GENERAL ASSEMBLY – DATA SCIENCE IMMERSIVE PROGRAM

LEVERAGING SOCIAL MEDIA TO MAP DISASTERS

Team Members

Belén Sanchez

Ryan Stewart

Hakob Avjyan

Jonathan Slapnik

Date

October 2018

1. PROBLEM STATEMENT

When responding to disasters, it is critical to map and identify locations of survivors needing assistance. In recent history, social media has grown rapidly, with new platforms coming out every year. Now that a majority of people use social media, it has becoming increasingly helpful when natural disasters hit.

Social media can help identify isolated communities at risk, locations of survivors and areas where assistance team should be sent for search and rescue, levels of damage, where more information needs to be collected, and where resources should be allocated. People will tweet that themselves or others need help, they will post pictures and videos to Twitter, Facebook, Instagram, Snapchat, and YouTube showing the conditions that they are currently dealing with.

Since these platforms are updated every minute, leveraging the different platforms can be very helpful in figuring out where help needs to be sent next. When these disasters hit, there are areas that need to receive help that have not so far. It is critical that we are able to map and identify the locations of survivors that need aid.

2. GATHERING THE DATA

As is normal with web scraping, there are some limitations to what data we can actually acquire. One of our goals for this project is to create either a map of locations in need or create a list of latitudes and longitudes of those locations. Facebook and Instagram have very restrictive APIs due to recent security issues, so we were unable to use them to gather geolocations. Snapchat has an API that provides marketing services, but not for data analysis.

When it comes to Twitter, a very low percentage of people turn on their location feature. Twitter does come with geolocations; however, the location is set to where the account was set up, and not where the tweets are coming from. Twitter offers three different APIs to developers.

The first is the Standard API. This is free and the most basic API. It searches against a sampling of recent tweets published in the past seven days. The Standard API only allows us to gather one hundred tweets at a time, making it a painstakingly long process to collect enough tweets to use for a model. The second is the Premium API. This API can be free or paid, and it offers either the previous thirty days of tweets or access to tweets from as early as 2006. The Premium API is built on the reliability and full-fidelity of enterprise data APIs and it provides the opportunity to upgrade access as your business grows. The third API offered is the Enterprise API. This is a paid and managed API that gives access to either the last thirty days of tweets or access to tweets from as early as 2006. The Enterprise API provides full-fidelity data, direct account management support, and technical support to assist with integration strategy.

As a team we faced two limitations during this project. Firstly, the standard API access to twitter only allowed us to pull 100 tweets at a time and only provides access to tweets from the past 7 days. We worked under the assumption that an organization working for FEMA or FEMA itself could use our code through an Enterprise Level API access which would give them much greater access to twitter data.

A second limitation is with regard to the geo-data. While twitter collects location data about tweets, over 99.9% of users opt out of sharing that data with third parties. This means that the geolocation field will be 'Empty' for each tweet. We thought about using location tied to the users account, as opposed to the location from which they posted a tweet, but that location actually just represents the place where the account was created and was often more hurtful than helpful.

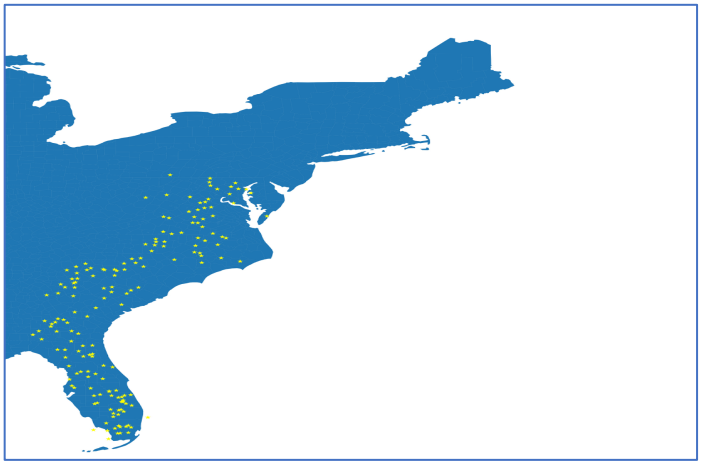
During our research we found that social media companies like Facebook have provided FEMA with access to restricted geolocation data during emergencies[1]. As an alternative for the future, FEMA might negotiate with Twitter to have access under similar terms.

[1] Techcrunch, 2017. "Facebook will share anonymized location data with disaster relief organizations". https://techcrunch.com/2017/06/07/facebook-will-share-anonymized-location-data-with-disaster-relief-organizations/

For the purposes of this project, we used a third party database that allowed us to pull almost 10,000 tweets related to the hashtag --> #HurricaneFlorence2018[2]. To solve the geo location limitation, we artificially simulated geolocation data for those tweets by creating a random distribution of longitudes and latitudes centered somewhere in the southeastern US. Theoretically, it would be real location data tied to the tweets that an emergency response organization is pulling.

In order to map the geolocation data, we decided to use GeoPandas, which is an opensource library created specifically to make it easy to work with geolocation data within Pandas.

Graph 1. Tweets reported during Huricane Florence 2018

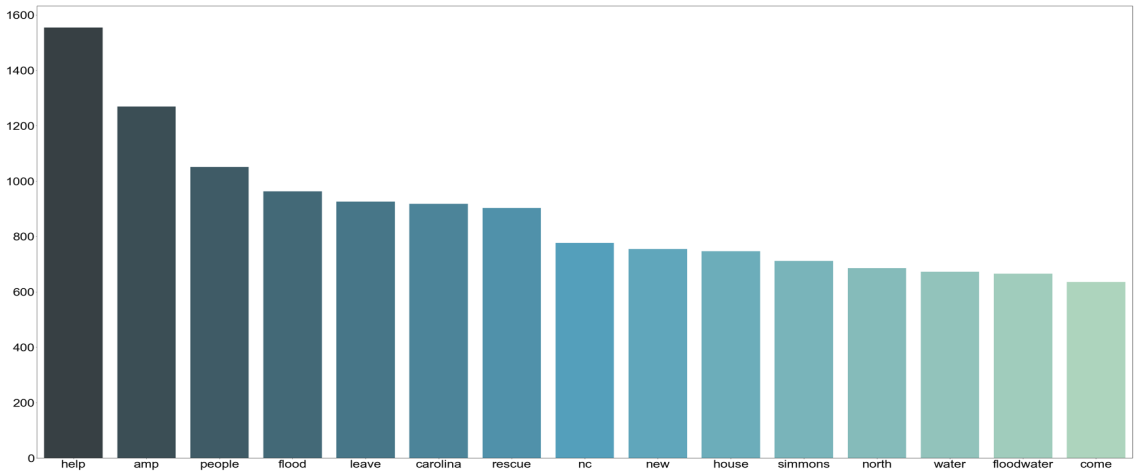


3. MODELING

There are two distinct elements to modeling a solution to this problem. One being the processing of the text, and the second being the actual mapping of the relevant tweets. To tackle the first, we used some common Natural Language Processing techniques (tokenizing, stemming and removing punctuation and stopwords) to extract the meaningful words from each tweet while ignoring the words that do not provide any additional analytical value. This allowed us to identify the most frequently words across all tweets.

[2] TAGS, 2018. https://tags.hawksey.info/

Graph 2. Most Common Words in Tweets



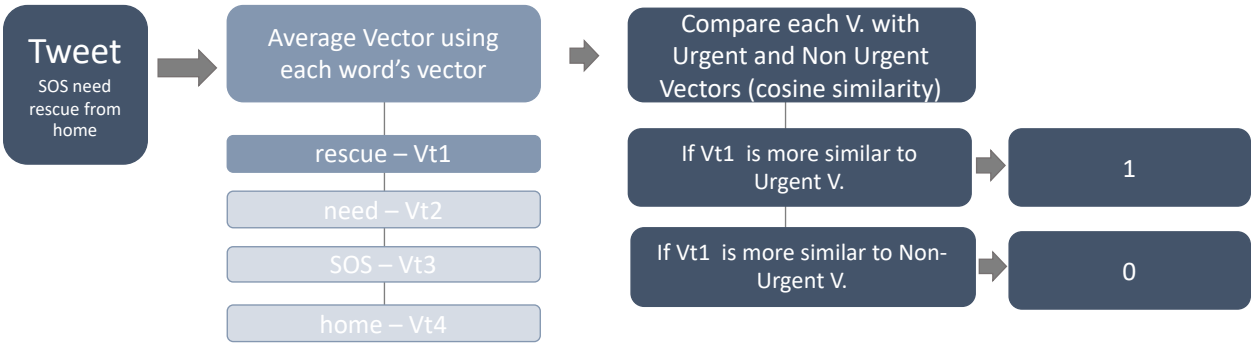
However, frequency of a commonly used word does not alone indicate that a tweet is important. We decided to create an unsupervised learning model that would allow us to classify our tweets among categories.

To define a framework that would allow us to make a relevant classification of our tweet, we used as a guideline the categories of emergency and permanent work defined by the Damage Assessment Operations Manual of FEMA[3]. To the emergency work, which is categorized as debris removal and emergency protective measures, we added relevant words that might be used by people in Twitter when reporting emergencies and renamed this bag of words as Urgent Response. The second bag of words took as a base the permanent work category, which is categorized as roads and bridges, water control facilities, buildings and equipment, utilities, and parks, recreation, and others. We added other relevant words that might be used by people using Twitter in this bag of words and renamed it as Non-Urgent Response.

[3] FEMA, 2016. "Damage Assessment Operations Manual"

In order to identify whether a tweet warranted action, we used a Word2Vec. model Word2Vec is a neural network model that takes words and converts them into vectors. The idea behind Word2Vec is that it takes something that the computer can not understand (i.e. human language) and turns them into something it can, in this case, vectors in 300-dimensional space. By taking the average vector of all of the words contained in a sentence, tweet, or any list of words, we can identify the "average vector" to determine the overall sentiment of the message, something we commonly call sentiment analysis.
Word2Vec can make highly accurate guesses about a word's meaning based on past appearances, and words that have more similar meaning usually have closer vector approximations. In order to train our model, we used 3Mil words from Google news4.

Graph 3. Word2Vec explained



We started the process by defining vectors that were associated with our urgent or non-urgent bag of words. Then we took our tweets and each tweet is broken down into its component words (tokens). Each of those words is assigned a vector based off on our trained Word2Vec model that was trained on our Google News vectorized words.

After each word was assigned with a vector, we calculated the average vector for each tweet. Using dot product, we compare the angle between the average vector of the tweet and the vectors of each of the bag of words that we created (Urgent Response and Non-

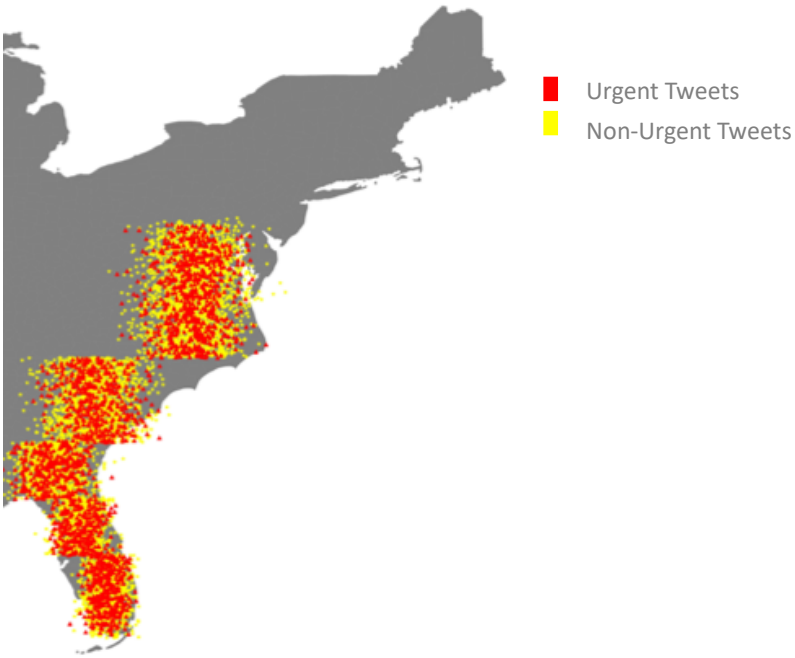
[4] Google, 2018. "Google News Vectors". https://groups.google.com/forum/#!topic/word2vec-toolkit/z0Aw5powUco

Urgent Response). We used cosine similarity to do this comparison. If the average vector of the tweet is closer to that of the urgent words, it is assigned a 1. If the average vector is closer to that of the non-urgent words, the tweet is assigned a 0.

4. RESULTS

Once we have filtered the tweets into the urgent and non-urgent categories, we map our tweets based on their location. The map below show our tweets classified as urgent and non-urgent tweets.

Graph 5. Urgent and Non-Urgent Tweets



5. CONCLUSIONS AND NEXT STEPS

The main take away of this project are:

- Accessing geo location in social media can be difficult, but not impossible. Public agencies like FEMA might have the political capital to negotiate better access to this type of information from tech companies in times of emergency.
- Word2vec and the use of cosine similarities between words can be used to classify your tweets or posts from any source in a meaningful way.

- As next steps, we would like to work with emergency experts to optimize our bag of words for hurricane response, but also to think about bag of words that might be applicable to the different type of disasters.
- It would also be important to impute tweets with real geo location to see the results of our model.