“Crash and Burn”

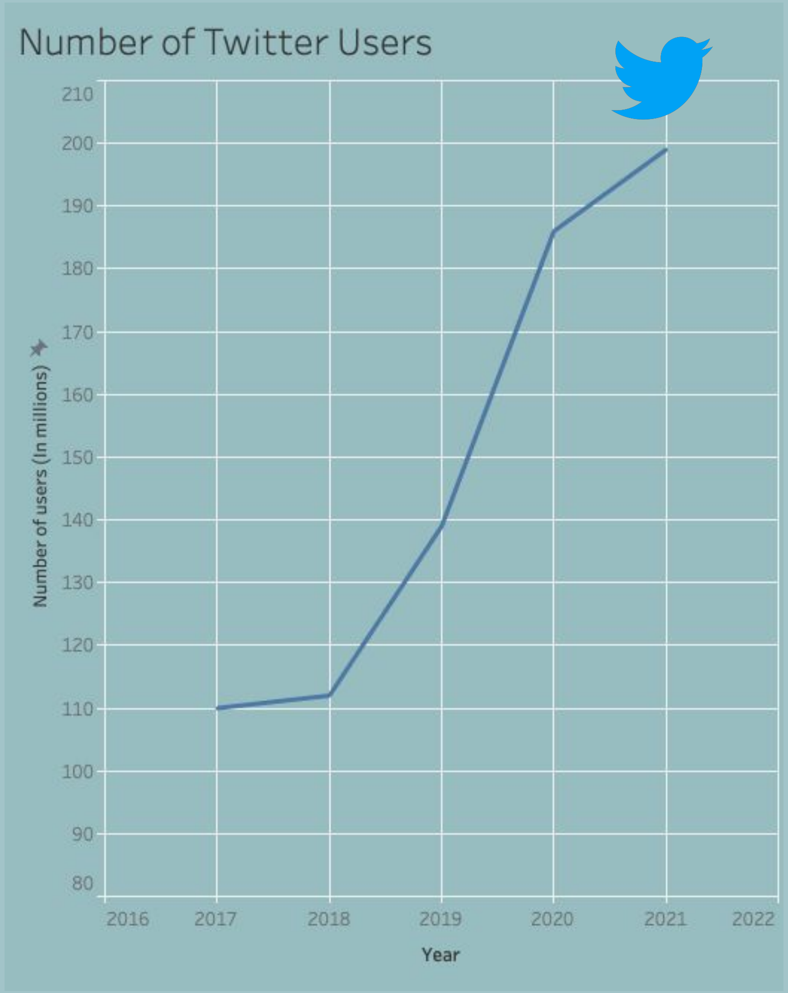
Categorizing “disaster” tweets using Unsupervised Learning models

The Ocean of Twitter

500+ million tweets/day

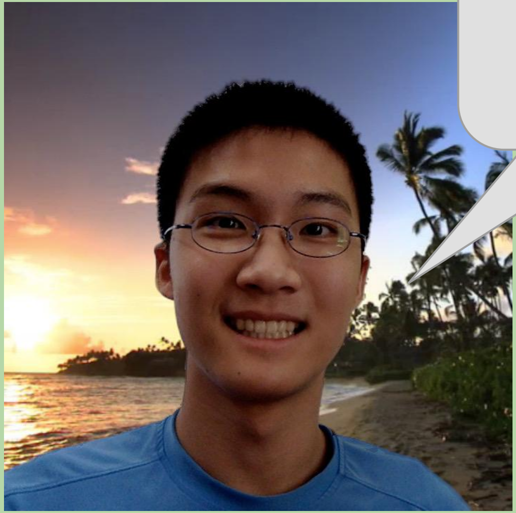
200+ million users

Countless categories that can be modeled



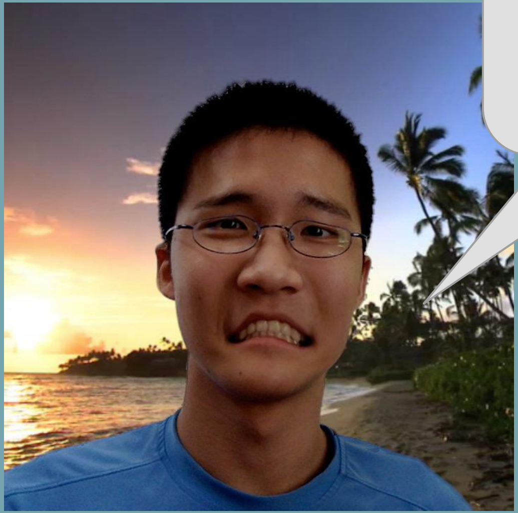
Source: https://www.omnicoreagency.com/twitter-statistics/

Matt when he submitted the proposal:

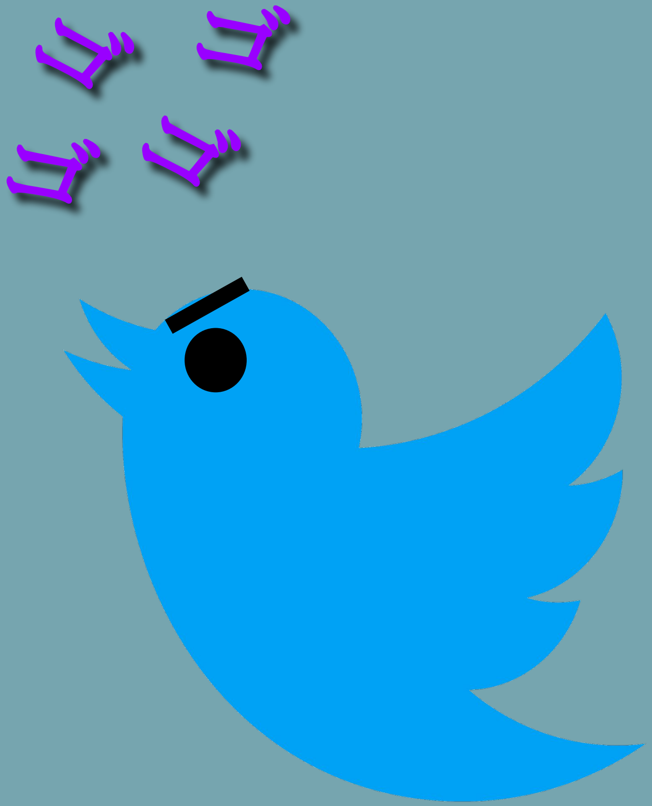


"Disasters" is somewhat specific, so it should be pretty easy to model

Matt on Friday Morning:



Uh oh

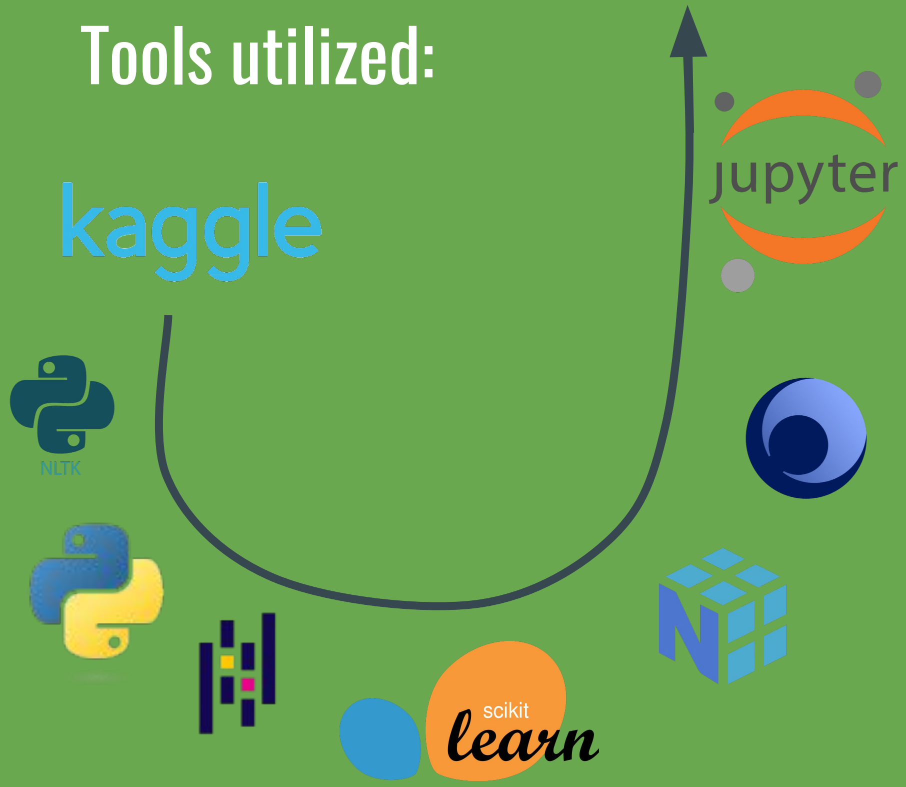


Objective: Create an unsupervised model to categorize tweets into topics

Data

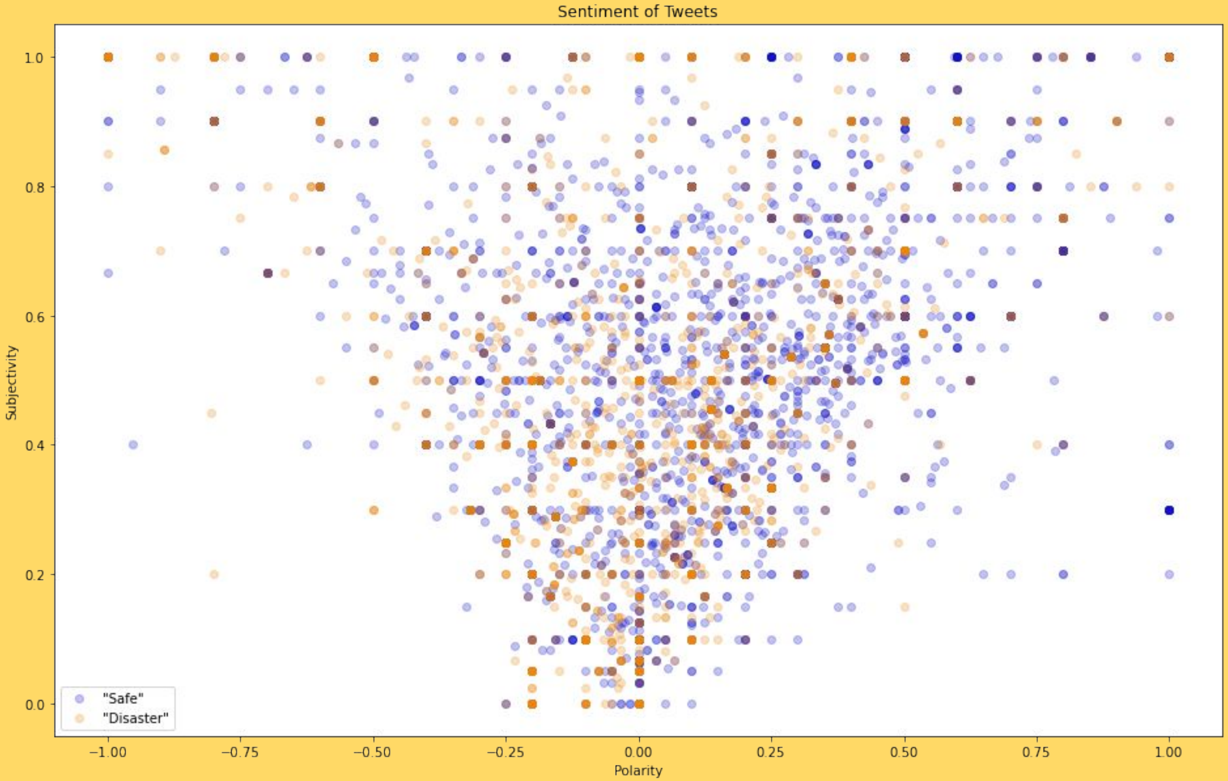
- 7,613 tweets
- 4,342 “safe”
- 3,271 “disaster”

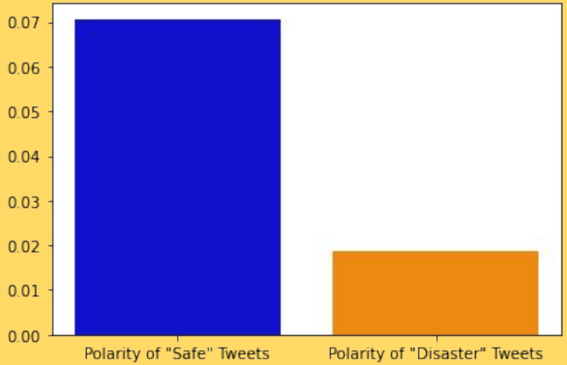
Tools utilized:

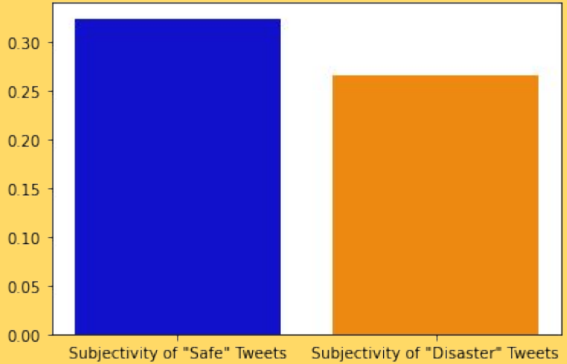


https://www.kaggle.com/c/nlp-getting-started

Preliminary Findings







Preprocessing Techniques

Default Tokenizer

Lemmatization

Cleaning non-Unicode characters

Using bigrams as tokens

NLTK

Model

LDA Topic Modeling

K-Means Clustering

Max Frequency = 7.5%

Min Frequency = 0.5%

8 topics

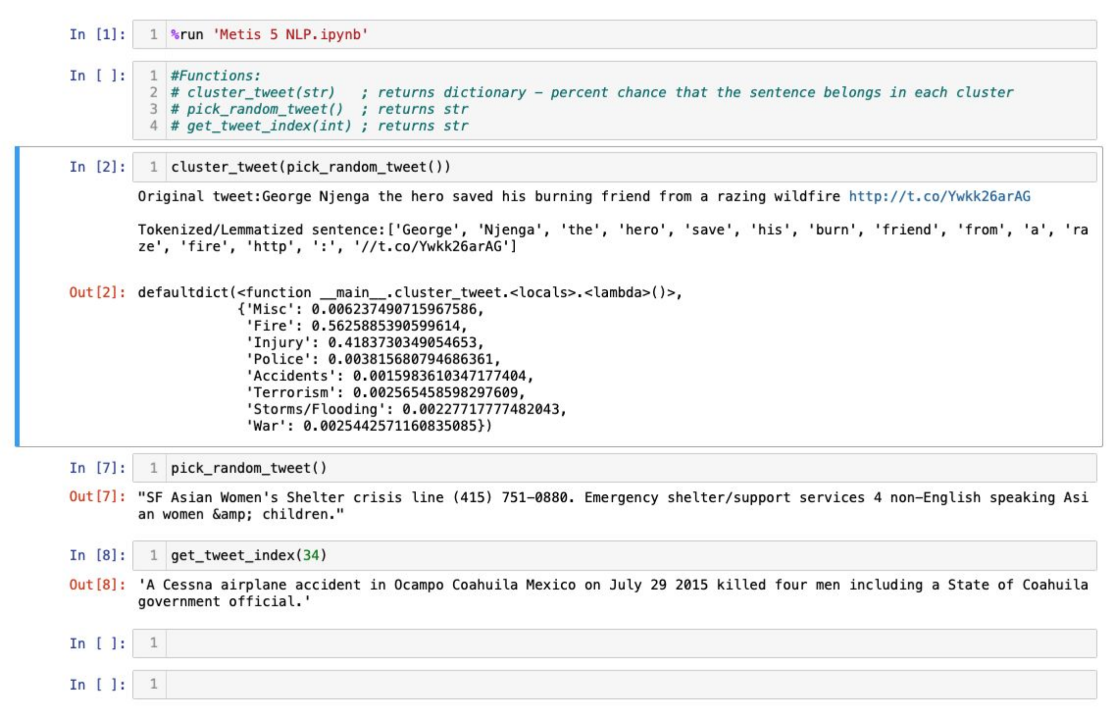
Topics - most common words

'Accidents': help, top, near, keep, collide, person	112	
'Fire': burn, see, wild, wild fire, school, car, high	235	
'Injury': wound, blood, movie, fatal, building	208	(“movie”? Mislabeled tweets contaminating)
'Police': police, officer, home, collapse, area, line	249	
'Storms/Flooding': wreckage, weapon, injury, hiroshima, nuclear, damage	198	(Japan-related overlap)
'Terrorism': man, attack, bomb, tragedy, city, fall	302	
'War': war, emergency, kill, talk, loud, service	131	
'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.



```
In [1]:  %run 'Metis 5 NLP.ipynb'

In [ ]:  #Functions:
         # cluster_tweet(str)    ; returns dictionary - percent chance that the sentence belongs in each cluster
         # pick_random_tweet()   ; returns str
         # get_tweet_index(int)  ; returns str

In [2]:  cluster_tweet(pick_random_tweet())

         Original tweet:George Njenga the hero saved his burning friend from a razing wildfire http://t.co/Ywkk26arAG

         Tokenized/Lemmatized sentence:['George', 'Njenga', 'the', 'hero', 'save', 'his', 'burn', 'friend', 'from', 'a', 'ra
         ze', 'fire', 'http', ':', '//t.co/Ywkk26arAG']


Out[2]:  defaultdict(<function __main__.cluster_tweet.<locals>.<lambda>()>,
                     {'Misc': 0.006237490715967586,
                      'Fire': 0.5625885390599614,
                      'Injury': 0.4183730349054653,
                      'Police': 0.003815680794686361,
                      'Accidents': 0.0015983610347177404,
                      'Terrorism': 0.002565458598297609,
                      'Storms/Flooding': 0.00227717777482043,
                      'War': 0.0025442571160835085})

In [7]:  pick_random_tweet()

Out[7]:  "SF Asian Women's Shelter crisis line (415) 751-0880. Emergency shelter/support services 4 non-English speaking Asi
         an women &amp; children."

In [8]:  get_tweet_index(34)

Out[8]:  'A Cessna airplane accident in Ocampo Coahuila Mexico on July 29 2015 killed four men including a State of Coahuila
         government official.'

In [ ]:  

In [ ]:  
```

Future Work

- More data, without unrelated topics
- Seed topic weights to reduce topic overlap
- Explore other models

“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”

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

Questions?