NLP
DISASTER TWEETS

A binary classification problem

By Cristina Sahoo

PROBLEM STATEMENT

Kaggle: In this competition, you're challenged *to build a machine learning model that predicts which Tweets are about real disasters and which one's aren't*. You'll have access to a dataset of 10,000 tweets that were hand classified.

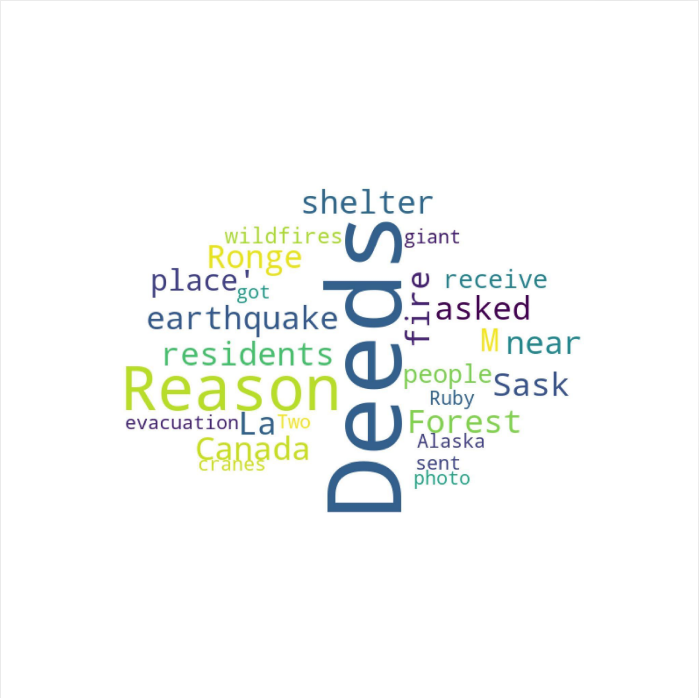
WHY IS THIS IMPORTANT?

Twitter has become an important communication channel in times of emergency. The ubiquitousness of smartphones enables people to announce an emergency they're observing in real-time. Because of this, more agencies are interested in programatically monitoring Twitter (i.e. disaster relief organizations and news agencies).

DATA DICTIONARY

Feature	Type	Description
id	int64	a unique identifier for each tweet
text	object	the text of the tweet
location	object	the location the tweet was sent from (may be blank)
keyword	object	a particular keyword from the tweet (may be blank)
target	int64	in train.csv only, this denotes whether a tweet is about a real disaster (1) or not (0)

WORD CLOUDS


shelter
wildfires
giant
Ronge
place'
got
earthquake
residents
Reason
evacuation
La
Two
Canada
cranes
Deeds
fire
receive
asked
M
near
people
Sask
Ruby
Forest
Alaska
sent
photo


fast
fruits
Cramer
wrecked
lovely
man
Iger
Li
Disney
explode
ready
engineshed
goooooooaaaaaal
British
words
love
car
Great
boxes
Summer
s
atmosphere

MODELS

Models used:

- Dense Neural Network with three layers, 16 units each, activation function 'relu', and output function 'sigmoid'; optimizer 'rmsprop', loss function 'binary crossentropy', and metric 'accuracy'. Accuracy 79%.
- Multinomial Naive Bayes. Accuracy 80%.