

REU 2012

Visualizing Twitter Trends

by Ahmed Al-Asadi

Mentors: James Abello and Yifan Hu

Data

- tweets collected from the Twitter API
- Only 140 characters long
- Contains slang, abbreviations, misspelled words, and poor grammar

Challenges

- Creating a tokenizer to extract useful information from tweets
- Finding a method to visualize twitter data
- Making the clusters viewable by removing any “useless” tweets (i.e. that don’t help in predicting trends)

Tokenizing

Sample tweet:

@kwong @mpop @heftyleft @neel Federer won his 7th Wimbledon title. Federer is the best tennis player in history: on.wsj.com/pcrqiz #bbctennis

Tokens: federer won wimbledon title tennis player history

Tokenizing

- Removes:
 1. usernames, URLs, hashtags, duplicate words
 2. very common words and phrases
 3. numbers, and words with non-latin characters
- Incorporates a language detector to keep only English tweets (created by Nakatani Shuoyo)
- Shortens elongations: coooooool -> cool

Graph

- Each node is a tweet
- Every tweet contains a vector of term frequency scores for every word in the total list of tweets

Graph

Tokenized tweet (1) obama healthcare

Tokenized tweet (2) college graduation

Tokenized tweet (3) obama speaker graduation

Create vectors with dimension n , where n is the number of unique words in the corpus after tokenization

<obama, healthcare, college, graduation, speaker>

The value in each position is the frequency of the word in the corpus, but 0 if that word is not contained in the tweet

Vector for (1): <2, 1, 0, 0, 0>

Vector for (2): <0, 0, 1, 2, 0>

Vector for (3): <2, 0, 0, 2, 1>

Graph

- The edge weight between (tweet1, tweet2) is the cosine similarity of the tweet's term frequency vectors

$$\text{cosine similarity} = \frac{\sum_{i=1}^n A_i \times B_i}{\sqrt{\sum_{i=1}^n (A_i)^2} \times \sqrt{\sum_{i=1}^n (B_i)^2}}$$

Note that $0 \leq \text{cosine similarity} \leq 1$, where 1 means the tweets are the same and 0 means the tweets are very different (i.e. no words in common)

Graph

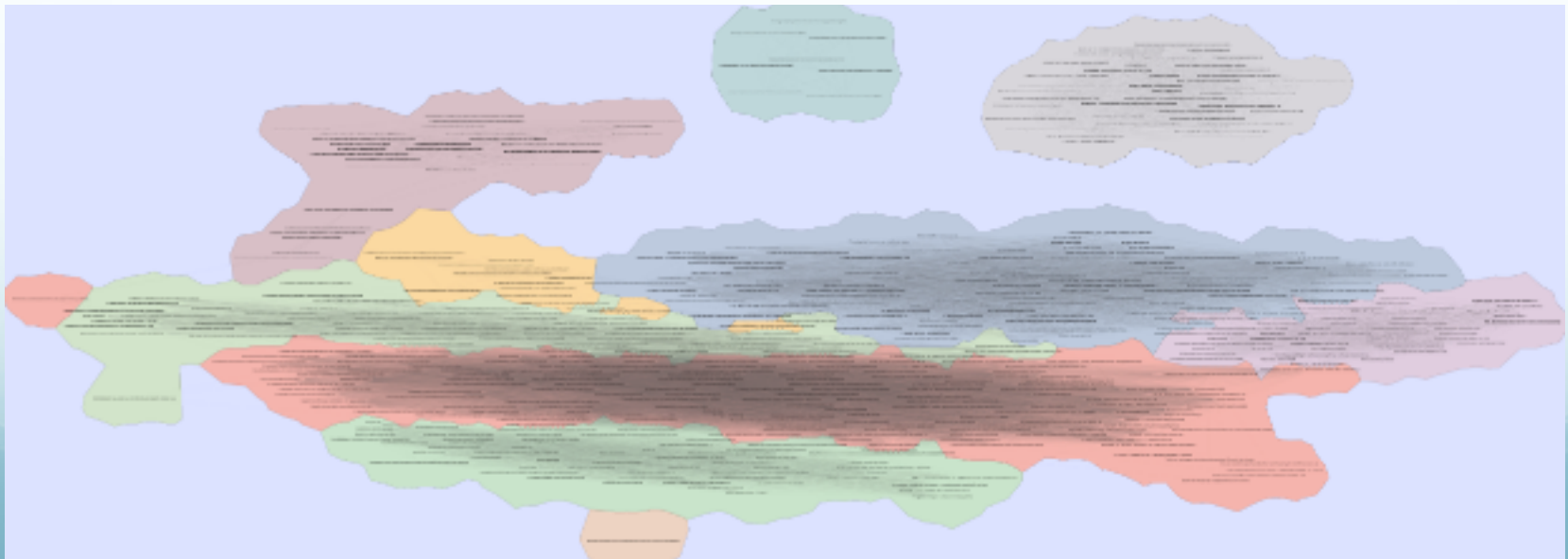
- We add an additional criterion: if a pair of tweets don't have at least two words in common, we give the edge weight between the tweets a value of 0
- Afterwards, we prune all edges with edge weight 0 and additionally we remove all isolated nodes
- In the end, we keep all the nodes that have degrees that are more than two standard deviations away from the mean, and prune the rest

Visualization

- We input the graph into graphviz, a graph visualization program, and cluster the graph
- Clustering groups sets of related nodes
- We then input the resulting graph into gvmap, which emphasizes clustering relationships on a map

Results

- This is a map of ~40,000 tweets from July 8th, 2012. Each color corresponds to a different cluster.
- In the following slides, there is a more close-up picture of each cluster.



Results

RT @annabellegrant: I don't understand how people can think Nicki Minaj is attractive.

nicki minaj looked a

i even buy tickets to see her?

RT @tr4ngpak: wait is nicki minaj shitting into the microphone or is that what she sounds like live?!

Nicki minaj is amazing can't wait to go and see her :-) #excited

i havnt been so jelouse this weekend as i am right now watchin nicki minaj !!!

RT @HazzaP_10: Sorry but Nicki Minaj is so rotten #howler

op laughing hahaha

g nicki minaj at T in the park On the telly has got me so excited to see her live!!!

Gin: RT @He_CuteeAsFuck: Yoo WTF.??? :O RT @You_Fuckface: Holly shittt Nicki Minaj :O!!!! <http://t.co/...>

one who has nicki minaj in their #10FavouriteVocalists get some real music in your system for fuck' ...

Results



Nicki Minaj at T In The Park 2012

Artist: [Nicki Minaj](#), Venue: [Balado, Kinross, Scotland](#)

Roman's Revenge Did it on em I Am Your Leader Beez
Moment 4 Life Starships Pound the Alarm Turn Me On



[View](#)



[Edit](#)

Results

star Ernest Borgnine dies in LA at age 95

Greatly saddened by the news of the passing of

R.I.P to the legend that is Ernest Borgnine

Wow thats old-very cool. RT @ABCWorldNews: BREAKING: Oscar-winning film

RT @msnbc_breaking: Actor Ernest Borgnine dies at 95, AP reports <http://t.co/6ILSWsMo>

RT @GMA: RT @ABCWorldNews #breakingnews: Ernest Borgnine has died

Ernest Borgnine - A legendary actor. #RIP

RIP Ernest Borgnine *Mouse Ears* #Mart

RT @CBSLA: #BREAKINGNEWS A spokesman for Ernest Borgnine says the

RT @swirsk054: â@BreakingNews: Spokesman: Oscar-winning film star Ernest Borgnine has died in Los Angeles at age 95 - @APâMcHales Navy. ...

st Borgnine, the legendary actor perh... <http://t.co/3MVYLulg> #celebrity #gossip

I miss the Great Circus Parade, btw. #Milwaukee Loved watching Ernest Borgn

RT @washingtonpost: #BREAKINGNEWS: Oscar-winning actor Ernest Borgnine dies at 95 <http://t.co/GkvNWvBQ>

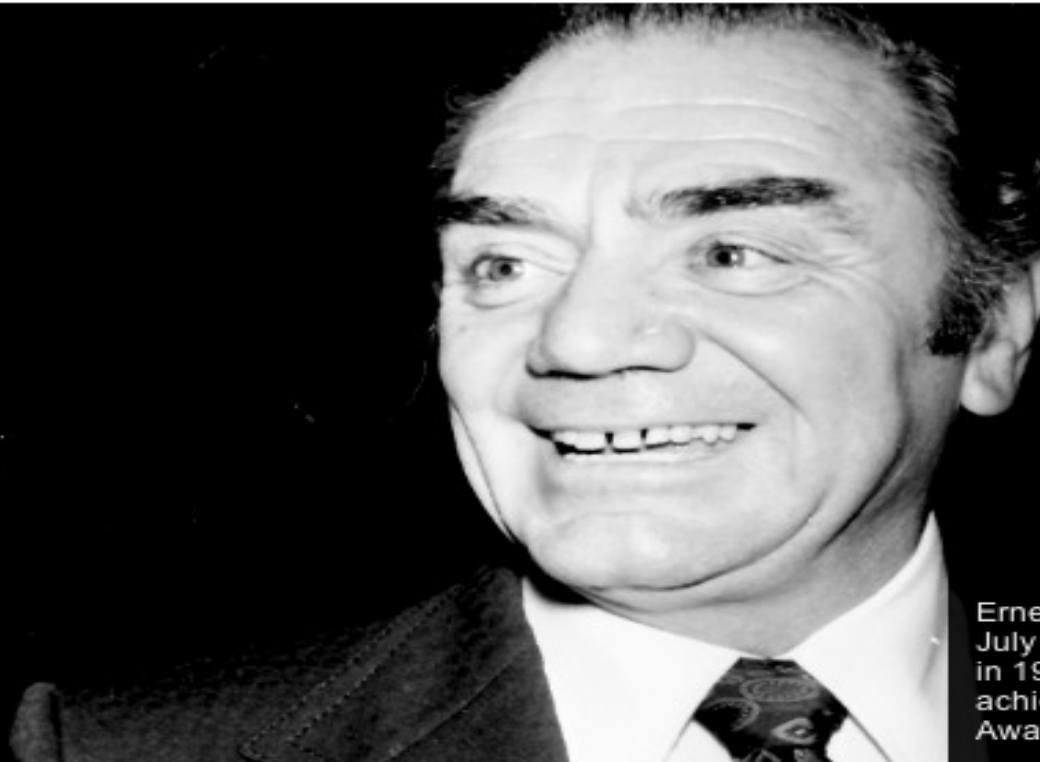
Osc

Results

Actor Ernest Borgnine dead at 95

By the CNN Wire Staff

updated 7:20 PM EDT, Sun July 8, 2012



COVER/GETTY IMAGES

Ernest Borgnine died on Sunday, July 8, at age 95. He's pictured here in 1973. In his greatest acting achievement, he won an Academy Award for his role in 1955's "Marty."

HIDE CAPTION

Borgnine through the years

Results

One direction to 5 directions

tbh one direction SHOULD split up so me and zayn can focus more

One Direction. Favorite for The Wanted.

if you believe one direction have split youre silly, very silly.

@British_Airways for the competition with

LOL MAYBE THE USA ONE DIRECTION SPLITTED UP LOLILOLOL LAFFIN 4 EFA

RT @teenagedirrtbag: looks like they're all going their sperate ways. so much for One Direction

@1DCentral_ @1Dnews @ @The1DScene ONE DIRECTION ARE SCHEDULED TO BE ON THE ELLEN SHOW ON JULY 16

litting up...there was obviously more than one direction then #haha

Results

**PEREZ HILTON: ONE
DIRECTION WILL
BREAK UP**



07 Jul 12 By: laurencf Liked it?:

SQUEEZE IT



Like

124



Tweet



+1

0



Sub

Results

ation should've made you realize that love is love. You related to his songs because you relate to ...

RT @kissmyspanish_ : Frank Ocean and The Weeknd should

Frank Ocean On Full Blast ! #CantStopWontStop !

RT @Hot_Dxmn: #np Frank

RT @IAmHollo: Frank Ocean And His BoyFriendâ: Awww â» >>> <http://t.co/OjT5yPTx>

RT @lois_wilkesss: Frank Ocean con

egend: #10FavoriteVocalists Frank Fucking Ocean.

Ash RT fake Frank Ocean but that was that real though I must say.

I canât even sing frank ocean âthinking about youâ no more. ð

Real Niggas listen to Frank Ocean <http://t.co/w8L6yG9h>

ee: #NP Frank Ocean - Thinking About You <3

#4FavoriteVocalists adele, beyonce, frank ocean, melanie fiona, i can't name 10

mes out with a song called sweet life lol double entendre and irony at the same dame time

All @DrunkBarney tweets be retweets of that fake ass Frank Ocean

onjour_ImDejjj: People Be Making Up Some Wack Ass Jokes About Frank Ocean Being Gay . .

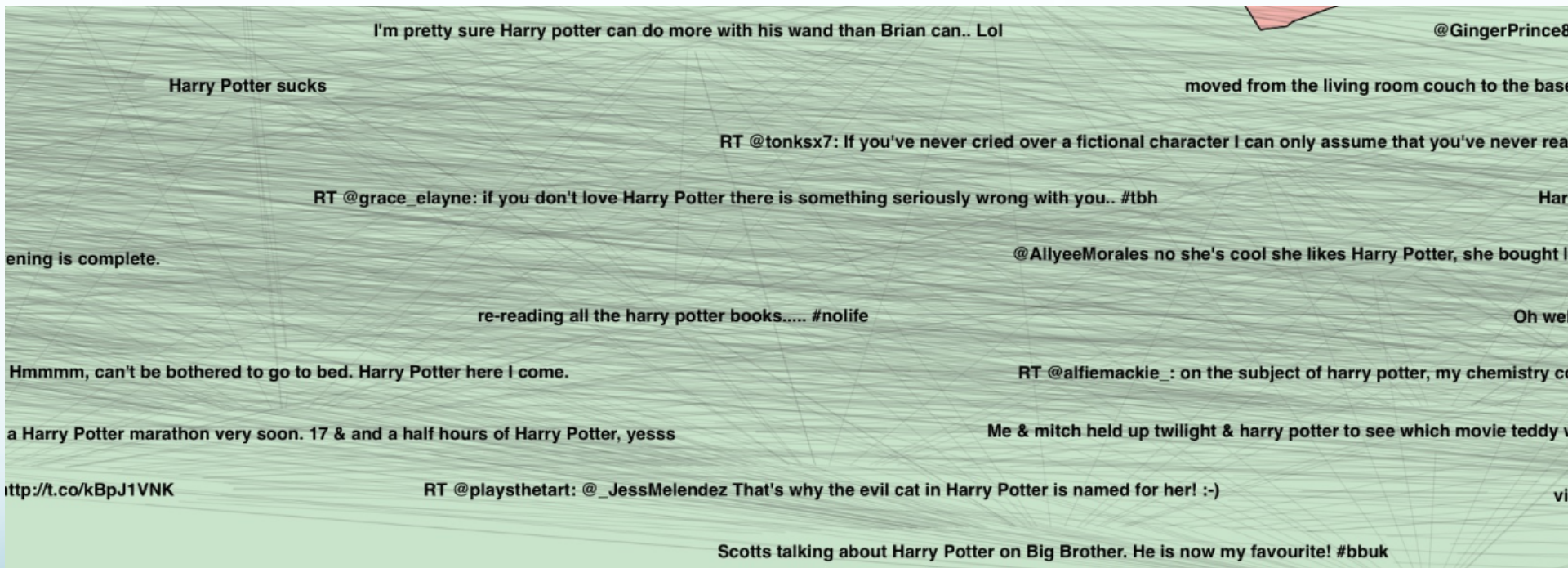
i only like 3 songs by frank ocean: novacane, thinking about you & we all try.

Results

R&B star Frank Ocean makes media waves with admission he's gay

By THOMAS CONNER tconner@suntimes.com July 11, 2012 5:28PM

Results



Results

2:00 pm	Harry Potter and the Goblet of Fire TV-PG D, L, S, V Signs of Voldemort's return emerge as Harry's friends help him prepare for the Triwizard Tournament. Based on a book by J.K. Rowling.
3:45 pm	On Air Only Add to calendar
5:30 pm	Harry Potter and the Order of the Phoenix TV-PG V When a new professor's bureaucratic methods leave Hogwarts unprepared, a dedicated group of students how to defend against the dark arts. Based on a book by J.K. Rowling.
7:00 pm	On Air Only Add to calendar
8:30 pm	Harry Potter and the Half-Blood Prince TV-PG L, V Dumbledore tries to prepare Harry for the final battle with Voldemort, but the wizard's death is inevitable. Based on the novel by J.K. Rowling.
10:15 pm	

Results

Federer beats Andy Murray to win men's final <http://t.co/88iSwfDp>

ly Murray, maybe if you had kissed Luke you might have won

tional pride soaring thanks to Andy Murray and Wimbledon <http://t.co/3nTBcVGV> via @p

New post: Roger Federer Defeats Andy Murray For His 7th Wimbledon Championship

Results

Roger Federer wins 7th Wimbledon title



ASSOCIATED PRESS

Roger Federer of Switzerland celebrates with the trophy after winning the men's singles final against Andy Murray of Britain at the All England Lawn Tennis Championships at Wimbledon, England, Sunday, July 8, 2012.

Chris Lehourites, Associated Press
Sunday, July 8, 2012 | 11:16 a.m.

Results

An amazing new weight loss product! Lose 10lbs a week,I didnt even change my diet! <http://t.co/gwa6FJta>

Amazing new weight loss product! Lose 10lbs a week,I didnt even change my diet! <http://t.co/PKLzI9Jj>

An amazing new weight loss product! Lose 10lbs a week,I didnt even change m

An amazing new weight loss product! Lose 10lbs a week,I didnt even change my diet! <http://t.co/1cY1Hmoh>

loss product! Lose 10lbs a week,I didnt even change my diet! <http://t.co/G1gqicFE>

An amazing new weight loss product! Lose 10lbs a week,I didnt ev

Results

◀ [Drone hijacked by hackers from Tex...](#)

[Predictive policing brings burglary n...](#) ▶

Twitter hit by amazing weight loss spam attack [VIDEO]

Join thousands of others, and sign up for Naked Security's newsletter

you@example.com

Do it!

[Don't show me this again](#) X

by [Graham Cluley](#) on July 2, 2012 | [Comments \(1\)](#)

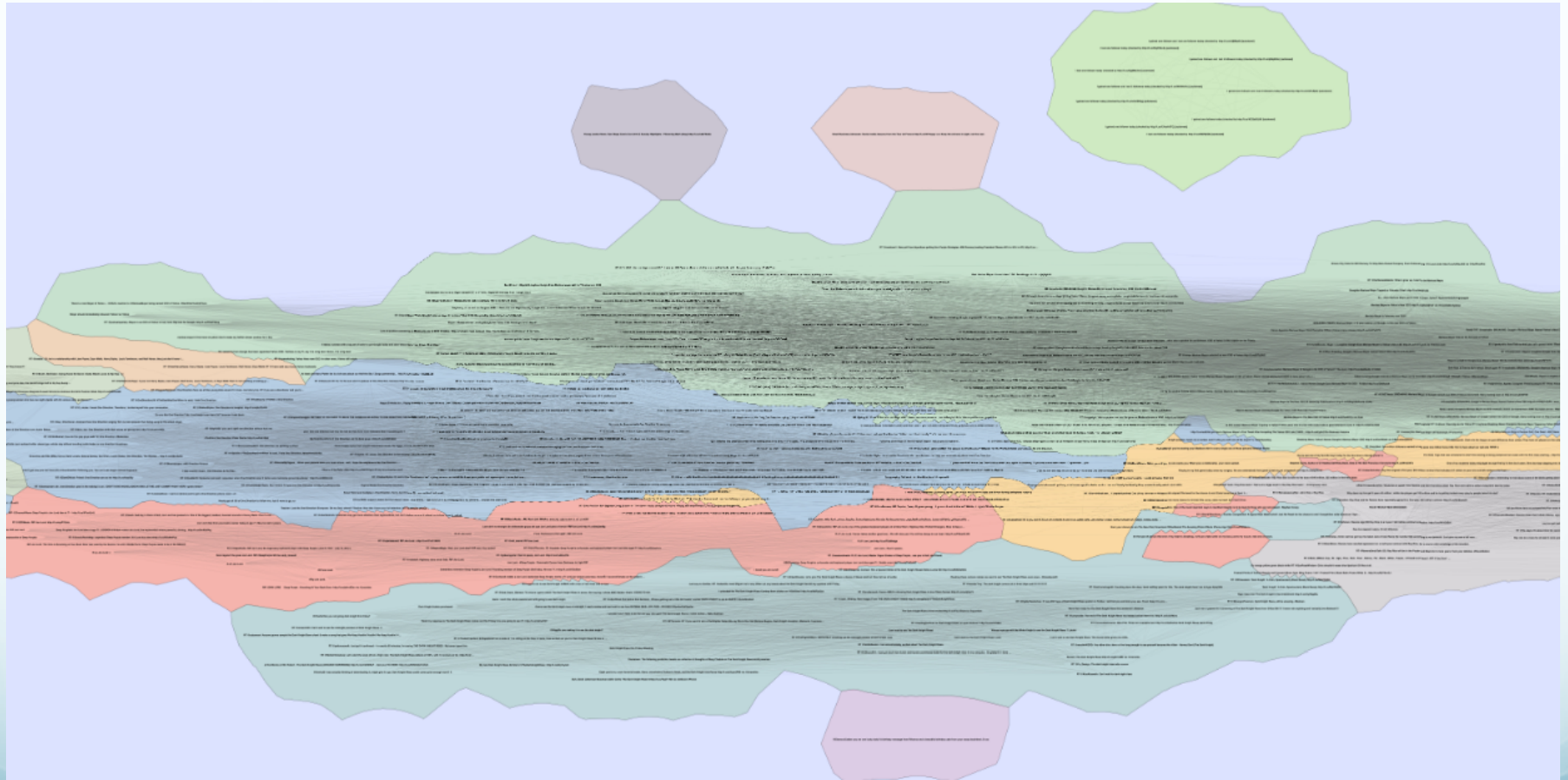
FILED UNDER: [Featured](#), [Social networks](#), [Spam](#), [Video](#)

Has your Twitter account been spewing out messages about an amazing weight loss product?

Many users' accounts have been posting messages like this, pointing to a website promoting a "miracle" Acai Berry diet:

An amazing new weight loss product! Lose 10lbs a week,I didnt even change my diet!

Given a cluster of events, can you tell what day it was?



Given a cluster of events, can you tell what day it was?

- Events

Marissa Meyer becomes CEO of Yahoo -> July 16th

Love and Hip Hop airs that day -> July 16th

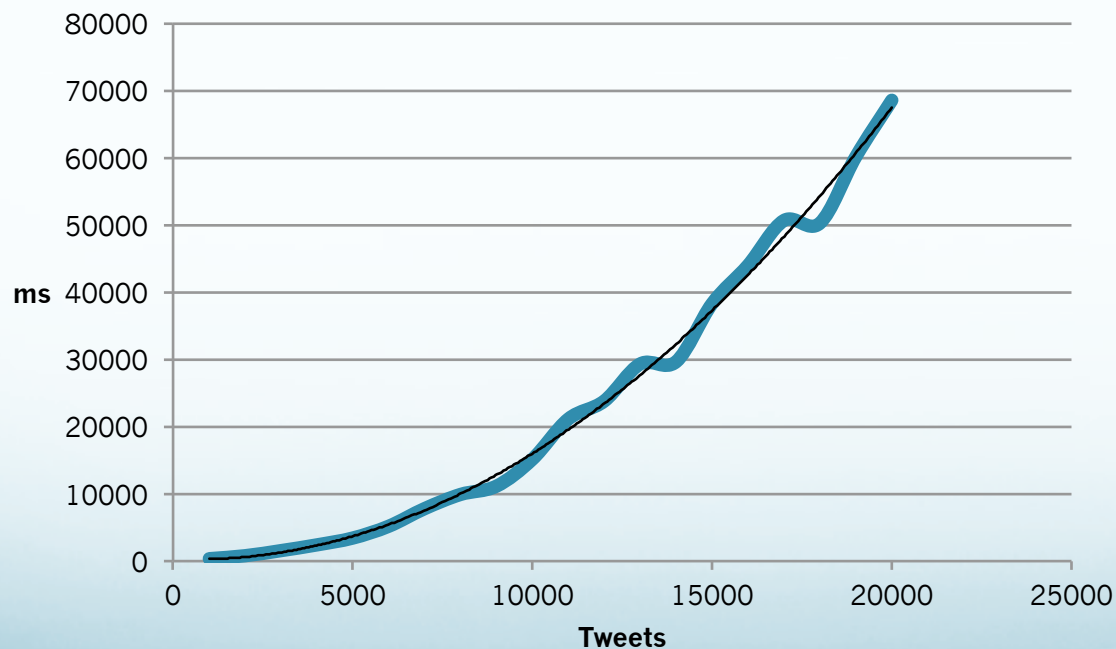
Jon Lord dies -> July 16th

Ray Rice gets signed to the Ravens -> July 16th

Easy to tell!

Complexity

- The complexity is $O(n^2)$, because we consider all possible edges between tweets in the beginning



Future work

- How do we name the clusters so that it best represents what people are talking about in that cluster?
- How can we see what trends change with time?
- Compare different ways of clustering

Questions and Comments



Acknowledgements

- Thanks to:

James Abello, Yifan Hu, Neel Parikh, Michael Poplavski, and Kevin Wong for their help

References

- Nakatani Shuoyo, author of the language detector <http://shuyo.wordpress.com/>
- Graphviz <http://www.graphviz.org/>
- Gvmap http://www.graphviz.org/Gallery/undirected/gd_1994_2007.html
- <http://www.setlist.fm/setlists/nicki-minaj-53d4e769.html>
- <http://www.cnn.com/2012/07/08/showbiz/obit-borgnine/index.html>
- <http://www.sugarscape.com/main-topics/celebrities/739070/perez-hilton-one-direction-will-break>
- <http://www.suntimes.com/entertainment/conner/13706153-452/rb-star-frank-ocean-makes-media-waves-with-admission-hes-gay.html>
- <http://beta.abcfamily.go.com/schedule>
- <http://www.lasvegassun.com/news/2012/jul/08/roger-federer-wins-7th-wimbledon-title/>
- <http://nakedsecurity.sophos.com/2012/07/02/twitter-weight-loss-spam-video/>