Client Project

Mike Laccavole, Patrick McCaul, Kathryn Shupe

Data Dictionary

File Name	Description
raw_df	Unedited `raw` tweets pulled through Twitterscraper
df.csv	Cleaned tweets, ready for EDA and modeling
tweets_lemmatized.csv	Lemmatized Tweets
tweets_stemmed.csv	Stemmed Tweets
tweets_tokenized.csv	Tokenized Tweets
df_urgent.csv	`urgent` classified tweets with assigned coordinates
df_nonrguent.csv	`non_urgent` classified tweets with assigned coordinates
disaster_response_messages_training.csv	Data the `Word2Vec` classifier model was trained on

Twitter: Data Collection and Cleaning

Due to logistical constraints of obtaining Twitter Developer API, the Twitterscraper web-scraping tool was used to collect tweets. The benefit of this particular tool was the ability to use a customizable `query` for filtering tweets relevant to location and timeline. Due to the importance of location for our project, location was particularly difficult aspect to account for. Location of tweet is not a possible output in the query. Instead, location, as an input, is only used to search for tweets by city, area, or geo-coordinates.

Although all tweets have the same limitation on size, tweets can vary in a number of ways, and have a wide range of characters they can contain. In measuring statistically relevant textual information within a given tweet, we removed non-alphabetical characters, removed `url`'s, and stripped away all columns except for `text`. This process was automated using a custom function for further use.

EDA

In order to make our models, we had to further process the social media posts properly and conduct some exploratory data analysis. We created functions that tokenize all the documents within the corpus, that lemmatize the tweets, and that stems the tweets. All 3 were saved to csv files so we could conduct seperate model generation processes to see which method would be the most accurate. We also visualized the most commonly used words (both accounting and not accounting for stop words) to see which words were being used the most frequently in the relevant social media

posts.

Model 1: Preparation for Bag of Words

Countless tweets are posted during disaster situations, in order to understand and provide greater context for emergency responders, we decided to model using a bag of words approach to classify tweets are being urgent or non-urgent.

In order to create a bag of words for each vector (urgent and non-urgent) a multi-lingual disaster response message database from open source database website Appen was utilized. A logistic regression was run on this database using the language of the disaster message to classify if it was a direct call for help or not. The purpose of this classification was analyze the coefficients created through this language processing to better fit each bag of words vector with true disaster-related terms.

The second approach taken in creating the bag of words vectors was to apply industry knowledge to each term. According to FEMA's public assistance program guide, public assistance is divided into subcategories grouped by if the work is emergency work or permanent work. Terms were extracted from each of the various categories and applied to the appropriate bag of words vector.

Model 2: words2vec with Google News Words

An advanced Neural Network (words2vec) was applied to classify if the language in a tweet made a tweet more similar to the urgent bag of words vector or more similar to the non-urgent bag of words vector. The model was trained using a bin file of Google News Words and Phrases. This file contains about 100 billion words that are populated in 300-dimensional vectors of words and phrases.

This model allows us to compare similarities between language vectors to provide deeper understanding of our documents and corpus. Using this model, cosine similarity scores are extracted, these scores represent the measure of the angle between the two compared language vectors. For example, one vector was created for each tweet, one vector was created for the urgent bag of words, and one for the non-urgent bag of words. During analysis, a cosine similarity score is created for the angle between the tweet vector and the urgent bag of words vector as well as another cosine similarity score to measure the angle between that same tweet and the non-urgent bag of words vector.

Next, the two cosine similarity scores for each tweet are compared, and the tweet is then classified into the bag of words vector which it is most similar to. After modeling, a dataframe is populated with each tweet, it's two associated cosine similarity scores, and it's classification (1 for urgent, and 0 for non-urgent)

Creating Coordinates for Mapping

Because of the limitations on acquiring tweet location from either the Twitter API or Twitterscraper tool, we opted to randomly generate geo-coordinates for simulating the mapping of tweet locations (by geo-coordinate location) for our proof of concept.

A random cluster of coordinate positions was created around the coordinates for Paradise, Ca. Any cluster can be created by entering into the function the `number of coordinates to generate`, the `center around which to cluster`, and the `radius of the cluster`. The `coords_dict` dictionary contains several latitude / longitude coordinates wihtin the area of the Camp Fire wildfire, of which can be updated to reflect a clustering in another location.

ArcGIS is a powerful visualization tool that works well with `geopandas`. Once the coordinates were generated and assigned to the dataframes containing the `urgent` and `non_urgent` classified tweets, the respective dataframes were passed into a `GeoDataFrame`. This transformed the randomly generated coordinates into a geopandas `geometry` compatible for layering in ArcGIS.

Summary

The power of social media to connect and amplify messages can never be underestimated, especially in disaster situations. The ability to collect, process, classify by urgency and map tweets during situations can lead to more efficient and successful emergency response efforts.

Moving forward, twitter should set a realistic goal to provide relief organizations with geolocation data for tweets during disaster situations. This project and model were based upon the assumption that emergency responders would have access to geo-located tweets in disaster situations. The nexts steps would include applying industry knowledge through each step of the process, but crucially, while creating the bag of words vectors.

The ArcGIS platform provides a valuable tool for responders to visualize their own data, as well as applying data from the vast amount of organizations that are also responding and mapping their area of focus during disasters. For example, many city agencies working together to map traffic, fire, commodities, shelters, social media posts and other important resources during disaster situations. These kinds of partnerships are crucial in effective emergency management and utilizing ArcGIS provides a platform for each commodity to work independently and collectively.