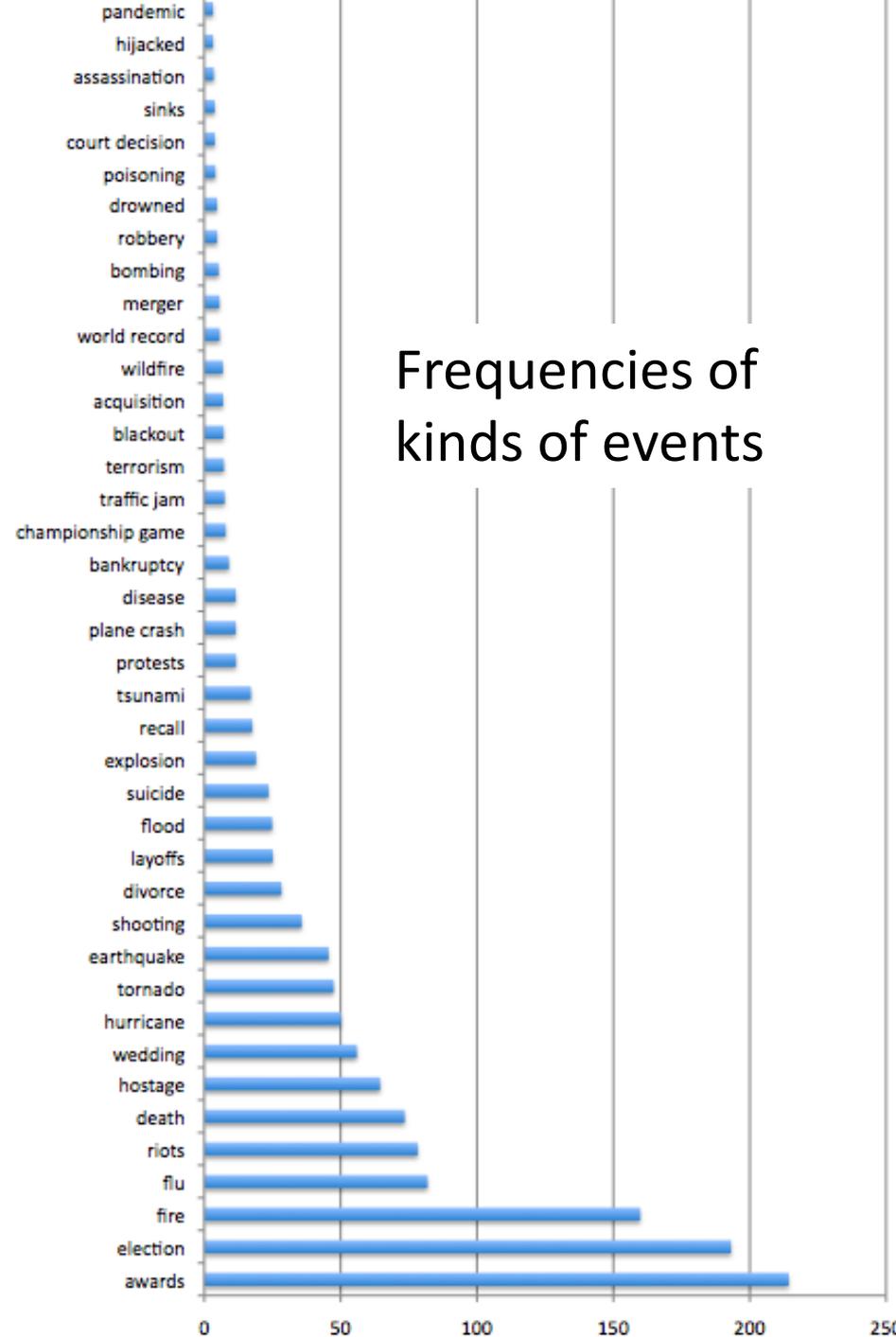
Location words



term frequency
location word plots

Finding kinds of events

- To learn topic signatures, we manually selected some good keywords, and (using 'burstiness') learned more co-occurring words automatically
- Studied 50 topics



0: ['flood']

signature: flood(0.074695), river(0.013374), water(0.012089), flooding(0.004749), areas(0.007776), control(0.011872), insurance(0.011557), relief(0.005406), levees(0.001876)...

* * * * * timeintervals ranked in 5.613292sec

(1281826800;1281830400),0.019086, August 14 2010 at 23 PST,1 hours,
(1280062800;1280070000),0.018256, July 25 2010 at 13 PST,2 hours,
(1280908800;1280912400),0.017363, August 04 2010 at 08 PST,1 hours,
(1287435600;1287439200),0.014191, October 18 2010 at 21 PST,1 hours, ...

1: ['tornado']

signature: tornado(0.027791), storm(0.003742), twister(0.001389), weather(0.003887), homes(0.002850), oklahoma(0.002345), damage(0.001924), kansas(0.002177)...

* * * * * timeintervals ranked in 5.565819sec

(1287399600;1287403200),0.014456, October 18 2010 at 11 PST,1 hours,
(1279922400;1279944000),0.004438, July 23 2010 at 22 PST,6 hours,
(1279850400;1279857600),0.002292, July 23 2010 at 02 PST,2 hours,
(1290459600;1290466800),0.001988, November 22 2010 at 21 PST,2 hours, ...

2: ['shooting']

signature: shooting(0.037280), police(0.009983), gunman(0.001702), school(0.005369), officers(0.002311), shooters(0.001069), shootings(0.000836), shooter(0.000851), ...

* * * * * timeintervals ranked in 5.552338sec

(1293649200;1293652800),0.000000, December 29 2010 at 19 PST,1 hours,
(1293033600;1293037200),0.000000, December 22 2010 at 16 PST,1 hours, ...

Measuring the 'burstiness'

- $Burstiness = P(w|D) / P(w)$
where $P(w|D)$ is the word frequency in the current time window and $P(w)$ is the background word frequency
- $P(w|D) = [tf(w,D) + \mu * cf(w)/N] / [|D| + \mu]$
- $P(w) = cf(w)/N$
where μ is a smoothing parameter, $\mu = 0.2$ current setting
- or $P(w) = (cf(w) + K) / (N + K*|V|)$
where K is the smoothing parameter and $|V|$ is the total number of words in the vocabulary. When $K=1$, this is known as Laplacian (or add one) smoothing. This adds a fixed number of observations (K) to the collection frequency of each term in the vocabulary; we use a relatively large K (e.g., 1,000,000)

Event evolution: Learning timelines

	NEW ZEALAND		MARIANA ISLANDS		ECUADOR		VANUATU	
1 day before	pabl,sommore,tectonics,hughley,andreanof,aleutian,1918,earthworms,tho..rt,yokohama	32	8:17pm,lavell,kuril,azores,3cm,ecuador,jacinto,15m,bumpy,pretended	19	9:56pm,somethingand,2:10am,ktika,amounted,jolted,abon,wakin',18:26,kuasa	30	lamberts,thourt,cautions,salem,earthquake-unpredictable,unicef,cyclones,exceeds,14m,wyclef	38
10 minutes after	aftershocks,christchurch,drawers,woken,powerful,monster,earthquake,opened,realize,child	2		0		0		0
1 hour after	ilam,mightnt,snoozes,nzi,tson,7.4,christchurch,chch,aftershocks,usgs	39	saipan,guam,mariana,215,216,predicted,widespread,tsunami,floods,islands	3	7.2m,186km,ambato,139km,6:54am,20km,ecuador,10m,6.9, revised	4	oce,vanuatu,ij,7.6,vila,7.5,kashmir,seismic,tsunami,combined	7
12 hours after	7.0),tvnz7,00:06,retweet.my,2010an,7.2-magnitude,hahaglad,destructddamage,rellies,photo.rt	288	jayapura,pb2,www.njslea.org,08:47,bonin,workgroup,mariana,guam,kulon,conce	30	119km,6:53am,dreampt,7.2m,ambato,186km,6.9-magnitude,139km,6:54am,#science	26	f16s,50km,grumbly,12:48am,vanuatu,honshu,oce,precuation,ij,35km	26
1 day after	shariff,yellow,4:55pm,kleff,9:52pm,_hands,n.z.rt,tsunami.my,thread],jesus.twitter	404	01:47,08:47,www.njslea.orgrt,night.knights,zinfandel=best,34:01,pb2,l-rd,jayapura,crackhouse	39	dinecuador,ecuador-colombia,azaab,hogi...,kehti,seekho!,ruknay,15:36,ecua,earthquake-affected	37	funniest/,uncle*,297km,50km,grumbly,portvila,11:39pm,65km,okvanuatu,12:48am	38
2 days after	devastates,anderton,helpsocial,haiti?,3news,earthquake-unpredictable,declaired,#quotes,thread],wikipedians	499	guitar*newly,quakefactor,recter..rect,4aftershocks,wellbrace,earthquakes/half,01:47,08:47,www.njslea.orgrt,night.knights	67	12/08/10,ecuador-colombia,dinecuador,info<,ruknay,azaab,hogi...,kehti,seekho!,15:36	75	14million,03:35,underw,lillyawork,mannington,sometimesxd,kmt...its,uncle*,funniest/,297km	64
1 week after	anatahan,faultline,99m,pb1,5:41am,earth(,vou,1356,earthquakei,delice	844	hagatna,19km,jaw-droppingly,epicent,6.3-magnitude,halmahera,kamchatka,07:35,taxesthe,2:28am	205	12/08/10,info<,311km,4:28am,www.garyowen.com,lavell,ecuador-colombia,dinecuador,ecua,15:36	217	churches<,church<,somethingeverynight,sindongnya,sissoalnya,crowdet,terorism,21:56,khorasan,agoyes	228

So what can we do?

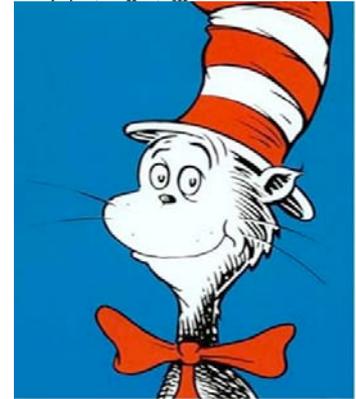
- Given some event of interest, we can learn its topic signature (if it's a large enough event)
 - **'Pulling' usage:** We can go and find specific events if we know when it occurred and roughly what it was
 - **'Pushing' usage:** We can build an Alert system that monitors the Twitter stream and tells you when something occurs
- **Exploration:** We can go and search for what other events reliably occur together with it
- **Other sources:** We can apply the techniques to websites (backpage.com, FaceBook...)

Example 1: Human Trafficking

- Question: can we identify instances of:
 - Child prostitution
 - Forced labor / slavery ?
- Activity:
 - We're building a 'daily broadsheet' that lists all recent activity from various sources (Backpage.com, Twitter, MySpace...)
- Interest from several collaborators:
 - Mark Latonero, USC School of Communications
 - FBI (Los Angeles Division): various people
 - Long Beach Police Department
 - US Equal Employment Opportunity Commission, LA District Office
 - DHS Investigations in Human Trafficking, LA

Example 2: Campus Crisis

- Question: How do people self-organize in the face of anomalies?
- Activity: The Mad Hatter experiment:
 - Competitions on 2 campuses: Find the (unknown) anomalous event + report it
 - Analyze students' SMS and twitter acts:
 - Propagation of info through network
 - Observation -> confirmation -> certainty
- Collaborators:
 - CCICADA team members at Rutgers U
 - CCICADA team members at RPI



Discussion

- Open problems:
 - Abnormal words: neologisms, weird spelling, foreign words, slang, etc.
 - Determining location of tweets / messages
 - Which events are of interest?
 - Events have structure: subevents 'within' events may require special treatment
 - Time window granularity: which is best? Same for all events?
- Do you know of any other uses?