NLP with Disaster Tweets

Name: Jonathan Lawrence
Semester: Spring 2020
https://github.com/Jonathan813/NLP---Disaster-Tweets

Which Domain?

What domain is this data going to come from? Please list 10 references (with a brief annotation) to use to make sense of what you're doing with these data.

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. The data comes from a Kaggle competition (https://www.kaggle.com/c/nlp-getting-started/data). 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.

References:

Shaikh, R. (2018, October 24). Gentle Start to Natural Language Processing using Python. Retrieved from https://towardsdatascience.com/gentle-start-to-natural-language-processing-using-python-6e46c07addf3

Bansal, S. (2019, September 3). Ultimate Guide to Understand & Implement Natural Language Processing. Retrieved from https://www.analyticsvidhya.com/blog/2017/01/ultimate-guide-to-understand-implement-natural-language-processing-codes-in-python/

5 Heroic Python NLP Libraries. (2018, February 8). Retrieved from https://elitedatascience.com/python-nlp-libraries

Tseng, G. (2018, July 21). Summarizing Tweets in a Disaster. Retrieved from https://towardsdatascience.com/summarizing-tweets-in-a-disaster-e6b355a41732

Kumar, V. (2020, February 24). Real or Not? NLP with Disaster Tweets (A Data science Capstone Project). Retrieved from https://medium.com/real-or-not-nlp-with-disaster-tweets/real-or-not-nlp-with-disaster-tweets-a-data-science-capstone-project-fafa6c35c16f

Martinez, V. R. (2019, May 15). Identifying disaster-related tweets using deep learning and natural language processing with Fast... Retrieved from https://medium.com/datadriveninvestor/identifying-disaster-related-tweets-using-deep-learning-and-natural-language-processing-with-fast-e0dfb790b57a

Stowe, K., Paul, M., Palmer, M., Palen, L., & Anderson, K. (n.d.). Identifying and Categorizing Disaster Related Tweets. Retrieved from https://cmci.colorado.edu/~mpaul/files/socialnlp16_disasters.pdf

Geitgey, A. (2019, September 30). Natural Language Processing is Fun! Retrieved from https://medium.com/@ageitgey/natural-language-processing-is-fun-9a0bff37854e

Ma, G. (n.d.). Tweets Classification with BERT in the Field of Disaster Management. Retrieved from https://pythonprogramming.net/tokenizing-words-sentences-nltk-tutorial/

Hernandez-Suarez, A., Sanchez-Perez, G., Toscano-Medina, K., Perez-Meana, H., Portillo-Portillo, J., And Luis, V. S., & Javier García Villalba, L. (2019, April 11). Using Twitter Data to Monitor Natural Disaster Social Dynamics: A Recurrent Neural Network Approach with Word Embeddings and Kernel Density Estimation. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6484392/

Which Data?

What is the dataset you'll be examining? Please provide a codebook if there is one or a link to the dataset as well as a detailed description.

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)

Research Questions? Benefits? Why analyze these data?

How are you proposing to analyze this dataset? This is about your approach. Here, you'll be proposing your research questions as well as justifications for why you'd offer these data in this way.

We intend to analyze 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?

We'll be looking at data from approximately 20,000 tweets in order to study a wide variety of examples.

What Method?

What methods will you be using? What will those methods provide in terms of analysis? How is this useful?

We will start by cleaning the data. There are quite a few null values and features that we won't be dealing with. Then we will clean the tweets themselves by removing stop words, punctuation, URL's, hashtags, etc. Once the text is clean, we can tokenize it.

Next we will do some exploratory data analysis. This will help us learn a little more about the tweets that we are dealing with. We can see if there are any clear differences between disaster and non-disaster tweets such as their number of characters or words.

Lastly, we will do some modeling. We will need to vectorize the text so that it can be used by regression models. Then we can build a ridge regression model to determine the qualities of disaster tweets

compared to non-disaster tweets. Finally, we can use a test dataset to predict if the tweets are based on a disaster.

Potential Issues?

What challenges do you anticipate having? What could cause this project to go off schedule?

Natural language processing is something that we haven't done a lot of work within this course beyond organizing unstructured data. We are comfortable dealing with stopwords and tokenizing but when it comes to interpreting the tweets, we expect to have complications. Thankfully, there are many resources to help us move forward and try different methods.

Concluding Remarks

Tie it all together. Think of this section as your final report's abstract.

Context can be difficult to interpret in text-based communication. It puts you in a position where you can't pick up on a person's mood or their body language in order to determine their meaning. This is particularly difficult when trying to interpret phrases like "This day has been a total disaster!" Add to this problem the vagueness and incoherent nature of messages sent over Twitter, combined with hastags and special characters, and you end up with a very difficult challenge for any human. The purpose of this project is to find out if a computer, given enough data, can determine whether a Tweet is about a real disaster, or just a figure of speech.