

# Keepr

Algorithm for Extracting Entities, Eyewitnesses and Amplifiers

Hong Qu  
September 19, 2013

# About me



2003



2006



2012



2013

twitter

Login Join Twitter!

<http://twitpic.com/135xa> - There's a plane in the Hudson. I'm on the ferry going to pick up the people. Crazy.

*12:36 PM Jan 15th from TwitPic*




**jkrooms**  
Janis Krums




# New York plane crash: Twitter breaks the news, again

Twitter has once again led the media and the blogosphere in breaking news.



share photos on twitter

Rotate left → ← Rotate right



Posted on January 15, 2009  
by jkaums


View full size

Share this photo

Put this photo on your website




Views: 107122

Tags



Do you tweet about working out and getting fit? Join [www.TwitFilter.com](http://www.TwitFilter.com) to meet people just like you!

More photos by jkaums



Print this article

Share 29

Facebook 5

Twitter 24

Email

LinkedIn 0

+1 0

---

**Twitter**

Travel » World News  
Major News »  
Travel News »  
Technology »

Twitter lead the media and the blogosphere in breaking news about US Airways flight 1549 crashing in New York's Hudson river



# Teaching Applied Natural Language Processing: Triumphs and Tribulations

**Marti Hearst**

School of Information Management & Systems  
University of California, Berkeley  
Berkeley, CA 94720  
hearst@sims.berkeley.edu



## Abstract

In Fall 2004 I introduced a new course called Applied Natural Language Processing, in which students acquire an understanding of which text analysis techniques are currently feasible for practical applications. The class was intended for interdisciplinary students with a somewhat technical background. This paper describes the topics covered and the programming exercises, emphasizing which aspects were successful and which problematic, and makes recommendations for future versions of the course.

foundations and core NLP algorithms. Several computer science students took both courses, and thus learned both the theoretical and the applied sides of NLP. Dan and I discussed the goals and content of our respective courses in advance, but developed the courses independently.

## 2 Course Role within the SIMS Program

The primary target audience of the Applied NLP course were masters students, and to a lesser extent, PhD students, in the School of Information Management and Systems. (Nevertheless, PhD students in computer science and other fields also took the course.) MIMS students (as the SIMS masters students are known) pursue a professional degree studying information at the intersection of tech-

# Automated Blog Classification: Challenges and Pitfalls

Hong Qu, Andrea La Pietra, Sarah Poon

University of California at Berkeley  
School of Information Management & Systems  
314 South Hall, Berkeley, CA, USA 94720-4600  
{hqu, lapietra, sspoon}@sims.berkeley.edu

## Abstract

Blogs are difficult to categorize by humans and machines alike, because they are written in a capricious style. In the early days of web, directories maintained by humans could not keep up with millions of websites; likewise, blog directories cannot keep up with the explosive growth of the blogosphere. This paper investigates the efficacy of using machine learning to categorize blogs. We design a text classification experiment to categorize one hundred and twenty blogs into four topics: personal diary, news, political, and sports. The baseline feature is unigrams weighed by TF-IDF, which yielded 84% accuracy. We analyze the corpus, features, and result data. Our analysis leads us to believe that blog taxonomies need to support polyhierarchy—a given blog may be correctly classified under more than one category.

because blogs do not fit neatly in mutually exclusive categories: a particular blog can fall into multiple categories.

## 2. Previous Work

There are many definitions for what constitutes a blog. For our purpose a blog is a website for personal expression composed of “webpages that are constantly updated with new commentary and links about a particular topic. Often very personal [3].” However, it may be too early to group blogs into a directory based on an arbitrary taxonomy because “our collective conceptions of weblogs are changing too quickly to realistically capture them in such frameworks [4].”



**Hong**

@hqu



Tweets with links inadvertently tag the webpages and media they point to with contextual metadata #search #algorithm #metadata #socialmedia

← Reply ↻ Retweet ★ Favorite ●●● More

5:14 AM - 9 May 09

<https://twitter.com/hqu/status/1745171763>



# Analyzing Big Data with Twitter

*A special UC Berkeley iSchool course*



UC Berkeley School of Information



[Home](#) [About](#) [Projects](#) [Syllabus](#) [Assignments](#)

## UC Berkeley Course Lectures: Analyzing Big Data With Twitter

Posted on [13 December, 2012](#) by [Marti Hearst](#)

Thank you all for a wonderful semester. Here is a summary, in chronological order, of our recorded lectures. You can also [view the entire playlist](#) on youtube.

### Archives

- [December 2012](#)
- [November 2012](#)
- [October 2012](#)

# Nieman Application

"My proposal to is to automatically summarize highlights of any live-tweet event. I intend to apply computational linguistic and wisdom of the crowd heuristics to identify and highlight tweets that has the ~~strongest resonance~~. This algorithm would "filter" thousands of tweets to automatically and distill meaningful discourse out of noisy chatter."

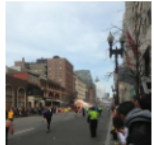


CNN Breaking News @cnnbrk Follow

Explosion reported near Boston Marathon finish line, CNN affiliate WCVB reports. [bit.ly/117FlHz](http://bit.ly/117FlHz)

7:07 PM - 15 Apr 13 (UTC)

**Explosions reported near Boston Marathon finish line**



Two explosions have been reported near the Boston Marathon finish line on Boylston Street, and there are reports of injuries.



WCVB-TV Boston @WCVB

13,363 RETWEETS 713 FAVORITES



The Associated Press @AP Follow

**BREAKING:** Intelligence official: 2 more explosive devices found at Boston Marathon; being dismantled

8:19 PM - 15 Apr 13 (UTC)



boston

texas

police

General Pervez

Former Pakistan President

McLennan County

Texas Gov

Twitter Music

Kimberley Hainey

obama

suspects

fertilizer

AP

The Associated Press

@AP

Follow

**BREAKING: Law enforcement official: Arrest imminent in Boston Marathon bombing, suspect to be brought to court.**

1:42 PM - 17 Apr 2013

7,176 RETWEETS 498 FAVORITES



CNN

CNN Breaking News

@cnnbrk

Follow

**Law enforcement sources: Arrest made in the Boston bombings investigation. [on.cnn.com/15fOGEd](http://on.cnn.com/15fOGEd)**

1:56 PM - 17 Apr 2013

**Sources: Possible suspects sought in Boston blasts**

After law enforcement sources told CNN that an arrest was made in Boston Marathon bombings, two senior administration officials and another federal official told CNN contributor Fran Townsend that no...



CNN @CNN

3,809 RETWEETS 392 FAVORITES



Tweets

CNN

CNN Breaking News @cnnbrk

9m

Police in #Watertown announced on bullhorn that a suspect is in custody, according to CNN photographer on scene.

Expand

BBC NEWS

BBC Breaking News @BBCBreaking

13m

1 person arrested in #Watertown, close to Boston, as major police operation continues after officer killed at #MIT [bbc.in/1514fKG](http://bbc.in/1514fKG)

Expand

CNN

CNN Breaking News @cnnbrk

14m

Dozens of police rushed to area of Watertown, Mass., about 2 miles from Cambridge. Reports that explosives were involved.

Expand

BREAKING NEWS

Breaking News @BreakingNews

28m

Scanner: Police told to power down cellphones following reports of explosive devices, shooting in Watertown - Boston Globe's @Billy\_Baker

Expand

BBC NEWS

BBC Breaking News @BBCBreaking

37m

Multiple witnesses report gunfire & booms in #Watertown, about 10 miles west of Boston, after officer dies at #MIT [bbc.in/15kZL6T](http://bbc.in/15kZL6T)

Expand

BBC NEWS

BBC Breaking News @BBCBreaking

1h

Update: Police officer dies after being shot several times at #MIT campus near Boston in US [bbc.in/1ZBxYd1](http://bbc.in/1ZBxYd1)



Sunil

Search for News

Pete Williams tripathi

boston suspect

watertown Martin Richard

police scanner

@stoolpresidente @BuzzFeedJack

@MichaelSkolnik @kmattio

http://x.co/1272N

 **Anonymous**  
@YourAnonNews Follow

Police on scanner identify the names of #BostonMarathon suspects in gunfight, Suspect 1:



Pete Williams of MSNBC: sources say suspect #2 NOT Sunil Tripathi. #Watertown 3 minutes ago



Pete Williams at NBC is saying categorically that Sunil Tripathi is NOT suspect 2. Another huge Twitter bandwagon busted? 4 minutes ago



Pete Williams going all in saying suspect 2 (white hat) is NOT Sunil Tripathi #Watertown #boston #BostonMarathon #p2 #p2by #dems 4 minutes ago



NBC's Pete Williams: Speculation that one of suspects (white-hat, Sunil Tripathi) is a missing student is NOT correct. 4 minutes ago



This Sunil kid is going to come back from his unannounced extended vacay

Sunil → Pete Williams



suspects watertown

Search for News

police

boston

custody

Police Battle Suspects

OFFICER ASSIST

Current Watertown

Police Radio

marathon

bostonmarathon

@YourAnonNews

@universalhub

@aravosis

@megturney

<https://twitter.com/sethmnookin/>

<http://on.wcvb.com/XPieWS>



WCVB-TV Boston

@WCVB

Follow

**#BREAKING:** Gunfight between police, suspects reported on Dexter Street in Watertown, Mass. Live Video: [on.wcvb.com/XPieWS](http://on.wcvb.com/XPieWS)

1:00 AM - 19 Apr 2013

**MIT police officer killed; reports of gunfight, explosions**

An MIT police officer was shot and killed Thursday night. In addition, police were led on a chase into Watertown, where a gunfight and explosives were detonated.



[@WardCrap](#) [@itsjuststarla](#) Timeline: MIT shooting, carjacking nearby, grenades/explosions/gunfire reported in Watertown, 1-3 suspects thus far  
[18 seconds ago](#)



[@bostonradio](#) the guy doesn't look like one of the suspects [#MITShooting](#) [#watertown](#) [#nesn](#)  
[24 seconds ago](#)



Police Battle Suspects with Grenades, Automatic Weapons in Watertown Mass. <http://t.co/mpStC7J46w>  
[34 seconds ago](#)

Tweets All / No replies

 **K 80 Michael Skolnik** @MichaelSkolnik 50s  
There is NO confirmation at this moment of any connection between the Boston Marathon bombings, the MIT shooting and the Watertown shootout.  
Expand


 **K 80 Michael Skolnik** @MichaelSkolnik 6m  
"we do not have a 2nd suspect in custody" just heard from police. #WTF  
Expand

 **K 80 Michael Skolnik** @MichaelSkolnik 7m  
loud boom just heard in Watertown.  
Expand Reply Retweet Favorite More

 **K 80 Michael Skolnik** @MichaelSkolnik 7m  
No shots fired at emergency area of Boston Children's Hospital, however the hospital is still lockdown.  
Expand


 **K 71 GlobalGrind Politics** @GGPolitics 9m  
JUST IN: Photos of the shootout & explosions in Watertown after one officer is killed at MIT [bit.ly/YyMaSp](http://bit.ly/YyMaSp)  
Retweeted by Michael Skolnik  
Expand

 **K 80 Michael Skolnik** @MichaelSkolnik 11m  
picture of one of the suspects in custody. #watertown [twitpic.com/ckcg3j](http://twitpic.com/ckcg3j)  
View photo

 **K 80 Michael Skolnik** @MichaelSkolnik 13m  
no active shooter at this time  
Expand

 **K 80 Michael Skolnik** @MichaelSkolnik 13m  
Soldiers have arrived and walking around perimeter. Probably National Guard.  
Expand


 **K 80 Michael Skolnik** @MichaelSkolnik 11m  
picture of one of the suspects in custody. #watertown [twitpic.com/ckcg3j](http://twitpic.com/ckcg3j)  
View photo


 **K 80 Michael Skolnik** @MichaelSkolnik 13m  
no active shooter at this time  
Expand Reply Retweet Favorite More


 **K 80 Michael Skolnik** @MichaelSkolnik 13m  
Soldiers have arrived and walking around perimeter. Probably National Guard.  
Expand


 **K 80 Michael Skolnik** @MichaelSkolnik 16m  
active shooter.  
Expand

 **K 80 Michael Skolnik** @MichaelSkolnik 18m  
air support has been called in  
Expand

 **K 80 Michael Skolnik** @MichaelSkolnik 19m  
The person they just handcuffed and put into the vehicle was stripped naked. They have NOT ID'd him yet as 2nd suspect.  
Expand

 **K 80 Michael Skolnik** @MichaelSkolnik 19m  
Boston PD now reporting 2nd suspect may still be at large. They may not have 2nd suspect. Oh shit.  
Expand

 **K 80 Michael Skolnik** @MichaelSkolnik 22m  
"vehicle has been cleared, waiting on ID of 2nd suspect."  
Expand

 **K 80 Michael Skolnik** @MichaelSkolnik 23m  
Boston Chief Linskey on scene.  
Expand

 **K 80 Michael Skolnik** @MichaelSkolnik 23m  
"This is still extremely dangerous," - FBI officers -- could be explosives still on streets, in backyards, etc.

## Tweets



**K** 81 **Seth Mnookin** @sethmnookin

3m

More police loudspeaker deep into perimeter; can't make out what said over sound of helicopters.

Expand



**K** 81 **Seth Mnookin** @sethmnookin

3m

Not sure if this is why police wanted to use my phone but they were told to shut theirs down b/c of detonation concern.

Expand



**K** 81 **Seth Mnookin** @sethmnookin

7m

Some perspective for our of towners: MIT shooting scene 15 min drive away from this scene in Watertown. Likely over 100 officers on scene.

Expand



**K** 81 **Seth Mnookin** @sethmnookin

9m

When I got here, police were asking to use my phone to navigate neighborhood. Two mins later, told in no uncertain terms to leave area.

Expand



**K** 81 **Seth Mnookin** @sethmnookin

10m

I didn't hear any gunshots at scene (Dexter St & Nichols). Lots of talk on scanner, though.

Expand



**K** 81 **Seth Mnookin** @sethmnookin

12m

Actual perimeter just now set up. My car one block beyond this.

[pic.twitter.com/SvnyC1nbVz](https://pic.twitter.com/SvnyC1nbVz)

[View photo](#)



**K** 81 **Seth Mnookin** @sethmnookin

13m

.@taylordobbs @brianjdamico & I only reporters on scene for 10 mins. No way to see the suspect. Reports of skin color pure speculation.

Expand

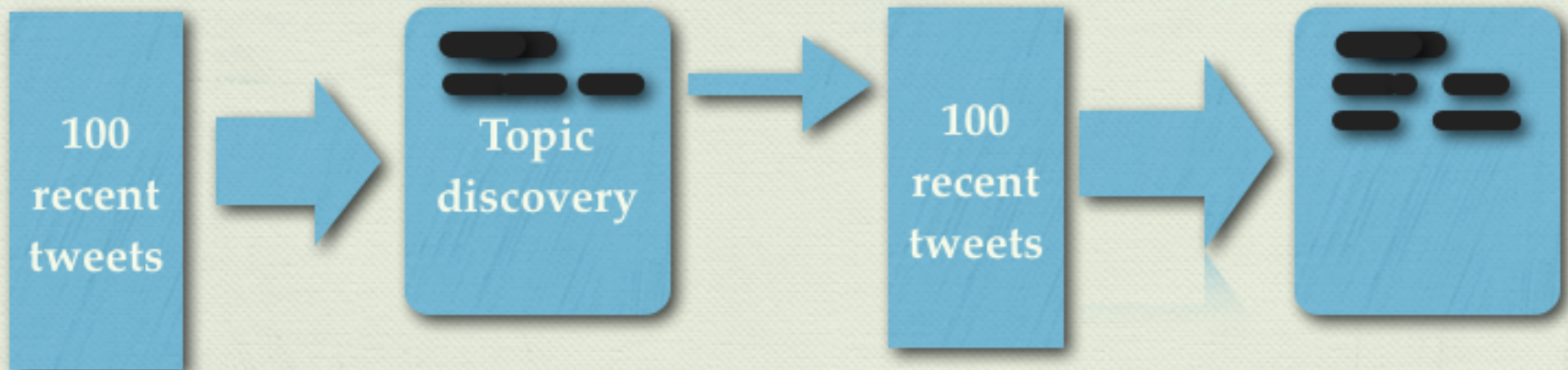


# Algorithm

<http://www.keepr.com/demo>

Search any topic

OR "from:cnnbrk OR from:breakingnews OR from:BBCBreaking OR from:AP+BREAKING:"



Pattern  
Recognition

Abnormality  
Detection

Predictive  
Models

boston

custody

bostonbombing

Retweet This

Jake Tepper

CRAZY SHIT

Videos Said

Boston Police

police

mitshooting

@CNN

@LuckyCharms321

@JohnnyHopkins91

http://on.cnn.com/175CKDq



**CNN Breaking News** ✓

@cnnbrk

Follow

Federal law enforcement official says the 2 suspects stayed at scene to watch the #Boston carnage unfold.

[on.cnn.com/175CKDq](http://on.cnn.com/175CKDq)

8:54 PM - 18 Apr 2013

**FBI: Help us ID Boston bomb suspects**

Boston will pause to mourn, and heal, Thursday after the fatal terror attack on the city's beloved marathon.

**CNN** @CNN



871 RETWEETS 110 FAVORITES



**Stay up to date**

Enter your email

Tweets containing

2 suspects



@cookie\_lass @xpeanutgalleryx

there are 2 suspects.

26 seconds ago



Chaos in Watertown, MA after fatal shooting of officer at #MIT. 2 suspects heavily armed chased by PD. One suspect in custody. #GreatBritain

27 seconds ago



Officer dead in #mit #watertown shooting. 2 Suspects threw explosives out car window as they were being chased=1 suspect poss still on loose

36 seconds ago



@Millsy11374 And M.I.T. officer was shot & killed, suspects on loose. Then 2 suspects carjacked a mercedes, police in pursuit. suspects.

45 seconds ago



So they are looking for 2 suspects for the marathon bombing and they are also looking for 2 suspects for MIT/WaterTown shooting. Hmm

49 seconds ago

boston

mits shooting

boston bombing

police

Current Watertown

both

@CNN

@MichaelSkolnik

@aravosis

@WCVB

@TwitPic

http://twitpic.com/ckcg3j



Anonymous

@YourAnonNews

Follow

Recap: IED explosion in #Watertown, near Boston. 2 Suspects in custody. Unconfirmed if related to #BostonBombing or #MITShooting.

1:35 AM - 19 Apr 2013

359 RETWEETS 26 FAVORITES



Michael Skolnik

@MichaelSkolnik

Follow

picture of one of the suspects in custody. #watertown  
twitpic.com/ckcg3j

1:34 AM - 19 Apr 2013



TwitPic @TwitPic

168 RETWEETS 20 FAVORITES



#Watertown MIT One of the suspects going to Beth Israel Hospital, The other in custody now. The Police Officer injuries were fatal.

25 seconds ago



Recap: IED explosion in #Watertown, near Boston. 2 Suspects in custody. Unconfirmed if related to #BostonBombing or #MITShooting.

#Anonymous

25 seconds ago



Recap: IED explosion in #Watertown, near Boston. 2 Suspects in custody. Unconfirmed if related to

#BostonBombing or #MITShooting.

30 seconds ago



Two suspects in custody in Watertown, Mass.

32 seconds ago



Officers down. Suspects stole police SUV. One naked man in custody. Other suspect status unknown.

#Watertown

33 seconds ago



Shooting at MIT; suspects found in Watertown. Both were taken into custody. One was naked. Reports of explosives FBI/DHS there. Now CNN

35 seconds ago



: Recap: IED explosion in #Watertown, near Boston. 2 Suspects in custody. Unconfirmed if related to #BostonBombing or #MITShooting."



police

twitter

watertown

listening

Auburn Hospital

people

time

@sethmookin



**Ginger Gibson**

@GingerGibson

Follow

I've listened to a lot of scanners in my time (cops reporter) and I would caution everyone to be careful what you report off the scanner.

12:56 AM - 19 Apr 2013

73 RETWEETS 16 FAVORITES



**Stephanie McMaster**

@Smethanie

Follow

What I heard on scanners: Two dudes carjacked a Mercedes in Cambridge at a gas station, then stole police SUV, were chased to Watertown...

1:36 AM - 19 Apr 2013

76 RETWEETS 6 FAVORITES



**Les Perreux**

@perreux

Follow

Remembering the disjointed way police scanners work, I'd say combining them with twitter pretty much guarantees headsplisions.

1:32 AM - 19 Apr 2013

42 RETWEETS 9 FAVORITES



Stay up to date

Enter your email



I'm waiting to hear The Joker hijack the police scanners.

22 seconds ago



@Jim\_Gardner what/who are scanners?

24 seconds ago



Listening to police scanners always reminds me of farva in supertroopers.

#doyouneedmyassistance #carramrod  
24 seconds ago



@watson5j I'm used to having four police scanners playing in the newsroom at once! Just not on my laptop!

25 seconds ago



@ckanal Seems more complicated. Twitter & Reddit mainly echoing accounts of local TV (stream) & newspaper reporters (on Twitter) & scanners

26 seconds ago



@CathyBrowne Glued to TV/twitter/scanners.

28 seconds ago



I'm listening to 2 police scanners at once.. so many points of data, shit

28 seconds ago



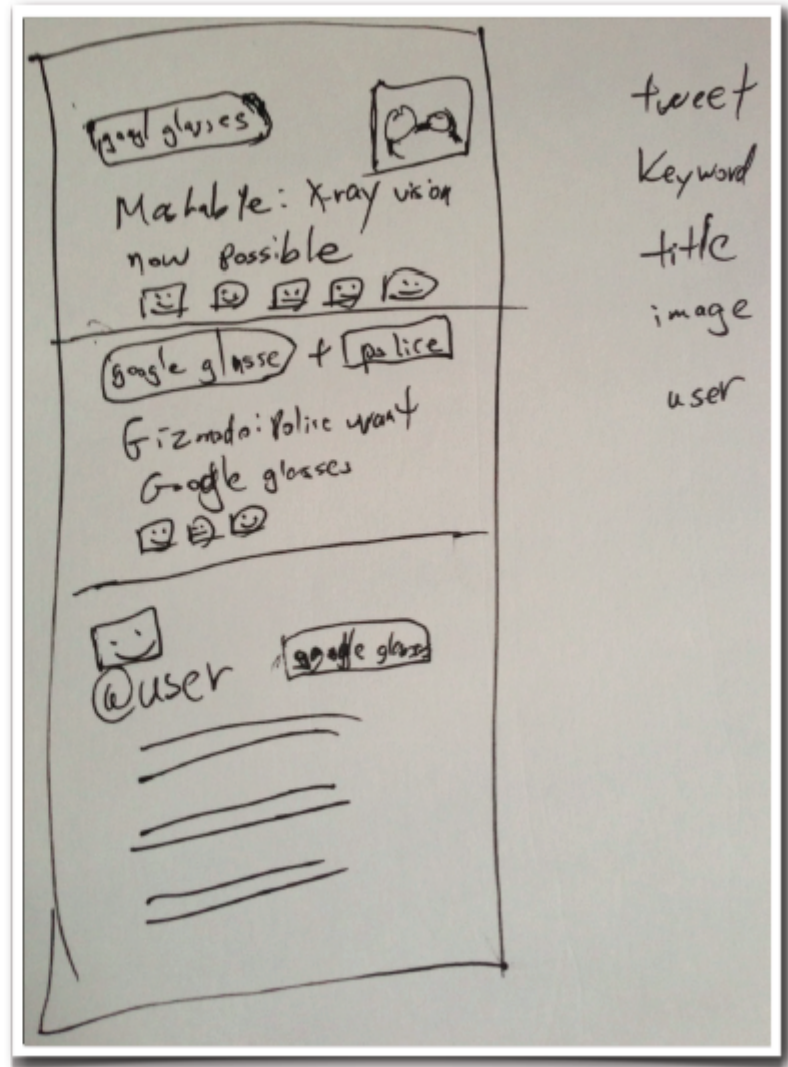
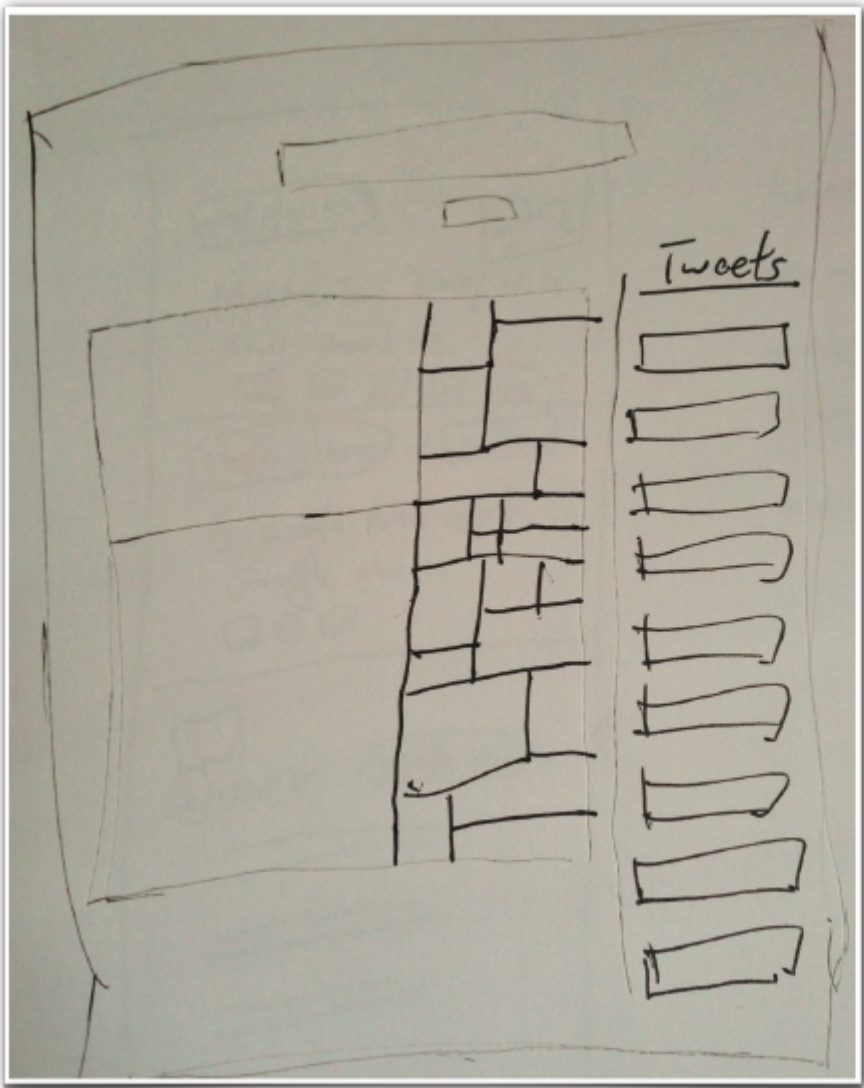
@nfdraftscout where are you listening to the scanners?

29 seconds ago



They got 2 suspects, one at gunpoint, the other captured.. but there are





tweet  
 Keyword  
 title  
 image  
 user

watertown

Search for News

- boston suspect police
- One Marathon Boston Marathon
- Army Lacrosse News Update
- boat arrest obama
- @stoolpresidente @sethmnookin
- <http://bit.ly/11tldPs>

 **The Boston Globe**   [Follow](#)  
@BostonGlobe

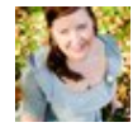
One Marathon bombing suspect has been caught, and another is on the loose in Watertown after a firefight with police, officials said.



Hat's off to @WCVB crew! Been great coverage all day. Photographers getting great video all week. #wcvb #bostonmanhunt #watertown  
16 seconds ago



In critical but stable condition. Bleeding in the lungs. Which means he could still drown, despite being in a boat, on land, in Watertown.  
16 seconds ago



My favorite line from the night: "If you're going to escape on a boat, make sure it's in water, not Watertown."  
19 seconds ago



Boston Marathon bomber manhunt: Police nab suspect alive | The Lookout - Yahoo! News <http://t.co/5rUb0vPsOx> via @YahooNews  
20 seconds ago



To bad it takes something of this magnitude to bring our country united .  
#Boston #Watertown  
20 seconds ago





## watertown

A public list by Hong

21

MEMBERS

30

SUBSCRIBERS

Subscribe

Tweets >

List members >

List subscribers >

Recently added members · [View all](#)



**Donald R. Winslow** @donaldrwin...

Follow



**Brian D'Amico** @brianjdamico

Follow



**Victoria Warren** @vwarrenon7

Follow



**Matthew Gregoire** @BreakngNe...

Follow

More lists by [@hqu](#) · [View all](#)

[Egypt](#)  
[Coupons](#)  
[readings](#)  
[watertown](#)  
[senators NO vote](#)  
[reporters](#)  
[breakingnews](#)  
[Boston Startup](#)  
[appsPriceDrop](#)  
[politics](#)  
[fun](#)  
[sf](#)

© 2013 Twitter [About](#) [Help](#) [Terms](#) [Privacy](#)  
[Blog](#) [Status](#) [Apps](#) [Resources](#) [Jobs](#) [Ads](#)  
[Advertisers](#) [Businesses](#) [Media](#) [Developers](#)

## List members



**Donald R. Winslow** @donaldrwinslow

Editor of News Photographer magazine for NPPA. Writer, photojournalist, runner, teaches at John Cabot University & John Felice Rome Center. [magazine@nppa.org](mailto:magazine@nppa.org)



Follow



**Brian D'Amico** @brianjdamico

Breaking news enthusiast & freelance photographer. Chemistry undergrad @ Northeastern. Posting mostly breaking news & photography. (Other me: [@damicobrian](#))



Follow



**Victoria Warren** @vwarrenon7

Victoria Warren is a 7NEWS General Assignment Reporter and volunteer with the [@FpiesFoundation](#)



Follow



**Matthew Gregoire** @BreakngNewsPhtg

Previously [@ProvFireVids](#). Freelance News Photographer for Fox 25 & LNS WBZ Boston. Retired from Woonsocket, RI Fire Dept L732. Breaking News Alerts



Follow



**Alexander Rowsell** @AureliusR

A guy who's crazy into computers and electronics. Going to try and make a career out of it, in fact!



Follow



**Steve** @alertnewengland

Posting breaking police/fire/EMS news heard on the scanner from across New England (but mostly Eastern Mass).



Follow



**Eric Moskowitz** @GlobeMoskowitz

Boston Globe Metro reporter/features writer. Tweeting mostly about work. Somerville resident, Needham native, Celtics ticket-holder.



Follow



**Andrew Kitzenberg** @AKitz

Founder of [@GetOnHand](#) - Instagram - a\_kitz  
You can call me Kitz



Follow



**Sean Kelly** @SeanKellyTV

Reporter for WCVB Newscenter 5. Opinion tweets are my own. Follow me on Instagram [@SeanKellyTV](#)



Follow



**Seth Mnookin** @sethmnookin

I teach at MIT's Grad Program in Science Writing & wrote The Panic Virus, Feeding the Monster & Hard News. Site <http://bit.ly/fw3Z0t> blog <http://bit.ly/mPVOCY>



Follow



# Humans vs Machines

## HUMANS ARE GOOD AT

Meaning

Feeling

Body sensory experience

Opinion

Emotion

Creative

Compelling

Exchange

Value Judgment

## MACHINES ARE GOOD AT

memory

matching

messaging

Networked

mining

Indexing structure data

Collaborative filtering

NLP Computational linguistics

Pattern Recognition

# Computers count really, really fast!

## TFIDF

For a term  $i$  in document  $j$ :

$$w_{i,j} = tf_{i,j} \times \log \left( \frac{N}{df_i} \right)$$

$tf_{ij}$  = number of occurrences of  $i$  in  $j$

$df_i$  = number of documents containing  $i$

$N$  = total number of documents

# What is a Tweet?

---

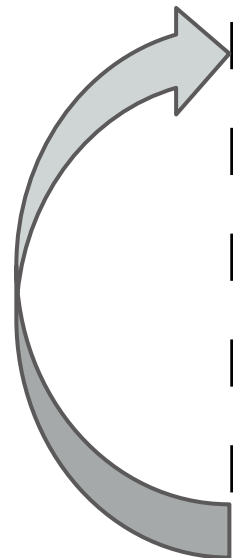
140 characters:

- words
  - @user mentions
  - #hashtags
  - links
-

# How does Keepr process tweets?

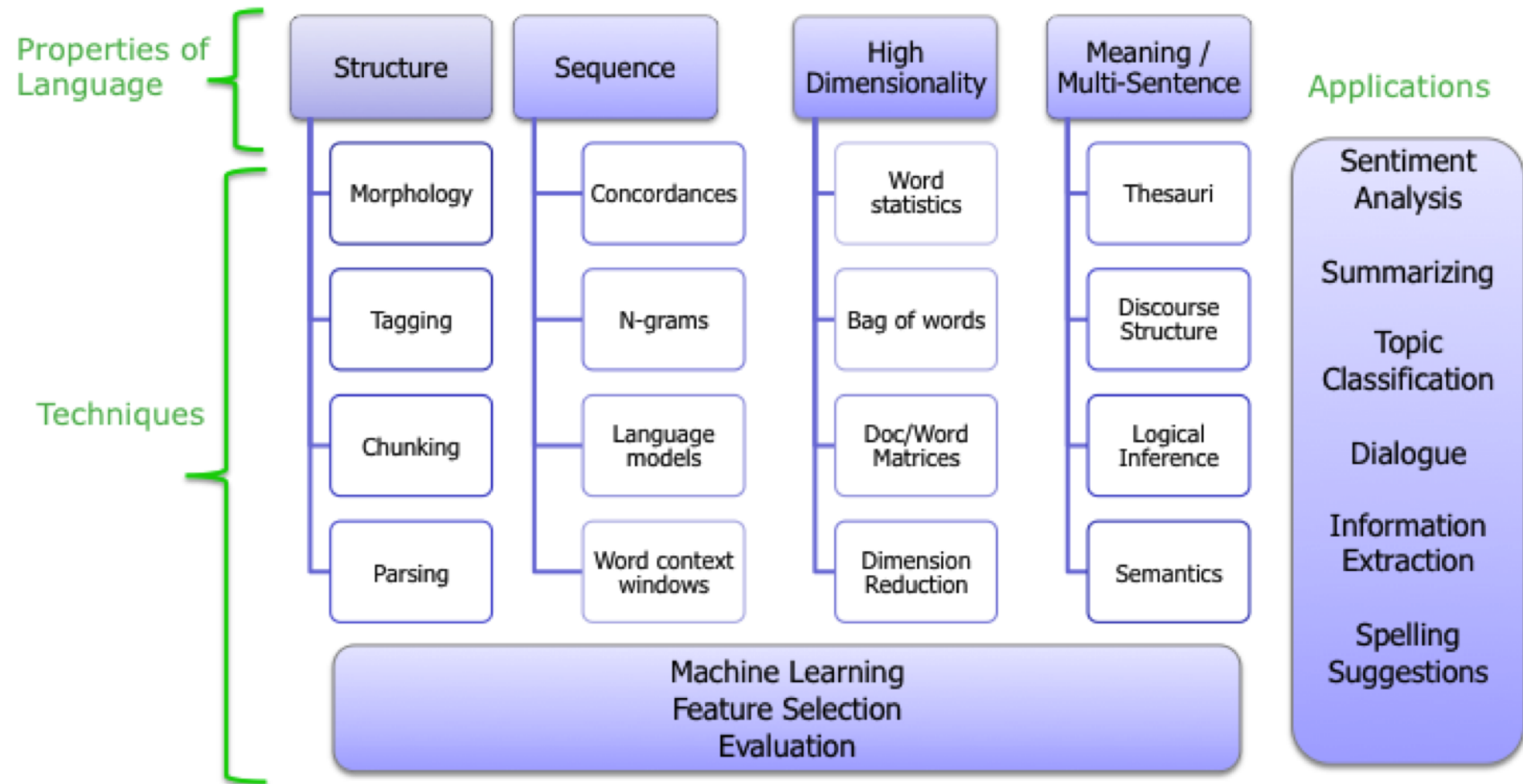
---

140 character \* 100 tweets = 14,000 characters

- 
- ❑ Parse it
  - ❑ Count it
  - ❑ Visualize it
  - ❑ Zoom in
  - ❑ Archive it
-



# Natural Language Processing



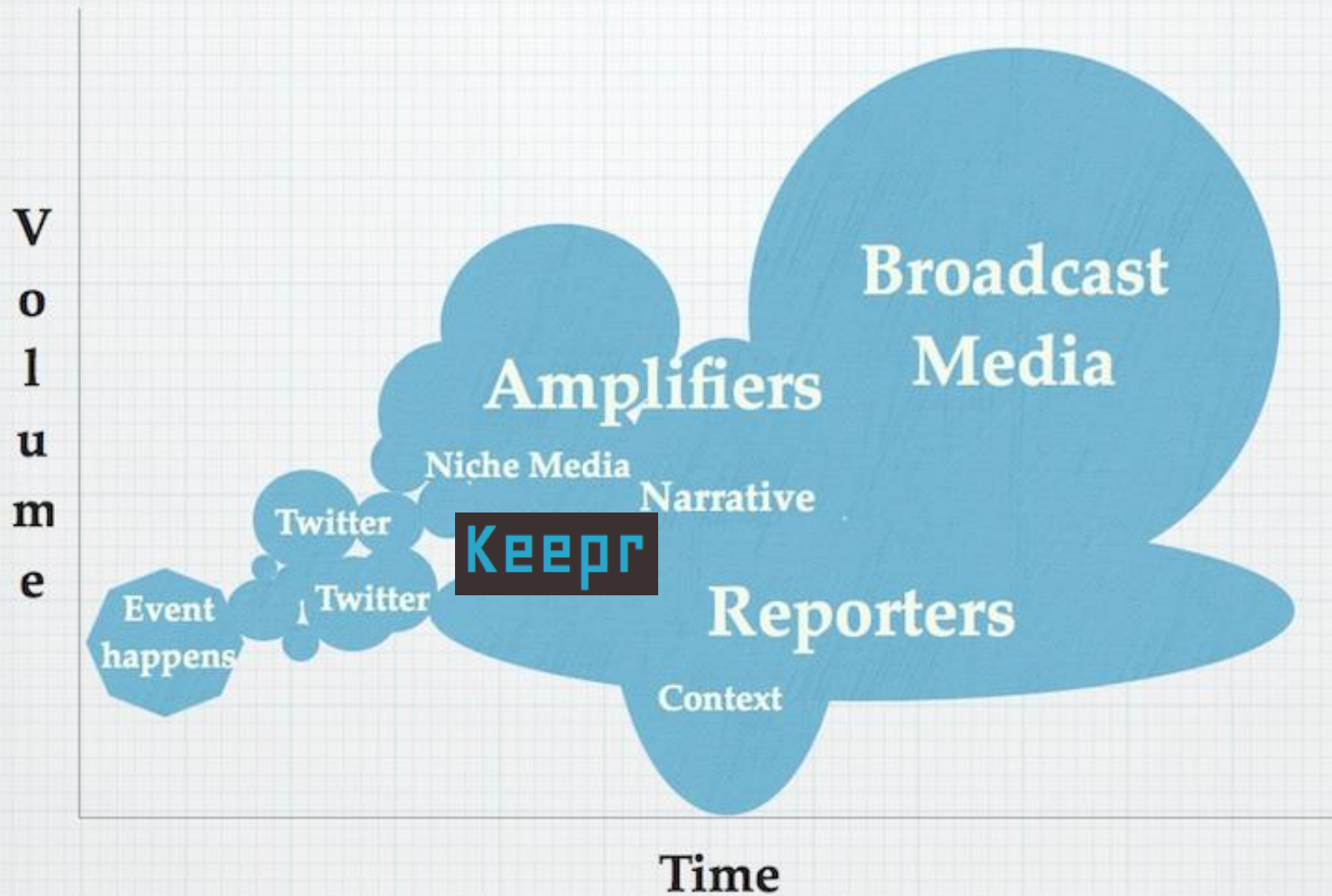
**BUT**



**Humans are way better at  
at making value judgements  
and telling stories**

---

# Breaking News in Social Media



[Social media and the Boston bombings: When citizens and journalists cover the same story](#)

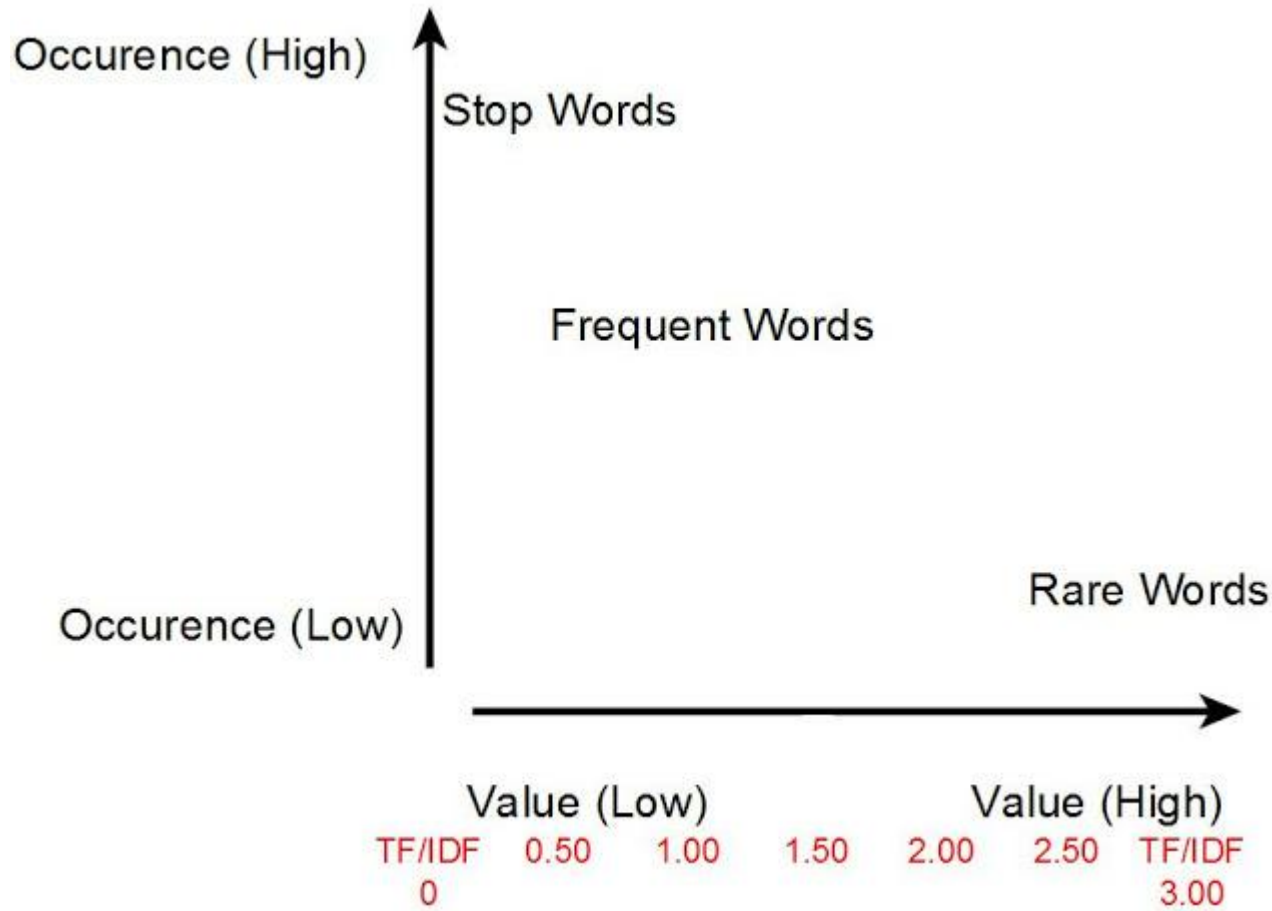


# Keepr's Algorithm



- Entity extraction
  - ◆ Topics
- Media extraction
  - ◆ images, videos
- Link expansion
  - ◆ articles
- Conversation analysis
  - ◆ @ mentions
  - ◆ source discovery
  - ◆ amplification velocity
- Source verification
  - ◆ geo-location
  - ◆ social media profiles

# Topic Extraction by Term Frequency



# Bursty and Hierarchical Structure in Streams \*

Jon Kleinberg †

## Abstract

A fundamental problem in text data mining is to extract meaningful structure from document streams that arrive continuously over time. E-mail and news articles are two natural examples of such streams, each characterized by topics that appear, grow in intensity for a period of time, and then fade away. The published literature in a particular research field can be seen to exhibit similar phenomena over a much longer time scale. Underlying much of the text mining work in this area is the following intuitive premise — that the appearance of a topic in a document stream is signaled by a “burst of activity,” with certain features rising sharply in frequency as the topic emerges.

# Journalists want

- ❑ Source discovery and curation
- ❑ Passive monitoring and alerts
- ❑ Saving and archiving
- ❑ Visualizations
- ❑ Parity with TweetDeck user interface

# People want

“I want to catch up with a summary of key information about the breaking news story.”

*I want to get a list of Twitter accounts who are official organizations related to that story.*



# My Musing

NOVEMBER 6, 2009

## SOCIAL MEDIA IN POSTMODERNITY

Mass media **homogenizes**. Social media **democratizes**, countering the culture industry's profit maximizing tendencies, thereby unwinding the **merchants of cool's** hegemonic grip on emerging and fleeting **spectacles** that perpetuate our postmodern economy's natural propensity to fabricate **somatic hyper-reality**.

# VERIFICATION JUNKIE

A growing directory of tools for verifying, fact checking and assessing the validity of social media and user generated content.

Managed by Josh Stearns.  
Send tips to @jcstearns on Twitter.

ABOUT

SUBSCRIBE VIA RSS

ARCHIVE

Telpher by Theme Static

The screenshot shows a Twitter search interface for the term "Watertown". At the top, there are search filters for "boston", "mitshooting", "bostonbombing", "police", and "Current Watertown". A red circle highlights the word "both" in the search filters. Below the filters, there are several tweets. The first tweet is from "Anonymous" (@YourAnonNews) and says "Recap: IED explosion in #Watertown, near Boston. 2 Suspects in custody. Unconfirmed if related to #BostonBombing or #MITShooting." The second tweet is from "Michael Skolnik" (@MichaelSkolnik) and says "picture of one of the suspects in custody. #watertown twitpic.com/ckeg3j". Below this tweet is a photo of a person lying on the ground. To the right of the main tweet, there is a list of related tweets. Several of these tweets have red boxes around the phrase "2 Suspects in custody" or "both". For example, one tweet says "#Watertown MIT One of the suspects going to Beth Israel Hospital. The other in custody now. The Police Officer injuries were fatal." Another tweet says "Recap: IED explosion in #Watertown, near Boston. 2 Suspects in custody. Unconfirmed if related to #BostonBombing or #MITShooting." A third tweet says "Recap: IED explosion in #Watertown, near Boston. 2 Suspects in custody. Unconfirmed if related to #BostonBombing or #MITShooting." A fourth tweet says "Two suspects in custody in Watertown, Mass." A fifth tweet says "Officers down. Suspects stole police SUV. One naked man in custody. Other suspect status unknown. #Watertown". A sixth tweet says "Shooting at MIT; suspects found in Watertown. Both were taken into custody. One was naked. Reports of explosives FBI/DHS there. Now CNN". A seventh tweet says "Recap: IED explosion in #Watertown, near Boston. 2 Suspects in custody. Unconfirmed if related to #BostonBombing or #MITShooting".

Screenshot of Keepr search for "Watertown" shows examples of misinformation stating that both suspects are in custody. **View larger image »**

Tool: Keepr

# What's next for keepr?

---

- refine algorithm
- source classification
- conversation analysis and visualization
- archiving search results and tweets

Rolling out a Beta program for newsrooms

Sign up at [www.keepr.com/beta](http://www.keepr.com/beta)

---

# ONA13

CONFERENCE AND AWARDS

OCTOBER 17-19, 2013  
ATLANTA, GA



**Online News Assn.** ✓  
@ONA



Follow

Has your large news org. done great  
breaking news reporting online this year?  
Submit to #OJA2013. [bit.ly/JruGzg](http://bit.ly/JruGzg)

Reply Retweet Favorite Pocket More

1:57 PM - 7 Jun 13

Reply to @ONA



hqu 2 months ago add entire site 1

1 contributor



file | 491 lines (406 sloc) | 18.828 kb

```
1 <!DOCTYPE html>
2 <!--[if IE 8]>                                <html class="no-js lt-ie9" lang="en"> <![endif]
3 <!--[if gt IE 8]><!--> <html class="no-js" lang="en"> <!--<![endif]-->
4 <head>
5     <link href='http://fonts.googleapis.com/css?family=Geo|Lato:400,900,700,300|Ubu
6
7     <meta charset="utf-8" />
8     <meta name="viewport" content="width=device-width" />
9     <title>Keepr - data mining social media chatter</title>
10    <link rel="stylesheet" href="css/normalize.css" />
11    <link rel="stylesheet" href="css/foundation.css" />
12    <script src="js/vendor/custom.modernizr.js"></script>
13 </head>
14 <body style="font-family: 'Lato', sans-serif;">
15 <style>
16 a:hover {
17     text-decoration:underline;
18 }
19 </style>
20 <script async src="//platform.twitter.com/widgets.js" charset="utf-8"></script>
21
22 <?php
23 $raw_query_string = htmlspecialchars($_POST['q']);
24 if (!$raw_query_string) {
25     $raw_query_string = htmlspecialchars($_GET['q']);
26 }
27 $raw_query_string = str_replace("breakingnews", "", $raw_query_string);
28 if (!$raw_query_string) {
29     $raw_query_string = "from:@breakingnews";
30     $query_string = "from:@breakingnews";     https://github.com/hqu/keepr/blob/master/index.php
```

<https://github.com/hqu/keepr>



# Verification Resources

- [Verifying Social Media Content](#)
- [verificationjunkie.tumblr.com](#)
- [BBC processes for verifying social media content](#)
- [Storyful's validation process](#)
- [InformaCam](#)