Social Media & Text Analysis

lecture 2 - Twitter API

CSE 5539-0010 Ohio State University

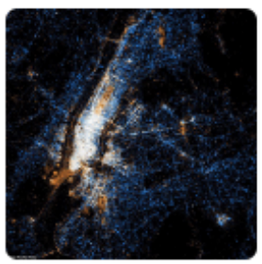
Instructor: Wei Xu

Website: socialmedia-class.org

Course Website

socialmedia-class.org

Social Media & Text Analytics Syllabus Twitter API Tutorial Homework



A visualization showing the location of Twitter messages (blue) and Flickr photos (orange) in New York City by Eric Fischer

Social media provides a massive amount of valuable information and shows us how language is actually used by lots of people. This course will give an overview of prominent research findings on language use in social media. The course will also cover several machine learning algorithms and the core natural language processing techniques for obtaining and processing Twitter data.

Instructor

Wei Xu is an assistant professor in the Department of Computer Science and Engineering at the Ohio State University. Her research interests lie at the intersection of machine learning, natural language processing, and social media. She holds a PhD in Computer Science from New York University. Prior to joining OSU, she was a postdoc at the University of Pennsylvania. She is organizing the ACL/COLING Workshop on Noisy User-generated Text, serving as a workshop co-chair for ACL 2017, an area chair for EMNLP 2016 and the publicity chair for NAACL 2016.

Time/Place new

Fall 2016, CSE 5539-0010 **The Ohio State University**
Cockins Hall **Room 218 | Wednesday 2:20PM – 4:10PM**
dual-listed undergraduate and graduate course

Prerequisites

In order to succeed in this course, you should know basic probability and statistics, such as the chain rule of probability and Bayes' rule. On the programming side, all projects will be in Python. You should understand basic computer science concepts (like recursion), basic data structures (trees, graphs), and basic algorithms (search, sorting, etc).

Course Readings

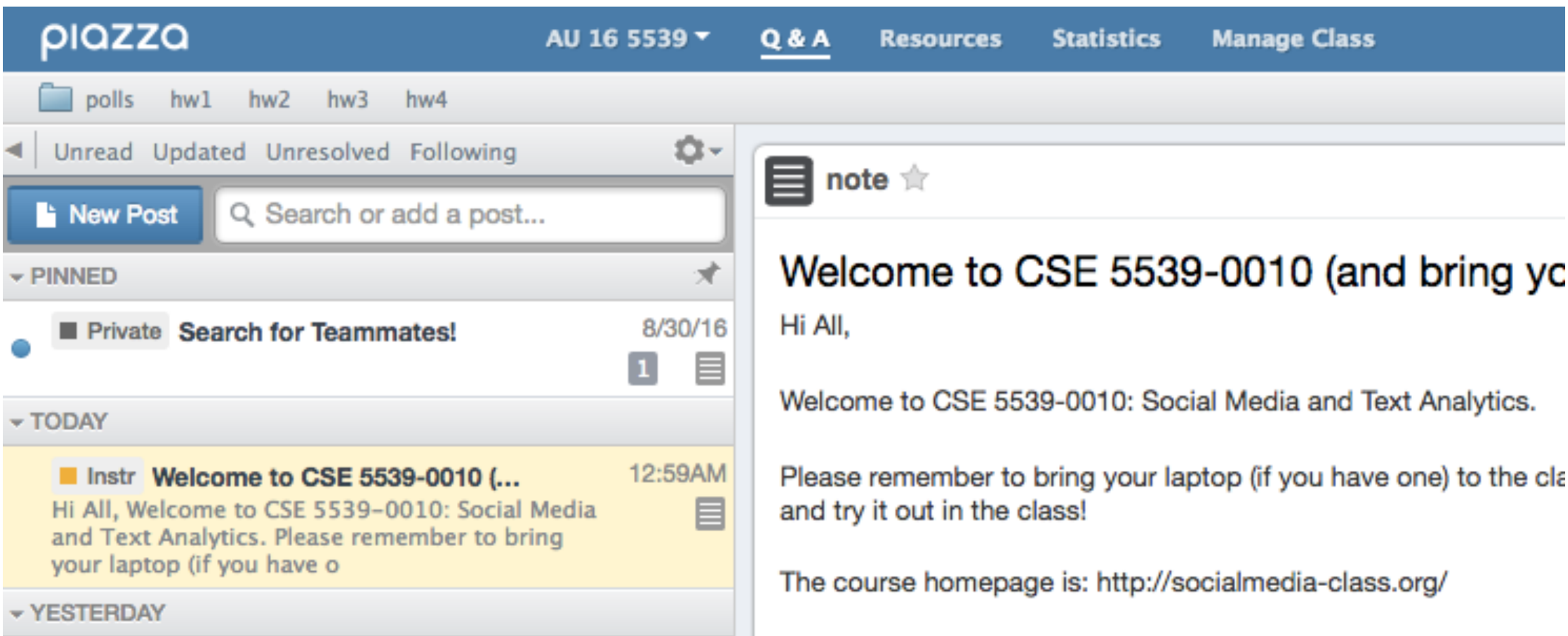
Various academic papers

Previous Offerings

Summer 2016, The North American Summer School on Logic, Language, and Information (NASSLLI)
Teaching evaluation was 5.72 out of 6 at NASSLLI; average across all instructors was 5.23.
Summer 2015, University of Pennsylvania (where this course was first designed and taught)

Have a Question?

- **Ask in class!**
- **Office Hour:** Tue 4:15 pm — 5:15 pm, Dreese 495
- **Piazza Q&A Board** (a Module within OSU Canvas)



This is a **Special Topic** Class

- It is about NLP **research**, not programming. (pre-requirements: familiar with Python programming)

- Homework #2 can be **difficult** (not about software engineering, but machine learning algorithm — difficult to debug).

- Students are required to think hard and **independently** for solutions. No direct answer or help (no spoilers!) will be given to direct questions about homework.

This is a **Special Topic** Class

Not logged in Talk Contributions Preferences Create account Log in

Entry Discussion Citations Read Edit History Search Wiktionary

Wiktionary
The free dictionary

Main Page
Community portal
Preferences
Requested entries
Recent changes
Random entry
Help
Glossary
Donations
Contact us

Tools
What links here
Related changes
Upload file
Special pages
Permanent link
Page information
Cite this page

Visibility
Show translations

give a man a fish and you feed him for a day; teach a man to fish and you feed him for a lifetime

Contents [hide]

1 English
1.1 Etymology
1.2 Proverb
1.2.1 Translations

English [edit]

Etymology [edit]

The oldest English-language use of the proverb has been found in Anne Isabella Thackeray Ritchie's (1837–1919) novel, *Mrs. Dymond* (1885), in a slightly different form:

" […] *if you give a man a fish he is hungry again in an hour. If you teach him to catch a fish you do him a good turn.*

The proverb has been attributed to many others, but no solid evidence has been produced.

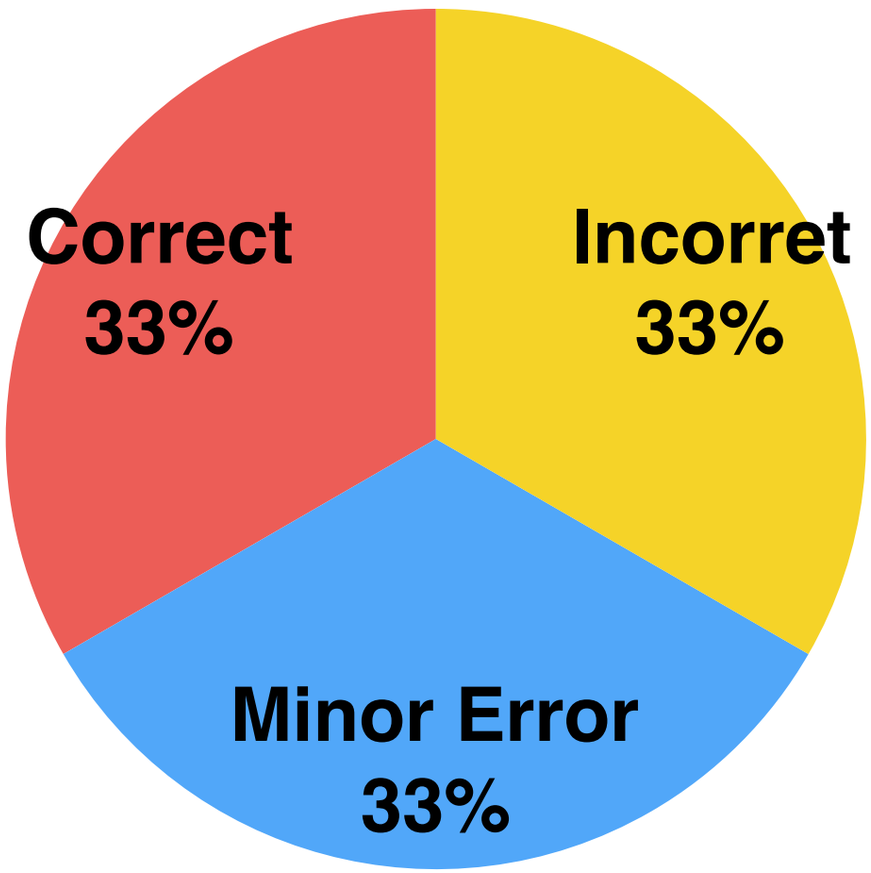
Proverb [edit]

give a man a fish and you feed him for a day; teach a man to fish and you feed him for a lifetime

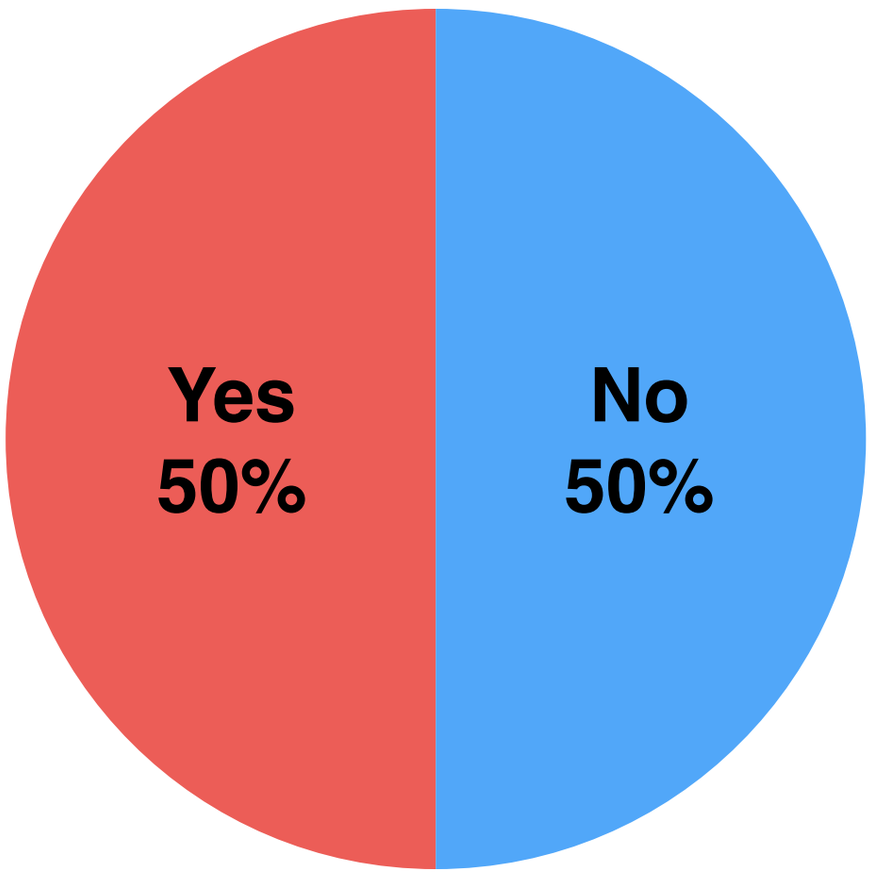
1. It is more worthwhile to teach someone to do something (for themselves) than to do it for them (on an ongoing basis).

Homework #2 (last year)

HW#2
(Main Algorithm)



HW#2
(Axillary Algorithm)



Alternatives

- **audit** the course or take LING 5801 (Computational Linguistics I)
- **more background**: CSE 3521, 5521, 3522, Stat 3460, 3470
- **other related courses**:
  - CSE 5525 Foundations of Speech and Language Processing
  - CSE 5523 Machine Learning
  - CSE 5522 Survey of Artificial Intelligence II: Advanced Techniques
  - CSE 5526 Introduction to Neural Networks

Quiz #1

- For events A and B, prove

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Quiz #1

- What does this regular expression mean?

```
147    Hashtag = "#[a-zA-Z0-9_]+"
```

Quiz #1

- Softmax function is defined as $softmax(\mathbf{x})_i = \frac{e^{x_i}}{\sum_j e^{x_j}}$
- prove

$$softmax(\mathbf{x}) = softmax(\mathbf{x} + c)$$

Useful for improving the numerical stability of the computation!

Quiz #1

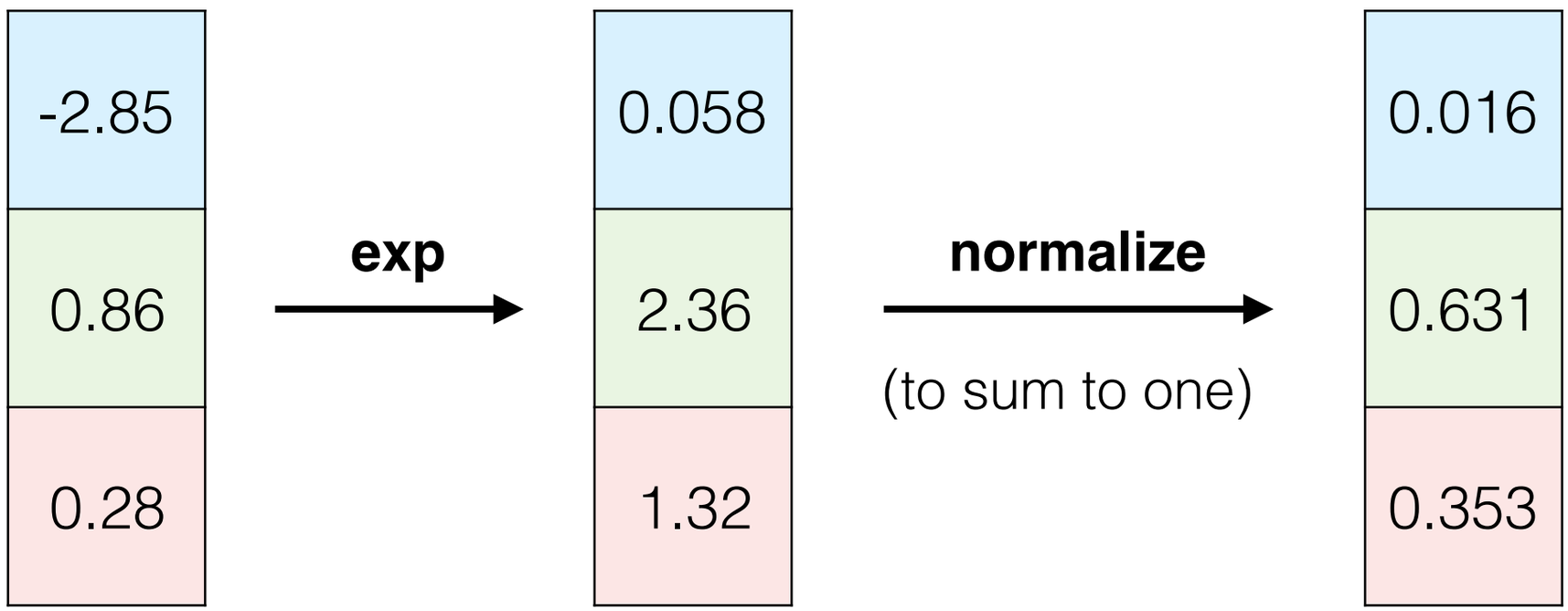
- implement Softmax function in Python (need to be computationally efficient)

```python
import numpy as np

def softmax(x):

    # Handle the special case: when vector x is only 1-dimensional
    if x.ndim <= 1:
        x = x - np.max(x)
        ex = np.exp(x)
        return ex / np.sum(ex)


    ### YOUR CODE HERE  -- for vectors that are not 1-dimensional




    ### END YOUR CODE

    return dist


# Check your Softmax implementation

print softmax(np.array([[101,102],[-1,-2]]))
```

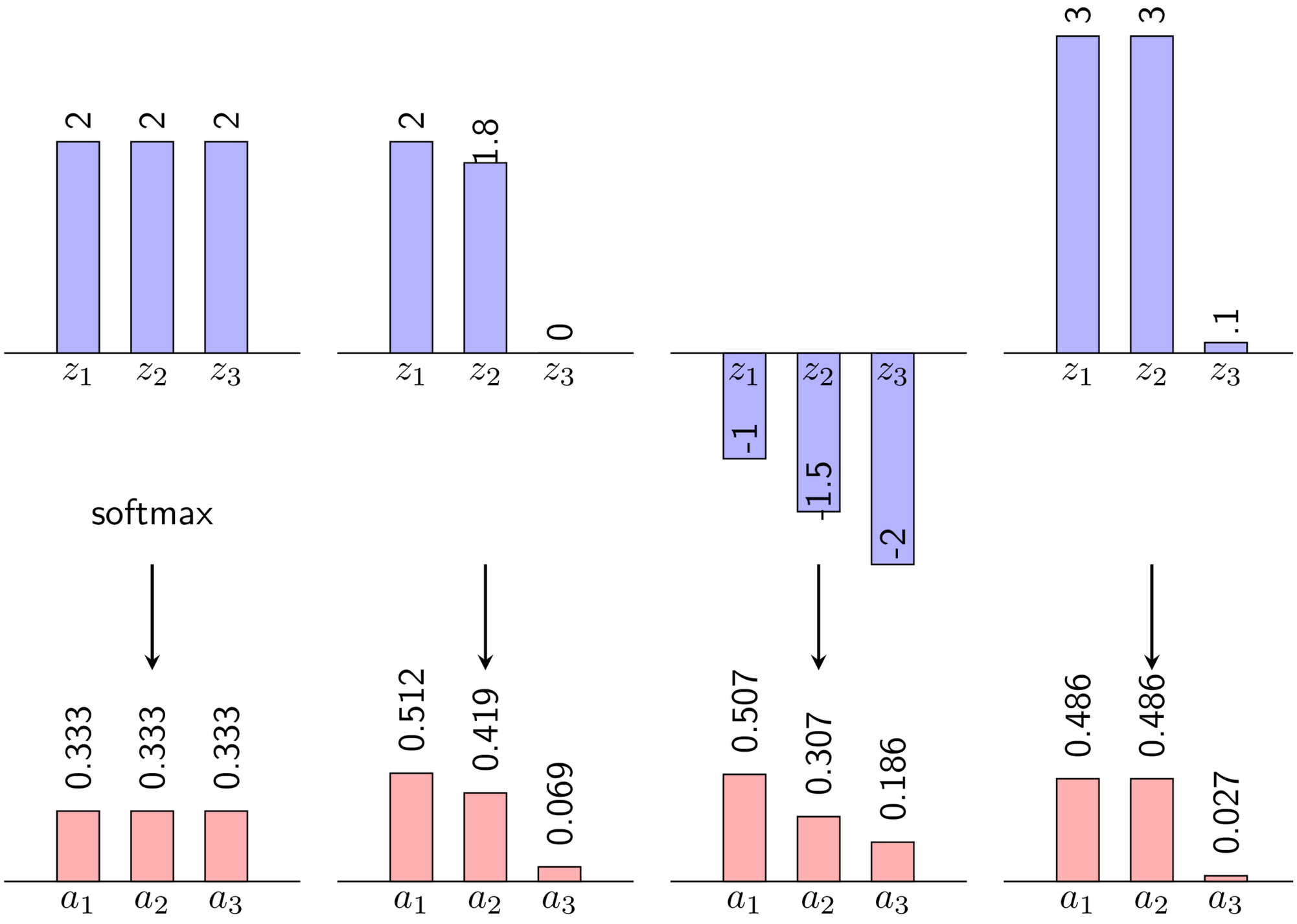
A normalization trick for numerical stability!
(highest value in the vector becomes 0)

Softmax Function

$$softmax(\mathbf{x})_i = \frac{e^{x_i}}{\sum_j e^{x_j}}$$



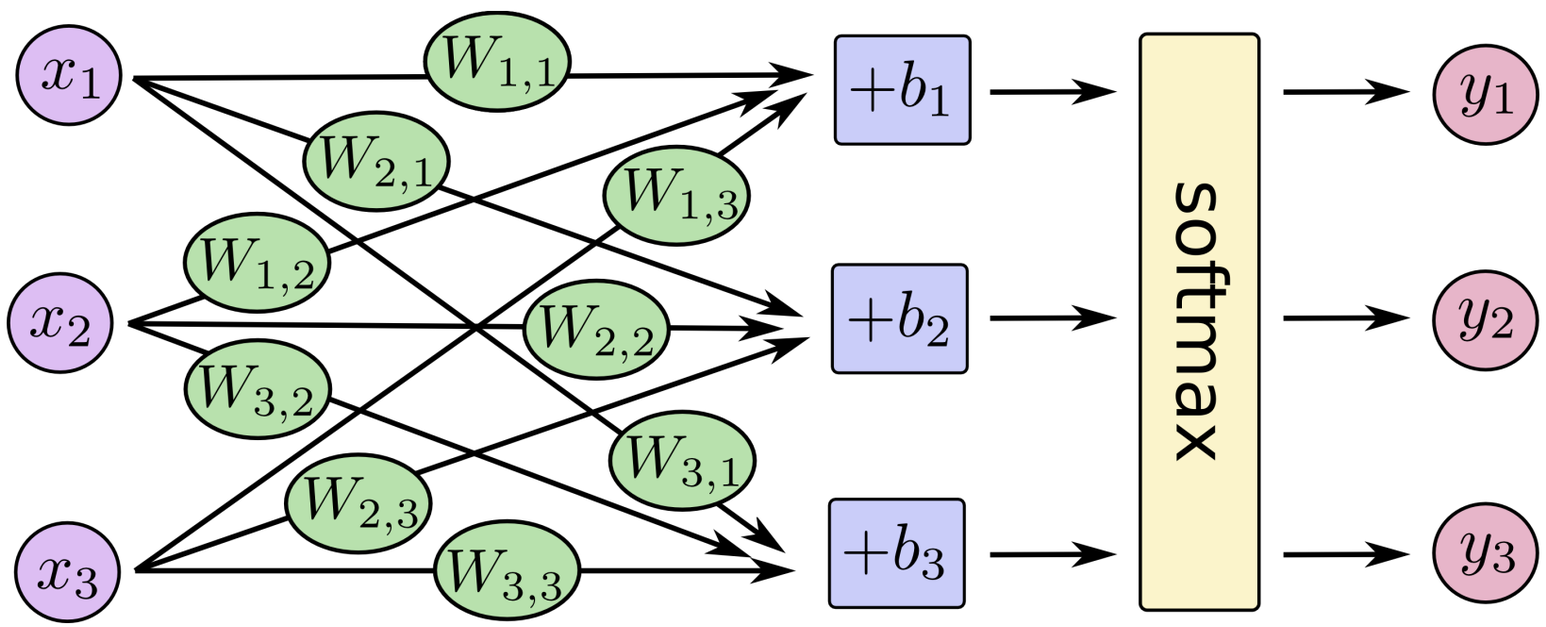
Softmax



see also: http://cs231n.github.io/linear-classify/#softmax

Softmax Function

- softmax regression (multinominal logistic regression)
- often used as the output layer in neural networks



- We will learn later in the class

Quiz #2

- derivative of the Sigmoid function:

- use the chain rule:

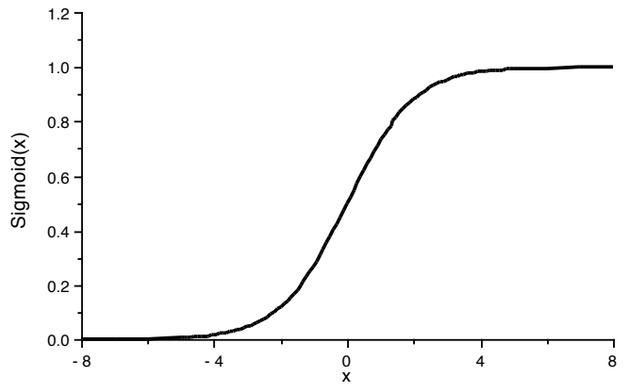
if $f = g(u)$ and $u = h(x)$, i.e. $f(x) = g(h(x))$, then:

$$\frac{df}{dx} = \frac{df}{du}\frac{du}{dx} = \frac{dg(u)}{du}\frac{dh(x)}{dx}$$

The Derivative of a Sigmoid

We noted earlier that the Sigmoid is a smooth (i.e. differentiable) threshold function:

$$f(x) = \text{Sigmoid}(x) = \frac{1}{1+e^{-x}}$$

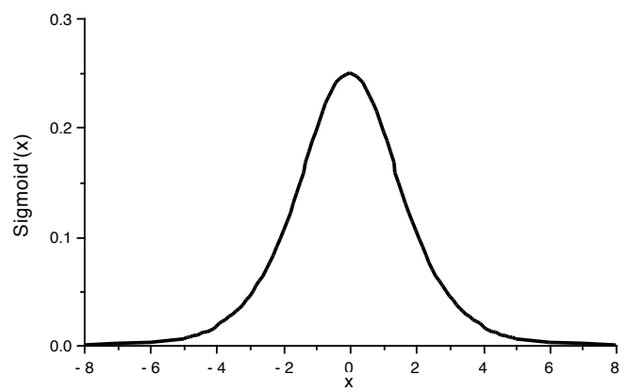


We can use the chain rule by putting $f(x) = g(h(x))$ with $g(h) = h^{-1}$ and $h(x) = 1 + e^{-x}$ so

$$\frac{\partial g(h)}{\partial h} = -\frac{1}{h^2} \quad \text{and} \quad \frac{\partial h(x)}{\partial x} = -e^{-x}$$

$$\frac{\partial f(x)}{\partial x} = -\frac{1}{(1+e^{-x})^2} \cdot (-e^{-x}) = \left(\frac{1}{1+e^{-x}}\right).\left(\frac{1+e^{-x}-1}{1+e^{-x}}\right)$$

$$f'(x) = \frac{\partial f(x)}{\partial x} = f(x).(1-f(x))$$



This simple relation will make our equations much easier and save a lot of computing time!

Source: John A. Bullinaria

Twitter API Tutorial: socialmedia-class.org

Social Media & Text Analytics | Syllabus | Twitter API Tutorial | Homework Assignments

Twitter's 404 error page -- the Fail Whale

Twitter API tutorial

by Wei Xu (July 1, 2015)

Follow @cocoweixu

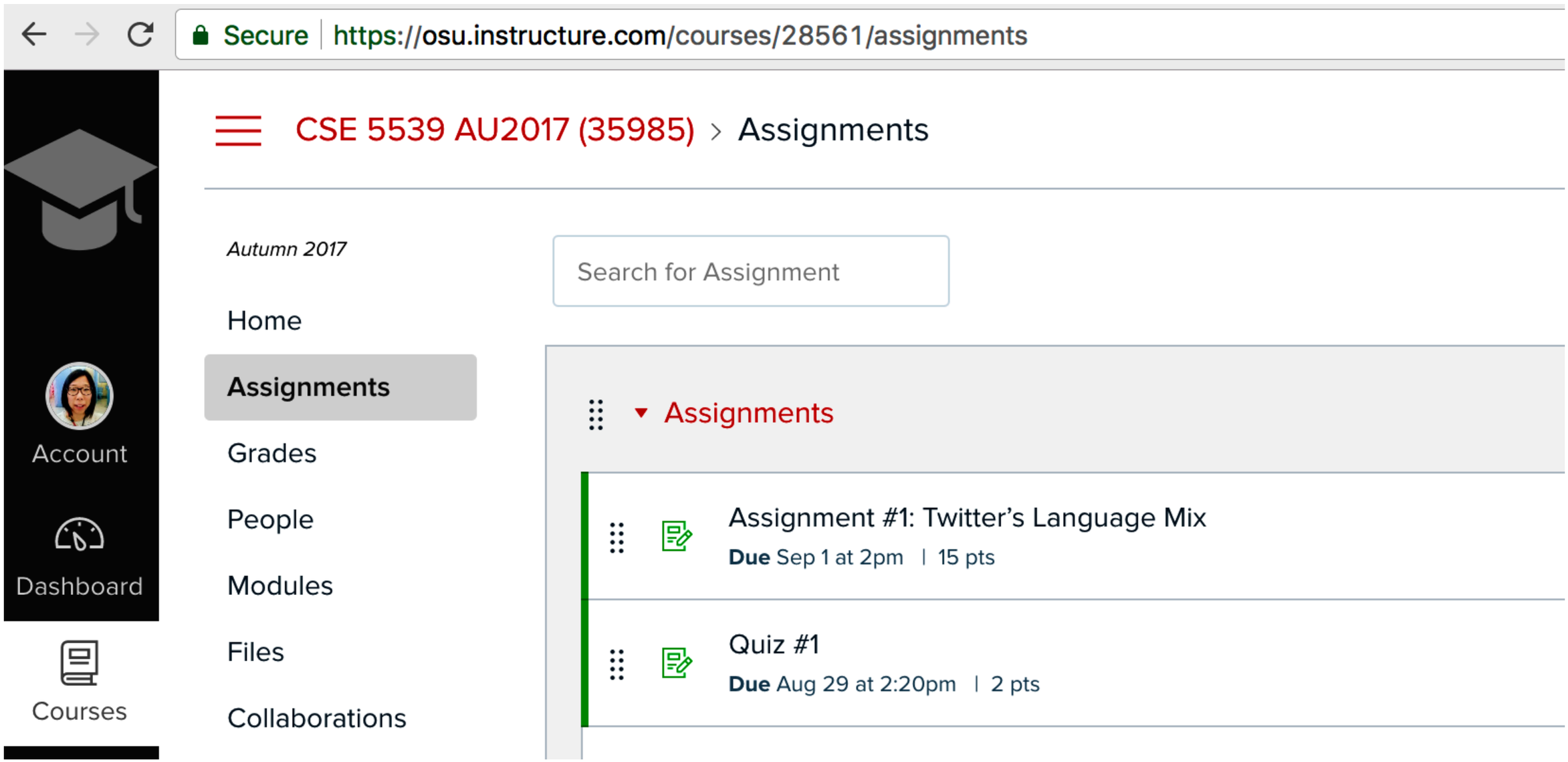
1. Getting Twitter API keys

To start with, you will need to have a Twitter account and obtain credentia
Twitter developer site to access the Twitter API, following these steps:

- Create a Twitter user account if you do not already have one.
- Go to https://apps.twitter.com/ and log in with your Twitter user acc
  your user account.
- Click "Create New App"

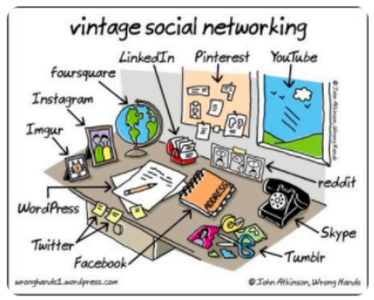
Homework #1 is out

Due next Tuesday (Sep 5)



Reading #1 is out
Due Sep 12

Social Media & Text Analytics | Syllabus | Twitter API Tutorial | Homework



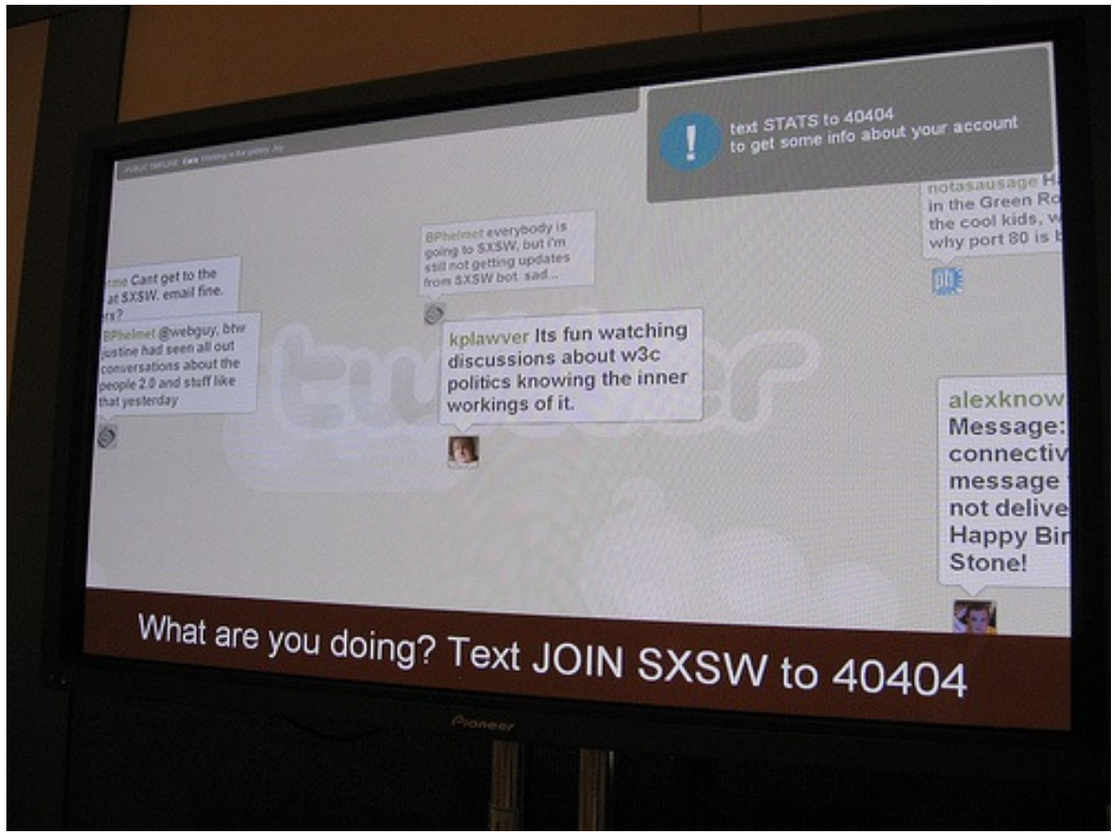
Vintage Social Media

Subject to change as the Fall 2017 term progresses. ★ marks the required reading.

Lecture	Topic	Readings
August 22, 2017	Introduction • Introduction of social media and natural language processing research • Overview of the course	★ Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength Across Diverse Cultures by Golder and Macy (Science 2011) Psychological Language on Twitter Predicts County-Level Heart Disease Mortality by Eichstaedt et al. (Psychological Science 2015) Google's Python class by Nick Parlante Intro to NLP in Python by Bird, Klein, Loper
August 29, 2017	Twitter and Twitter API Tutorial [Quiz1 due] • Brief history of Twitter • Key features of Twitter • Hands-on instructions on obtaining Twitter data via APIs	★ Twitter API Tutorial by the instructor Wei Xu ★ What is Twitter, a Social Network or a News Media? by Kwak, Lee, Park and Moon (WWW 2010)
TBA	AI Seminar Talk by Wuwei Lan • • A Continuously Growing Dataset of Sentential Paraphrases	A Continuously Growing Dataset of Sentential Paraphrases by Wuwei Lan, Siyu Qiu, Hua He, Wei Xu (EMNLP 2017)

Twitter History

- Jack Dorsey's idea
  (a NYU undergraduate then)
- 1st tweet on March 21, 2006
- exploded at SXSW 2007
  (20k→60k tweets/day)
- 100m tweets/quarter in 2008,
  50m tweets/day in 2010,
  400m tweets/day in 2013
- Huge API usage was
  unexpected as was the rise of
  the @ sign for replies

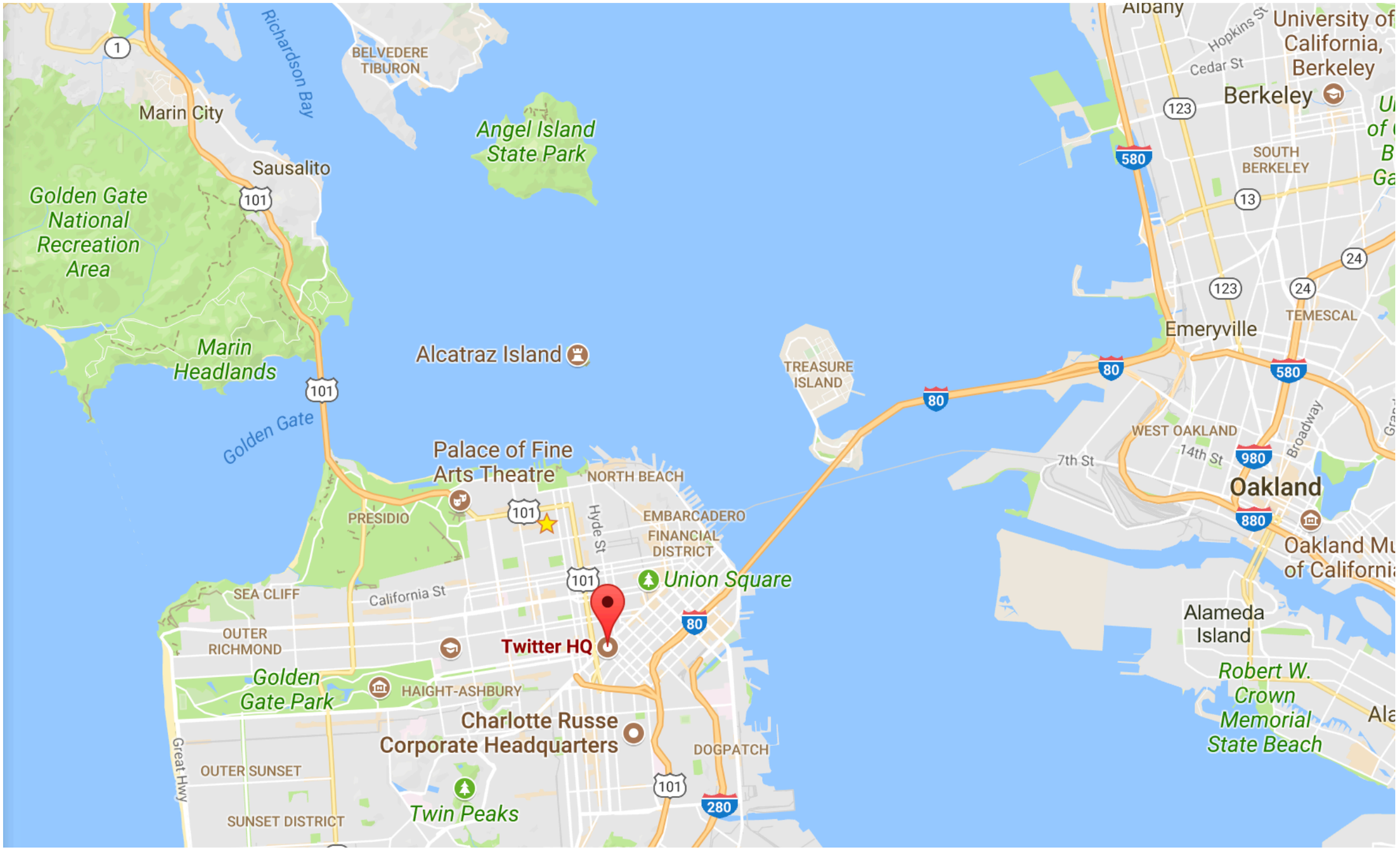


Twitter staff received the festival's Web Award prize with the remark "we'd like to thank you in 140 characters or less. And we just did!"

Twitter History

- IPO in 2013 Q4
- market value $24b, revenue $435m, net loss $162m in 2015 Q1
- CEO Dick Costolo resigned July 1st, 2015

Twitter HQ (since 2012)

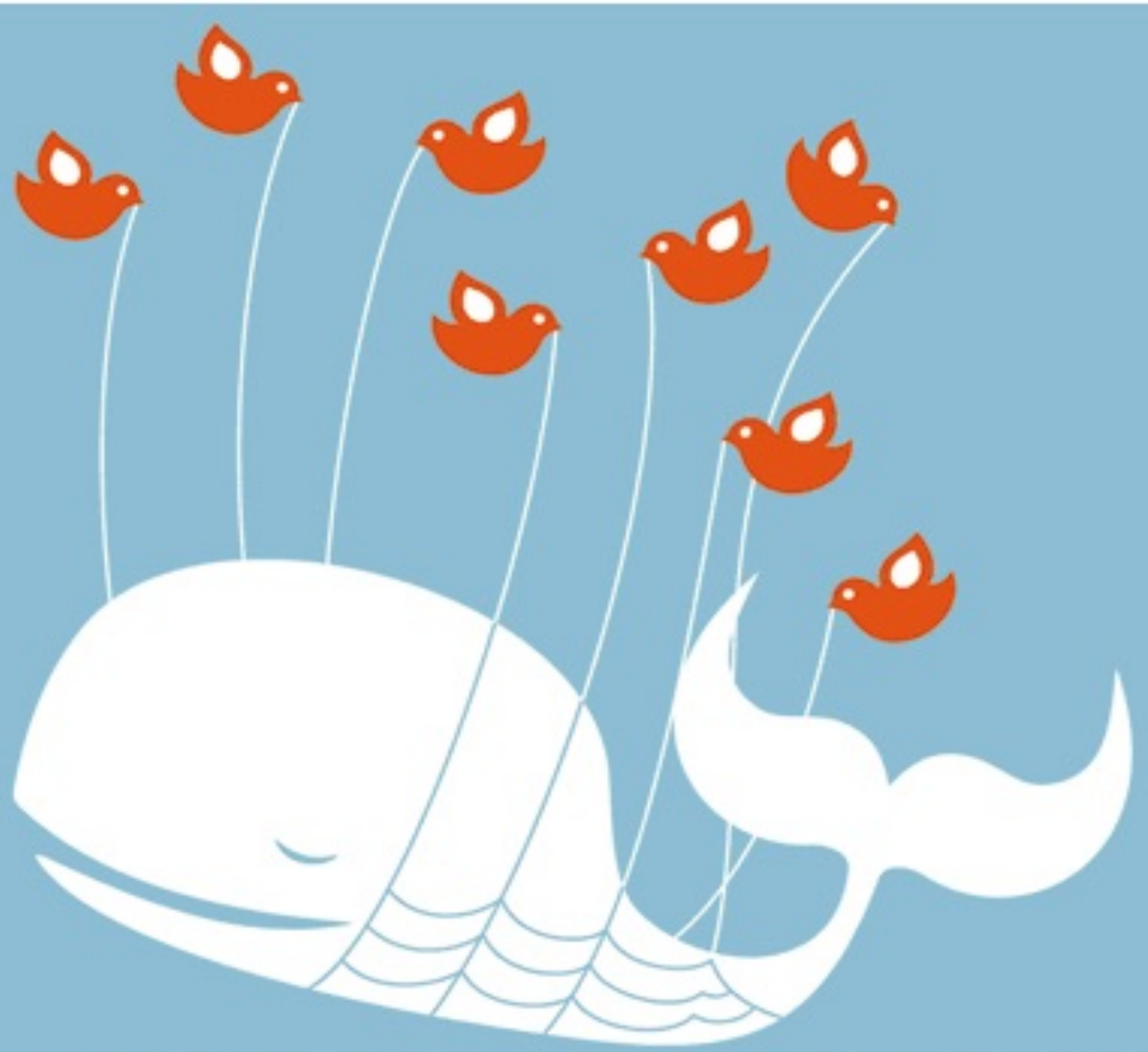


twitter

Home ›

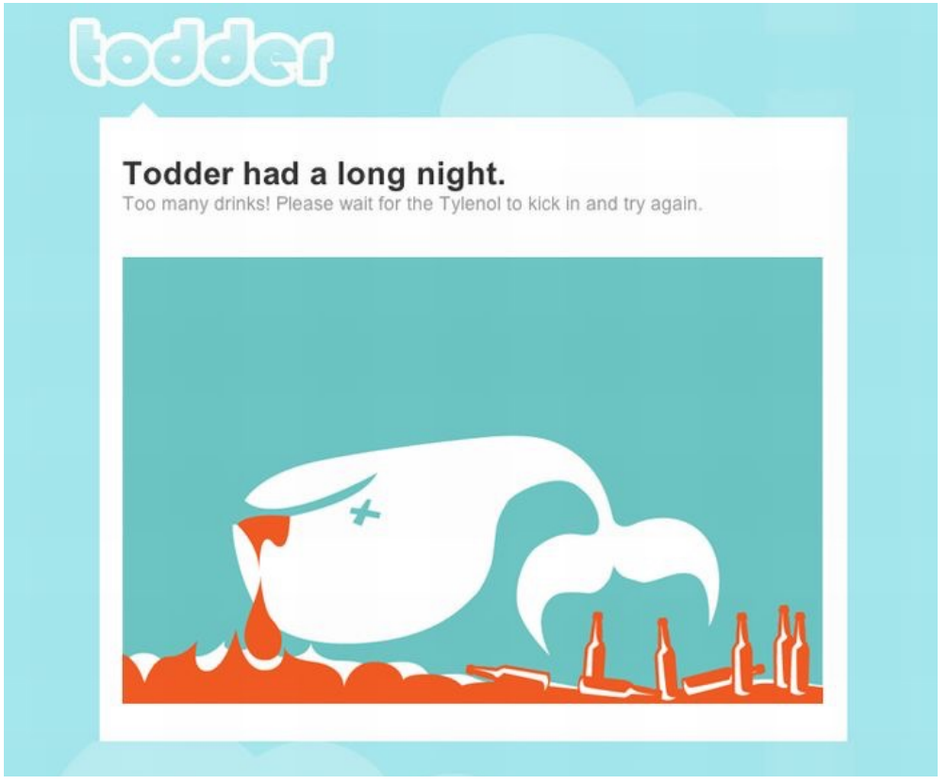
Twitter is over capacity.

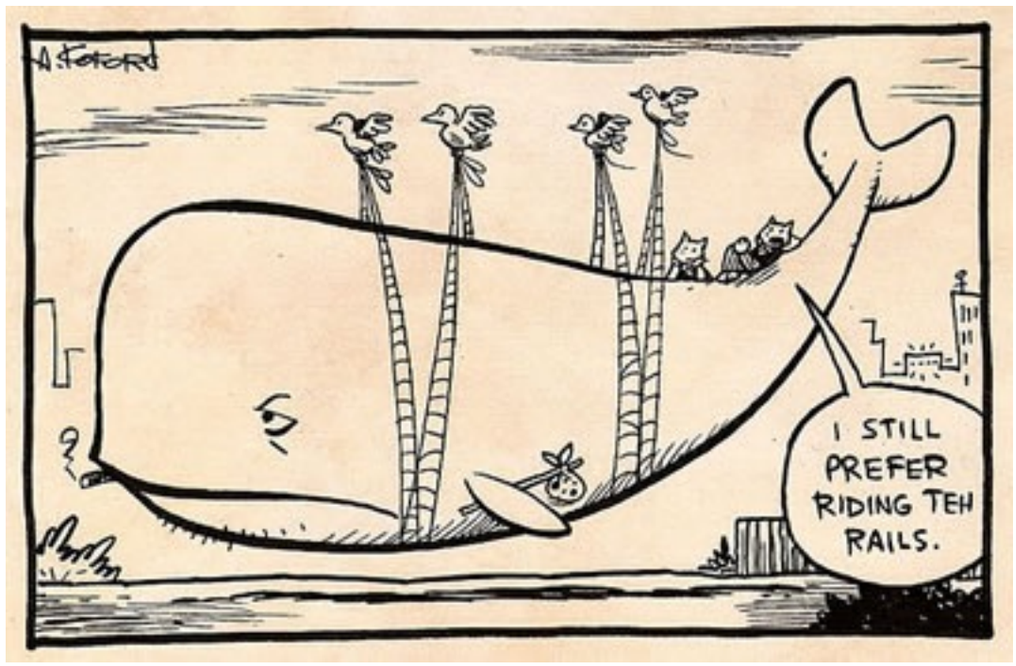
Please wait a moment and try again. For more information, check out Twitter Status »



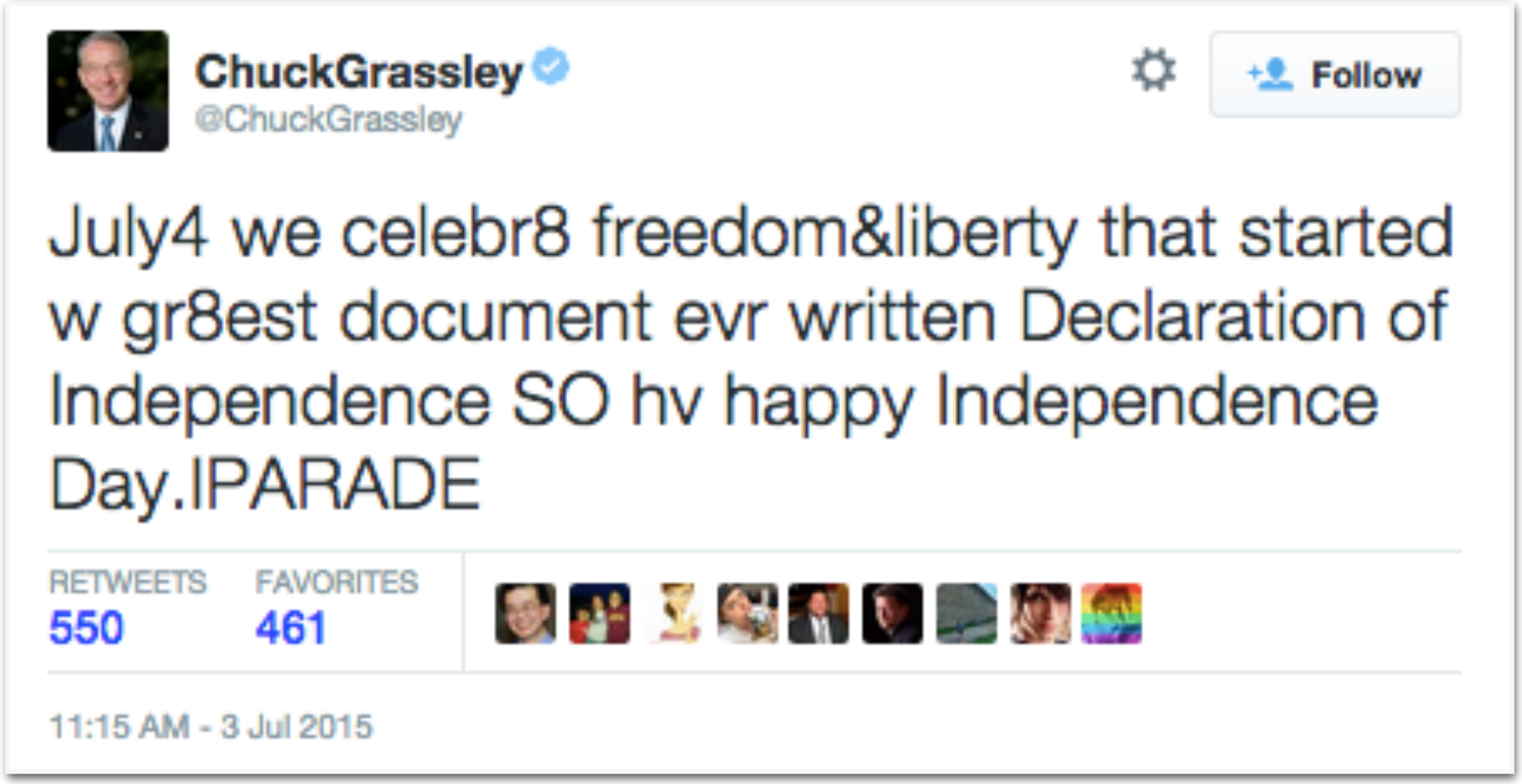
English Deutsch Español Français Italiano 日本語

© 2011 Twitter About Us Contact Blog Status API Help Jobs TOS Privacy

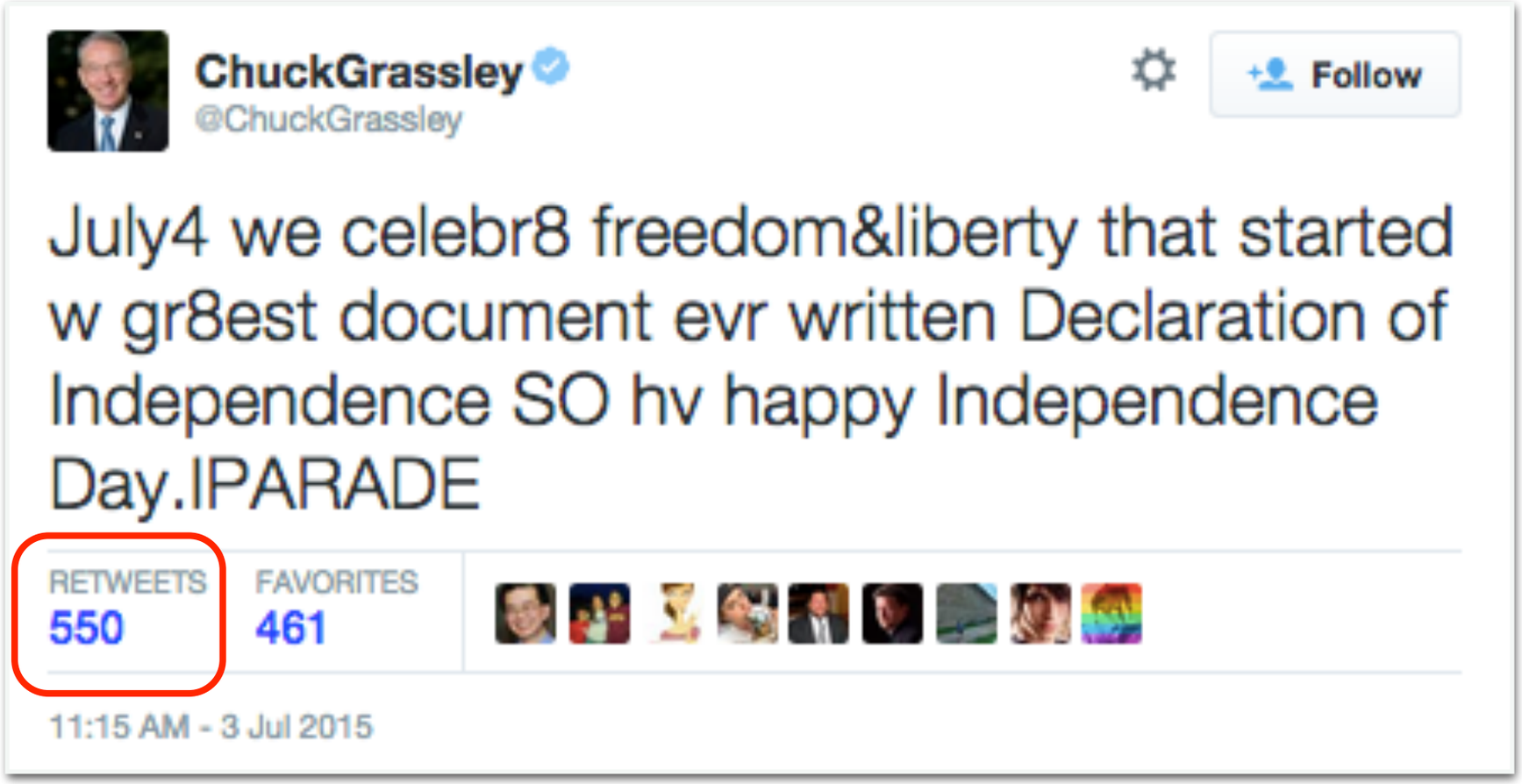

todder
Todder had a long night.
Too many drinks! Please wait for the Tylenol to kick in and try again.


I STILL PREFER RIDING TEH RAILS.

Tweets

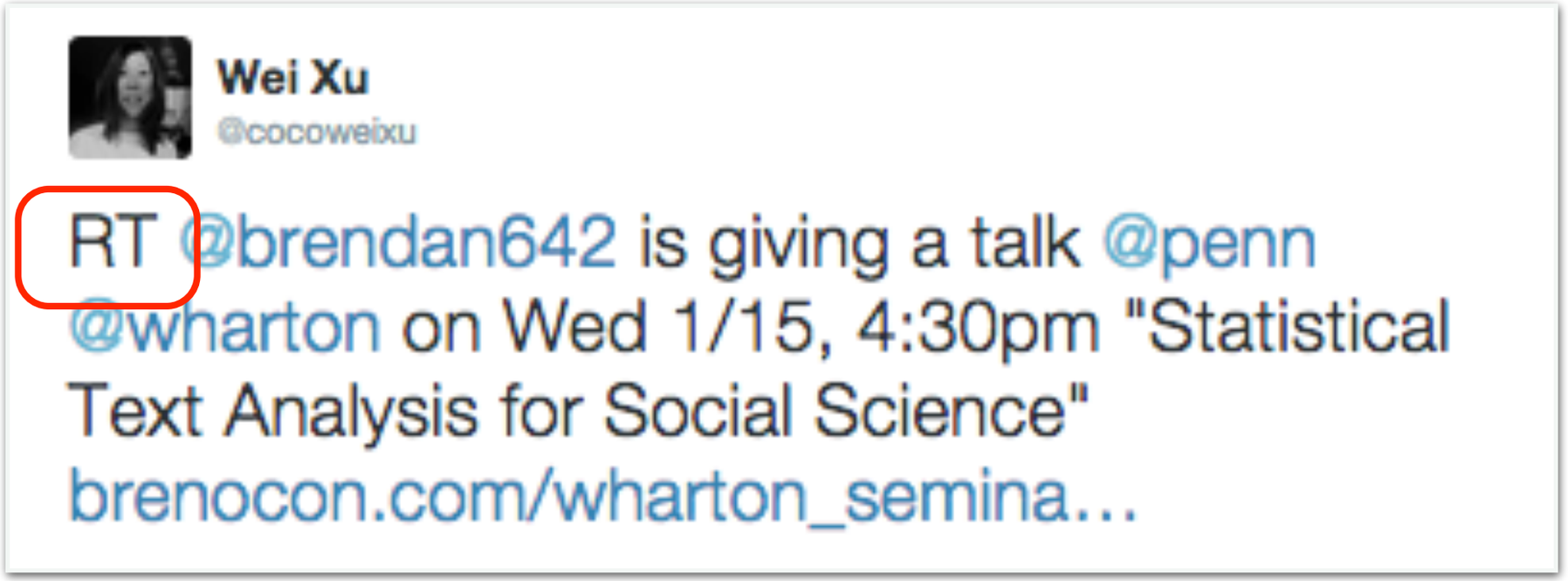

ChuckGrassley
@ChuckGrassley
Follow
July4 we celebr8 freedom&liberty that started w gr8est document evr written Declaration of Independence SO hv happy Independence Day.IPARADE
RETWEETS
550
FAVORITES
461
11:15 AM - 3 Jul 2015

ReTweets



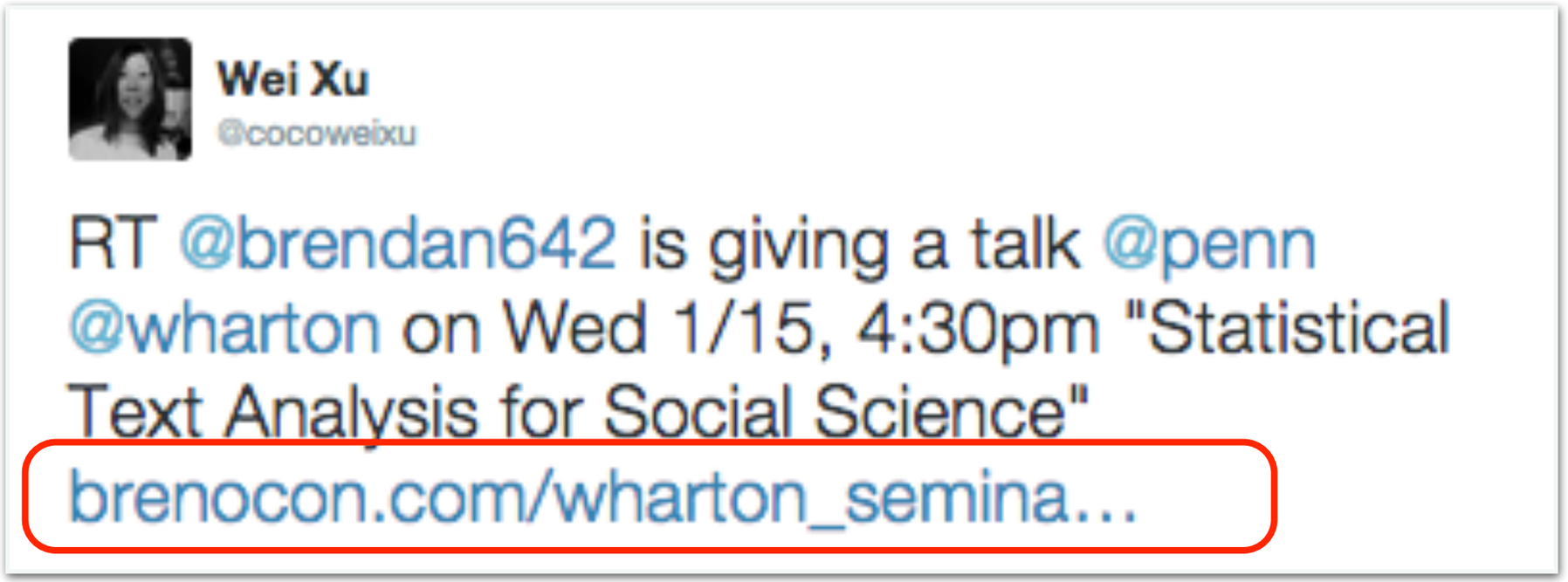
a re-posting of someone else's Tweet

ReTweets



- not an official Twitter feature
- often signifies quoting another user
- sometimes creates problems for data analytics

Embedded Links

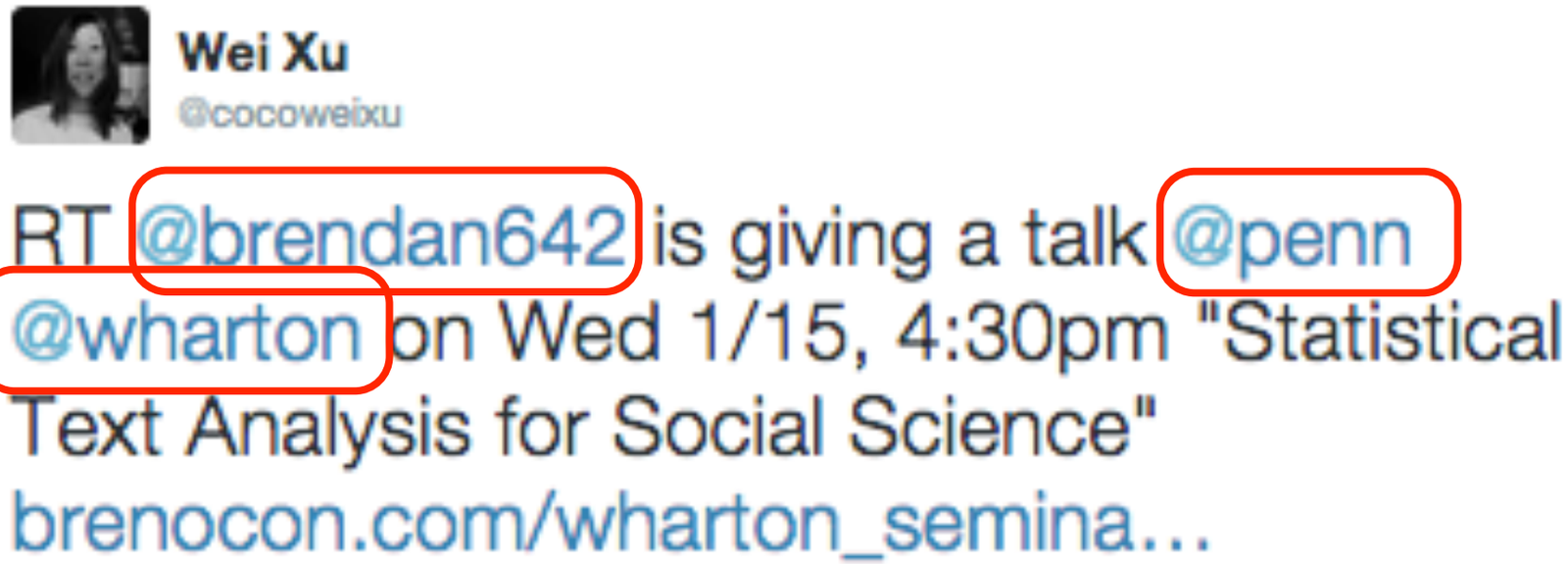


- shortened for display

Embedded Links

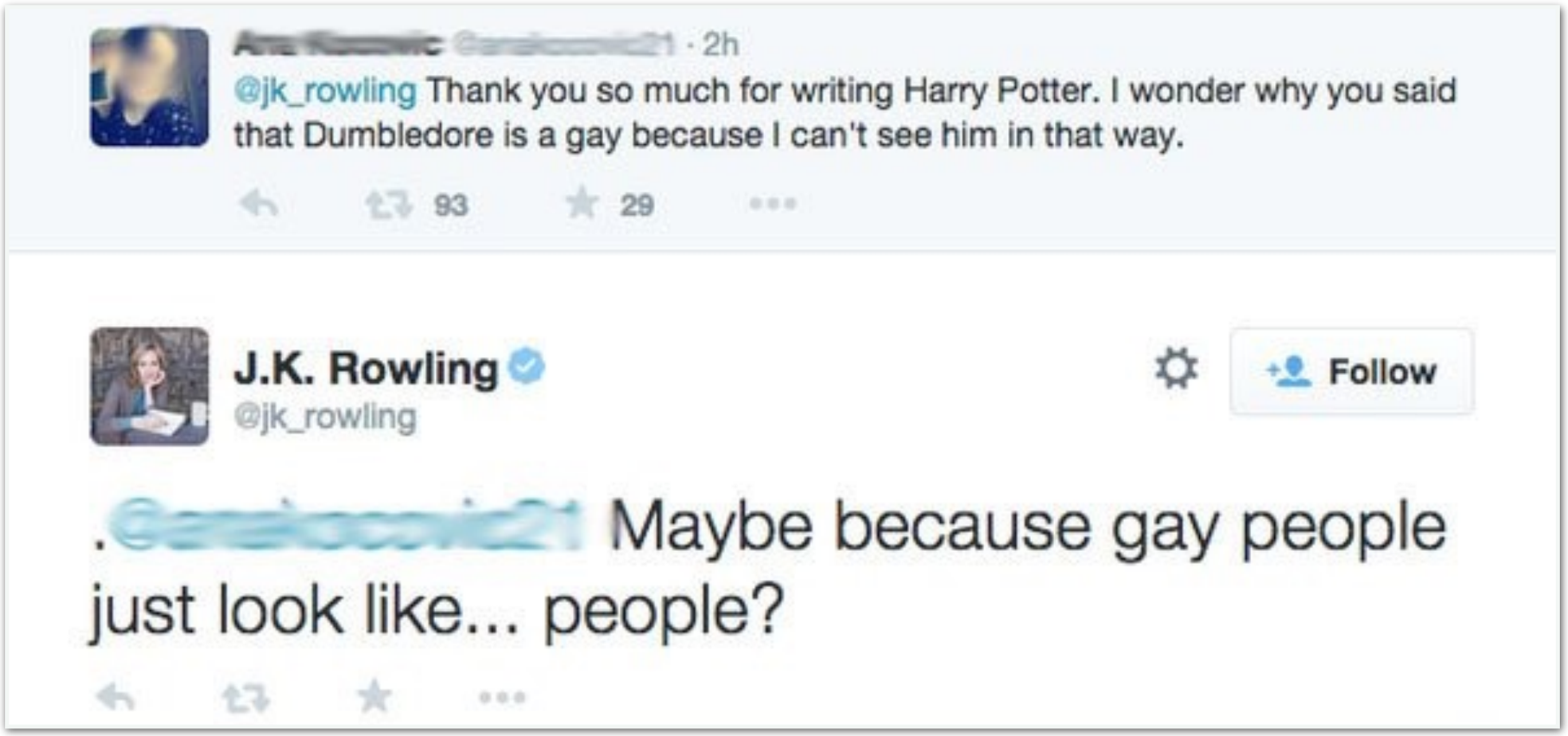
- can provide extra external information for text processing

Mentions



- user's @username anywhere in the body of the Tweet

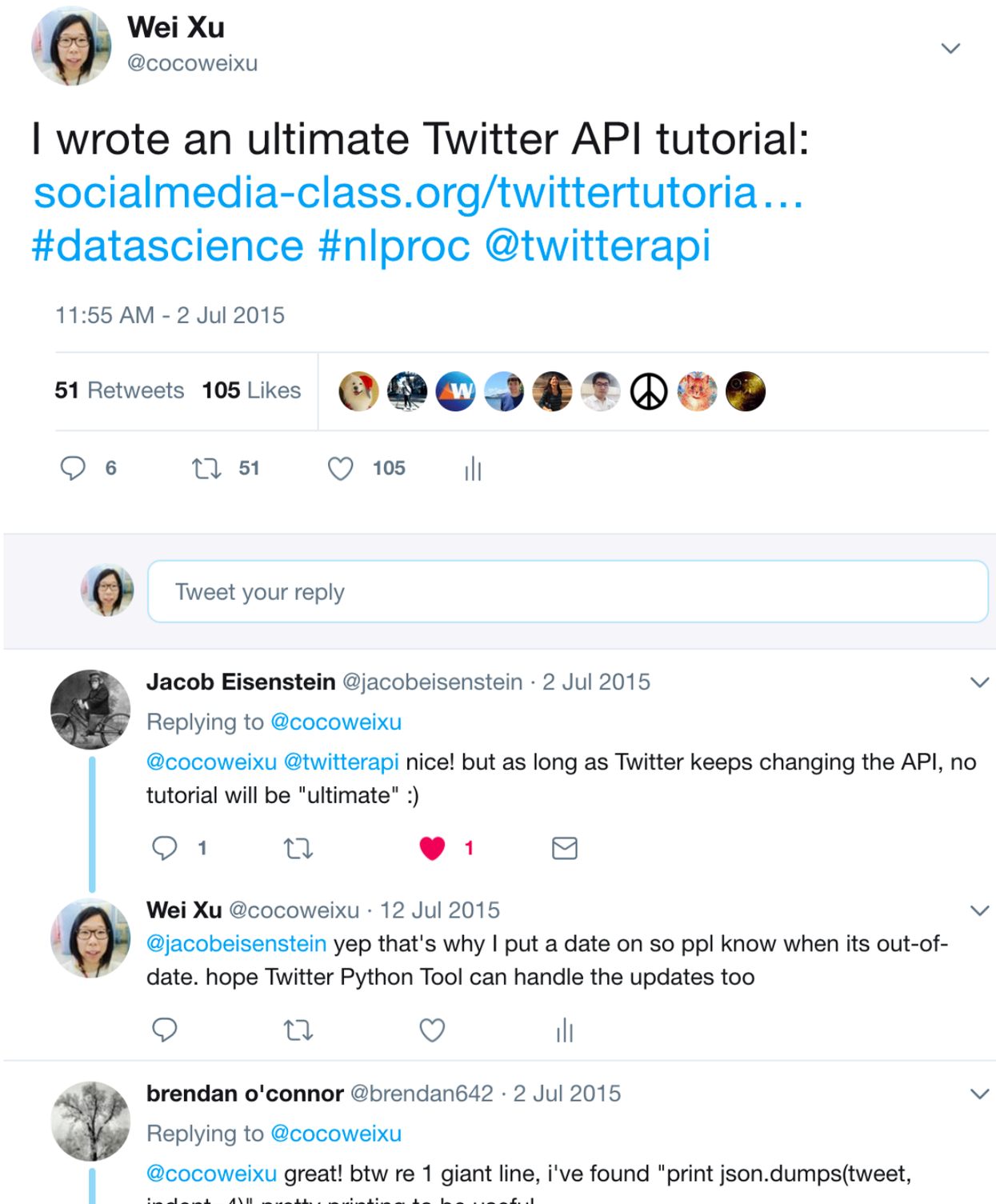
Replies/Conversations



- Tweet starts with a @username

Replies/Conversations

- can have multi-round conversations



What are the top forums or discussion websites where leading researchers in the field of Natural Language Processing interact?

1 Answer

Jordan Boyd-Graber, answering questions on Quora because the stakes are so low

Answered Mar 10

It seems to be Twitter (and to a lesser extent, Facebook). Follow your favorite researchers and often technical questions come up.

A random sampling of people I follow on Twitter (as sorted by Twitter):

- Alex Smola (@smolix) | Twitter
- Forough (@fpoursabzi) | Twitter
- Alice Zheng (@RainyData) | Twitter
- Thomas G. Dietterich
- Aaron Clauset (@aaronclauset) | Twitter
- UMD CLIP lab (@umdclip)
- Hugo Larochelle (@hugo_larochelle) | Twitter
- Russ Salakhutdinov
- Tom M Mitchell (@tommmitchell)
- Karl Moritz Hermann
- Edward Grefenstette
- Bert Huang (@berty38) | Twitter
- Tim Vieira (@xtimv) | Twitter
- Yoav Artzi (@yoavartzi) | Twitter
- Omer Levy (@omerlevy_) | Twitter
- Wei Xu (@cocoweixu) | Twitter
- Anima Anandkumar
- Naomi Saphra (@nsaphra) | Twitter
- Dirk Hovy (@dirk_hovy) | Twitter

Quora Home Answer Notifications 1 Search Quora Ask Question

Jason Eisner

computer science professor at Johns Hopkins

You can learn more about me and my research at http://cs.jhu.edu/~jason. On Quora, I typically answer technical questions about natural language processing and machine learning. Sometimes I also... (more)

Follow 23.2k Turn On Notifications Ask Question

Feeds

- Answers 216
- Questions 0
- Activity
- Posts 0
- Blogs 0
- Followers 23,283
- Following 5
- Topics 46
- Edits 1,269

216 Answers Most Recent / 30-Day Views

What are the topics in computer science?

Jason Eisner, computer science professor at Johns Hopkins
Answered Jul 24

You're off to a good start, but yes, there's plenty more! To get a sense of the breadth of CS, you can have a look through the ACM's curriculum guidelines for undergraduate CS education (last updat... (more)

Upvote 75 Downvote

What are the things I should know as a new CS PhD student?

Jason Eisner, computer science professor at Johns Hopkins
Answered Jun 15, 2015

[A2A] There's lots of advice on the web. Search for "how to be a good grad student" to get some of it.

How to be a Successful Graduate Student, by Mark Dredze (my colleague) and Hanna Wallach, is a good guide with a long list of links at the end, including a link to my own advice page.

2.4k Views · 24 Upvotes · Answer requested by Hao WU

Upvote 24 Downvote

Credentials & Highlights More

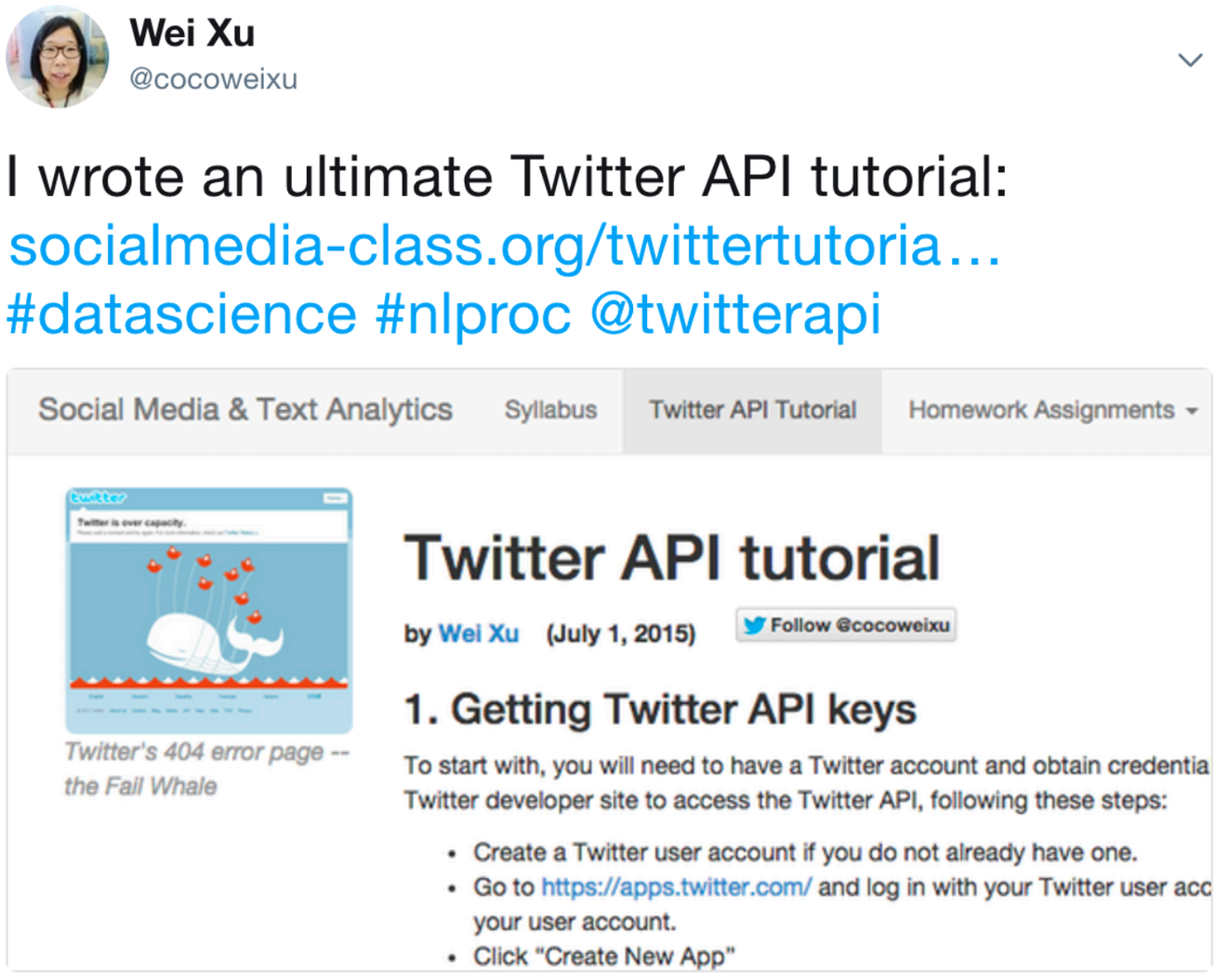
- Professor at Johns Hopkins University
  2001-present
- Studied at University of Pennsylvania
- Lives in Baltimore
- 2.7m answer views
  37.7k this month
- Top Writer
  2017 and 2016

Knows About

- Graduate School Education
  40 answers
- Academia
  28 answers
- Higher Education
  21 answers
- Machine Learning
  19 answers
- Natural Language Processing
  18 answers

View More

Images


Wei Xu
@cocoweixu
I wrote an ultimate Twitter API tutorial:
socialmedia-class.org/twittertutoria…
#datascience #nlproc @twitterapi
Social Media & Text Analytics
Syllabus
Twitter API Tutorial
Homework Assignments
Twitter is over capacity.
Twitter's 404 error page -- the Fail Whale
Twitter API tutorial
by Wei Xu (July 1, 2015)
Follow @cocoweixu
1. Getting Twitter API keys
To start with, you will need to have a Twitter account and obtain credentia
Twitter developer site to access the Twitter API, following these steps:
Create a Twitter user account if you do not already have one.
Go to https://apps.twitter.com/ and log in with your Twitter user acc
your user account.
Click "Create New App"
11:55 AM - 2 Jul 2015
51 Retweets
105 Likes
6
51
105

Hashtags

Wei Xu
@cocoweixu

I wrote an ultimate Twitter API tutorial:
socialmedia-class.org/twittertutoria…
#datascience #nlproc @twitterapi

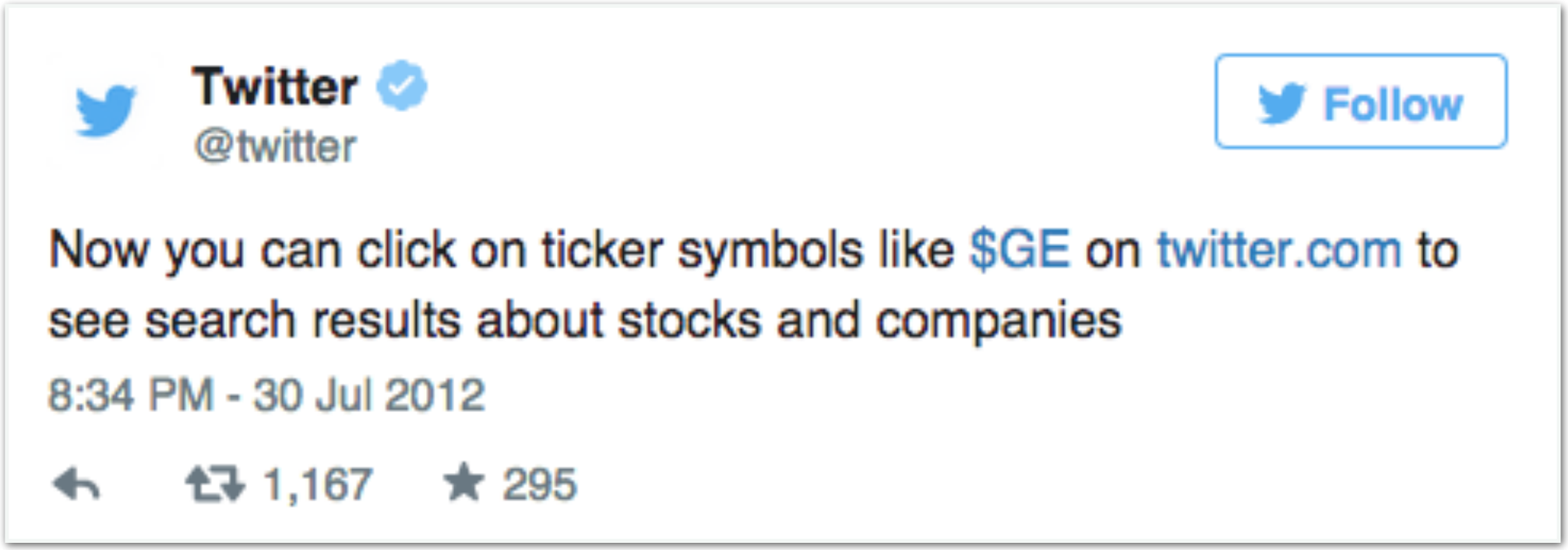
RETWEETS 45 LIKES 79

11:55 AM - 2 Jul 2015

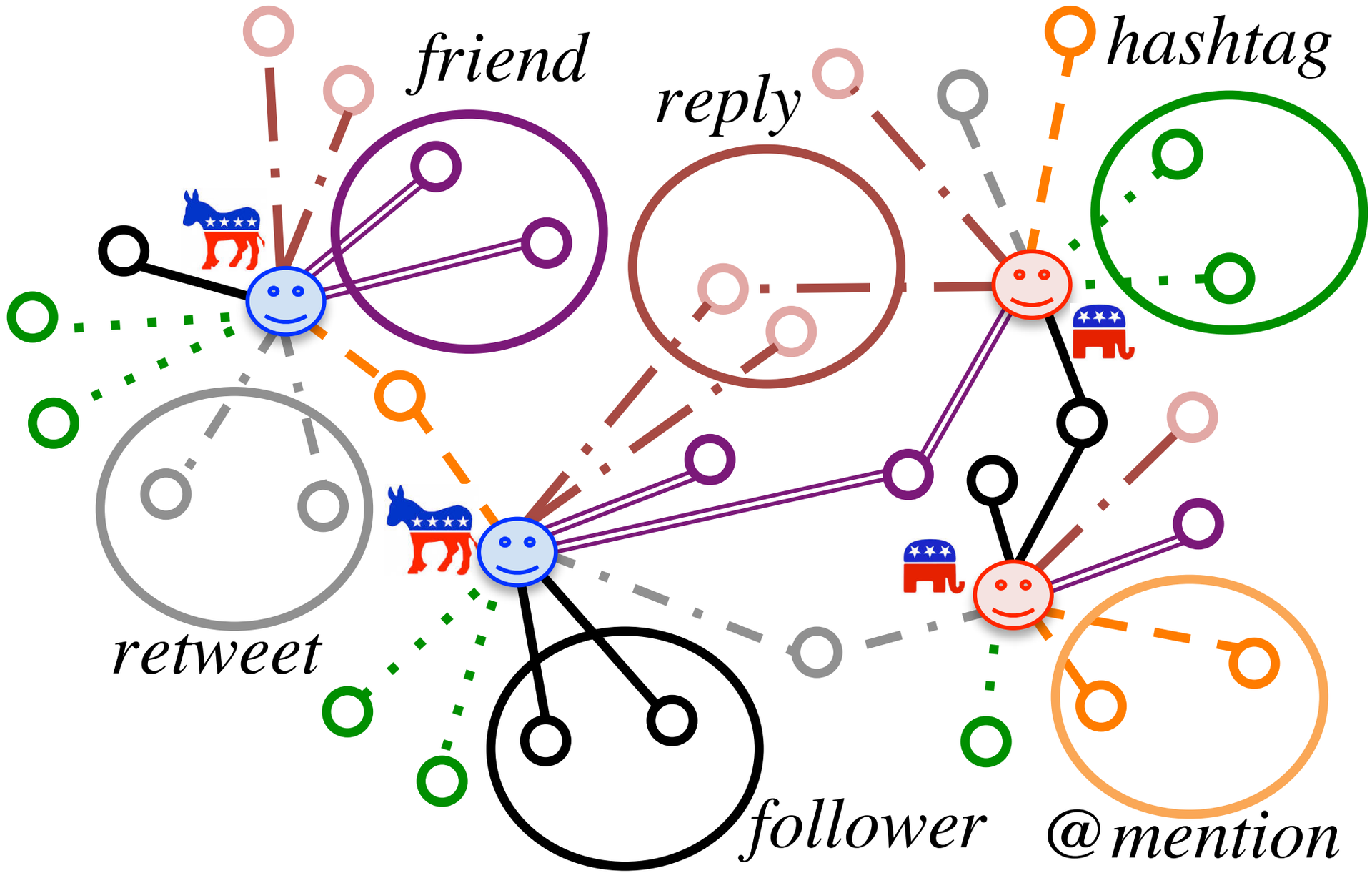
45 79

hashtags are powerful

Cashtags


Twitter
@twitter
Follow
Now you can click on ticker symbols like $GE on twitter.com to see search results about stocks and companies
8:34 PM - 30 Jul 2012
1,167
295

Twitter's Social Graph



Source: Volkova, Van Durme, Yarowsky, Bachrach
"Tutorial on Social Media Predictive Analytics" NAACL 2015

Twitter API

What is an API?

Application **P**rogramming **I**nterface

API is a set of protocols that specify how software programs communicate with each other.

What is an API?

> **Without API:**
> An app finds the current weather in London by opening http://www.weather.com/ and reading the webpage like a human does, interpreting the content.

> **With API:**
> An app finds the current weather in London by sending a message to the weather.com API (in a structured format like XML). The weather.com API then replies with a structured response.

Source: Chris Beach @ Quora

Twitter API

- Twitter is recognized for having one of the most open and powerful developer APIs of any major technology company.

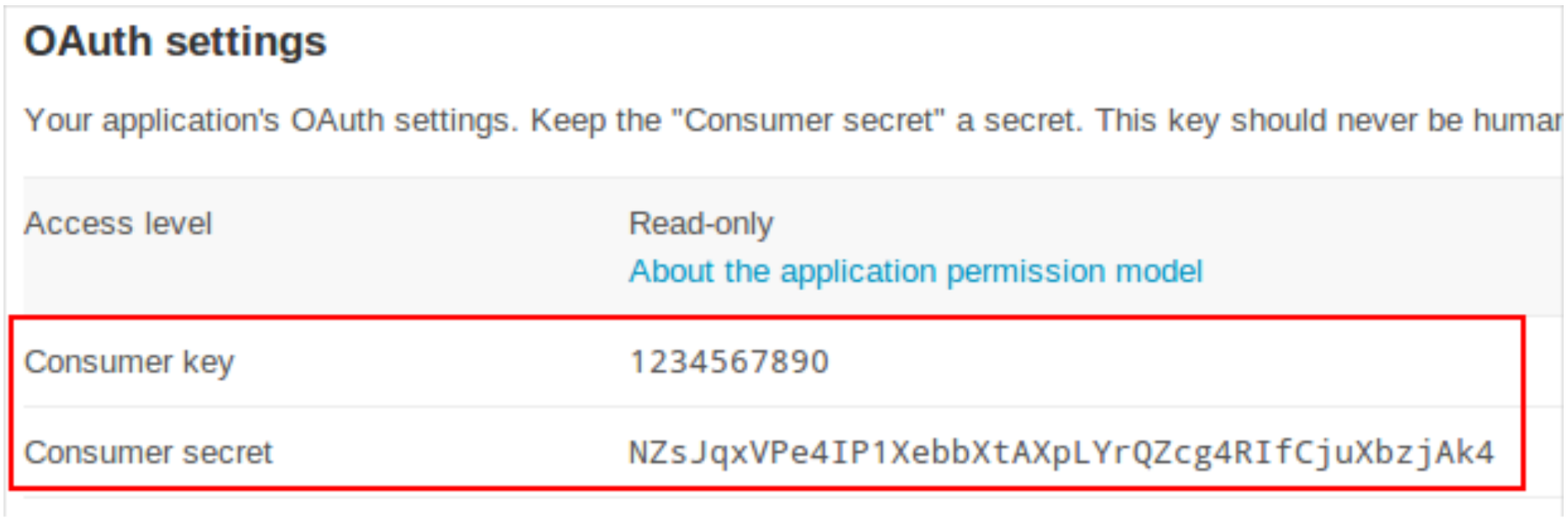
- The first version of its public API was released in September 2006.

Two Most Popular APIs

Streaming API	**REST API**
a sample of public tweets and events as they published on Twitter (can specify search terms or users)	- search - trends - read author profile and follower data - post / modify
only real-time data	historical data up to a week
continuous net connection	one-time request
no limit	rate limit (varies for different requests)

OAuth

- Twitter uses OAuth to provide authorized access to its API.
- which means, to start with needs:
  - a Twitter account
  - OAuth access tokens from apps.twitter.com



Python Twitter Tools

Streaming API

```
# Import the necessary package to process data in JSON format
try:
    import json
except ImportError:
    import simplejson as json

# Import the necessary methods from "twitter" library
from twitter import Twitter, OAuth, TwitterHTTPError, TwitterStream

# Variables that contains the user credentials to access Twitter API
ACCESS_TOKEN = 'YOUR ACCESS TOKEN"'
ACCESS_SECRET = 'YOUR ACCESS TOKEN SECRET'
CONSUMER_KEY = 'YOUR API KEY'
CONSUMER_SECRET = 'ENTER YOUR API SECRET'

oauth = OAuth(ACCESS_TOKEN, ACCESS_SECRET, CONSUMER_KEY, CONSUMER_SECRET)

# Initiate the connection to Twitter Streaming API
twitter_stream = TwitterStream(auth=oauth)

# Get a sample of the public data following through Twitter
iterator = twitter_stream.statuses.sample()
```

OAuth →

connection →

JSON

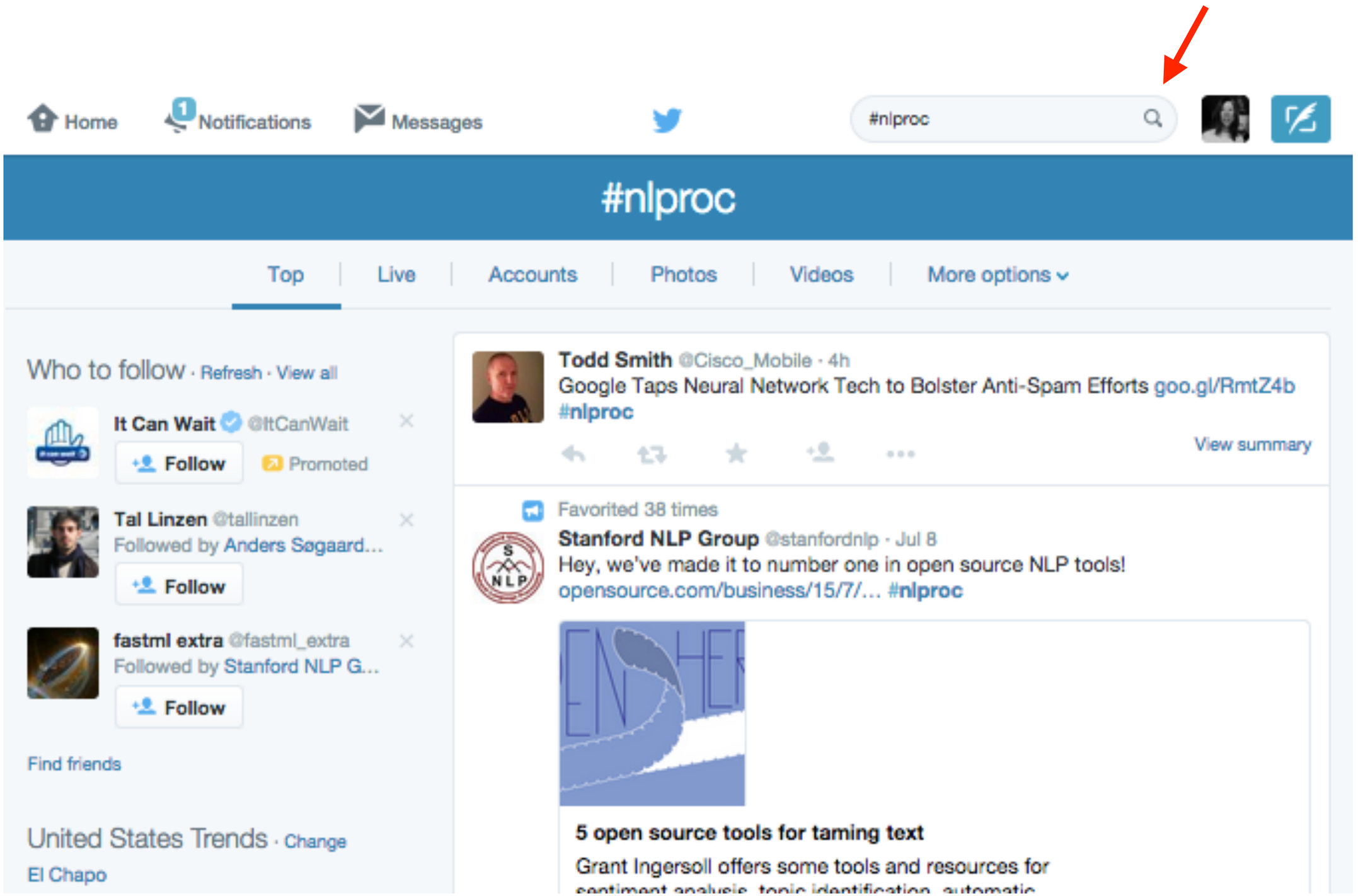
JavaScript Object Notation

JSON is a minimal, readable format for structuring data.

A Tweet in JSON

```
{
   "favorited": false,
   "contributors": null,
   "truncated": false,
   "text": "#CFP Workshop on Noisy User-generated Text at ACL - Beijing 31 July 2015. Papers due: 11 May 2015. http://t.co/rcygyEowqH   #NLProc #WNUT15",
   "possibly_sensitive": false,
   "in_reply_to_status_id": null,
   "user": {
      "follow_request_sent": null,
      "profile_use_background_image": true,
      "default_profile_image": false,
      "id": 237918251,
      "verified": false,
      "profile_image_url_https": "https://pbs.twimg.com/profile_images/527088456967544832/Dn
```

Search


Home
Notifications
Messages
#nlproc
#nlproc
Top
Live
Accounts
Photos
Videos
More options
Who to follow · Refresh · View all
It Can Wait @ItCanWait
Follow
Promoted
Tal Linzen @tallinzen
Followed by Anders Søgaard...
Follow
fastml extra @fastml_extra
Followed by Stanford NLP G...
Follow
Find friends
United States Trends · Change
El Chapo
Todd Smith @Cisco_Mobile · 4h
Google Taps Neural Network Tech to Bolster Anti-Spam Efforts goo.gl/RmtZ4b #nlproc
View summary
Favorited 38 times
Stanford NLP Group @stanfordnlp · Jul 8
Hey, we've made it to number one in open source NLP tools!
opensource.com/business/15/7/... #nlproc
5 open source tools for taming text
Grant Ingersoll offers some tools and resources for

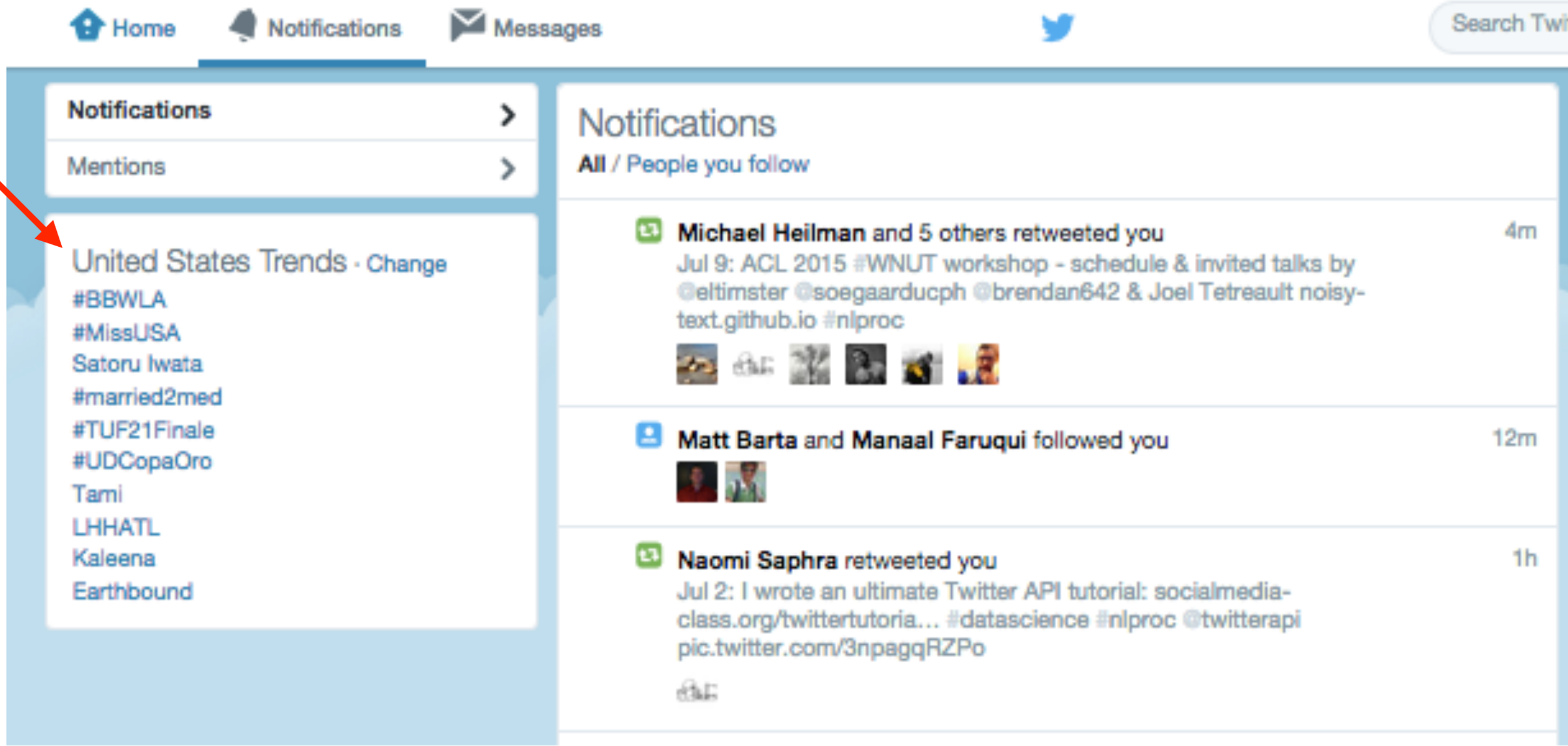
Search API

```
# Initiate the connection to Twitter REST API
twitter = Twitter(auth=oauth)
```

```
# Search for latest tweets about "#nlproc"
twitter.search.tweets(q='#nlproc')
```

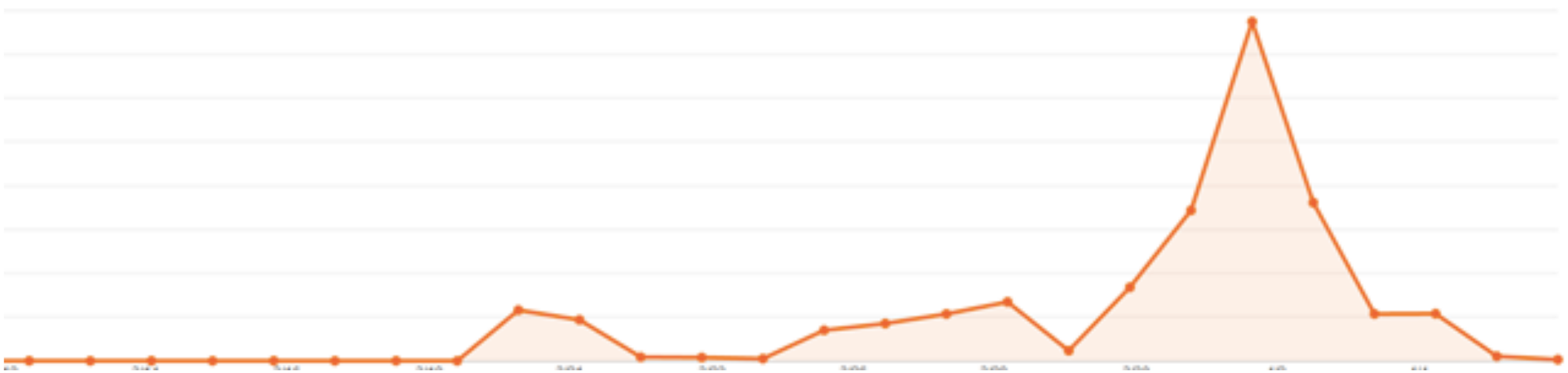
```
twitter.search.tweets(q='#nlproc', result_type='recent', lang='en', count=10)
```

Trends


Home
Notifications
Messages
Search Twi
Notifications
Mentions
United States Trends · Change
#BBWLA
#MissUSA
Satoru Iwata
#married2med
#TUF21Finale
#UDCopaOro
Tami
LHHATL
Kaleena
Earthbound
Notifications
All / People you follow
Michael Heilman and 5 others retweeted you
4m
Jul 9: ACL 2015 #WNUT workshop - schedule & invited talks by @eltimster @soegaarducph @brendan642 & Joel Tetreault noisy-text.github.io #nlproc
Matt Barta and Manaal Faruqui followed you
12m
Naomi Saphra retweeted you
1h
Jul 2: I wrote an ultimate Twitter API tutorial: socialmedia-class.org/twittertutoria… #datascience #nlproc @twitterapi pic.twitter.com/3npagqRZPo

Trends

trending topics are determined by an unpublished algorithm, which finds words, phrases and hashtags that have had a sharp increase in popularity, as opposed to overall volume.



Trends API

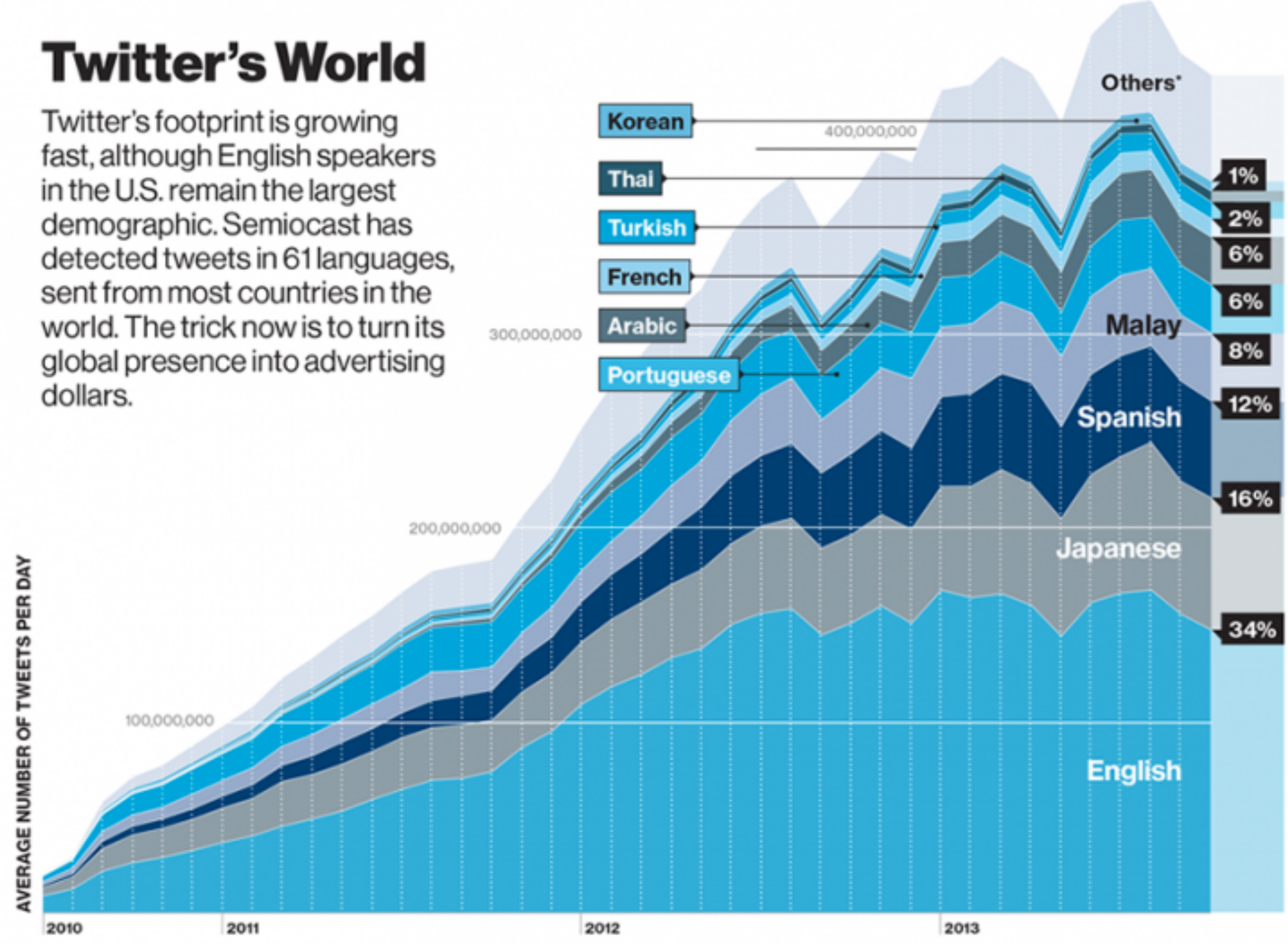
Where On Earth ID

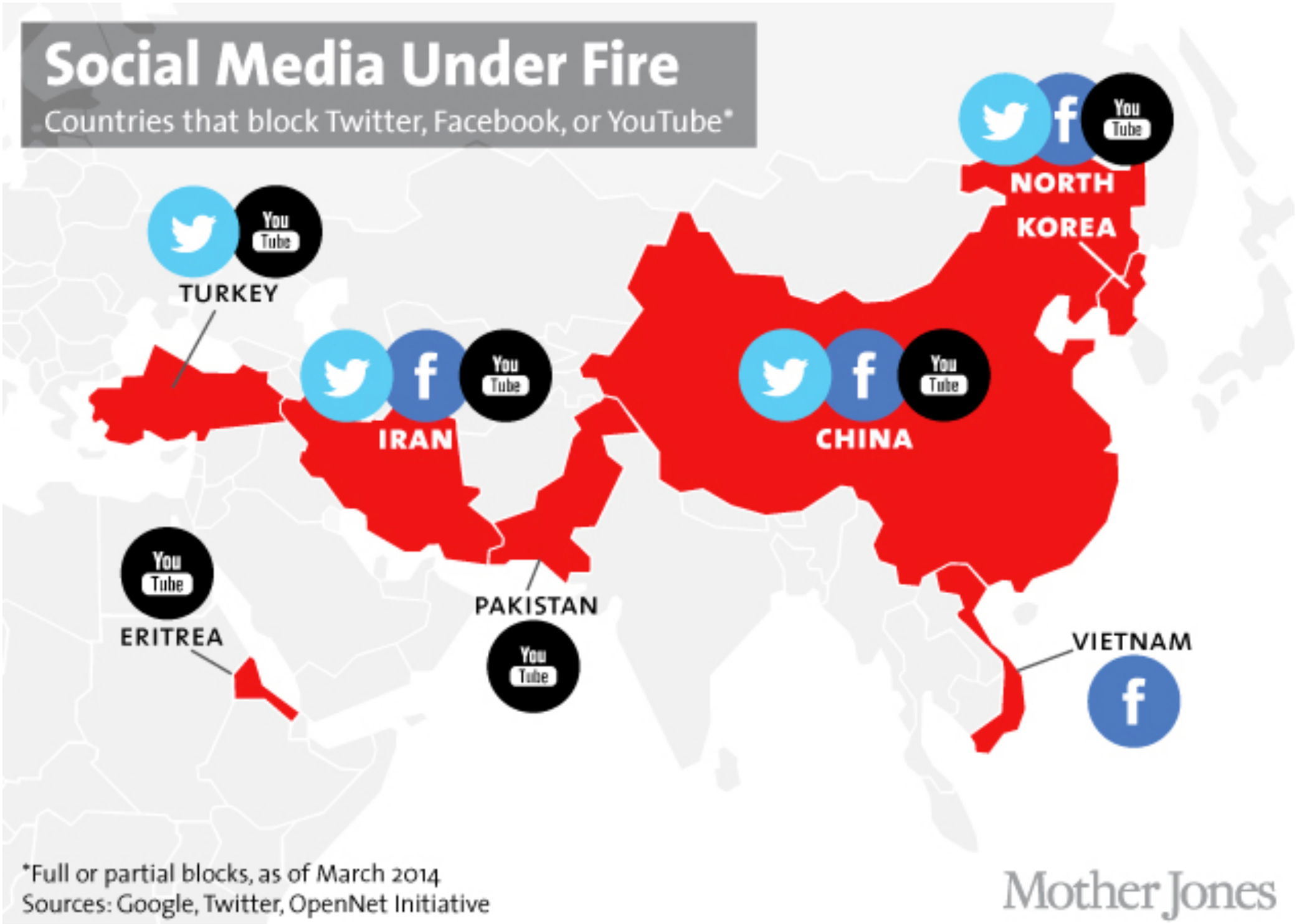
```
# Get all (it's always 10) trending topics in San Francisco (its WOEID is 2487956)
sfo_trends = twitter.trends.place(_id = 2487956)
```


```
{
   "created_at":"2015-07-01T22:09:55Z",
   "trends":[
      {
         "url":"http://twitter.com/search?q=%23LiesIveToldMyParents",
         "query":"%23LiesIveToldMyParents",
         "name":"#LiesIveToldMyParents",
         "promoted_content":null
      },
      {
         "url":"http://twitter.com/search?q=%22Kevin+Love%22",
         "query":"%22Kevin+Love%22",
         "name":"Kevin Love",
         "promoted_content":null
      },

      ... [and another 8 trends omitted here to save space]
```

Twitter's World

Twitter's footprint is growing fast, although English speakers in the U.S. remain the largest demographic. Semiocast has detected tweets in 61 languages, sent from most countries in the world. The trick now is to turn its global presence into advertising dollars.





新浪微博
weibo.com

known as the “Chinese Twitter”
120 Million Posts / Day

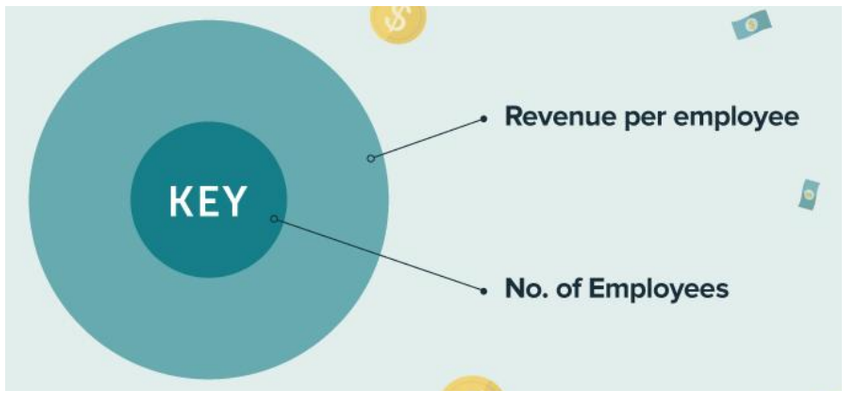
Twitter Demographics

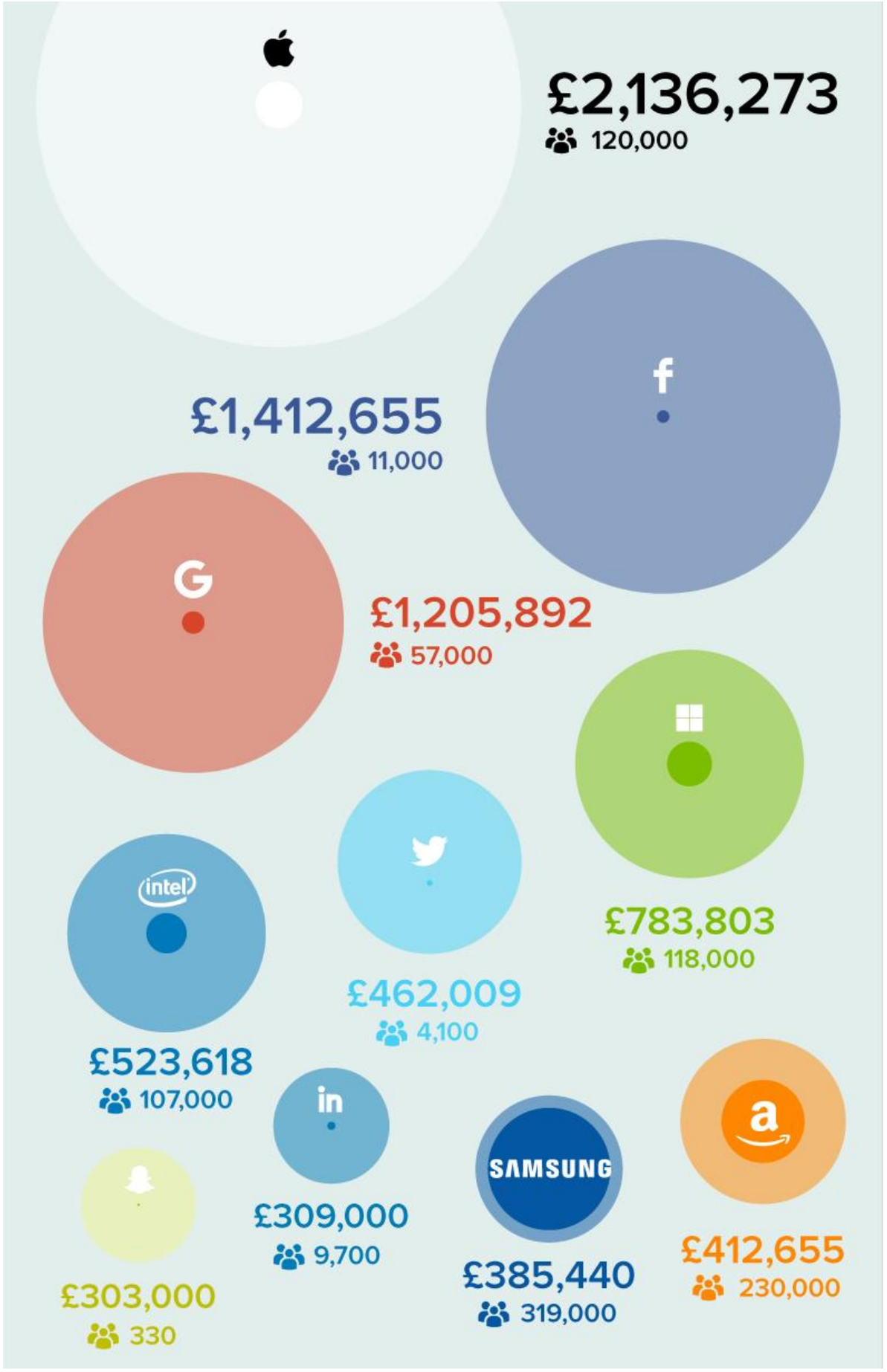
- 24% of All Internet male users use Twitter, whereas 21% of All Internet Female users use Twitter.
- 79% of Twitter accounts are based outside the United States
- There are over 67 million Twitter users in US.
- Total number of Twitter users in UK is 13 million.
- 37% of Twitter users are between ages of 18 and 29, 25% users are 30-49 years old.
- 54% of Twitter users earn more than $50,000 a year at least.
- The top three countries by user count outside the U.S. are Brazil (27.7 million users), Japan (25.9 million), and Mexico (23.5 million).

Fun Facts about Twitter

- More than 100 million tweets contained GIFs in 2015.
- Saudi Arabia has the highest percent of internet users who are active on Twitter.
- Number of Twitter timeline views in 2014 is 200 billion.
- 83% of 193 UN member countries have Twitter presence.
- Twitter's revenue per employee is $488,913.

RPE





Source: http://www.ecardshack.com/blog/top-tech-companies-revenue-per-employee

No Class next week

emnlp2017.net

EMNLP 2017: Conference on Empirical Methods in Natural Language Processing — September 7–11, 2017 — Copenhagen, Denmark.

SIGDAT, the Association for Computational Linguistics special interest group on linguistic data and corpus-based approaches to NLP, invites you to participate in EMNLP 2017.

Conference Program Copenhagen Registration Anti-harassment policy

No Class next week

noisy-text.github.io/2017/

W-NUT | Home | 2017 | 2016 | 2015 | Call | Shared-tasks | Committee

2017 The 3rd Workshop on Noisy User-generated Text (W-NUT)

September 7th, Copenhagen (at EMNLP 2017)

The WNUT workshop focuses on Natural Language Processing applied to noisy user-generated text, such as that found in social media, online reviews, crowdsourced data, web forums, clinical records and language learner essays. This year, there will be one shared task on Entity Recognition - details below.

The workshop hashtag is #wnut.

Workshop Organizers

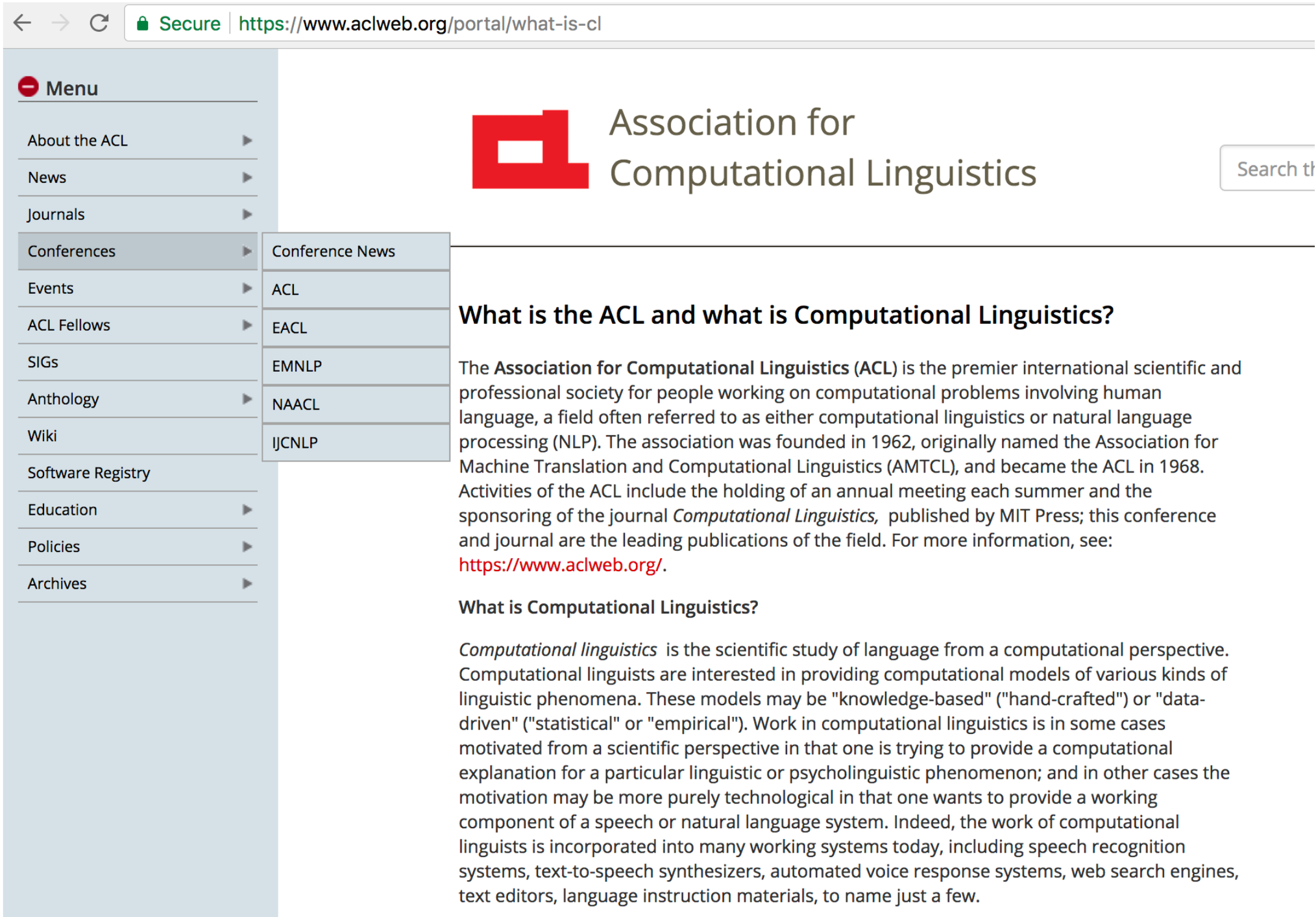
- Leon Derczynski (The University of Sheffield)
- Wei Xu (The Ohio State University)
- Alan Ritter (The Ohio State University)
- Tim Baldwin (The University of Melbourne)

Natural Language Processing 101

a.k.a.

- Natural Language Processing (NLP)
- Text Analysis
- Computational Linguistics

ACL


Secure
https://www.aclweb.org/portal/what-is-cl
Menu
About the ACL
News
Journals
Conferences
Events
ACL Fellows
SIGs
Anthology
Wiki
Software Registry
Education
Policies
Archives
Conference News
ACL
EACL
EMNLP
NAACL
IJCNLP
Association for Computational Linguistics
Search th
What is the ACL and what is Computational Linguistics?
The Association for Computational Linguistics (ACL) is the premier international scientific and professional society for people working on computational problems involving human language, a field often referred to as either computational linguistics or natural language processing (NLP). The association was founded in 1962, originally named the Association for Machine Translation and Computational Linguistics (AMTCL), and became the ACL in 1968. Activities of the ACL include the holding of an annual meeting each summer and the sponsoring of the journal Computational Linguistics, published by MIT Press; this conference and journal are the leading publications of the field. For more information, see: https://www.aclweb.org/.
What is Computational Linguistics?
Computational linguistics is the scientific study of language from a computational perspective. Computational linguists are interested in providing computational models of various kinds of linguistic phenomena. These models may be "knowledge-based" ("hand-crafted") or "data-driven" ("statistical" or "empirical"). Work in computational linguistics is in some cases motivated from a scientific perspective in that one is trying to provide a computational explanation for a particular linguistic or psycholinguistic phenomenon; and in other cases the motivation may be more purely technological in that one wants to provide a working component of a speech or natural language system. Indeed, the work of computational linguists is incorporated into many working systems today, including speech recognition systems, text-to-speech synthesizers, automated voice response systems, web search engines, text editors, language instruction materials, to name just a few.

NLP Publications

- top NLP-specific venues:
  - ACL, NAACL, EACL, EMNLP, COLING (conference)
  - TACL (journal+conference model)
  - CL (journal)

- other venues:
  - NLP: CoNLL, *Sem, WMT, LREC, IJNLP, Workshops …
  - related CS fields: WWW, KDD, AAAI, WSDM, NIPS, ICWSM, CIKM, ICML …
  - related non-CS fields: psychology, linguistics, …

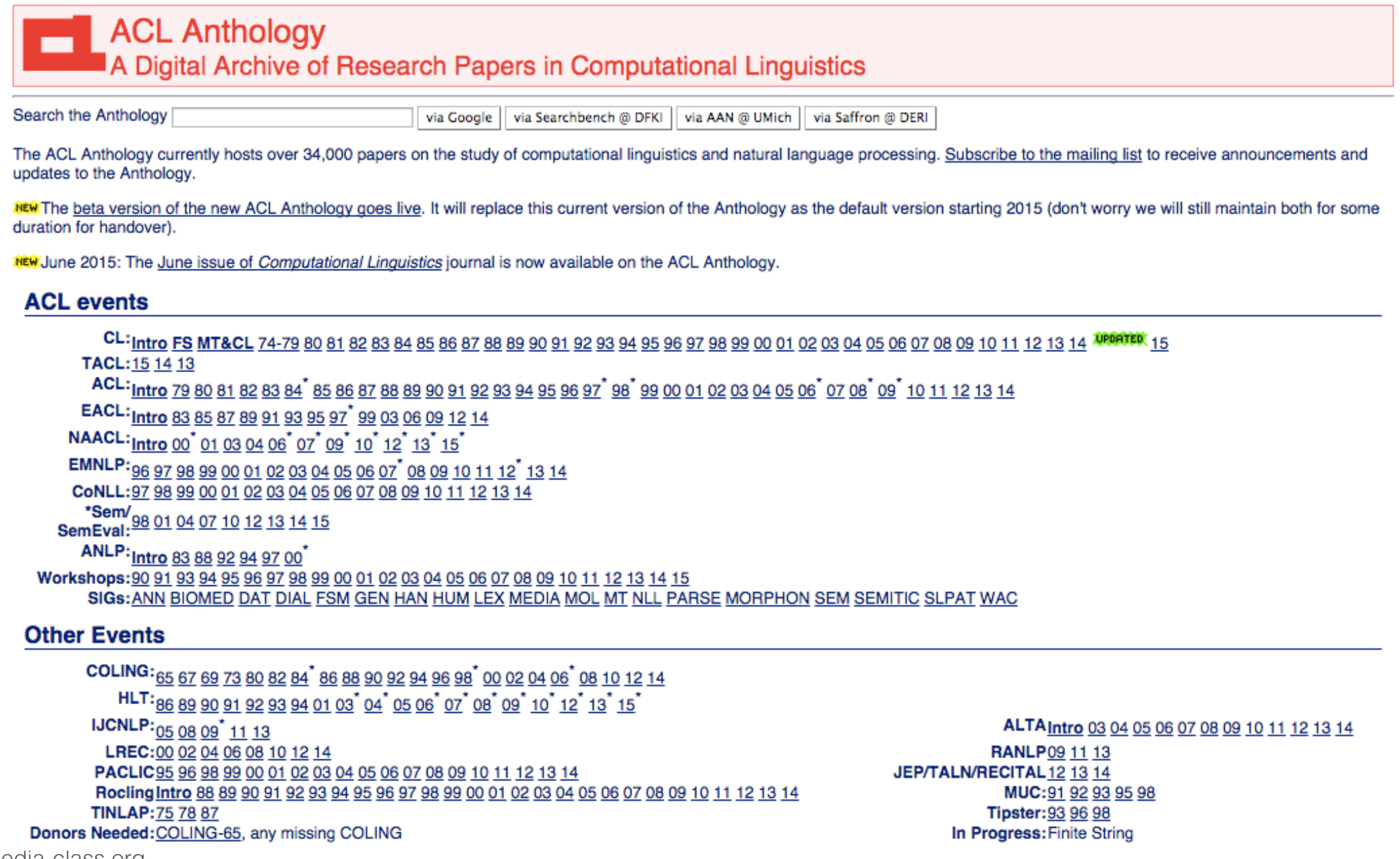
Conference Rotation

- ACL (and/or NAACL, EACL), EMNLP / COLING

NLP Publications

- ACL Anthology (http://aclweb.org/anthology/)

  all NLP conference and journal papers (free!)



ACL Anthology
A Digital Archive of Research Papers in Computational Linguistics

Search the Anthology [ ] via Google | via Searchbench @ DFKI | via AAN @ UMich | via Saffron @ DERI

The ACL Anthology currently hosts over 34,000 papers on the study of computational linguistics and natural language processing. Subscribe to the mailing list to receive announcements and updates to the Anthology.

NEW The beta version of the new ACL Anthology goes live. It will replace this current version of the Anthology as the default version starting 2015 (don't worry we will still maintain both for some duration for handover).

NEW June 2015: The June issue of *Computational Linguistics* journal is now available on the ACL Anthology.

ACL events

CL: Intro FS MT&CL 74-79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 00 01 02 03 04 05 06 07 08 09 10 11 12 13 14 UPDATED 15
TACL: 15 14 13
ACL: Intro 79 80 81 82 83 84* 85 86 87 88 89 90 91 92 93 94 95 96 97* 98* 99 00 01 02 03 04 05 06* 07 08* 09* 10 11 12 13 14
EACL: Intro 83 85 87 89 91 93 95 97* 99 03 06 09 12 14
NAACL: Intro 00* 01 03 04 06* 07* 09* 10* 12* 13* 15*
EMNLP: 96 97 98 99 00 01 02 03 04 05 06 07* 08 09 10 11 12* 13 14
CoNLL: 97 98 99 00 01 02 03 04 05 06 07 08 09 10 11 12 13 14
***Sem/SemEval:** 98 01 04 07 10 12 13 14 15
ANLP: Intro 83 88 92 94 97 00*
Workshops: 90 91 93 94 95 96 97 98 99 00 01 02 03 04 05 06 07 08 09 10 11 12 13 14 15
SIGs: ANN BIOMED DAT DIAL FSM GEN HAN HUM LEX MEDIA MOL MT NLL PARSE MORPHON SEM SEMITIC SLPAT WAC

Other Events

COLING: 65 67 69 73 80 82 84* 86 88 90 92 94 96 98* 00 02 04 06* 08 10 12 14
HLT: 86 89 90 91 92 93 94 01 03* 04* 05 06* 07* 08* 09* 10* 12* 13* 15*
IJCNLP: 05 08 09* 11 13
LREC: 00 02 04 06 08 10 12 14
PACLIC 95 96 98 99 00 01 02 03 04 05 06 07 08 09 10 11 12 13 14
Rocling Intro 88 89 90 91 92 93 94 95 96 97 98 99 00 01 02 03 04 05 06 07 08 09 10 11 12 13 14
TINLAP: 75 78 87
Donors Needed: COLING-65, any missing COLING

ALTA Intro 03 04 05 06 07 08 09 10 11 12 13 14
RANLP 09 11 13
JEP/TALN/RECITAL 12 13 14
MUC: 91 92 93 95 98
Tipster: 93 96 98
In Progress: Finite String

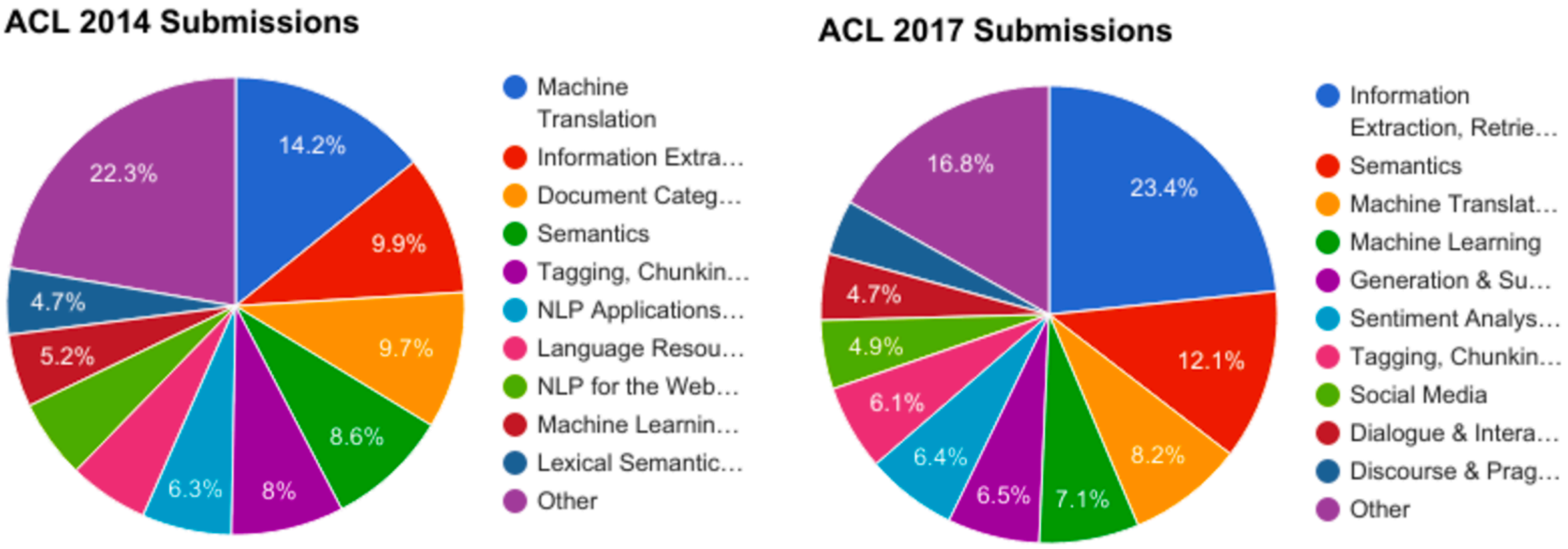
ACL'14 at A Glance

- The Annual Meeting of the Association for Computational Linguistics
- Duration:
  - tutorials (1 day)
  - main conference (3 days)
  - workshops (2 days)
- Attendance of 1300+ people
- Papers:
  - 1,123 submissions
  - 146 long papers and 129 short papers accepted
  - + 19 TACL papers
  - 159 oral and 145 poster presentations

ACL'17 at A Glance

- The Annual Meeting of the Association for Computational Linguistics
- Duration:
  - tutorials (1 day)
  - main conference (3 days)
  - workshops (2 days)
- Attendance of 1800+ people
- Papers:
  - 1,318 submissions
  - 195 long papers and 107 short papers accepted
  - + 21 TACL papers
  - 151 oral and 151 poster presentations

ACL'14 vs. ACL'17



Some shifts: e.g. summarization and generation is now in top 5 areas, while in 2014 it didn't even make top 10

read more: https://chairs-blog.acl2017.org/

Research Areas

	2017 actual total submissions (after cleanup)	2017 actual # long submitted (after cleanup)	2017 actual # short submitted (after cleanup)
Information Extraction, Retrieval, Question Answering, Document Analysis and NLP Applications	308	192	116
Semantics	159	100	59
Machine Translation	108	60	48
Machine Learning	93	55	38
Generation and Summarization	86	52	34
Sentiment Analysis and Opinion Mining	85	54	31
Tagging, Chunking, Syntax and Parsing	80	40	40
Social Media	64	25	39
Dialogue and Interactive Systems	62	36	26
Discourse and Pragmatics	51	26	25
Phonology, Morphology and Word Segmentation	43	23	20
Resources and Evaluation	39	17	22
Multidisciplinary	32	12	20
Vision, Robotics and Grounding	31	19	12
Cognitive Modeling and Psycholinguistics	28	14	14
Multilinguality	28	15	13
Biomedical	12	6	6
Speech	9	5	4
Total	1318	751	567

How to Do Research

William Wang

UCSB Computer Science
10/06/2016

What is research?

- Investigate and understand the known unknowns and unknown unknowns in the scientific world.

- In our lab, we are specifically interested in:
  - designing accurate, robust, and scalable **machine learning** algorithms;
  - advancing **natural language processing** models;
  - combining **learning and reasoning** for better AI.

How's research different from taking courses?

- Taking courses: instructor tells you **exactly** what to do.

- Research:
  - define an **open** research problem with your advisor;
  - you **(students) take the initiatives**;
  - discuss and refine the technical approaches;
  - you (students) implement the approach and perform experiments to verify the idea.

How to make good progress in research activities

- **Clearly define the problem / task** that you want to solve;
- Understand the **literature**: what other people have done, and what you can learn from them;
- Work out the algorithm first, find a suitable dataset, and put theories into practice: **write some code**;
- Start with **smaller subset of data** for debugging, and move on to larger datasets.
- **Document the results** carefully in spreadsheet / docs.

How to measure the effectiveness of ideas?

- Use **mathematical** tools to clearly define the problem and your solutions;

- Look at the theoretical properties of your **algorithms**;

- Define good **metric**(s), and perform experiments on **multiple datasets**;

- Report results and compare with state-of-the-arts **baselines**.

Why is publication important?

- Publication is the most important formal method for **scholarly communications**.

- Presenting your research and attending leading conferences will create **impacts**, get **inspirations**, and facilitate the **exchange of thoughts** and good ideas.

- Peer-review is a good way to get **feedback** from top researchers in your field.

- And it is a *relatively* **objective** way to claim the effectiveness of your research.

What is in a good research (paper)?

- Is the problem **new**?
- Is your approach **new**?
- How good are the results **comparing to prior work**?
- Can you contribute any new **open-source** datasets/code?
- Is this paper well-structured and **well-written**?

Research is hard

- They are open problems that no one has a perfect solution!
- Implementing ideas and debugging code could be challenging.
- Performing good experiments are not easy.
- Writing papers against deadlines..

Research is rewarding

- You helped to advance science!
- When your first top conference full paper is accepted… (acceptance rates typically 10-30%);
- Other people attend your talk, read/cite your papers, and use your code/approaches;
- You are now the world's expert in this area.