



Tweet Mining: Is It Useful and Should We Bother?

Nils C. Newman

Alan L. Porter & Jon Garner



www.TheVantagePoint.com

Background

Science and Social Media – The New Frontier

The New York Times **Science**

WORLD U.S. N.Y. / REGION BUSINESS TECHNOLOGY SCIENCE HEALTH SPORTS
ENVIRONMENT SPACE & C

New Frontier for Topics in Science: Social Media
By MARY ANN GIORDANO
Published: December 31, 2012

Understanding how Twitter is used to spread scientific messages

Julie Letierce, Alexandre Passant,
Stefan Decker
Digital Enterprise Research Institute
National University of Ireland, Galway
Galway, Ireland
firstname.lastname@deri.org

John G. Breslin
School of Engineering and Informatics
National University of Ireland, Galway
Galway, Ireland
john.breslin@nuigalway.ie

Citation Analysis in Twitter: Approaches for Defining and Measuring Information Flows within Tweets during Scientific Conferences

Katrin Weller¹, Evelyn Dröge¹ and Cornelius Puschmann²,

Heinrich-Heine-University Düsseldorf,
¹ Dept. of Information Science & ² Dept. for English Language and Linguistics,
Universitätsstr. 1, 40225 Düsseldorf, Germany
{katrin.weller, evelyn.droege, cornelius.puschmann}@uni-duesseldorf.de

Social networking in scientific conferences – Twitter as tool for strengthen a scientific community

Martin Ebner¹ and Wolfgang Reinhardt²

¹ Graz University of Technology, Faculty of Computer Science, Inffeldgasse 16b,
8010 Graz, Austria martin.ebner@tugraz.at

² University of Paderborn, Institute of Computer Science, Fuerstenallee 11,
33102 Paderborn Germany wolle@upb.de



The Premise of our Pilot Project:

Treat Twitter as a new data source for S&T analysis

- Think of Twitter in terms of any traditional data source – Patents, Scientific Publications, etc..
- Use our standard analysis techniques (VantagePoint) to look at search results on Graphene and Nano Enhanced Drug Delivery

The only difference is...

- Every abstract is only 140 characters long

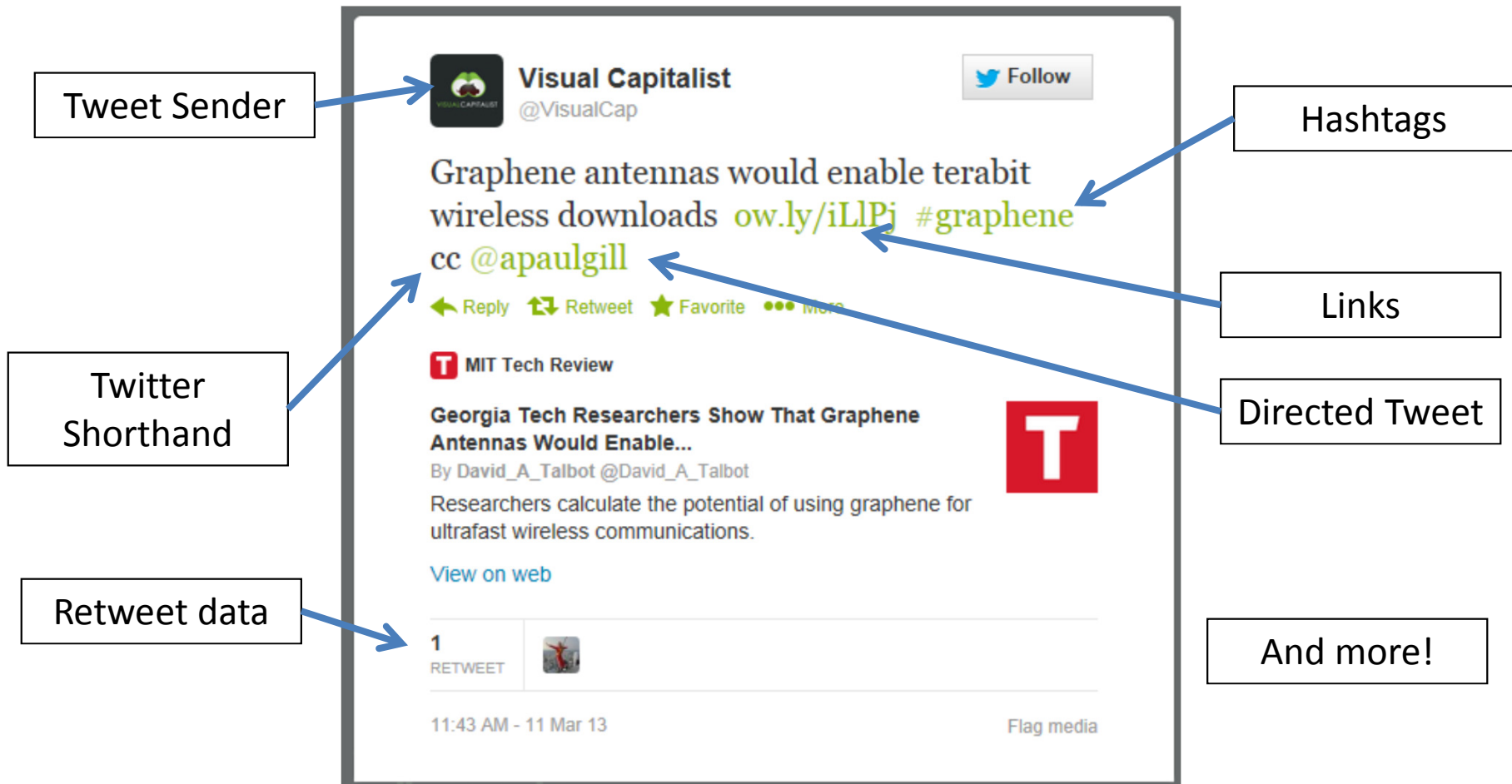


But....

There is a bit more than 140 characters
of content to work with



Anatomy of Tweet



Things we could do....

Given the combinations of names, links, re-tweet information, and other Twitter data, in theory we could:

- Find key influence leaders
- Discover emerging terminology
- Track geographic spread
- Track time trends
- Etc...



However....

Now for the messy bits



Twitter Data:

Now you see it, now you don't

With the Twitter API you can search all of Twitter but the API only provides access to the last 8 days of data.

If you want more data, you can

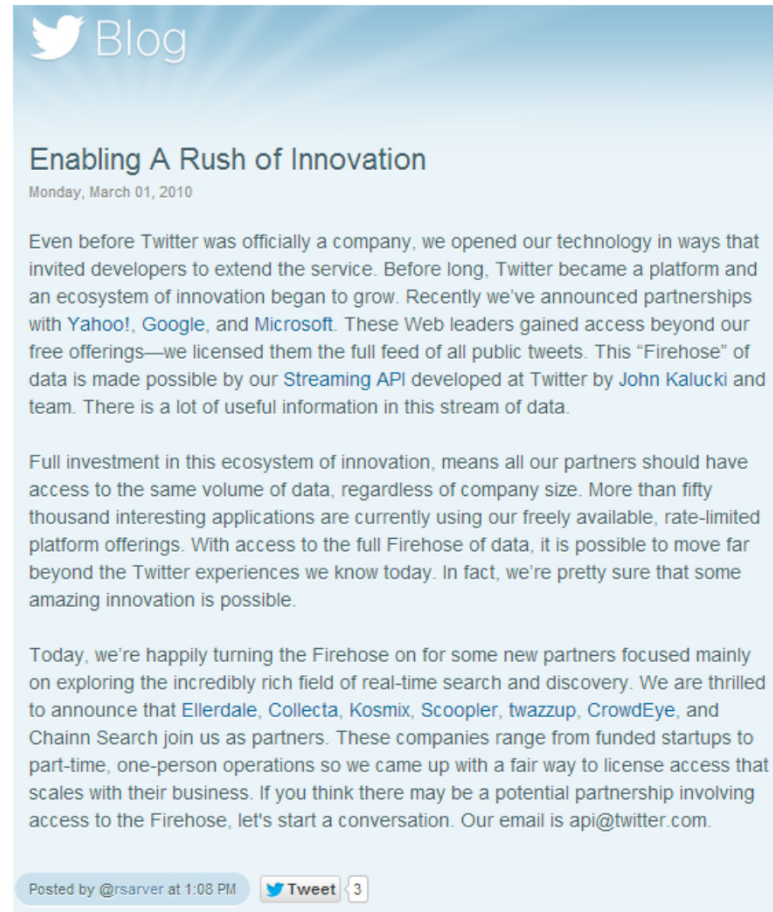
- Build your own twitter database going forward
- Purchase access to the Twitter “Firehose” to go back in time



The Quest for the Twitter Firehose

Who actually has access to the Twitter Firehose?

- Yahoo, Google, MS
- In 2010, seven companies were given access to the Firehose
- In 2013, of those seven companies, none are still around



Twitter Blog

Enabling A Rush of Innovation

Monday, March 01, 2010

Even before Twitter was officially a company, we opened our technology in ways that invited developers to extend the service. Before long, Twitter became a platform and an ecosystem of innovation began to grow. Recently we've announced partnerships with [Yahoo!](#), [Google](#), and [Microsoft](#). These Web leaders gained access beyond our free offerings—we licensed them the full feed of all public tweets. This "Firehose" of data is made possible by our [Streaming API](#) developed at Twitter by [John Kalucki](#) and team. There is a lot of useful information in this stream of data.

Full investment in this ecosystem of innovation, means all our partners should have access to the same volume of data, regardless of company size. More than fifty thousand interesting applications are currently using our freely available, rate-limited platform offerings. With access to the full Firehose of data, it is possible to move far beyond the Twitter experiences we know today. In fact, we're pretty sure that some amazing innovation is possible.

Today, we're happily turning the Firehose on for some new partners focused mainly on exploring the incredibly rich field of real-time search and discovery. We are thrilled to announce that [Ellerdale](#), [Collecta](#), [Kosmix](#), [Scoopler](#), [twazzup](#), [CrowdEye](#), and [Chainn Search](#) join us as partners. These companies range from funded startups to part-time, one-person operations so we came up with a fair way to license access that scales with their business. If you think there may be a potential partnership involving access to the Firehose, let's start a conversation. Our email is api@twitter.com.

Posted by @rsarver at 1:08 PM [Tweet](#) 3



Our Firehose Odyssey

After a bit of digging, we finally found current firehose providers who were still in business

- One wouldn't respond to inquiries
- One has embedded it into their own analysis products
- But finally, one did respond

TOPSY



Graphene Pilot



Graphene: Progress with Topsy

With Topsy, we were able to

- Get a key to access their Otter API
- Use their search interface to search for “Graphene”
- Successfully download 34,586 Tweets with coverage back to 2006
- Import the Tweets into VantagePoint for analysis



We were happy!



Then we looked at the data and found
more messy bits...



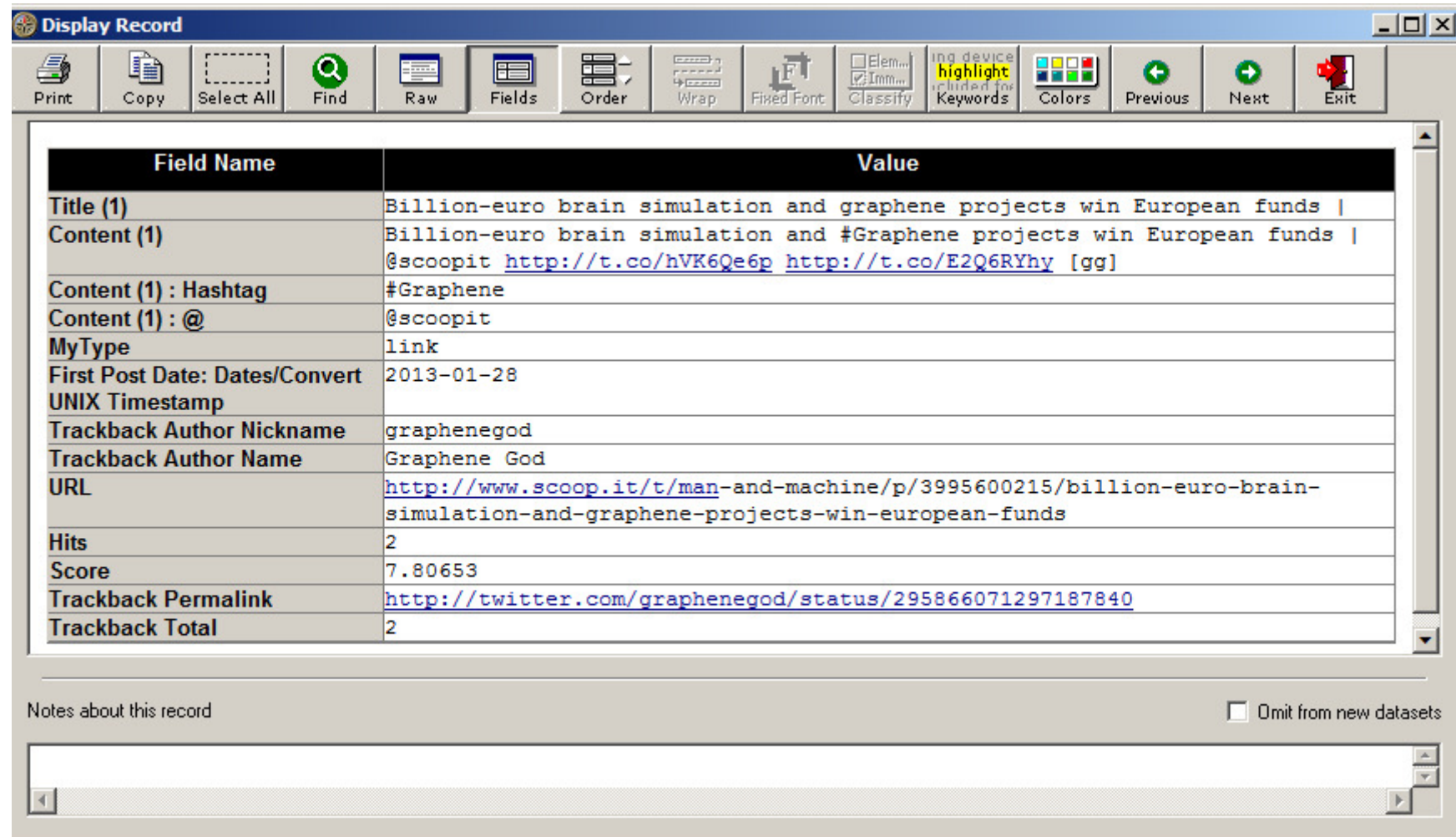
You call this documentation?

The first issue we ran into was translating Topsy Twitterese into something we could understand

- Twitter specific jargon
 - Hashtag, directed tweet, RT, etc...
- Date codes
 - In Unix Timestamp format
- Topsy specific jargon and vaguely defined indicators
 - Hits
 - Score
 - Trackback totals



But eventually we sorted most of it out



The screenshot shows a software window titled "Display Record" with a toolbar containing icons for Print, Copy, Select All, Find, Raw, Fields, Order, Wrap, Fixed Font, Element Classifier, Highlight Keywords, Colors, Previous, Next, and Exit. The main area displays a table with the following data:

Field Name	Value
Title (1)	Billion-euro brain simulation and graphene projects win European funds
Content (1)	Billion-euro brain simulation and #Graphene projects win European funds @scoopit http://t.co/hVK6Qe6p http://t.co/E2Q6RYhy [gg]
Content (1) : Hashtag	#Graphene
Content (1) : @	@scoopit
MyType	link
First Post Date: Dates/Convert UNIX Timestamp	2013-01-28
Trackback Author Nickname	graphenegod
Trackback Author Name	Graphene God
URL	http://www.scoop.it/t/man-and-machine/p/3995600215/billion-euro-brain-simulation-and-graphene-projects-win-european-funds
Hits	2
Score	7.80653
Trackback Permalink	http://twitter.com/graphenegod/status/295866071297187840
Trackback Total	2

Notes about this record Omit from new datasets



And we were able to do actual analysis

The screenshot displays the VantagePoint software interface for analyzing Twitter data. The main window is titled "Map::Trackback Author Name X Content (1) : Hashtag".

Left Panel: Titles
165 Titles, 0 Selected
#FlakeGraphite Graphene Breakout: Lomiko ...
#Graphene coating stops #corrosion 100x b...
#Graphene Laboratories Deal LMR. v http://t...
#Graphene Labs CEO Dr Polyakova will spea...
#Graphene Labs CTO Dr. Stolyarov will be p...
#graphene magic! Invisible barrier set to re...
#Graphene vs Carbon Nanotubes Possible I...
#Graphene's #solar cell potential reve...
#Graphite http://t.co/7a1Ps3civs #Lomiko M...
#News Graphene Breakout: Lomiko Metals B...
#NYSE #NASDAQ Graphene Breakout: Lomi...
#NYSE #NASDAQ Graphene plasmonics bea...
#Silicene, graphene-like silicon http://t.co/3...
\$1.35 Billion US Dollars... You a believer yet...
@GrapheneLabs now shows up in the first s...
@Indmin Free: #Graphite miners target pat...

Center Panel: Cross-Correlation Map
Trackback Author Name (20+) Content (1) : Hashtag
Tag links shown: > 0.75 19 (23), 0.50 - 0.75 2 (4), 0.25 - 0.50 0 (0), < 0.25 2 (94)
The map shows a network of nodes connected by red lines. Nodes include "World Patents clipped", "Lomiko", "Graphene 10/11/12", "Graphene 10/11/13", "Graphene 10/11/14", "Graphene 10/11/15", "Graphene 10/11/16", "Graphene 10/11/17", "Graphene 10/11/18", "Graphene 10/11/19", "Graphene 10/11/20", "Graphene 10/11/21", "Graphene 10/11/22", "Graphene 10/11/23", "Graphene 10/11/24", "Graphene 10/11/25", "Graphene 10/11/26", "Graphene 10/11/27", "Graphene 10/11/28", "Graphene 10/11/29", "Graphene 10/11/30".

Right Panel: Content (1) (2): NLP/Phrases
28 |||| #graphene
23 |||| graphene
7 |||| Graphite Mania
7 |||| Lomiko Metals Banking
7 |||| Tech Surge
5 |||| #nanotech
5 |||| bilayer #graphene
4 |||| lomiko
4 |||| publication
4 |||| RT @maxmaterials
3 |||| Acoustics @ASME_FutureME
3 |||| Check
3 |||| Engineer
3 |||| J
3 |||| Vib
3 |||| waves
2 |||| #cleantech
2 |||| #graphene #nanoplatelets
2 |||| #graphene #nanotech
2 |||| #graphene business directory W
2 |||| #graphene formation
2 |||| #Graphene Labs
2 |||| #graphene quantum dots
2 |||| #graphene sheets
2 |||| #Lomiko Deal SLMR SLMRMF
2 |||| #MoS2
2 |||| 2.7 million shares
2 |||| 45.5%
2 |||| added Lomiko Metals
2 |||| advanced 2D materials
2 |||| bad
2 |||| BN
2 |||| Business Opportunities
2 |||| challenges #tech #news
2 |||| comment
2 |||| commercial potentials
2 |||| conductance measurements
2 |||| coverage
2 |||| dr

Bottom Right Panel: MyType
Pie chart showing MyType distribution: tweet (green), image (blue), link (red).

Bottom Right Panel: First Post Date: Dates/Convert UNIX 1
Bar chart showing First Post Date distribution from 2006 to 2013. The x-axis is labeled "First Post Date: Dates/Convert UNIX 1" and the y-axis is labeled "Date/Extract Year". The chart shows a significant increase in posts starting in 2011, peaking in 2013.

Bottom Right Panel: URL
List of URLs:
2 |||| http://arxiv.org/abs/1301.4106
1 |||| http://apl.aip.org/resource/1/applab
1 |||| http://arxiv.org/abs/1208.0064
1 |||| http://arxiv.org/abs/1209.6364
1 |||| http://arxiv.org/abs/1212.2581
1 |||| http://arxiv.org/abs/1301.0673
1 |||| http://arxiv.org/abs/1301.0697
1 |||| http://arxiv.org/abs/1301.0944
1 |||| http://arxiv.org/abs/1301.1192
1 |||| http://arxiv.org/abs/1301.1257

Bottom Panel: VantagePoint Analyst Guide
Choose from:
[Analyst's handbook](#)
[Walkthroughs of common analyses](#)
[Frequently Asked Questions](#)
Go to:
[VantagePoint Help](#)
<Contents>

For Help, press F1

Windows taskbar at the bottom shows the Start button, system tray, and the time 6:18 PM on 4/10/2013.



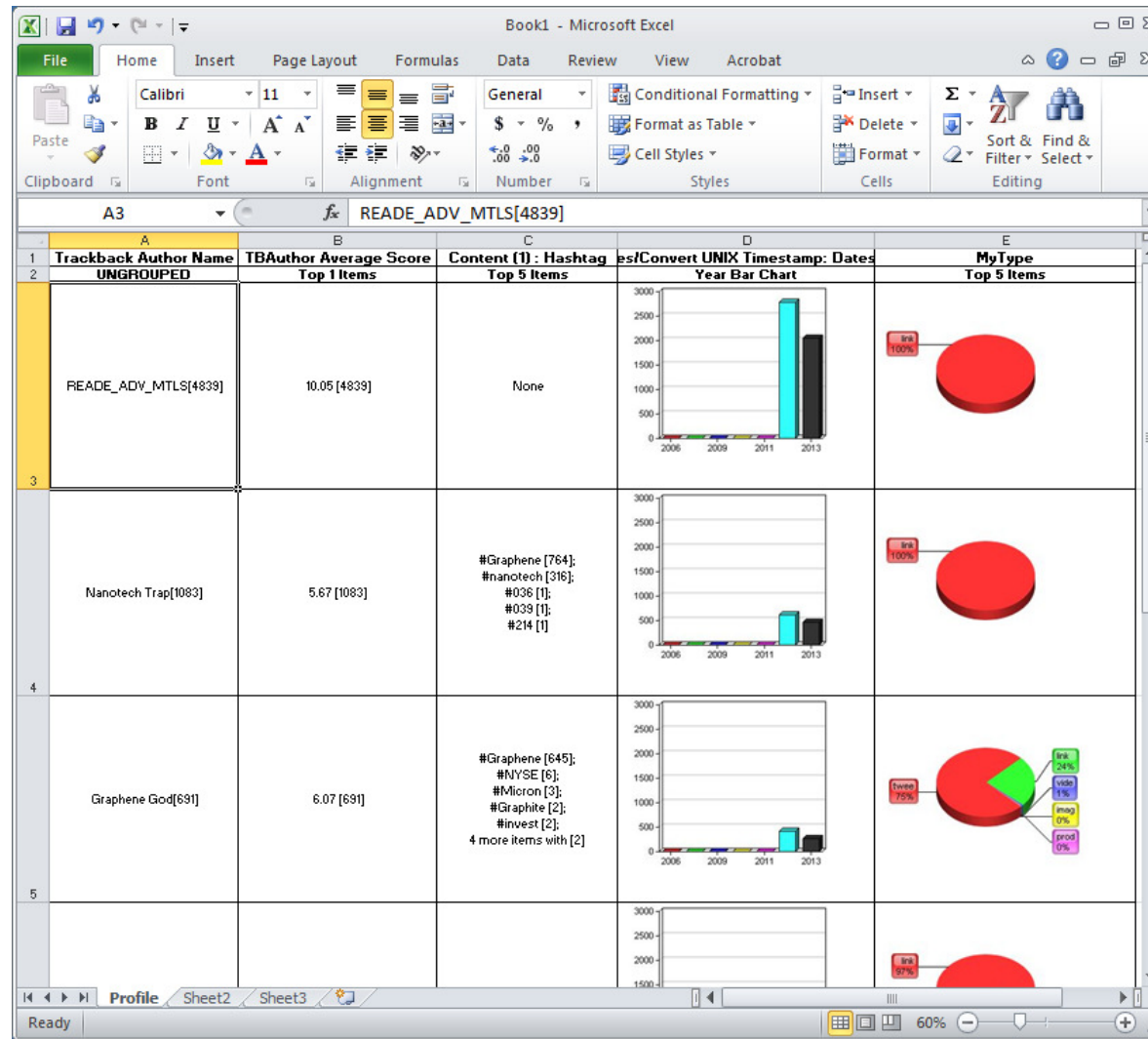
And create analytical output

The screenshot shows a Microsoft Excel spreadsheet with the following data:

Year	2006	2008	2009	2010	2011	2012	2013
Records	1	4	163	243	553	13874	19484
Terms	1	14	28	49	499	704	
	#Graphene [1 of 4029]	#tech [1 of 162]	#NYSE [20 of 366]	#1 [2 of 4]	#DailyArticle [3 of 17]	#Graphite [71 of 146]	#Sharapova [22 of 22]
			#gadget [3 of 7]	#David [2 of 37]	#material [3 of 5]	#daily [24 of 38]	#Greatvote [20 of 20]
			#networkwifi [2 of 4]	#physics [2 of 74]	#nanotech [3 of 422]	#msf12 [22 of 22]	#Brain [14 of 14]
			#Applications [1 of 1]	#140Friday [1 of 1]	#iphone [2 of 6]	#Solar [20 of 45]	#Industrial [14 of 14]
			#Bollington [1 of 2]	#1554679 [1 of 1]	#nuclear [2 of 4]	#Nanomaterials [17 of 23]	#PR [12 of 12]
			#Danny [1 of 36]	#8211 [1 of 5]	#aare2012 [1 of 1]	#CO1B [13 of 29]	#Europe [10 of 10]
			#FET [1 of 23]	#BritSciAssoc [1 of 11]	#appforthat [1 of 1]	#Electronics [13 of 31]	#Patent [10 of 10]
			#flexibleelectronics [1 of 3]	#EC [1 of 2]	#AusOpen [1 of 2]	#SoundCloud [12 of 18]	#PDAC2013 [10 of 10]
			#franken [1 of 2]	#EU [1 of 42]	#bbcnews [1 of 10]	#dental [11 of 11]	#FlakeGraphite [9 of 9]
			#goocus [1 of 2]	#graphenelove [1 of 2]	#berita [1 of 4]	#RT [11 of 14]	#investor [9 of 9]
			#ixzz0t0W19nSk [1 of 1]	#graphenetheme [1 of 1]	#CityCon [1 of 1]	#engineering [10 of 21]	#constantcontact [8 of 8]
			#nyt [1 of 2]	#greg [1 of 29]	#composites [1 of 25]	#Green [10 of 25]	#Flagship [7 of 7]
			#Silicon [1 of 10]	#hamr [1 of 3]	#dailyhempbuzz [1 of 1]	#materials [10 of 15]	#H01M [7 of 7]
			#toocool [1 of 2]	#Innovation [1 of 28]	#endless [1 of 1]	#Nano [10 of 18]	#magnetic [7 of 7]
				#IPHONE6 [1 of 1]	#euregionalfunds [1 of 2]	#Technology [10 of 50]	#UAB [7 of 7]
				#Magic [1 of 1]	#future [1 of 28]	#Grafoid [9 of 10]	#wearabletechnology [7 of 7]
				#marvelous [1 of 1]	#grafeno [1 of 22]	#Lithium [9 of 14]	#WiFi [7 of 7]
				#nanotechnology [1 of 127]	#grapheneisawesome [1 of 1]	#microbiology [9 of 9]	#Electronic [6 of 6]
				#nieuws [1 of 1]	#HEAD [1 of 23]	#physique [9 of 17]	#fbemcr [6 of 6]
				#PhysicsHistory [1 of 2]	#HoCo [1 of 1]	#reddit [9 of 22]	#ff [6 of 6]
				#Samsung [1 of 7]	#HTML [1 of 1]	#039 [8 of 14]	#fukushima [6 of 6]
				#stemcommunication [1 of 1]	#investing [1 of 4]	#CENews [8 of 8]	#globaldev [6 of 6]
				#tennis [1 of 23]	#japancc [1 of 2]	#Chemistry [8 of 19]	#intiled [6 of 6]
				#Water [1 of 15]	#job [1 of 52]	#Chimie [8 of 10]	#ittwt [6 of 6]
				#web [1 of 5]	#JobsAndGrowth [1 of 2]	#freetoread [8 of 17]	#kof1 [6 of 6]
				#WordPress [1 of 9]	#les [1 of 1]	#physicsvideos [8 of 8]	#Metals [6 of 6]
				#wtf [1 of 1]	#love [1 of 2]	#Arquillian [7 of 10]	#Pakistan [6 of 6]
				#yulbcn [1 of 1]	#MustWatch [1 of 1]	#MoS2 [7 of 10]	#sciencejobs [6 of 6]
					#nanotubes [1 of 3]	#PHOTONICS [7 of 7]	#TheMBC2013 [6 of 6]
					#nature [1 of 11]	#aerographite [6 of 6]	#think20 [6 of 6]
					#NBN [1 of 1]	#H01L [6 of 10]	#zeronautbe [6 of 6]



And produce reportable output



But is it meaningful?

- The data have a lot of noise
 - Order your Graphene t-shirt today!
 - Maria Sharapova wins with Graphene Instinct racket
- There is a lot of gray data that are challenging to interpret
 - Graphene jobs available
- There is also a reasonable amount of interesting stuff
 - Research funding announcements
 - Business information
 - Technical content



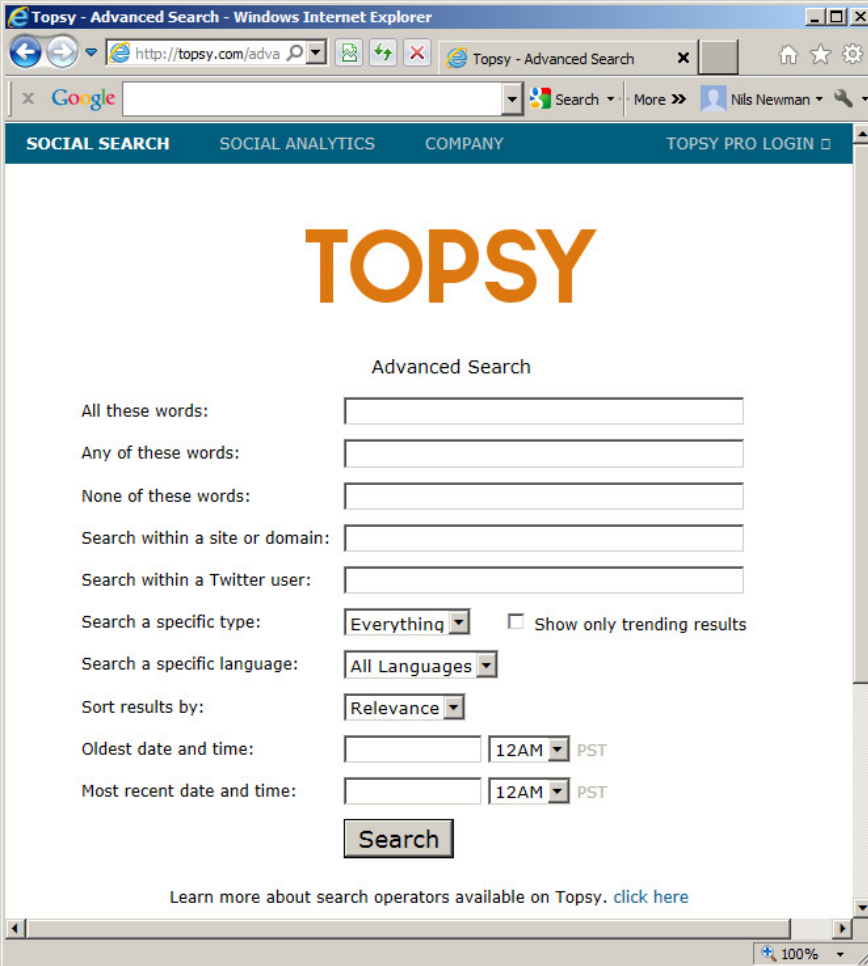
NEDD Pilot



NEDD: Progress with Topsy

NEDD presented more of a issue.

- The Topsy search interface is a bit limited
- Our Nano Enhanced Drug Delivery search strategy required complex Boolean, wildcards and nesting strategies
- Topsy only allows simple boolean



The screenshot shows the Topsy Advanced Search page in a Windows Internet Explorer browser. The browser's address bar displays the URL <http://topsy.com/adva>. The page features a navigation bar with links for SOCIAL SEARCH, SOCIAL ANALYTICS, COMPANY, and TOPSY PRO LOGIN. The main content area is titled "TOPSY" in large orange letters, followed by "Advanced Search". Below this, there are several search filters and input fields:

- All these words:
- Any of these words:
- None of these words:
- Search within a site or domain:
- Search within a Twitter user:
- Search a specific type: Show only trending results
- Search a specific language:
- Sort results by:
- Oldest date and time:
- Most recent date and time:

A "Search" button is located at the bottom of the form. At the bottom of the page, there is a link: "Learn more about search operators available on Topsy. [click here](#)". The browser's status bar at the bottom right shows a zoom level of 100%.



Our attempt

Was more than messy...



NEDD “Results”

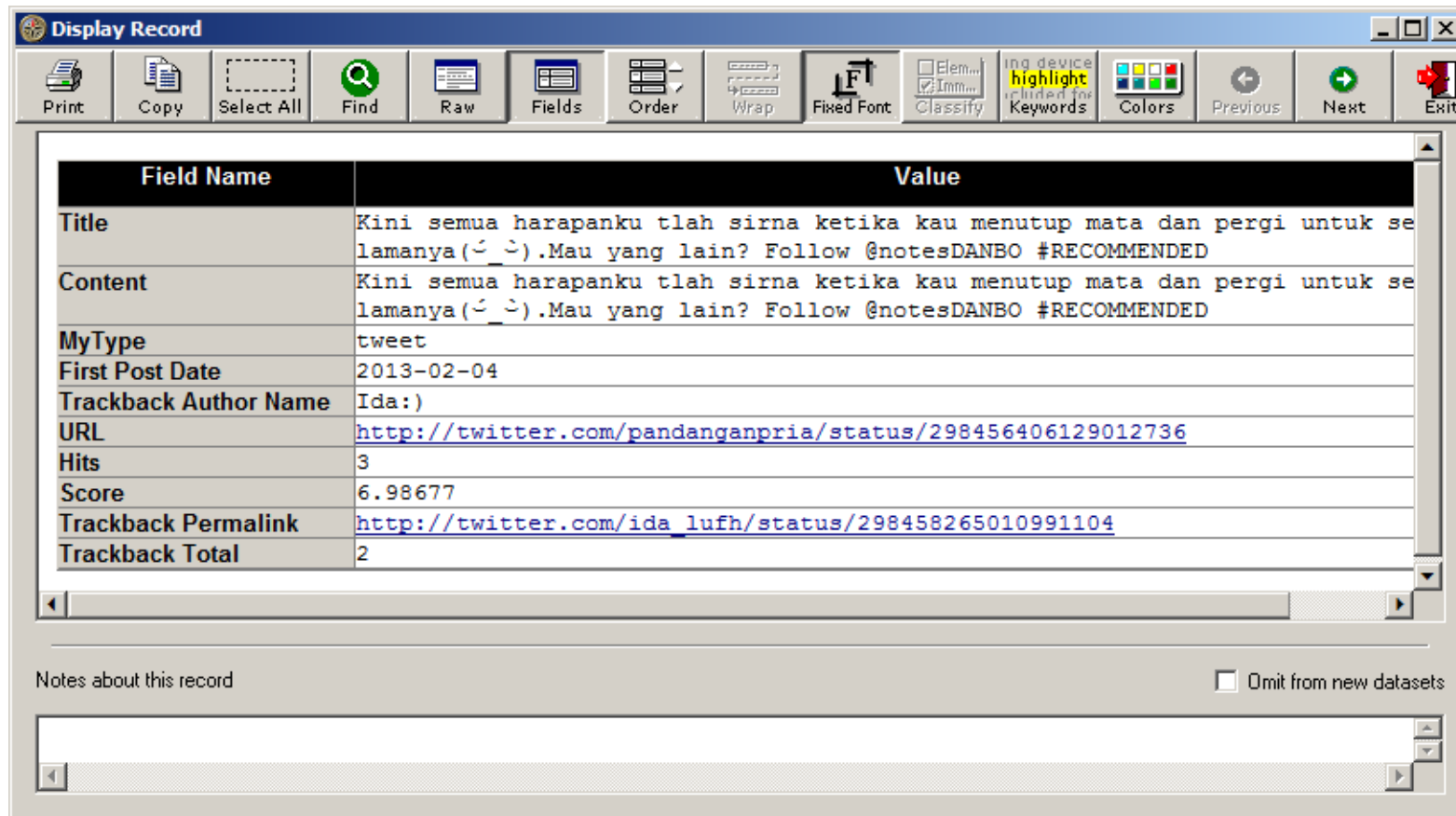
The NEDD experiment produced a number of major issues

- Search terms such as RNAi are words in that have other meanings in other languages so you have to control for language (which doesn't always seem to work)
- No wildcards or truncation - presented problems
- Limitations on the Boolean was an issue



The Result

The NEDD search was basically unusable without significant additional effort



The screenshot shows a software window titled "Display Record" with a toolbar containing icons for Print, Copy, Select All, Find, Raw, Fields, Order, Wrap, Fixed Font, Element, Imm, Classify, Highlight, Colors, Previous, Next, and Exit. The main area displays a table with the following data:

Field Name	Value
Title	Kini semua harapanku tlah sirna ketika kau menutup mata dan pergi untuk se lamanya(☺ ☺).Mau yang lain? Follow @notesDANBO #RECOMMENDED
Content	Kini semua harapanku tlah sirna ketika kau menutup mata dan pergi untuk se lamanya(☺ ☺).Mau yang lain? Follow @notesDANBO #RECOMMENDED
MyType	tweet
First Post Date	2013-02-04
Trackback Author Name	Ida:)
URL	http://twitter.com/pandanganpria/status/298456406129012736
Hits	3
Score	6.98677
Trackback Permalink	http://twitter.com/ida_lufh/status/298458265010991104
Trackback Total	2

Below the table, there is a section for "Notes about this record" with a checkbox labeled "Omit from new datasets".



Conclusions

The results of the pilot were a little more than mixed

- The Graphene pilot was a positive experience
- The NEDD pilot was pretty negative
- We can see the potential but it is going to take a bit more work
- However, the difficulty in accessing the data, the unknown cost, and the weakness in the search interface are major issues



So, Is Twitter mining useful?

There is potential.



So, Is it worth it?

Not yet.



Questions?

Thank you!

