From Text to Map: A Case Study of Twitter as a Tool for Disaster Response During Hurricane Harvey

Rachel Hausmann

Abstract

Twitter has become an instrumental source of news in emergencies where efficient access, dissemination of information, and immediate reactions are critical. Nevertheless, due to several challenges, the current fully-automated processing methods are not yet mature enough for deployment in real scenarios. One of those challenges is converting the needed aid expressed in tweets onto a map for emergency workers to act effectively. Unfortunately, only about 2% of tweets are geotagged by the user, posing challenges to the quantity of data available for quick mapping. This project contains an exploratory data analysis of 400,000 tweets and how they can be geoparsed before mapped in a geographic information system. The results highlight challenges for real-time disaster management. First, the place names mentioned are not strictly within the region of interest (ROI). Future work towards an automatic NLP/Geoparsing program of twitter data for disaster response should cluster the geoparsed place names associated with lat/longs and filter a specified distance in a circular manner from the centroid of the ROI latitude and longitude. For the sake of this project, Houston, TX is our ROI. Second, for effective disaster response a geoparsed place name or latitude and longitude should be accompanied by a disaster keyword or short description that could be fed into a supervised classification model to categorize each lat/long point with a needed response: e.g. “Needed Evacuation”, “Trapped inside home”, “Children need help”, “Fire”, the list goes on.

Design

This project was designed to explore: (1) An unsupervised machine learning algorithm to model the topics discussed in the dataset and (2) A geoparser tool to map the places mentioned in the dataset.

Data

Roughly 400,000 tweets with hurricane harvey hashtags were downloaded from Kaggle. The data was cleaned and thus reduced to 398,406 tweets. The word matrix contained 104,000 words by 398,406 users.

Algorithms

An unsupervised learning model was successfully built and tuned using SKlearn. Tweets were modeled using TF-IDF statistical analysis and an nMF dimensionality reduction method.

Tools

- Python pandas for EDA and data cleaning
- SKlearn was used for model building
- Spacy’s Name Entity Recognition (NER) was used for extracting place names out of corpus.
- ArcGIS Online was used for mapping