Disaster Tweets

Milestone 3

Taylor Anderson

INTRODUCTION

Twitter is a place people go to talk about things that are going on in their life. If an emergency or disaster happens, Twitter will often be the first place that they go to comment on it. It allows people to report that emergency in real time. There are many reasons that it would be useful to be able to quickly identify and filter out tweets that relate to a disaster. We could receive immediate information from primary sources. The tweets could include pictures, videos, or descriptions of what happened. After the fact, we could reach out to twitter users that posted about the disaster for follow-up information. The list goes on and on. We will be using Natural Language Processing techniques to try and identify these tweets that are related to a disaster.

BACKGROUND

The data comes from a Kaggle competition (<https://www.kaggle.com/c/nlp-getting-started/data>) and was originally created by the machine intelligence company, Figure Eight. The data contains a training set with 10,000 tweets that were hand classified and a test set with 10,000 more tweets that are not labeled.

The goal of the competition was to identify which tweets are about real disasters and which one’s are not. This is the same goal of our project. We have analyzed this dataset by determining the structure of the tweets, cleaning them to account for special symbols, stopwords, and any other content that we believe might misconstrue the context of the tweet, and tokenizing the phrases to obtain the meaning. Our research questions include:

* Can we tell the difference between a tweet that involves a natural disaster, and one that is a figure of speech?
* Which words or phrases are helpful in understanding the context?
* Which words or phrases are detrimental to understanding the context?
* Do special symbols contribute to the meaning of the sentence, or steer the meaning off course?
* Does the length of a tweet have any correlation to its meaning?
* Does punctuation (‘!’ vs ‘.’) make a difference?

DATA CLEANING

We started by inspecting the data to see what we were working with. The dataset has the following fields:

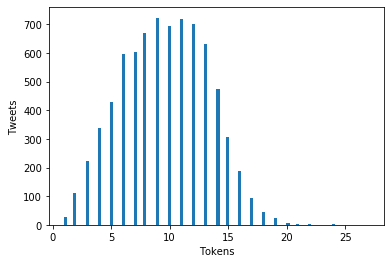
* id: a unique identifier for reach tweet
* text: the text of the tweet
* location: the location the tweet was sent from (may be blank)
* keyword: a particular keyword from the tweet (may be blank)
* target: in train.csv only, this denotes whether a tweet is about a real disaster (1) or not (0)

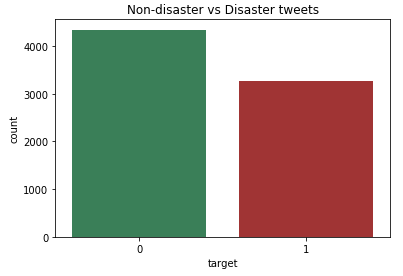


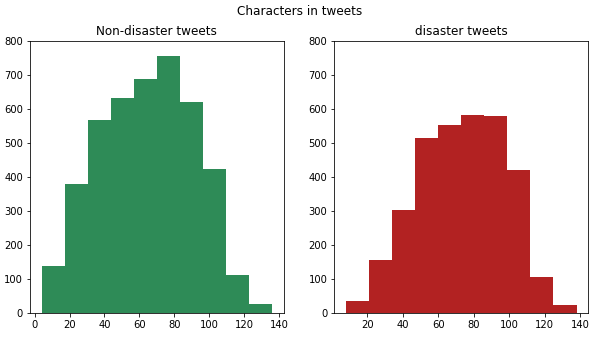
We removed the id, location, and keyword features because we wanted to focus on only the text of the tweet.

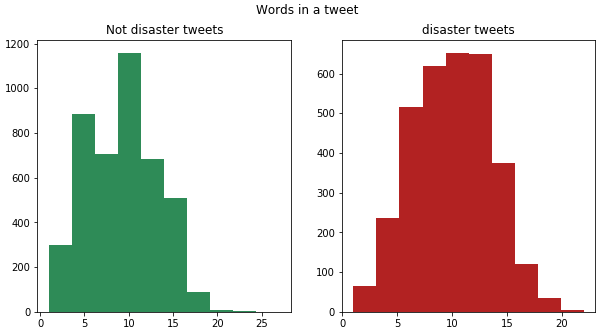
Next, we had to clean the tweets. To do this, we created a function that ran the text through five different functions which cleaned the data in different ways. The first function used BeautifulSoup to remove html language that may exist in the tweets, the second function removes stopwords and punctuation, the third removes brackets, the fourth removes URLs, and the final function removes hashtags. We apply each of these functions to the data which then returns the cleaned text which we add as a new column to the dataset.

EDA (Visualizations)









CONCLUSIONS

ACKNOWLEDGEMENTS

We are indebted to the communities behind the multiple open-source software packages and statistical reporting on which we depend.

REFERENCES