Thinking Outside the Bubble
Yet another blog by a liker of technology

Looking At The World Through Twitter Data

Posted on **May 8, 2012**

Me and my friend Kang, also an MIT CS undergrad started playing with some data from Twitter a little while ago. I wrote this post to give a summary of some of the challenges we faced, some things we learned along the way, and some of our results so far. I hope they'll show to you, as they did to us, **how valuable social data can be.**

Scaling With Limited Resources

Thanks to AWS, companies these days are not as hard to scale as they used to be. However, college dorm-room projects still are. One of the less important things a company has to worry about is paying for its servers. That's not the case for us, though since we've been short on money and pretty greedy with data.

Here are some rough numbers about the volume of the data that we analyzed and got our results from.

User data: ~ 70 GB
Tweets: > 1 TB and growing
Analysis results: ~ 300 GB

> 10 billion rows in our databases.

Given the fact that we use almost all of this data to run experiments everyday, there was no way we could possibly afford putting it up on Amazon on a student budget. So we had to take a trip down to the closest hardware store and put together two desktops. That's what we're still using now.

We did lots of research about choosing the right database. Since we are only working on two nodes and are mainly bottlenecked by insertion speeds, we decided to go with good old MySQL. All other solutions were too slow on a couple nodes, or were too restrictive for running diverse queries. We wanted flexibility so we could experiment more easily.

Dealing with the I/O limitations on a single node is not easy. We had to use all kind of different of tricks to get around our limitations. SSD Caching, Bulk insertions, MySQL partitioning, dumping to files, extensive use of bitmaps, and the list goes on.

If you have advice or questions regarding all that fun stuff, we're all ears. :)

Now on to the more interesting stuff, our results.

Word-based Signal Detection

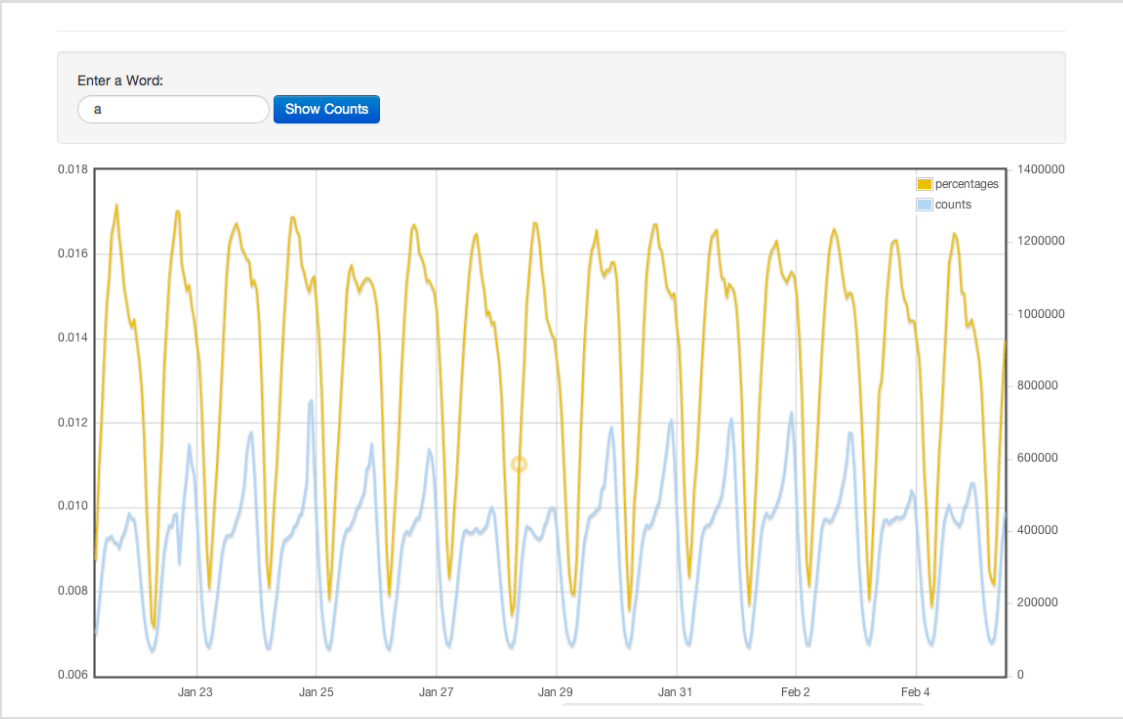
Our first step towards signal and event detection was counting single words. We counted the number of occurrences of every word in our tweets during every hour. Sort of like Google's Ngrams, but for tweets.

Make sure you play around with some of our experimental results here. The times are EST/EDT. *

Here are some cool stuff we found from looking at the counts. If you also find anything interesting from looking at different words, please share it with us!

Daily and Weekly Fluctuations

If you look for a very common word like 'a' to see an estimate of the total volume of tweets, you clearly see a daily and weekly pattern. Tweeting peaks at around 7 pm PST (10 pm EST) and hits a low at around 3 am PST every day. There's also generally less tweeting during Fridays and Saturdays, probably because people have better things to do with their lives than to tweet!

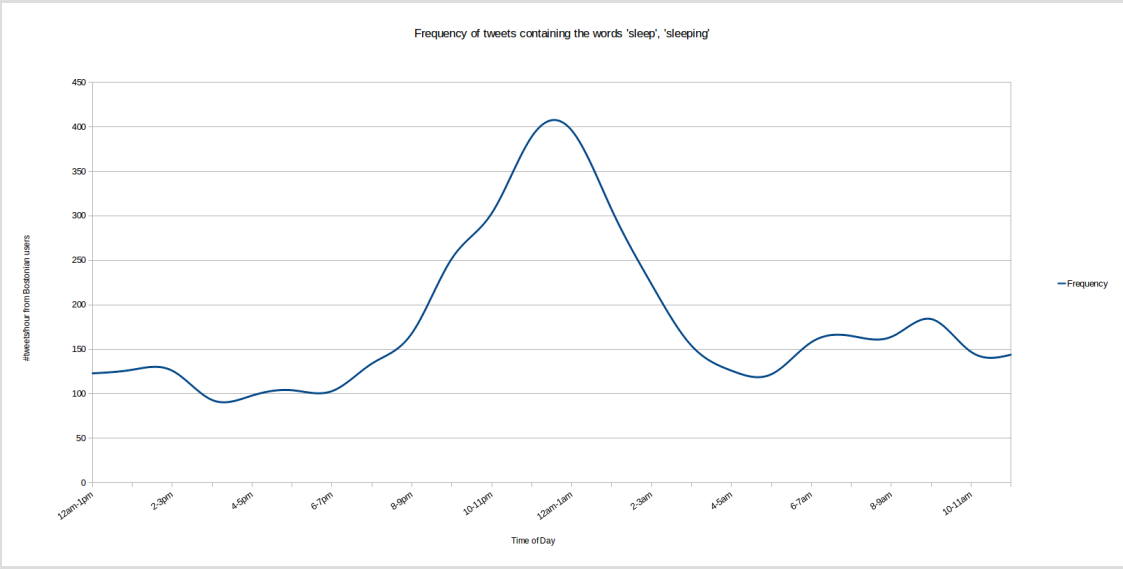


A side note:

We've tried to focus on English speaking tweeters within the States. Note that the percentage of tweets containing 'a' also fluctuates during the day, which is surprising at first. But, this is because non-English tweets that we have discarded are much more frequent during the night in our time zone, and they often don't contain the word 'a' as often as English tweets do.

Sleep

I'm a night owl myself and I had always been curious to know at exactly **what time the average person goes to sleep** or at least thinks about it! I looked for the words "sleep", "sleeping", and "bed". You can do this yourself, but the only problem you'll see is that not all the tweets have the same time zones. To solve this issue, we isolated several million tweets which had users who had set their locations to Boston or Cambridge, MA. Then, we created a histogram of their average sleeping hours. Here's the result:



It seems like the average Bostonian sleeps at around midnight! Of course, that's probably not the average everywhere. After all, a fourth of our city are nocturnal college students!

You can look at all kinds of words relating to recurring events like 'lunch', 'class', 'work', 'hungry' and whatever you can imagine. I promise you, you'll be fascinated.

Here are some suggestions:
Coke, Valentine, Hugo and other oscars-related words, IPO.
(please suggest other interesting things I should add to this list)

I'm Obsessed With Linguistics

As we were looking at different words, we noticed that the words Monday, Tuesday, etc show very interesting weekly patterns. They reflect a signal that has its peak on the respective day, as you'd expect, and which rises as you get closer to that day. This means that people have more anticipation for days that are closer, more or less linearly. But if you pay closer attention, you'll see that the day immediately before the search term corresponds to a clear valley in the curve. This points to a very interesting linguistic phenomena. That in English, we never refer to the next day with the name of that weekday, and instead use the word 'tomorrow'.

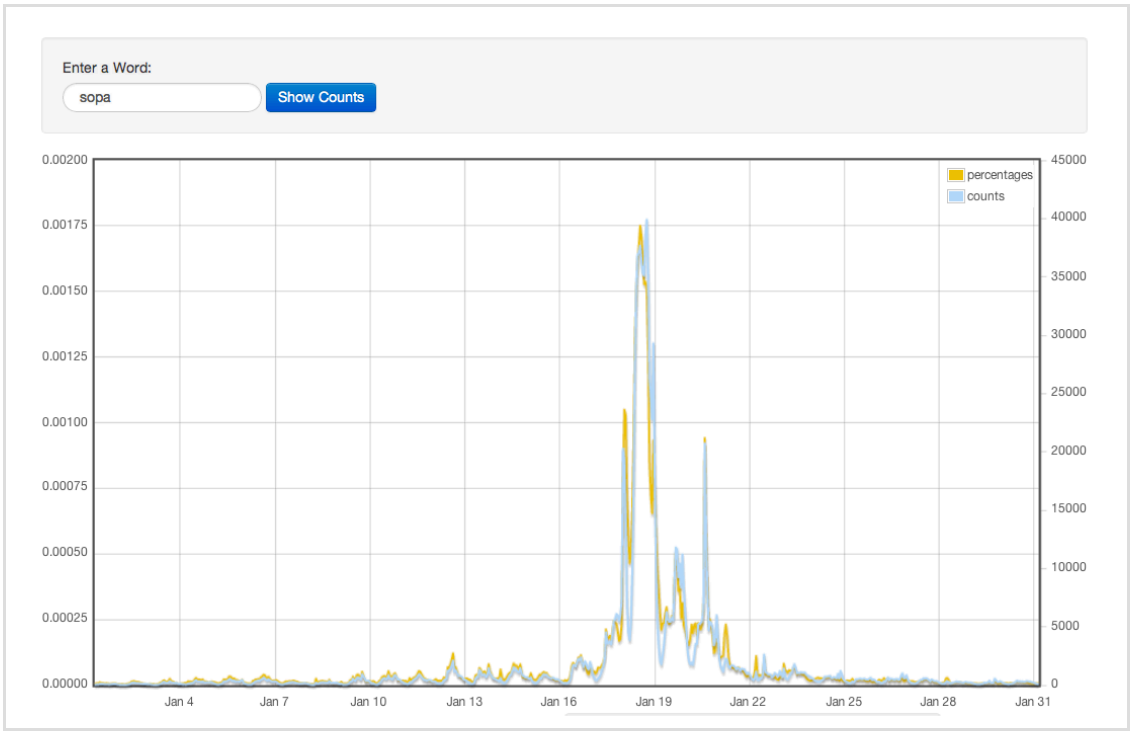
On a Wednesday, people don't say 'Thursday'

Events

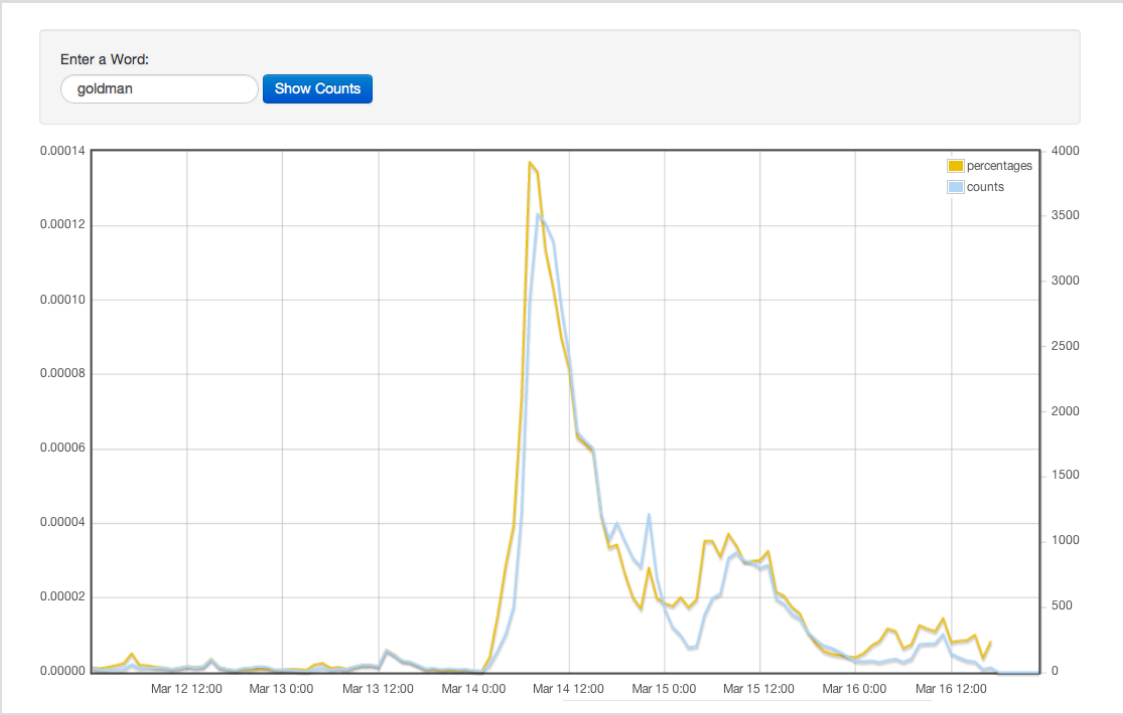
We tried to find events that we thought would have a strong reflection in the Twitter sphere. ‘superbowl’, ‘sopa’, and ‘goldman’ were pretty interesting. Here are the graphs for those three, which you can also recreate yourself.

Tweets about ‘Superbowl’ during each hour

Tweets about ‘Sopa’ during each hour



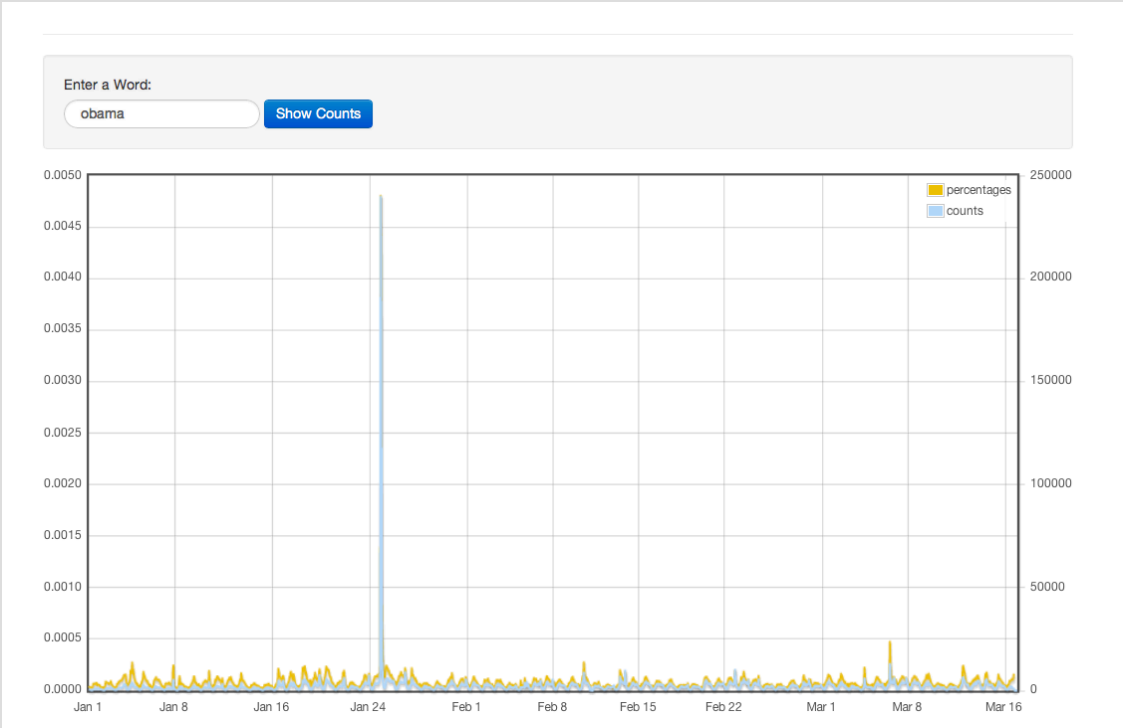
Tweets about ‘Goldman Sachs’ during each hour



We'll post more about our attempt to exactly dissect what happened on Twitter during these events as time progressed. In the Goldman Sachs case, for example, the peak happens on the day of the controversial public exit of a GS employee, which was reflected in the NYTimes. The earliest news release time was at 7am GMT which is the same as the first signs of a rise in our signal.

Politics and Public Sentiment

If you query the word Obama, this is what you'll see:



When we first saw this spike, we were very suspicious. The spike seemed way too prominent to be associated with an event. (~25 times the average signal amplitude) But guess what. The spike was at 9 PM on Jan 24th when the state of the union speech happened!

We were curious to see some of the 250 K sample tweets containing 'obama' from that hour. Here are a few of them along with some self-declared descriptions from users:

ok. the obama / giffords embrace made me choke up a little.
A teacher from Killeen, Texas. ~200 followers.

I love that Obama starts out with a tribute to our military. #SOTU
A liker of progressive politics from Utah. ~100 followers.

Great Speech Obama #SOTU
A CEO from NYC. ~700 followers.

There were both positive and negative tweets. But we wanted to know whether the tweets were positive or negative because that's what really matters in a context like this. Here are some results and an explanation of how our sentiment analysis works in general.

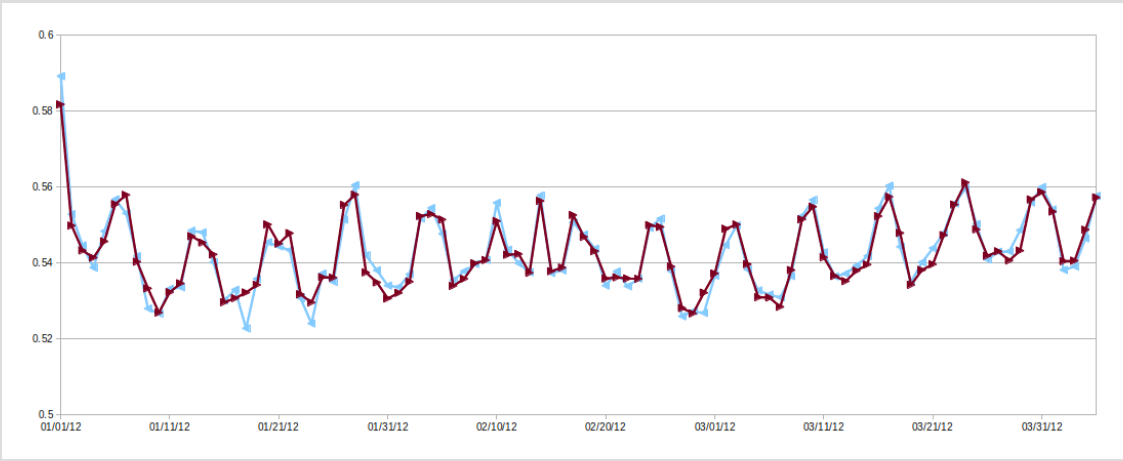
Sentiment Analysis

Our sentiment analysis is done by training our model using several thousand manually classified sample tweets. It reaches very good prediction accuracy according to our tests and as I'll explain below.

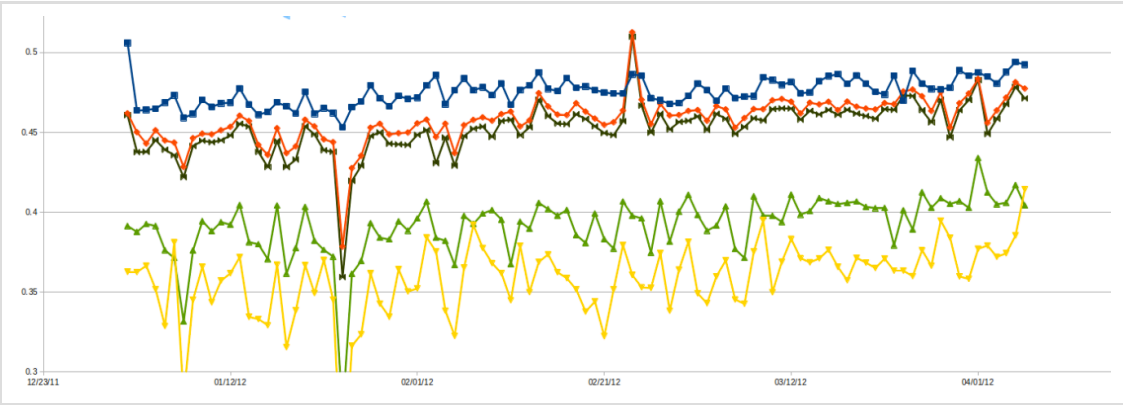
The graph below shows the normalized sentiments of several different sets of tweets for each day during a 3 month period.

The two graphs here are sentiments of millions of independent randomly chosen tweets from our set. The fact that they follow each other so closely is the important achievement of our system. It means that the signal to noise ratio is so high that the sentiment is clearly measurable.

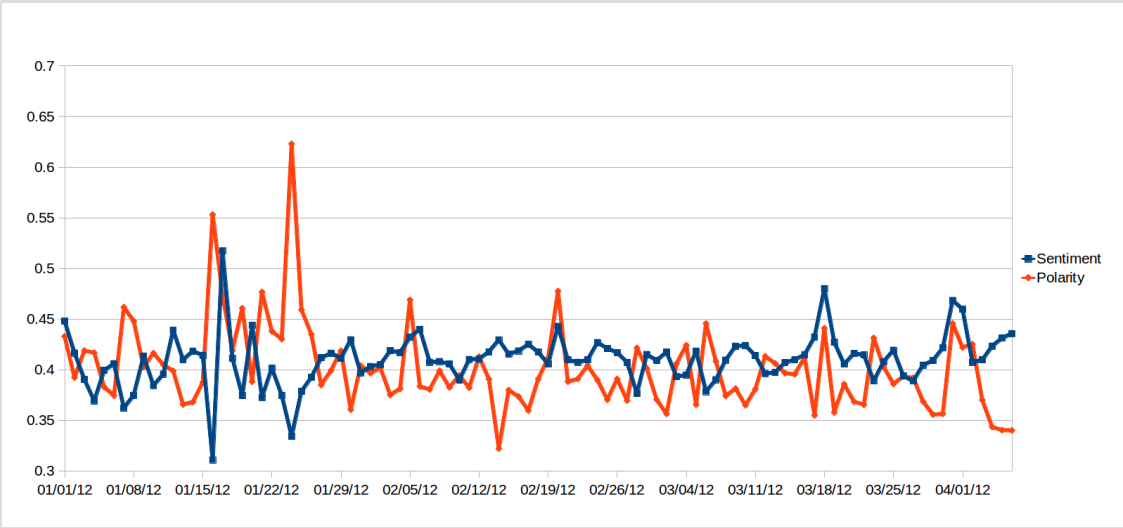
You can also observe a weekly periodicity in the general sentiment. Interestingly, it shows that people have happier and more positive tweets during the weekends compared to the rest of the week! In addition, the signal acts sort of unusually around January 1st and February 14th!



The graphs below, on the other hand, are indicative of the sentiments of tweets in which some combination of keywords related to ‘economy’ or ‘energy’ were talked about. As you can see, the patterns in the graphs are fairly stable other than at a single day, January 24th, where the sentiment significantly drops. That’s when Obama’s state of the union speech was, and it looks like his speech triggered a lot of negative tweets related to energy and the economy.



We were curious to see whether the sentiment was only negative for this bag of tweets (those that contain energy, economy), or if tweets about obama during the state of the union speech were negative in general. Here are our results of running sentiment analysis on all the data containing the word ‘obama’ in the past 4 months.



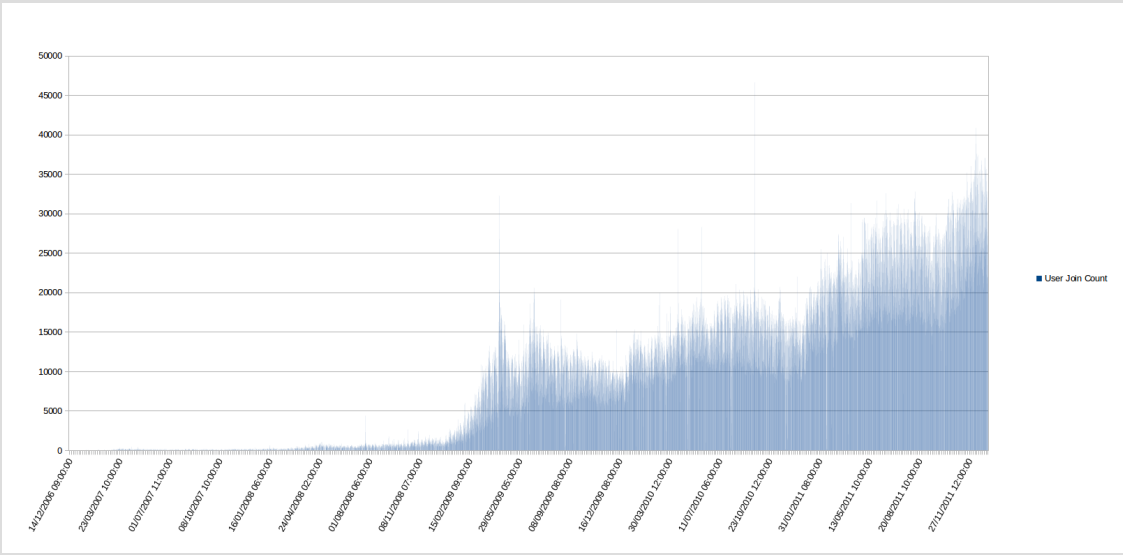
Here, the blue curve is the average sentiment of tweets for each day and the red curve shows the amount of variance in the sentiment. (If it's high, then there were more happy and sad tweets and when it's low, tweets were more neutral)

The graph clearly shows that there was some heated debate about Obama on exactly January 24th. On the same day, the sentiment has dropped and so we can tell that there was more negative tweeting than there was positive. It looks like people weren't too happy during the #sotu speech.

As we looked at the graph, it was also hard to ignore the other peak in variance (polarity) that seems to appear around the 15th to 17th of January. It seems like sentiment about Obama fell and raised rapidly in only a few days in that time span. This is likely a result of tweets about SOPA/PIPA and Obama's disagreement with the bill which happened during those days.

The Growth of Twitter

Twitter has had periods of slow and rapid growth since its inception on March 21st of 2006 up until now. We tried to capture the growth of Twitter, since its very first user until earlier this year. Here is the result showing the number of people joining Twitter during every hour since day one:



As you can see, there are some very abnormally large numbers of people joining during specific hours throughout Twitter's lifetime. One is apparently in April 2009, when they released their search feature. Another is probably when they rolled out the then New Twitter for everyone.

Summing up…

As we started looking at the data for the first time, **we were absolutely blown away** by all the cool insights you can extract from it. It makes you wonder why people aren't doing these sorts of things more often with all this cheap data and computing power that they have these days.

This was our first blog post and we hope you liked what you saw. You're awesome if you're still reading. Stay tuned for more and please give us any feedback you may have!

Also, you can reach us at twitwi at mit dot edu.

**The count results aren't perfectly accurate. There is a general upward trend because twitter deletes history before 3200 tweets.There is also a discontinuity on around Feb 11th which is because of a temporary glitch we had.*

This entry was posted in **Uncategorized** by **arashd**. Bookmark the **permalink [http://arashd.scripts.mit.edu/blog/?p=34]** .

26 THOUGHTS ON "LOOKING AT THE WORLD THROUGH TWITTER DATA"

Doug Feigelson

on **May 8, 2012 at 8:37 pm** said:

Sick post. Looking forward to hearing more about this.

arashd
on **May 8, 2012 at 10:22 pm** said:

Thanks Doug. Glad you liked it!

Atin
on **May 8, 2012 at 8:38 pm** said:

I guess people are already doing it. I am sure enterprises are writing applications for this already. What will be more interesting it to do this on realtime, as in rather than capturing data and doing analysis on stale date, if we can do this on a streaming data…

something like

tail data | grep “my search query”

I am sure that with minimum changes you will be able to achieve with your system.

Great read

kang
on **May 8, 2012 at 10:17 pm** said:

Realtime for gigabytes of data is trivial if we have as much money as enterprises, but it is rather challenging to achieve on one or two commodity PCs. We are still in the process of developing very efficient algorithms for this purpose.

Bruno Gonçalves
on **May 8, 2012 at 10:49 pm** said:

You might want to take a look at http://truthy.indiana.edu/.

We analyze the twitter stream in to detect and track memes in real time.

Tommy Garver
on **May 8, 2012 at 8:45 pm** said:

I thought it was cool to do search for the word “studying” and look at it over the course of a week. It peaks at around 8pm every night, except friday nights where it’s almost completely gone. On the other hand, a search of “partying” peaks at about 11pm every night, but the overall magnitude of the peaks increases through Saturday night and drops off on Sunday night, when people get back to their studying 

arashd
on **May 9, 2012 at 1:57 am** said:

hahah, thanks Tommy. great observation :) Yeah there are so many of these interesting things you can see!

Tony
on **May 8, 2012 at 9:37 pm** said:

Great work. I’ll be following your updates. I’m building a platform for developers to crunch such APIs / data sets.

chris
on **May 8, 2012 at 11:10 pm** said:

very interesting stuff, I saw your post on Hacker News and was really interested since I've been working on a similar side project. I've been working on an algorithm efficient enough for real-time sentiment analysis and have made some significant progress with it. Shoot me an email if you get a chance and we can chat.

-chris

Susan Beebe
on **May 8, 2012 at 11:46 pm** said:

Wow this is really fantastic! What a fun project. The data findings are quite provocative. I need to play around with this some. thanks for doing this.

@SusanBeebe

arashd
on **May 9, 2012 at 1:59 am** said:

Glad you like it! Do let us know here if you find cool stuff as you play around with it! :)

John Childerspoon
on **May 9, 2012 at 12:44 am** said:

Can you share the technology details? Other than mysql, what other tools are you using (for sentiment analysis etc…)

rouli
on **May 9, 2012 at 12:44 am** said:

very cool!
one feature request, if I may, it could be nice if when one clicks on the graph, the tool showed some selected tweets from that day or hour that match the search query, so you could figure the reason for the peaks.

Pingback: MIT: Looking At The World Through Twitter Data | Data story

sadeq
on **May 9, 2012 at 5:21 am** said:

Thats really Cool. Is it possible to use the tweets as a kind of evidence for disambiguation of keywords ? Have you try that?!

Matt Alcock
on **May 9, 2012 at 8:03 am** said:

Arash, Kang great post. You'd be surprised actually at how many companies do use twitter data to drive insight. Working for one I can tell you that its not uncommon at all. Its an exciting space!

I was wondering what the parameters you used to constrain the twitter population. You mentioned sampling 1TB of tweets what universe does this represent? This might cause a sampling bias on your results or provide a more focused insight on the results.

Also I'd love to know how you used MySQL to generate the analysis results and what data-structures the results where stored in and how you converted them into the MySQL relational store?

"Dealing with the I/O limitations on a single node is not easy. We had to use all kind of different of tricks to get around our limitations. SSD Caching, Bulk insertions, MySQL partitioning, dumping to files, extensive use of bitmaps, and the list goes on"

Please tell use more about the extensive use of bitmaps e.t.c it sounds very interesting.

Regards
Matt

Patrick
on **May 9, 2012 at 8:21 am** said:

Awesome, well done guys  And thanks again for your insights with MySQL 
Patrick

Patrick
on **May 9, 2012 at 11:26 am** said:

btw: I tried entering AAPL: Theres a peak at the last earnings conference of Apple and I am pretty sure that there is one again on april 24th 

Nico
on **May 9, 2012 at 9:41 am** said:

Cool! Keep up the great work!

I have been analysing Twitter data for 3 years. I've gathered couple a billion tweets since then, willing to share if you're interested.

I'm interested to know more details about your sentiment analysis for our new startup PeerReach. Is this something you are willing to share?

Nico

Carl Youngblood
on **May 9, 2012 at 9:50 am** said:

You should have a look at tokutek's new indexing engine for mysql. It's supposed to be ideal for databases with a lot of inserts and they claim two orders of magnitude performance improvements.

Rocco
on **May 9, 2012 at 2:12 pm** said:

Really good read. Are you guys going to be releasing the code you've been working with as well as describing how you've actually collected the data, by chance?

puyol5
on **May 9, 2012 at 2:14 pm** said:

I also used some cheap Twiter data for soziometrie.ch (sociometry in english). It really is fun to gain all those insights that Twitter won't give you. The biggest restriction i have experienced are the Twitter API limits (150 hits per hour). How did you handle this?

srikanth
on **May 9, 2012 at 3:00 pm** said:

very good post. Would you be able to share any code related information? What open source components you used, etc.

Nabs
on **May 9, 2012 at 11:46 pm** said:

Very interesting stuff! Can you share how you are doing the sentiment analysis?

Thanks and keep up the good work!

Nabs

Sue Hepworth
on **May 10, 2012 at 1:23 am** said:

This is very interesting and I enjoyed reading it and thinking about your analyses. II have just one worry – namely, that you will forget that not everyone tweets. There are large subsections of the population who don't tweet, and there are a lot of people who don't even understand what a tweet is. Believe me! I know some of these people! Therefore, you must guard against making statements like "It seems like the average Bostonian sleeps at around midnight" because that statement is not justified by your data. What you should say be saying is "It seems like the average Bostonian who uses Twitter sleeps at around midnight."

arashd
on **May 10, 2012 at 2:06 am** said:

Thanks! That's actually a great observation. Even though Twitter's user base is becoming incredibly large, surpassing 500M recently, you're absolutely right about the user demographics being extremely non-uniform. In fact, some of the researchers and companies doing this sort of analysis sometimes ignore this, but it is a reality. I think it also strongly biases results about political opinions. For example, republicans and democrats probably don't have an equal representation on Twitter.

> I should edit my post to mention this wherever I'm making these sorts of claims.