

Social Media & Text Analysis

lecture 2 - Twitter API



CSE 5539-0010 Ohio State University
Instructor: Wei Xu
Website: socialmedia-class.org

Course Website

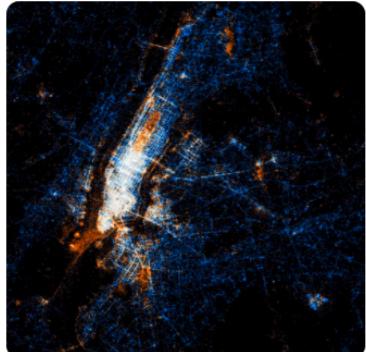
socialmedia-class.org

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 2017, CSE 5539-0010](#) The Ohio State University

[Bolz Hall Room 318 | Tuesday 2:20PM – 4:10PM](#)

dual-listed undergraduate and graduate course

[Office Hour] Dreese 495 | Tuesday 4:15PM – 5:15PM

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](#)

Discussion Board

[Piazza](#) (TBA)

Have a Question?

- **Ask in class!**
- **Office Hour:** Tue 4:30 pm — 5:30 pm, Dreese 495
- **Piazza Q&A Board**

The screenshot shows the Piazza Q&A platform interface for a course named "CSE 5539-0010". The top navigation bar includes links for "Q & A", "Resources", "Statistics", and "Manage Class". The user profile "Wei Xu" is visible on the right.

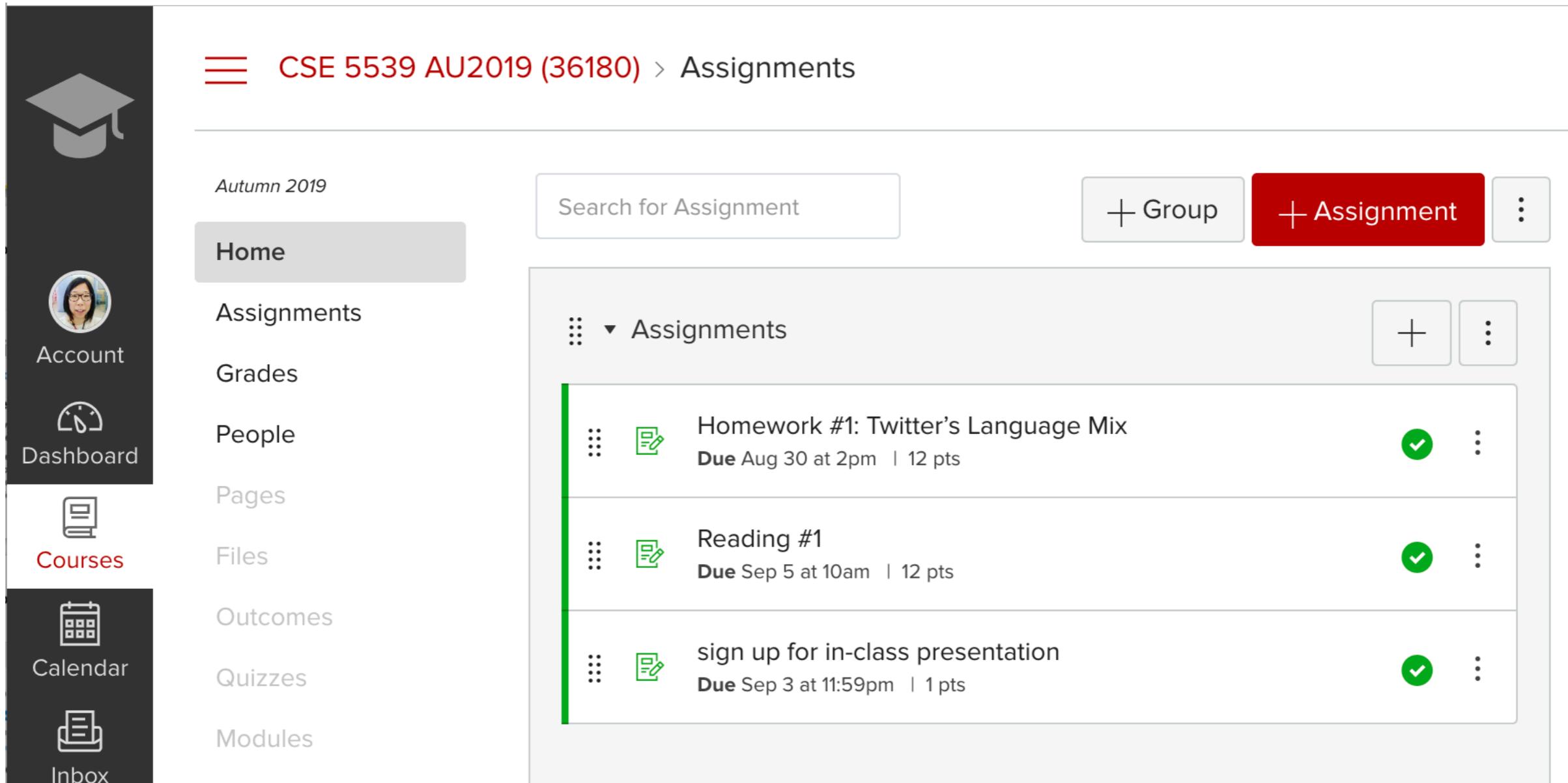
The main area displays a "Question History" section. A specific question titled "Assignment 1" is highlighted, asking how to get at least 10000 tweets if there is a rate limit smaller than 10000. This question has 30 views and was posted by "Karthik Pillalamarri" 12 hours ago. It is tagged with "hw1".

Below the question, a student's answer is shown, stating: "I think you can try some other Streaming API, like user_timeline, which returns you the tweets of a specific user. Or using StreamListener to get some tweets from specific location, key words, or something else." The student also provides a link to a page on the course website for writing a StreamListener.

Other visible posts include a pinned post about searching for teammates and a post about submitting HW 1 as a .ipynb file. There is also a post about presenting a dataset rather than a .py file.

Homework #1 is out

Due this Friday (Aug 30)



The screenshot shows a learning management system interface for the course CSE 5539 AU2019 (36180). The left sidebar has a dark background with icons for Account, Dashboard, Courses, Calendar, and Inbox. The main area shows the course title "CSE 5539 AU2019 (36180) > Assignments". A navigation bar includes "Autumn 2019", "Search for Assignment", "+ Group", "+ Assignment", and a more button. The "Assignments" tab is selected. Three assignments are listed:

- Homework #1: Twitter's Language Mix (Due Aug 30 at 2pm | 12 pts)
- Reading #1 (Due Sep 5 at 10am | 12 pts)
- sign up for in-class presentation (Due Sep 3 at 11:59pm | 1 pt)

Each assignment has a green checkmark and a more button.

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 credentials from the 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 account.
- Click "Create New App"

Twitter API Tutorial: socialmedia-class.org

Social Media & Text Analytics

Syllabus

Twitter API Tutorial

Homework▼

High School Outreach



Twitter's 404 error
page -- the Fail
Whale

Twitter API tutorial

by Wei Xu [Follow @cocoweixu](#)

and Jeniya Tabassum [Follow @JeniyaTabassum](#)

(Ohio State University)

Last updated March 20, 2018 (added a script for obtaining all followers of a Twitter user; updated with tweepy p

[[download the Jupyter notebook for this tutorial](#)]

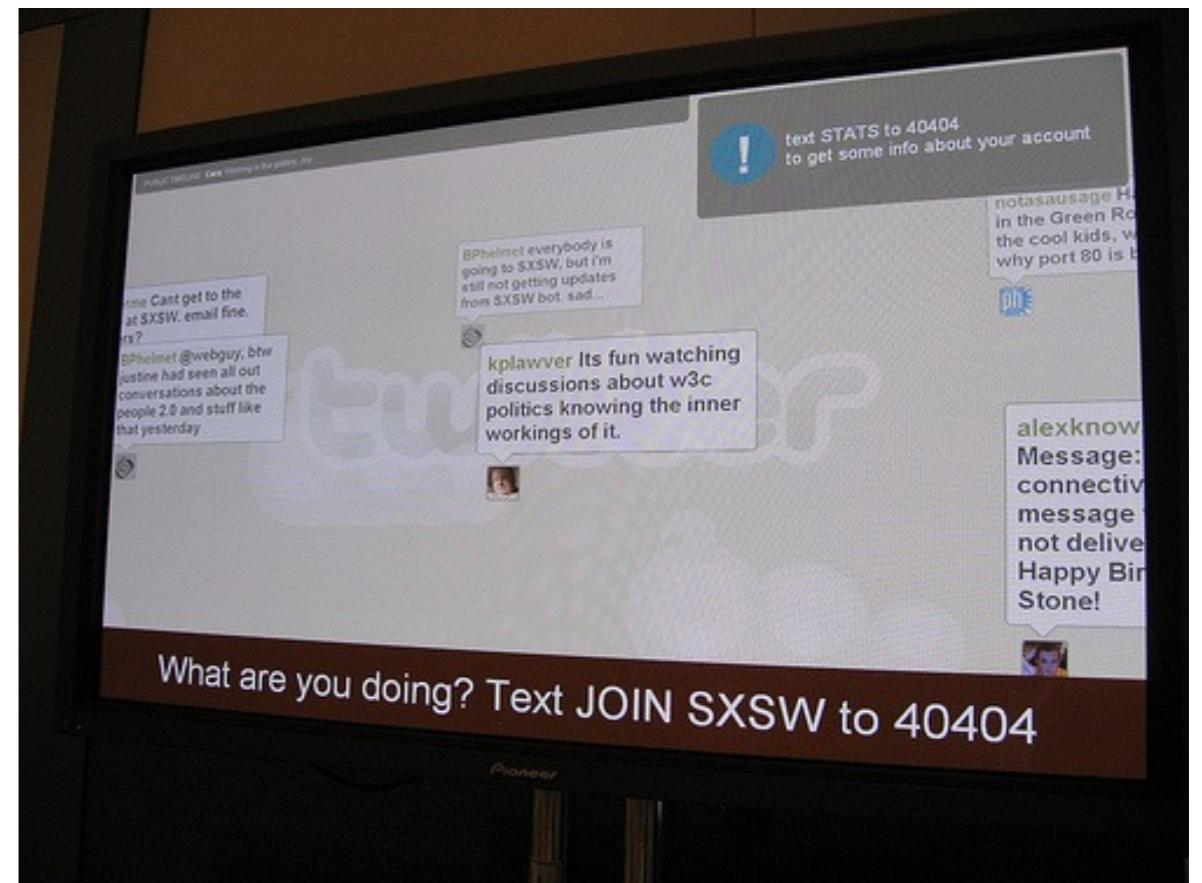
1. Getting Twitter API keys

To start with, you will need to have a Twitter developer account and obtain credentials (i.e. API key, API secret, and Access token secret) on the to access the Twitter API, following these steps:

- Create a Twitter developer account if you do not already have one from : <https://developer.twitter.com/>
- Go to <https://developer.twitter.com/en/apps> and log in with your Twitter user account.
- Click “Create an app”
- Fill out the form, and click “Create”
- A pop up window will appear for reviewing Developer Terms. Click the “Create” button again.
- In the next page, click on “Keys and Access Tokens” tab, and copy your “API key” and “API secret” from [Consumer API keys](#) section.

Twitter History

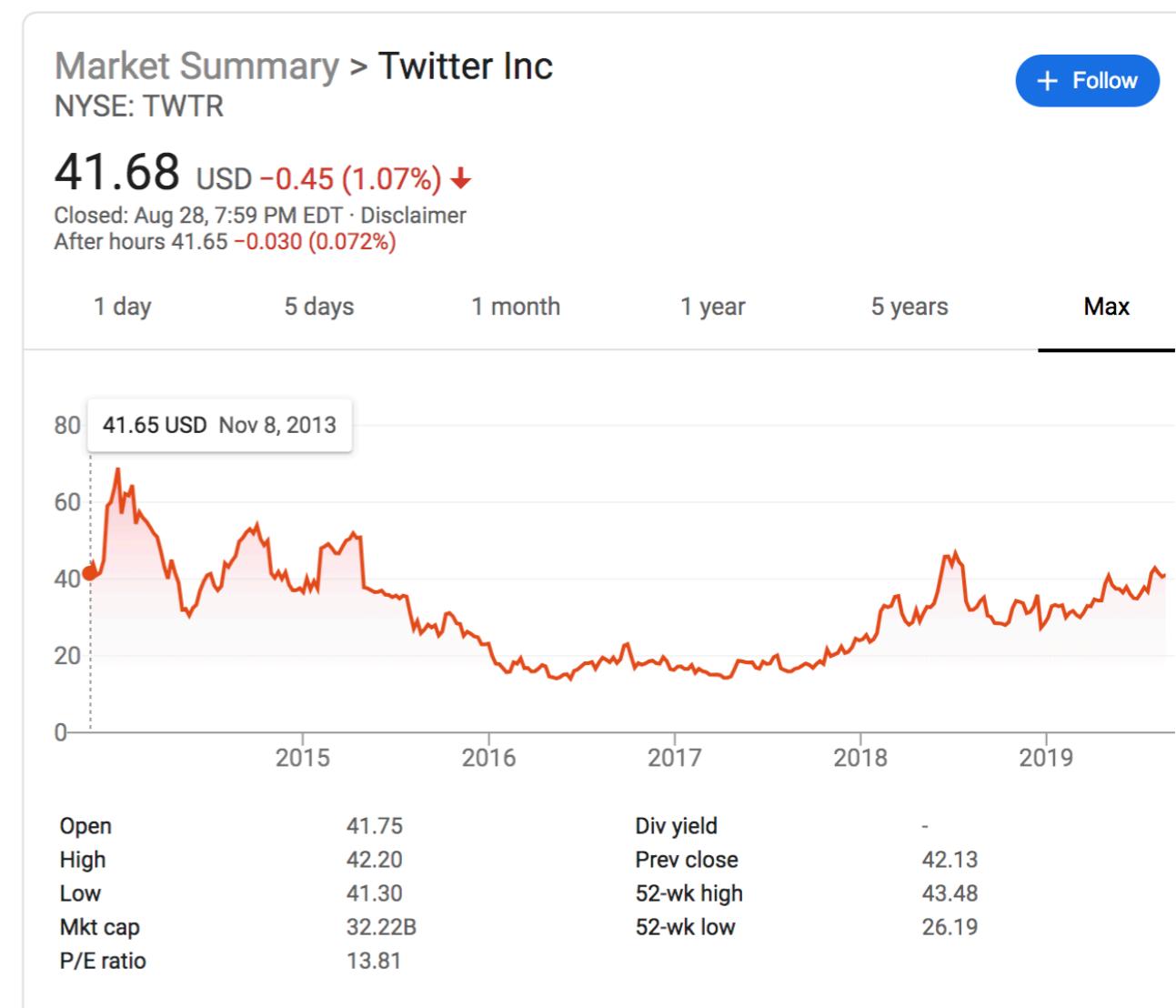
- Jack Dorsey's idea
(a NYU undergraduate then)
- 1st tweet on March 21, 2006
- exploded at SXSW 2007
(20k→60k tweets/day)
- 100m tweets/quarter in 2008,
50m tweets/day in 2010,
400m tweets/day in 2013
- Huge API usage was
unexpected as was the rise of
the @ sign for replies



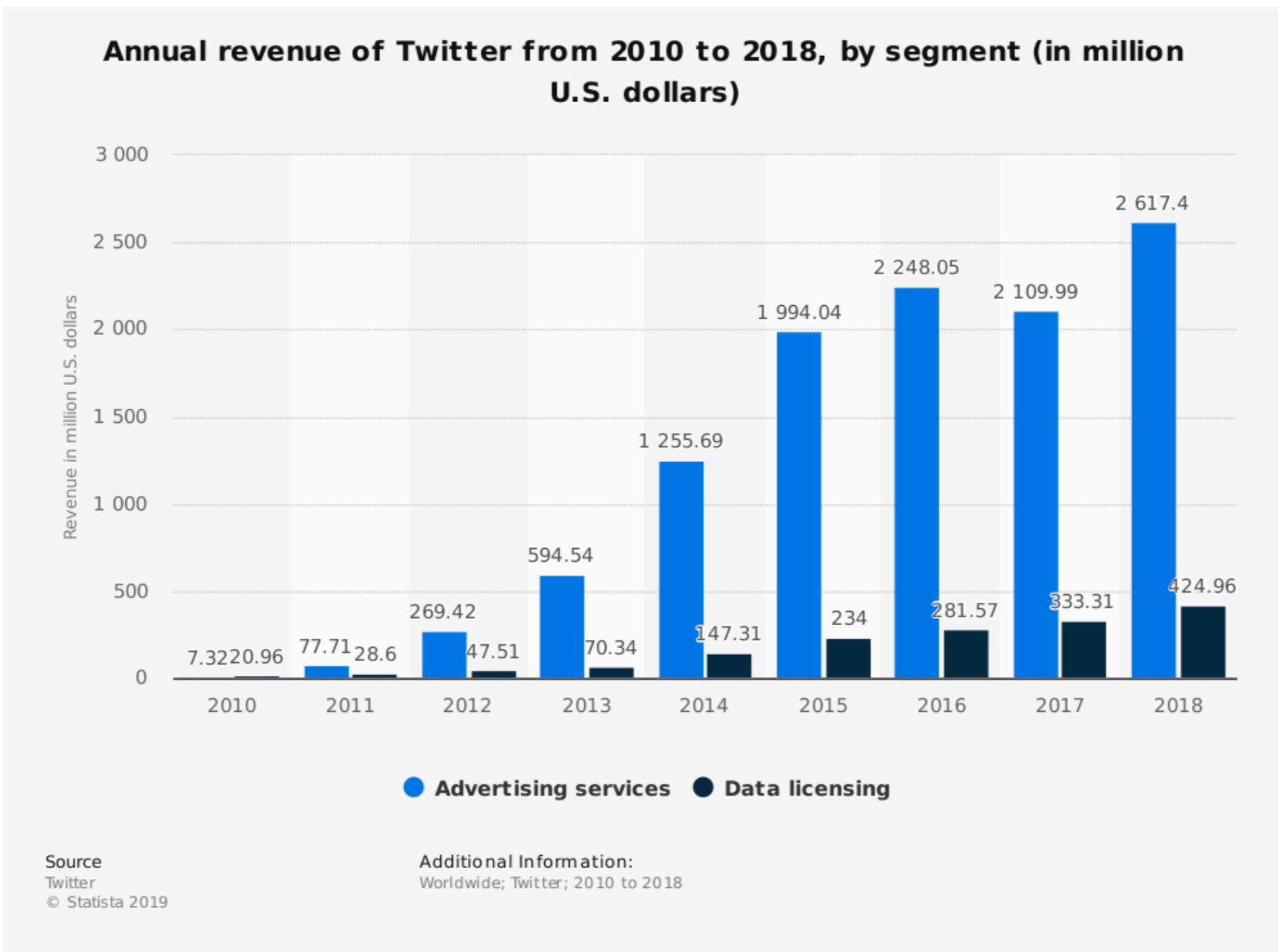
Twitter staff received the festival's Web Award prize with the remark "we'd like to thank you in 140 characters or less. And we just did!"

Twitter History

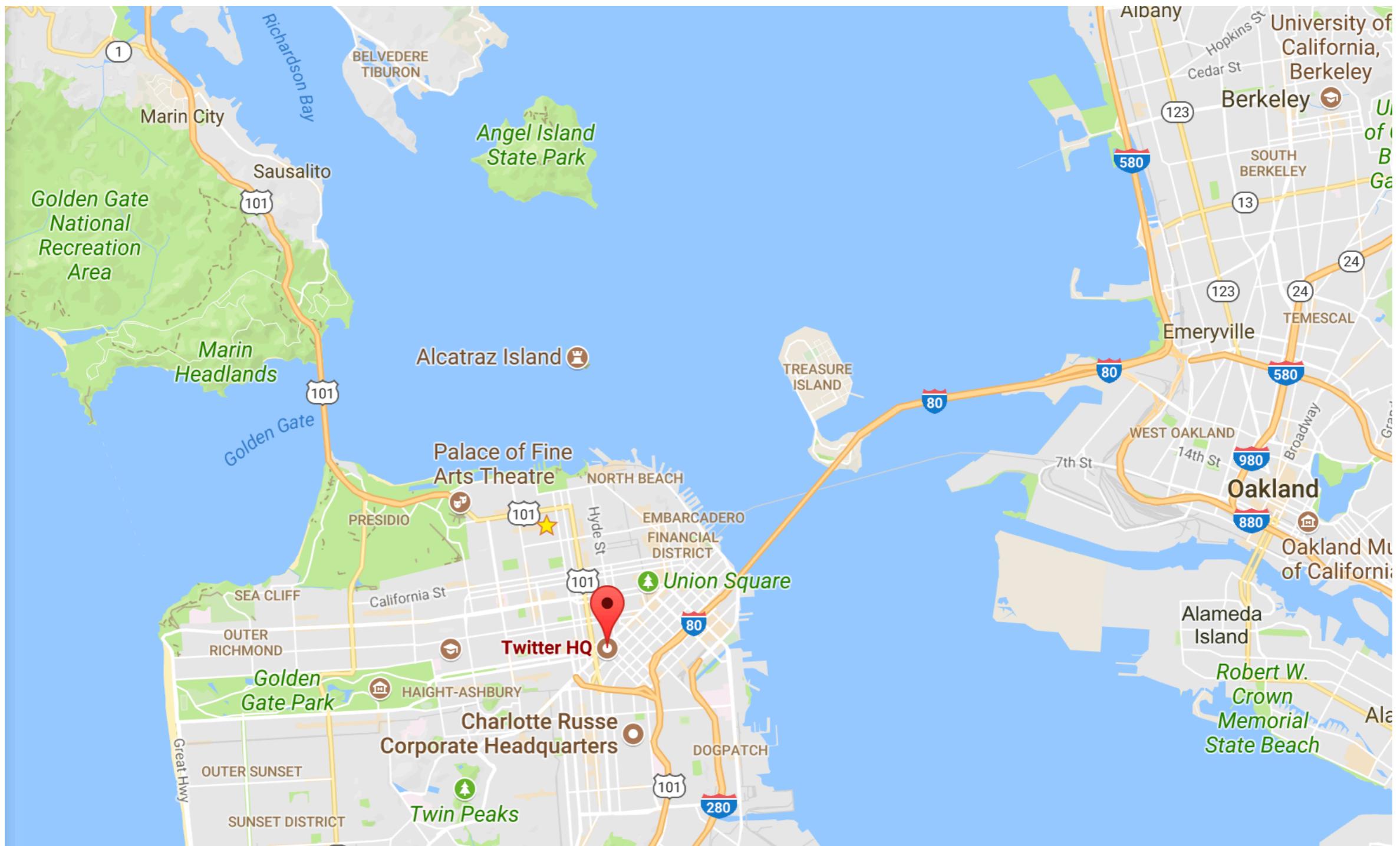
- IPO in 2013 Q4
- market value \$24b, revenue \$435m, net loss \$162m in 2015 Q1
- CEO Dick Costolo resigned July 1st, 2015
- Dorsey was named permanent CEO of Twitter on October 5, 2015



Twitter Revenue

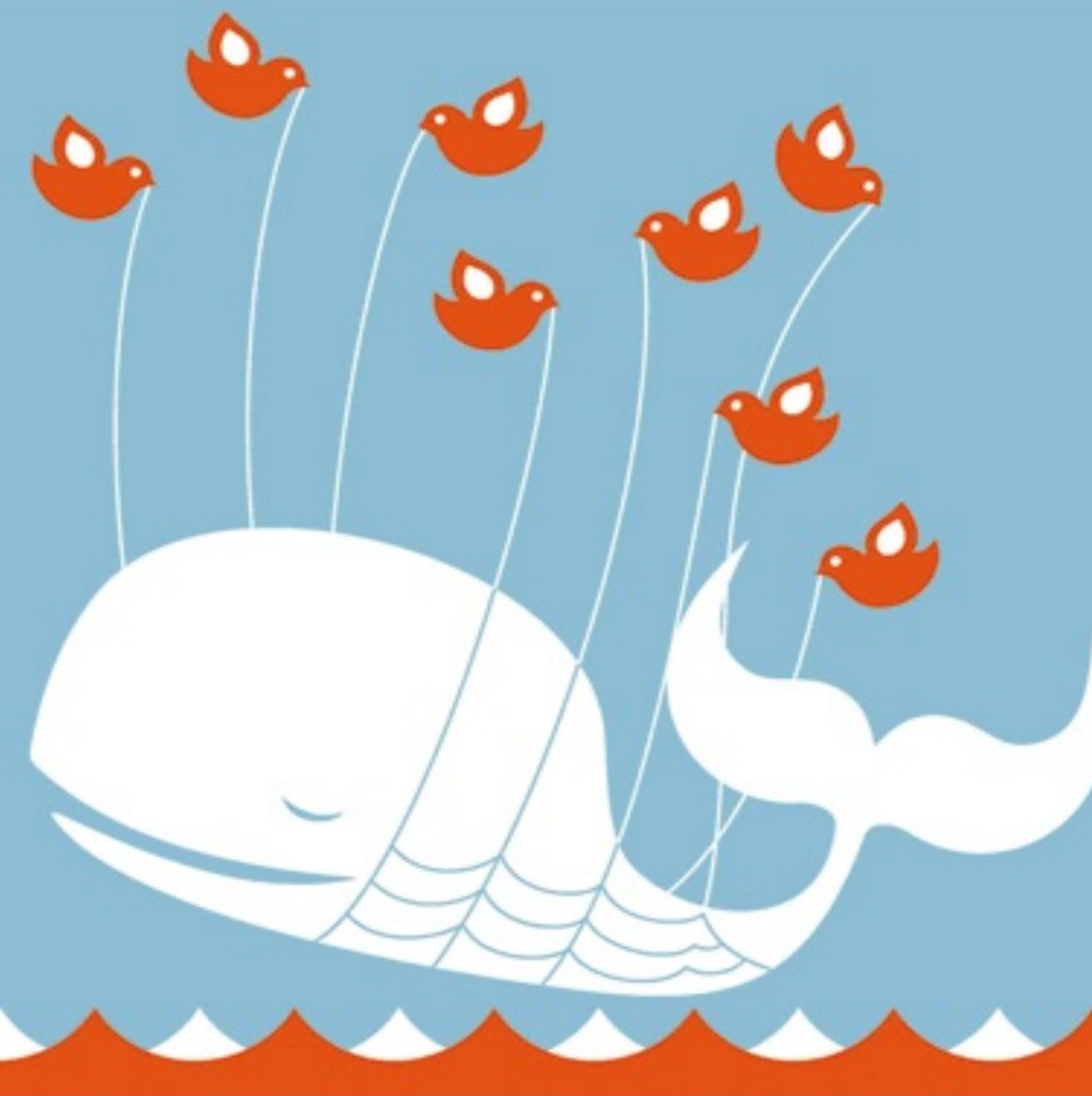


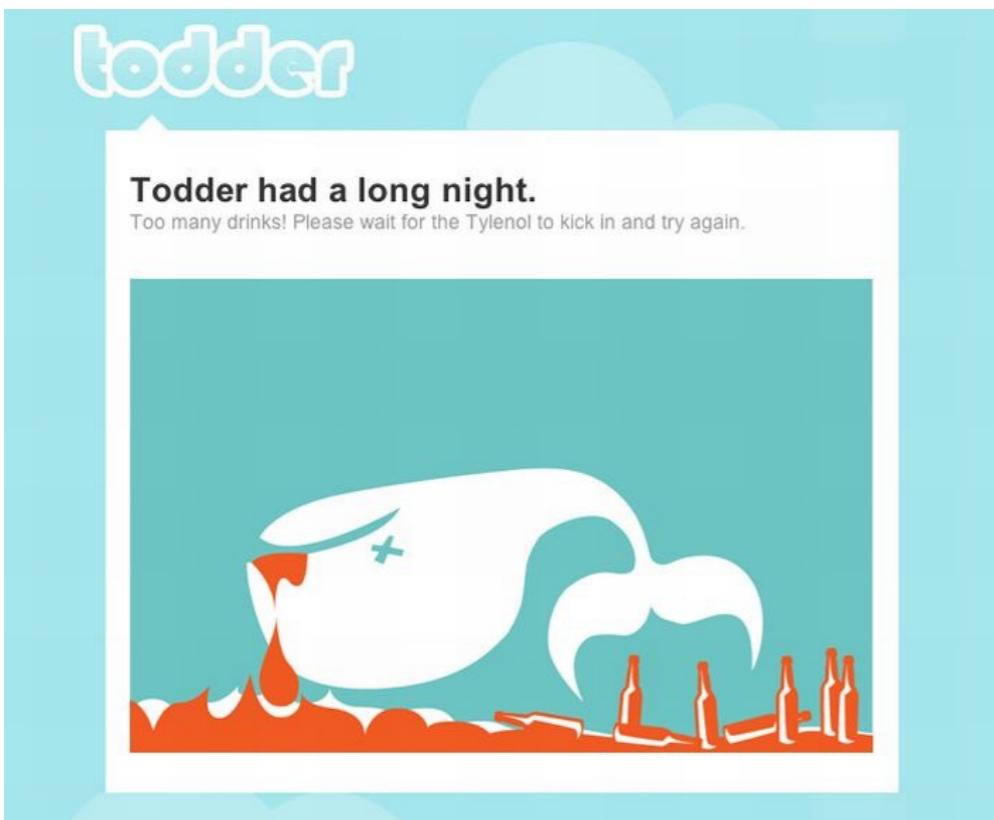
Twitter HQ (since 2012)



Twitter is over capacity.

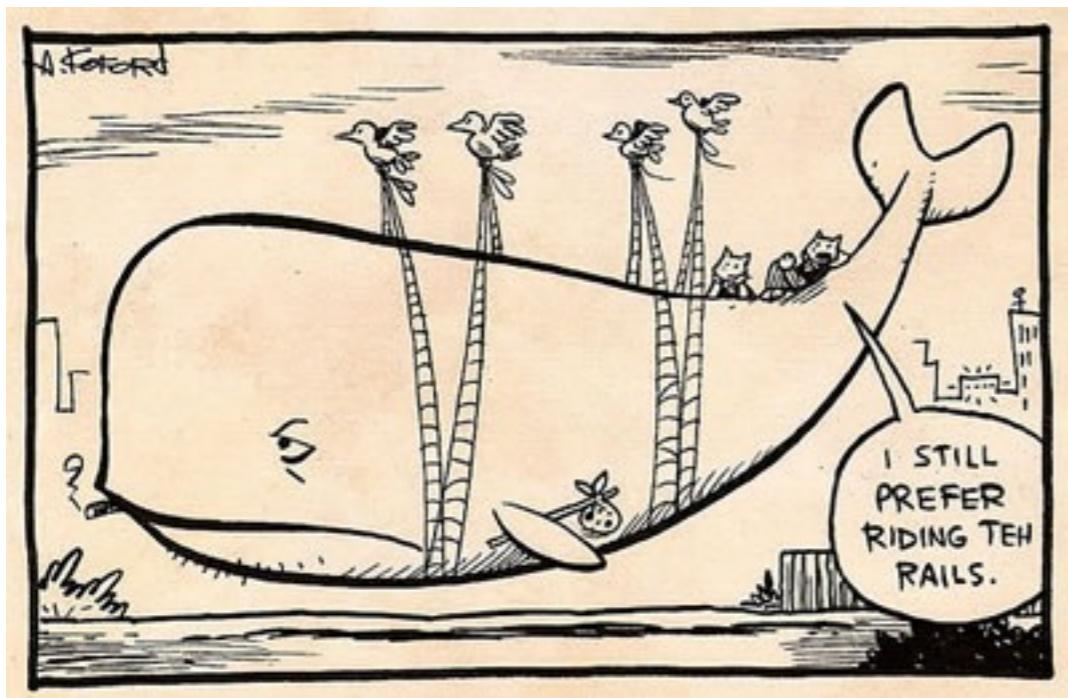
Please wait a moment and try again. For more information, check out [Twitter Status »](#)

[English](#)[Deutsch](#)[Español](#)[Français](#)[Italiano](#)[日本語](#)



Todder had a long night.

Too many drinks! Please wait for the Tylenol to kick in and try again.



Tweets

 **ChuckGrassley** 
@ChuckGrassley

July4 we celeb8 freedom&liberty that started
w gr8est document evr written Declaration of
Independence SO hv happy Independence
Day.IPARADE

RETWEETS FAVORITES
550 **461**

11:15 AM - 3 Jul 2015



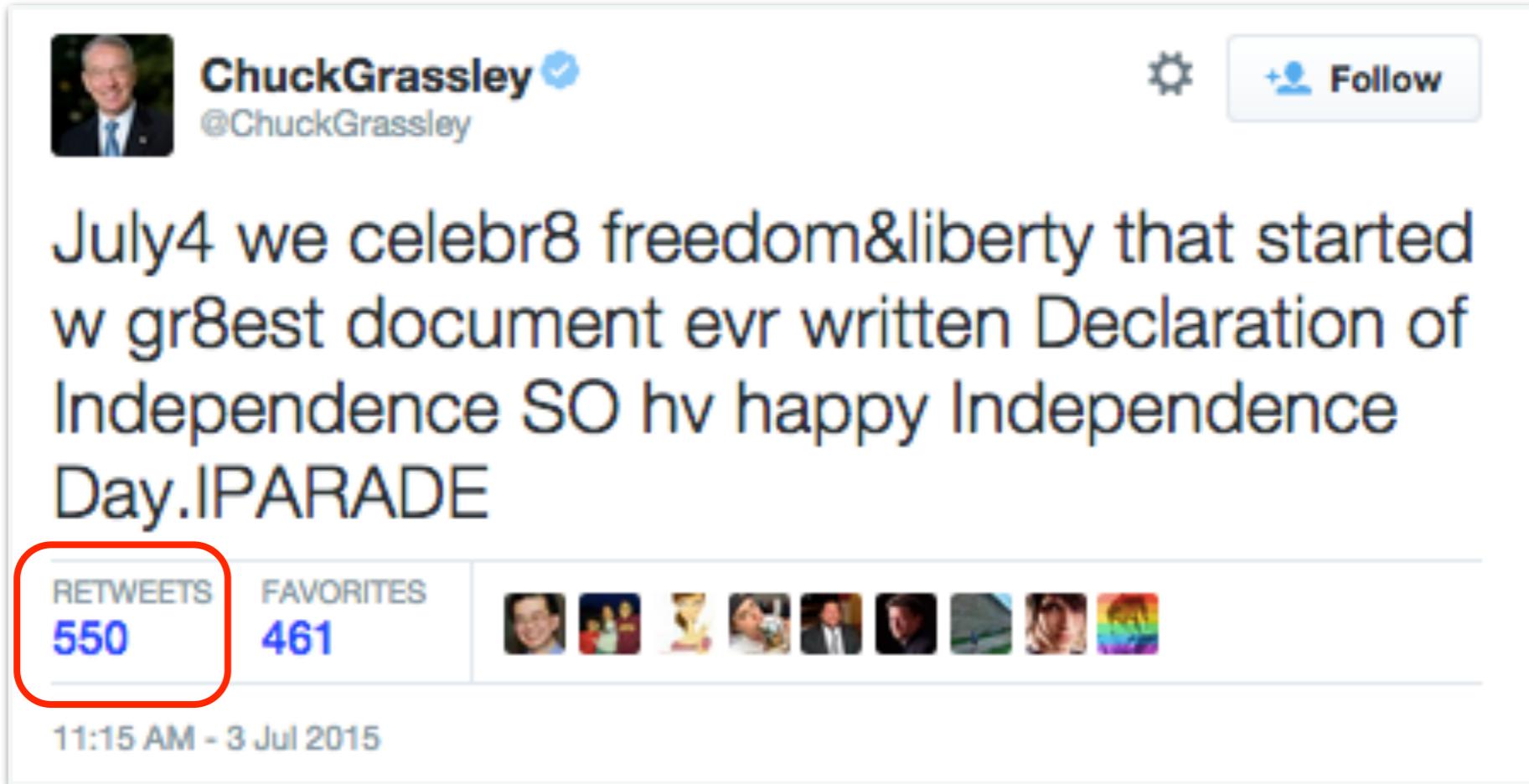
ReTweets

 **ChuckGrassley** 
@ChuckGrassley

July4 we celeb8 freedom&liberty that started
w gr8est document evr written Declaration of
Independence SO hv happy Independence
Day.IPARADE

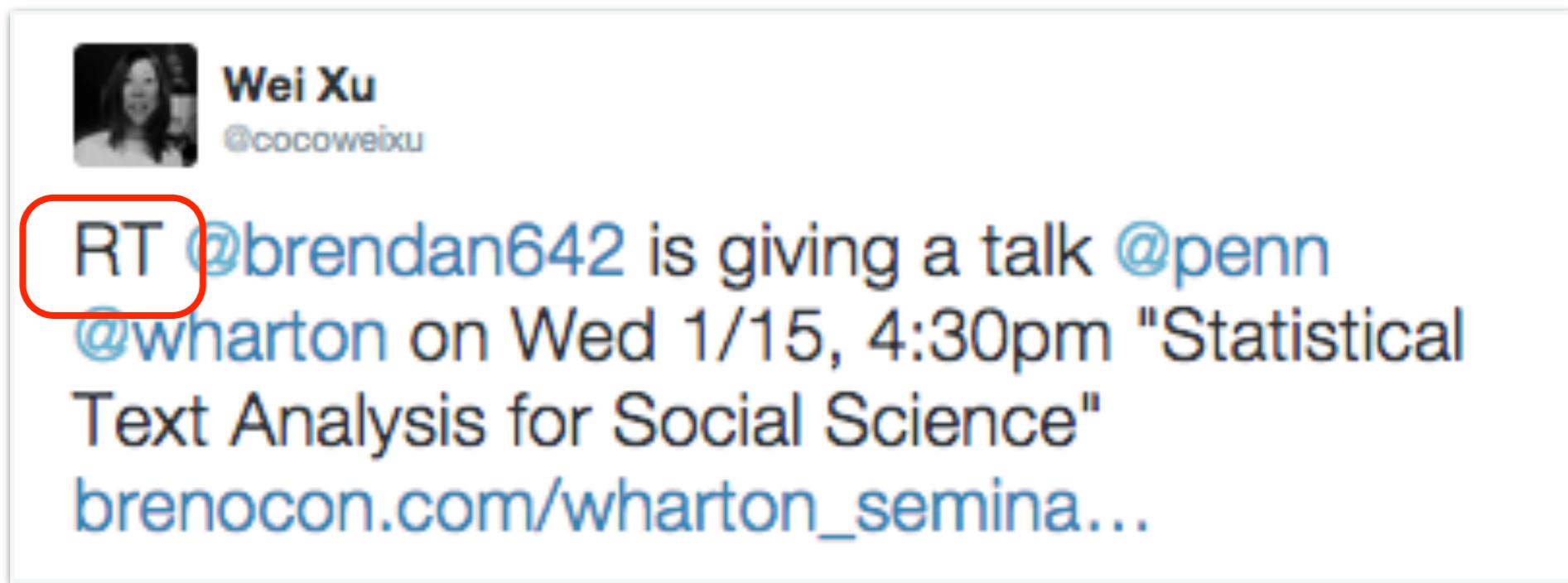
RETWEETS 550 FAVORITES 461

11:15 AM - 3 Jul 2015



a re-posting of someone else's Tweet

ReTweets



Wei Xu
@cocoweixu

RT @brendan642 is giving a talk @penn @wharton on Wed 1/15, 4:30pm "Statistical Text Analysis for Social Science" brenocon.com/wharton_semina...

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

Embedded Links



Wei Xu
@cocoweixu

RT @brendan642 is giving a talk @penn
@wharton on Wed 1/15, 4:30pm "Statistical
Text Analysis for Social Science"
brenocon.com/wharton_semina...

- shortened for display

Embedded Links

 **Richard Henry** @richardhenry 4m
That's a whole lot of people... nyti.ms/yzg6Wq

[Hide summary](#) [Reply](#) [Retweet](#) [Favorite](#)

Parade of Fans for Houston's Funeral

By Sarah Maslin Nir @SarahMaslinNir

NEWARK — The guest list and the parade of limousines with celebrities emerging from them seemed a little more suited to a red carpet event in Hollywood or ...

 **The New York Times** @nytimes



9:03 PM Sep 27 via Twitter for Mac · View Tweet page

- can provide extra external information for text processing

Mentions



Wei Xu

@cocoweixu

RT @brendan642 is giving a talk @penn
@wharton on Wed 1/15, 4:30pm "Statistical
Text Analysis for Social Science"
brenocon.com/wharton_semina...

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

Replies/Conversations

The image shows a Twitter interaction. A user (@.) asks J.K. Rowling (@jk_rowling) why Dumbledore is described as gay. J.K. Rowling replies that maybe it's because gay people just look like... people? The tweet has 93 likes and 29 retweets.

Amberleec @. · 2h
@jk_rowling Thank you so much for writing Harry Potter. I wonder why you said that Dumbledore is a gay because I can't see him in that way.

93 29

J.K. Rowling @jk_rowling

. Maybe because gay people just look like... people?

- Tweet starts with a @username

Replies/Conversations



Wei Xu

@cocoweixu

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

11:55 AM - 2 Jul 2015

51 Retweets 105 Likes



6



51



105



Tweet your reply



Jacob Eisenstein @jacobeisenstein · 2 Jul 2015

Replies to @cocoweixu

@cocoweixu @twitterapi nice! but as long as Twitter keeps changing the API, no tutorial will be "ultimate" :)



1



1



1



Wei Xu @cocoweixu · 12 Jul 2015

@jacobeisenstein yep that's why I put a date on so ppl know when its out-of-date. hope Twitter Python Tool can handle the updates too



brendan o'connor @brendan642 · 2 Jul 2015

Replies to @cocoweixu

@cocoweixu great! btw re 1 giant line, i've found "print json.dumps(tweet, indent=4)" pretty printing to be useful

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

[Answer](#)[Request ▾](#)

Follow 9 Comment Share Downvote

...

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 \(@tommitchell\) ↗](#)
- [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 ↗](#)



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

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

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 update... [\(more\)](#))

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

...

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](#)

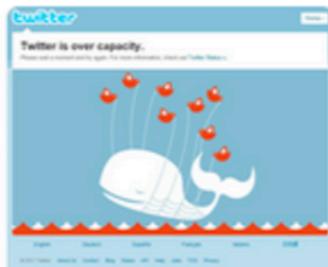
Images



Wei Xu
@cocoweixu

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

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 credentials from the 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 account.
- Click "Create New App"

11:55 AM - 2 Jul 2015

51 Retweets 105 Likes

6 51 105

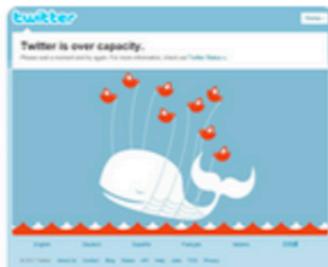
Hashtags



Wei Xu
@cocoweixu

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

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 credentials from the 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 account.
- Click "Create New App"

11:55 AM - 2 Jul 2015

51 Retweets 105 Likes



6

51

105

|||



hashtags are powerful

Cashtags

 Twitter 
@twitter

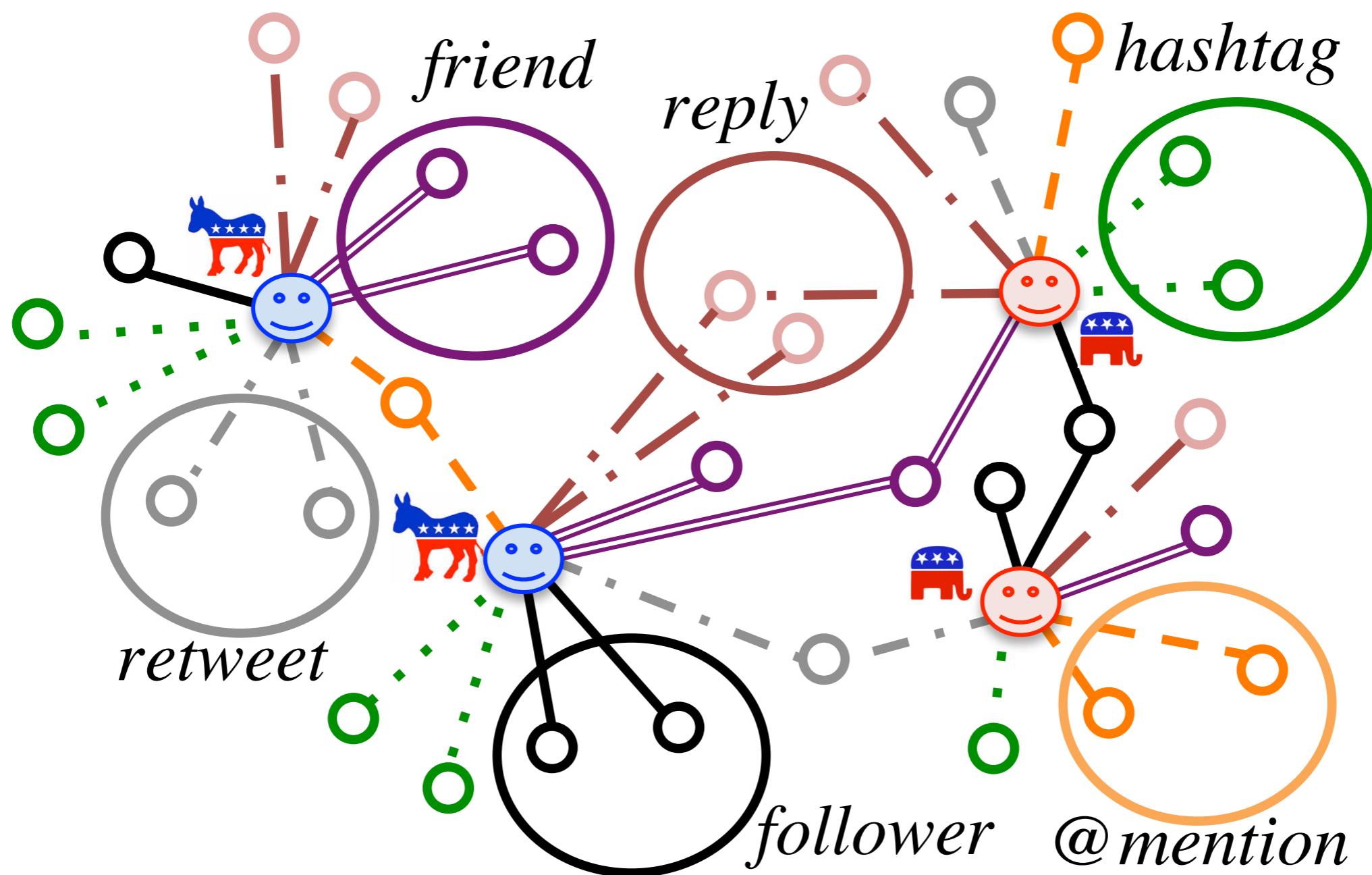


Now you can click on ticker symbols like \$GE on [twitter.com](#) to see search results about stocks and companies

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.

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 or locations)	- 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

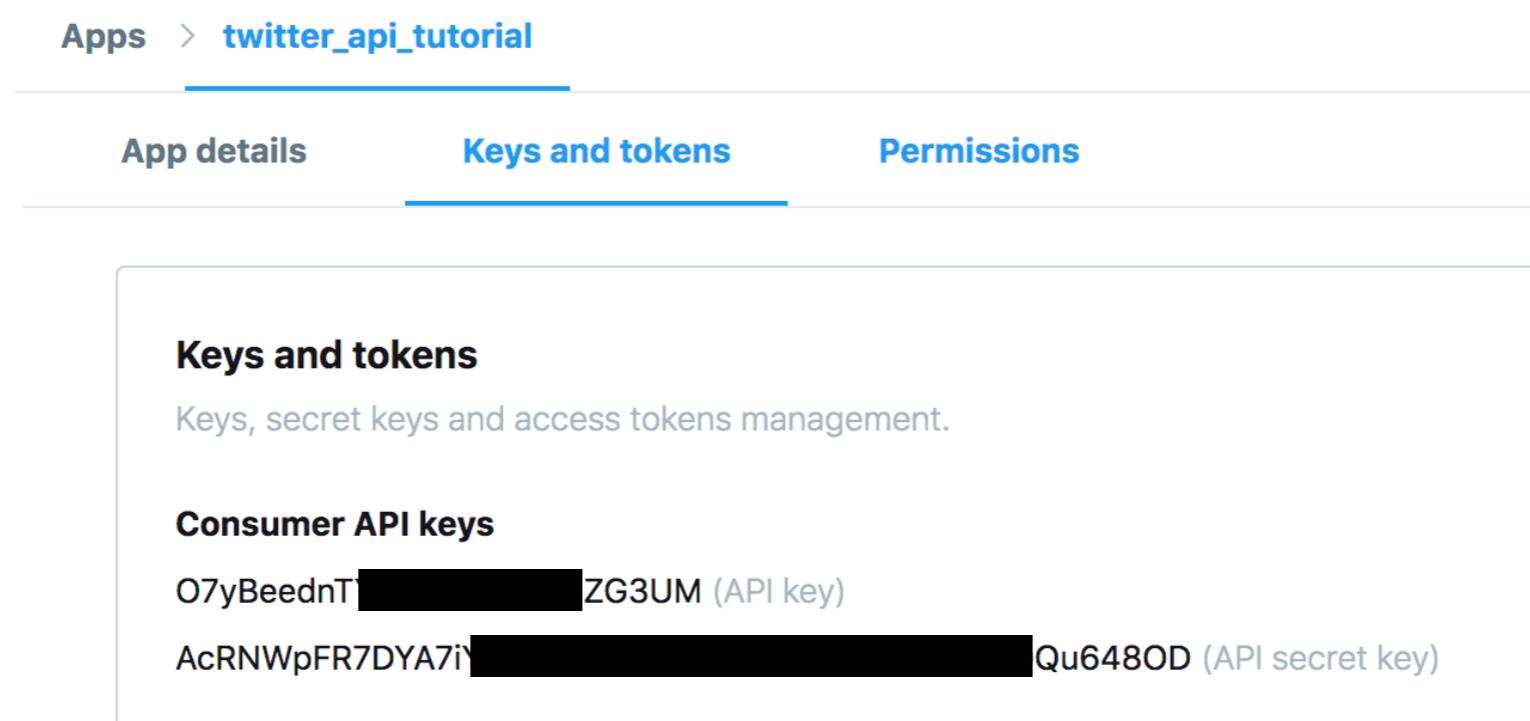
- Twitter uses OAuth to provide authorized access to its API.
- which means, to start with needs:
 - a Twitter account
 - OAuth access tokens from developer.twitter.com/

Apps > [twitter_api_tutorial](#)

[App details](#) [Keys and tokens](#) [Permissions](#)

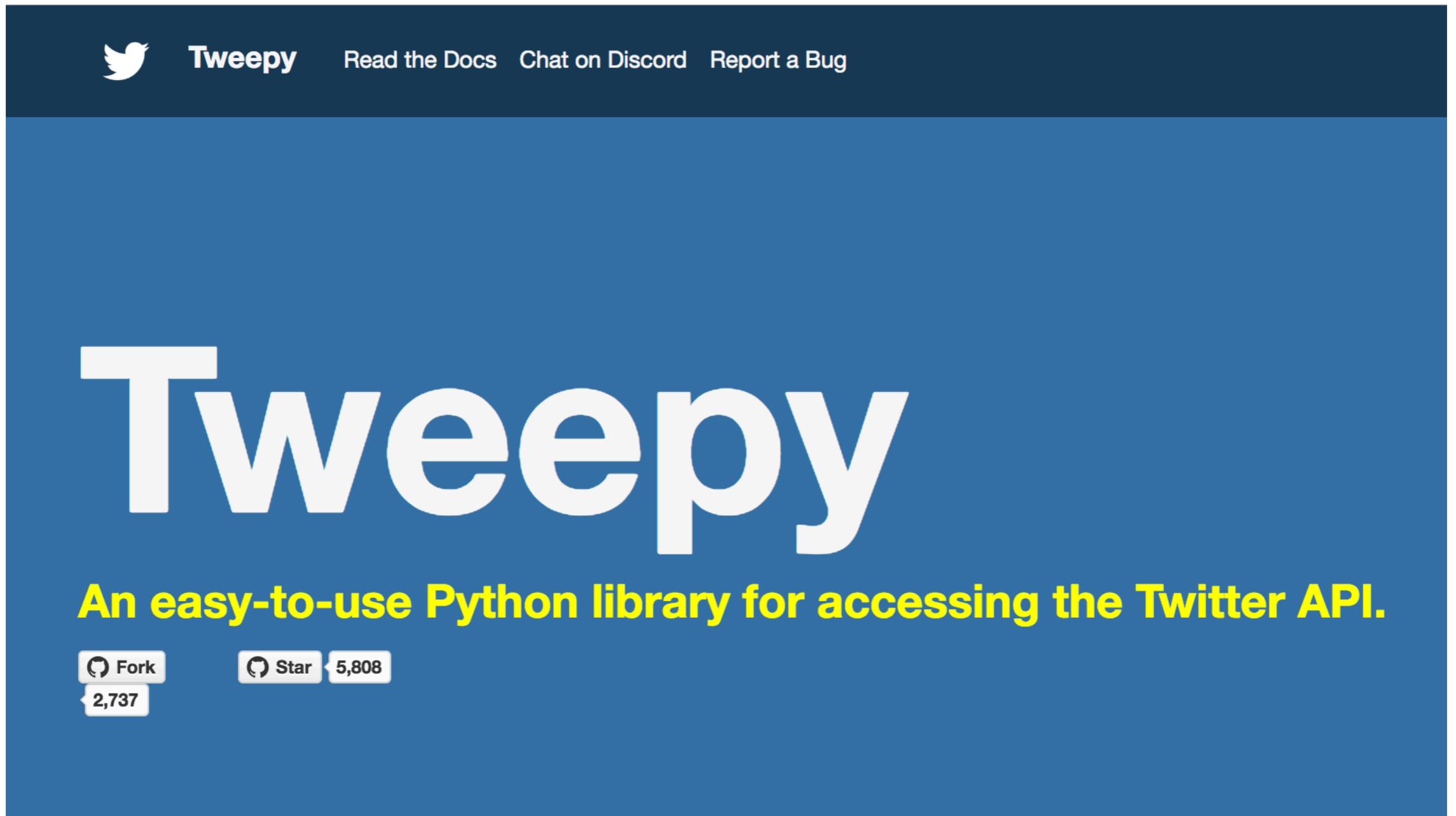
Keys and tokens
Keys, secret keys and access tokens management.

Consumer API keys
O7yBeednT [REDACTED] ZG3UM (API key)
AcRNWpFR7DYA7iY [REDACTED] Qu648OD (API secret key)



Python Twitter Tools

www.tweepy.org



The screenshot shows the GitHub repository page for Tweepy. At the top, there's a dark header bar with the Twitter logo, the word "Tweepy", and links for "Read the Docs", "Chat on Discord", and "Report a Bug". Below the header is a large blue background area. In the center, the word "Tweepy" is written in a large, white, sans-serif font. Below it, the text "An easy-to-use Python library for accessing the Twitter API." is displayed in a yellow font. At the bottom left, there are two GitHub statistics: "Fork" (2,737) and "Star" (5,808). The overall theme is blue and white.

OAuth Authentication

Twitter uses OAuth to provide authorized access to the API.

```
[ ] # Import the tweepy library
import tweepy
from tweepy.streaming import StreamListener

# 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'

# Setup tweepy to authenticate with Twitter credentials:
auth = tweepy.OAuthHandler(CONSUMER_KEY, CONSUMER_SECRET)
auth.set_access_token(ACCESS_TOKEN, ACCESS_SECRET)
```

Streaming API

```
[ ] # This is a basic listener that just prints received tweets to stdout.  
class StdOutListener(StreamListener):  
  
    def on_data(self, data):  
        print(data)  
        return True  
  
    #def on_status(self, status):  
    #    print(json.dumps(status._json))  
    #    return True  
  
    def on_error(self, status_code):  
        print(status_code)  
        return False  
  
# tweepy.Stream.sample() will give a live stream (~1% sample) of all public tweets  
# Warning: it will continue to run indefinitely until you stop it.  
  
listener = StdOutListener()  
twitterStream = tweepy.Stream(auth, listener)  
twitterStream.sample()
```

JSON

JavaScript Object Notation

JSON is a minimal, readable format for structuring data.

A Tweet in JSON



Wei Xu
@cocoweixu

#CFP Workshop on Noisy User-generated Text at ACL - Beijing 31 July 2015. Papers due: 11 May 2015. noisy-text.github.io
#NLProc #WNUT15

```
{  
    "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 Moments Notifications Messages  #nlproc   Tweet

ACL2019 and 4 others liked

Manaal Faruqui @manaalfar · 6h
As a junior AC for [@ACL2019_Italy](#) I have to write meta-reviews for only 10 papers in the morphology/phonology track. So much better than ACL 2017 when I had to do it for 25 papers in Semantics track. Refreshing to be an AC of a smaller area! :D [#NLProc](#)

1 13

Stanford NLP Group @stanfordnlp · 6h

 Today, Mar 20—[@Stanford](#) CS224N NLP with Deep Learning Poster Session 5–9pm Arrillaga Alumni Center. Free parking after 4pm in A/C spots on Galvez, lots, garages. Come talk with 500 amazing Stanford students about question answering, dialog, MT etc [#NLProc](#) [facebook.com/events/1218481...](#)



2 13

Machine Learning and NLP @ML_NLP · 6h

GluonNLP 0.6: Closing the Gap in Reproducible Research with BERT
[medium.com/apache-mxnet/g...](#) [#NLProc](#)

Search API

```
[ ] # Search for latest tweets about "#nlproc"
tweets = tweepy.Cursor(api.search, q='#nlproc')

# Print out the latest 10 tweets that contain "#nlproc" hashtag
for item in tweets.items(10):
    print(item._json)
```

```
> {'name': 'Dwayne Haskins', 'url': 'http://twitter.com/search?q=%22Dwayne+Haskins%22', 'promoted_content': None, 'query': '%22Dwayne+Haskins%22', 'twe
{'name': 'McCain', 'url': 'http://twitter.com/search?q=McCain', 'promoted_content': None, 'query': 'McCain', 'tweet_volume': 381900}
{'name': 'Lima', 'url': 'http://twitter.com/search?q=Lima', 'promoted_content': None, 'query': 'Lima', 'tweet_volume': 70085}
{'name': '#firstdayofspring', 'url': 'http://twitter.com/search?q=%23firstdayofspring', 'promoted_content': None, 'query': '%23firstdayofspring', 't
{'name': 'Daniel Caesar', 'url': 'http://twitter.com/search?q=%22Daniel+Caesar%22', 'promoted_content': None, 'query': '%22Daniel+Caesar%22', 'tweet_
{'name': '#InternationalDayOfHappiness', 'url': 'http://twitter.com/search?q=%23InternationalDayOfHappiness', 'promoted_content': None, 'query': '%23
{'name': 'AirPods', 'url': 'http://twitter.com/search?q=AirPods', 'promoted_content': None, 'query': 'AirPods', 'tweet_volume': 131265}
{'name': 'Pro Day', 'url': 'http://twitter.com/search?q=%22Pro+Day%22', 'promoted_content': None, 'query': '%22Pro+Day%22', 'tweet_volume': 19808}
{'name': '#SpringEquinox', 'url': 'http://twitter.com/search?q=%23SpringEquinox', 'promoted_content': None, 'query': '%23SpringEquinox', 'tweet_volum
{'name': 'Flume', 'url': 'http://twitter.com/search?q=Flume', 'promoted_content': None, 'query': 'Flume', 'tweet_volume': 37576}
{'name': '#HappinessInOneWord', 'url': 'http://twitter.com/search?q=%23HappinessInOneWord', 'promoted_content': None, 'query': '%23HappinessInOneWord'
{'name': '#StrangerThings3', 'url': 'http://twitter.com/search?q=%23StrangerThings3', 'promoted_content': None, 'query': '%23StrangerThings3', 'tweet
{'name': 'Eloy', 'url': 'http://twitter.com/search?q=Eloy', 'promoted_content': None, 'query': 'Eloy', 'tweet_volume': None}
{'name': 'Happy Spring', 'url': 'http://twitter.com/search?q=%22Happy+Spring%22', 'promoted_content': None, 'query': '%22Happy+Spring%22', 'tweet_vo
{'name': 'Bill & Ted 3', 'url': 'http://twitter.com/search?q=%22Bill+%26+Ted+3%22', 'promoted_content': None, 'query': '%22Bill+%26+Ted+3%22', 'tweet
```

Trends

Home Moments Notifications Messages  Search Twitter  Tweet

 **Wei Xu**
@cocoweixu
Tweets 423 Following 510 Followers 2,767

Columbus trends · Change

#ApexSeason1
Apex Legends Season 1 is Here
↗ Promoted by Apex Legends

Terry McLaurin

Ohio State
13.4K Tweets

Haskins
9,893 Tweets

Pro Day
21.9K Tweets

Johnnie Dixon

Nick Bosa
1,853 Tweets

Giants
35.9K Tweets

#firstdayofspring
It is officially the first day of spring! 🌸🌼

Lima
63.6K Tweets

What's happening?

 **Women in Analytics Conference (WIA)** @wia_conference · 4s
We are so excited to see everyone tomorrow at the 2019 Women in Analytics Conference! Keep us in the loop by using hashtag #WIA2019 when sharing your favorite conference moments on [@LinkedIn](#), [@Twitter](#), [@Facebook](#), and [@Instagram](#).
womeninanalytics.org



2019 Women in Analytics Conference | Ethics in Algorithms
womeninanalytics.org

Travelfuntu @travel_funtu · Mar 10
This Photo Has Not Been Edited, Look Closer



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.



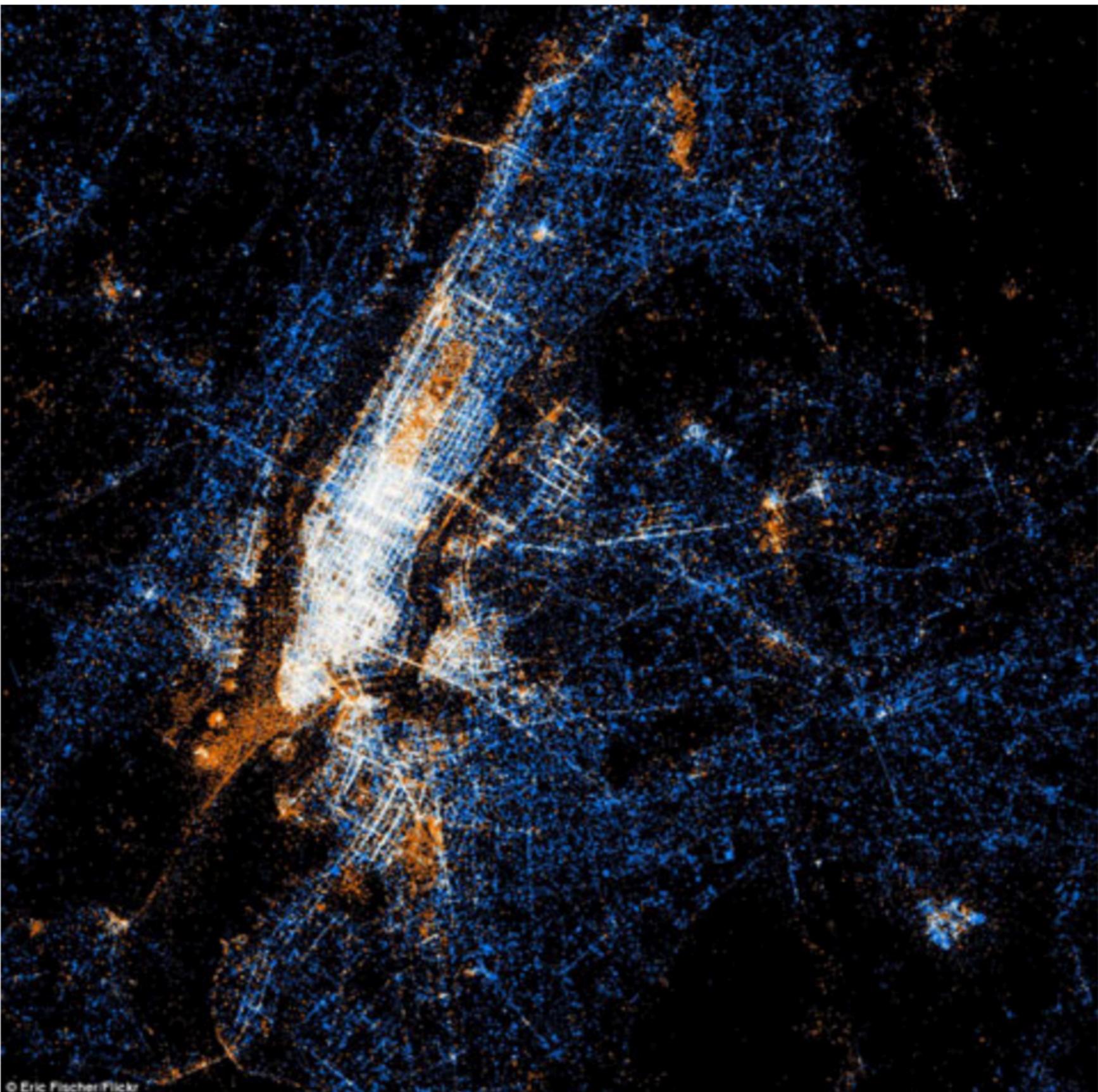
Trends API

Where On Earth ID



```
# Where On Earth ID for Columbus, Ohio is 2383660.  
COLUMBUS_WOE_ID = 2383660  
  
columbus_trends = api.trends_place(COLUMBUS_WOE_ID)  
  
trends = json.loads(json.dumps(columbus_trends, indent=1))  
  
for trend in trends[0]["trends"]:  
    print (trend)
```





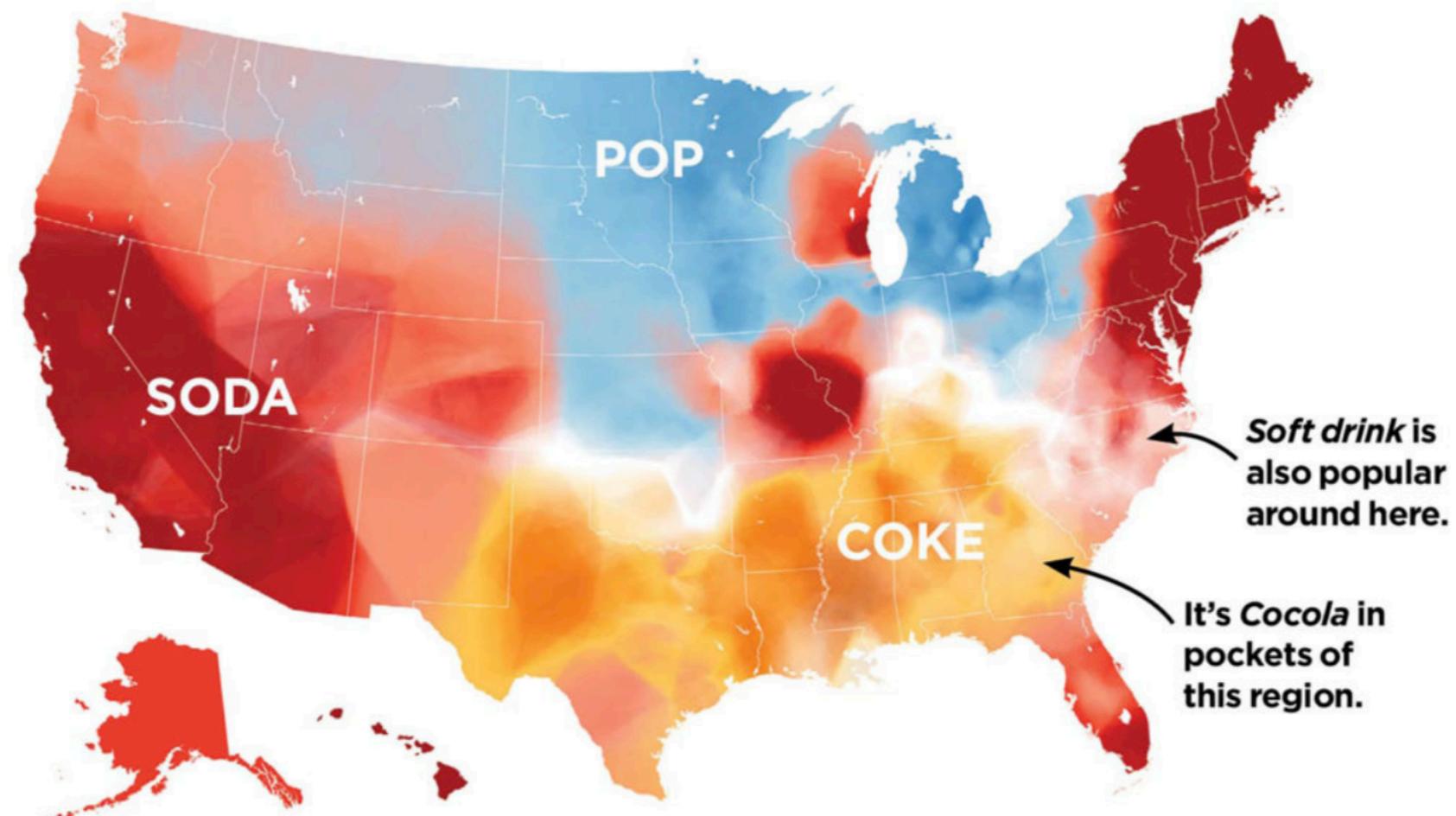
© Eric Fischer/Flickr

visualizations by [Eric Fischer](#) A visualization showing the location of Twitter messages and Flickr photos in New York City.

What do you call carbonated beverages?

- Soda
- Pop
- Coke
- other ways?

What do you call carbonated beverages?



JOSH KATZ

What do you call a sale of household items

- Garage Sale
- Yard Sale
- other ways?

What do you call a sale of household items

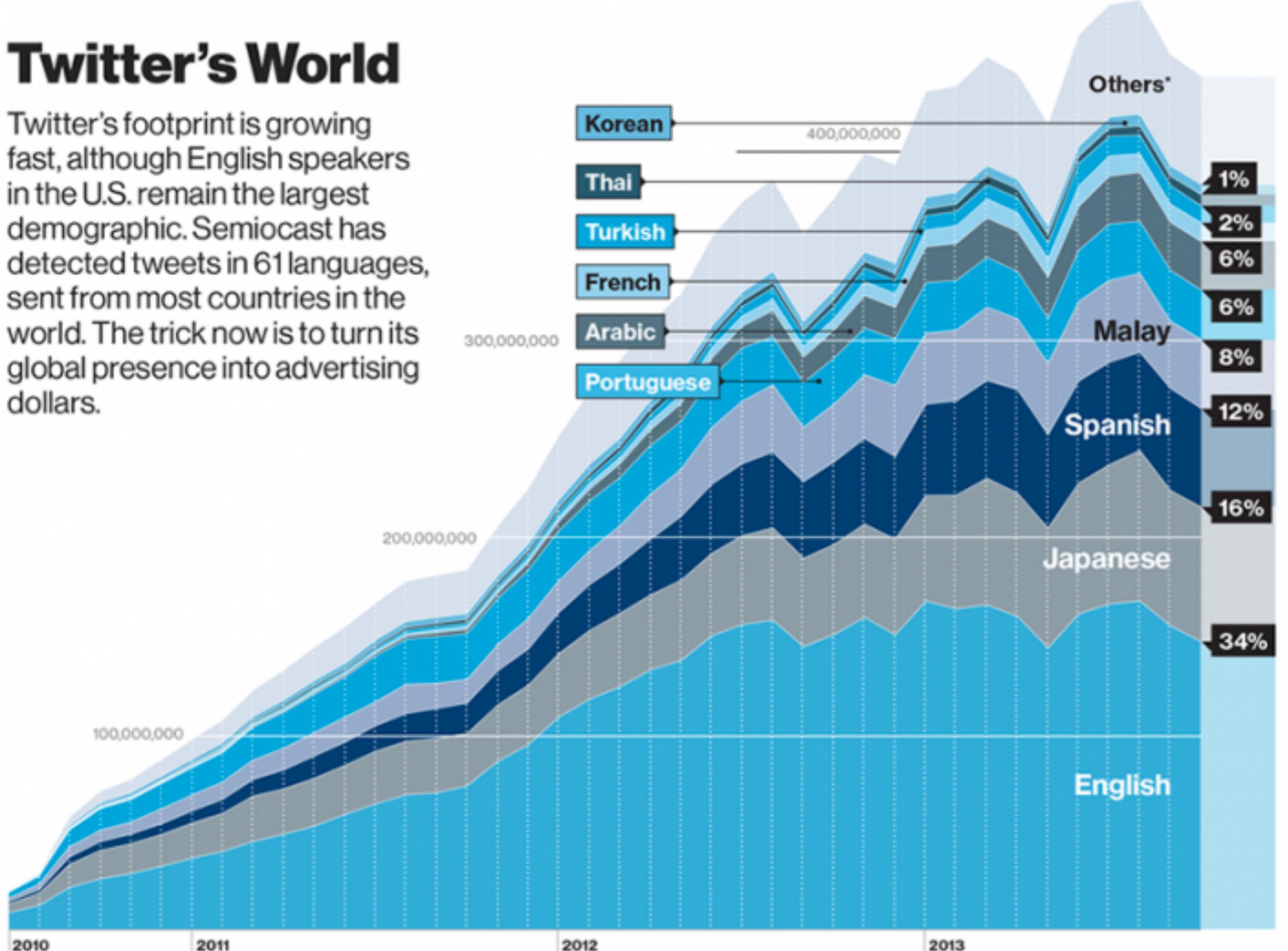


JOSH KATZ

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.

AVERAGE NUMBER OF TWEETS PER DAY



Social Media Under Fire

Countries that block Twitter, Facebook, or YouTube*



*Full or partial blocks, as of March 2014
Sources: Google, Twitter, OpenNet Initiative

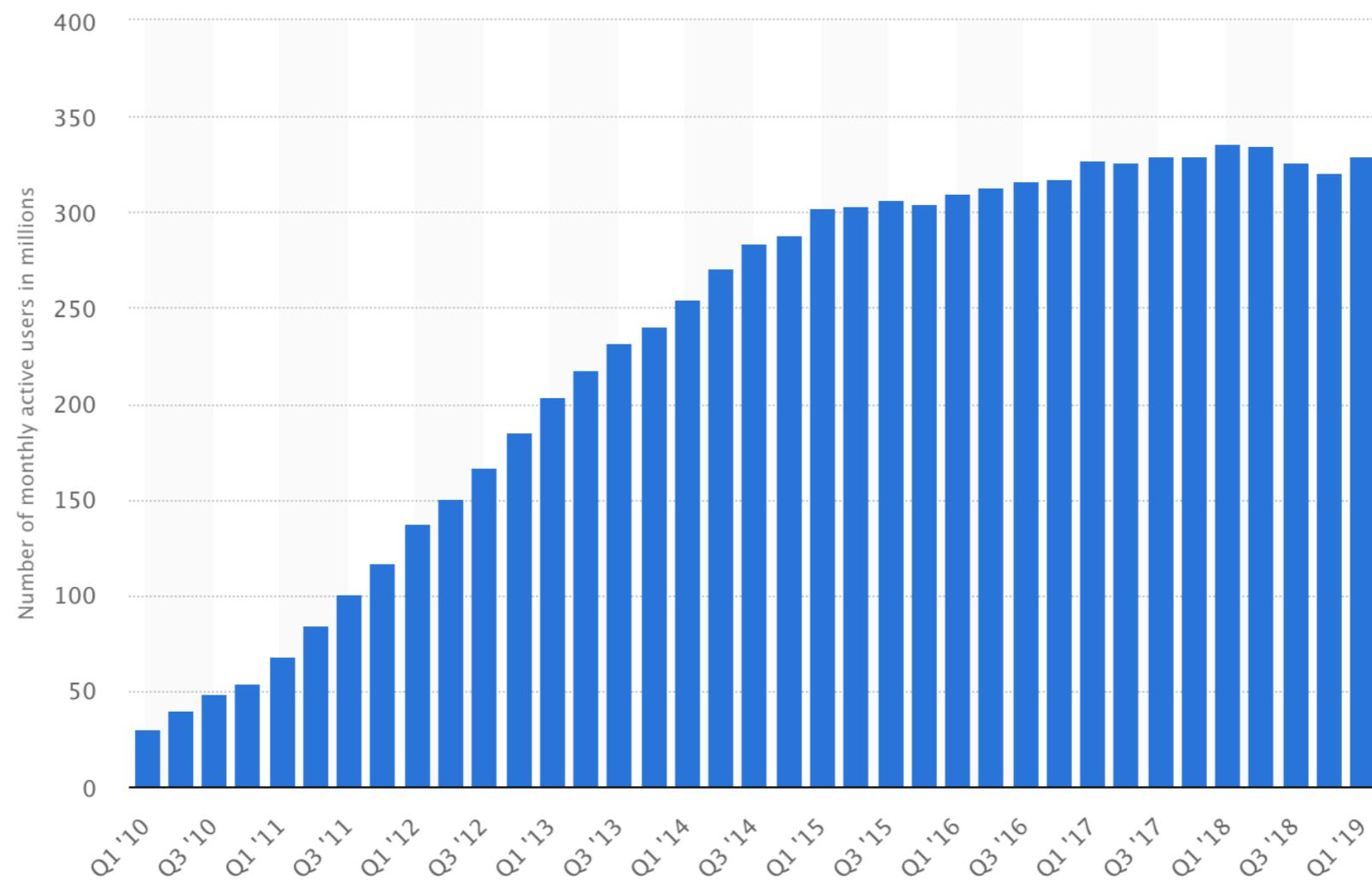
Mother Jones



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

Twitter Demographics

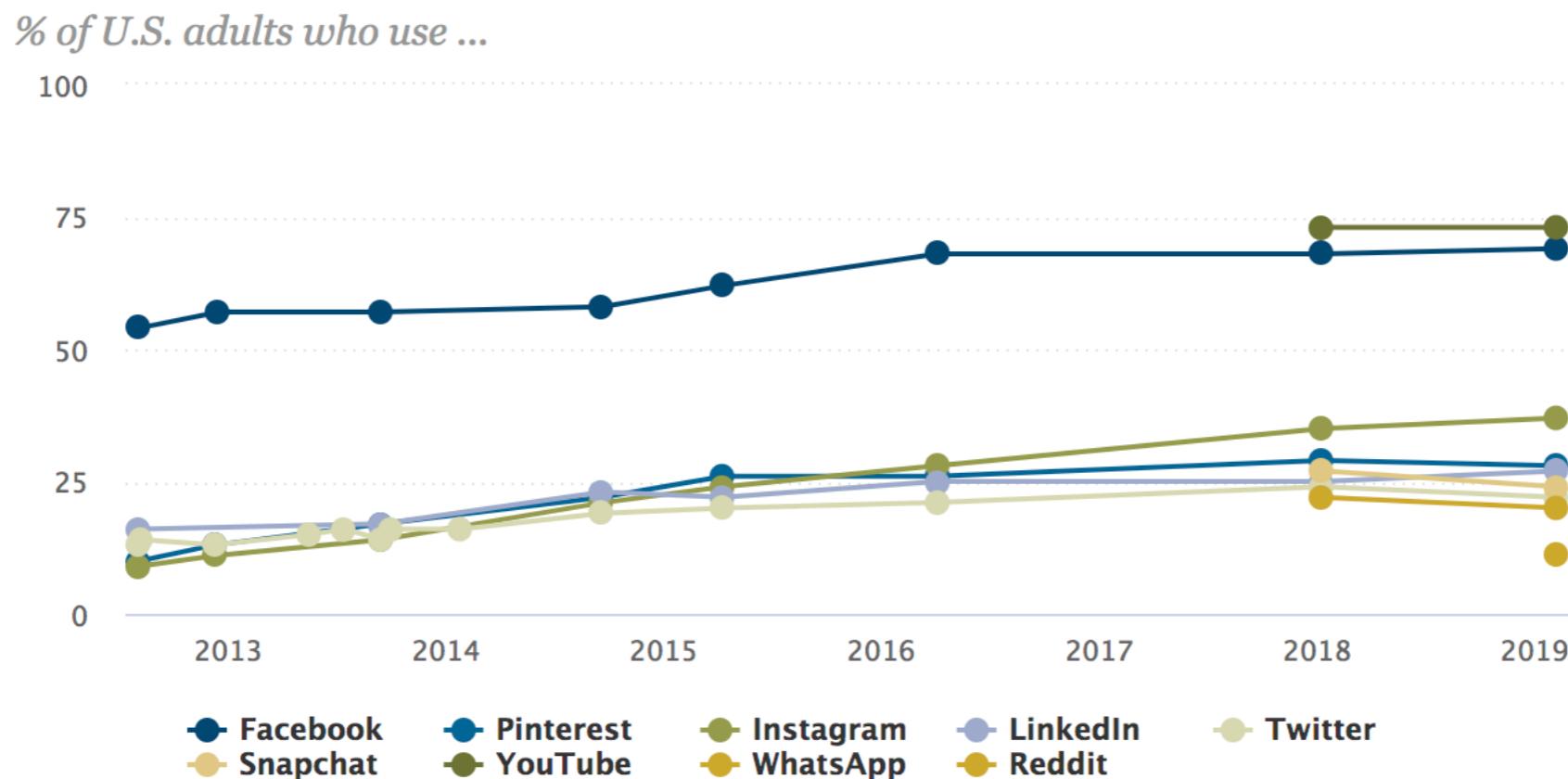
- As of the 1st quarter of 2019, Twitter averaged 330 million monthly active users, and 139 million monetizable daily active Twitter users worldwide.



Twitter Demographics

- About 80% (262 million) of all monthly active Twitter users live outside the United States.
- 72 million monthly active users live within the United States.
- The top countries on Twitter outside the U.S. are Japan (50.9 million users), the United Kingdom (18.6 million users), and Saudi Arabia (13.8 million users).

The Most Popular Social Media Platforms



Source: Surveys conducted 2012-2019.

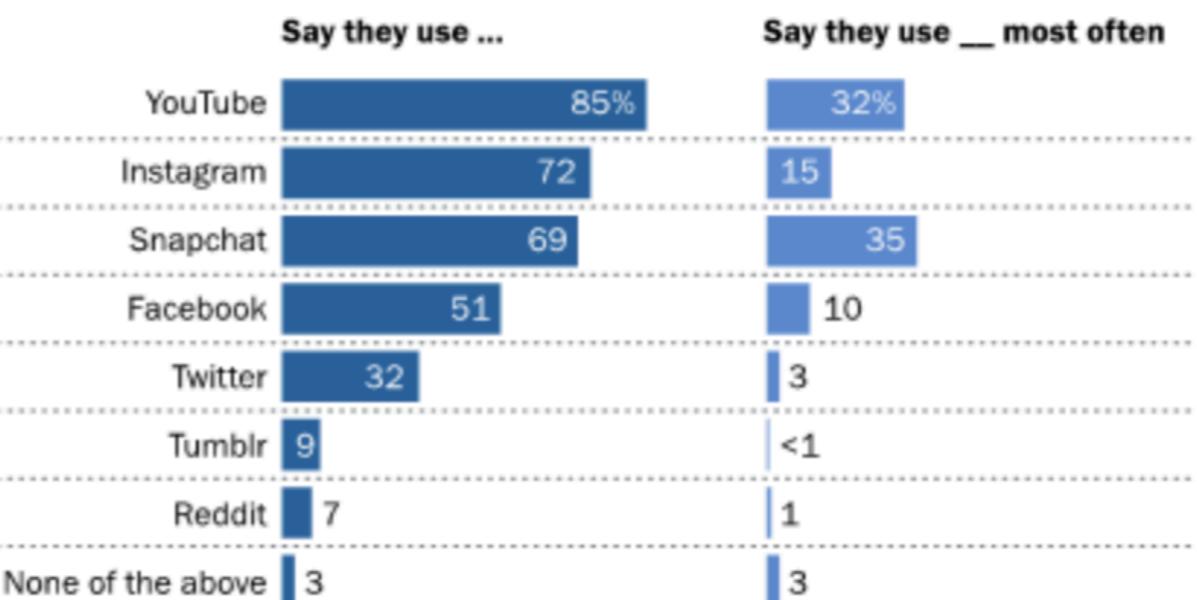
PEW RESEARCH CENTER

The Most Popular Social Media Platforms

- 95% of teens (age 13-17) now report they have a smartphone or access to one. 45% of teens now say they are online on a near-constant basis.

YouTube, Instagram and Snapchat are the most popular online platforms among teens

% of U.S. teens who ...



Note: Figures in first column add to more than 100% because multiple responses were allowed. Question about most-used site was asked only of respondents who use multiple sites; results have been recalculated to include those who use only one site. Respondents who did not give an answer are not shown.

Source: Survey conducted March 7-April 10, 2018.

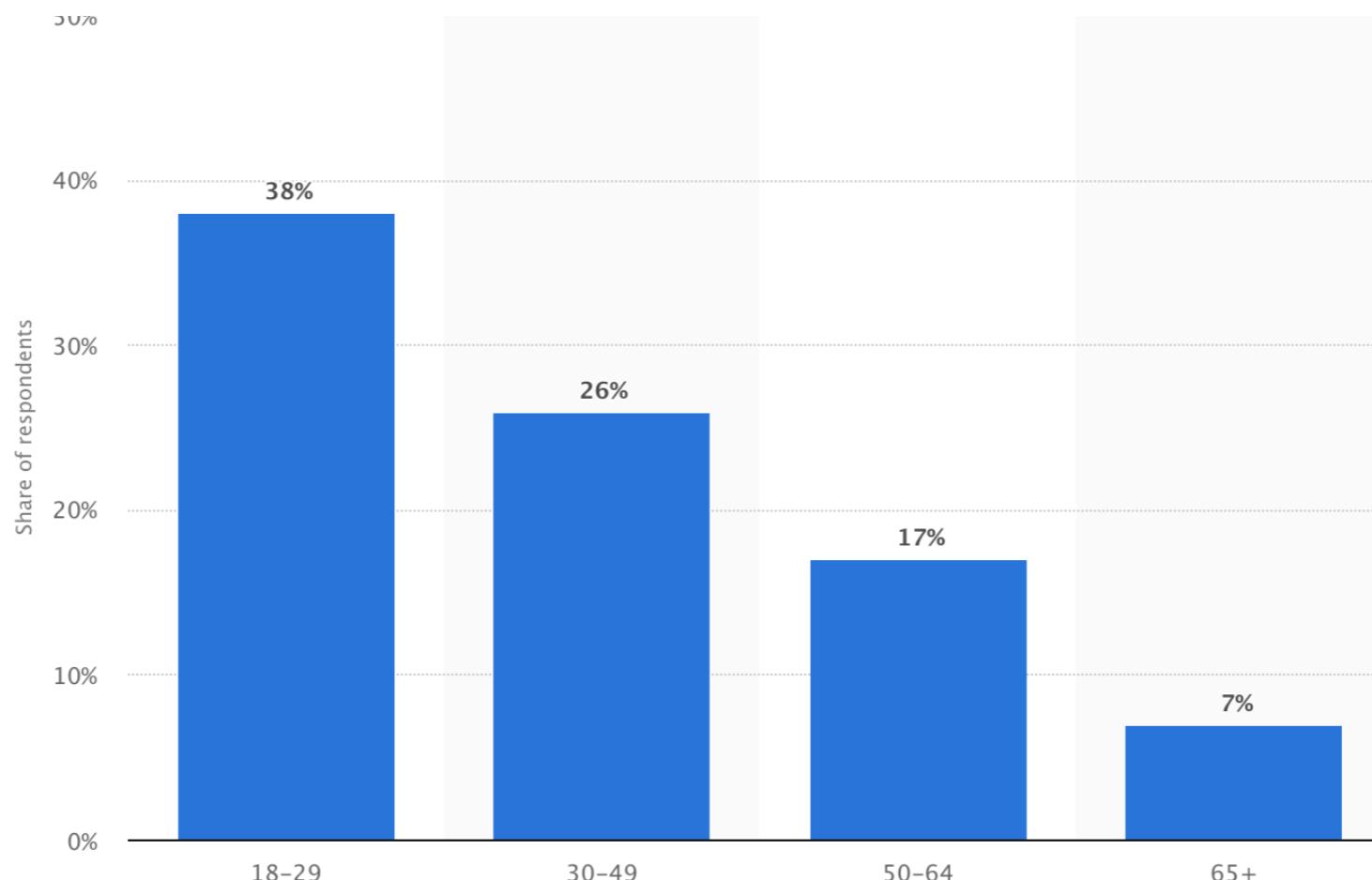
"Teens, Social Media & Technology 2018"

PEW RESEARCH CENTER

Twitter Demographics

- How old are Twitter users? Not very old, it seems. But also not very young.

Percentage of U.S. adults who use Twitter as of February 2019, by age group



Twitter Demographics

- Globally, more men use Twitter than women. But in the United States, more women use Twitter than men.

% of U.S. adults who say they use ...

	Facebook	YouTube	Pinterest	Instagram	Snapchat	LinkedIn	Twitter	WhatsApp
Total	68%	73%	29%	35%	27%	25%	24%	22%
Men	62	75	16	30	23	25	23	20
Women	74	72	41	39	31	25	24	24
White	67	71	32	32	24	26	24	14
Black	70	76	23	43	36	28	26	21
Hispanic	73	78	23	38	31	13	20	49
Ages 18-29	81	91	34	64	68	29	40	27
18-24	80	94	31	71	78	25	45	25
25-29	82	88	39	54	54	34	33	31
30-49	78	85	34	40	26	33	27	32
50-64	65	68	26	21	10	24	19	17
65+	41	40	16	10	3	9	8	6

Note: Whites and blacks include only non-Hispanics. Hispanics are of any race.

Source: Survey conducted Jan. 3-10, 2018

"Social Media Use in 2018"

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**
 - Conference News
 - ACL**
 - EACL
 - EMNLP
 - NAACL**
 - IJCNLP
- Events
- ACL Fellows
- SIGs
- Anthology
- Wiki
- Software Registry
- Education
- Policies
- Archives



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.

Wei Xu o socialmedia-class.org

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, ...

NLP Publications

- ACL Anthology (<http://aclweb.org/anthology/>)
all NLP conference and journal papers (free!)



The ACL Anthology currently hosts 52251 papers on the study of computational linguistics and natural language processing.

[Subscribe to the mailing list](#) to receive announcements and updates to the Anthology.

The Anthology can archive your poster or presentation! Please submit them in PDF format by [filling out this form](#). Attachments will be distributed under the terms of the [CC-BY-4.0 license](#).

[Full Anthology as BibTeX \(7.16 MB\)](#)

[Give feedback](#)

ACL Events

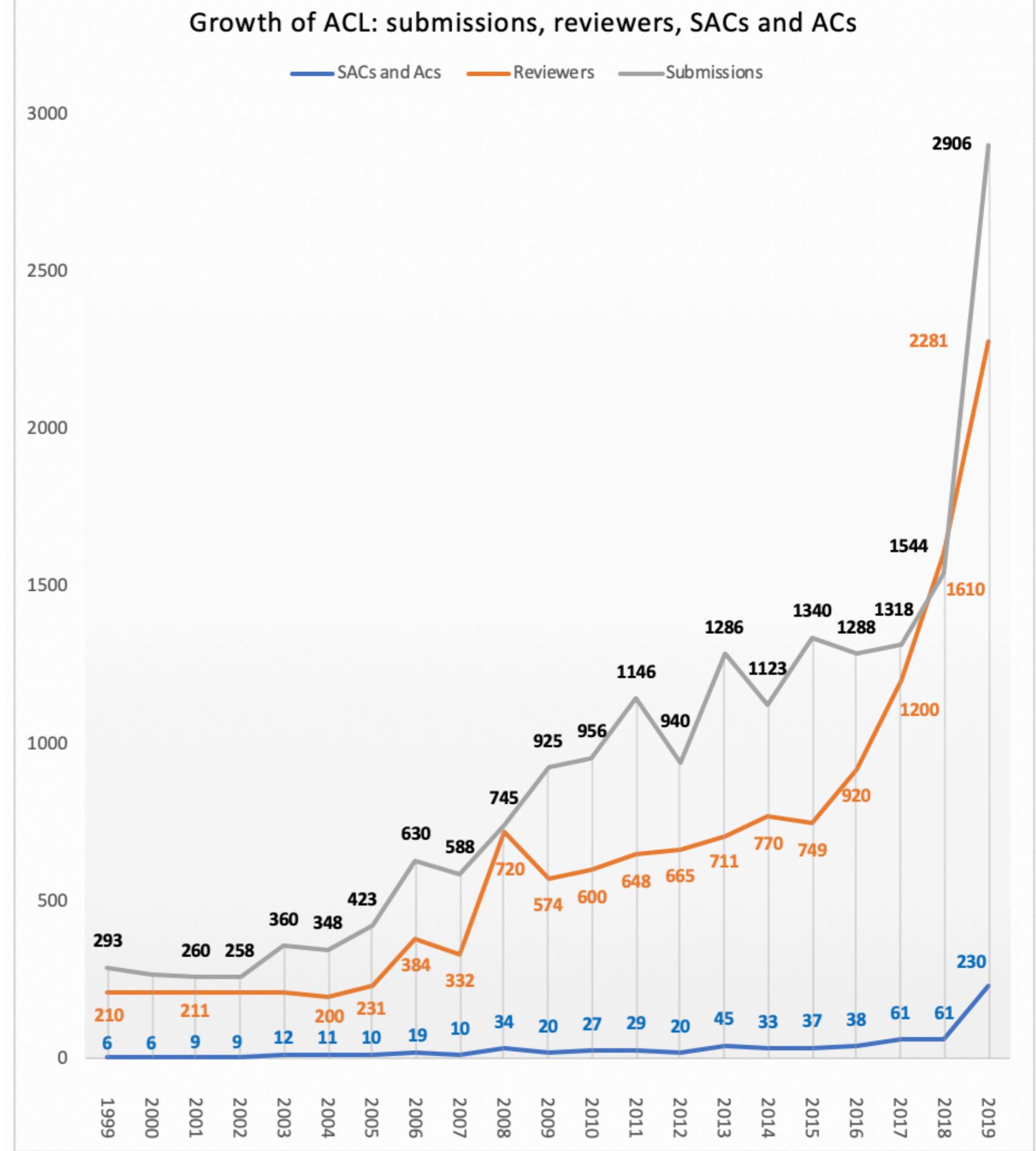
Venue	Present – 2010										2009 – 2000										1999 – 1990										1989 and older																	
ACL	19	18	17	16	15	14	13	12	11	10	09	08	07	06	05	04	03	02	01	00	99	98	97	96	95	94	93	92	91	90	89	88	87	86	85	84	83	82	81	80	79							
ANLP																					00			97			94			92			88			83			83			83						
CL	19	18	17	16	15	14	13	12	11	10	09	08	07	06	05	04	03	02	01	00	99	98	97	96	95	94	93	92	91	90	89	88	87	86	85	84	83	82	81	80	78							
CoNLL	18	17	16	15	14	13	12	11	10											09	08	07	06	05	04	03	02	01	00	99	98	97																
EACL											09			06			03			02			01			00	99	97	95	93	91	89	87	85	83	83	83											
EMNLP	18	17	16	15	14	13	12	11	10											09	08	07	06	05	04	03	02	01	00	99	98	97	96															
NAACL	19	18	16	15	13	12	10											09	07	06	04	03	01	00																								
*SEMEVAL	19	18	17	16	15	14	13	12	10											07	04	01											98															
TACL	19	18	17	16	15	14	13																																									

Conference Rotation

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



Growth of ACL



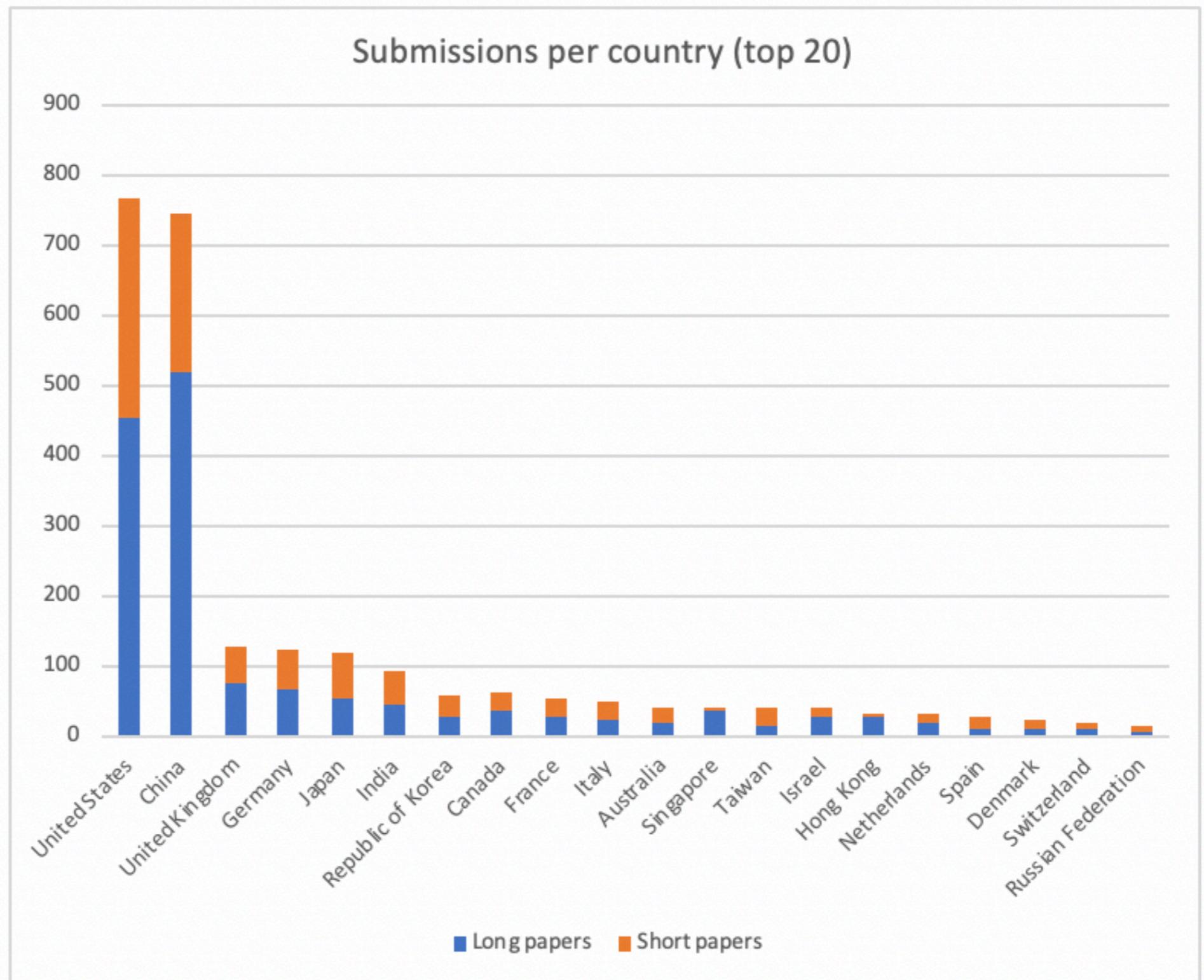
ACL'19 at A Glance

- ▶ The Annual Meeting of the Association for Computational Linguistics
- ▶ Duration:
 - tutorials (1 day)
 - main conference (3 days)
 - workshops (2 days)
- ▶ Attendance of 2000+ people
- ▶ Papers:
 - 2,905 submissions
 - 447 long papers and 213 short papers accepted
 - + ?? TACL papers
 - 151 oral and 151 poster presentations

Research Areas

	Area	Long submissions	Accepts	Accept rate (%)
1.	Applications	65	14	28.8
2.	Dialogue and Interactive Systems	126	38	30.2
3.	Discourse and Pragmatics	33	7	21.2
4.	Document Analysis	48	8	16.7
5.	Generation	96	32	33.3
6.	Information Extraction and Text Mining	155	37	23.9
7.	Linguistic Theories, Cognitive Modeling and Psycholinguistics	39	9	23.1
8.	Machine Learning	148	38	25.7
8.	Machine Translation	102	27	26.5
10.	Multidisciplinary and Area Chair COI	69	21	30.4
11.	Multilinguality	43	11	25.6
12.	Phonology Morphology and Word Segmentation	26	7	26.9
13.	Question Answering	99	32	32.3
14.	Resources and Evaluation	70	26	37.1
15.	Sentence-level semantics	69	14	20.3
15.	Sentiment Analysis and Argument Mining	91	24	26.4
17.	Social Media	51	14	27.5
18.	Summarization	48	11	22.9
19.	Tagging Chunking Syntax and Parsing	50	17	34.0
20.	Textual Inference and Other Areas of Semantics	44	16	36.4
21.	Vision Robotics Multimodal Grounding and Speech	56	20	35.7
22.	Word-level Semantics	78	20	25.6

By Country



By Country

Country or Region	All submissions			Long submissions			Short submissions		
	Sub.	Acc.	Rate (%)	Sub.	Acc.	Rate (%)	Sub.	Acc.	Rate (%)
Australia	46	11	23.9	22	4	18.2	24	7	29.2
Austria	5	0	0.0	2	0	0.0	3	0	0.0
Belgium	8	1	12.5	3	1	33.3	5	0	0.0
Brazil	11	0	0.0	6	0	0.0	5	0	0.0
Canada	74	16	21.6	44	12	27.3	30	4	13.3
Chile	2	0	0.0	2	0	0.0	0	0	N/A
China	817	155	19.0	567	118	20.8	250	37	14.8
Czech Republic	12	2	16.7	5	0	0.0	7	2	28.6
Denmark	25	4	16.0	11	1	9.1	14	3	21.4
Egypt	2	0	0.0	1	0	0.0	1	0	0.0
Estonia	2	0	0.0	2	0	0.0	0	0	N/A
Finland	6	0	0.0	2	0	0.0	4	0	0.0
France	60	11	18.3	32	4	12.5	28	7	25.0
Germany	136	39	28.7	73	26	35.6	63	13	20.6
Greece	7	4	57.1	1	1	100.0	6	3	50.0
Hong Kong	34	10	29.4	26	9	34.6	8	1	12.5
Hungary	7	1	14.3	3	1	33.3	4	0	0.0
India	107	18	16.8	54	16	29.6	53	2	3.8
Iran	3	0	0.0	2	0	0.0	1	0	0.0
Ireland	10	1	10.0	4	1	25.0	6	0	0.0
Israel	41	14	34.1	30	11	36.7	11	3	27.3
Italy	50	6	12.0	25	3	12.0	25	3	12.0
Japan	125	23	18.4	58	13	22.4	67	10	14.9
Luxembourg	2	0	0.0	2	0	0.0	0	0	N/A
Macau	5	1	20.0	3	1	33.3	2	0	0.0
Malta	2	0	0.0	0	0	N/A	2	0	0.0
Mexico	2	0	0.0	0	0	N/A	2	0	0.0
Netherlands	36	9	25.0	22	8	36.4	14	1	7.1
Norway	6	2	33.3	4	1	25.0	2	1	50.0
Pakistan	2	0	0.0	1	0	0.0	1	0	0.0
Peru	2	0	0.0	1	0	0.0	1	0	0.0
Poland	7	1	14.3	5	1	20.0	2	0	0.0
Portugal	8	3	37.5	4	2	50.0	4	1	25.0
Qatar	4	0	0.0	2	0	0.0	2	0	0.0
Republic of Korea	72	7	9.7	36	4	11.1	36	3	8.3
Romania	2	1	50.0	2	1	50.0	0	0	N/A
Russian Federation	14	4	28.6	7	2	28.6	7	2	28.6
Singapore	46	16	34.8	39	13	33.3	7	3	42.9
Slovakia	2	0	0.0	1	0	0.0	1	0	0.0
South Africa	2	1	50.0	1	0	0.0	1	1	100
Spain	29	6	20.7	12	1	8.3	17	5	29.4
Sri Lanka	5	0	0.0	1	0	0.0	4	0	0.0
Sweden	9	0	0.0	4	0	0.0	5	0	0.0
Switzerland	23	4	17.4	12	2	16.7	11	2	18.2
Taiwan	46	6	13.0	18	3	16.7	28	3	10.7
Thailand	2	0	0.0	1	0	0.0	1	0	0.0
Turkey	7	0	0.0	3	0	0.0	4	0	0.0
United Arab Emirates	4	2	50.0	1	1	100.0	3	1	33.3
United Kingdom	138	41	29.7	84	30	35.7	54	11	20.4
United States	820	236	28.8	485	154	31.8	335	82	24.5
Others	18	2		12	0		6	3	
TOTAL	2905	660	22.7	1737	447	25.7	1168	213	18.2

THE WALL STREET JOURNAL.

U.S. Edition ▾ | August 29, 2019 | Print Edition | Video

[Subscribe](#) | [Sign In](#)**Try 2 months for \$1**[Home](#) [World](#) [U.S.](#) [Politics](#) [Economy](#) [Business](#) [Tech](#) [Markets](#) [Opinion](#) [Life & Arts](#) [Real Estate](#) [WSJ. Magazine](#)**BREAKING NEWS**

The Trump administration is moving to end regulation of the oil-and-gas industry's methane emissions



MuleSoft Your guide to API and service management [Free eBook](#)

SHARE

TEXT

[IDEAS](#) | [ESSAY](#)

How the U.S. Surrendered to China on Scientific Research

Washington could show its seriousness about key technologies like AI, gene editing and quantum computing by making big federal investments

By *Ezekiel Emanuel, Amy Gadsden and Scott Moore*

April 19, 2019 5:30 am ET



BEHAVIOR & SOCIETY

One Reason Young People Don't Go Into Science? We Don't Fail Well

A single project failure drives many students to switch to other majors

By Sara Whitlock, STAT on March 27, 2017



READ THIS NEXT

BEHAVIOR & SOCIETY

The Personality of Academic Majors

April 19, 2016 — Anna Vedel

MIND

The Need for Belonging in Math and Science

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.

Reading #1 is out

Due Sep 5

Social Media & Text Analytics

Syllabus

Twitter API Tutorial

Homework▼

High School Outreach

August 28,
2019
(Wednesday)

AI Seminar by Mounica
Maddela

- 4:00 -- 5:00pm, Dreese
480

[Multi-task Pairwise Neural Ranking for Hashtag Segmentation](#) by
Mounica Maddela, Wei Xu, Daniel Preoțiuc-Pietro (ACL 2019)

August 29,
2019

Twitter and Twitter API Tutorial

- Brief history of Twitter
- Key features of Twitter
- Hands-on instructions on obtaining Twitter data via APIs

[★ Twitter API Tutorial](#) by Wei Xu
[★ What is Twitter, a Social Network or a News Media?](#) by Kwak, Lee, Park and Moon (WWW 2010)
[How to Do Research with a Professor](#) by Jason Eisner
[How to Read a Technical Paper](#) by Jason Eisner

September
5, 2019

Language Identification and
Naïve Bayes [Reading 1 due]

- Language Identification
- Supervised Learning
- Classification
- Naïve Bayes Algorithm + feature selection (Information Gain)

[★ Cross-domain Feature Selection for Language Identification](#) by
Lui, Baldwin (IJCNLP 2011)

[Evaluating language identification performance](#) by Mitja T @tm
[langid.py: An Off-the-shelf Language Identification Tool](#) by Lui, Baldwin (ACL 2012)
[6 Easy Steps to Learn Naïve Bayes Algorithm](#) by Sunil Ray
[Text Classification using Naive Bayes](#) by Hiroshi Shimodaira

In-class Presentation

5539 Presentations (2019AU)  

File Edit View Insert Format Data Tools Add-ons Help

    | 100%  .0  .00  123  | Arial  10  |             | ...

 Date

	A	B	C	D	
1	Date	Name of NLP Researcher/Paper, or Social Media Platform/Dataset	Student Presentation Group #1	Student Presentation Group #2	
2	8/22/2019	1st class - no student presentation	-----	-----	
3	8/29/2019				
4	9/5/2019		Emin & Njoki	Chao Jiang & Bohan Zhang	
5	9/12/2019		Yukun Feng & Sam Lin		
6	9/19/2019		Jack Dubbs & David Van Drei		
7	9/26/2019				
8	10/3/2019		Andrew Jivoin, Andrew Everman		
9	10/10/2019	Autumn Break	-----	-----	
10	10/17/2019		Andrew Davis, Neel Mansukhani	Qi Song, looking for teammate.	
11	10/24/2019		Colin Voisard, JP Dahms	Fengze Wu, Anybody who wants to join	
12	10/31/2019		Hanwei Peng, Zhengqi Dong	Umar Jara, Faris Rehman	
13	11/7/2019				
14	11/14/2019		Neil, Max, John		
15	11/21/2019			Wenjie Bai,	
16	11/28/2019	Thanksgiving	-----	-----	
17	12/5/2019	12/4 is the last day of classes	-----	-----	
18					
19					
20	It would be great if some one can present: (1) Gnip/Datashift - two social media related companies; https://www.programmableweb.com/news/how-datasift-survived-twitters-merciless-business-behavior-api-economy/native-case-study/2018/12/11 ; (2) the MTNT dataset: https://www.cs.cmu.edu/~pmichel1/mtnt/ ; (3) Google Shoelace and Google+ ; (4) and this research paper on online trolling behavior: https://files.clr3.com/papers/2017_anyone.pdf				
21					
22					