“Crash and Burn”

•••

Categorizing “disaster” tweets using Unsupervised Learning models

Matthew Kwee

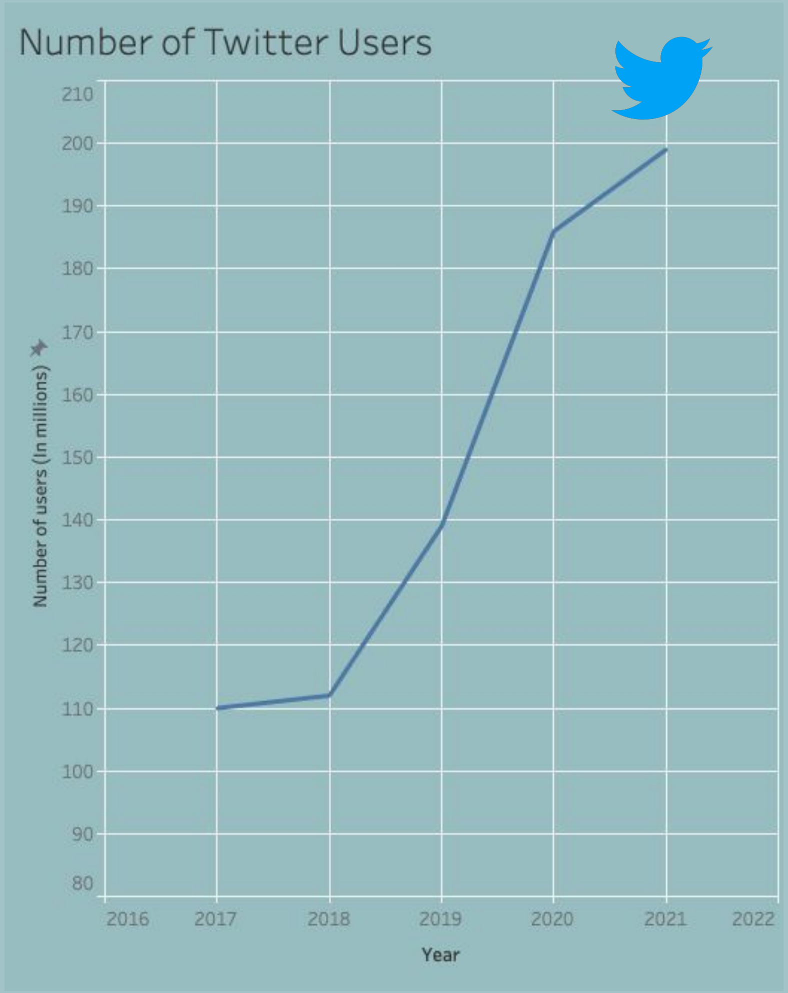
12 November 2021

The Ocean of Twitter

500+ million tweets/day

200+ million users

Countless categories that can be modeled

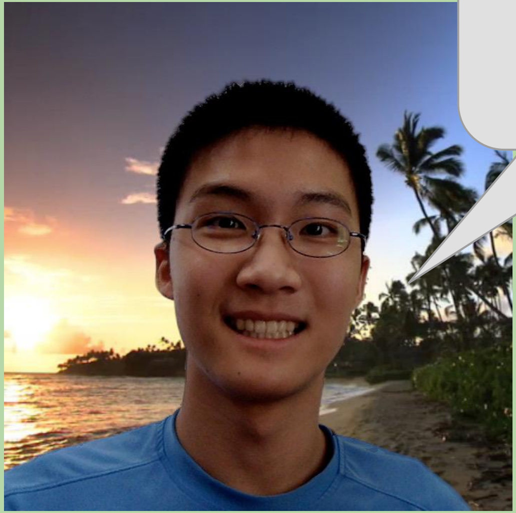


Source: https://www.omnicoreagency.com/twitter-statistics/

Objective:

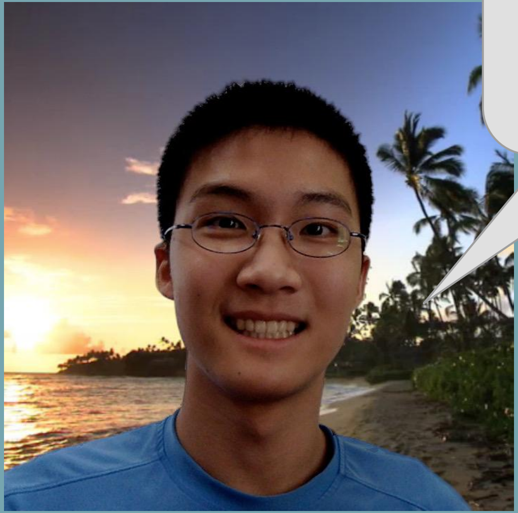
Create an unsupervised model to categorize “disaster” labeled tweets into topics

Matt when he submitted the proposal

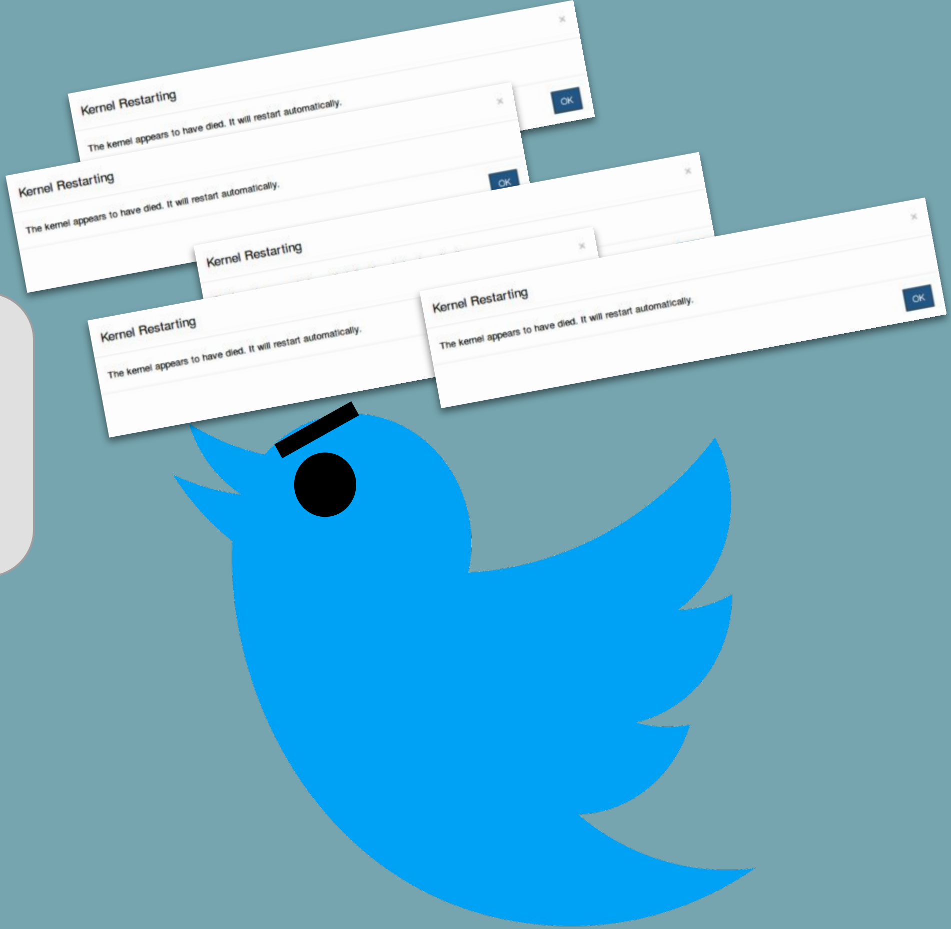


"Disasters" is somewhat specific, so it should be pretty easy to model

Matt while working on his project



Uh oh



Kernel Restarting

The kernel appears to have died. It will restart automatically.

OK

Kernel Restarting

The kernel appears to have died. It will restart automatically.

OK

Kernel Restarting

Kernel Restarting

The kernel appears to have died. It will restart automatically.

Kernel Restarting

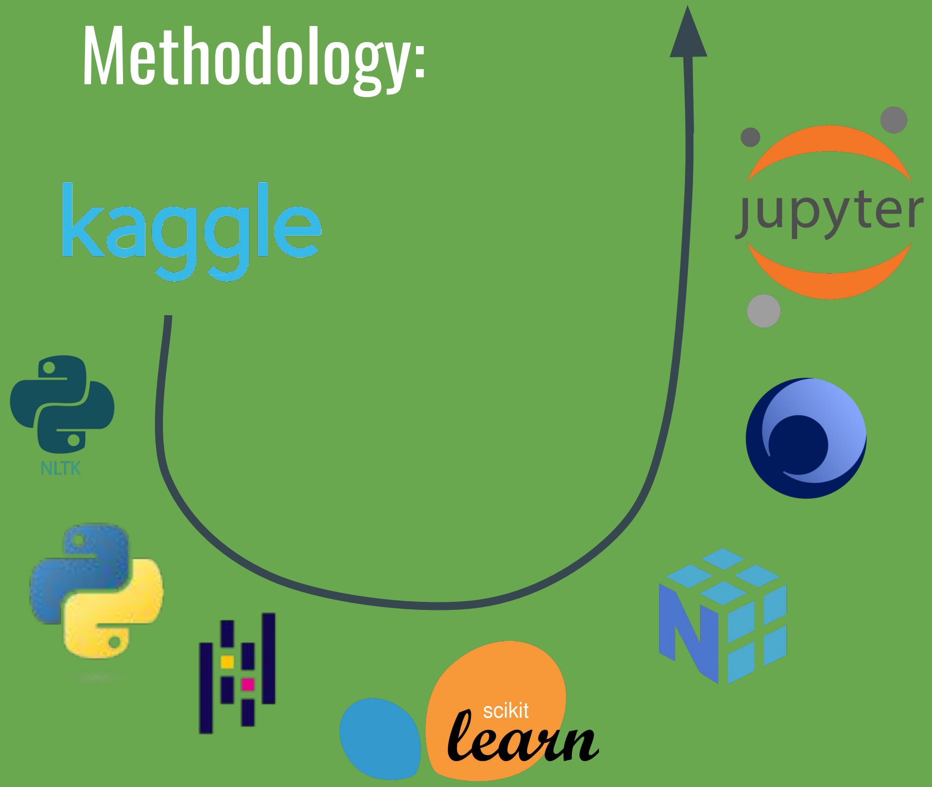
The kernel appears to have died. It will restart automatically.

OK

Corpus

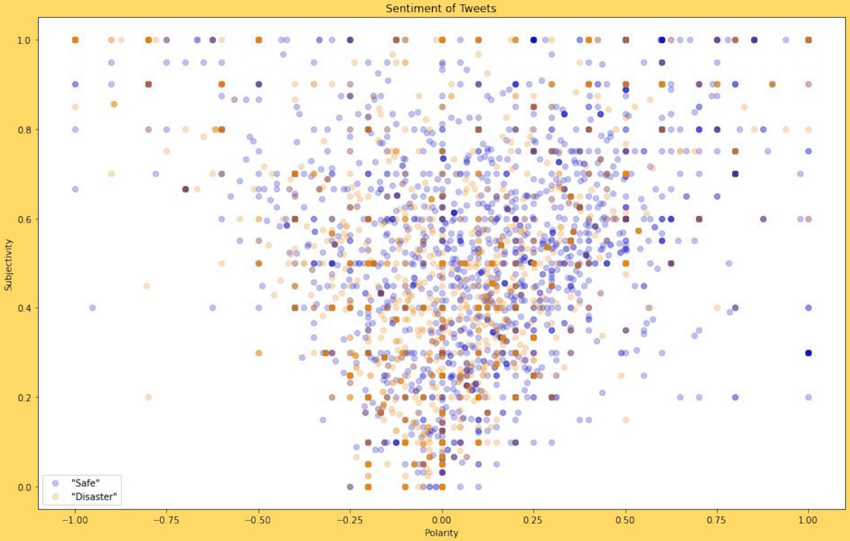
- 7,613 tweets
- 4,342 “safe”
- 3,271 “disaster”

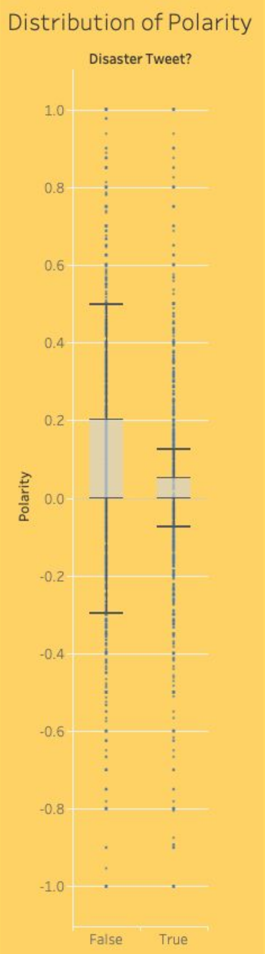
Methodology:

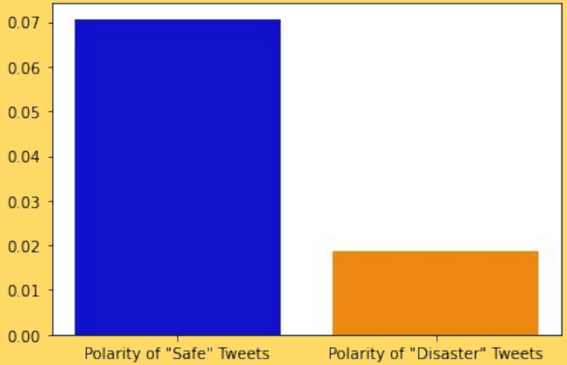


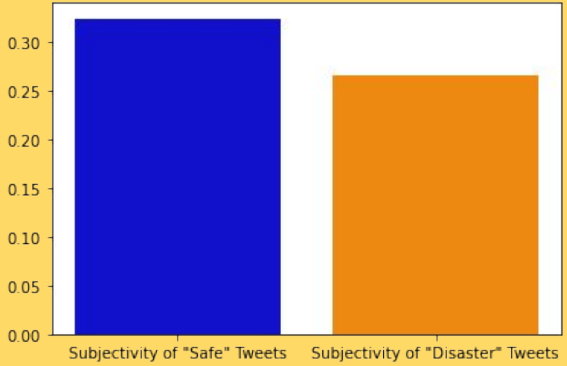
https://www.kaggle.com/c/nlp-getting-started

Preliminary Findings









Preprocessing Techniques

3,271 “disaster” tweets

Tokenization/Lemmatization

Ignoring words shorter than 3 letters

Using bigrams as tokens

Left out tweets with fewer than 4 usable tokens

1,866 tweets remaining

NLTK

Mislabeled Tweets

Not all disaster-labeled tweets are actually about disaster.

Data will be selected more rigorously in future versions.

??????

Probably randomly generated

“4 equipment ego break upon dig your family internet hoke excepting versus a sinking term: dfLJEV”

“@[redacted] haha so would you say its so hot your [redacted] are burning off????”

"@[redacted] @[redacted] remember when u were up like 4-0 and blew it in one game? U probs don't because it was before the kings won the cup"

Model

LDA Topic Modeling

K-Means Clustering

Max Frequency = 7.5%

Min Frequency = 0.5%

>>> 307 Features

8 topics

Topics - most common words

'Accidents': help, top, collide, near, keep, person	112	(Mostly vehicular collisions)
'Fire': burn, see, wild, wild fire, school, car, high	235	(Wildfires and buildings burning)
'Injury': wound, blood, movie, fatal, building	208	("movie"? Mislabeled tweets are blurring topics)
'Police': police, officer, home, collapse, area, line	249	(Police are often involved in urban disasters)
'Storms/Flooding': wreckage, injury, hiroshima, nuclear, damage, weapon	198	(Japan-related overlap)
'Terrorism': man, attack, bomb, tragedy, city, fall	302	(Somewhat straightforward)
'War': war, emergency, kill, talk, service, survivor	131	(Fairly small cluster - well defined)
'Misc': Everything else	432	
Total:	1866	

Using the model

The model can be used to categorize any text into the eight main clusters.

It can be easily accessed from a secondary notebook in the GitHub repository.

“George Njenga the hero saved his burning friend from a razing wildfire”

Model results:

Fire	56.3%
Injury	41.9%
Misc.	0.6%
Police	0.4%
Accidents	0.2%
Storms/Flooding	0.2%
War	0.2%
Terrorism	0.2%

Future Work

- More diverse data, without unrelated topics
- Seed topic weights to reduce topic overlap
- Explore other models
- Capture model semantics to improve coherence

> “going to redo my nails and watch behind the scenes of desolation of smaug ayyy”

> "Next May I'll be free...from school from obligations like family.... Best of all that damn curfew..."

> “OMG?? Didnt expect Drag Me Down to be the first song Pandora played OMG I SCREAMED SO LOUD My coworker is scared”

> “@[redacted] @[redacted] you can keep this please!!!!! Arachys [2 Pieces] - Now deals 4000% weapon damage (up from 2500%)”

Questions?