CAPSTONE PROJECT PROPOSAL

PABLO BACHO

Domain Background

After finishing this program, I would like to tackle some Kaggle competitions to practice and keep learning. Thus, I picked one of Kaggle's "Getting Started" competitions on Natural Language Processing for my capstone project.

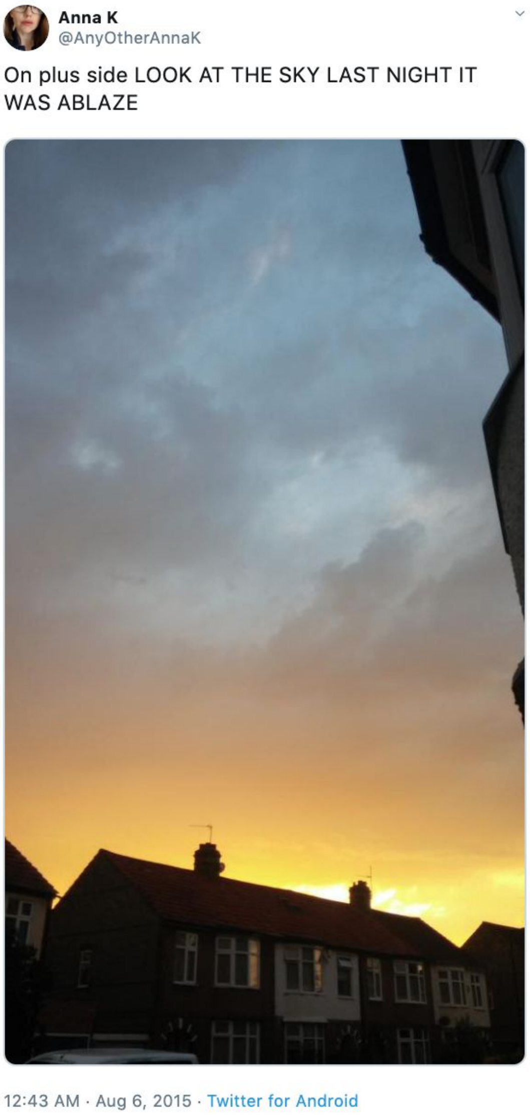
The competition is **Real or Not? NLP with Disaster Tweets**, and can be found on Kaggle's website at https://www.kaggle.com/c/nlp-getting-started/overview/evaluation

Problem Statement

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).

But, it's not always clear whether a person's words are actually announcing a disaster. Look at the example in the picture. The author explicitly uses the word "ABLAZE" but means it metaphorically. This is clear to a human right away, especially with the visual aid. But it's less clear to a machine.

The goal of this project is to build a machine learning model that predicts which Tweets are about real disasters and which ones aren't.



Anna K
@AnyOtherAnnaK

On plus side LOOK AT THE SKY LAST NIGHT IT WAS ABLAZE

12:43 AM · Aug 6, 2015 · Twitter for Android

Datasets and Inputs

The dataset is "Disasters on social media", collected by the company Figure Eight and publicly available to download on their website:

https://www.figure-eight.com/data-for-everyone/

It comprises over 10,000 tweets that were hand classified from searches like "ablaze", "quarantine" and "pandemonium" and then noted whether they related to a catastrophic event or not.

Solution Statement

Using a classification algorithm, the model to be built will tell apart tweets about disasters and others. During the program, I have come to learn how Amazon Sagemaker can be a good tool to solve this type of problem.

Benchmark Model

There is a leaderboard on Kaggle that can be used to check how my solution is performing against my peers'. The dataset is labeled, and the results can be checked against a subset of tweets.

Evaluation Metrics

The performance of the solution can be evaluated using sklearn's F-score (or F1) method. The F1 score can be interpreted as a weighted average of the precision and recall, where an F1 score reaches its best value at 1 and worst score at 0. The relative contribution of precision and recall to the F1 score are equal. The formula for the F1 score is:

$$F1 = 2 * (precision * recall) / (precision + recall)$$

Project Design

The dataset must be analyzed and pre-processed before working with it.

I will extract different features from the dataset such as different n-grams, and to do that stop words and punctuation marks will be removed.

I might explore other features such as the number of exclamation marks to see if they make a difference, or the use of uppercase letters.

As of right now, I have not decided which of Sagemaker's algorithms I will use to build this model.